5

Nonparametric Mixed Membership Models

Daniel Heinz

Department of Mathematics and Statistics, Loyola University of Maryland, Baltimore, MD 21210, USA

CONTENTS

5.1 Introduction 90
5.2 The Dirichlet Mixture Model 91
5.2.1 Finite Chinese Restaurant Process 92
5.3 The Dirichlet Process Mixture Model 93
5.3.1 The Dirichlet Process 94
5.3.2 Chinese Restaurant Process 95
5.3.3 Comparison of Dirichlet Mixtures and Dirichlet Process Mixtures 96
5.4 Mixed Membership Models 97
5.5 The Hierarchical Dirichlet Process Mixture Model 98
5.5.1 Chinese Restaurant Franchise 100
5.5.2 Comparison of GoM and HDPM models 101
5.6 Simulated HDPM Models 102
5.6.1 Size of the Population-Level Mixture 102
5.6.2 Size of Individual-Level Mixtures 102
5.6.3 Similarity among Individual-Level Mixtures 103
5.7 Inference Strategies for HDP Mixtures 106
5.7.1 Markov Chain Monte Carlo Techniques 107
5.7.2 Variational Inference 108
5.7.3 Hyperparameters 109
5.8 Example Applications of Hierarchical DPs 109
5.8.1 The Infinite Hidden Markov Model 110
5.9 Other Nonparametric Mixed Membership Models 111
5.9.1 Multiple-Level Hierarchies 111
5.9.2 Dependent Dirichlet Processes 112
5.9.3 Pitman-Yor Processes 112
5.10 Conclusion 113
References 113

One issue with parametric latent class models, regardless of whether or not they feature mixed memberships, is the need to specify a bounded number of classes a priori. By contrast, nonparametric models use an unbounded number of classes, of which some random number are observed in the data. In this way, nonparametric models provide a method to infer the correct number of classes based on the number of observations and their similarity.

The following chapter seeks to provide mathematical and intuitive understanding of nonparametric mixed membership models, focusing on the hierarchical Dirichlet process mixture model (HDPM). This model can be understood as a nonparametric extension of the Grade of Membership model (GoM) described by Erosheva et al. (2007). To elucidate this relationship, the Dirichlet mixture model (DM) and Dirichlet process mixture model (DPM) are first reviewed; many of the interesting properties of these latent class models carry over to the GoM and HDPM models.

After describing these four models, the HDPM model is further explored through simulation studies, including an analysis of how the model parameters affect the model's clustering behavior.

An overview of inference procedures is also provided with a focus on Gibbs sampling and variational inference. Finally, some example applications and model extensions are briefly reviewed.

5.1 Introduction

Choosing the appropriate model complexity is a problem that must be solved in almost any statistical analysis, including latent class models. Simple models efficiently describe a small set of behaviors, but are not flexible. Complex models describe a wide variety of behaviors, but are subject to overfitting the training data. For latent class models, complexity refers to the number of groups used to describe the distribution of observed and/or predicted data. One strategy is to fit multiple models of varying complexity then decide among them with a post-hoc analysis (e.g., penalized likelihood). Nonparametric mixture models provide an alternate strategy which bypasses the need to choose the correct number of classes. The hallmark of nonparametric models is that their complexity increases stochastically as more data are observed. The rate of accumulation is determined by various tuning parameters and the similarity of observed data.

One of the best-known examples of nonparametric Bayesian inference is the Dirichlet process mixture model (DPM), a nonparametric version of the Dirichlet mixture model (DM). The DM model assumes that the population consists of a fixed and finite number of classes and it therefore bounds the number of classes used to represent any sample. By contrast, a DPM posits that the population consists of an infinite number of classes. Of these, some finite but *unbounded* number of classes are observed in the data. Because the number of classes is unbounded, the model always has a positive probability of assigning a new observation to a new class.

Both Dirichlet mixtures and Dirichlet process mixtures assume that observations are fully exchangeable. Extensions for both models exist for situations in which full exchangeability is inappropriate. This may be the case when multiple measurements are made for individuals in the sample. For example, one may consider a survey analysis in which each individual responds to several items. In this case, one expects two responses to be more similar if they come from the same individual. The Grade of Membership model (GoM) adapts the DM model for partial exchangeability (Erosheva et al., 2007). In the GoM model, two responses are exchangeable if and only if they are measured from the same individual. Like the DM model, it bounds the number of classes. The GoM model is known as a *mixed membership model* or *individual-level mixture model* because each individual in the sample is associated with unique mixing weights for the various classes.

The hierarchical Dirichlet process mixture model (HDPM) extends the GoM model in the same way that the Dirichlet process mixture model extends the DM model. As with the GoM model, the HDPM model assumes that responses are exchangeable if and only if they come from the same individual. Whereas the GoM assumes that the population consists of a fixed and finite number of classes, the HDPM posits an infinite number of classes. Thus, the HDPM model does not bound the number of classes used to represent the sample.

All four models (DM, DPM, GoM, and HDPM) cluster observations into various classes where class memberships are unobserved. They are distinguished by the type of exchangeability (full or partial) and whether or not the number of classes in the population is bounded a priori.

Nonparametric mixture models, such as the DPM and HDPM models, have several intuitive advantages. Because the number of classes is not fixed, they provide a posterior distribution over the model complexity. Posterior inference includes a natural weighting of high-probability models of varying complexity. Hence, uncertainty about the "true" number of classes is measurable. Furthermore, because the number of classes is unbounded, nonparametric models always include a positive probability that the next observation belongs to a previously unobserved class. This property is

especially nice when considering predictive distributions. If the number of classes is unknown, it is possible that the next observation will be unlike any of the previous observations.

This chapter aims to provide an intuitive understanding of the hierarchical Dirichlet mixture model. Sections 5.2 and 5.3 begin with the fully exchangeable models, showing how the DM model is built into the nonparametric HDPM model by removing the bound on the number of classes. Properties of these mixtures are illustrated and compared using the Chinese restaurant process. This relationship forms the foundation for exploring properties of the GoM and HDPM models in Sections 5.4 and 5.5. In Section 5.6, the role of tuning parameters for the HDPM model is explored intuitively and illustrated through simulations. Section 5.7 provides an overview of inference strategies for DPM and HDPM models. Section 5.8 reviews some example applications with Section 5.9 devoted to brief descriptions of some model extensions.

5.2 The Dirichlet Mixture Model

Suppose a sample contains n observations, $(x_1, \ldots, x_n)$, where x_i is possibly vector-valued. A latent class model assumes that each observation belongs to one of K possible classes, where K is a finite constant. Observations are conditionally independent given their class memberships, but dependence arises because class memberships are not observed. As a generative model, each x_i is drawn by randomly choosing a class, say z_i, then sampling from the class-specific distribution.

Denote the population proportions of these K classes by $\pi = (\pi_1, \ldots, \pi_K)$ and the distribution of class k by F_k. For simplicity, assume that these distributions belong to some parametric family, $\{F(\cdot|\theta) : \theta \in \Theta\}$. Therefore, $F_k = F(\cdot|\theta_k)$, where $\theta_k \in \Theta$ denotes the class-specific parameter for class k. Given the class proportions and parameters, the latent class model is described by a simple hierarchy:

$$\begin{aligned} z_i|\pi &\overset{\text{i.i.d.}}{\sim} \text{Mult}(\pi) && i = 1 \ldots n. \\ x_i|z_i, \theta &\sim F(\cdot|\theta_{z_i}) && i = 1 \ldots n. \end{aligned}$$

Here, Mult(π) is the multinomial distribution satisfying $\mathbb{P}(z_i = k) = \pi_k$ for $k = 1 \ldots K$.

Inferential questions include learning the mixing proportions (π_k), class parameters (θ_k), and possibly the latent class assignments (z_i). Uncertainty about the class proportions and parameters may be expressed through prior laws. In the Dirichlet mixture model (DM), the class proportions have a symmetric Dirichlet prior, $\pi \sim \text{Dir}(\alpha/K)$. This distribution is specified by the precision $\alpha > 0$ and has the density function

$$f(\pi|\alpha) = \frac{\Gamma(\alpha)}{[\Gamma(\alpha/K)]^K} \prod_{k=1}^{K} \pi_k^{\alpha/K-1}, \tag{5.1}$$

wherever π is a K-dimensional vector whose elements are non-negative and sum to 1. The range of possible values for π is known as the $(K-1)$-*dimensional simplex* or more simply the $(K-1)$-simplex. The expected value of the symmetric Dirichlet distribution is the uniform probability vector $E[\pi] = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$. The precision specifies the concentration of the distribution about this mean, with larger values of α translating to less variability.

More generally, an asymmetric Dirichlet distribution is defined by a precision $\alpha > 0$ and a mean vector $\pi_0 = E[\pi] = (\pi_{01}, \ldots, \pi_{0K})$ in the $(K-1)$-simplex. If $\pi \sim \text{Dir}(\alpha, \pi_0)$, then its distribution function is

$$f(\pi|\alpha,\pi_0) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha\pi_{0k}\right)}{\prod_{k=1}^{K}\Gamma(\alpha\pi_{0k})}\prod_{k=1}^{K}\pi_{0k}^{\alpha\pi_{0k}-1}. \tag{5.2}$$

Note that the symmetric Dirichlet $\text{Dir}(\alpha/K)$ is equivalent to the distribution $\text{Dir}\big(\alpha, \frac{1}{K}\mathbf{1}\big)$, where $\mathbf{1}$ denotes the vector of ones.

The symmetric Dirichlet prior influences the way in which the DM model groups observations into the K possible classes. To finish specifying the model, each class is associated with its class parameter, θ_k. The class parameters are assumed to be i.i.d. from some prior distribution $H(\lambda)$, where λ is a model-specific hyperparameter. This results in the following model:

Dirichlet Mixture Model

$$\begin{aligned}
\pi|\alpha &\sim \text{Dir}\left(\frac{\alpha}{K}\right). \\
\theta_k|\lambda &\overset{\text{i.i.d.}}{\sim} H(\lambda) && k = 1\ldots K. \\
z_i|\pi &\overset{\text{i.i.d.}}{\sim} \text{Mult}(\pi) && i = 1\ldots n. \\
x_i|z_i,\theta &\sim F(\cdot|\theta_{z_i}) && i = 1\ldots n.
\end{aligned}$$

Because class memberships are dependent only on α, the Dirichlet precision fully specifies the prior clustering behavior of the model. The hyperparameter λ only influences class probabilities during posterior inference. A priori, observations are expected to be uniformly dispersed across the K classes and α measures how strongly the prior insists on uniformity.

5.2.1 Finite Chinese Restaurant Process [1]

Imagine a restaurant with an infinite number of tables, each of which has infinite capacity. Observations are represented by customers and class membership is defined by the customer's choice of dish. All customers at a particular table eat the same dish. When a customer sits at an unoccupied table, he selects one of the K possible dishes for his table with uniform probabilities of $1/K$. Because multiple tables may serve the same dish, the class membership of an observation must be defined by the customer's *dish* rather than his table.

