Hierarchical Dirichlet Processes

Yee Whye Teh ywteh@eecs.berkeley.edu
Computer Science Division, University of California at Berkeley,
Berkeley CA 94720-1776, USA

Michael I. Jordan jordan@eecs.berkeley.edu
Computer Science Division and Department of Statistics,
University of California at Berkeley, Berkeley CA 94720-1776, USA

Matthew J. Beal mbeal@cse.buffalo.edu
Department of Computer Science & Engineering,
State University of New York at Buffalo, Buffalo NY 14260-2000, USA

David M. Blei blei@eecs.berkeley.edu
Computer Science Division, University of California at Berkeley,
Berkeley CA 94720-1776, USA

October 8, 2004

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 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, and describe applications to problems in information retrieval and text modelling.

Keywords: clustering, mixture models, nonparametric Bayesian statistics, hierarchical models, Markov chain Monte Carlo

1 INTRODUCTION

A recurring theme in statistics is the need to separate observations into groups, and yet allow the groups to remain linked—to "share statistical strength." In the Bayesian formalism such sharing is achieved naturally via hierarchical modeling—parameters are shared among groups, and the randomness of the parameters induces dependencies among the groups. Estimates based on the posterior distribution exhibit "shrinkage."

In the current paper we explore a hierarchical approach to the problem of model-based clustering. We assume that the data are subdivided into a set of J groups, and that within each group we wish to find clusters that capture latent structure in the data assigned to that group. The number of clusters within each group is unknown and is to be inferred. Moreover, in a sense that we make precise, we wish to allow clusters to be shared among the groups.

An example of the kind of problem that motivates us can be found in genetics. Consider a set of k binary markers (e.g., single nucleotide polymorphisms or "SNPs") in a localized region of the human genome. While an individual human could exhibit any of 2^k different patterns of markers on a single chromosome, in real populations only a small subset of such patterns—*haplotypes*—are actually observed (Gabriel et al. 2002). Given a meiotic model for the combination of a pair of haplotypes into a *genotype* during mating, and given a set of observed genotypes in a sample from a human population, it is of great interest to identify the underlying haplotypes (Stephens et al. 2001). Now consider an extension of this problem in which the population is divided into a set of groups; e.g., African, Asian and European subpopulations. We may not only want to discover the sets of haplotypes within each subpopulation, but we may also wish to discover which haplotypes are shared between subpopulations. The identification of such haplotypes would have significant implications for the understanding of the migration patterns of ancestral populations of humans.

As a second example, consider the problem of the modeling of relationships among sets of documents in the field of information retrieval (IR). In IR, documents are generally modeled under an exchangeability assumption—the so-called "bag of words assumption"—in which the order of words in a document is ignored (Salton and McGill 1983). It is also common to view the words in a document as arising from a number of latent clusters or "topics," where a topic is generally modeled as a probability distribution on words from some basic vocabulary (Blei et al. 2003). Thus, in a document concerned with university funding the words in the document might be drawn from the topics "education" and "finance." If we now consider a corpus of such documents, we may wish to allow topics to be shared among the documents in the corpus. For example, if the corpus also contains a document concerned with university football, the topics may be "education" and "sports," and we would want the former topic to be related to that discovered in the analysis of the document on university funding.

Moreover, we may want to extend the model to allow for multiple corpora. For example, documents in scientific journals are often grouped into themes (e.g., "empirical process theory," "multivariate statistics," "survival analysis"), and it would be of interest to discover to what extent the latent topics that are shared among documents are also shared across these groupings. Thus in general we wish to consider the sharing of clusters across multiple, nested groupings of data.

Our approach to the problem of sharing clusters among multiple, related groups is a nonparametric Bayesian approach, reposing on the *Dirichlet process* (Ferguson 1973). The Dirichlet process $\mathrm{DP}(\alpha_0, G_0)$ is a measure on measures. It has two parameters, a *scaling parameter* $\alpha_0 > 0$ and a *base measure* G_0. An explicit representation of a draw from a Dirichlet process (DP) was given by

Sethuraman (1994), who showed that if $G \sim \mathrm{DP}(\alpha_0, G_0)$, then with probability one:

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

where the θ_k are independent random variables distributed according to G_0, where δ_{θ_k} is an atom at θ_k, and where the "stick-breaking weights" β_k are also random and depend on the parameter α_0 (the definition of the β_k is provided in Section 3.1).

The representation in (1) shows that draws from a DP are discrete (with probability one). The discrete nature of the DP makes it unsuitable for general applications in Bayesian nonparametrics, but it is well suited for the problem of placing priors on mixture components in mixture modeling. The idea is basically to associate a mixture component with each atom in G. Introducing indicator variables to associate data points with mixture components, the posterior distribution yields a probability distribution on partitions of the data. A number of authors have studied such *Dirichlet process mixture models* (Antoniak 1974; 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 setting in which the data are subdivided into J groups. Given our goal of solving a clustering problem within each group, we consider a set of random measures, G_j for $j = 1, \ldots, J$, where G_j is distributed according to a group-specific Dirichlet process $\mathrm{DP}(\alpha_{0j}, G_{0j})$. To link these clustering problems, we link the group-specific DPs. Many authors have considered ways to induce dependencies among multiple DPs via links among the parameters G_{0j} and/or α_{0j} (Cifarelli and Regazzini 1978; MacEachern 1999; Tomlinson and Escobar 2003; Müller et al. 2004; De Iorio et al. 2004; Kleinman and Ibrahim 1998; Mallick and Walker 1997; Ishwaran and James 2004). Focusing on the G_{0j}, one natural proposal is a hierarchy in which the measures G_j arise as conditionally independent draws from a single underlying Dirichlet process $\mathrm{DP}(\alpha_0, G_0(\tau))$, where $G_0(\tau)$ is a parametric distribution with random parameter τ (Carota and Parmigiani 2002; Fong et al. 2002; Muliere and Petrone 1993). Integrating over τ induces dependencies among the DPs.

That this simple hierarchical approach will not solve our problem can be observed by considering the case in which $G_0(\tau)$ is absolutely continuous with respect to Lebesgue measure for almost all τ (e.g., G_0 is Gaussian with mean τ). In this case, given that the draws G_j arise as conditionally independent draws from $G_0(\tau)$, they necessarily have no atoms in common (with probability one). Thus, although clusters arise *within* each group via the discreteness of draws from a DP, the atoms associated with the different groups are different and there is no sharing of clusters *between* groups. This problem can be skirted by assuming that G_0 lies in a discrete parametric family, but such an assumption would be overly restrictive.

Our proposed solution to the problem is straightforward—to force G_0 to be discrete and yet have broad support we consider a nonparametric hierarchical model in which G_0 is itself a draw from a Dirichlet process $\mathrm{DP}(\gamma, H)$. This restores flexibility in that the modeler can choose H to be continuous or discrete. In either case, with probability one, G_0 is discrete and has a stick-breaking representation as in (1). The atoms θ_k are shared among the multiple DPs, yielding the desired sharing of atoms among groups. In summary, we consider the hierarchical specification:

$$G_0 \mid \gamma, H \quad \sim \quad \mathrm{DP}(\gamma, H) \tag{2}$$

$$G_j \mid \alpha_0, G_0 \quad \sim \quad \mathrm{DP}(\alpha_0, G_0) \ , \tag{3}$$

which we refer to as a *hierarchical Dirichlet process*. The immediate extension to *hierarchical*

Dirichlet process mixture models yields our proposed formalism for sharing clusters among related clustering problems.

Related nonparametric approaches to linking multiple DPs have been discussed by a number of authors. Our approach is a special case of a general framework for "dependent Dirichlet processes" due to MacEachern (1999) and MacEachern et al. (2001). In this framework the random variables β_k and θ_k in (1) are general stochastic processes (i.e., indexed collections of random variables); this allows very general forms of dependency among DPs. Our hierarchical approach fits into this framework—endow the stick-breaking weights β_k in (1) with a second subscript indexing the groups $j = 1, \ldots, J$, and view the weights β_{kj} as dependent for each fixed value of k. Indeed, as we show in Section 4, the definition in (3) yields a specific, canonical form of dependence among the weights β_{kj}.

Our approach is also a special case of a framework referred to as *analysis of densities* (AnDe) by Tomlinson and Escobar (2003). The AnDe model is a hierarchical model for multiple DPs in which the common base measure G_0 is random, but rather than treating G_0 as a draw from a DP, as in our case, it is treated as a draw from a mixture of DPs. The resulting G_0 is continuous in general (Antoniak 1974), which, as we have discussed, is ruinous for our problem of sharing clusters. It is an appropriate choice, however, for the problem addressed by Tomlinson and Escobar (2003), which is that of sharing statistical strength among multiple sets of density estimation problems. Thus, while the AnDe framework and our hierarchical DP framework are closely related formally, the inferential goal is rather different. Moreover, as we will see, our restriction to discrete G_0 has important implications for the design of efficient MCMC inference algorithms.

The terminology of "hierarchical Dirichlet process" has also been used by Müller et al. (2004) to describe a different notion of hierarchy than the one discussed here. These authors consider a model in which a coupled set of random measures G_j are defined as $G_j = \epsilon F_0 + (1 - \epsilon) F_j$, where F_0 and the F_j are draws from Dirichlet processes. This model provides an alternative approach to sharing clusters, one in which the shared clusters are given the same stick-breaking weights (those associated with F_0) in each of the groups. By contrast, in our hierarchical model, the draws G_j are based on the same underlying base measure G_0, but each draw assigns different stick-breaking weights to the shared atoms associated with G_0. Atoms can be "partially shared."

Finally, the terminology of "hierarchical Dirichlet process" has been used in yet a third way by Beal et al. (2002) in the context of a model known as the *infinite hidden Markov model*—a hidden Markov model with a countably infinite state space. The "hierarchical Dirichlet process" of Beal et al. (2002) is not, however, a hierarchy in the Bayesian sense—involving a distribution on the parameters of a distribution—but is instead an algorithmic description of a coupled set of urn models. We discuss this model in more detail in Section 7, where we show that the notion of hierarchical Dirichlet process presented here yields an elegant treatment of the infinite hidden Markov model.

