

Hierarchical Dirichlet Processes

Yee Whye Teh⁽¹⁾

ywteh@eecs.berkeley.edu

Matthew J. Beal⁽³⁾

beal@cs.utoronto.ca

Michael I. Jordan^(1,2)

jordan@eecs.berkeley.edu

David M. Blei⁽¹⁾

blei@eecs.berkeley.edu

⁽¹⁾Computer Science Division
University of California, Berkeley
Berkeley CA 94720-1776, USA

⁽²⁾Department of Statistics
University of California, Berkeley
Berkeley CA 94720-3860, USA

⁽³⁾Department of Computer Science
University of Toronto
Toronto M5S 3G4, Canada

March 6, 2004

Technical Report 653
Department of Statistics
University of California, Berkeley

Abstract

We consider problems involving groups of data, where each observation within a group is a draw from a mixture model, and where it is desirable to share mixture components both within and between groups. We assume that the number of mixture components is unknown a priori and is to be inferred from the data. In this setting it is natural to consider sets of Dirichlet processes, one for each group, where the well-known clustering property of the Dirichlet process provides a nonparametric prior for the number of mixture components within each group. Given our desire to tie the mixture models in the various groups, we consider a hierarchical model, specifically one in which the base measure for the child Dirichlet processes is itself distributed according to a Dirichlet process. Such a base measure being discrete, the child Dirichlet processes necessarily share atoms. Thus, as desired, the mixture models in the different groups necessarily share mixture components. We discuss representations of hierarchical Dirichlet processes in terms of a stick-breaking process, and a generalization of the Chinese restaurant process that we refer to as the “Chinese restaurant franchise.” We present Markov chain Monte Carlo algorithms for posterior inference in hierarchical Dirichlet process mixtures.

1 Introduction

Mixture models provide a flexible tool that links parametric and nonparametric statistics—the mixture components are generally taken from a parametric family, but the number of mixture components is allowed to grow without bound as the number of data points grows. This poses a challenging, perennial problem—how to choose the number of components? One approach to this problem reposes on a nonparametric prior known as the *Dirichlet process*. The Dirichlet process prior induces probability on the parameters of a distribution in such a way that particular values of the parameter tend to recur, with the number of distinct values growing slowly (logarithmically) with the number of draws. Using these distinct values to index distinct mixture components, this yields a mixture model with a random number of mixture components. Markov chain Monte Carlo algorithms are available to sample from the posterior distribution associated with the Dirichlet process prior, yielding a posterior distribution on the number of mixture components and on their parameters (Escobar and West, 1995, MacEachern and Müller, 1998, Neal, 1998).

In this paper we consider an elaboration of the classical mixture model setting in which the data are subdivided into groups, and in which the overall model consists of a set of mixture models, one for each group. We want the number of mixture components to be allowed to grow within each group, and we make use of the Dirichlet process to achieve this. Moreover, as will be clear from the examples that we consider, it is also desirable to allow mixture components to be shared between groups. Thus, we want to allow the possibility that data points in different groups come from the same mixture component. Thus we wish to achieve a “combinatorial” notion of the sharing of statistical strength between sets of mixture models.

The Dirichlet process is a random measure—a measure on measures (Ferguson, 1973). It is defined by considering partitions of the underlying sample space. Specializing to measures on the real line, let $(A_i)_{i=1}^r$ be a partition of \mathcal{X} . The distribution $\text{DP}(\alpha_0, G_0)$ is a Dirichlet process if the probability that it assigns to such a partition is distributed as $\text{Dir}(\alpha_0 G_0(A_1), \alpha_0 G_0(A_2), \dots, \alpha_0 G_0(A_r))$, for any partition. Intuitively, the Dirichlet process yields measures that are variations on the *base measure* G_0 , where the *concentration parameter* α_0 provides control over the variability around G_0 .

This definition does not make clear the discrete, combinatorial nature of the Dirichlet process. A seminal paper by Sethuraman (1994) laid bare the inherent discreteness of the Dirichlet process, via a characterization in terms of a so-called “stick-breaking” process. Sethuraman (1994) showed that a measure sampled from a Dirichlet process can be written explicitly as:

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k},$$

where the θ_k are independent random variables distributed according to G_0 , where δ_{θ_k} is a delta function at θ_k , and where the β_k are also random (the definition of the β_k is provided in Section 3.1). This shows explicitly that draws from a Dirichlet process are discrete (with probability one).

The stick-breaking representation shows that successive draws from G can yield exactly the same value with positive probability, even if G_0 assigns zero probability to such an event. The fact that values tend to cluster, however, with the number of distinct values growing relatively slowly with successive draws, is not immediately apparent from the stick-breaking characterization. (It is an implicit consequence of the particular way in which the random variables β_k are defined). It is apparent, however, from yet another perspective on the Dirichlet process—the Pólya urn model of Blackwell and MacQueen (1973). Blackwell and MacQueen (1973) showed that having observed

n values (ϕ_1, \dots, ϕ_n) sampled from a measure G distributed according to a Dirichlet process, the probability of the $(n + 1)^{\text{th}}$ value is given by:

$$\phi_{n+1} \mid \phi_1, \dots, \phi_n, \alpha_0, G_0 \sim \sum_{l=1}^n \frac{1}{n + \alpha_0} \delta_{\phi_l} + \frac{\alpha_0}{n + \alpha_0} G_0. \quad (1)$$

Here we explicitly see the clustering—values that have been sampled more frequently in the past have higher probability of being sampled again.

The clustering phenomenon and the discrete nature of the Dirichlet process make it unsuitable for general applications in Bayesian nonparametrics, but they are ideally suited for the problem of placing priors on mixture components in mixture modeling. The idea is simply to associate a mixture component with each cluster. That is, each distinct value of ϕ_i defines a mixture component, with ϕ_i as its parameter. A number of authors have studied such “Dirichlet process mixture models” (Escobar and West, 1995, MacEachern and Müller, 1998). These models provide an alternative to methods that attempt to select a particular number of mixture components, or methods that place an explicit parametric prior on the number of components.

Let us now consider the problem of modeling a set of J groups of data, where each group is modeled as a set of repeated draws from a group-specific mixture model. To allow the number of mixture components within each group to grow, we associate a draw from a Dirichlet process with each group. To link the groups, we take a hierarchical Bayesian point of view, and assume that these draws are a conditionally independent set of draws from the same underlying Dirichlet process $\text{DP}(\alpha_0, G_0)$, where the base measure G_0 is itself a random measure. Conditional on G_0 , we have a stick-breaking representation for each of the groups:

$$G_j = \sum_{k=1}^{\infty} \beta_{jk} \delta_{\theta_{jk}}.$$

Now suppose that we repeatedly draw parameter values and data points within each group. If G_0 is absolutely continuous with respect to Lebesgue measure—i.e., it has no atoms—then although clusters arise *within* each group, the atoms associated with the different groups are different and there is no sharing of clusters *between* groups.

The way to achieve sharing of mixture components between groups is straightforward: The base measure G_0 should not be an arbitrary random measure, but it should itself be distributed according to a Dirichlet process. In this case, G_0 contains *only* atoms (with probability one), and these atoms will be shared by the measures G_j at the next level of the hierarchy. Thus we define a *hierarchical Dirichlet process* in which the underlying base measure for a set of draws from a Dirichlet process is itself distributed according to a Dirichlet process. A *hierarchical Dirichlet process mixture model* will allow sharing of mixture components within and between groups of mixture models.

Having given ourselves the ability to share mixture components among a set of mixture models, we expect to face challenging computational problems in managing this sharing when computing posterior distributions under the hierarchical Dirichlet process. To organize such computations, and to provide a general framework for designing procedures for posterior inference that parallel those available for the Dirichlet process, it is necessary to develop analogs for the hierarchical Dirichlet process of some of the representations that have proved useful in the Dirichlet process setting. We provide these analogs in Section 4—in particular, we discuss a stick-breaking representation of the hierarchical Dirichlet process, an analog of the Pólya urn model that we refer to as the “Chinese

restaurant franchise,” and a representation of the hierarchical Dirichlet process in terms of an infinite limit of finite mixture models. With these representations as background, we present Markov chain Monte Carlo algorithms for posterior inference under hierarchical Dirichlet process mixtures in Section 5. We discuss related work in Section 6 and present our conclusions in Section 7.

2 Setting

We are interested in problems in which observations are organized into *groups*, and where the *observations* are assumed exchangeable within groups. In particular, letting $j \in \{1, 2, \dots, J\}$ index the J groups, and letting $\mathbf{x}_j = (x_{ji})_{i=1}^{n_j}$ denote the n_j observations in group j , we assume that each observation x_{ji} is a conditionally independent draw from a mixture model, where the parameters of the mixture model are drawn once per group. We will also assume that $\mathbf{x}_1, \dots, \mathbf{x}_J$ are exchangeable at the group level. Let $\mathbf{x} = (\mathbf{x}_j)_{j=1}^J$ denote the entire data set.

If each observation is drawn independently from a mixture model, then there is a different mixture component associated with each observation. Let ϕ_{ji} denote a parameter specifying the mixture component associated with the observation x_{ji} . We will refer to the variables ϕ_{ji} as “factors.” Note that these variables are not generally distinct—we will develop a different notation for the distinct values of factors. Let $F(\phi_{ji})$ denote the distribution of x_{ji} given the factor ϕ_{ji} . Let G_j denote a prior distribution for the factors $\phi_j = (\phi_{ji})_{i=1}^{n_j}$ associated with group j . We assume that the factors are conditionally independent given G_j . Thus we have the following probability model:

$$\begin{aligned} \phi_{ji} \mid G_j &\sim G_j && \text{for each } j \text{ and } i, \\ x_{ji} \mid \phi_j &\sim F(\phi_{ji}) && \text{for each } j \text{ and } i. \end{aligned} \tag{2}$$

2.1 Examples

Grouped data of these kind arise in a number of problem domains. Here we describe three examples.

In the field of information retrieval, documents are often modeled under the so-called “bag-of-words assumption”—the assumption that the words in a document are exchangeable (Salton and McGill, 1983). Thus, in our nomenclature, the documents are groups. Blei et al. (2003) presented a model in which the words in a document are drawn from a mixture, where each mixture component is viewed as a “topic.” A “topic” is a probability distribution on words from some basic vocabulary. Thus, in a document concerned with university funding the words in the document might be drawn from the topics “education” and “finance.” In another document concerned with university football the words might be drawn from the topics “education” and “sports.” The mixing proportions for these mixture models are document-specific, but the mixture components are shared across documents. That is, the topics are characteristic of the corpus as a whole, while each individual document is associated with a probability distribution on the available topics. In our nomenclature, the topics are factors, while the distributions on the available topics are the measures G_j .

An analogous problem arises in bioinformatics. Genes are regulated by transcription factors that bind to the regulatory region of the genes and trigger the transcription of the genes. Each transcription factor has an associated DNA footprint that appears as a short sequence known as a “motif” (Davidson, 2001). Each gene has many motifs in its regulatory region, and the same set of motifs are used by many genes. We can thus view a gene as a “group,” analogous to the documents in the information retrieval domain, and the motifs are analogous to the words in the document. A

“factor”—a probability distribution across motifs—can be viewed as a representation of a regulatory circuit.