The first customer sits at the first table and randomly chooses one of the K dishes. The second customer joins the first table with probability $1/(1+\alpha)$ or starts a new table with probability $\alpha/(1+\alpha)$. As subsequent customers enter, the probability that they join an occupied table is proportional to the number of people already seated there. Alternatively, they may choose a new table with probability proportional to α.

Mathematically, let T denote the number of occupied tables when the nth customer arrives and let t_n denote the table that he chooses. Given the seating arrangement of the previous customers, the probability function for t_n is

$$f(t_n|\alpha,\mathbf{t}_{\widetilde{n}}) \propto \begin{cases} \sum_{i<n}\mathbf{1}\left(t_i = t_n\right), & t_n \leq T \\ \alpha, & t_n = T+1 \end{cases}, \tag{5.3}$$

where $\mathbf{t}_{\widetilde{n}}$ denotes the table assignments of all but the nth customer and $\mathbf{1}\left(t_i = t_n\right)$ is the indicator function, which is equal to 1 if $t_i = t_n$ and 0 otherwise.

Since all customers at a table eat the same dish, if a customer joins an occupied table, he eats

[1] The typical Chinese restaurant process, as described in Section 5.3.2, illustrates the clustering behavior of the Dirichlet process mixture model after integrating out the unknown vector of class proportions (π). Here, the modified finite version describes a Dirichlet mixture by fixing the number of possible dishes at $K < \infty$.

whatever dish was previously chosen for that table. When a customer starts a new table, he must select a dish for that table by randomly choosing one of the K menu items with uniform probabilities. Therefore, the distribution for the nth customer's dish is

$$f(z_n|\alpha, \mathbf{z}_{\tilde{n}}) = \frac{\sum_{i<n} \mathbf{1}(z_i = z_n) + \alpha/K}{n-1+\alpha} \qquad k \leq K, \tag{5.4}$$

where z_i is the dish (class membership) for the ith customer and $\mathbf{z}_{\tilde{n}}$ is the vector of dishes for all but the nth customer.

The Chinese restaurant analogy depicts the clustering behavior of the Dirichlet mixture model. Notably, tables with many customers are more likely to be chosen by subsequent customers. This creates a "rich-get-richer" effect. The marginal distribution of z_n is uniform due to the symmetric Dirichlet prior, but the conditional distribution given $(z_1, \ldots, z_{n-1})$ is skewed toward the class memberships of previous customers. Let $f_{\text{EDF}}(k) = \frac{\sum_{i<n} \mathbf{1}(z_i=k)}{n-1}$ denote the empirical distribution of the first $n-1$ class memberships. Equation 5.4 can be written as a weighted combination of the empirical distribution and the uniform prior:

$$f(z_n|\alpha, \mathbf{z}_{\tilde{n}}) \propto (n-1) f_{\text{EDF}}(z_n) + \alpha \frac{1}{K}. \tag{5.5}$$

Equation (5.5) shows the smoothing behavior of the Dirichlet mixture when the class proportions (π) are integrated out. Specifically, the class weights for the nth observation are smoothed toward the average value of $\frac{1}{K}$. The Dirichlet precision α controls the degree of smoothing. It has the effect of adding α prior observations spread evenly across all K classes. Because the class memberships are fully exchangeable, this equation expresses the conditional distribution for any z_i based on the other class memberships by treating x_i as the last observation.

Recall that the customers' dishes represent their class memberships. To completely specify the mixture distribution, each dish is associated with a parameter value, $\theta_k \overset{\text{i.i.d.}}{\sim} H$. In other words, each customer that eats dish k represents an observation from the kth class with class parameter θ_k. The class parameters are mutually independent and independent of the latent class memberships.

An important property of the DM model is that z_i is bounded by K. At most, K classes will be used to represent the n observations in the sample. The next section explores the behavior of this model when the bound is removed.

5.3 The Dirichlet Process Mixture Model

Consider the issue of deciding how many classes are needed to represent a given sample. One method is to fit latent class models for several values and use diagnostics to compare the fits. Such methods include, among others, cross-validation techniques (Hastie et al., 2009) and penalized likelihood scores such as Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978). Though AIC and BIC are popular choices, their validity for latent class models has been criticized (McLachlan and Peel, 2000). Instead of choosing the single best model complexity, one can use a prior distribution over the number of classes to calculate posterior probabilities (Roeder and Wasserman, 1997). Given a suitable prior, reversible jump Markov chain Monte Carlo techniques can sample from a posterior distribution which includes models of varying dimensionality (Green, 1995; Giudici and Green, 1999). An alternate strategy is to assume that the number of latent classes in the population is unbounded. The Dirichlet process mixture model (DPM) arises as the limiting distribution of Dirichlet mixture models when K approaches infinity. This limit uses a Dirichlet process as the prior for class proportions and parameters. This

section reviews properties of the the Dirichlet process and its relationship to the finite Dirichlet mixture model. These properties elucidate the hierarchical Dirichlet process (Section 5.5), which uses multiple Dirichlet process priors to construct a nonparametric mixed membership model.

5.3.1 The Dirichlet Process

The Dirichlet process is a much-publicized nonparametric process formally introduced by Ferguson (1973). It is a prior law over probability distributions whose finite-dimensional marginals have Dirichlet distributions. Dirichlet processes have been used for modeling Gaussian mixtures when the number of components is unknown (Escobar and West, 1995; MacEachern and Müller, 1998; Rasmussen, 1999), survival analysis (Kim, 2003), hidden Markov models with infinite state-spaces (Beal et al., 2001), and evolutionary clustering in which both data and clusters come and go as time progresses (Xu et al., 2008).

The classical definition of a Dirichlet process constructs a random measure P in terms of finite-dimensional Dirichlet distributions (Ferguson, 1973). Let α be a positive scalar and let H be a probability measure with support Θ. If $P \sim \text{DP}(\alpha, H)$ is a Dirichlet process with precision α and base measure H, then for any natural number K,

$$\big(P(A_1), \ldots, P(A_K)\big) \sim \text{Dir}\big(\alpha H(A_1), \ldots, \alpha H(A_K)\big), \tag{5.6}$$

whenever $(A_k)_{k=1}^K$ is a measurable finite partition of Θ.

Sethuraman (1994) provides a constructive definition of P based on an infinite series of independent beta random variables. Let $\phi_k \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha)$ and $\theta_k \overset{\text{i.i.d.}}{\sim} H$ be independent sequences. Define $\pi_1 = \phi_1$, and set $\pi_k = \phi_k \prod_{j=1}^{k-1}(1-\phi_j)$ for $k > 1$. The random measure $P = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ has distribution $\text{DP}(\alpha, H)$, where δ_x is the degenerate distribution with $f(x) = 1$. This definition of the Dirichlet process is called a stick-breaking process. Imagine a stick of unit length which is divided into an infinite number of pieces. The first step breaks off a piece of length $\pi_1 = \phi_1$. After $k-1$ steps, the remaining length of the stick is $\prod_{i=1}^{k-1}(1-\phi_i)$. The kth step breaks off a fraction ϕ_k of this length, which results in a new piece of length π_k.

The stick-breaking representation shows that $P \sim \text{DP}(\alpha, H)$ is discrete with probability 1. The measure P is revealed to be a mixture of an infinite number of point masses. Hence, there is a positive probability that a finite sample from P will contain repeated values. This leads to the clustering behavior of the following Dirichlet process mixture model (Antoniak, 1974):

$$\begin{aligned} P &\sim \text{DP}(\alpha, H). \\ \theta_i^* | P &\overset{\text{i.i.d.}}{\sim} P & i &= 1 \ldots n. \\ x_i | \theta_i^* &\sim F(\cdot | \theta_i^*) & i &= 1 \ldots n. \end{aligned}$$

Let $(\theta_1, \ldots, \theta_K)$ denote the unique values of the sequence $(\theta_1^*, \ldots, \theta_n^*)$, where K is the random number of unique values. Set z_i such that $\theta_i^* = \theta_{z_i}$. Given z_i and θ, the distribution of the ith observation is

$$F(x_i | z_i, \theta) = F(\cdot | \theta_{z_i}). \tag{5.7}$$

By comparing Equation (5.7) to the Dirichlet mixture model, one can interpret z_i as a class membership, θ as the class parameters, and K as the number of classes represented in the sample. Note that K is random in this model, whereas it is a constant in the DM model. Therefore, the DPM model provides an implicit prior over the number of classes in the sample. Antoniak (1974) specifies this prior explicitly.

To make direct comparisons between Dirichlet process mixtures and Dirichlet mixtures, it is useful to disentangle the distributions of π and θ_k. Let $\pi \sim \text{SBP}(\alpha)$ denote the vector of weights

based on the stick-breaking process. Extend the notation Mult(π) to include infinite multinomial distributions such that $\mathbb{P}(z_i = k) = \pi_k$ for all positive integers k. The DPM model is equivalent to the following hierarchy:

Dirichlet Process Mixture Model

$$\begin{aligned} \pi|\alpha &\sim \text{SBP}(\alpha). \\ \theta_k|\lambda &\overset{\text{i.i.d.}}{\sim} H(\lambda) && k = 1, 2, \ldots. \\ z_i|\pi &\overset{\text{i.i.d.}}{\sim} \text{Mult}(\pi) && i = 1 \ldots n. \\ x_i|z_i, \theta &\sim F(\cdot|\theta_{z_i}) && i = 1 \ldots n. \end{aligned}$$

The above hierarchy directly shows the relationship between the Dirichlet mixture model and the DP mixture model. Where the Dirichlet mixture uses a symmetric Dirichlet prior for π, the DP mixture uses the stick-breaking process to generate an infinite sequence of class weights. In fact, Ishwaran and Zarepour (2002) shows that the marginal distribution induced on $x_1, \ldots, x_n$ by the DM model approaches that of the DPM model as the number of classes increases to infinity. Thus, the Dirichlet process mixture model may be interpreted as the infinite limit of finite Dirichlet mixture models.

5.3.2 Chinese Restaurant Process

A Chinese restaurant process illustrates the clustering behavior of a Dirichlet process mixture model when the unknown class proportions (π) are integrated out (Aldous, 1985). Customers arrive and choose tables as in the finite version for Dirichlet mixtures, but the menu in the full Chinese restaurant process contains an infinite number of dishes.

Recall that in the finite Chinese restaurant process, a customer who sits at an empty table chooses one of the K available dishes using uniform probabilities. For a DP mixture, the menu has an unlimited number of dishes. The discrete uniform distribution is not defined over infinite sets, but this technicality can be sidestepped. Class parameters are assigned independently of each other and the enumeration of the dishes is immaterial. Therefore, dishes do not need to be labeled until after they are sampled. Whatever dish happens to be selected first can be labeled 1, the second dish to be chosen can be labeled 2, and so on. In other words, when sampling a dish, there is no need to distinguish between any of the unsampled dishes. Because there are finitely many sampled dishes and infinitely many unsampled dishes, a "uniform" distribution implies that, with probability 1, the customer selects a new dish from the distribution $H(\lambda)$. Note that if the distribution H has any points with strictly positive probability, there is a chance that the "new" dish chosen by the customer will be the same as an already observed dish. To avoid this technicality and simplify wording, one may assume that H is continuous. The mathematics are the same in either case.