In summary, the notion of hierarchical Dirichlet process that we explore here is a specific example of a dependency model for multiple Dirichlet processes, one specifically aimed at the problem of sharing clusters among related groups of data. It involves a simple Bayesian hierarchy—the base measure for a set of Dirichlet processes is itself distributed according to a Dirichlet process. While there are many ways to couple Dirichlet processes, we view this simple, canonical Bayesian hierarchy as particularly worthy of study. Note in particular the appealing recursiveness of the definition—a hierarchical Dirichlet process can be readily extended to multiple hierarchical levels. This is natural in applications. For example, in our application to document modeling, one level of hierarchy is needed to share clusters among multiple documents within a corpus, and second level of hierarchy is needed to share clusters among multiple corpora. Similarly, in the genetics example, it is of interest to consider nested subdivisions of populations according to various criteria (geographic, cultural, economic), and to consider the flow of haplotypes on the resulting tree.

As is the case with other nonparametric Bayesian methods, a significant component of the challenge in working with the hierarchical Dirichlet process is computational. To provide a general framework for designing procedures for posterior inference for the hierarchical Dirichlet process 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 present experimental results in Section 6 and present our conclusions in Section 8.

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 $j \in \{1, 2, \ldots, J\}$ index the J groups, and letting $\boldsymbol{x}_j = (x_{ji})_{i=1}^{n_j}$ 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J$ are exchangeable at the group level. Let $\boldsymbol{x} = (\boldsymbol{x}_j)_{j=1}^{J}$ denote the entire data set.

If each observation is drawn independently from a mixture model, then there is a 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 $\boldsymbol{\phi}_j = (\phi_{ji})_{i=1}^{n_j}$ associated with group j. We assume that the factors are conditionally independent given G_j. Thus we have the following probability model:

$$\begin{aligned} \phi_{ji} \mid G_j &\sim G_j && \text{for each } j \text{ and } i, \\ x_{ji} \mid \boldsymbol{\phi}_j &\sim F(\phi_{ji}) && \text{for each } j \text{ and } i, \end{aligned} \tag{4}$$

to augment the specification given in (3).

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

of Θ, the random vector $(G(A_1), \ldots, G(A_r))$ is distributed as a finite-dimensional Dirichlet distribution with parameters $(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r))$:

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)) \,. \tag{5}$$

We write $G \sim \mathrm{DP}(\alpha_0, G_0)$ 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 independent sequences of independent random variables $(\pi'_k)_{k=1}^{\infty}$ and $(\theta_k)_{k=1}^{\infty}$:

$$\pi'_k \mid \alpha_0, G_0 \sim \mathrm{Beta}(1, \alpha_0) \qquad\qquad \theta_k \mid \alpha_0, G_0 \sim G_0 \,, \tag{6}$$

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

$$\pi_k = \pi'_k \prod_{l=1}^{k-1} (1 - \pi'_l) \qquad\qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \,, \tag{7}$$

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 $\mathrm{DP}(\alpha_0, G_0)$.

It is important to note that the sequence $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ constructed by (6) and (7) satisfies $\sum_{k=1}^{\infty} \pi_k = 1$ with probability one. Thus we may interpret $\boldsymbol{\pi}$ as a random probability measure on the positive integers. For convenience, we shall write $\boldsymbol{\pi} \sim \mathrm{Stick}(\alpha_0)$ if $\boldsymbol{\pi}$ is a random probability measure defined by (6) and (7).

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 $\phi_1, \phi_2, \ldots$ be a sequence of i.i.d. random variables distributed according to G. That is, the variables $\phi_1, \phi_2, \ldots$ are conditionally independent given G, and hence exchangeable. Let us consider the successive conditional distributions of ϕ_i given $\phi_1, \ldots, \phi_{i-1}$, where G has been integrated out. Blackwell and MacQueen (1973) showed that these conditional distributions have the following simple form:

$$\phi_i \mid \phi_1, \ldots, \phi_{i-1}, \alpha_0, G_0 \sim \sum_{l=1}^{i-1} \frac{1}{i-1+\alpha_0} \delta_{\phi_l} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \,. \tag{8}$$

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 together with another ball of the

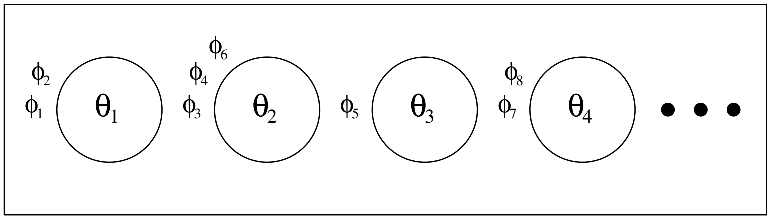

Figure 1: A depiction of a Chinese restaurant after eight customers have been seated. Customers (ϕ_i's) are seated at tables (circles) which correspond to the unique values θ_k.

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 $\theta_1, \ldots, \theta_K$ to be the distinct values taken on by $\phi_1, \ldots, \phi_{i-1}$, 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 (8) as

$$\phi_i \mid \phi_1, \ldots, \phi_{i-1}, \alpha_0, G_0 \;\sim\; \sum_{k=1}^{K} \frac{n_k}{i-1+\alpha_0} \delta_{\theta_k} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \,. \tag{9}$$

Using a somewhat different metaphor, the Pólya urn scheme is closely related to a distribution on partitions known as the *Chinese restaurant process* (Aldous 1985). This metaphor has turned out to 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 3.2.

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:

$$\begin{aligned} \phi_i \mid G &\;\sim\; G \\ x_i \mid \phi_i &\;\sim\; F(\phi_i) \,, \end{aligned} \tag{10}$$

where $F(\phi_i)$ denotes the distribution of the observation x_i given ϕ_i. The *factors* ϕ_i are conditionally independent given G, and 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 (7), 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 $\boldsymbol{\pi}$ (interpreting $\boldsymbol{\pi}$ as a random probability measure on

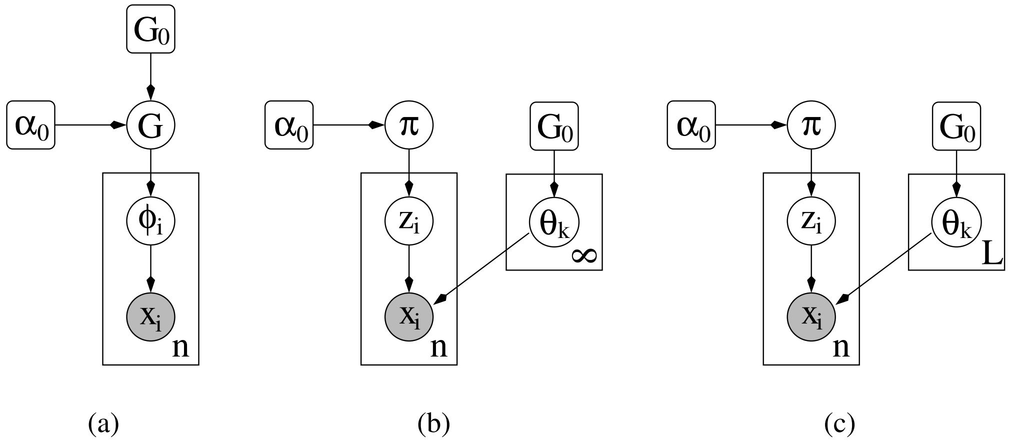

Figure 2: (a) A representation of a Dirichlet process mixture model as a graphical model. In the graphical model formalism, each node in the graph is associated with a random variable and joint probabilities are defined as products of conditional probabilities, where a conditional probability is associated 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 (note the L in place of the ∞).

the positive integers). Hence an equivalent representation of a Dirichlet process mixture is given by Figure 2(b), where the conditional distributions are:

$$\begin{aligned} \boldsymbol{\pi} \mid \alpha_0 &\sim \mathrm{Stick}(\alpha_0) & \qquad z_i \mid \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\ \theta_k \mid G_0 &\sim G_0 & \qquad x_i \mid z_i, (\theta_k)_{k=1}^{\infty} &\sim F(\theta_{z_i})\,. \end{aligned} \tag{11}$$

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 $\boldsymbol{\pi} = (\pi_1, \ldots \pi_L)$ denote the mixing proportions. Note that we previously 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 $L \to \infty$ these vectors are equivalent up to a random *size-biased permutation* of their entries (Patil and Taillie 1977).

We place a Dirichlet prior on $\boldsymbol{\pi}$ with symmetric parameters $(\alpha_0/L, \ldots, \alpha_0/L)$. 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 following model:

$$\begin{aligned} \boldsymbol{\pi} \mid \alpha_0 &\sim \mathrm{Dir}(\alpha_0/L, \ldots, \alpha_0/L) & z_i \mid \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\ \theta_k \mid G_0 &\sim G_0 & x_i \mid z_i, (\theta_k)_{k=1}^L &\sim F(\theta_{z_i}) \, . \end{aligned} \tag{12}$$

The corresponding 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 $L \to \infty$:

$$\int f(\phi) \, dG^L(\phi) \xrightarrow{\mathcal{D}} \int f(\phi) \, dG(\phi) \, . \tag{13}$$

A consequence of this is that the marginal distribution induced on the observations $x_1, \ldots, x_n$ 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 $(\Theta, \mathcal{B})$. 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:

$$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H) \, , \tag{14}$$

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:

$$G_j \mid \alpha_0, G_0 \sim \mathrm{DP}(\alpha_0, G_0) \, . \tag{15}$$

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 $(\boldsymbol{\phi}_j)_{j=1}^J$. 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 $\boldsymbol{\phi}_j$ 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. In this paper, 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. Eacah ϕ_{ji} is a factor corresponding to a single observation x_{ji}. The likelihood is given by:

$$\begin{aligned} \phi_{ji} \mid G_j &\sim G_j \\ x_{ji} \mid \phi_{ji} &\sim F(\phi_{ji}) \, . \end{aligned} \tag{16}$$

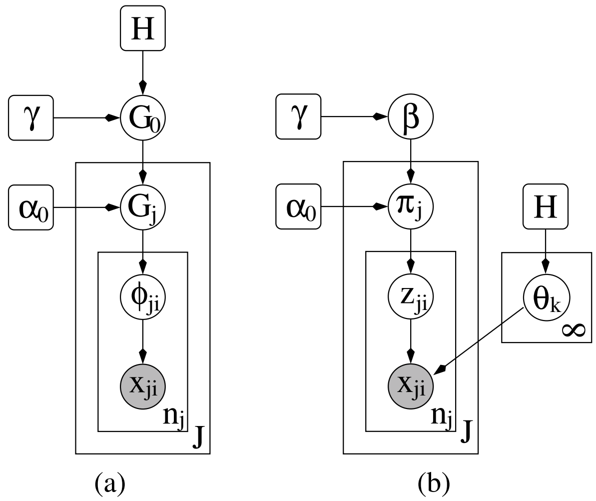

Figure 3: (a) A hierarchical Dirichlet process mixture model. (b) A alternative representation of a hierarchical Dirichlet process mixture model in terms of the stick-breaking construction.