Finally, we consider a problem in which the group structure is not known a priori. A hidden Markov model is a doubly stochastic Markov chain in which a sequence of “state” variables (v_1, v_2, \dots, v_T) are linked via a Markov chain, and each element y_t in a sequence of “observations” (y_1, y_2, \dots, y_T) is drawn independently of the other observations conditional on v_t (Rabiner, 1989). This model can be treated within our framework by letting the state variables define the groups. Thus, we link as one group all observations y_{t+1} with the same value of the current state v_t . Conditioned on v_t , the observations y_{t+1} are draws from a mixture model with v_{t+1} indicating the mixture component while the mixing proportions are given by the transition probabilities $T(v_{t+1}|v_t)$ of the Markov chain. In other words, the next state v_{t+1} is the “factor” governing the distribution over the observation y_{t+1} . We will further describe the hidden Markov model in Section 6.1.

Each of these problems are thus characterized by a set of mixture distributions, one per group (i.e., one per document, gene, or current state). The mixing proportions associated with these mixtures are group-specific. The mixture components, on the other hand, are shared across the groups. We want to have the same set of topics available to each of the documents, the same set of motifs available to each of the genes, and the same set of next states available to each of the current states.

Thus the problem is that of linking the mixture components among a set of mixture models. We want to do this in the setting in which the number of mixture components is unknown, and where the machinery of Dirichlet processes is used to allow a potentially unbounded set of mixture components, and to provide a representation of prior and posterior uncertainty over these mixture components. Thus we have multiple Dirichlet processes, one per group, and we want to link these processes. In the remainder of the paper we present a hierarchical Bayesian approach to this problem.

3 Dirichlet Processes

In order to make the paper self-contained, we provide a brief overview of Dirichlet processes in this section. After a discussion of basic definitions, we present three different perspectives on the Dirichlet process—one based on the stick-breaking construction, one based on a Pólya urn model, and one based on a limit of finite mixture models. Each of these perspectives will have an analog in the hierarchical Dirichlet process to be introduced in Section 4.

Let (Θ, \mathcal{B}) be a measurable space, with G_0 a probability measure on the space. Let α_0 be a positive real number. A *Dirichlet process* $DP(\alpha_0, G_0)$ is defined to be the distribution of a random probability measure G over (Θ, \mathcal{B}) such that, for any finite measurable partition (A_1, A_2, \dots, A_r) of Θ , the random vector $(G(A_1), \dots, G(A_r))$ is distributed as a finite-dimensional Dirichlet distribution with parameters $(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_r))$:

$$(G(A_1), \dots, G(A_r)) \sim \text{Dir}(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_r)). \quad (3)$$

We write $G \sim DP(\alpha_0, G_0)$ if G is a random probability measure with distribution given by the Dirichlet process. The existence of the Dirichlet process was established by Ferguson (1973).

3.1 The stick-breaking construction

Measures drawn from a Dirichlet process turn out to be discrete with probability one (Ferguson, 1973). This property is made explicit in the *stick-breaking construction* due to Sethuraman (1994). The stick-breaking construction is based on sequences of random variables π'_1, π'_2, \dots and $\theta_1, \theta_2, \dots$. These sequences are i.i.d. and have the following distributions:

$$\pi'_k \mid \alpha_0, G_0 \sim \text{Beta}(1, \alpha_0) \qquad \theta_k \mid \alpha_0, G_0 \sim G_0, \quad (4)$$

where $\text{Beta}(a, b)$ is the Beta distribution with parameters a and b . Now define a random measure G as

$$\pi_k = \pi'_k \prod_{l=1}^{k-1} (1 - \pi'_l) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \quad (5)$$

where δ_θ is a probability measure concentrated at θ . Sethuraman (1994) showed that G as defined in this way is a random probability measure distributed according to $\text{DP}(\alpha_0, G_0)$.

It is important to note that the sequence $\pi = (\pi_k)_{k=1}^{\infty}$ constructed by (4) and (5) satisfies $\sum_{k=1}^{\infty} \pi_k = 1$ with probability one. Thus we may interpret π as a random probability measure on the positive integers. For convenience, we shall write $\pi \sim \text{Stick}(\alpha_0)$ if π is a random probability measure defined by (4) and (5).

3.2 The Chinese restaurant process

A second perspective on the Dirichlet process is provided by the *Pólya urn scheme* due to Blackwell and MacQueen (1973). The Pólya urn scheme shows that not only are draws from the Dirichlet process discrete, but also that they exhibit a clustering property.

The Pólya urn scheme refers not to G directly, but rather to draws from G . Thus, let ϕ_1, ϕ_2, \dots be a sequence of i.i.d. random variables distributed according to G . That is, the variables ϕ_1, ϕ_2, \dots are conditionally independent given G , and hence exchangeable. Let us consider the successive conditional distributions of ϕ_i given $\phi_1, \dots, \phi_{i-1}$, where G has been integrated out. Blackwell and MacQueen (1973) showed that these conditional distributions have the following simple form:

$$\phi_i \mid \phi_1, \dots, \phi_{i-1}, \alpha_0, G_0 \sim \sum_{l=1}^{i-1} \frac{1}{i-1+\alpha_0} \delta_{\phi_l} + \frac{\alpha_0}{i-1+\alpha_0} G_0. \quad (6)$$

This expression shows that ϕ_i has positive probability of being equal to one of the previous draws, and that there is a positive reinforcement effect—the more often a point is drawn, the more likely it is to be drawn in the future. We can interpret the conditional distributions in terms of a simple urn model in which a ball of a distinct color is associated with each atom. The balls are drawn equiprobably; when a ball is drawn it is placed back in the urn with another ball of the same color. In addition, with probability proportional to α_0 a new atom is created by drawing from G_0 and a ball of a new color is added to the urn.

To make the clustering property explicit, it is helpful to introduce a new set of variables that represent distinct values of the atoms. Define $\theta_1, \dots, \theta_K$ to be the distinct values taken on by

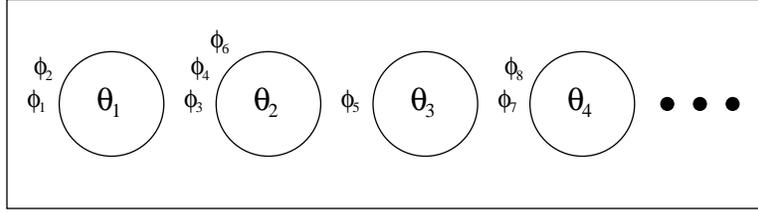


Figure 1: A depiction of a Chinese restaurant after eight customers have been seated. Customers (ϕ_i 's) are seated at tables (circles) which correspond to the unique values θ_k .

$\phi_1, \dots, \phi_{i-1}$, and let n_k be the number of values $\phi_{i'}$ that are equal to θ_k for $1 \leq i' < i$. We can re-express (6) as

$$\phi_i \mid \phi_1, \dots, \phi_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^K \frac{n_k}{i-1+\alpha_0} \delta_{\theta_k} + \frac{\alpha_0}{i-1+\alpha_0} G_0. \quad (7)$$

Using a somewhat different metaphor, the Pólya urn scheme is also known as the *Chinese restaurant process* (Aldous, 1985). This metaphor has turned out to be useful in considering various generalizations of the Dirichlet process (Pitman, 2002), and it will be useful in this paper. The metaphor is as follows. Consider a Chinese restaurant with an unbounded number of tables. Each ϕ_i corresponds to a customer who enters the restaurant, while the distinct values θ_k correspond to the tables at which the customers sit. The i^{th} customer sits at the table indexed by θ_k with probability proportional to n_k (in which case we set $\phi_i = \theta_k$), and sits at a new table with probability proportional to α_0 (set $\phi_i \sim G_0$). An example of a Chinese restaurant is depicted graphically in Figure 1.

3.3 Dirichlet process mixture models

One of the most important applications of the Dirichlet process is as a nonparametric prior distribution on the components of a mixture model. In particular, suppose that observations x_i arise as follows:

$$\begin{aligned} \phi_i \mid G &\sim G \\ x_i \mid \phi_i &\sim F(\phi_i), \end{aligned} \quad (8)$$

where $F(\phi_i)$ denotes the distribution of the observation x_i given ϕ_i . The *factors* ϕ_i are conditionally independent given G , while the observation x_i is conditionally independent of the other observations given the factor ϕ_i . When G is distributed according to a Dirichlet process, this model is referred to as a *Dirichlet process mixture model*. A graphical model representation of a Dirichlet process mixture model is shown in Figure 2(a).

Since G can be represented using a stick-breaking construction (5), the factors ϕ_i take on values θ_k with probability π_k . We may denote this using an indicator variable z_i , which takes on positive integral values and is distributed according to π (interpreting π as a random probability measure on the positive integers). Hence an equivalent representation of a Dirichlet process mixture is given by

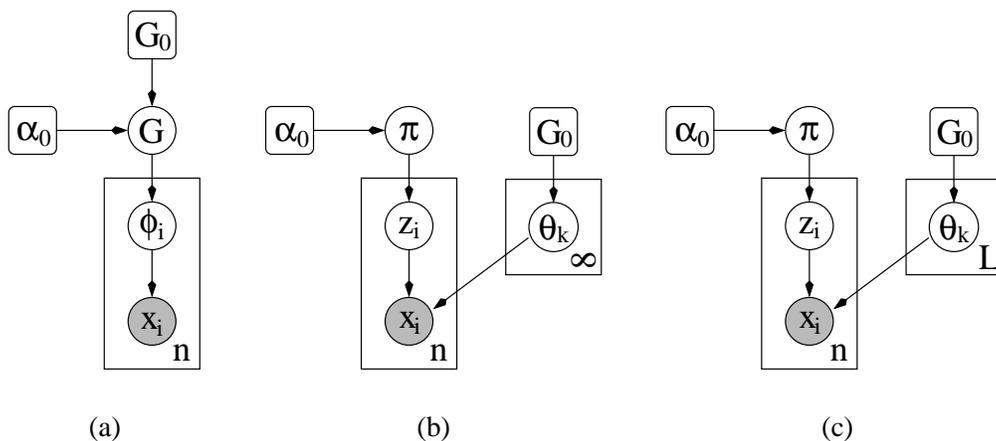


Figure 2: (a) A representation of a Dirichlet process mixture model as a graphical model. In the graphical model formalism, each node in the graph is associated with a random variable and joint probabilities are defined as products of conditional probabilities, where a conditional probability is associated with a node and its parents. Rectangles (“plates”) denote replication, with the number of replicates given by the number in the bottom right corner of the rectangle. We also use a square with rounded corners to denote a variable that is a fixed hyperparameter, while a shaded node is an observable. (b) An equivalent representation of a Dirichlet process mixture model in terms of the stick-breaking construction. (c) A finite mixture model (notice the L instead ∞).