In the Chinese restaurant process, the first customer sits at the first table and randomly chooses a random dish, which is labeled 1. The second customer joins the first table with probability $1/(1+\alpha)$ or starts a new table with probability $\alpha/(1 + \alpha)$. As subsequent customers enter, the probability that they join an occupied table is proportional to the number of people already seated there. Alternatively, they may choose a new table with probability proportional to α.

Suppose that there are T occupied tables when the nth customer enters. The probability distribution for the nth customer's table, t_n, is the same as in the finite Chinese restaurant process (Equation 5.3). In contrast, the distribution for his dish, z_n, is slightly different because each table has a unique dish. Let K be the current number of unique dishes:

$$\mathbb{P}(z_n = k|\alpha, \mathbf{z}_{\widetilde{n}}) = \begin{cases} \frac{\sum_{i<n} \mathbf{1}(z_i=k)}{n-1+\alpha}, & k \leq K \\ \frac{\alpha}{n-1+\alpha}, & k = K+1 \end{cases}. \tag{5.8}$$

Again, $\mathbf{z}_{\widetilde{n}}$ denotes the vector of dishes (class assignments) for all but the nth customer.

To finish specifying the mixture, each dish is associated with a class parameter drawn independently from the base measure H. As in the finite Chinese restaurant process, the class parameter for the nth observation can also be written as a weighted combination of the current empirical distribution and the prior distribution:

$$F(\theta_n) \propto (n-1) f_{\text{EDF}} + \alpha H, \tag{5.9}$$

where $f_{\text{EDF}} = \sum_{i<n} \delta_{\theta_{z_i}} / (n-1)$ denotes the empirical distribution of the first $n-1$ class memberships. Note that the only difference from the finite Chinese restaurant is that Equation (5.9) uses the prior distribution H in place of the prior uniform probability $\frac{1}{K}$ of Equation (5.5).

The Chinese restaurant process illustrates how the population (restaurant) has an infinite number of classes (dishes), but only a finite number are represented (ordered by a customer). Note that each customer selects a random table with probabilities that depend on the precision α, but not the base measure H. Hence, the choice of α amounts to an implicit prior on the number of classes. Antoniak (1974) specifies this prior explicitly. Notably, the number of classes increases stochastically with both n and α. In the limit, as α approaches infinity, each customer chooses an unoccupied table. As a result, there is no clustering and each observation belongs to its own unique class. The distribution of $(\theta_{z_1}, \ldots, \theta_{z_n})$ approaches an i.i.d. sample from H. In the other extreme, as α approaches zero, each customer chooses the first table, resulting in a single class. In effect, the population distribution is no longer a mixture distribution.

5.3.3 Comparison of Dirichlet Mixtures and Dirichlet Process Mixtures

Both the DM model and the DPM model assume that observed data are representatives of a finite number of latent classes. The chief difference being that the DM model places a bound on the number of classes while the DPM model does not. Ishwaran and Zarepour (2002) makes this relationship explicit: as the number of classes increases to infinity, the DM model converges in distribution to the DPM model. Because the number of classes is unbounded, there is always a positive probability that the next response represents a previously unobserved class. The DPM model is an example of a nonparametric Bayesian model, which allows model complexity to increase as more data are observed.

A comparison of Equations (5.8) and (5.4) reveals how the nonparametric DPM model differs from the bounded-complexity DM model. In both models, the distribution of an observation's class is simply the empirical distribution of the previous class memberships, plus additional α prior observations. However, the prior weight is distributed differently. In the DM model, the α prior observations are placed uniformly over the K classes. Once all K classes have been observed, there is no chance of observing a novel class. In the DPM model, the α prior observations are placed on the next unoccupied table, which will serve a new dish with probability 1 (if the base measure H is continuous.) Hence, there is always a non-zero probability that the next observation belongs to a previously unobserved class, though this probability decreases as the sample size increases. While the DPM model allows greater flexibility in clustering, both models yield the same distribution for the observations and class parameters when conditioned on the vector of class memberships.

DP mixtures, and other nonparametric Bayesian models, are one strategy for determining the appropriate model complexity given a set of data. The theory behind these mixtures states that there is the possibility that some classes have not been encountered yet. For prediction, as opposed to estimation, this flexibility may be especially attractive since the new observation may not fit well into any of the current classes.

Recall that the precision α amounts to a prior over the number of classes. The posterior distribution of the latent class memberships provides a way to learn about the complexity from the observations that does not require choosing a specific value. Furthermore, it is possible to expand the DPM model to include a hyperprior for α (Escobar and West, 1995).

5.4 Mixed Membership Models

In the mixture models of Sections 5.2 and 5.3, each individual is assumed to belong to one of K underlying classes in the population. In a mixed membership model, each individual may belong to multiple classes with varying degrees of membership. In other words, each observation is associated with a *mixture* of the K classes. For this reason, mixed membership models are also called *individual-level mixture models*. Mixed membership models have been used for survey analysis (Erosheva, 2003; Erosheva et al., 2007), language models (Blei et al., 2003; Erosheva et al., 2004), and analysis of social and protein networks (Airoldi et al., 2008). This section focuses on models where K is a finite constant, which bounds the number of classes that may be observed.

Consider a population with K classes. Let $H(\lambda)$ denote the prior over class parameters where λ is a hyperparameter. In the DM model, individual i has membership in a single class, denoted z_i. Alternatively, the ith individual's class may be represented as the K-dimensional vector $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$, where π_{ik} is 1 if $z_i = k$ and 0 otherwise. By contrast, a mixed membership model allows π_i to be any non-negative vector whose elements sum to 1. The range of π_i is called the $(K-1)$-simplex. Geometrically, the simplex is a hyper-tetrahedron in $\mathbb{R}^K$. A mixed membership model allows π_i to take any value in the simplex while the DM model constrains π_i to be one of the K vertices.

The Grade of Membership model (GoM) extends the Dirichlet mixture model to allow for mixed membership (Erosheva et al., 2007). Both models can be understood as mixtures of the K possible classes. The DM model has a single population-level mixture for all individuals. In the GoM model, the population-level mixture provides typical values for the class weights, but the actual weights vary between individuals. As with the DM model, the population-level mixture in the GoM model has a symmetric Dirichlet prior. This mixture serves as the expected value for the individual-level mixtures, which also have a symmetric Dirichlet distribution. Denote the Dirichlet precision at the population level by α_0 and the precision at the individual level by α. The GoM model can be expressed by the following hierarchy:

$$
\begin{aligned}
\pi_0|\alpha_0 &\sim \text{Dir}(\alpha_0/K). \\
\theta_k|\lambda &\overset{\text{i.i.d.}}{\sim} H(\lambda) && k = 1 \ldots K. \\
\pi_i|\alpha, \pi_0 &\overset{\text{i.i.d.}}{\sim} \text{Dir}(\alpha\pi_{01}, \ldots, \alpha\pi_{0K}) && i = 1 \ldots n. \\
x_i|\pi_i, \theta &\sim \sum_{k=1}^{K} \pi_{ij} F(\cdot|\theta_k) && i = 1 \ldots n.
\end{aligned}
$$

This model is the same as the DM model, except for the individual-level mixture proportions. The Dirichlet mixture model constrains π_{ik} to zero for all but one class, so the distribution of x_i is a "mixture" of one randomly selected component. By contrast, in a mixed membership model, π_i can take any value in the $(K-1)$-simplex resulting in a true mixture of the K components.

Clearly, the GoM model generalizes the K-dimensional DM model by allowing more flexibility in individual-level mixtures. Conversely, the GoM model can also be described as a special case of a larger DM model. Suppose each individual is measured across J different items. (For simplicity of notation, assume J is constant for all individuals; removing this restriction is trivial.) Erosheva et al. (2007) provides a representation theorem to express a K-class GoM model as a constrained DM model with K^J classes. Therefore, this theorem will assist in building the GoM model into a nonparametric model in much the same way that Section 5.3 built the DM model into the nonparametric DPM model.

Erosheva et al. describe their representation theorem in the context of survey analysis. In this

case, a sample of n people each respond to a series of J survey items and an observation is the collection of one person's responses to all of the items. Let $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$ be the membership vector and let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})$ be the response vector for the ith individual. (Henceforth, the ith observation shall be explicitly denoted as a vector because scalar observations do not naturally fit with the representation theorem.) According to the GoM model, the distribution of $\mathbf{x}_i$ is a mixture of the K class distributions with mixing proportions given by π_i. Alternatively, the GoM model can be interpreted as a Dirichlet mixture model in which individuals can move among the classes for each response. That is, individual i may belong to class z_{ij} in response to item j, but belong to a different class z_{ij^*} in response to item j^*. The probability that individual i behaves as class k for a particular item is π_{ik}. Note that this probability depends on the individual, but is constant across all items. Let z_{ij} denote the class membership for the ith individual in response to the jth item. The distribution of the ith individual's response is determined by $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ})$ and the class parameters (θ). Therefore, individual i may be considered a member of the latent class $\mathbf{z}_i$ with class parameter $(\theta_{z_{i1}}, \ldots, \theta_{z_{iJ}})$. Each of the J components takes on one of K possible classes, making the GoM model a constrained DM model with K^J possible classes. The constraints arise because π_i is constant across all items in the GoM model, whereas the DM model allows all probabilities to vary freely. Thus, the probability of class $\mathbf{z}_i$ is constant under permutation of its elements in the GoM model but not the DM model.

The representation theorem suggests augmenting the GoM model with the collection of latent individual-per-item class memberships:

Grade of Membership Model

$$
\begin{aligned}
\pi_0|\alpha_0 &\sim \text{Dir}(\alpha_0/K). \\
\theta_k|\lambda &\overset{\text{i.i.d.}}{\sim} H(\lambda) && k = 1 \ldots K. \\
\pi_i|\alpha, \pi_0 &\overset{\text{i.i.d.}}{\sim} \text{Dir}(\alpha\pi_{01}, \ldots, \alpha\pi_{0K}) && i = 1 \ldots n. \\
z_{ij}|\pi_i &\sim \text{Mult}(\pi_i) && i = 1 \ldots n,\ j = 1 \ldots J. \\
x_{ij}|z_{ij}, \theta &\sim F(\cdot|\theta_{z_{ij}}) && i = 1 \ldots n,\ j = 1 \ldots J.
\end{aligned}
$$

As with the DM model, each measurement (x_{ij}) is generated by randomly choosing a latent class membership then sampling from the class-specific distribution. In both models, responses are assumed to be independent given the class memberships. The responses in the DM model are fully exchangeable because each one uses the same vector of class weights. The GoM model includes individual-level mixtures that allow for the individual's class weights to vary from the population average. Therefore, responses are exchangeable only if they belong to the same individual. Note that z_{ij} is a positive integer less than or equal to K. The next section builds this latent class representation into a nonparametric model by removing this bound on the value of z_{ij}.