This completes the definition of a *hierarchical Dirichlet process mixture model*. The corresponding graphical model is shown in Figure 4(a).

Notice that $(\phi_{ji})_{i=1}^{n_j}$ are exchangeable random variables if we integrate out G_j. Similarly, $(\boldsymbol{\phi}_j)_{j=1}^{J}$ 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 $(\boldsymbol{x}_j)_{j=1}^{J}$ is not violated by the hierarchical Dirichlet process mixture model.

The hierarchical Dirichlet process can readily be extended to more than two levels. That is, the base measure H can itself be a draw from a DP, and the hierarchy can be extended for as many levels as are deemed useful. In general, we obtain a tree in which a DP is associated with each node, in which the children of a given node are conditionally independent given their parent, and in which the draw from the DP at a given node serves as a base measure for its children. The atoms in the stick-breaking representation at a given node are thus shared among all descendant nodes, providing notion of shared clusters at multiple levels of resolution. The software for hierarchical Dirichlet process mixtures that we describe in Section 6—software which is publicly available—provides an implementation for arbitrary trees of this kind.

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:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \ , \tag{17}$$

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

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \ . \tag{18}$$

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

Let $(A_1, \ldots, A_r)$ be a measurable partition of Θ and let $K_l = \{k : \theta_k \in A_l\}$ for $l = 1, \ldots, r$. Note that $(K_1, \ldots, K_r)$ 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:

$$\begin{aligned}
(G_j(A_1), \ldots, G_j(A_r)) &\sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)) \\
\Rightarrow \left(\sum_{k \in K_1} \pi_{jk}, \ldots, \sum_{k \in K_r} \pi_{jk} \right) &\sim \mathrm{Dir}\left(\alpha_0 \sum_{k \in K_1} \beta_k, \ldots, \alpha_0 \sum_{k \in K_r} \beta_k \right) ,
\end{aligned} \tag{19}$$

for every finite partition of the positive integers. Hence each $\boldsymbol{\pi}_j$ is independently distributed according to $\mathrm{DP}(\alpha_0, \boldsymbol{\beta})$, where we interpret $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_j$ 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 takes 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 4(b) gives an equivalent representation of the hierarchical Dirichlet process mixture, with conditional distributions summarized here:

$$\begin{aligned}
\boldsymbol{\beta} \mid \gamma &\sim \mathrm{Stick}(\gamma) \\
\boldsymbol{\pi}_j \mid \alpha_0, \boldsymbol{\beta} &\sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta}) & z_{ji} \mid \boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\
\theta_k \mid H &\sim H & x_{ji} \mid z_{ji}, (\theta_k)_{k=1}^{\infty} &\sim F(\theta_{z_{ji}}) .
\end{aligned} \tag{20}$$

We now derive an explicit relationship between the elements of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_j$. Recall that the stick-breaking construction for Dirichlet processes defines the variables β_k in (17) as follows:

$$\beta'_k \sim \mathrm{Beta}(1, \gamma) \qquad\qquad \beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) . \tag{21}$$

Using (19), we show that the following stick-breaking construction produces a random probability measure $\boldsymbol{\pi}_j \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$:

$$\pi'_{jk} \sim \mathrm{Beta}\left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{l=1}^{k} \beta_l \right) \right) \qquad\qquad \pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl}) . \tag{22}$$

To derive (22), first notice that for a partition $(\{1, \ldots, k-1\}, \{k\}, \{k+1, k+2, \ldots\})$, (19) gives:

$$\left(\sum_{l=1}^{k-1} \pi_{jl}, \pi_{jk}, \sum_{l=k+1}^{\infty} \pi_{jl} \right) \sim \mathrm{Dir}\left(\alpha_0 \sum_{l=1}^{k-1} \beta_l, \alpha_0 \beta_k, \alpha_0 \sum_{l=k+1}^{\infty} \beta_l \right) . \tag{23}$$

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

$$\frac{1}{1 - \sum_{l=1}^{k-1} \pi_{jl}} \left(\pi_{jk}, \sum_{l=k+1}^{\infty} \pi_{jl} \right) \sim \mathrm{Dir}\left(\alpha_0 \beta_k, \alpha_0 \sum_{l=k+1}^{\infty} \beta_l \right) . \tag{24}$$

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 (22). Together with (21), (17) and (18), 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 $\theta_1, \ldots, \theta_K$ denote K i.i.d. random variables distributed according to H, and, for each j, we let $\psi_{j1}, \ldots, \psi_{jT_j}$ 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 $G_0 \sim \mathrm{DP}(\gamma, H)$), we are denoting them as distinct random variables.

First consider the conditional distribution for ϕ_{ji} given $\phi_{j1}, \ldots, \phi_{j\,i-1}$ and G_0, where G_j is integrated out. From (9), we have:

$$\phi_{ji} \mid \phi_{j1}, \ldots, \phi_{j\,i-1}, \alpha_0, G_0 \;\sim\; \sum_{t=1}^{T_j} \frac{n_{jt}}{i-1+\alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \,, \tag{25}$$

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 (9) again and writing the conditional distribution of ψ_{jt} directly:

$$\psi_{jt} \mid \psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{j\,t-1}, \gamma, H \;\sim\; \sum_{k=1}^{K} \frac{m_k}{\sum_k m_k + \gamma} \delta_{\theta_k} + \frac{\gamma}{\sum_k m_k + \gamma} H \,. \tag{26}$$

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 (25). If a new sample from G_0 is needed, we use (26) to obtain a new sample ψ_{jt} and set $\phi_{ji} = \psi_{jt}$.

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.

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.

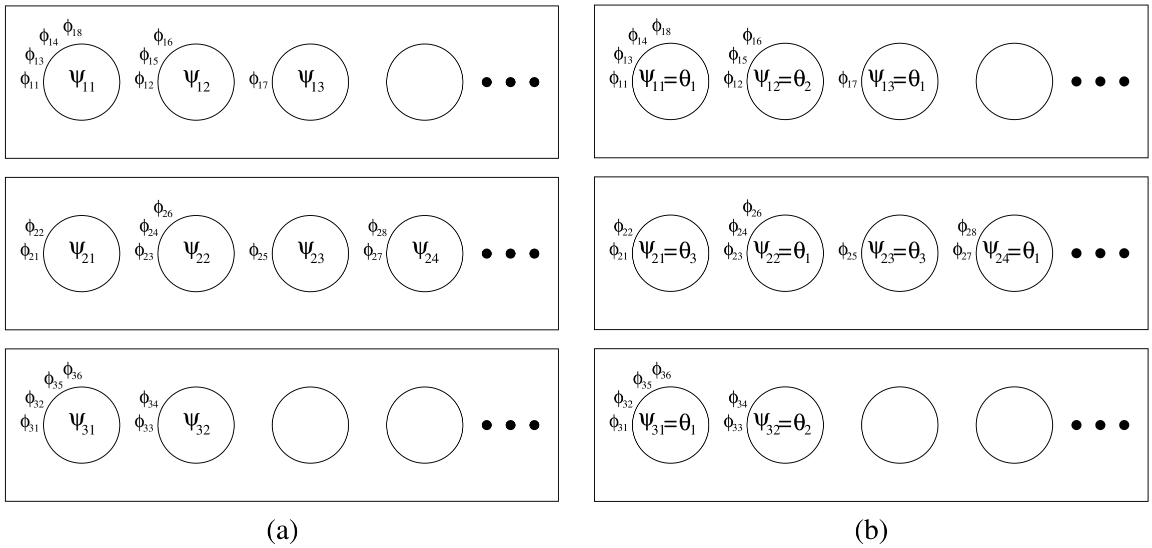

Figure 4: (a) A depiction of a hierarchical Dirichlet process as a Chinese restaurant. Each rectangle is a restaurant (group) with a number of tables. Each table is associated with a parameter ψ_{jt} which is distributed according to G_0, and each ϕ_{ji} sits at the table to which it has been assigned in (25). (b) Integrating out G_0, each ψ_{jt} is assigned some dish (mixture component) θ_k.

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 (25). 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 (26).