Figure 2(b), where the conditional distributions are:

$$\begin{aligned}
 \pi \mid \alpha_0 &\sim \text{Stick}(\alpha_0) & z_i \mid \pi &\sim \pi \\
 \theta_k \mid G_0 &\sim G_0 & x_i \mid z_i, (\theta_k)_{k=1}^{\infty} &\sim F(\theta_{z_i}).
 \end{aligned}
 \tag{9}$$

Here $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ and $\phi_i = \theta_{z_i}$.

3.4 The infinite limit of finite mixture models

A Dirichlet process mixture model can be derived as the limit of a sequence of finite mixture models, where the number of mixture components is taken to infinity (Neal, 1992, Rasmussen, 2000, Green and Richardson, 2001, Ishwaran and Zarepour, 2002). This limiting process provides a third perspective on the Dirichlet process.

Suppose we have L mixture components. Let $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)$ denote the mixing proportions.¹ We place a Dirichlet prior on $\boldsymbol{\pi}$ with symmetric parameters $(\alpha_0/L, \dots, \alpha_0/L)$. Let θ_k denote the parameter vector associated with mixture component k , and let θ_k have prior distribution G_0 . Drawing an observation x_i from the mixture model involves picking a specific mixture component with probability given by the mixing proportions; let z_i denote that component. We thus have the

¹Previously we used the symbol $\boldsymbol{\pi}$ to denote the weights associated with the atoms in G . We have deliberately overloaded the definition of $\boldsymbol{\pi}$ here; as we shall see later, they are closely related. In fact, in the limit $L \rightarrow \infty$ they will be equivalent up to a random *size-based permutation* of their entries (Patil and Taillie, 1977).

following model:

$$\begin{aligned} \boldsymbol{\pi} \mid \alpha_0 &\sim \text{Dir}(\alpha_0/L, \dots, \alpha_0/L) & z_i \mid \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\ \theta_k \mid G_0 &\sim G_0 & x_i \mid z_i, (\theta_k)_{k=1}^L &\sim F(\theta_{z_i}). \end{aligned} \quad (10)$$

The graphical model is shown in Figure 2(c). Let $G^L = \sum_{k=1}^L \pi_k \delta_{\theta_k}$. Ishwaran and Zarepour (2002) show that for every measurable function f integrable with respect to G_0 , we have, as $L \rightarrow \infty$:

$$\int f(\phi) dG^L(\phi) \xrightarrow{\mathcal{D}} \int f(\phi) dG(\phi). \quad (11)$$

A consequence of this is that the marginal distribution induced on the observations x_1, \dots, x_n approaches that of a Dirichlet process mixture model. This limiting process is unsurprising in hindsight, given the striking similarity between Figures 2(b) and 2(c).

4 Hierarchical Dirichlet Processes

We propose a nonparametric Bayesian approach to the modeling of grouped data, where each group is associated with a mixture model, and where we wish to link these mixture models. By analogy with Dirichlet process mixture models, we first define the appropriate nonparametric prior, which we refer to as the *hierarchical Dirichlet process*. We then show how this prior can be used in the grouped mixture model setting. We present analogs of the three perspectives presented earlier for the Dirichlet process—a stick-breaking construction, a Chinese restaurant process representation, and a representation in terms of a limit of finite mixture models.

A hierarchical Dirichlet process is a distribution over a set of random probability measures over (Θ, \mathcal{B}) . The process defines a set of random probability measures $(G_j)_{j=1}^J$, one for each group, and a global random probability measure G_0 . The global measure G_0 is distributed as a Dirichlet process with concentration parameter γ and base probability measure H :

$$G_0 \mid \gamma, H \sim \text{DP}(\gamma, H), \quad (12)$$

and the random measures $(G_j)_{j=1}^J$ are conditionally independent given G_0 , with distributions given by a Dirichlet process with base probability measure G_0 :

$$G_j \mid \alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0). \quad (13)$$

The hyperparameters of the hierarchical Dirichlet process consist of the baseline probability measure H , and the concentration parameters γ and α_0 . The baseline H provides the prior distribution for the parameters $(\phi_j)_{j=1}^J$. The distribution G_0 varies around the prior H , with the amount of variability governed by γ . The actual distribution G_j over the parameters ϕ_j in the j^{th} group deviates from G_0 , with the amount of variability governed by α_0 . If we expect the variability in different groups to be different, we can use a separate concentration parameter α_j for each group j . Finally, following Escobar and West (1995), we put vague gamma priors on γ and α_0 .

A hierarchical Dirichlet process can be used as the prior distribution over the factors for grouped data. For each j let $(\phi_{ji})_{i=1}^{n_j}$ be i.i.d. random variables distributed as G_j . These ϕ_{ji} are factors each

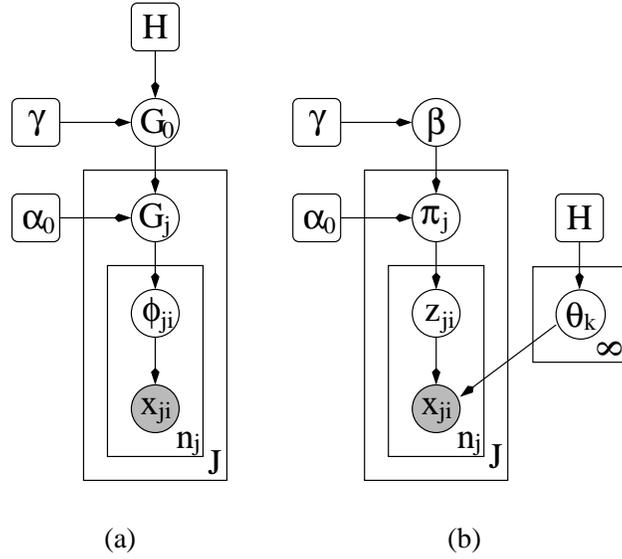


Figure 3: (a) A hierarchical Dirichlet process mixture model. (b) An alternative representation of a hierarchical Dirichlet process mixture model in terms of the stick-breaking construction.

corresponding to one observation x_{ji} . The likelihood is given by:

$$\begin{aligned} \phi_{ji} | G_j &\sim G_j \\ x_{ji} | \phi_{ji} &\sim F(\phi_{ji}). \end{aligned} \quad (14)$$

This completes the definition of a *hierarchical Dirichlet process mixture model*. The corresponding graphical model is shown in Figure 3(a).

Notice that $(\phi_{ji})_{i=1}^{n_j}$ are exchangeable random variables if we integrate out G_j . Similarly, $(\phi_j)_{j=1}^J$ are exchangeable at the group level. Since each x_{ji} is independently distributed according to $F(\phi_{ji})$, our exchangeability assumption for the grouped data $(\mathbf{x}_j)_{j=1}^J$ is not violated by the hierarchical Dirichlet process mixture model.

4.1 The stick-breaking construction

Given that the global measure G_0 is distributed as a Dirichlet process, it can be expressed using a stick-breaking representation:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}, \quad (15)$$

where $\theta_k \sim H$ independently and $\beta = (\beta_i)_{i=1}^{\infty} \sim \text{Stick}(\gamma)$ are mutually independent. Since G_0 has support at the points $\theta = (\theta_i)_{i=1}^{\infty}$, each G_j necessarily has support at these points as well, and can thus be written as:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}. \quad (16)$$

Let $\pi_j = (\pi_{jk})_{k=1}^\infty$. Note that the weights π_j are independent given β (since G_j are independent given G_0). We now describe how the weights π_j are related to the global weights β .

Let (A_1, \dots, A_r) be a measurable partition of Θ and let $K_l = \{k : \theta_k \in A_l\}$ for $l = 1, \dots, r$. Note that (K_1, \dots, K_r) is a finite partition of the positive integers. Further, assuming that H is non-atomic, the θ_k 's are distinct with probability one, so any partition of the positive integers corresponds to some partition of Θ . Thus, for each j we have:

$$\begin{aligned} (G_j(A_1), \dots, G_j(A_r)) &\sim \text{Dir}(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_r)) \\ &= \left(\sum_{k \in K_1} \pi_{jk}, \dots, \sum_{k \in K_r} \pi_{jk} \right) \sim \text{Dir} \left(\alpha_0 \sum_{k \in K_1} \beta_k, \dots, \alpha_0 \sum_{k \in K_r} \beta_k \right), \end{aligned} \quad (17)$$

for every finite partition of the positive integers. Hence each π_j is independently distributed according to $\text{DP}(\alpha_0, \beta)$, where we interpret β and π_j as probability measures on the positive integers.

As in the Dirichlet process mixture model, since each factor ϕ_{ji} is distributed according to G_j , it will take on the value θ_k with probability π_{jk} . Again let z_{ji} be an indicator variable such that $\phi_{ji} = \theta_{z_{ji}}$. Given z_{ji} we have $x_{ji} \sim F(\theta_{z_{ji}})$. Thus Figure 3(b) gives an equivalent representation of the hierarchical Dirichlet process mixture, with conditional distributions summarized here:

$$\begin{aligned} \beta &| \gamma \sim \text{Stick}(\gamma) \\ \pi_j &| \alpha_0, \beta \sim \text{DP}(\alpha_0, \beta) & z_{ji} &| \pi_j \sim \pi_j \\ \theta_k &| H \sim H & x_{ji} &| z_{ji}, (\theta_k)_{k=1}^\infty \sim F(\theta_{z_{ji}}). \end{aligned} \quad (18)$$

Given the relations between π_j and β , we now derive an explicit construction for the elements of β and π_j . Recall that the stick-breaking construction for Dirichlet processes defines the variables β_k in (15) as follows:

$$\beta'_k \sim \text{Beta}(1, \gamma) \quad \beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l). \quad (19)$$

Using (17), we will show that the following stick-breaking construction produces a random probability measure $\pi_j \sim \text{DP}(\alpha_0, \beta)$:

$$\pi'_{jk} \sim \text{Beta} \left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{l=1}^k \beta_l \right) \right) \quad \pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl}). \quad (20)$$

To derive (20), first notice that for a partition $(\{1, \dots, k-1\}, \{k\}, \{k+1, k+2, \dots\})$, (17) gives:

$$\left(\sum_{l=1}^{k-1} \pi_{jl}, \pi_{jk}, \sum_{l=k+1}^{\infty} \pi_{jl} \right) \sim \text{Dir} \left(\alpha_0 \sum_{l=1}^{k-1} \beta_l, \alpha_0 \beta_k, \alpha_0 \sum_{l=k+1}^{\infty} \beta_l \right). \quad (21)$$

Removing the first element, and using standard properties of the finite Dirichlet distribution, we have:

$$\frac{1}{1 - \sum_{l=1}^{k-1} \pi_{jl}} \left(\pi_{jk}, \sum_{l=k+1}^{\infty} \pi_{jl} \right) \sim \text{Dir} \left(\alpha_0 \beta_k, \alpha_0 \sum_{l=k+1}^{\infty} \beta_l \right). \quad (22)$$

Finally, define $\pi'_{jk} = \frac{\pi_{jk}}{1 - \sum_{l=1}^{k-1} \pi_{jl}}$ and observe that $1 - \sum_{l=1}^k \beta_l = \sum_{l=k+1}^{\infty} \beta_l$ to obtain (20). Together with (19), (15) and (16), this completes the description of the stick-breaking construction for hierarchical Dirichlet processes.

