Variational inference for Dirichlet process mixtures

David M. Blei
School of Computer Science
Carnegie-Mellon University
blei@cs.cmu.edu

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California, Berkeley
jordan@stat.berkeley.edu

October 5, 2004

Abstract

Dirichlet process (DP) mixture models are the cornerstone of nonparametric Bayesian statistics, and the development of Monte-Carlo Markov chain (MCMC) sampling methods for DP mixtures has enabled their applications to a variety of practical data analysis problems. However, MCMC sampling can be prohibitively slow, and it is important to explore alternatives. One class of alternatives is provided by variational methods, a class of deterministic algorithms that convert inference problems into optimization problems (Opper and Saad, 2001; Wainwright and Jordan, 2003). Thus far, variational methods have mainly been explored in the parametric setting, in particular within the formalism of the exponential family (Attias, 2000; Ghahramani and Beal, 2001; Blei et al., 2003). In this paper, we present a variational inference algorithm for DP mixtures. We present experiments that compare the algorithm to Gibbs sampling algorithms for DP mixtures of Gaussians and present an application to a large-scale image analysis problem.

1 Introduction

The methodology of Monte Carlo Markov chain (MCMC) sampling has energized Bayesian statistics during the past decade, providing a systematic approach to the computation of likelihoods and posterior distributions, and permitting the deployment of Bayesian methods in a rapidly growing number of applied problems. However, while an unquestioned success story, MCMC is not an unqualified success story—MCMC methods can be slow to converge and their convergence can be difficult to diagnose. While further research on sampling is needed, it is also important to explore alternatives, particularly in the context of large-scale problems.

One such class of alternatives is provided by *variational inference methods* (Ghahramani and Beal, 2001; Jordan et al., 1999; Opper and Saad, 2001; Wainwright and Jordan, 2003; Wiegerinck, 2000). Like MCMC, variational inference methods have their roots in statistical physics, and, in contradistinction to MCMC methods, they are deterministic. The basic idea of variational inference is to formulate the computation of a marginal or conditional probability in terms of an optimization problem. This (generally intractable) problem is then "relaxed," yielding a simplified optimization problem that depends on a number of free parameters, known as variational parameters. Solving for the variational parameters gives an approximation to the marginal or conditional probabilities of interest.

Variational inference methods have been developed principally in the context of the exponential family, where the convexity properties of the natural parameter space and the cumulant generating function yield an elegant general variational formalism (Wainwright and Jordan, 2003). For example, variational methods have been developed for parametric hierarchical Bayesian models based on general exponential family specifications (Ghahramani and Beal, 2001). MCMC methods have seen much wider application. In particular, the development of MCMC algorithms for nonparametric models such as the Dirichlet process has led to increased interest in nonparametric Bayesian methods. In the current paper, we aim to close this gap and indicate how variational methods can be used in the Dirichlet process setting.

The Dirichlet process (DP), introduced in Ferguson (1973), is parameterized by a base measure G_0 and positive scaling parameter α. Writing $G \,|\, \{G_0, \alpha\} \sim \mathrm{DP}(G_0, \alpha)$ for a draw from the Dirichlet process, suppose that $\{\eta_1, \ldots, \eta_N\}$ are subsequently drawn independently from G: $\eta_n \,|\, G \sim G$. Marginalizing out the random measure G, the joint distribution of $\{\eta_1, \ldots, \eta_N\}$ turns out to follow a Pólya urn scheme (Blackwell and MacQueen, 1973). Thus, positive probability is assigned to configurations in which different η_n take on identical values, and the underlying random measure G is discrete with probability one. This is seen most directly in the stick-breaking representation of the DP, in which G is represented explicitly as an infinite sum of atomic measures (Sethuraman, 1994).

The Dirichlet process mixture model (Antoniak, 1974) adds a level to the hierarchy, treating η_n as the parameter of the distribution of the nth observation. Given the discreteness of G, the DP mixture has an interpretation as a mixture model with an unbounded number of mixture components.