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 4.3(a). Here the number of mixture components L is a positive integer, and the mixing proportions $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_j$ are vectors of length L. The conditional distributions are given by

$$\begin{aligned} \boldsymbol{\beta} \mid \gamma &\sim \mathrm{Dir}(\gamma/L, \ldots, \gamma/L) & & \\ \boldsymbol{\pi}_j \mid \alpha_0, \boldsymbol{\beta} &\sim \mathrm{Dir}(\alpha_0 \boldsymbol{\beta}) & z_{ji} \mid \boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\ \theta_k \mid H &\sim H & x_{ji} \mid z_{ji}, (\theta_k)_{k=1}^L &\sim F(\theta_{z_{ji}}) \,. \end{aligned} \tag{27}$$

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

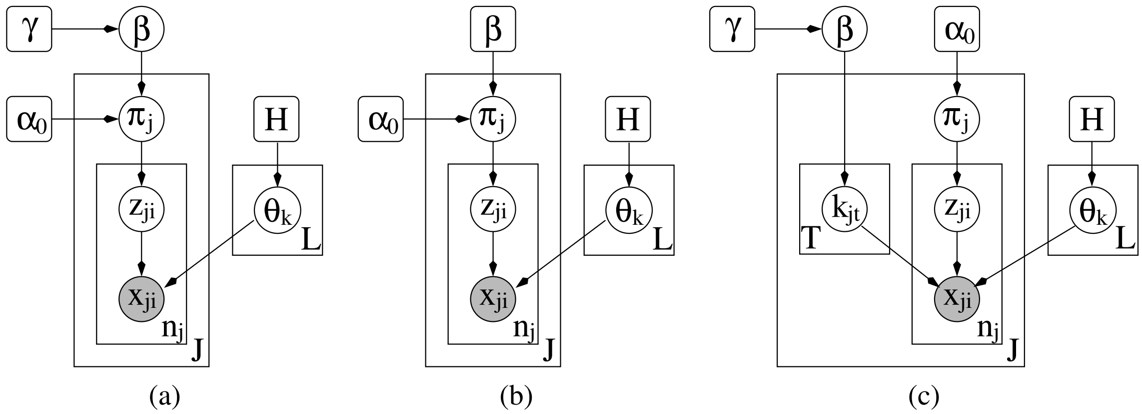

Figure 5: Finite models. (a) A finite hierarchical multiple mixture model whose infinite limit yields the hierarchical Dirichlet process mixture model. (b) The finite model with symmetric $\boldsymbol{\beta}$ weights. The various mixture models are independent of each other given $\alpha_0, \boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and thus cannot capture dependencies between the groups. (c) Another finite model that yields the hierarchical Dirichlet process in the infinite limit.

in Section 3.4, for every measurable function f integrable with respect to H we have

$$\int f(\phi)\, dG_0^L(\phi) \xrightarrow{\mathcal{D}} \int f(\phi)\, dG_0(\phi) \,, \tag{28}$$

as $L \to \infty$. Further, using standard properties of the Dirichlet distribution, we see that (19) still holds for the finite case for partitions of $\{1, \ldots, L\}$; hence we have:

$$G_j^L \;\sim\; \mathrm{DP}(\alpha_0, G_0^L) \,. \tag{29}$$

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

By way of comparison, it is interesting to consider what happens if we set $\boldsymbol{\beta} = (1/L, \ldots, 1/L)$ symmetrically instead, and take the limit $L \to \infty$ (shown in Figure 4.3(b)). Let k be a mixture 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 $\boldsymbol{\pi}_{j'}$ is independent of $\boldsymbol{\pi}_j$, and $\boldsymbol{\beta}$ is symmetric, this probability is:

$$p(\exists\, i' : z_{j'i'} = k \mid \alpha_0 \boldsymbol{\beta}) \leq \sum_{i'} p(z_{j'i'} = k \mid \alpha_0 \boldsymbol{\beta}) = \frac{n_j}{L} \to 0 \qquad \text{as } L \to \infty \,. \tag{30}$$

Since group j can use at most n_j mixture components (there are only n_j observations), as $L \to \infty$ 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., $\boldsymbol{\beta}$ is symmetric). Thus one possible direct way to deal with the problem would be to assume asymmetric weights for $\boldsymbol{\beta}$. In order that the parameter set does not grow as $L \to \infty$, we need to place a prior on $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ (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 4.3(c), where:

$$\begin{aligned} \boldsymbol{\beta} \mid \gamma &\sim \mathrm{Dir}(\gamma/L, \ldots, \gamma/L) & k_{jt} \mid \boldsymbol{\beta} &\sim \boldsymbol{\beta} \\ \boldsymbol{\pi}_j \mid \alpha_0 &\sim \mathrm{Dir}(\alpha_0/T, \ldots, \alpha_0/T) & t_{ji} \mid \boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\ \theta_k \mid H &\sim H & x_{ji} \mid t_{ji}, (k_{jt})_{t=1}^T, (\theta_k)_{k=1}^L &\sim F(\theta_{k_{jt_{ji}}}) \,. \end{aligned} \tag{31}$$

As $T \to \infty$ and $L \to \infty$, 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

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 4.3(a). We also describe a sampling scheme for the concentration parameters α_0, and γ based on extensions of analogous techniques for Dirichlet processes (Escobar and West 1995).

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 $\boldsymbol{\beta} = (\beta_k)_{k=1}^\infty$, and the group weights are $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^\infty$. 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 $\{1, \ldots, T_j\}$, and each k_{jt} takes values in $\{1, \ldots, K\}$.

Let $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jn_j})$, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$, $\boldsymbol{t} = (t_{ji} : \text{ all } j, i)$, $\boldsymbol{k} = (k_{jt} : \text{ all } j, t)$, $\boldsymbol{z} = (z_{ji} : \text{ all } j, i)$ $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and $\boldsymbol{m} = (m_{jk} : \text{ all } j, k)$. When a superscript is attached to a set of variables or an occupancy number, e.g., $\boldsymbol{\theta}^{-k}$, $\boldsymbol{k}^{-jt}$, 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, $\boldsymbol{\theta}^{-k} = \boldsymbol{\theta} \backslash \theta_k$, $\boldsymbol{k}^{-jt} = \boldsymbol{k} \backslash k_{jt}$ 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 $\boldsymbol{x}$.

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 (25) and (26) can be easily adapted to be expressed in terms of t_{ji} and k_{jt}.

The state space consists of values of $\boldsymbol{t}, \boldsymbol{k}$ 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 consisting of a countably infinite number of θ_k and k_{jt}. Only finitely many are actually associated to data and represented explicitly.

Sampling $\boldsymbol{t}$. 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 (25) and (26). 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 (25), 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 (26) (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:

$$k_{jt^{\text{new}}} \mid \boldsymbol{k} \;\sim\; \sum_{k=1}^{K} \frac{m_k}{\sum_k m_k + \gamma}\delta_k + \frac{\gamma}{\sum_k m_k + \gamma}\delta_{k^{\text{new}}} \qquad\qquad \theta_{k^{\text{new}}} \;\sim\; H\,, \tag{32}$$

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

$$p(t_{ji} = t|\boldsymbol{t}^{-ji}, \boldsymbol{k}, \boldsymbol{\theta}, \boldsymbol{x}) \propto \begin{cases} \alpha_0 f(x_{ji}|\theta_{k_{jt}}) & \text{if } t = t^{\text{new}}, \\ n_{jt}^{-i} f(x_{ji}|\theta_{k_{jt}}) & \text{if } t \text{ previously used.} \end{cases} \tag{33}$$

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 $\boldsymbol{k}$. 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

$$p(k_{jt} = k|\boldsymbol{t}, \boldsymbol{k}^{-jt}, \boldsymbol{\theta}, \boldsymbol{x}) \propto \begin{cases} \gamma \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k = k^{\text{new}}, \\ m_k^{-t} \prod_{i:t_{ji}=t} f(x_{ji}|\theta_k) & \text{if } k \text{ is previously used.} \end{cases} \tag{34}$$

Sampling $\boldsymbol{\theta}$. Conditioned on the indicator variables $\boldsymbol{k}$ and $\boldsymbol{t}$, 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:

$$p(\theta_k|\boldsymbol{t},\boldsymbol{k},\boldsymbol{\theta}^{-k},\boldsymbol{x}) \propto h(\theta_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\theta_k) \tag{35}$$

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 present an alternative 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 (27) and Figure 4.3(a). Taking the infinite limit, the model approaches a hierarchical Dirichlet process mixture model, and our sampling scheme 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). A similar scheme can be obtained starting from the stick-breaking representation.

Suppose we have L mixture components. For our sampling scheme to be computationally feasible when we take $L \to \infty$, 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 instead pool them together and use a single *unrepresented* component. Whenever the unrepresented component gets chosen to model an observation, we increment K and instantiate a new component from this pool.

The variables of interest in the finite model are $\boldsymbol{z}$, $\boldsymbol{\pi}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We integrate out $\boldsymbol{\pi}$, and Gibbs sample $\boldsymbol{z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. By permuting the indices we may assume that the represented components are $1,\ldots,K$. 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 take $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K,\beta_u)$. Let $\gamma_r = \gamma/L$ and $\gamma_u = \gamma(L-K)/L$ so that we have $\boldsymbol{\beta} \sim \mathrm{Dir}(\gamma_r,\ldots,\gamma_r,\gamma_u)$. 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 integrate over $\boldsymbol{\pi}$ analytically, giving the following conditional probability of $\boldsymbol{z}$ given $\boldsymbol{\beta}$:

$$p(\boldsymbol{z}|\boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_0\beta_k+n_{jk})}{\Gamma(\alpha_0\beta_k)} \,. \tag{36}$$

Sampling $\boldsymbol{z}$. From (36), the prior probability for $z_{ji}=k$ given $\boldsymbol{z}^{-ji}$ and $\boldsymbol{\beta}$ is simply $\alpha_0\beta_k+n_{jk}^{-ji}$ for each $k=1,\ldots,K,u$. Combined with the likelihood of x_{ji} we get the conditional probability for z_{ji}:

$$p(z_{ji}=k|\boldsymbol{z}^{-ji},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{x}) \propto (\alpha_0\beta_k+n_{jk}^{-ji}) f(x_{ji}|\theta_k) \qquad \text{for } k=1,\ldots,K,u. \tag{37}$$

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.

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

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

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 $b \sim \mathrm{Beta}(1,\gamma)$ 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 $k = K+1,\ldots,L$. Notice, however, that $(\beta_{K+1}/\beta_u,\ldots,\beta_L/\beta_u) \sim \mathrm{Dir}(\gamma_r,\ldots,\gamma_r)$. It is now an exercise in standard properties of the Dirichlet distribution to show that $\beta_{k_u}/\beta_u \sim \mathrm{Beta}(1+\gamma_r, \gamma_u-\gamma_r)$. As $L \to \infty$ this is $\mathrm{Beta}(1,\gamma)$. Hence this new component has weight $b\beta_u$ where $b \sim \mathrm{Beta}(1,\gamma)$, while the weights of the unrepresented components sum to $(1-b)\beta_u$.