4.2 The Chinese restaurant franchise

In this section we describe an analog of the Chinese restaurant process for hierarchical Dirichlet processes that we refer to as the ‘‘Chinese restaurant franchise.’’ In the Chinese restaurant franchise, the metaphor of the Chinese restaurant process is extended to allow multiple restaurants which share a set of dishes.

Recall that the factors ϕ_{ji} are random variables with distribution G_j . In the following discussion, we will let $\theta_1, \dots, \theta_K$ denote K i.i.d. random variables distributed according to H , and, for each j , we let $\psi_{j1}, \dots, \psi_{jT_j}$ denote T_j i.i.d. variables distributed according to G_0 .

Each ϕ_{ji} is associated with one ψ_{jt} , while each ψ_{jt} is associated with one θ_k . Let t_{ji} be the index of the ψ_{jt} associated with ϕ_{ji} , and let k_{jt} be the index of θ_k associated with ψ_{jt} . Let n_{jt} be the number of ϕ_{ji} 's associated with ψ_{jt} , while m_{jk} is the number of ψ_{jt} 's associated with θ_k . Define $m_k = \sum_j m_{jk}$ as the number of ψ_{jt} 's associated with θ_k over all j . Notice that while the values taken on by the ψ_{jt} 's need not be distinct (indeed, they are distributed according to a discrete random probability measure $G_0 \sim \text{DP}(\gamma, H)$), we are denoting them as distinct random variables.

First consider the conditional distribution for ϕ_{ji} given $\phi_{j1}, \dots, \phi_{ji-1}$ and G_0 , where G_j is integrated out. From (7), we have:

$$\phi_{ji} \mid \phi_{j1}, \dots, \phi_{ji-1}, \alpha_0, G_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{i-1 + \alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1 + \alpha_0} G_0, \quad (23)$$

This is a mixture, and a draw from this mixture can be obtained by drawing from the terms on the right-hand side with probabilities given by the corresponding mixing proportions. If a term in the first summation is chosen, then we set $\phi_{ji} = \psi_{jt}$ and let $t_{ji} = t$ for the chosen t . If the second term is chosen, then we increment T_j by one, draw $\psi_{jT_j} \sim G_0$ and set $\phi_{ji} = \psi_{jT_j}$ and $t_{ji} = T_j$. The various pieces of information involved are depicted as a ‘‘Chinese restaurant’’ in Figure 4(a).

Now we proceed to integrate out G_0 . Notice that G_0 appears only in its role as the distribution of the variables ψ_{jt} . Since G_0 is distributed according to a Dirichlet process, we can integrate it out by using (7) again and writing the conditional distribution of ψ_{jt} directly:

$$\psi_{jt} \mid \psi_{11}, \psi_{12}, \dots, \psi_{21}, \dots, \psi_{jt-1}, \gamma, H \sim \sum_{k=1}^K \frac{m_k}{\sum_k m_k + \gamma} \delta_{\theta_k} + \frac{\gamma}{\sum_k m_k + \gamma} H. \quad (24)$$

If we draw ψ_{jt} via choosing a term in the summation on the right-hand side of this equation, we set $\psi_{jt} = \theta_k$ and let $k_{jt} = k$ for the chosen k . If the second term is chosen, we increment K by one, draw $\theta_K \sim H$ and set $\psi_{jt} = \theta_K$, $k_{jt} = K$.

This completes the description of the conditional distributions of the ϕ_{ji} variables. To use these equations to obtain samples of ϕ_{ji} , we proceed as follows. For each j and i , first sample ϕ_{ji} using (23). If a new sample from G_0 is needed, we use (24) to obtain a new sample ψ_{jt} and set $\phi_{ji} = \psi_{jt}$. This procedure is summarized in Algorithm 1.

Note that in the hierarchical Dirichlet process the values of the factors are shared between the groups, as well as within the groups. This is a key property of hierarchical Dirichlet processes.

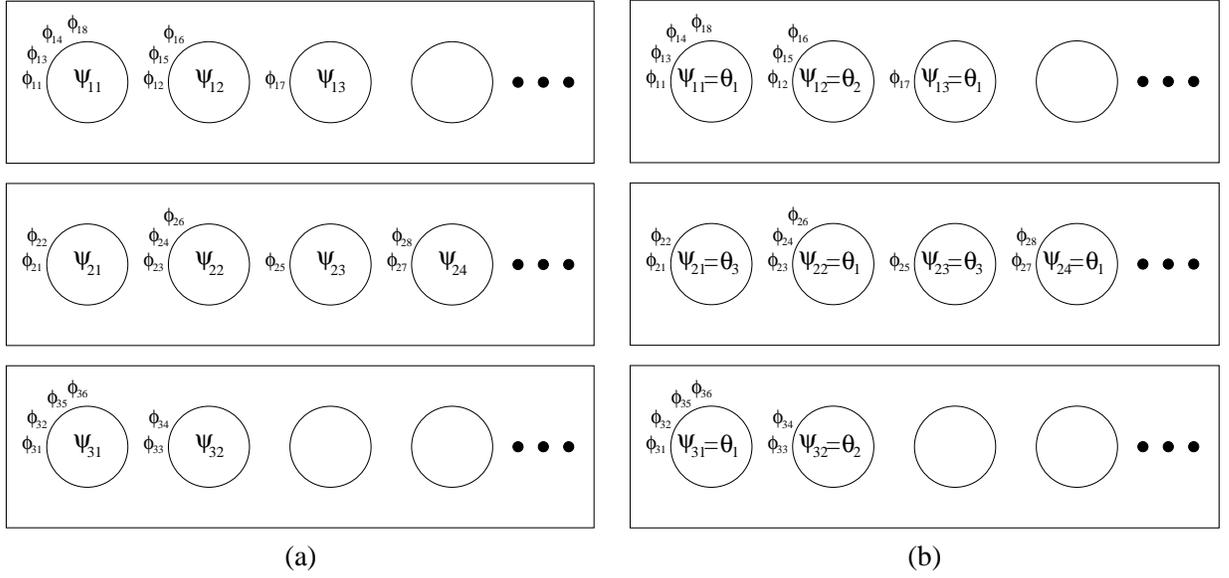


Figure 4: (a) A depiction of a hierarchical Dirichlet process as a Chinese restaurant. Each rectangle is a restaurant (group) with a number of tables. Each table is associated with a parameter ψ_{jt} which is distributed according to G_0 , and each ϕ_{ji} sits at the table to which it has been assigned in (23). (b) Integrating out G_0 , each ψ_{jt} is assigned some dish (mixture component) θ_k .

We call this generalized urn model the *Chinese restaurant franchise* (see Figure 4(b)). The metaphor is as follows. We have a franchise with J restaurants, with a shared menu across the restaurants. At each table of each restaurant one dish is ordered from the menu by the first customer who sits there, and it is shared among all customers who sit at that table. Multiple tables at multiple restaurants can serve the same dish. The restaurants correspond to groups, the customers correspond to the ϕ_{ji} variables, the tables to the ψ_{jt} variables, and the dishes to the θ_k variables.

A customer entering the j^{th} restaurant sits at one of the occupied tables with a certain probability, and sits at a new table with the remaining probability. This is the Chinese restaurant process and corresponds to (23). If the customer sits at an occupied table, she eats the dish that has already been ordered. If she sits at a new table, she gets to pick the dish for the table. The dish is picked according to its popularity among the whole franchise, while a new dish can also be tried. This corresponds to (24).

4.3 The infinite limit of finite mixture models

As in the case of a Dirichlet process mixture model, the hierarchical Dirichlet process mixture model can be derived as the infinite limit of finite mixtures. In this section, we present two apparently different finite models that both yield the hierarchical Dirichlet process mixture in the infinite limit, each emphasizing a different aspect of the model. We also show how a third finite model fails to yield the hierarchical Dirichlet process; the reasons for this failure will provide additional insight.

Consider the first finite model, shown in Figure 5(a). Here the number of mixture components L is a positive integer, and the mixing proportions β and π_j are vectors of length L . The conditional

Algorithm 1: Obtaining samples from a sample of a hierarchical Dirichlet process.

- **Initialize.**

Set $K = 0, T_j = 0$ for each j .

- **Generate samples ϕ_{ji} .**

For each j and i , with probability:

$$\frac{n_{jt}}{\sum_t n_{jt} + \alpha_0} \text{ for } t = 1, 2, \dots, T_j:$$

Assign item i to “table” t .

Set $t_{ji} \leftarrow t, \phi_{ji} \leftarrow \psi_{jt}, n_{jt} \leftarrow n_{jt} + 1$.

$$\frac{\alpha_0}{\sum_t n_{jt} + \alpha_0} :$$

Generate a new “table” and sample from G_0 .

Set $T_j \leftarrow T_j + 1, n_{jT_j} \leftarrow 0$. With probability:

$$\frac{m_k}{\sum_k m_k + a} \text{ for } k = 1, 2, \dots, K:$$

Assign “table” T_j to mixture component k .

Set $k_{jT_j} \leftarrow k, \psi_{jT_j} \leftarrow \theta_k, m_k \leftarrow m_k + 1$.

$$\frac{a}{\sum_k m_k + a} :$$

Generate a new mixture component and sample from H .

Set $K \leftarrow K + 1, m_K \leftarrow 0, \theta_K \sim H$.

Assign “table” T_j to mixture component K .

Set $k_{jT_j} \leftarrow K, \psi_{jT_j} \leftarrow \theta_K, m_K \leftarrow m_K + 1$.

Assign item i to “table” T_j .

Set $t_{ji} \leftarrow T_j, \phi_{ji} \leftarrow \psi_{jT_j}, n_{jT_j} \leftarrow n_{jT_j} + 1$.

distributions are given by

$$\begin{aligned} \beta &| \gamma \sim \text{Dir}(\gamma/L, \dots, \gamma/L) \\ \pi_j &| \alpha_0, \beta \sim \text{Dir}(\alpha_0 \beta) & z_{ji} &| \pi_j \sim \pi_j \\ \theta_k &| H \sim H & x_{ji} &| z_{ji}, (\theta_k)_{k=1}^L \sim F(\theta_{z_{ji}}). \end{aligned} \quad (25)$$

Let us consider the random probability measures $G_0^L = \sum_{k=1}^L \beta_k \delta_{\theta_k}$ and $G_j^L = \sum_{k=1}^L \pi_{jk} \delta_{\theta_k}$. As in Section 3.4, for every measurable function f integrable with respect to H we have

$$\int f(\phi) dG_0^L(\phi) \xrightarrow{\mathcal{D}} \int f(\phi) dG_0(\phi), \quad (26)$$

as $L \rightarrow \infty$. Further, using standard properties of the Dirichlet distribution, we see that (17) still holds for the finite case for partitions of $\{1, \dots, L\}$; hence we have:

$$G_j^L \sim \text{DP}(\alpha_0, G_0^L). \quad (27)$$

It is now clear that as $L \rightarrow \infty$ the marginal distribution this finite model induces on \mathbf{x} approaches the hierarchical Dirichlet process mixture model.