Our goal will be to compute the predictive density:

$$p(x \mid x_1, \ldots, x_N) = \int p(x \mid \eta) p(\eta \mid x_1, \ldots, x_N) d\eta, \tag{1}$$

under the DP mixture, given a sample $\{x_1, \ldots, x_N\}$. As in many hierarchical Bayesian models, the posterior distribution $p(\eta \mid x_1, \ldots, x_N)$ is complicated and difficult to characterize in a closed form in the DP mixture setting. MCMC provides one class of approximations for this posterior and the predictive density (Escobar and West, 1995; Neal, 2000).

In this paper, we present a variational inference algorithm for DP mixtures based on the stick-breaking representation of the underlying DP. The algorithm involves two probability distributions—the posterior distribution p and a variational distribution q. The latter is endowed with free variational parameters, and the algorithmic problem is to adjust these parameters so that q approximates p. We also use a stick-breaking representation for q, but in this case we truncate the representation to yield a finite-dimensional representation. While in principle we could also truncate p, turning the model into a finite-dimensional model, it is important to emphasize at the outset that this is not our approach—we only truncate the variational distribution and approximate the posterior of an infinite-dimensional model.

The paper is organized as follows. In Section 2 we provide basic background on DP mixture models, focusing on the case of exponential family mixtures. Section 3 overviews MCMC algorithms for the DP mixture, discussing algorithms based both on the Pólya urn representation and the stick-breaking representation. In Section 4 we present a variational inference algorithm for DP mixtures. Section 5 presents the results of experimental comparisons and Section 7 presents our conclusions.

2 Dirichlet process mixture models

Let η be a continuous random variable, let G_0 be a non-atomic probability distribution for η, and let α be a positive, real-valued scalar. A random measure G is distributed according to a *Dirichlet process* (DP) (Ferguson, 1973), with scaling parameter α and base measure G_0, if for all natural numbers k and k-partitions

$\{B_1, \ldots, B_k\}$:

$$(G(\eta \in B_1), G(\eta \in B_2), \ldots, G(\eta \in B_k)) \sim \mathrm{Dir}(\alpha G_0(B_1), \alpha G_0(B_2), \ldots, \alpha G_0(B_k)).$$

Integrating out G, the joint distribution of the collection of variables $\{\eta_1, \ldots, \eta_n\}$ exhibits a clustering effect; conditioned on $n-1$ draws, the nth value is, with positive probability, exactly equal to one of those draws:

$$p(\eta \,|\, \eta_1, \ldots, \eta_{n-1}) \propto \alpha p(\eta \,|\, G_0) + \sum_{i=1}^{n-1} \delta_{\eta_i}(\eta). \tag{2}$$

Thus, $\{\eta_1, \ldots, \eta_{n-1}\}$ are randomly partitioned according to which variables are equal to the same value, with the distribution of the partition obtained from a Pólya urn scheme (Blackwell and MacQueen, 1973). Let $\{\eta_1^*, \ldots, \eta_{|\mathbf{c}|}^*\}$ denote the distinct values of $\{\eta_1, \ldots, \eta_{n-1}\}$, let $\mathbf{c} = \{c_1, \ldots, c_{n-1}\}$ denote the partition such that $\eta_i = \eta_{c_n}^*$, and let $|\mathbf{c}|$ denote the number of groups in that partition. The distribution of η_n follows the urn distribution:

$$\eta_n = \begin{cases} \eta_i^* & \text{with prob } \frac{|\mathbf{c}|_i}{n-1+\alpha} \\ \eta, \eta \sim G_0 & \text{with prob } \frac{\alpha}{n-1+\alpha}, \end{cases} \tag{3}$$

where $|\mathbf{c}|_i$ is the number of times the value η_i^* occurs in $\{\eta_1, \ldots, \eta_{n-1}\}$.