Sampling $\boldsymbol{\beta}$. We use an auxiliary variable method for sampling $\boldsymbol{\beta}$. Notice that in the likelihood term (36) for $\boldsymbol{\beta}$, 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:

$$\frac{\Gamma(n_{jk}+\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k)} = \prod_{m_{jk}=1}^{n_{jk}} (m_{jk}-1+\alpha_0\beta_k) = \sum_{m_{jk}=0}^{n_{jk}} s(n_{jk},m_{jk})(\alpha_0\beta_k)^{m_{jk}} , \tag{38}$$

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 $\boldsymbol{m} = (m_{jk} : \text{ all } j,k)$ as auxiliary variables to the model. Plugging (38) into (36) and including the prior for $\boldsymbol{\beta}$, the distribution over $\boldsymbol{z}$, $\boldsymbol{m}$ and $\boldsymbol{\beta}$ is:

$$q(\boldsymbol{z},\boldsymbol{m},\boldsymbol{\beta}) = \frac{\Gamma(\gamma)}{\Gamma(\gamma_r)^K\Gamma(\gamma_u)} \left(\prod_{j=1}^{J} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)}\right) \beta_u^{\gamma_u-1} \prod_{k=1}^{K} \beta_k^{\gamma_r-1} \prod_{j=1}^{J} (\alpha_0\beta_k)^{m_{jk}} s(n_{jk},m_{jk}) . \tag{39}$$

It can be verified that $\sum_{\boldsymbol{m}} q(\boldsymbol{z},\boldsymbol{m}|\boldsymbol{\beta}) = p(\boldsymbol{z}|\boldsymbol{\beta})$. Finally, to obtain $\boldsymbol{\beta}$ given $\boldsymbol{z}$, we simply iterate sampling between $\boldsymbol{m}$ and $\boldsymbol{\beta}$ using the conditional distributions derived from (39). In the limit $L \to \infty$ the conditional distributions are simply:

$$q(m_{jk}=m|\boldsymbol{z},\boldsymbol{m}^{-jk},\boldsymbol{\beta}) \propto s(n_{jk},m)(\alpha_0\beta_k)^m \tag{40}$$

$$q(\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{m}) \propto \beta_u^{\gamma-1} \prod_{k=1}^{K} \beta_k^{\sum_j m_{jk}-1} . \tag{41}$$

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 $\boldsymbol{m}$ the conditional distribution of $\boldsymbol{\beta}$ is simply a Dirichlet distribution $\mathrm{Dir}(\sum_j m_{j1}, \ldots, \sum_j m_{jK}, \gamma)$.

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

$$p(\theta_k|\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\theta}^{-k}, \boldsymbol{x}) \propto h(\theta_k) \prod_{ji:z_{ji}=k} f(x_{ji}|\theta_k) \qquad \text{for } k = 1, \ldots, K. \tag{42}$$

5.3 Conjugacy between $\boldsymbol{\beta}$ and $\boldsymbol{m}$

The derivation of the auxiliary variable sampling scheme reveals an interesting conjugacy between the weights $\boldsymbol{\beta}$ and the auxiliary variables $\boldsymbol{m}$. First notice that the posterior for $\boldsymbol{\pi}$ given $\boldsymbol{z}$ and $\boldsymbol{\beta}$ is

$$p((\pi_{j1}, \ldots, \pi_{jK}, \pi_{ju})_{j=1}^J|\boldsymbol{z}, \boldsymbol{\beta}) \propto \prod_{j=1}^J \pi_{ju}^{\alpha_0\beta_u-1} \prod_{k=1}^K \pi_{jk}^{\alpha_0\beta_k+n_{jk}-1}, \tag{43}$$

where $\pi_{ju} = \sum_{k=K+1}^{\infty} \pi_{jk}$ is the total weight for the unrepresented components. This describes the basic conjugacy between $\boldsymbol{\pi}_j$ 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 (41) suggests that the $\boldsymbol{\beta}$ 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 (40) 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 $\boldsymbol{z}$), and we know $\boldsymbol{\beta}$ (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

$$p(t_{ji} = t|k_{jt} = k, \boldsymbol{m}_{ji}, \boldsymbol{\beta}, \alpha_0) \propto n_{jt}^{-i}, \tag{44}$$

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

$$p(t_{ji} = t^{\text{new}}|k_{jt^{\text{new}}} = k, \boldsymbol{m}_{ji}, \boldsymbol{\beta}, \alpha_0) \propto \alpha_0\beta_k. \tag{45}$$

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:

$$p(\#\text{ components} = m|n_{jk} \text{ samples}, \alpha_0\beta_k) = s(n_{jk}, m)(\alpha_0\beta_k)^m \frac{\Gamma(\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k + n_{jk})}, \tag{46}$$

which is exactly (40). 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 $\boldsymbol{\beta}$.

5.4 Comparison of sampling schemes

We have described two different sampling schemes for hierarchical Dirichlet process mixture models. In Section 6 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.

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 $\boldsymbol{\beta}$, instead of sampling the number of tables under each component directly using (40), 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 (8), and is not a Markov chain). This follows from the conjugacy arguments in Section 5.3. A consequence is that we now have the number of tables in that component, which can be used to update $\boldsymbol{\beta}$. 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 Posterior sampling for concentration parameters

MCMC samples from the posterior distributions for the concentration parameters γ and α_0 of the hierarchical Dirichlet process can be obtained 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 5.3 this is given by:

$$p(T_1, \ldots, T_J | \alpha_0, n_1, \ldots, n_J) = \prod_{j=1}^{J} s(n_j, T_j) \alpha_0^{T_j} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} . \tag{47}$$

Further, α_0 does not govern other aspects of the joint distribution, hence given T_j the observations are independent of α_0. Therefore (47) gives the likelihood for α_0. Together with the prior for α_0 and the current sample for T_j we can now derive MCMC 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 (47) 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

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} = \int_0^1 w_j^{\alpha_0}(1 - w_j)^{n_j - 1}\left(1 + \frac{n_j}{\alpha_0}\right) dw_j \; . \tag{48}$$

In particular, we define auxiliary variables $\boldsymbol{w} = (w_j)_{j=1}^J$ and $\boldsymbol{s} = (s_j)_{j=1}^J$ 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:

$$q(\alpha_0, \boldsymbol{w}, \boldsymbol{s}) \propto \alpha_0^{a-1+\sum_{j=1}^J T_j} e^{\alpha_0 b} \prod_{j=1}^J w_j^{\alpha_0}(1 - w_j)^{n_j - 1}\left(\frac{n_j}{\alpha_0}\right)^{s_j} \; . \tag{49}$$

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 $\boldsymbol{w}$ and $\boldsymbol{s}$ we have:

$$q(\alpha_0 | \boldsymbol{w}, \boldsymbol{s}) \propto \alpha_0^{a-1+\sum_{j=1}^J T_j - s_j} e^{\alpha_0 (b - \sum_{j=1}^J \log w_j)} \; , \tag{50}$$

which is a gamma distribution 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:

$$q(w_j | \alpha_0) \propto w_j^{\alpha_0}(1 - w_j)^{n_j - 1} \tag{51}$$

$$q(s_j | \alpha_0) \propto \left(\frac{n_j}{\alpha_0}\right)^{s_j} \; , \tag{52}$$

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:

$$p(K | \gamma, T) = s(T, K)\gamma^K \frac{\Gamma(\gamma)}{\Gamma(\gamma + T)} \; . \tag{53}$$

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.

6 EXPERIMENTS

We describe three experiments in this section to highlight various aspects of the hierarchical Dirichlet process: its nonparametric nature, its hierarchical nature, and the ease with which we can extend the framework to more complex models, specifically a hidden Markov model with a countably infinite state space.

The software that we used for these experiments is available at http://www.cs.berkeley.edu/∼ywteh/research/npbayes. The core sampling routines are written in C, with a MATLAB interface for interactive control and graphical display. The software implements a hierarchy of Dirichlet processes of arbitrary depth.

6.1 Document modeling

Recall the problem of document modeling discussed in Section 1. Following standard methodology in the information retrieval literature (Salton and McGill 1983), we view a document as a "bag-of-words"; that is, we make an exchangeability assumption for the words in the document. Moreover, we model the words in a document as arising from a mixture model, in which a mixture component—a "topic"—is a probability distribution over words from some basic vocabulary. The goal is to model a corpus of documents in such a way as to allow the topics to be shared among the documents in a corpus.

A parametric approach to this problem is provided by the *latent Dirichlet allocation* (LDA) model of Blei et al. (2003). This model involves a finite mixture model in which the mixing proportions are drawn on a document-specific basis from a Dirichlet distribution. Moreover, given these mixing proportions, each word in the document is an independent draw from the mixture model. That is, to generate a word, a mixture component (i.e., a topic) is selected, and then a word is generated from that topic.

Note that the assumption that each word is associated with a possibly different topic differs from a model in which a mixture component is selected once per document, and then words are generated i.i.d. from the selected topic. Moreover, it is interesting to note that the same distinction arises in population genetics, where multiple words in a document are analogous to multiple markers along a chromosome. Pritchard et al. (2000). One can consider mixture models in which marker probabilities are selected once per chromosome or once per marker. The latter are referred to as "admixture" models by Pritchard et al. (2000), who develop an admixture model that is essentially identical to LDA.

A limitation of the parametric approach include the necessity of estimating the number of mixture components in some way. This is a particularly difficult problem in areas such as information retrieval and genetics, in which the number of components is expected to be large. It is natural to consider replacing the finite mixture model with a DP, but, given the differing mixing proportions for each document, this requires a different DP for each document. This then poses the problem of sharing mixture components across multiple DPs, precisely the problem that the hierarchical DP is designed to solve.