By way of comparison, it is interesting to consider what happens if we set $\beta = (1/L, \dots, 1/L)$ symmetrically instead, and take the limit $L \rightarrow \infty$ (shown in Figure 5(b)). Let k be a mixture

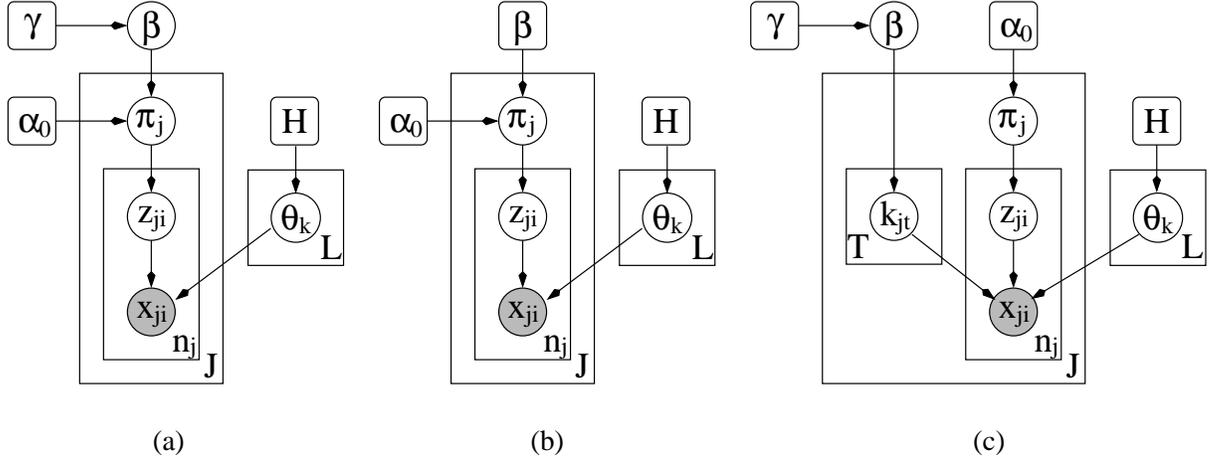


Figure 5: Finite models. (a) A finite hierarchical multiple mixture model whose infinite limit yields the hierarchical Dirichlet process mixture model. (b) The finite model with symmetric β weights. The various mixture models are independent of each other given α_0, β and H so cannot capture dependencies between the groups. (c) Another finite model that yields the hierarchical Dirichlet process in the infinite limit.

component used in group j ; i.e., suppose that $z_{ji} = k$ for some i . Consider the probability that mixture component k is used in another group $j' \neq j$; i.e., suppose that $z_{j'i'} = k$ for some i' . Since $\pi_{j'}$ is independent of π_j , and β is symmetric, this probability is:

$$p(\exists i' : z_{j'i'} = k \mid \alpha_0 \beta) \leq \sum_{i'} p(z_{j'i'} = k \mid \alpha_0 \beta) = \frac{n_j}{L} \rightarrow 0 \quad \text{as } L \rightarrow \infty. \quad (28)$$

Since group j can use at most n_j mixture components (there are only n_j observations), as $L \rightarrow \infty$ the groups will have zero probability of sharing a mixture component. This lack of overlap among the mixture components in different groups is the behavior that we consider undesirable and wish to avoid.

The lack of overlap arises when we assume that each mixture component has the same prior probability of being used in each group (i.e., β is symmetric). Thus one possible direct way to deal with the problem would be to assume asymmetric weights for β . In order that the parameter set does not grow as $L \rightarrow \infty$, we need to place a prior on β and integrate over these values. The hierarchical Dirichlet process is in essence an elegant way of imposing this prior.

A third finite model solves the lack-of-overlap problem via a different method. Instead of introducing dependencies between the groups by placing a prior on β (as in the first finite model), each group can instead choose a subset of T mixture components from a model-wide set of L mixture components. In particular consider the model given in Figure 5(c), where:

$$\begin{aligned}
 \beta \mid \gamma &\sim \text{Dir}(\gamma/L, \dots, \gamma/L) & k_{jt} \mid \beta &\sim \beta \\
 \pi_j \mid \alpha_0 &\sim \text{Dir}(\alpha_0/T, \dots, \alpha_0/T) & t_{ji} \mid \pi_j &\sim \pi_j \\
 \theta_k \mid H &\sim H & x_{ji} \mid t_{ji}, (k_{jt})_{t=1}^T, (\theta_k)_{k=1}^L &\sim F(\theta_{k_{jt_{ji}}}).
 \end{aligned} \quad (29)$$

As $T \rightarrow \infty$ and $L \rightarrow \infty$, the limit of this model is the Chinese restaurant franchise process; hence the infinite limit of this model is also the hierarchical Dirichlet process mixture model.

5 Inference for the hierarchical Dirichlet process mixture model

In this section we describe two Markov chain Monte Carlo sampling schemes for the hierarchical Dirichlet process mixture model. The first one is based on the Chinese restaurant franchise, while the second one is an auxiliary variable method based upon the infinite limit of the finite model in Figure 5(a).

We first recall the various variables and quantities of interest. The variables x_{ji} are the observed data. Each x_{ji} comes from a distribution $F(\phi_{ji})$ where the parameter is the factor ϕ_{ji} . Let $F(\theta)$ have density $f(\cdot|\theta)$. Let the factor ϕ_{ji} be associated with the table t_{ji} in the restaurant representation, and let $\phi_{ji} = \psi_{jt_{ji}}$. The random variable ψ_{jt} is an instance of mixture component k_{jt} ; i.e., we have $\psi_{jt} = \theta_{k_{jt}}$. The prior over the parameters θ_k is H , with density $h(\cdot)$. Let $z_{ji} = k_{jt_{ji}}$ denote the mixture component associated with the observation x_{ji} . Finally the global weights are $\beta = (\beta_k)_{k=1}^{\infty}$, and the group weights are $\pi_j = (\pi_{jk})_{k=1}^{\infty}$. The global distribution of the factors is $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}$, while the group-specific distributions are $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}$.

For each group j , define the occupancy numbers n_j as the number of observations, n_{jt} the number of ϕ_{ji} 's associated with ψ_{jt} , and n_{jk} the number of ϕ_{ji} 's indirectly associated with θ_k through ψ_{jt} . Also let m_{jk} be the number of ψ_{jt} 's associated with θ_k , and let $m_k = \sum_j m_{jk}$. Finally let K be the number of θ_k 's, and T_j the number of ψ_{jt} 's in group j . By permuting the indices, we may always assume that each t_{ji} takes on values in $\{1, \dots, T_j\}$, and each k_{jt} takes values in $\{1, \dots, K\}$.

Let $\mathbf{x}_j = (x_{j1}, \dots, x_{jn_j})$, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_J)$, $\mathbf{t} = (t_{ji} : \text{all } j, i)$, $\mathbf{k} = (k_{jt} : \text{all } j, t)$, $\mathbf{z} = (z_{ji} : \text{all } j, i)$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ and $\mathbf{m} = (m_{jk} : \text{all } j, k)$. When a superscript is attached to a set of variables or an occupancy number, e.g., $\boldsymbol{\theta}^{-k}$, \mathbf{k}^{-jt} , n_{jt}^{-i} , this means that the variable corresponding to the superscripted index is removed from the set or from the calculation of the occupancy number. In the examples, $\boldsymbol{\theta}^{-k} = \boldsymbol{\theta} \setminus \theta_k$, $\mathbf{k}^{-jt} = \mathbf{k} \setminus k_{jt}$ and n_{jt}^{-i} is the number of observations in group j whose factor is associated with ψ_{jt} , except item x_{ji} .

5.1 Posterior sampling in the Chinese restaurant franchise

The Chinese restaurant franchise presented in Section 4.2 can be used to produce samples from the prior distribution over the ϕ_{ji} , as well as intermediary information related to the tables and mixture components. This scheme can be adapted to yield a Gibbs sampling scheme for posterior sampling given observations \mathbf{x} .

Rather than dealing with the ϕ_{ji} 's and ψ_{jt} 's directly, we shall sample their index variables t_{ji} and k_{jt} as well as the distinct values θ_k . The ϕ_{ji} 's and ψ_{jt} 's can be reconstructed from these index variables and the θ_k . This representation makes the Markov chain Monte Carlo sampling scheme more efficient (cf. Neal, 2000). Notice that the t_{ji} and the k_{jt} inherit the exchangeability properties of the ϕ_{ji} and the ψ_{jt} —the conditional distributions in (23) and (24) can be easily adapted to be expressed in terms of t_{ji} and k_{jt} .

The state space consists of values of \mathbf{t} , \mathbf{k} and $\boldsymbol{\theta}$. Notice that the number of k_{jt} and θ_k variables represented explicitly by the algorithm is not fixed. We can think of the actual state space as con-

sisting of a countably infinite number of θ_k and k_{jt} . Only finitely many are actually associated to data and represented explicitly.

Sampling t . To compute the conditional distribution of t_{ji} given the remainder of the variables, we make use of exchangeability and treat t_{ji} as the last variable being sampled in the last group in (23) and (24). We can then easily compute the conditional prior distribution for t_{ji} . Combined with the likelihood of generating x_{ji} , we obtain the conditional posterior for t_{ji} .

Using (23), the prior probability that t_{ji} takes on a particular previously seen value t is proportional to n_{jt}^{-i} , whereas the probability that it takes on a new value (say $t^{\text{new}} = T_j + 1$) is proportional to α_0 . The likelihood of the data given $t_{ji} = t$ for some previously seen t is simply $f(x_{ji}|\theta_{k_{jt}})$. To determine the likelihood if t_{ji} takes on value t^{new} , the simplest approach would be to generate a sample for $k_{jt^{\text{new}}}$ from its conditional prior (24) (Neal, 2000). If this value of $k_{jt^{\text{new}}}$ is itself a new value, say $k^{\text{new}} = K + 1$, we may generate a sample for $\theta_{k^{\text{new}}}$ as well:

$$k_{jt^{\text{new}}} | \mathbf{k} \sim \sum_{k=1}^K \frac{m_k}{\sum_k m_k + \gamma} \delta_k + \frac{\gamma}{\sum_k m_k + \gamma} \delta_{k^{\text{new}}} \quad \theta_{k^{\text{new}}} \sim H, \quad (30)$$

The likelihood for x_{ji} given $t_{ji} = t^{\text{new}}$ is now simply $f(x_{ji}|\theta_{k_{jt^{\text{new}}}})$. Combining all this information, the conditional distribution of t_{ji} is then

$$p(t_{ji} = t | \mathbf{t}^{-ji}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{x}) \propto \begin{cases} \alpha_0 f(x_{ji}|\theta_{k_{jt}}) & \text{if } t = t^{\text{new}}, \\ n_{jt}^{-i} f(x_{ji}|\theta_{k_{jt}}) & \text{if } t \text{ previously used.} \end{cases} \quad (31)$$

If the sampled value of t_{ji} is t^{new} , we insert the temporary values of $k_{jt^{\text{new}}}, \theta_{k_{jt^{\text{new}}}}$ into the data structure; otherwise these temporary variables are discarded. The values of n_{jt}, m_k, T_j and K are also updated as needed. In our implementation, rather than sampling $k_{jt^{\text{new}}}$, we actually consider all possible values for $k_{jt^{\text{new}}}$ and sum it out. This gives better convergence.