In the *Dirichlet process mixture model*, the DP is used as a nonparametric prior in a hierarchical Bayesian model (Antoniak, 1974):

$$\begin{aligned} G \,|\, \{\alpha, G_0\} &\sim \mathrm{DP}(\alpha, G_0) \\ \eta_n \,|\, G &\sim G \\ X_n \,|\, \eta_n &\sim p(x_n \,|\, \eta_n). \end{aligned}$$

Data generated from this model can be partitioned according to those values drawn from the same parameter. Thus, the DP mixture has a natural interpretation as a flexible mixture model in which the number of components (i.e., the number of groups in the partition) is random and grows as new data are observed.

The urn scheme in Equation (3) provides an implicit definition of the DP. Sethuraman (1994) provides an explicit definition via a *stick-breaking construction* of G. Consider two infinite collections of independent random variables, $V_i \sim \mathrm{Beta}(1, \alpha)$

and $\eta_i^* \sim G_0$, for $i = \{1, 2, \ldots\}$. We can write G as:

$$\theta_i = V_i \prod_{j=1}^{i-1}(1 - V_j) \tag{4}$$

$$G(\eta) = \sum_{i=1}^{\infty} \theta_i \delta_{\eta_i^*}(\eta). \tag{5}$$

Thus the support of G consists of a countably infinite set of atoms, drawn iid from G_0. The mixing proportions θ_i are given by successively breaking a unit length "stick" into an infinite number of pieces. The size of each successive piece, proportional to the rest of the stick, is given by an independent draw from a $\text{Beta}(1, \alpha)$ distribution.

In the DP mixture, the vector θ comprises the infinite vector of mixing proportions and $\{\eta_1^*, \eta_2^*, \ldots\}$ are the infinite number of mixture components. Let Z_n denote the mixture component with which x_n is associated.[1] The data can thus be described as arising from the following process:

1. Draw $V_i \,|\, \alpha \sim \text{Beta}(1, \alpha)$, $i = \{1, 2, \ldots\}$
2. Draw $\eta_i^* \,|\, G_0 \sim G_0$, $i = \{1, 2, \ldots\}$
3. For each data point n:
   (a) Draw $Z_n \,|\, \{v_1, v_2, \ldots\} \sim \text{Mult}(\theta)$.
   (b) Draw $X_n \,|\, z_n \sim p(x_n \,|\, \eta_{z_n}^*)$.

2.1 Exponential family mixtures

In this paper, we restrict ourselves to DP mixtures for which the observable data are drawn from an exponential family distribution, and where the base measure for the DP is the corresponding conjugate prior.

A DP mixture using the stick-breaking construction is illustrated as a graphical model in Figure 1. The distributions of V_k and Z_n are as described above. The distribution of X_n conditional on Z_n and $\{\eta_1^*, \eta_2^*, \ldots\}$ is:

$$p(x_n \,|\, z_n, \eta_1^*, \eta_2^*, \ldots) = \prod_{i=1}^{\infty} \left(h(x_n) \exp\{\eta_i^{*T} x_n - a(\eta_i^*)\}\right)^{z_n^i},$$

[1] We represent multinomial random vectors as indicator vectors consisting of a single component equal to one and the remaining components equal to zero.

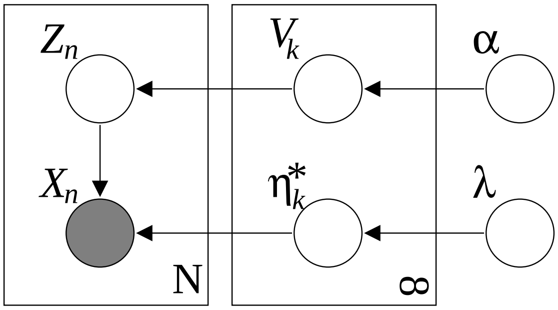

Figure 1: Graphical model representation of an exponential family DP mixture. Nodes denote random variables, edges denote possible dependence, and plates denote replication.

where $a(\eta_i^*)$ is the appropriate cumulant generating function and we assume for simplicity that x is the sufficient statistic for the canonical parameter η.