5.5 The Hierarchical Dirichlet Process Mixture Model

Table 5.1 illustrates two analogies that may help elucidate the hierarchical Dirichlet process mixture model (HDPM). Comparing the columns reveals that the relationship between the HDPM model and the DPM model is similar to the relationship between the GoM model and the DM model. Recall that the GoM model introduces mixed memberships to the DM model by introducing priors for individual-level mixtures that allow them to vary from the overall population mixture. In the same way, the HDPM adds mixed memberships to the DPM model through individual-level priors. In both the GoM and HDPM models, the population-level mixture provides the expected value for the

individual-level mixtures. Comparing the *rows* of Table 5.1 shows that the relationship between the HDPM and GoM models is similar to the relationship between the DPM and DM models. In both cases, the former model is a nonparametric version of the latter model that arises as a limiting distribution when the number of classes is unbounded. The HDPM and DPM models are specified mathematically by replacing the symmetric Dirichlet priors of the GoM and DM models with Dirichlet *process* priors.

	Number of Classes	
Exchangeability	Bounded	Unbounded
Full	DM	DPM
Partial	GoM	HDPM

TABLE 5.1
The relationship among the four main models of this chapter.

The hierarchical Dirichlet process mixture model (HDPM) incorporates a Dirichlet process for each individual, $P_i \sim \text{DP}(\alpha, P_0)$, where the base measure P_0 is itself drawn from a Dirichlet process, $P_0 \sim \text{DP}(\alpha_0, H)$. Thus, the model is parametrized by a top-level base measure, H, and two precision parameters, α_0 and α.

$$
\begin{aligned}
P_0|\alpha_0, H &\sim \text{DP}(\alpha_0, H). \\
P_i|\alpha, P_0 &\overset{\text{i.i.d.}}{\sim} \text{DP}(\alpha, P_0) && i = 1 \ldots n. \\
\theta^*_{ij}|P_i &\sim P_i && i = 1 \ldots n,\ j = 1 \ldots J. \\
x_{ij}|\theta^*_{ij} &\sim F(\cdot|\theta^*_{ij}) && i = 1 \ldots n,\ j = 1 \ldots J.
\end{aligned}
$$

Note that $E[P_i] = P_0$. Thus the population-level mixture, P_0, provides the expected value for the individual-level mixtures and the precision α influences how closely the P_is fall to this mean.

A stick-breaking representation of the HDPM model allows it to be expressed as a latent class model. Since P_0 has a Dirichlet process prior, it can be written as a random stick-breaking measure $P_0 = \sum_{k=1}^{\infty} \pi_{0k}\delta_{\theta_{0k}}$, where $\pi_0 \sim \text{SBP}(\alpha_0)$ and θ_0 is an infinite i.i.d. sample with distribution H. Likewise, each individual mixture P_i can be expressed as $P_i = \sum_{k=1}^{\infty} \pi^*_{ik}\delta_{\theta_{ik}}$, where $\pi^*_i \sim \text{SBP}(\alpha)$ and θ_i is an infinite i.i.d. sample from P_0. Because $\theta_{ik} \sim P_0$, it follows that each $\theta_{ik} \in \theta_0$. Therefore, $P_i = \sum_{k=1}^{\infty} \pi_{ik}\delta_{\theta_{0k}}$, where $\pi_{ik} = \sum_{j=1}^{\infty} \pi^*_{ij}\mathbf{1}\left(\theta_{ij} = \theta_{0k}\right)$. P_0 specifies the set of possible class parameters and the expected class proportions; P_i allows individual variability in class proportions. Since the class parameters are the same for all individuals, the notation θ_{0k} may be replaced by the simpler θ_k. While π_0 may be generated using the same stick-breaking procedure used in Section 5.3, the individual-level π_is require a different procedure given by Teh et al. (2006). Given π_0 and θ, let $\phi_{ik} \sim \text{Beta}\Big(\alpha\pi_{0k}, \alpha\left(1 - \sum_{j=1}^{k} \pi_{0j}\right)\Big)$. Define $\pi_{i1} = \phi_{i1}$, and set $\pi_{ik} = \phi_{ik}\prod_{j=1}^{k-1}(1 - \phi_{0j})$ for $k > 1$. The random measure $P_i = \sum_{k=1}^{\infty} \pi_{ik}\delta_{\theta_k}$ has distribution $\text{DP}(\alpha, P_0)$. Denote the conditional distribution of $\pi_i|\pi_0$ by $\text{SBP}_2(\alpha, \pi_0)$. The latent class representation of the HDPM model is as follows:

Hierarchical DP Mixture Model

$$
\begin{aligned}
\pi_0|\alpha_0, H &\sim \text{SBP}(\alpha_0). \\
\theta_k|\lambda &\overset{\text{i.i.d.}}{\sim} H(\lambda) && k = 1 \ldots K. \\
\pi_i|\alpha, \pi_0 &\overset{\text{i.i.d.}}{\sim} \text{SBP}_2(\alpha, \pi_0) && i = 1 \ldots n.
\end{aligned}
$$

$$z_{ij}|\pi_i \sim \text{Mult}(\pi_i) \qquad i = 1 \ldots n,\ j = 1 \ldots J.$$
$$x_{ij}|z_{ij}, \theta \sim F(\cdot|\theta_{z_{ij}}) \qquad i = 1 \ldots n,\ j = 1 \ldots J.$$

The three lowest levels in the HDPM model (pertaining to π_i, z_{ij}, and x_{ij}) represent the item-level distribution. Individual i, in response to item j, chooses a class according to its unique mixture: $z_{ij} \sim \text{Mult}(\pi_i)$. Its response is then given according to the distribution for that class: $x_{ij} \sim F(\cdot|\theta_{z_{ij}})$. This behavior is the same as the Grade of Membership model. Each individual is associated with a unique mixture over a common set of classes. The individual's mixture defines the probability that individual i behaves as class k in response to item j. Since this probability does not depend on j, two responses are exchangeable if and only if they come from the same individual. Unlike the GoM model, the HDPM model does not bound the number of classes a priori.

5.5.1 Chinese Restaurant Franchise

Teh et al. (2006) uses the analogy of a Chinese restaurant *franchise* to describe the clustering behavior of the hierarchical Dirichlet process. As each individual has a unique class mixture, each is represented by a distinct restaurant. Each of the customers represents one of the individual's features. For example, in the context of survey analysis, a customer is the individual's response to one of the survey items. The restaurants share a common menu with an infinite number of dishes to represent the various classes in the population. Each restaurant operates as an independent Chinese restaurant process with respect to seating arrangement, but dishes for each table are chosen by a different method which depends on the entire collection of restaurants.

Let x_{ij} denote the jth customer at the ith restaurant. When the customer enters, he chooses a previous table based on how many people are sitting there, or else starts a new table. The distribution for the table choice is the same as the Chinese restaurant process. It is given by Equation (5.3), taking t_n to denote the new customer's table and $t_1, \ldots, t_{n-1}$ to denote the previous customers' tables, where the numbering is confined to tables at restaurant i.

If the customer sits at a new table, he must select a dish. As with table choice, the customer will choose a previously selected dish with a probability that depends on how popular it is. Specifically, the probability is proportional to the number of other tables currently serving the dish across the entire franchise. Alternatively, with probability proportional to α_0, the customer will choose a new dish.

Suppose there are T tables currently occupied in the entire franchise when a customer decides to sit at a new table, becoming the first person at table $T+1$. Denote the dish served at table t by d_t and let K denote the current count of unique dishes. If a new table is started, the distribution for the next dish, d_{T+1} is

$$\mathbb{P}(d_{T+1} = k|d_1, \ldots, d_T) = \begin{cases} \frac{\sum_{t=1}^{T} \mathbf{1}(d_t=k)}{T+\alpha_0}, & k \leq K \\ \frac{\alpha_0}{T+\alpha_0}, & k = K+1 \end{cases}. \tag{5.10}$$

Note that the customer has three choices: he may join an already occupied table (i.e., choose locally from the dishes already being served at restaurant i); start a new table and choose a previous dish (i.e., choose a dish from the global menu); or start a new table and select a new dish. Let z_{iJ} denote the dish chosen by the Jth customer at restaurant i. The Chinese restaurant franchise shows that the distribution of z_{iJ} is comprised of three components:

$$\mathbb{P}(z_{iJ} = k|\alpha_0, \alpha, \mathbf{z}_{i\widetilde{J}}) = \begin{cases} 5\frac{\sum_{j=1}^{J-1} \mathbf{1}(z_{ij}=k)}{J-1+\alpha} + \frac{\alpha}{J-1+\alpha} \frac{\sum_{t=1}^{T} \mathbf{1}(d_t=k)}{T+\alpha_0}, & k \leq K \\ \frac{\alpha}{J-1+\alpha} \cdot \frac{\alpha_0}{T+\alpha_0}, & k = K+1 \end{cases}, \tag{5.11}$$

where T is the current number of occupied tables, K is the current number of distinct dishes across the entire franchise, and $\mathbf{z}_{i\widetilde{J}}$ is the set of all dish assignments except for z_{iJ}. Note that the weight

for a dish $k \leq K$ is the sum of the number of *customers at restaurant* i eating that dish plus α times the number of *tables in the franchise* serving that dish. In other words, dishes are chosen according to popularity, but popularity within restaurant i is weighted more heavily. In practical terms, measurements from the same individual share information more strongly than measurements from multiple individuals. The precision α specifies the relative importance of the local-level and global-level mixtures.

To finish specifying the HDPM model, each dish is associated with a class parameter drawn independently from the base measure H. The distribution of $\theta_{z_{iJ}}$ is a three-part mixture. Let $F_{\text{Ind}} = \sum_{j<J} \delta_{\theta_{z_{ij}}}/(J-1)$ be the empirical distribution of class parameters based on the customers at restaurant i. Let $F_{\text{Pop}} = \sum_t \delta_{\theta_{d_t}}/T$ be the empirical distribution based on the proportion of tables serving each dish across the entire restaurant. The distribution of $\theta_{z_{iJ}}$ is

$$F\left(\theta_{z_{iJ}}|\alpha, \alpha_0, \theta_{z_{\widetilde{iJ}}}\right) \propto (J-1)F_{Ind} + \alpha T F_{Pop} + \alpha\alpha_0 H, \tag{5.12}$$

where $\theta_{z_{\widetilde{iJ}}}$ denotes the class parameters for all customers except for the Jth customer of the ith restaurant. As with the other three mixture models, DM, DPM, and GoM, the responses are assumed to be independent given the class memberships. Hence, Equations (5.11) and (5.12) can be applied to any customer by treating him as the last one.