We fit both the LDA model and the hierarchical DP mixture model to a corpus of nematode biology abstracts (see http://elegans.swmed.edu/wli/cgcbib). There are 5838 abstracts in total. After removing standard stop words and words appearing fewer than 10 times, we are left with 476441 words in total and a vocabulary size of 5699.

Both models were as similar as possible beyond the distinction between the parametric or non-parametric Dirichlet distribution. Both models used a symmetric Dirichlet distribution with parameters of 0.5 for the prior H over topic distributions. The concentration parameters were integrated out using a vague gamma prior: $\gamma \sim \mathrm{Gamma}(1, .1)$ and $\alpha_0 \sim \mathrm{Gamma}(1, 1)$.

We evaluated the models via 10-fold cross-validation. The evaluation metric was the *perplexity*, a standard metric in the information retrieval literature. The perplexity of a held-out abstract

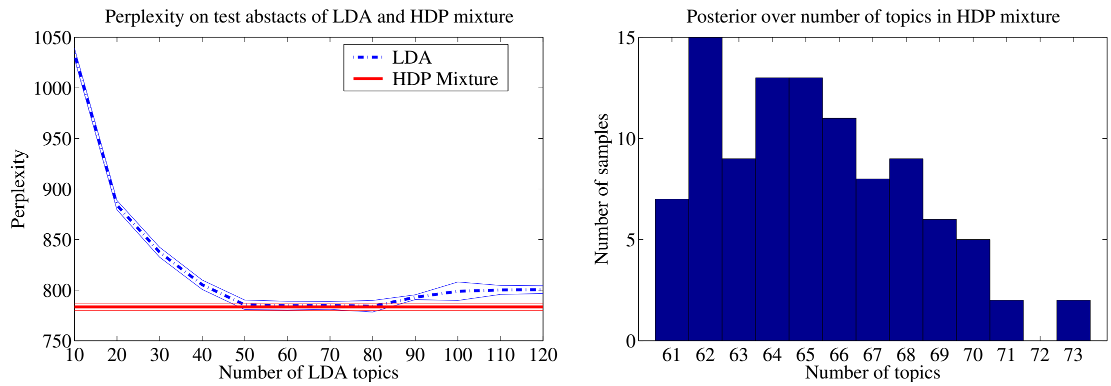

Figure 6: (Left) Comparison of latent Dirichlet allocation and the hierarchical Dirichlet process mixture. Results are averaged over 10 runs; the error bars are one standard error. (Right) Histogram of the number of topics for the hierarchical Dirichlet process mixture over 100 posterior samples.

consisting of words $w_1, \ldots, w_I$ is defined to be:

$$\exp\left(-\frac{1}{I}\log p(w_1, \ldots, w_I|\text{Training corpus})\right) \tag{54}$$

where $p(\cdot)$ is the probability mass function for a given model. The perplexity can be understood as the average inverse probability of single words given the training set.

The results are shown in Figure 6.1. For LDA we evaluated the perplexity for mixture component cardinalities ranging between 10 and 120. As seen in Figure 6.1(Left), the hierarchical DP mixture approach—which integrates over the mixture component cardinalities—performs as well as the best LDA model, doing so without any form of model selection procedure as would be required for LDA. Moreover, as shown in Figure 6.1(Right), the posterior over the number of topics obtained under the hierarchical DP mixture model is consistent with this range of the best-fitting LDA models.

6.2 Multiple corpora

We now consider the problem of sharing clusters among the documents in multiple corpora. We approach this problem by extending the hierarchical Dirichlet process to a third level. A draw from a top-level DP yields the base measure for each of a set of corpus-level DPs. Draws from each of these corpus-level DPs yield the base measures for DPs associated with the documents within a corpus. Finally, draws from the document-level DPs provide a representation of each document as a probability distribution across "topics," which are distributions across words. The model allows topics to be shared both within corpora and between corpora.

The documents that we used for these experiments consist of articles from the proceedings of the *Neural Information Processing Systems* (NIPS) conference for the years 1988-1999. The original articles are available at http://books.nips.cc; we use a preprocessed version available at http://www.cs.utoronto.ca/~roweis/nips. The NIPS conference deals with a range of topics covering both human and machine intelligence. Articles are separated into nine prototypical sections: algorithms and architectures (AA), applications (AP), cognitive science (CS), control and navigation (CN), implementations (IM), learning theory (LT), neuroscience (NS), signal processing (SP), vision sciences (VS). (These are the sections used in the years 1995-1999. The sectioning in earlier

Table 2: Summary statistics for the NIPS data set.

Sections	# Papers	Total # words	Distinct # words
Cognitive science (CS)	72	83798	4825
Neuroscience (NS)	157	170186	5031
Learning theory (LT)	226	225217	5114
Algorithms and architectures (AA)	369	388402	5365
Implementations (IM)	195	199366	5334
Signal processing (SP)	71	75016	4618
Vision sciences (VS)	104	114231	4947
Applications (AP)	146	158621	5276
Control and navigation (CN)	107	116885	4783
All sections	1447	1531722	5570

years differed slightly; we manually relabeled sections from the earlier years to match those used in 1995-1999.) We treat these sections as "corpora," and are interested in the pattern of sharing of topics among these corpora.

There were 1447 articles in total. Each article was modeled as a "bag-of-words," i.e., each word was modeled as a multinomial variate and the words were modeled as conditionally i.i.d. given the underlying draw from the DP. We culled standard stop words as well as words occurring more than 4000 or fewer than 50 times in the whole corpus. This left us with on average slightly more than 1000 words per article. Some summary statistics for the data set are provided in Table 2.

We considered the following experimental setup. Given a set of articles from a single NIPS section that we wish to model (the VS section in the experiments that we report below), we wish to know whether it is of value (in terms of prediction performance) to include articles from other NIPS sections. This can be done in one of two ways: we can lump all of the articles together without regard for the division into sections, or we can use the hierarchical DP approach to link the sections. Thus we consider three models (see Figure 7 for graphical representations of these models):

- **M1**: This model ignores articles from the other sections and simply uses a hierarchical DP mixture of the kind presented in Section 6.1 to model the VS documents. This model serves as a baseline. We used $\gamma \sim \mathrm{Gamma}(5, 0.1)$ and $\alpha_0 \sim \mathrm{Gamma}(0.1, 0.1)$ as prior distributions for the concentration parameters.

- **M2**: This model incorporates articles from other sections, but ignores the distinction into sections, using a single hierarchical DP mixture model to model all of the articles. Priors of $\gamma \sim \mathrm{Gamma}(5, 0.1)$ and $\alpha_0 \sim \mathrm{Gamma}(0.1, 0.1)$ were used.

- **M3**: This model takes a full hierarchical approach and models the NIPS sections as multiple corpora, linked via the hierarchical DP mixture formalism. The model is a tree, in which the root is a draw from a single DP for all articles, the first level is a set of draws from DPs for the NIPS sections, and the second level is set of draws from DPs for the articles within sections. Priors of $\gamma \sim \mathrm{Gamma}(5, 0.1)$, $\alpha_0 \sim \mathrm{Gamma}(5, 0.1)$, and $\alpha_1 \sim \mathrm{Gamma}(0.1, 0.1)$ were used.

We conducted experiments in which sets of 80 articles were chosen uniformly at random from each of the sections other than VS (this was done to balance the sections, which are of different

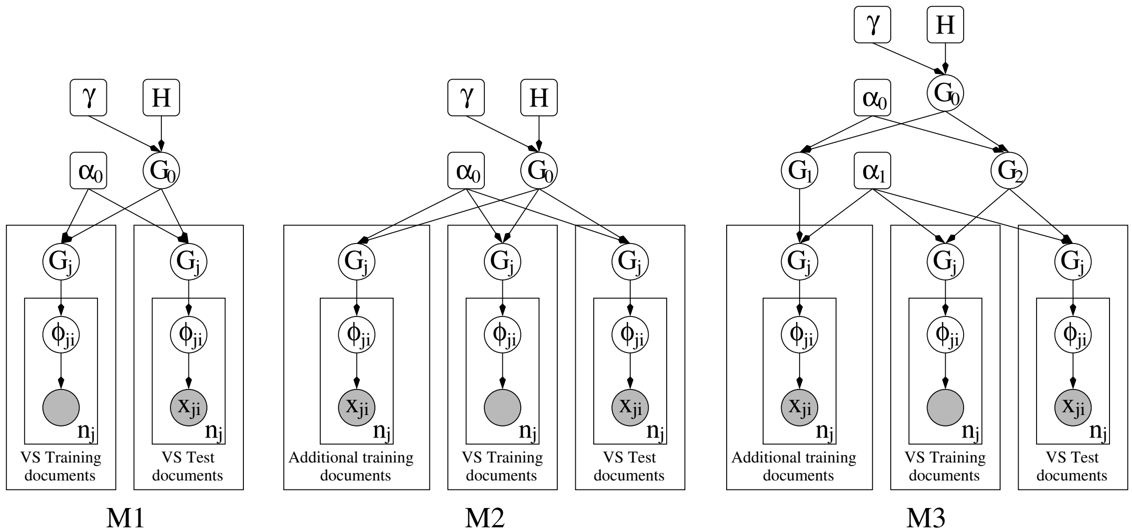

Figure 7: Three models for the NIPS data. From left to right: M1, M2 and M3.

sizes). A "training set" of 80 articles were also chosen uniformly at random from the VS section, as were an additional set of 47 test articles.

Figure 6.2(Left) presents the predictive performance for all three models as the number N of articles used for training in the VS section ranged from 0 to 80. The performance is measured in terms of the perplexity of single words from the test articles given the training articles. As seen in the figure, the fully hierarchical model M3 performs best, with perplexity decreasing rapidly with modest values of N. For small values of N, the performance of M1 is quite poor, but the performance approaches that of M3 when more than 20 articles are included in the VS training set. The performance of the partially-hierarchical M2 was poorer than the fully-hierarchical M3 throughout the range of N. M2 dominated M1 for small N, but yielded poorer performance than M1 for N greater than 14. Our interpretation is that the sharing of strength based on other articles is useful when little other information is available (small N), but that eventually (medium to large N) there is crosstalk between the sections and it is preferable to model them separately and share strength via the hierarchy.