The vector of sufficient statistics of the corresponding conjugate family is $(\eta^{*T}, -a(\eta^*))^T$. The base measure is thus:

$$p(\eta^* \mid \lambda) = h(\eta^*) \exp\{\lambda_1^T \eta^* + \lambda_2(-a(\eta^*)) - a(\lambda)\},$$

where we decompose the hyperparameter λ such that λ_1 contains the first $\dim(\eta^*)$ components and λ_2 is a scalar.

2.2 The truncated Dirichlet process

Ishwaran and James (2001) have discussed the *truncated Dirichlet process* (TDP), in which V_{K-1} is set equal to one for some fixed value K. This yields $\theta_i = 0$ for $i \geq K$, and thus converts the infinite sum in Equation (4) into a finite sum. Ishwaran and James (2001) show that a TDP closely approximates a true Dirichlet process when the truncation level K is chosen large enough relative to the number of data points. Thus, they can justify substituting a TDP mixture model for a full DP mixture model.

3 MCMC for DP mixtures

The posterior distribution under both the DP and TDP mixture models cannot be computed efficiently in any direct way. It must be approximated, and Markov chain

Monte Carlo (MCMC) methods are the method of choice for approximating these posteriors (Escobar and West, 1995; Neal, 2000; Ishwaran and James, 2001).

As in the parametric setting, the idea behind MCMC for approximate posterior inference in the DP mixture is to construct a Markov chain for which the stationary distribution is the posterior of interest. One collects samples from the sample path of this Markov chain to construct an estimate of the posterior. Such an estimate can then be used to compute an approximation of the predictive distribution (see Equation 1).

The simplest MCMC algorithm is the Gibbs sampler, in which the Markov chain is defined by iteratively sampling each latent variable conditional on the data and the most recently sampled values of the other latent variables. This yields a chain with the desired stationary distribution (Geman and Geman, 1984; Gelfand and Smith, 1990; Neal, 1993). Below, we review the Gibbs sampling algorithms for DP and TDP mixtures.

3.1 Collapsed Gibbs sampling

In the *collapsed Gibbs sampler* for a DP mixture with conjugate base measure (Escobar and West, 1995), we integrate out the random measure G and distinct parameter values $\{\eta_1^*, \ldots, \eta_{|\mathbf{c}|}^*\}$. The Markov chain is thus defined only on the latent partition of the data, $\mathbf{c} = \{c_1, \ldots, c_N\}$.

Denote the data by $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$. For $n \in \{1, \ldots, N\}$, the algorithm iteratively samples each group assignment C_n, conditional on the partition of the rest of the data $\mathbf{c}_{-n}$. Note that C_n can be assigned to one of $|\mathbf{c}_{-n}| + 1$ values: either the nth data point is in a group with other data points, or in a group by itself.

By exchangeability, C_n is drawn from the following multinomial distribution:

$$p(c_n^k = 1 \,|\, \mathbf{x}, \mathbf{c}_{-n}, \lambda, \alpha) \propto p(x_n \,|\, \mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n^k = 1, \lambda) p(c_n^k = 1 \,|\, \mathbf{c}_{-n}, \alpha). \tag{6}$$

The first term is a ratio of normalizing constants of the posterior distribution of the kth parameter, one including and one excluding the nth data point:

$$p(x_n \,|\, \mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n^k = 1, \lambda) = \frac{\exp\left\{a(\lambda_1 + \sum_{m \neq n} c_m^k X_m + X_n, \lambda_2 + \sum_{m \neq n} c_m^k + 1)\right\}}{\exp\left\{a(\lambda_1 + \sum_m c_m^k X_m, \lambda_2 + \sum_{m \neq n} c_m^k)\right\}}. \tag{7}$$

The second term is given by the Pólya urn scheme:

$$p(c_n^k = 1 \,|\, \mathbf{c}_{-n}) \propto \begin{cases} |\mathbf{c}_{-n}|_k & \text{if } k \text{ is an existing group in the partition} \\ \alpha & \text{if } k \text{ is a new group in the partition,} \end{cases} \tag{8}$$

where $|\mathbf{c}_{-n}|_k$ denotes the number of data in the kth group of the partition.