Equation (5.12) illustrates the role of the two precision parameters. Larger values of α place more emphasis on the population-level mixture, so that individual responses will tend to be closer to the overall mean. Meanwhile, larger values of α_0 place more emphasis on the base measure H. This specifies the prior distribution of $\theta_{z_{iJ}}$. After observing the response x_{iJ}, the class weights are updated by the likelihood of observing x_{iJ} within each class. Unfortunately, this model requires fairly complicated bookkeeping as to track the number of customers at each table and the number of tables serving each menu item. Blunsom et al. (2009) proposes a strategy to reduce this overhead. Inference is discussed more fully in Section 5.7.

5.5.2 Comparison of GoM and HDPM models

The relationship between the HDPM and GoM models is very similar to the relationship between the DP and DPM models described in Section 5.3.3. Both the GoM and HDPM models assume that observed data are representatives of a finite number of classes. Whereas the DM and DPM models assume that the observations are fully exchangeable, the GoM and HDPM models are mixed membership models that treat some observations as more similar than others. For example, in survey analysis, two responses from one individual are assumed to be more alike than responses from two different individuals. In text analysis, the topics within one document are assumed to be more similar than topics contained in different documents.

The HDPM model is similar to the GoM model as both combine individual-level and population-level mixtures. The chief difference between the GoM and HDPM models is that the GoM model bounds the number of classes a priori while the HDPM model does not. Teh et al. (2006) makes this relationship explicit. It shows that the HDPM model is the limiting distribution of GoM models when the number of classes approaches infinity. Because the number of classes is unbounded, there is always a positive probability that the next response represents a previously unobserved class.

The clustering behaviors of the HDPM and GoM models are very similar. In the GoM model, the individual-level mixtures (π_i) are shrunk toward the overall population-level mixture (π_0), which is itself shrunk toward the uniform probability vector, $\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$. The Dirichlet process priors in the HDPM model exhibit a similar property, where the individual-level P_is are shrunk toward the overall population-level mixture, P_0. Whereas the GoM model shrinks π_0 toward the uniform prior over the K classes, the HDPM model shrinks P_0 toward a prior base measure H. The result is that the HDPM model always maintains a strictly positive probability that a new observation is assigned to a new class. This is illustrated by the Chinese restaurant franchise, with the exact probability

of a new class given by Equation (5.11). In other words, the DM and GoM models place a finite bound on the observed number of classes, but the DPM and HDPM models are nonparametric models that allow the number of classes to grow as new data are observed. While the HDPM allows greater flexibility in clustering than the GoM model, both models yield the same distribution for the observations and class parameters, when conditioned on the vector of class memberships.

Recall that the precision α in the DPM model amounts to a prior over the number of observed classes. In the HDPM model, this prior is specified by two precision parameters: α_0 at the population level and α at the individual level. The posterior distribution of the latent class memberships provides a way to learn about the complexity of the data without choosing a specific value. Teh et al. (2006) extends the HDPM model with hyperpriors for both α_0 and α to augment the model's ability to infer complexity from the data. Section 5.6 explores the role of α_0 and α intuitively with simulation studies for illustration.

5.6 Simulated HDPM Models

The Chinese restaurant process can be used to construct simple simulations of HDPM models. Such simulations can reveal how the the model is affected by changes in the tuning parameters or sample size. Specifically, the simulations in this section illustrate the behavior of mixtures at both the population level and individual level, as well as the similarity between different individuals.

5.6.1 Size of the Population-Level Mixture

Figure 5.1 shows how the number of population classes in the HDPM model is affected by sample size and the two precision parameters. The values result from simulation of a Chinese restaurant franchise in which each restaurant receives 16 customers (e.g., each individual responds to 16 survey items.) With α and n held fixed, the average size (number of components) of the population mixture increases as α_0 increases from 1 to 100. The average mixture also increases with α when n and α_0 are fixed, but the difference is not significant except when $\alpha_0 = 100$. Thus, both precisions affect the expected number of classes, but α_0 may limit or dampen the effect of α. Intuitively, a large value of α causes more customers to choose a new table, but a low value for α_0 means that they frequently choose a previously ordered dish. Hence, new classes will be encountered infrequently. Indeed, as α_0 approaches 0, the limit in the number of classes is 1, regardless of the value of α. Standard errors in mixture size were also estimated by repeating each simulation 100 times. The effect of α_0 and α on the standard error is similar to the effect on the mean mixture size (see Figure 5.2).

5.6.2 Size of Individual-Level Mixtures

Figures 5.3–5.4 show that the effect of the HDPM parameters on the size of the individual mixtures is similar to their effect on the population mixture size. As seen in Figure 5.3, the average number of classes in each individual mixture increases with both α and α_0. The first third of the chart comes from simulations in which each individual responds to 16 survey items. For the second third, there are 64 items per individual, and there are 100 items per individual in the last third. There is a clear interaction between the precision parameters and the number of survey items. The size of individual mixtures increases with the number of responses per individual. In terms of the Chinese restaurant franchise, more tables are needed as more customers enter each restaurant. Interestingly, this effect is influenced by the two precision parameters. The mixture size increases more dramatically for large values of α and α_0. Figure 5.4 shows how the standard deviation among the size of the individual mixtures is affected by the number of survey items and the precision parameters. The

effect on variability is similar to, but weaker than, the effect on average mixture size. In most cases, the variability in the size of individual mixtures is quite small compared to the average. Note that this analysis compares only the *number* of classes represented by the individuals. This does not take into account the number of shared dishes between restaurants nor the variability in their class proportions (the π_is).

5.6.3 Similarity among Individual-Level Mixtures

In the HDPM model, each individual is associated with a unique mixture of the population classes. The similarity among the individuals is influenced by both α_0 and α, though in opposite manners. Large values of α lead to individual mixtures being closer to the population, and hence, each other. Therefore, similarity among the individuals tends to increase with α. On the other hand, similarity tends to decrease as α_0 increases. Intuitively, the number of classes in the entire sample tends to be small when α_0 is small. Hence, individual mixtures select from a small pool of potential classes. This leads to high similarity between individuals. When α is large, individual mixtures select from a large pool of potential classes and tend to be less similar. Indeed, as α_0 approaches infinity, the class parameters tend to behave as i.i.d. draws from the base measure H. If H is non-atomic, then the individual mixtures will not have any components in common. In the other extreme, as α_0 approaches zero, the number of classes in the sample tends to 1. This results in every individual "mixture" being exactly the same, with 100% of the weight on the sole class.

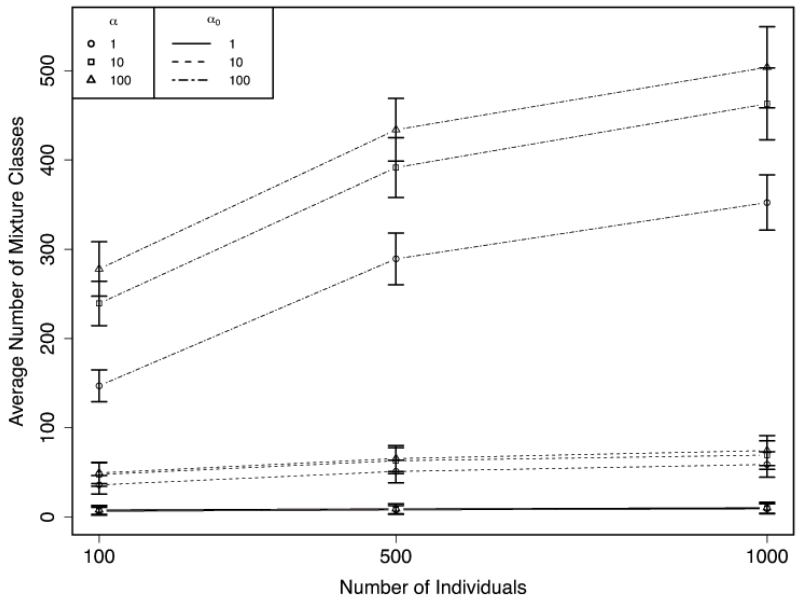

FIGURE 5.1
The effect of prior precisions and the number of individuals on the expected population mixture size in HDPM models. The population mixture size is the number of classes represented across all individuals in response to $J = 16$ measurements. Error bars represent two standard errors. Estimates are based on 100 simulations of each model.

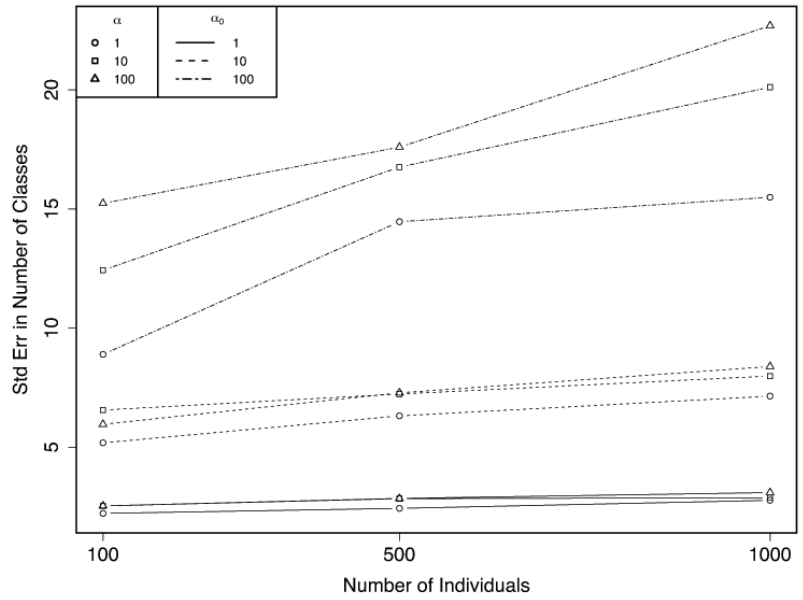

FIGURE 5.2
Standard errors for the population mixture size for various sample sizes and precisions in the HDPM model, based on 100 simulations per model.

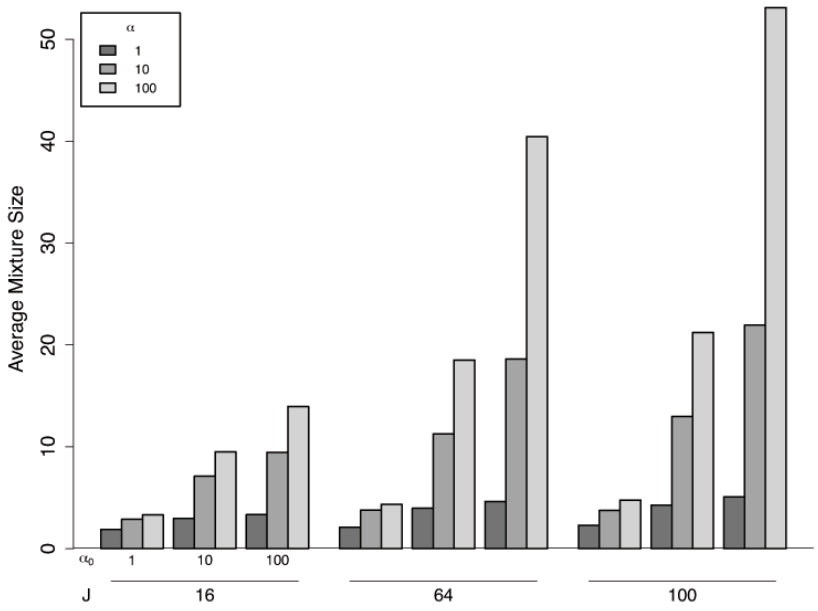

FIGURE 5.3
The effect of prior precisions and number of measurements per individual on the average individual-level mixture size. The mixture size represents the number of classes an individual represented during J measurements. Based on 100 simulations per model.