While the results in Figure 6.2(Left) are an average over the sections, it is also of interest to see which sections are the most beneficial in terms of enhancing the prediction of the articles in VS. Figure 6.2(Right) plots the predictive performance for model M3 when given data from each of three particular sections: LT, AA and AP. While articles in the LT section are concerned mostly with theoretical properties of learning algorithms, those in AA are mostly concerned with models and methodology, and those in AP are mostly concerned with applications of learning algorithms to various problems. As seen in the figure, we see that predictive performance is enhanced the most by prior exposure to articles from AP, less by articles from AA, and still less by articles from LT. Given that articles in the VS tend to be concerned with the practical application of learning algorithms to problems in computer vision, this pattern of transfer seems reasonable.

Finally, it is of interest to investigate the subject matter content of the topics discovered by the hierarchical DP model. We did so in the following experimental setup. For a given section other than VS (e.g., AA), we fit a model based on articles from that section. We then introduced articles from the VS section and continued to fit the model, while holding the topics found from the earlier fit

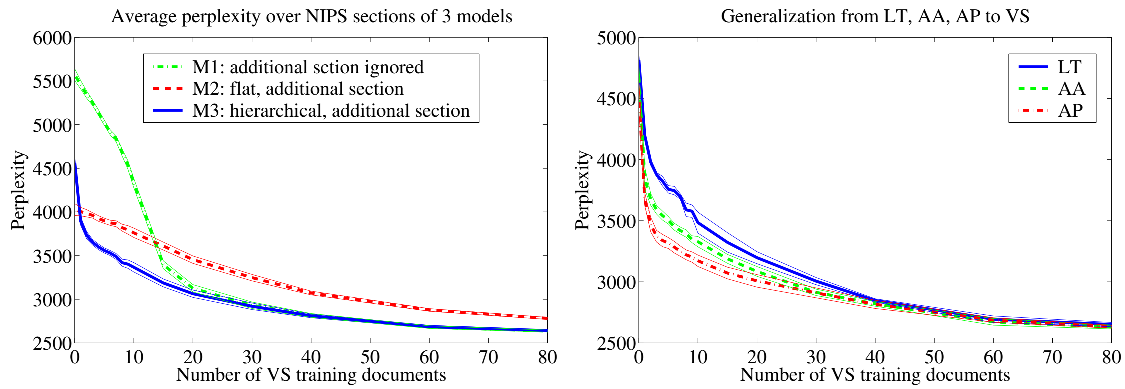

Figure 8: (Left) Perplexity of test VS documents given training documents from VS and another section for 3 different models. Curves shown are averaged over the other sections and 5 runs. (Right) Perplexity of test VS documents given LT, AA and AP documents respectively, using M3, averaged over 5 runs. In both plots, the error bars represent one standard error.

fixed, and recording which topics from the earlier section were allocated to words in the VS section. Table 3 displays the two most frequently occurring topics in this setup (a topic is represented by the set of words which have highest probability under that topic). We also show some of the new topics created by the model while fitting the VS data in Table 4. Both sets of topics provide qualitative confirmation of our expectations regarding the overlap between VS and other sections.

7 HIDDEN MARKOV MODELS

The simplicity of the hierarchical DP specification—the base measure for a DP is distributed as a DP—makes it straightforward to exploit the hierarchical DP as a building block in more complex models. In this section we demonstrate this in the case of the hidden Markov model.

Recall that a hidden Markov model (HMM) is a doubly stochastic Markov chain in which a sequence of multinomial "state" variables $(v_1, v_2, \ldots, v_T)$ are linked via a state transition matrix, and each element y_t in a sequence of "observations" $(y_1, y_2, \ldots, y_T)$ is drawn independently of the other observations conditional on v_t (Rabiner 1989). This is essentially a dynamic variant of a finite mixture model, in which there is one mixture component corresponding to each value of the multinomial state. As with classical finite mixtures, it is interesting to consider replacing the finite mixture underlying the HMM with a Dirichlet process.

Note that the HMM involves not a single mixture model, but rather a set of mixture models—one for each value of the current state. That is, the "current state" v_t indexes a specific row of the transition matrix, with the probabilities in this row serving as the mixing proportions for the choice of the "next state" v_{t+1}. Given the next state v_{t+1}, the observation y_{t+1} is drawn from the mixture component indexed by v_{t+1}. Thus, to consider a nonparametric variant of the HMM which allows an unbounded set of states, we must consider a set of DPs, one for each value of the current state. Moreover, these DPs must be linked, because we want the same set of "next states" to be reachable from each of the "current states." This amounts to the requirement that the atoms associated with the state-conditional DPs should be shared—exactly the framework of the hierarchical DP.

Thus, we can define a nonparametric hidden Markov model by simply replacing the set of conditional finite mixture models underlying the classical HMM with a hierarchical Dirichlet process. We

Table 3: Topics shared between VS and the other NIPS sections. These topics are the most frequently occurring in the VS fit, under the constraint that they are associated with a significant number of words (greater than 2500) from the other section.

CS	task representation pattern processing trained representations three process unit patterns examples concept similarity bayesian hypotheses generalization numbers positive classes hypothesis
NS	cells cell activity response neuron visual patterns pattern single fig visual cells cortical orientation receptive contrast spatial cortex stimulus tuning
LT	signal layer gaussian cells fig nonlinearity nonlinear rate eq cell large examples form point see parameter consider random small optimal
AA	algorithms test approach methods based point problems form large paper distance tangent image images transformation transformations pattern vectors convolution simard
IM	processing pattern approach architecture single shows simple based large control motion visual velocity flow target chip eye smooth direction optical
SP	visual images video language image pixel acoustic delta lowpass flow signals separation signal sources source matrix blind mixing gradient eq
AP	approach based trained test layer features table classification rate paper image images face similarity pixel visual database matching facial examples
CN	ii tree pomdp observable strategy class stochastic history strategies density policy optimal reinforcement control action states actions step problems goal

Table 4: Novel topics (not shared with another NIPS section) that arose during the fit of the VS data.

CS	matching correspondence match point points transformation line matches object objective
NS	matching correspondence match point points transformation object matches line constraints
LT	matching correspondence match point points transformation object matches line scene
AA	depth grossberg contrast stage gray perception boundaries classification regions patch
IM	face facial images image view faces expression gesture action representation
SP	motion visual cells orientation field receptive stimulus cortex direction spatial
AP	disparity stereo layer match left surfaces depth energy constraints constraint
CN	motion visual direction velocity moving stimulus stage signals directions second

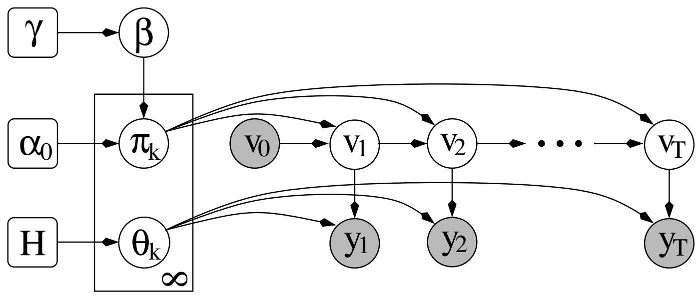

Figure 9: A hierarchical Bayesian model for the infinite hidden Markov model.

refer to the resulting model as a *hierarchical Dirichlet process hidden Markov model* (HDP-HMM). The HDP-HMM provides an alternative to methods that place an explicit parametric prior on the number of states or make use of model selection methods to select a fixed number of states (Stolcke and Omohundro 1993).

In work that served as an inspiration for the HDP-HMM, Beal et al. (2002) discussed a model 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. Indeed, Beal et al. (2002) defined a notion of "hierarchical Dirichlet process" for this model, 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) considered the following 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 relate this two-level urn model to the hierarchical DP framework, let us describe a representation of the latter using the stick-breaking formalism. In particular, consider the hierarchical Dirichlet process representation shown in Figure 9. The parameters in this representation have the following distributions:

$$\boldsymbol{\beta} \mid \gamma \sim \mathrm{Stick}(\gamma) \qquad \boldsymbol{\pi}_k \mid \alpha_0, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta}) \qquad \theta_k \mid H \sim H\,, \tag{55}$$

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

$$v_t \mid v_{t-1}, (\boldsymbol{\pi}_k)_{k=1}^{\infty} \sim \boldsymbol{\pi}_{v_{t-1}} \qquad y_t \mid v_t, (\theta_k)_{k=1}^{\infty} \sim F(\theta_{v_t})\,, \tag{56}$$

where we assume for simplicity that there is a distinguished initial state v_0. If we now consider the Chinese restaurant franchise representation of this model as discussed in Section 5, it turns out that the result is equivalent to the coupled urn model of Beal et al. (2002).

Unfortunately, posterior inference using the Chinese restaurant franchise representation is awkward for this model, involving involving substantial bookkeeping. 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.

On the other hand, both the stick-breaking and the infinite limit of finite models representation lead directly to a Markov chain Monte Carlo sampling scheme involving auxiliary variables that is straightforward to implement. In the experiments reported in the following section we used the auxiliary variable representation.

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. This setup is readily accommodated within the hierarchical DP framework by simply considering an additional level of the Bayesian hierarchy, letting a master Dirichlet process couple each of the HDP-HMMs, each of which is a set of Dirichlet processes.

7.1 Alice in Wonderland

In this section we report experimental results for the problem of predicting strings of letters in sentences taken from Lewis Carroll's *Alice's Adventures in Wonderland*, comparing the HDP-HMM to other HMM-related approaches.

Each sentence is treated as a sequence of letters and spaces (rather than as a sequence of words). There are 27 distinct symbols—26 letters and space—cases and punctuation marks are ignored. The emission distributions are again multinomial. There are 20 training sentences, with average length of 51 symbols, while there are 40 test sentences with an average length of 100.