Once this chain has reached its stationary distribution, we collect B samples $\{\mathbf{c}_1, \ldots, \mathbf{c}_B\}$ to approximate the posterior. The approximate predictive distribution is an average of the predictive distributions for each of the collected samples:

$$p(x_{N+1} \,|\, x_1, \ldots, x_N, \alpha, \lambda) = \frac{1}{B} \sum_{b=1}^{B} p(x_{N+1} \,|\, \mathbf{c}_b, \mathbf{x}, \alpha, \lambda).$$

For a particular sample, that distribution is:

$$p(x_{N+1} \,|\, \mathbf{c}, \mathbf{x}, \alpha, \lambda) = \sum_{k=1}^{|\mathbf{c}|+1} p(c^k_{N+1} = 1 \,|\, \mathbf{c}) p(x \,|\, \mathbf{c}, \mathbf{x}, c^k_{N+1} = 1).$$

When G_0 is not conjugate, the integral in Equation (7) does not have a simple closed form. Effective algorithms for handling this case are given in Neal (2000).

3.2 Blocked Gibbs sampling

In the collapsed Gibbs sampler, the distribution of each partition group variable C_n depends on the most recently sampled values of the other variables. Thus, these variables must be updated one at a time, which could potentially slow down the algorithm when compared to a blocking strategy. To this end, Ishwaran and James (2001) developed an inference algorithm based on the TDP described in Section 2. By explicitly sampling an approximation of G, this model allows for a blocked Gibbs sampler, in which collections of variables can be updated simultaneously.

The state of the Markov chain consists of the beta variables $\mathbf{V} = \{V_1, \ldots, V_{K-1}\}$, the component parameters $\boldsymbol{\eta}^* = \{\eta^*_1, \ldots, \eta^*_K\}$, and the component assignment variables $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$. The blocked Gibbs sampler iterates between the following three steps:

1. For $n \in \{1, \ldots, N\}$, independently sample Z_n from:

   $$p(z^k_n = 1 \,|\, \mathbf{v}, \boldsymbol{\eta}^*, \mathbf{x}) = \theta_k p(x_n \,|\, \eta^*_k),$$

   where θ_k is the function of $\mathbf{v}$ given in Equation (4).

2. For $k \in \{1, \ldots, K\}$, independently sample V_k from $\text{Beta}(\gamma_{k,1}, \gamma_{k,2})$, where:

   $$\begin{aligned} \gamma_{k,1} &= 1 + \textstyle\sum_{n=1}^{N} z^k_n \\ \gamma_{k,2} &= \alpha + \textstyle\sum_{i=k+1}^{K} \sum_{n=1}^{N} z^i_n. \end{aligned}$$

This step follows from the conjugacy between the multinomial data $\mathbf{z}$ and the truncated stick-breaking construction, which is a generalized Dirichlet distribution (Connor and Mosimann, 1969).

3. For $k \in \{1, \ldots, K\}$, independently sample η_k^* from $p(\eta_k^* \mid \tau_k)$. This distribution is in the same family as the base measure, with parameters:

$$\begin{array}{rcl} \tau_{k,1} & = & \lambda_1 + \sum_{i \neq n} z_i^k x_i \\ \tau_{k,2} & = & \lambda_2 + \sum_{i \neq n} z_i^k. \end{array} \tag{9}$$

After the chain has reached its stationary distribution, we collect B samples and construct an approximate predictive distribution. Again, this distribution is an average of the predictive distributions for each of the collected samples. The predictive distribution for a particular sample is:

$$p(x_{N+1} \mid \mathbf{z}, \mathbf{x}, \alpha, \lambda) = \sum_{k=1}^{K} \mathrm{E}\left[\theta_i \mid \gamma_1, \ldots, \gamma_k\right] p(x_{N+1} \mid \tau_k), \tag{10}$$

where $\mathrm{E}\left[\theta_i \mid \gamma_1, \ldots, \gamma_k\right]$ is the expectation of the product of independent beta variables given in Equation (4). This distribution only depends on $\mathbf{z}$; the other variables are needed in the Gibbs sampling procedure, but can be integrated out here.