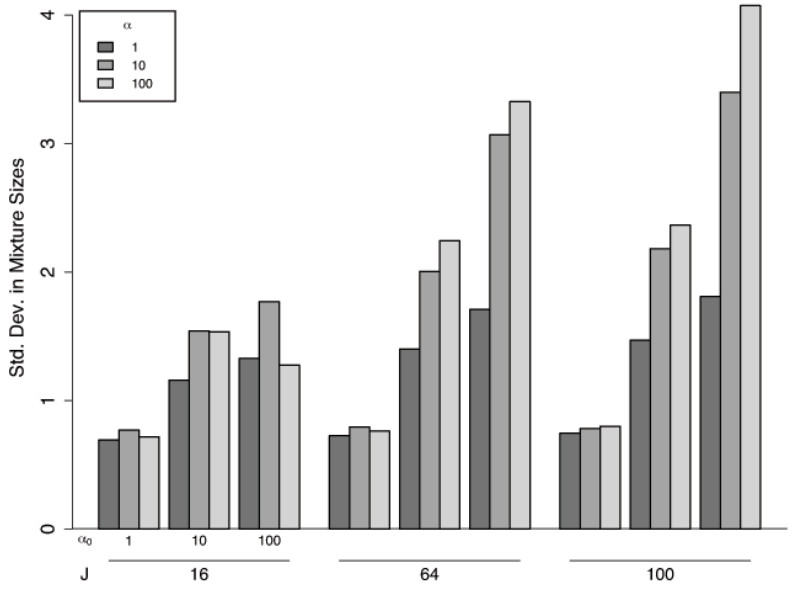

FIGURE 5.4
Standard deviation in individual-level mixture size for various HDPM models. Mixture size measures the number of classes an individual represented during J measurements. Based on 100 simulations per model.

The effect of the two precision parameters on similarity is shown in Figure 5.5 based on 100 simulations of $n = 50$ individuals responding to $J = 16$ items. There are several reasonable choices for measuring similarity. Here, the similarity between two individuals is defined by

$$\text{Sim}(i, i') = \frac{\sum_{k=1}^{K} \min\left(n_{ik}, n_{i'k}\right)}{J}, \tag{5.13}$$

where J is the number of items per individual (16 in this case), K is the total number of classes in the sample, and n_{ik} is the number of times individual i responded to a survey item as a member of class k. In effect, $\text{Sim}(i, i')$ counts the number of times both individuals represented the same class, after arranging the second individual's responses to maximize the overlap with the first individual. Rearrangement is valid because responses from each individual are exchangeable.

Note that none of the heatmaps in Figure 5.5 exhibit any strong structure. This is expected under the HDPM model since the individuals are conditionally independent given the population-level mixture. On the other hand, one can easily see that the similarity among individuals in a particular model increases as α increases and as α_0 decreases. This exactly matches the intuition explained above.

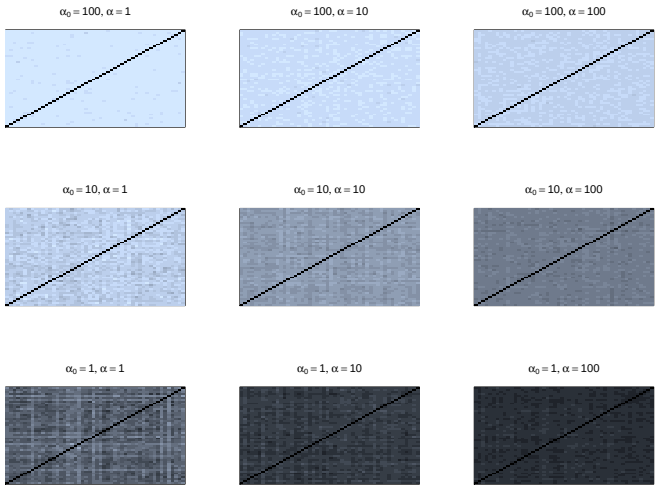

FIGURE 5.5
The effect of HDPM precision parameters on the similarity of individual-level mixtures. Darker areas correspond to higher similarity. Similarity is averaged across 100 simulations.

5.7 Inference Strategies for HDP Mixtures

Two broad categories of inference strategies for nonparametric mixture models are Markov chain Monte Carlo (MCMC) sampling and variational inference. Sampling techniques have the advantage of converging to the correct answer, at least under certain circumstances. Unfortunately, these techniques often require a great deal of computation time and it can be very difficult to assess convergence. Convergence for variational inference can be achieved quickly and assessed easily,

but at the cost of some bias. Some simulation experiments have shown that the bias is not too drastic, at least when H is in the exponential family (Blei and Jordan, 2006).

5.7.1 Markov Chain Monte Carlo Techniques

Escobar and West (1995) demonstrate a Gibbs sampling scheme to estimate the posterior distribution for the DPM model, including inference for the precision term α. They directly sample from $f(\theta_{z_n}|\theta_{z_{\tilde{n}}}, \alpha)$, where $\theta_{z_{\tilde{n}}}$ denotes the class parameters for all observations except the nth one. Unfortunately, Markov chains built on this representation are slow to converge. In order for a class parameter to change, each member of that class must move to a new or different class one at a time. Thus, in order to remove a class or create a new one, there are low-probability intermediate states in which observations are in their own class. A more efficient strategy is to represent θ_{z_n} as the class parameters (θ_k) and class memberships (z_i) (MacEachern, 1994) . This strategy is sometimes called the "collapsed" Gibbs sampler. Class assignments can be updated by combining the prior probabilities from the Chinese restaurant process with the likelihood of x_i given the class parameters. Let K denote the current number of classes in the mixture. For $k \leq K + 1$, let $f_{\text{CRP}}(k)$ be the probability that $z_n = k$ conditioned on the rest of the class memberships under the Chinese restaurant process (Equation 5.8). The probability that observation n should be assigned to class $k \leq K$ is

$$\mathbb{P}(z_n = k|\alpha, \mathbf{z}_{\tilde{n}}) \propto f_{\text{CRP}}(k) f(x_n|\theta_k), \tag{5.14}$$

where $\mathbf{z}_{\tilde{n}}$ denotes all class assignments except for z_n. The probability that x_n should be assigned to a new class is

$$\mathbb{P}(z_n = K + 1|\alpha, \mathbf{z}_{\tilde{n}}) \propto f_{\text{CRP}}(K + 1) \int_{\Theta} f(x_i|\theta) dH(\theta). \tag{5.15}$$

Since the observations are exchangeable, these equations can be used for any z_i by treating x_i as the last observation. Once the class membership vectors are updated, the class parameter θ_k can be updated from the posterior distribution given the prior H and the set of observations currently assigned to class k, denoted by $A_k = \{i : z_i = k\}$:

$$f(\theta_k) \propto \prod_{x_i \in A_k} f(x_i|\theta_k) dH(\theta_k). \tag{5.16}$$

In cases where two or more classes share similar structure, Jain and Neal (2004) proposes a "split-merge algorithm" that allows larger jumps in MCMC updates. This algorithm uses a Metropolis-Hastings step to potentially split one class into two or merge two classes into one. For the DPM model, MCMC sampling is fairly straightforward if the base measure $H(\theta)$ is conjugate to $F(\cdot|\theta)$. This conjugacy is important for two reasons. First, the probability of moving x_i to a new class depends on the integral $\int_{\Theta} f(x_i|\theta) dH(\theta)$. Second, in the collapsed Gibbs sampler, conjugacy leads to simple updates of θ_k given the observations in class k. Strategies for non-conjugate H include the "no gaps" algorithm, which augments the latent class representation with empty classes (MacEachern and Müller, 1998), and a split-merge algorithm for non-conjugate base measures (Jain and Neal, 2007).

For the HDPM model, Gibbs sampling is more complex due to the larger amount of bookkeeping required. In order to update z_{ij}, it is necessary to keep track of how many tables have dish k, how many customers are at each of those tables, and which restaurant the tables are in. This can lead to heavy memory requirements in large datasets. Blunsom et al. (2009) proposes a more efficient representation based on the idea of histograms. For each dish k and each positive integer m, they simply maintain a count of how many tables with m customers are serving dish k. This representation takes advantage of the exchangeability properties of the HDPM model. Due to the fact

that responses are independent given the latent classes, it does not matter which table a customer actually sits at. When a customer joins a table, the appropriate bin count is decremented and the bin above is incremented. For example, if a customer is assigned to table 9, which has two previous customers, then there are now three customers at table 9. Thus, there is one fewer table with two customers and one more table with three customers. When a customer leaves a table, the opposite happens. The appropriate bin is decremented and the bin below is incremented.

Once the mechanism for implementing the Chinese restaurant franchise is decided, MCMC sampling can proceed as in the DPM model. That is, the latent class assignments (z_{ij}) and class parameters (θ_k) can be alternately updated. Let K be the current number of classes. For $k \leq K+1$, let $f_{\text{CRF}}(k)$ be the probability that $z_n = k$ given the rest of the class assignments under the Chinese restaurant franchise (Equation 5.11). The probability that observation n should be assigned class $k \leq K$ is

$$\mathbb{P}(z_n = k|\alpha, \alpha_0, z_{\widetilde{iJ}}) \propto f_{\text{CRF}}(k) f(x_n|\theta_k), \tag{5.17}$$

where $z_{\widetilde{iJ}}$ denotes all class assignments except for z_{iJ}. The probability that x_n should be assigned to a new class is

$$\mathbb{P}(z_n = K+1|\alpha, \alpha_0, z_{\widetilde{iJ}}) \propto f_{\text{CRF}}(k) \int_\Theta f(x_n|\theta) dH(\theta). \tag{5.18}$$

Note that the updates are the same as in the DPM model, except that $f_{\text{CRP}}(k)$ is replaced by $f_{\text{CRF}}(k)$.

Since the observations are independent given the latent class assignments, these equations can be used for any z_{ij} by treating x_{ij} as the last observation. Once the class parameters are updated, the class parameter θ_k can be updated in the same way as in the DPM model. Namely, the new value of θ_k is randomly generated from its posterior distribution given the prior H and the set of observations assigned to class k as in Equation (5.16).

5.7.2 Variational Inference

Variational inference can be viewed as an extension of the expectation maximization algorithm (EM) (Beal, 2003). Whereas EM uses an iterative approach to find a point estimate for some vector of unobserved variables (e.g., latent variables and parameters), variational inference attempts to approximate their entire posterior distribution.