If as a result of updating t_{ji} some table t becomes unoccupied, i.e., $n_{jt} = 0$, then the probability that this table will be occupied again in the future will be zero, since this is always proportional to n_{jt} . As a result, we may delete the corresponding k_{jt} from the data structure. If as a result of deleting k_{jt} some mixture component k becomes unallocated, we may delete this mixture component as well.

Sampling k . Sampling the k_{jt} variables is similar to sampling the t_{ji} variables. First we generate a new mixture parameter $\theta_{k^{\text{new}}} \sim H$. Since changing k_{jt} actually changes the component membership of all data items in table t , the likelihood of setting $k_{jt} = k$ is given by $\prod_{i:t_{ji}=t} f(x_{ji}|\theta_k)$, so that the conditional probability of k_{jt} is

$$p(k_{jt} = k | \mathbf{t}, \mathbf{k}^{-jt}, \boldsymbol{\theta}, \mathbf{x}) \propto \begin{cases} \gamma \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k = k^{\text{new}}, \\ m_k^{-t} \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k \text{ previously used.} \end{cases} \quad (32)$$

Sampling θ . Conditioned on the indicator variables \mathbf{k} and \mathbf{t} , the parameters θ_k for each mixture component are mutually independent. The posterior distribution is dependent only on the data items associated with component k , and is given by:

$$p(\theta_k | \mathbf{t}, \mathbf{k}, \boldsymbol{\theta}^{-k}, \mathbf{x}) \propto h(\theta_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\theta_k) \quad (33)$$

where $h(\theta)$ is the density of the baseline distribution H at θ . If H is conjugate to $F(\cdot)$ we have the option of integrating out $\boldsymbol{\theta}$.

5.2 Posterior sampling with auxiliary variables

In this section we will develop a sampling scheme for the hierarchical Dirichlet process mixture model based on auxiliary variables. We first develop the sampling scheme for the finite model given in (25) and Figure 5(a). Taking the infinite limit, the model approaches a hierarchical Dirichlet process mixture model, and the sampling scheme we developed approaches a sampling scheme for the hierarchical Dirichlet process mixture as well. For a similar treatment of the Dirichlet process mixture model, see Neal (1992) and Rasmussen (2000).

Suppose we have L mixture components. In order that our sampling scheme stays computationally feasible when we take $L \rightarrow \infty$, we need a representation of the posterior which does not grow with L . Suppose that out of the L components only K are currently used to model the observations. It is unnecessary to explicitly represent each of the unused components separately, so we will instead pool them together and use a single *unrepresented* component. Whenever the unrepresented component gets chosen to model an observation, we will increment K and instantiate a new component from this pool.

The variables of interest in the finite model are \mathbf{z} , $\boldsymbol{\pi}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We will integrate out $\boldsymbol{\pi}$, and Gibbs sample \mathbf{z} , $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. By permuting the indices we may assume that the represented components are $1, \dots, K$. Hence each $z_{ji} \leq K$, and we explicitly represent β_k and θ_k for $1 \leq k \leq K$. Define $\beta_u = \sum_{k=K+1}^L \beta_k$ to be the mixing proportion corresponding to the unrepresented component u . In this section we shall take $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K, \beta_u)$. Let $\gamma_r = \gamma/L$ and $\gamma_u = \gamma(L - K)/L$ so that we have $\boldsymbol{\beta} \sim \text{Dir}(\gamma_r, \dots, \gamma_r, \gamma_u)$. We also only need to keep track of the counts n_{jk} for $1 \leq k \leq K$, and set $n_{ju} = 0$.

Integrating out $\boldsymbol{\pi}$. Since $\boldsymbol{\pi}$ is Dirichlet distributed and the Dirichlet distribution is conjugate to the multinomial, we may analytically integrate out $\boldsymbol{\pi}$, giving the following conditional probability of \mathbf{z} given $\boldsymbol{\beta}$:

$$p(\mathbf{z}|\boldsymbol{\beta}) = \prod_{j=1}^J \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} \prod_{k=1}^K \frac{\Gamma(\alpha_0 \beta_k + n_{jk})}{\Gamma(\alpha_0 \beta_k)}. \quad (34)$$

Sampling \mathbf{z} . From (34), the prior probability for $z_{ji} = k$ given \mathbf{z}^{-ji} and $\boldsymbol{\beta}$ is simply $\alpha_0 \beta_k + n_{jk}^{-ji}$ for each $k = 1, \dots, K, u$. Combined with the likelihood of x_{ji} we get the conditional probability for z_{ji} :

$$p(z_{ji} = k | \mathbf{z}^{-ji}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}) \propto (\alpha_0 \beta_k + n_{jk}^{-ji}) f(x_{ji} | \theta_k) \quad \text{for } k = 1, \dots, K, u. \quad (35)$$

where θ_u is sampled from its prior H . If as a result of sampling z_{ji} a represented component is left with no observations associated with it, we may remove it from the represented list of components. If on the other hand the new value for z_{ji} is u , we need to instantiate a new component for it. To do so, we increment K by 1, set $z_{ji} \leftarrow K$, $\theta_K \leftarrow \theta_u$, and we draw $b \sim \text{Beta}(1, \gamma)$ and set $\beta_K \leftarrow b\beta_u$, $\beta_u \leftarrow (1 - b)\beta_u$.

The updates to β_K and β_u can be understood as follows. We instantiate a new component by obtaining a sample, with index variable k_u , from the pool of unrepresented components. That is, we choose component $k_u = k$ with probability $\beta_k / \sum \beta_k = \beta_k / \beta_u$ for each $k = K + 1, \dots, L$. Notice, however, that $(\beta_{K+1}/\beta_u, \dots, \beta_L/\beta_u) \sim \text{Dir}(\gamma_r, \dots, \gamma_r)$. It is now an exercise in standard properties of the Dirichlet distribution to show that $\beta_{k_u}/\beta_u \sim \text{Beta}(1 + \gamma_r, \gamma_u - \gamma_r)$. As $L \rightarrow \infty$

$s(n, m)$	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$
$n = 0$	1	0	0	0	0
$n = 1$	0	1	0	0	0
$n = 2$	0	1	1	0	0
$n = 3$	0	2	3	1	0
$n = 4$	0	6	11	6	1

Table 1: Table of the unsigned Stirling numbers of the first kind.

this is $\text{Beta}(1, \gamma)$. Hence this new component has weight $b\beta_u$ where $b \sim \text{Beta}(1, \gamma)$, while the weights of the unrepresented components sum to $(1 - b)\beta_u$.

Sampling β . We use an auxiliary variable method for sampling β . Notice that in the likelihood term (34) for β , the variables β_k appear as arguments of Gamma functions. However the ratios of Gamma functions are polynomials in $\alpha_0\beta_k$, and can be expanded as follows:

$$\frac{\Gamma(n_{jk} + \alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k)} = \prod_{m_{jk}=1}^{n_{jk}} (m_{jk} - 1 + \alpha_0\beta_k) = \sum_{m_{jk}=0}^{n_{jk}} s(n_{jk}, m_{jk})(\alpha_0\beta_k)^{m_{jk}}, \quad (36)$$

where $s(n_{jk}, m_{jk})$ is the coefficient of $(\alpha_0\beta_k)^{m_{jk}}$. In fact, the $s(n_{jk}, m_{jk})$ terms are unsigned Stirling numbers of the first kind. Table 1 presents some values of $s(n, m)$. We have by definition that $s(0, 0) = 1$, $s(n, 0) = 0$, $s(n, n) = 1$ and $s(n, m) = 0$ for $m > n$. Other entries of the table can be computed as $s(n + 1, m) = s(n, m - 1) + ns(n, m)$. We introduce $\mathbf{m} = (m_{jk} : \text{all } j, k)$ as auxiliary variables to the model. Plugging (36) into (34) and including the prior for β , the distribution over \mathbf{z} , \mathbf{m} and β is:

$$q(\mathbf{z}, \mathbf{m}, \beta) = \frac{\Gamma(\gamma)}{\Gamma(\gamma_r)^K \Gamma(\gamma_u)} \left(\prod_{j=1}^J \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} \right) \beta_u^{\gamma_u - 1} \prod_{k=1}^K \beta_k^{\gamma_r - 1} \prod_{j=1}^J (\alpha_0\beta_k)^{m_{jk}} s(n_{jk}, m_{jk}). \quad (37)$$

It can be verified that $\sum_{\mathbf{m}} q(\mathbf{z}, \mathbf{m} | \beta) = p(\mathbf{z} | \beta)$. Finally, to obtain β given \mathbf{z} , we simply iterate sampling between \mathbf{m} and β using the conditional distributions derived from (37). In the limit $L \rightarrow \infty$ the conditional distributions are simply:

$$q(m_{jk} = m | \mathbf{z}, \mathbf{m}^{-jk}, \beta) \propto s(n_{jk}, m)(\alpha_0\beta_k)^m \quad (38)$$

$$q(\beta | \mathbf{z}, \mathbf{m}) \propto \beta_u^{\gamma_u - 1} \prod_{k=1}^K \beta_k^{\sum_j m_{jk} - 1}. \quad (39)$$

The conditional distributions of m_{jk} are easily computed since they can only take on values between zero and n_{jk} , and $s(n, m)$ are easily computed and can optionally be stored at little cost. Given \mathbf{m} the conditional distribution of β is simply a Dirichlet distribution with weights $(\sum_j m_{j1}, \dots, \sum_j m_{jK}, \gamma)$.

Sampling θ in this scheme is the same as for the Chinese restaurant franchise scheme. Each θ_k is updated using its posterior given \mathbf{z} and \mathbf{x} :

$$p(\theta_k | \mathbf{z}, \beta, \theta^{-k}, \mathbf{x}) \propto h(\theta_k) \prod_{j: z_{ji}=k} f(x_{ji} | \theta_k) \quad \text{for } k = 1, \dots, K. \quad (40)$$

5.2.1 Conjugacy between β and m

The derivation of the auxiliary variable sampling scheme reveals an interesting conjugacy between the weights β and the auxiliary variables m . First notice that the posterior for π given z and β is

$$p((\pi_{j1}, \dots, \pi_{jK}, \pi_{ju})_{j=1}^J | z, \beta) \propto \prod_{j=1}^J \pi_{ju}^{\alpha_0 \beta_u - 1} \prod_{k=1}^K \pi_{jk}^{\alpha_0 \beta_k + n_{jk} - 1}, \quad (41)$$

where $\pi_{ju} = \sum_{k=K+1}^{\infty} \pi_{jk}$ is the total weight for the unrepresented components. This describes the basic conjugacy between π_j and n_{jk} 's in the case of the ordinary Dirichlet process, and is a direct result of the conjugacy between Dirichlet and multinomial distributions (Ishwaran and Zarepour, 2002). This conjugacy has been used to improve the sampling scheme for stick-breaking generalizations of the Dirichlet process (Ishwaran and James, 2001).