The TDP sampler readily handles non-conjugacy of G_0, provided that there is a method of sampling η_i^* from its posterior.

3.3 Placing a prior on the scaling parameter

A common extension to the DP mixture model involves placing a prior on the scaling parameter α, which determines how quickly the number of components grows with the data. For the urn-based samplers, Escobar and West (1995) place a Gamma(s_1, s_2) prior on α and implement the corresponding Gibbs updates with auxiliary variable methods.

In the TDP mixture, the gamma distribution is computationally convenient because it is conjugate to Beta$(1, \alpha)$ (see Appendix A). The Gibbs updates for α are thus:

$$\alpha \mid \{\mathbf{v}, s_1, s_2\} \sim \text{Gamma}\left(s_1 + K - 1, s_2 - \sum_{i=1}^{K-1} \log(1 - v_i)\right). \tag{11}$$

4 Variational inference for the DP mixture

Variational inference provides an alternative, deterministic methodology for approximating likelihoods and posteriors in an intractable probabilistic model (Wainwright and Jordan, 2003). We first review the basic idea in the context of the exponential family of distributions, and then turn to its application to DP mixture models.

Consider the exponential family indexed by the natural parameter θ:

$$p(z \,|\, \theta) = \exp\{\theta^T t(z) - a(\theta)\} h(z),$$

where $t(z)$ is the vector of sufficient statistics. The *cumulant generating function* $a(\theta)$, also known as the *log partition function*, is defined as follows:

$$a(\theta) = \log \int \exp\{\theta^T t(z)\} h(z) dz.$$

As discussed by Wainwright and Jordan (2003), this quantity can also be expressed variationally as:

$$a(\theta) = \sup_{\mu \in \mathcal{M}} \{\theta^T \mu - a^*(\mu)\}, \tag{12}$$

where $a^*(\mu)$ is the Fenchel-Legendre conjugate of $a(\theta)$ (Rockafellar, 1970), and $\mathcal{M}$ is the set of *realizable expected sufficient statistics*: $\mathcal{M} = \{\mu : \mu = \int t(z)p(z)h(z)dz, \text{for some } p\}$. There is a one-to-one mapping between parameters θ and the interior of $\mathcal{M}$ (Brown, 1986). Accordingly, the interior of $\mathcal{M}$ is often referred to as the set of *mean parameters.*

Let $\theta(\mu)$ be a natural parameter corresponding to the mean parameter $\mu \in \mathcal{M}$; thus $\mathrm{E}_\theta\left[t(Z)\right] = \mu$. Let $q(z \,|\, \theta(\mu))$ denote the corresponding density. Given $\mu \in \mathcal{M}$, a short calculation shows that $a^*(\mu)$ is the negative entropy of q:

$$a^*(\mu) = \mathrm{E}_{\theta(\mu)}\left[\log q(Z \,|\, \theta(\mu))\right]. \tag{13}$$

Given its definition as a Fenchel conjugate, the negative entropy is convex.

In many models of interest, $a(\theta)$ is not feasible to compute because of the complexity of $\mathcal{M}$ or the lack of any explicit form for $a^*(\mu)$. However, we can bound $a(\theta)$ using Equation (12):

$$a(\theta) \geq \mu^T \theta - a^*(\mu), \tag{14}$$

for any mean parameter $\mu \in \mathcal{M}$. Moreover, the tightness of the bound is measured by a Kullback-Leibler divergence expressed in terms of a mixed parameterization:

$$\begin{aligned} D(q(z \,|\, \theta(\mu)) \;||\; p(z \,|\, \theta)) &= \mathrm{E}_{\theta(\mu)}\left[\log q(z \,|\, \theta(\mu)) - \log p(z \,|\, \theta)\right] \\ &= \theta(\mu)^T \mu - a(\theta(\mu)) - \theta^T \mu + a(\theta) \\ &= a(\theta) - \theta^T \mu + a^*(\theta(\mu)). \end{aligned} \tag{15}$$