Let θ and $\mathbf{z}$ be the sets of model parameters and latent variables. In DPM and HDPM models, direct calculation of $f(\theta, \mathbf{z}|\mathbf{x})$ is impractical due to the intractable calculation of the data marginal. The intractability arises from the complex interactions among parameters and latent variables. The variational approach is to constrain the posterior to some simpler family of *variational functions* that treat these values as independent. The posterior is approximated by finding the variational function closest to the true posterior (e.g., in KL divergence). Because the variational functions break the dependence between some variables, it is possible to minimize the divergence by iteratively optimizing one piece of the function at a time, given the rest of the function. For example, one may constrain $f(\theta, \mathbf{z}|\mathbf{x})$ to be of the form $q_\theta(\theta|\mathbf{x}) \cdot q_z(\mathbf{z}|\mathbf{x})$. This can be optimized using coordinate ascent by iteratively updating q_θ and q_z based on the value of the other function.

Blei and Jordan (2006) provides an explicit algorithm for DP mixtures when the base measure H is exponential family . Teh et al. (2008) describes a variational approach for hierarchical models that can be used for mixed membership models. The latent variables in the DP mixture are the class proportions (π_k), class parameters (θ_k), and class assignments (z_k). Rather than work with the class proportions, Blei and Jordan work directly with (ϕ_k), the beta random variables from the stick-breaking process. In order to update the variational functions, they also limit the number of components in the variational function to a finite number, say T. However, they optimize the

KL-divergence between this truncated stick-breaking measure and the full DP posterior with infinite components. This yields a set of variational functions parametrized by:

$$q(\phi, \theta, \mathbf{z}) = \prod_{t=1}^{T-1} q_{\gamma_t}(\phi_t) \prod_{t=1}^{T} q_{\tau_t}(\theta_t) \prod_{i=1}^{n} q_{\rho_i}(z_i), \tag{5.19}$$

where each q_{γ_t} is a beta distribution, each q_{τ_t} is in the same family as the prior H, and each $q_{\rho_i}(z_i)$ is multinomial. Notice that the variational function for each variable is the same family as its marginal under the true posterior, however the variational function treats all variables as independent.

The variational function updates proceed like posterior updates given the data and the current value of the other functions. Blei and Jordan (2006) provide explicit updates for each function in the case where H is exponential family. They compared this variational algorithm to the collapsed Gibbs sampler and found that the log-probability of held-out data was similar, but that the variational approach required less computation time. Furthermore, the computation time for variational inference did not increase dramatically in the range of 5- to 40-dimensional observations.

Variational inference is even more efficient if some dimensions of the parameter space can be integrated out. For example, if inferential goals do not include recovering the full mixture posterior, it is possible to integrate out the mixing proportions (Kurihara et al., 2007). This still allows posterior analysis of class membership and parameters as well as calculation of a lower bound for the data marginal. Teh et al. (2008) extends this *collapsed* algorithm to hierarchical Dirichlet process mixtures.

One of the advantages of nonparametric models is that they allow the complexity of the model to grow as new data are observed. This property may be especially advantageous for *streaming* applications, for which new data continually arrive. Online variational inference algorithms have been developed for mixed membership models (Canini et al., 2009; Hoffman et al., 2010; Rodriguez, 2011) including the HDPM model (Wang et al., 2011).

5.7.3 Hyperparameters

The parameters for the DPM and HDPM models include the precisions for Dirichlet processes at each level of mixing and possible hyperparameters for the base distribution H. For example, if H is a normal distribution, a hyperprior may be used to learn about its mean and variance. Typically, hyperpriors are at least used for the precision parameters α_0 and α, since inference can be sensitive to these choices. For example, in one of the first practical applications of the DPM model, Escobar and West (1995) show that the posterior distribution over K is quite sensitive although the predictive distribution is robust. To decrease sensitivity, they recommend using diffuse gamma hyperpriors for precision parameters. Gamma hyperpriors are convenient because the induced posterior for α given the data and latent variables depends only on the number of classes. Thus, the value of α can be updated efficiently based on the current value of the other latent and observed variables.

5.8 Example Applications of Hierarchical DPs

Erosheva et al. (2007) applies the GoM model to data from the National Long Term Care Survey. Alternatively, the HDPM model provides a nonparametric approach to the same data. For each individual, the survey contains binary outcomes on 6 "Activities of Daily Living" (ADL) and 10 "Instrumental Activities of Daily Living" (IADL). ADL items include basic activities required for personal care, such as eating, dressing, and bathing. IADL items include basic activities necessary to reside in the community such as doing laundry, cooking, and managing money. Positive responses

(disabled) to each item signify that during the past week the activity was not completed or not expected to be completed without the assistance of another person or equipment. Each survey response is regarded as an independent Bernoulli random variable: $X_{ij} \sim \text{Bern}(\theta_{ij})$, where X_{ij} is the response of the ith individual on the jth item and θ_{ij} is the probability of a positive response. In this context, a mixture model asserts that the population consists of various sub-groups with varying probabilities of a positive (disabled) response. For example, the population may contain healthy, mildly disabled, and disabled cohorts with increasing probabilities of positive responses.

The GoM model combines mixture models for both the population and individual level. The individual-level mixture asserts, for example, that an individual may behave as a member of the healthy cohort in response to item 1, but behave as a member of the disabled cohort in response to item 2. Each individual is associated with unique mixture probabilities. The population mixture defines the overall proportions of the various cohorts across all individuals and items.

Replacing the GoM model with the HDPM model yields a similar structure, except that the number of classes does not need to be specified a priori.

Blei et al. (2003) presents a mixed membership model for modeling documents called latent Dirichlet allocation (LDA). The classes are various topics (e.g., computer science, operating systems, and machine learning). Each topic is considered a multinomial distribution over some finite vocabulary. The class parameters are the multinomial proportions, which are smoothed using a Dirichlet prior. Each document in the sample is associated with a unique mixture of topic proportions. A word is generated by selecting a topic from the document-level mixture, then choosing a word from the topic-specific multinomial. As with the Grade of Membership model, the number of classes (topics) must be specified a priori. Alternatively, one can use a hierarchical Dirichlet process mixture, in which the number of potential topics is countably infinite (Teh et al., 2006). Under this nonparametric mixed membership model, each new word has a positive probability of belonging to a new topic. Hoffman et al. (2008) uses a similar model to measure musical similarity, where the documents are musical pieces and the "topics" are features.

5.8.1 The Infinite Hidden Markov Model

In the hidden Markov model, a sequence of observations $(x_1, x_2, \ldots, x_n)$ are explained by a second sequence of latent variables $(y_1, y_2, \ldots, y_n)$. The latent sequence is modeled by a Markov chain and the observation (or *emission*) at time t is assumed to depend only on the state of the chain at time t. Hidden Markov models assume fixed finite numbers for both the number of latent states and the number of possible emissions. Each state s is associated with a vector of transition probabilities, $\pi_s^T = (\pi_{s1}^T, \ldots, \pi_{sK}^T)$, where $\pi_{sk}^E = \mathbb{P}(y_{t+1} = k | y_t = s)$; and a vector of emission probabilities, $\pi_s^E = (\pi_{s1}^T, \ldots, \pi_{sV}^T)$, where $\pi_{sv}^E = \mathbb{P}(x_t = v | y_t = s)$.

A hidden Markov model can be specified as a mixed membership model by taking the latent states as the possible classes. The vectors π_s^T and π_s^E define mixtures over the state-space and emission space; since y_{t+1} and x_t are conditionally independent given y_t, one may consider each mixture separately. Denoting the number of possible states by K, a mixed membership model for the transitions can be defined by Dirichlet priors:

$$
\begin{aligned}
\pi_0^T &\sim \text{Dir}(\alpha_0^T / K). \\
\pi_s^T &\overset{\text{i.i.d.}}{\sim} \text{Dir}(\alpha^T \cdot \pi_0^T) && s = 1 \ldots K. \\
y_{t+1} | y_t &\sim \text{Mult}(\pi_{y_t}^T) && t = 1 \ldots n.
\end{aligned}
$$

The state-dependent vectors, π_s^T, allow each state to have unique transition probabilities, which are

shrunk toward the population-level weights, π_0^T. Separate Dirichlet priors can be used to define a mixed membership model for emissions, with V denoting the number of possible values:

$$\begin{aligned}
\pi_0^E &\sim \text{Dir}(\alpha_0^E/V). \\
\pi_s^E &\overset{\text{i.i.d.}}{\sim} \text{Dir}(\alpha^E \cdot \pi_0^E) && s = 1 \dots K. \\
x_t|y_t &\sim \text{Mult}(\pi_{y_t}^E) && t = 1 \dots n.
\end{aligned}$$

As with the transition vectors, each state has unique emission probabilities, which are shrunk toward the population averages. Beal et al. (2001) developed a nonparametric version of HMMs by replacing the Dirichlet priors with Dirichlet process priors. As there are a countable infinite number of potential states and emissions, they call this model the *infinite* hidden Markov model (iHMM). The authors apply this model to a language processing problem. The latent state y_t denotes a topic which specifies a multinomial distribution for the tth word. Because both the transition and emission models are nonparametric, there is a non-zero probability that the Markov chain transitions to a new topic, or that a topic produces a previously unobserved word.

5.9 Other Nonparametric Mixed Membership Models

5.9.1 Multiple-Level Hierarchies

The four main models in this chapter: DM in Section 5.2, DPM in Section 5.3, GoM in Section 5.4, and HDPM in Section 5.5 all produce exchangeability within any given mixture. The individuals $(\mathbf{x}_i)$ are exchangeable in all models and the per-item responses (x_{ij}) are also exchangeable in the GoM and HDPM models. If this exchangeability structure is unrealistic or undesired, one way to introduce dependence is to include multiple levels of hierarchy. For example, in the National Long Term Care Survey, some responses concern "Activities of Daily Living" and others concern "Instrumental Activities of Daily Living." In theory, an individual's class membership probabilities could vary depending on the sub-category. This can be modeled by including an extra layer of Dirichlet process mixing with S denoting the number of sub-categories:

$$\begin{aligned}
P_0 &\sim \text{DP}(\alpha_0, H). \\
P_i|P_0 &\overset{\text{i.i.d.}}{\sim} \text{DP}(\alpha_1, P_0) && i = 1 \dots n. \\
P_{is}|P_i &\sim \text{DP}(\alpha_2, P_{ij}) && s = 1 \dots S. \\
\theta_{isj}|P_{is} &\sim P_{is} && i = 1 \dots n,\ s = 1 \dots S,\ j = 1 \dots J. \\
X_{isj}|\theta_{isj} &\sim F(X|\theta_{isj}) && i = 1 \dots n,\ s = 1 \dots S,\ j = 1 \dots J.
\end{aligned}$$

This model includes mixtures at the population level (P_0), at the individual level (P_i), and at the sub-category level for each individual (P_{is}). The degree to which any two responses share information is determined by how many hierarchy levels separate them (as well as the relevant precision parameters). As before, responses are fully exchangeable within each mixture.