Using the auxiliary variable sampling method for posterior predictive inference, we compared the HDD-HMM to a variety of other methods for prediction using hidden Markov models: (1) a classical HMM using maximum likelihood (ML) parameters obtained via the Baum-Welch algorithm (Rabiner 1989), (2) a classical HMM using maximum a posteriori (MAP) parameters, taking the priors to be were independent, symmetric Dirichlet distributions for both the transition and emission probabilities, and (3) a classical HMM trained using an approximation to a full Bayesian analysis—in particular, a variational Bayesian (VB) method due to Beal (2003). For each of these classical HMMs, we conducted experiments for each value of the state cardinality ranging from 1 to 30.

Again using the perplexity on test sentences to evaluate predictive performance (see (54)), we present the results in Figure 7.1. For VB, the predictive probability is intractable to compute, so the modal setting of parameters was used. Both MAP and VB models were given optimal settings of the hyperparameters found using the HDP-HMM. We see that the HDP-HMM has a lower perplexity than all of the models tested for ML, MAP, and VB.

8 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.

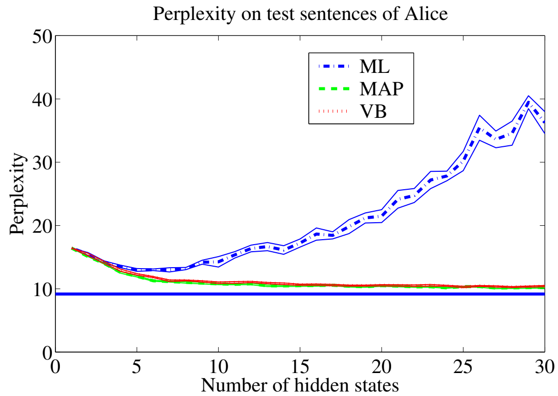

Figure 10: Comparing the infinite hidden Markov model (solid horizontal line) with ML, MAP and VB trained hidden Markov models. The error bars represent one standard error (those for the HDP-HMM are too small to see).

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 an 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 explicitly.

Clustering is an important activity in many large-scale data analysis problems in engineering and science, reflecting the heterogeneity that is often present when data are collected on a large scale. Clustering problems can be approached within a probabilistic framework via finite mixture models (and their dynamical cousins the HMM), and recent years have seen numerous examples of applications of finite mixtures and HMMs in areas such as bioinformatics (Durbin et al. 1998), speech recognition (Huang et al. 2001), information retrieval (Blei et al. 2003), computational vision (Forsyth and Ponce 2002) and robotics (Thrun 2000). These areas also provide numerous instances of data analyses which involve multiple, linked sets of clustering problems, for which classical clustering methods (model-based or non-model-based) provide little in the way of leverage. In bioinformatics we have already alluded to the problem of finding haplotype structure in subpopulations. Other examples in bioinformatics include the use of HMMs for amino acid sequences, where a hierarchical DP version of the HMM would allow motifs to be discovered and shared among different families of proteins. In speech recognition multiple HMMs are already widely used, in the form of word-specific and speaker-specific models, and adhoc methods are generally used to share statistical strength among models. We have discussed examples of grouped data in information retrieval; other examples include problems in which groups indexed by author or by language. Finally, computational vision and robotics problems often involve sets of descriptors or objects that are arranged in a taxonomy. Examples such as these, in which there is substantial uncertainty regarding appropriate numbers of clusters, and in which the sharing of statistical strength among groups is natural and desirable, suggest that the hierarchical nonparametric Bayesian approach to clustering presented here may provide a generally useful extension of model-based clustering.

References

Aldous, D. (1985), "Exchangeability and Related Topics," in *École d'Été de Probabilités de Saint-Flour XIII–1983*, Springer, Berlin, pp. 1–198.

Antoniak, C. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *Annals of Statistics*, 2(6), pp. 1152–1174.

Beal, M. (2003), "Variational Algorithms for Approximate Bayesian Inference," Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. (2002), "The Infinite Hidden Markov Model," in T. G. Dietterich, S. Becker, and Z. Ghahramani (eds.) *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, vol. 14, pp. 577–584.

Blackwell, D. and MacQueen, J. (1973), "Ferguson Distributions via Pólya Urn Schemes," *Annals of Statistics*, 1, pp. 353–355.

Blei, D., Jordan, M., and Ng, A. (2003), "Hierarchical Bayesian Models for Applications in Information Retrieval," in *Bayesian Statistics*, vol. 7, pp. 25–44.

Carota, C. and Parmigiani, G. (2002), "Semiparametric Regression for Count Data," *Biometrika*, 89(2), pp. 265–281.

Cifarelli, D. and Regazzini, E. (1978), "Problemi Statistici Non Parametrici in Condizioni di Scambiabilità Parziale e Impiego di Medie Associative," Tech. rep., Quaderni Istituto Matematica Finanziaria dell'Università di Torino.

De Iorio, M., Müller, P., and Rosner, G. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99(465), pp. 205–215.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998), *Biological Sequence Analysis—Probabilistic Models of Proteins and Nucleic Acids*, Cambridge, UK: Cambridge University Press.

Escobar, M. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, pp. 577–588.

Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1(2), pp. 209–230.

Fong, D., Pammer, S., Arnold, S., and Bolton, G. (2002), "Reanalyzing Ultimatum bargaining—Comparing Nondecreasing Curves Without Shape Constraints," *Journal of Business and Economic Statistics*, 20, pp. 423–440.

Forsyth, D. A. and Ponce, J. (2002), *Computer Vision—A Modern Approach*, Upper Saddle River, NJ: Prentice-Hall.

Gabriel, S. et al. (2002), "The Structure of Haplotype Blocks in the Human Genome," *Science*, 296, pp. 2225–2229.

Gilks, W. and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, pp. 337–348.

Green, P. and Richardson, S. (2001), “Modelling Heterogeneity with and without the Dirichlet Process,” *Scandinavian Journal of Statistics*, 28, pp. 355–377.

Huang, X., Acero, A., and Hon, H.-W. (2001), *Spoken Language Processing*, Upper Saddle River, NJ: Prentice-Hall.

Ishwaran, H. and James, L. (2001), “Gibbs Sampling Methods for Stick-Breaking Priors,” *Journal of the American Statistical Association*, 96(453), pp. 161–173.

—— (2004), “Computational Methods for Multiplicative Intensity Models using Weighted Gamma Processes: Proportional Hazards, Marked Point Processes and Panel Count Data,” *Journal of the American Statistical Association*, 99, pp. 175–190.

Ishwaran, H. and Zarepour, M. (2002), “Exact and Approximate Sum-Representations for the Dirichlet Process,” *Canadian Journal of Statistics*, 30, pp. 269–283.

Jain, S. and Neal, R. (2000), “A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model,” Tech. Rep. 2003, Department of Statistics, University of Toronto.

Kleinman, K. and Ibrahim, J. (1998), “A Semi-parametric Bayesian Approach to Generalized Linear Mixed Models,” *Statistics in Medicine*, 17, pp. 2579–2596.

MacEachern, S. (1999), “Dependent Nonparametric Processes,” in *Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association.

MacEachern, S., Kottas, A., and Gelfand, A. (2001), “Spatial Nonparametric Bayesian Models,” Tech. Rep. 01-10, Institute of Statistics and Decision Sciences, Duke University, http://ftp.isds.duke.edu/WorkingPapers/01-10.html.

MacEachern, S. and Müller, P. (1998), “Estimating Mixture of Dirichlet Process Models,” *Journal of Computational and Graphical Statistics*, 7, pp. 223–238.

Mallick, B. and Walker, S. (1997), “Combining Information from Several Experiments with Nonparametric Priors,” *Biometrika*, 84, pp. 697–706.

Muliere, P. and Petrone, S. (1993), “A Bayesian Predictive Approach to Sequential Search for an Optimal Dose: Parametric and Nonparametric Models,” *Journal of the Italian Statistical Society*, 2, pp. 349–364.

Müller, P., Quintana, F., and Rosner, G. (2004), “A Method for Combining Inference Across Related Nonparametric Bayesian Models,” *Journal of the Royal Statistical Society*, 66, pp. 735–749.

Neal, R. (1992), “Bayesian Mixture Modeling,” in *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, vol. 11, pp. 197–211.

—— (2000), “Markov Chain Sampling Methods for Dirichlet Process Mixture Models,” *Journal of Computational and Graphical Statistics*, 9, pp. 249–265.

Patil, G. and Taillie, C. (1977), “Diversity as a Concept and its Implications for Random Communities,” *Bulletin of the International Statistical Institute*, 47, pp. 497–515.

Pitman, J. (2002), “Combinatorial Stochastic Processes,” Tech. Rep. 621, Department of Statistics, University of California at Berkeley, lecture notes for St. Flour Summer School.

Pritchard, J., Stephens, M., and Donnelly, P. (2000), "Inference of Population Structure using Multilocus Genotype Data," *Genetics*, 155, pp. 945–959.

Rabiner, L. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77, pp. 257–285.

Rasmussen, C. (2000), "The Infinite Gaussian Mixture Model," in S. Solla, T. Leen, and K.-R. Müller (eds.) *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, vol. 12.

Salton, G. and McGill, M. (1983), *An Introduction to Modern Information Retrieval*, New York: McGraw-Hill.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, pp. 639–650.

Stephens, M., Smith, N., and Donnelly, P. (2001), "A New Statistical Method for Haplotype Reconstruction from Population Data," *American Journal of Human Genetics*, 68, pp. 978–989.

Stolcke, A. and Omohundro, S. (1993), "Hidden Markov Model Induction by Bayesian Model Merging," in C. Giles, S. Hanson, and J. Cowan (eds.) *Advances in Neural Information Processing Systems*, San Mateo CA: Morgan Kaufmann, vol. 5, pp. 11–18.

Thrun, S. (2000), "Probabilistic Algorithms in Robotics," *Artificial Intelligence Magazine*, 21(4), pp. 93–109.

Tomlinson, G. and Escobar, M. (2003), "Analysis of Densities," Talk given at the Joint Statistical Meeting.