On the other hand, the conditional distribution (39) suggests that the β weights are conjugate in some manner to the auxiliary variables m_{jk} . This raises the question of the meaning of the m_{jk} variables. The conditional distribution (38) of m_{jk} gives us a hint.

Consider again the Chinese restaurant franchise, in particular the probability that we obtain m tables corresponding to component k in mixture j , given that we know the component to which each data item in mixture j is assigned (i.e., we know z), and we know β (i.e., we are given the sample G_0). Notice that the number of tables in fact plays no role in the likelihood since we already know which component each data item comes from. Furthermore, the probability that i is assigned to some table t such that $k_{jt} = k$ is

$$p(t_{ji} = t | k_{jt} = k, \mathbf{m}_{ji}, \beta, \alpha_0) \propto n_{jt}^{-i}, \quad (42)$$

while the probability that i is assigned a new table under component k is

$$p(t_{ji} = t^{\text{new}} | k_{jt^{\text{new}}} = k, \mathbf{m}_{ji}, \beta, \alpha_0) \propto \alpha_0 \beta_k. \quad (43)$$

This shows that the distribution over the assignment of observations to tables is in fact equal to the distribution over the assignment of observations to components in an ordinary Dirichlet process with concentration parameter $\alpha_0 \beta_k$, given that n_{jk} samples are observed from the Dirichlet process. Antoniak (1974) has shown that this induces a distribution over the number of components:

$$p(\# \text{ components} = m | n_{jk} \text{ samples}, \alpha_0 \beta_k) = s(n_{jk}, m) (\alpha_0 \beta_k)^m \frac{\Gamma(\alpha_0 \beta_k)}{\Gamma(\alpha_0 \beta_k + n_{jk})}, \quad (44)$$

which is exactly (38). Hence m_{jk} is the number of tables assigned to component k in mixture j . This comes as no surprise, since the tables correspond to samples from G_0 so the number of samples equal to some distinct value (the number of tables under the corresponding component) should be conjugate to the weights β .

5.3 Posterior sampling for concentration parameters

We can update the concentration parameters γ and α_0 of the hierarchical Dirichlet process using straightforward extensions of analogous techniques for Dirichlet processes. Consider the Chinese

restaurant franchise representation. The concentration parameter α_0 governs the distribution over the number of ψ_{jt} 's in each mixture independently. As noted in Section 5.2.1 this is given by:

$$p(T_1, \dots, T_J | \alpha_0, n_1, \dots, n_J) = \prod_{j=1}^J s(n_j, T_j) \alpha_0^{T_j} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)}. \quad (45)$$

Further, α_0 does not govern other aspects of the joint distribution, hence given T_j the observations are independent of α_0 . Therefore (45) gives the likelihood term for α_0 . Together with the prior for α_0 and the current sample for T_j we can now derive updates for α_0 . In the case of a single mixture model ($J = 1$), Escobar and West (1995) proposed a gamma prior and derived an auxiliary variable update for α_0 , while Rasmussen (2000) observed that (45) is log-concave in α_0 and proposed using adaptive rejection sampling (Gilks and Wild, 1992) instead. Both can be adapted to the case $J > 1$.

The adaptive rejection sampler of Rasmussen (2000) can be directly applied to the case $J > 1$ since the conditional distribution of α_0 is still log-concave. The auxiliary variable method of Escobar and West (1995) requires a slight modification for the case $J > 1$. Assume that the prior for α_0 is a gamma distribution with parameters a and b . For each j we can write

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} = \int_0^1 w_j^{\alpha_0} (1 - w_j)^{n_j - 1} \left(1 + \frac{n_j}{\alpha_0}\right) dw_j. \quad (46)$$

In particular, we define auxiliary variables $\mathbf{w} = (w_j)_{j=1}^J$ and $\mathbf{s} = (s_j)_{j=1}^J$ where each w_j is a variable taking on values in $[0, 1]$, and each s_j is a binary $\{0, 1\}$ variable, define the following distribution:

$$q(\alpha_0, \mathbf{w}, \mathbf{s}) \propto \alpha_0^{a-1+\sum_{j=1}^J T_j} e^{\alpha_0 b} \prod_{j=1}^J w_j^{\alpha_0} (1 - w_j)^{n_j - 1} \left(\frac{n_j}{\alpha_0}\right)^{s_j}. \quad (47)$$

Now marginalizing q to α_0 gives the desired conditional distribution for α_0 . Hence q defines an auxiliary variable sampling scheme for α_0 . Given \mathbf{w} and \mathbf{s} we have:

$$q(\alpha_0 | \mathbf{w}, \mathbf{s}) \propto \alpha_0^{a-1+\sum_{j=1}^J T_j - s_j} e^{\alpha_0 (b - \sum_{j=1}^J \log w_j)}, \quad (48)$$

which is gamma distributed with parameters $a + \sum_{j=1}^J T_j - s_j$ and $b - \sum_{j=1}^J \log w_j$. Given α_0 , the w_j and s_j are conditionally independent, with distributions:

$$q(w_j | \alpha_0) \propto w_j^{\alpha_0} (1 - w_j)^{n_j - 1} \quad (49)$$

$$q(s_j | \alpha_0) \propto \left(\frac{n_j}{\alpha_0}\right)^{s_j}, \quad (50)$$

which are beta and binomial distributions respectively. This completes the auxiliary variable sampling scheme for α_0 . We prefer the auxiliary variable sampling scheme as it is easier to implement and typically mixes quickly (within 20 iterations).

Given the total number $T = \sum_j T_j$ of ψ_{jt} 's, the concentration parameter γ governs the distribution over the number of components K :

$$p(K | \gamma, T) = s(T, K) \gamma^K \frac{\Gamma(\gamma)}{\Gamma(\gamma + T)}. \quad (51)$$

Again the observations are independent of γ given T and K , hence we may apply the techniques of Escobar and West (1995) or Rasmussen (2000) directly to sampling γ .

5.4 Comparison of sampling schemes

We have described two different sampling schemes for hierarchical Dirichlet process mixture models. In the next section we present an example that indicates that neither of the two sampling schemes dominates the other. Here we provide some intuition regarding the dynamics involved in the sampling schemes.

In the Chinese restaurant franchise sampling scheme, we instantiate all the tables involved in the model, we assign data items to tables, and assign tables to mixture components. The assignment of data items to mixture components is indirect. This offers the possibility of speeding up convergence because changing the component assignment of one table offers the possibility of changing the component memberships of multiple data items. This is akin to split-and-merge techniques in Dirichlet process mixture modeling (Jain and Neal, 2000). The difference is that this is a Gibbs sampling procedure while split-and-merge techniques are based on Metropolis-Hastings updates.

Unfortunately, unlike split-and-merge methods, we do not have a ready way of assigning data items to tables within the same component. This is because the assignments of data items to tables is a consequence of the *prior* clustering effect of a Dirichlet process with n_{jk} samples. As a result, we expect that—with high-dimensional, large data sets, where tables will typically have large numbers of data items and components are well-separated—the probability that we have a successful reassignment of a table to another previously seen component is very small. However this intuition is still to be verified experimentally.

In the auxiliary variable sampling scheme, we have a direct assignment of data items to components, and tables are only indirectly represented via the number of tables assigned to each component in each mixture. As a result data items can only switch components one at a time. This is potentially slower than the Chinese restaurant franchise method. However, the sampling of the number of tables per component is very efficient, since it involves an auxiliary variable, and we have a simple form for the conditional distributions.

It is of interest to note that combinations of the two schemes may yield an even more efficient sampling scheme. We start from the auxiliary variable scheme. Given β , instead of sampling the number of tables under each component directly using (38), we may generate an assignment of data items to tables under each component using the Pólya urn scheme (this is a one shot procedure given by (6), and is not a Markov chain). This follows from the conjugacy arguments in Section 5.2.1. A consequence is that we now have the number of tables in that component, which can be used to update β . In addition, we also have the assignment of data items to tables, and tables to components, so we may consider changing the component assignment of each table as in the Chinese restaurant franchise scheme.

5.5 Bars problem

We examined the convergence properties of both the Chinese restaurant franchise and the auxiliary variable sampling methods on a small problem previously studied in Blei et al. (2004). The data set consists of 40 groups, each of which contains 100 observations. The observations in each group are generated from a mixture of between two and four components (from a total of 10 possible components). Each observation can take on one of 25 values, and each component distribution can be visualized as bars in a 5×5 image. Figure 6(a) shows the empirical distribution over ten of the groups, while Figure 6(b) shows the ten component distributions (in fact these were inferred from the data).

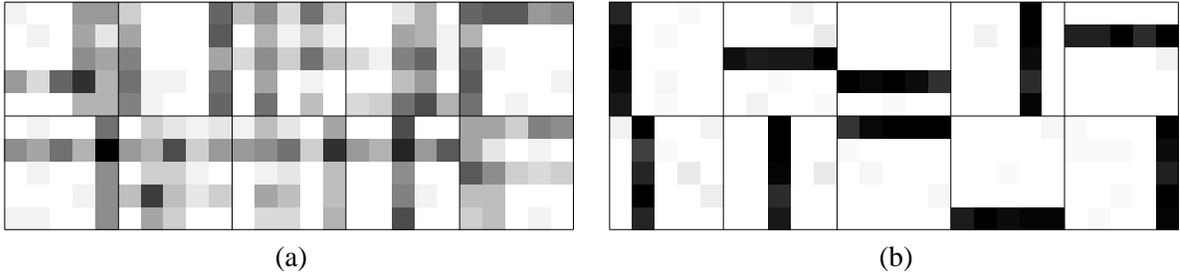


Figure 6: Distributions over observations can be visualized using a 5×5 image, with each pixel corresponding to each possible value of the observation. Black means a probability of at least 0.2, while white is probability of 0. (a) Ten examples of the empirical distribution in each group. (b) The component distributions for the ten components at the end of Markov chain sampling (using auxiliary variables).

We constructed a hierarchical Dirichlet process mixture model using a multinomial distribution for $F(\cdot)$, and a conjugate Dirichlet prior for H (with symmetric weights of $1/5$). The hyperparameters γ and α_0 are set equal to one. We ran both the Chinese restaurant franchise and the auxiliary variable Markov chain sampling methods on the same generated data set. We find that the Markov chains converge rapidly for this simple problem. After an initial burn-in period (200 iterations), we collected information for 5000 iterations. In particular, we kept track of the number of represented components K , and the total number of tables across the data set $m = \sum_{jk} m_{jk}$.