A double hierarchy may also be appropriate if individuals come from multiple sub-populations. For example, one could divide individuals based on type of residence: apartment, house, or nursing home. In this case, the HDPM model would include mixtures at the population level, sub-population

level (type of residence), and individual level. Furthermore, if items are divided into various categories, then it is possible to include a fourth level of the hierarchy to account for this. In theory, any number of levels is possible, although the number of latent variables needed to represent a mixture model grows with each new level. Teh et al. (2006) provides an example application of a three-level HDPM model that they use to analyze articles from the proceedings of the *Neural Information Processing Systems (NIPS)* conference from the years 1988 to 1999. The articles are divided into nine sections, such as "algorithms and architectures" and "applications." Documents from the same section are expected to have a similar distribution of topics. Their model incorporates topic mixtures at the document level, section level, and population level. Here, the population is the entire collection of documents. The section level mixture allows a document to share information about topics more closely with documents within the same section than with documents in other sections.

5.9.2 Dependent Dirichlet Processes

In some cases, the problem is not to create a desired exchangeability structure but to induce correlation between different mixture components. This may be the case when covariates are measured for each observation. Exchangeability implies that there is no a priori difference among the possible covariate values. This may be appropriate for nominal variables such as gender or ethnicity. In this case, the effects of the covariate can be accounted for using additional hierarchy levels as described above. On the other hand, exchangeability may not be appropriate for ordinal or continuous variables, such as years of experience or age. Dependent Dirichlet processes have been developed for these types of covariates. For example, spatial Dirichlet processes have been used when the "individuals" are points in space (Gelfand et al., 2005; Duan et al., 2007). Such models produce Dirichlet process mixtures at each point, such that the mixtures are more similar when points are closer together. Temporal versions of dependent Dirichlet processes have also been developed which allow a nonparametric mixture to evolve over time (Xu et al., 2008; Ahmed and Xing, 2008). Although applications of dependent Dirichlet processes have focused on extensions to the DPM model, they provide potential sources for new nonparametric mixture models when hierarchical versions are developed.

Exchangeability may also be undesirable if one believes that certain classes tend to co-occur more often than other classes. Teh et al. (2006) uses HDPM models to describe documents (the observations) as a mixture of various topics (the classes). In sufficiently broad collections of documents, one may find that certain topics often appear together. For example, a document that focuses on the topic "politics" may be more likely to include the topic "economics" and less likely to include the topic "baseball." In other words, the occurrence of politics and economics may be positively correlated whereas the occurrence of politics and baseball may be negatively correlated. Unfortunately, the exchangeability property of the HDPM model prevents it from explicitly describing this correlation. Paisley et al. (2012) replaces the hierarchical Dirichlet process with a prior that they call the discrete infinite logistic normal distribution. This prior produces a mixed membership model that is able to explicitly describe correlated topics. Paisley et al. uses this prior to model a collection of 10,000 documents from Wikipedia.

5.9.3 Pitman-Yor Processes

The Pitman-Yor process (Pitman and Yor, 1997), or two-parameter Poisson-Dirichlet process, provides more flexibility in the clustering behavior of Dirichlet process mixture models. In addition to the base measure (H) and precision (α), there is a discount parameter, $0 \leq d \leq 1$. The Pitman-Yor process allows negative values for α provided that $\alpha > -d$.

The Pitman-Yor process can be illustrated using a more general version of the Chinese restaurant process. Consider a hierarchical model with $\theta_1, \theta_2, \ldots$ being a sequence of i.i.d. random variables with random distribution P, where P has a Pitman-Yor process prior. Similar to the Chinese

restaurant process, when a customer arrives he either joins an existing table or begins a new table. Let K be the current number of occupied tables, z_i the dish for the ith customer, and $\mathbf{z}_{\widetilde{n}}$ the vector of dishes except for z_n:

$$\mathbb{P}(z_n = k|\alpha, d, \mathbf{z}_{\widetilde{n}}) = \begin{cases} \frac{\sum_{i<n} \mathbf{1}(z_i=k)-d}{n-1+\alpha}, & k \leq K \\ \frac{\alpha-dK}{n-1+\alpha}, & k = K+1 \end{cases} . \tag{5.20}$$

Notice that the discount parameter, d, reduces the clustering effect. The number of previous customers is reduced by d, and this weight is instead placed on the probability of a new table. For the limiting case with $d = 1$, $\theta_1, \ldots, \theta_n$ is an i.i.d. sample from H. On the other extreme, if $d = 0$, then the result is a Dirichlet process. As Teh (2006) shows, the number of unique values increases stochastically with both d and α. Recall that the stick-breaking process for the Dirichlet process constructs class proportions by setting $\pi_k = \phi_k \prod_{r=1}^{k-1}(1-\phi_r)$, where $\phi_k \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha)$. The Pitman-Yor process has a similar constructive definition using different beta marginals: $\phi_k \overset{\text{i.i.d.}}{\sim} \text{Beta}(1-d, \alpha+id)$ instead. It produces heavier tails than the Dirichlet process: π_k decreases stochastically with k for both processes, but this effect is more extreme with the Dirichlet process. Pitman-Yor processes have been used in applications such as natural language processing (Teh, 2006; Goldwater et al., 2006; Wallach et al., 2008) and image processing (Sudderth and Jordan, 2008).

Due to the stick-breaking construction, strategies for using Dirichlet processes can be adapted to Pitman-Yor processes. For example, it is straightforward to specify a hierarchical Pitman-Yor process by analogy to the hierarchical Dirichlet process. Teh (2006) constructs a MCMC sampling scheme for the hierarchical Pitman-Yor process, while Sudderth and Jordan (2008) develops a variational inference algorithm. The variational function updates for the Pitman-Yor process are similar to the Dirichlet process updates, since the stick-breaking proportions still have beta distributions. An open problem is to develop a more efficient collapsed strategy that integrates over the class proportions.

5.10 Conclusion

Nonparametric mixtures have been an active area of research since Sethuraman (1994) provided the seminal stick-breaking representation of the Dirichlet process. The Dirichlet process mixture model and its extensions have been used in many domains for modeling a population with an unbounded number of classes. The hierarchical Dirichlet process applies the same strategy for mixed membership models. Individual-level Dirichlet processes provide nonparametric mixtures for each individual, while a population-level Dirichlet process enables individuals to share statistical information. Such models have been used for survey analysis, document modeling, music models, and image analysis.

References

Ahmed, A. and Xing, E. P. (2008). Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: With applications to evolutionary clustering. In Wang, W. (ed), *Proceedings of the 2008 SIAM Conference on Data Mining (SDM '08)*. Philadelphia, PA, USA: SIAM, 219–230.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed-memberhsip stochastic blockmodels. *Journal of Machine Learning Research* 9: 1823–1856.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F. (eds), *Proceedings of the 2nd International Symposium on Information Theory*. Budapest, Hungary: Akadèmiai Kiadò, 267–281.

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

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

Beal, M. J. (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. E. (2001). The infinite hidden Markov model. In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds), *Advances in Neural Information Processings Systems 14*. Cambridge, MA: The MIT Press, 577–584.

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

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

Blunsom, P., Cohn, T., Goldwater, S., and Johnson, M. (2009). A note on the implementation of hierarchical Dirichlet processes. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*. Singapore: Association for Computational Linguistics, 337–340.

Canini, K. R., Shi, L., and Griffiths, T. L. (2009). Online inference of topics with latent Dirichlet allocation. In van Dyk, D. and Welling, M. (eds), *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*. *Journal of Machine Learning Research – Proceedings Track* 5 : 65–72.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika* 94: 809–825.

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. D. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* 101: 5220–5227.

Erosheva, E. A. (2003). Partial membership models with application to disability survey data. In Bozdogan, H. (ed), *Statistical Data Mining and Knowledge Discovery*. Chapman & Hall/CRC, 117–134.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics* 1: 502–537.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90: 577–588.

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

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet processes mixing. *Journal of the American Statistical Association* 100: 1021–1035.

Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* 86: 785–801.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In Weiss, Y., Schölkopf, B., and Platt, J. (eds), *Advances in Neural Information Processing Systems 18*. Cambridge, MA: The MIT Press, 459–466.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711–732.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer, 2nd edition.

Hoffman, M. D., Blei, D. M., and Bach, F. (2010). Online learning for latent Dirichlet allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds), *Advances in Neural Information Processing Systems 23*. Red Hook, NY: Curran Associates, Inc., 856–864.

Hoffman, M. D., Blei, D. M., and Cook, P. (2008). Content-based musical similarity computation using the hierarchical Dirichlet process. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*. ISMIR, 349–354.

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

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

Jain, S. and Neal, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis* 2: 445–472.

Kim, Y. (2003). On the posterior consistency of mixtures of Dirichlet process priors with censored data. *Scandinavian Journal of Statistics* 30: 535–547.

Kurihara, K., Welling, M., and Teh, Y. W. (2007). Collapsed variational Dirichlet process mixture models. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07)*. IJCAI, 2796–2801.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation* 23: 727–741.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* 7: 223–238.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York, NY: Wiley.

Paisley, J., Wang, C., and Blei, D. M. (2012). The discrete infinite logistic normal distribution. *Bayesian Analysis* 7 : 997–1034.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Statistics* 25: 855–900.

Rasmussen, C. E. (1999). The infinite Gaussian mixture model. In Solla, S. A., Leen, T. K., and Müller, K.-R. (eds), *Advances in Neural Information Processings Systems 12*. Cambridge, MA: The MIT Press, 554–560.

Rodriguez, A. (2011). On-line learning for the infinite hidden Markov model. *Communications in Statistics - Simulation and Computation* 40: 879–893.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* 92: 894–902.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* 4: 639–650.

Sudderth, E. B. and Jordan, M. I. (2008). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds), *Advances in Neural Information Processing Systems 21*. Red Hook, NY: Curran Associates, Inc., 1585–1592.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 985–992.

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

Teh, Y. W., Kurihara, K., and Welling, M. (2008). Collapsed variational inference for HDP. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds), *Advances in Neural Information Processing Systems 20*. Red Hook, NY: Curran Associates, Inc., 1481–1488.

Wallach, H. M., Sutton, C., and McCallum, A. (2008). Bayesian modeling of dependency trees using hierarchical Pitman-Yor priors. In *Proceedings of the Workshop on Prior Knowledge for Text and Language (held in conjunction with ICML/UAI/COLT)*. Helsinki, Finland, 15–20.

Wang, C., Paisley, J., and Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*. Palo Alto, CA, USA: AAAI, 752–760.

Xu, T., Zhang, Z. M., Yu, P. S., and Long, B. (2008). Evolutionary clustering by hierarchical Dirichlet process with hidden Markov state. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08)*. Los Alamitos, CA, USA: IEEE Computer Society, 658–667.