Mean-field variational methods are a special class of variational methods that are based on maximizing the bound in Equation (14) with respect to a subset $\mathcal{M}_{\text{tract}}$ of the space $\mathcal{M}$ of realizable mean parameters. In particular, $\mathcal{M}_{\text{tract}}$ is chosen so that $a^*(\theta(\mu))$ can be evaluated tractably and so that the maximization over $\mathcal{M}_{\text{tract}}$ can be performed tractably. Equivalently, given the result in Equation (15), mean-field variational methods minimize the KL divergence $D(q(z \,|\, \theta(\mu)) \;||\; p(z \,|\, \theta))$ with respect to its first argument.

If the distribution of interest is a posterior, then $a(\theta)$ is the log likelihood. Consider in particular a latent variable probabilistic model with hyperparameters θ, observed variables $\mathbf{x} = \{x_1, \ldots, x_N\}$, and latent variables $\mathbf{z} = \{z_1, \ldots, z_M\}$. The posterior can be written as:

$$p(\mathbf{z} \,|\, \mathbf{x}, \theta) = \exp\{\log p(\mathbf{z}, \mathbf{x} \,|\, \theta) - \log p(\mathbf{x} \,|\, \theta)\}, \tag{16}$$

and the bound in Equation (14) applies directly. We have:

$$\log p(\mathbf{x} \,|\, \theta) \geq \mathrm{E}_q \left[\log p(\mathbf{x}, \mathbf{Z} \,|\, \theta)\right] - \mathrm{E}_q \left[\log q(\mathbf{Z})\right]. \tag{17}$$

This equation holds for any q via Jensen's inequality, but, as our analysis has shown, it is useful specifically for q of the form $q(z \,|\, \theta(\mu))$ for $\mu \in \mathcal{M}_{\text{tract}}$.

A straightforward way to construct tractable subfamilies of exponential family distributions is to consider factorized families, in which each factor is an exponential family distribution depending on a so-called *variational parameter.* In particular, let us consider distributions of the form $q(\mathbf{z} \,|\, \boldsymbol{\nu}) = \prod_{i=1}^{M} q(z_i \,|\, \nu_i)$, where $\boldsymbol{\nu} = \{\nu_1, \nu_2, ..., \nu_M\}$ are variational parameters. Using this class of distributions, we simplify the likelihood bound using the chain rule:

$$\log p(\mathbf{x} \,|\, \theta) \geq \log p(\mathbf{x} \,|\, \theta) + \sum_{m=1}^{M} \mathrm{E}_q \left[\log p(Z_m \,|\, \mathbf{x}, Z_1, \ldots, Z_{m-1}, \theta)\right] - \sum_{m=1}^{M} \mathrm{E}_q \left[\log q(Z_m \,|\, \nu_m)\right]. \tag{18}$$

To obtain the best approximation available within the factorized subfamily, we now wish to optimize this expression with respect to ν_i.

To optimize with respect to ν_i, reorder $\mathbf{z}$ such that z_i is last in the list. The portion of Equation (18) depending on ν_i is:

$$\ell_i = \mathrm{E}_q \left[\log p(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \theta)\right] - \mathrm{E}_q \left[\log q(z_i \,|\, \nu_i)\right]. \tag{19}$$

Given our assumption that the variational distribution $q(z_i \,|\, \nu_i)$ is in the exponential family, we have:

$$q(z_i \,|\, \nu_i) = h(z_i) \exp\{\nu_i^T z_i - a(\nu_i)\},$$

and Equation (19) simplifies as follows:

$$\begin{aligned}\ell_i &= \mathrm{E}_q\left[\log p(Z_i \,|\, \mathbf{Z}_{-i}, \mathbf{x}, \theta) - \log h(Z_i) - \nu_i^T Z_i + a(\nu_i)\right] \\ &= \mathrm{E}_q\left[\log p(Z_i \,|\, \mathbf{Z}_{-i}, \mathbf{x}, \theta)\right] - \mathrm{E}_q\left[\log h(Z_i)\right] - \nu_i^T a'(\nu_i) + a(\nu_i),\end{aligned}$$