In Figures 7(a) and 7(b) we plot the autocorrelation function for K and m respectively. For the number of components the Chinese restaurant franchise method seems to produce uncorrelated samples faster than the auxiliary variable method. This conforms with the intuition presented in Section 5.4. Because we are sampling the component memberships of each individual table in the Chinese restaurant franchise, the component memberships of large numbers of observations can change all at once. This is reflected in the low autocorrelation of the number of components. On the other hand, in the auxiliary variable method, component memberships of observations can only be changed one at a time hence this method produces higher autocorrelation. However as discussed in 5, in higher dimensions and with more observations we expect the component memberships of tables to be very rigid and so we expect the auxiliary variable method to perform comparably to the Chinese restaurant franchise method.

For the number of tables, the auxiliary variable method produces uncorrelated samples faster. This is again expected because given β we are able to sample directly the number of tables m_{jk} assigned to each component in each mixture.

6 Related models

6.1 The infinite hidden Markov model

In work that served as an inspiration for the ideas developed here, Beal et al. (2002) discussed an architecture known as the *infinite hidden Markov model*, in which the number of hidden states of a hidden Markov model is allowed to be countably infinite via the formalism of the Dirichlet process.

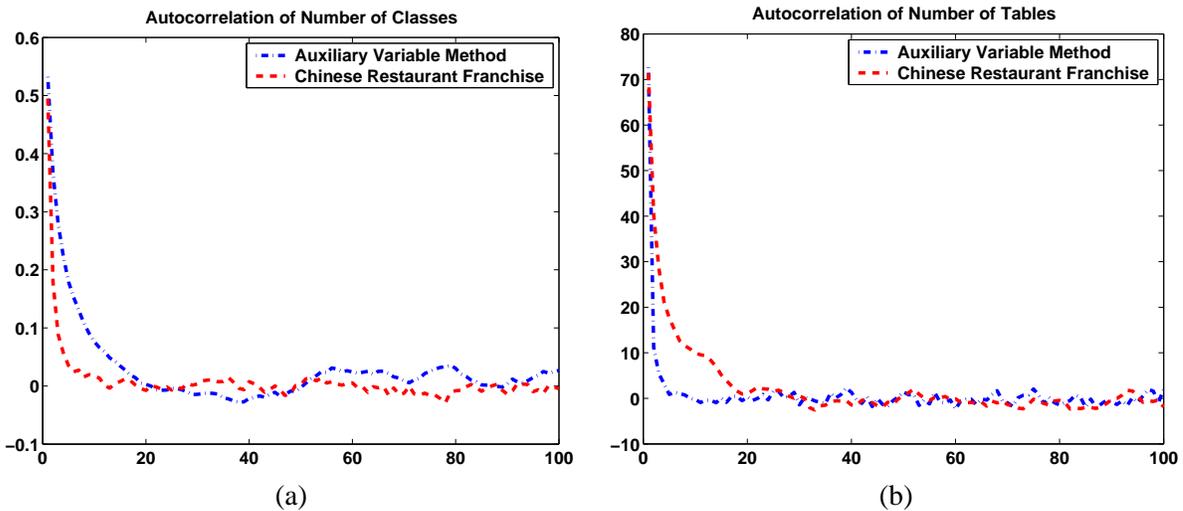


Figure 7: The autocorrelation for (a) the number of components and (b) the number of tables.

Indeed, Beal et al. (2002) defined a notion of “hierarchical Dirichlet process” for this architecture, but their “hierarchical Dirichlet process” is not hierarchical in the Bayesian sense—involving a distribution on the parameters of a Dirichlet process—but is instead a description of a coupled set of urn models. In this section we briefly review this construction, and relate it to our formulation.

Beal et al. (2002) described a two-level procedure for determining the transition probabilities of a Markov chain with an unbounded number of states. At the first level, the probability of transitioning from a state u to a state v is proportional to the number of times the same transition is observed at other time steps, while with probability proportional to α_0 an “oracle” process is invoked. At this second level, the probability of transitioning to state v is proportional to the number of times state v has been chosen by the oracle (regardless of the previous state), while the probability of transitioning to a novel state is proportional to γ . The intended role of the oracle is to tie together the transition models so that they have destination states in common, in much the same way that the baseline distribution G_0 ties together the group-specific mixture components in the hierarchical Dirichlet process.

To see how this two-level urn model can be justified, let us formulate the infinite hidden Markov model within the hierarchical Dirichlet process framework of the current paper. We do so by assigning observations to groups, where the groups are indexed by the value of the previous state variable. We thus treat the next-state and emission distributions as defining group-specific mixtures. This leads to the hierarchical Dirichlet process representation shown in Figure 8. The parameters in this representation have the following distributions:

$$\beta \mid \gamma \sim \text{Stick}(\gamma) \quad \pi_k \mid \alpha_0, \beta \sim \text{DP}(\alpha_0, \beta) \quad \theta_k \mid H \sim H, \quad (52)$$

for each $k = 1, 2, \dots$, while for each time step $t = 1, \dots, T$ the state and observation distributions are:²

$$v_t \mid v_{t-1}, (\pi_k)_{k=1}^{\infty} \sim \pi_{v_{t-1}} \quad y_t \mid v_t, (\theta_k)_{k=1}^{\infty} \sim F(\theta_{v_t}), \quad (53)$$

²We assume for simplicity that there is a distinguished initial state v_0 .

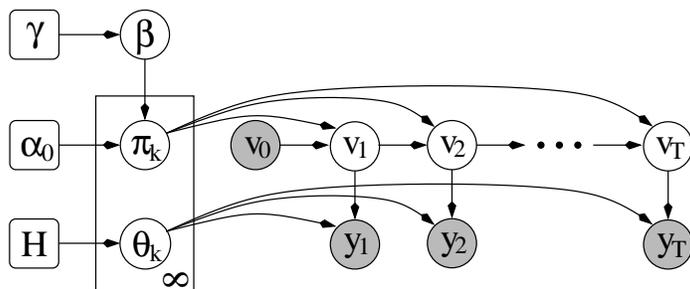


Figure 8: A hierarchical Bayesian model for the infinite hidden Markov model.

As discussed in Section 5, this stick-breaking representation leads to a Markov chain Monte Carlo sampling scheme involving auxiliary variables. On the other hand, one can also consider the Chinese restaurant franchise representation of this model. It turns out that this representation is equivalent to the coupled urn model of Beal et al. (2002). Unfortunately, the Chinese restaurant franchise representation is awkward for setting up an inference algorithm in this setting, involving substantial bookkeeping, and indeed Beal et al. (2002) did not present a Markov chain Monte Carlo inference algorithm for the infinite hidden Markov model, proposing instead a heuristic approximation to Gibbs sampling. The auxiliary variable approach that we presented in Section 5 turns out to lead to a more straightforward algorithm in this case, and it is this algorithm that we propose for inference in the infinite hidden Markov model.

6.2 Analysis of densities

Tomlinson and Escobar (2003) presented the *analysis of densities* (AnDe) model, a hierarchical Bayesian approach to modeling collections of densities. Formally their model is a more general model than the hierarchical Dirichlet process presented here, in that the model involves a collection of Dirichlet processes G_j in which the base density G_0 is drawn from a *mixture* of Dirichlet processes, not a single underlying Dirichlet process. Measures drawn from a mixture of Dirichlet processes are not discrete, which is clearly appropriate if the goal is to model densities. Our goal is different—we are explicitly interested in clustering. The discreteness of draws from the Dirichlet process is thus an essential feature of our approach. Moreover, while inference in the AnDe model is a relatively straightforward application of general Markov chain Monte Carlo methods for Dirichlet process mixture models, the clustering in the hierarchical Dirichlet process creates complications. Indeed, the thrust of our paper has been to address those complications head-on, via the stick-breaking representation, the Chinese restaurant franchise representation, and the resulting Markov chain Monte Carlo algorithms described in Section 5.

7 Discussion

We have described a nonparametric approach to the modeling of groups of data, where each group is characterized by a mixture model, and where it is desirable to allow mixture components to be shared between groups. We have proposed a hierarchical Bayesian solution to this problem, in

which a set of Dirichlet processes are coupled via their base measure, which is itself distributed according to a Dirichlet process.

We have described three different representations that capture aspects of the hierarchical Dirichlet process. In particular, we described a stick-breaking representation that describes the random measures explicitly, a representation of marginals in terms of an urn model that we referred to as the “Chinese restaurant franchise,” and a representation of the process in terms of the infinite limit of finite mixture models.

These representations led to the formulation of two Markov chain Monte Carlo sampling schemes for posterior inference under hierarchical Dirichlet process mixtures. The first scheme is based directly on the Chinese restaurant franchise representation, while the second scheme is an auxiliary variable method that represents the stick-breaking weights β explicitly. In a simple experiment, we saw that neither scheme dominates the other—both schemes have their strengths and weaknesses.

Is it of interest to consider additional levels of the hierarchy? Practical applications of hidden Markov models often consider sets of sequences, and treat these sequences as exchangeable at the level of sequences. Thus, in applications to speech recognition, a hidden Markov model for a given word in the vocabulary is generally trained via replicates of that word being spoken. If we wish to allow unbounded sets of states in such a setting, we are naturally led to a model in which multiple hidden Markov chains must be coupled. This is accommodated naturally in our formalism by considering an additional level of the Bayesian hierarchy, and allowing a master Dirichlet process to couple the chains.

Acknowledgments

We would like to acknowledge support for this project from Darpa under the CALO program and from a grant from the Intel Corporation. MJB is funded by a PREA grant to Professor Radford Neal at the University of Toronto.

References

- D. Aldous. Exchangeability and related topics. In *École d’été de probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin, 1985.
- C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 1974.
- M. J. Beal, Z. Ghahramani, and C.E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14, 2002.
- D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.
- D.M. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems*, 2004.
- D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

- E.H. Davidson. *Genomic Regulatory Systems*. Academic Press, New York, 2001.
- M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.
- T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2): 209–230, 1973.
- W.R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41: 337–348, 1992.
- P. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28:355–377, 2001.
- H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.
- H. Ishwaran and M. Zarepour. Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269–283, 2002.
- S. Jain and R.M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Technical Report 2003, Department of Statistics, University of Toronto, 2000.
- S.N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.
- R.M. Neal. Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, volume 11, pages 197–211, 1992.
- R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. Technical Report 9815, Department of Statistics, University of Toronto, 1998.
- R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.
- G.P. Patil and C. Taillie. Diversity as a concept and its implications for random communities. *Bulletin of the International Statistical Institute*, 47:497–515, 1977.
- J. Pitman. *Combinatorial Stochastic Processes: Notes for St. Flour Summer School*. 2002.
- L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 1989.
- C.E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, volume 12, 2000.
- G. Salton and M.J. McGill. *An Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
- G. Tomlinson and M. Escobar. Analysis of densities, 2003. Talk given at the Joint Statistical Meeting.