because $\mathrm{E}_q\left[Z_i\right] = a'(\nu_i)$. The derivative with respect to ν_i is:

$$\frac{\partial}{\partial \nu_i}\ell_i = \frac{\partial}{\partial \nu_i}\left(\mathrm{E}_q\left[\log p(Z_i \,|\, \mathbf{Z}_{-i}, \mathbf{x}, \theta)\right] - \mathrm{E}_q\left[\log h(Z_i)\right]\right) - \nu_i^T a''(\nu_i). \tag{20}$$

Thus the optimal ν_i satisfies:

$$\nu_i = \left[a''(\nu_i)\right]^{-1}\left(\frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log p(Z_i \,|\, \mathbf{Z}_{-i}, \mathbf{x}, \theta)\right] - \frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log h(Z_i)\right]\right). \tag{21}$$

The result in Equation (21) is general. In many applications of mean field methods, a further simplification is achieved. In particular, when the conditional distribution $p(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \theta)$ is an exponential family distribution[2], we have:

$$p(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \theta) = h(z_i)\exp\{g_i(\mathbf{z}_{-i}, \mathbf{x}, \theta)^T z_i - a(g_i(\mathbf{z}_{-i}, \mathbf{x}, \theta))\},$$

where $g_i(\mathbf{z}_{-i}, \mathbf{x}, \theta)$ denotes the natural parameter for z_i when conditioning on the remaining latent variables and the observations. This yields simplified expressions for the expected log probability of Z_i and its first derivative:

$$\begin{aligned}\mathrm{E}_q\left[\log p(Z_i \,|\, \mathbf{Z}_{-i}, \mathbf{x}, \theta)\right] &= \mathrm{E}\left[\log h(Z_i)\right] + \mathrm{E}_q\left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \theta)\right]^T a'(\nu_i) - \mathrm{E}_q\left[a(g_i(\mathbf{Z}_{-i}, \mathbf{x}, \theta))\right] \\ \frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log p(Z_i \,|\, \mathbf{Z}_{-i}, \mathbf{x}, \theta)\right] &= \frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log h(Z_i)\right] + \mathrm{E}_q\left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \theta)\right]^T a''(\nu_i).\end{aligned}$$

Using the first derivative in Equation (21), the maximum is attained at:

$$\nu_i = \mathrm{E}_q\left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \theta)\right]. \tag{22}$$

We define a coordinate ascent algorithm based on Equation (22) by iteratively updating ν_i for $i \in \{1, \ldots, N\}$. Such an algorithm finds a local maximum of Equation (17) by Proposition 2.7.1 of Bertsekas (1999), under the condition that the right-hand

[2] Examples in which $p(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \theta)$ is an exponential family distribution include Kalman filters, hidden Markov models, mixture models, hierarchical Bayesian models with conjugate and mixture of conjugate priors, and the hierarchical nonparametric Bayesian models which are the focus of this paper.

side of Equation (19) is strictly convex. Further perspectives on algorithms of this kind can be found in Xing et al. (2003) and Beal (2003).

Notice the interesting relationship of this algorithm to the Gibbs sampler. In Gibbs sampling, we iteratively draw the latent variables z_i from the distribution $p(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \theta)$. In mean-field variational inference, we iteratively update the variational parameters of z_i to be equal to the expected value of the parameter g_i of the conditional distribution $p(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \theta)$, where the expectation is taken under the variational distribution.[3]

4.1 Variational inference for DP mixtures

We develop a mean-field variational algorithm for the DP mixture based on the stick-breaking representation of the DP mixture in Figure 1. Using this representation, the bound on the likelihood given in Equation (17) becomes:

$$\begin{aligned}\log p(\mathbf{x} \,|\, \alpha, \lambda) \geq & \mathrm{E}_q\left[\log p(\mathbf{V} \,|\, \alpha)\right] + \mathrm{E}_q\left[\log p(\boldsymbol{\eta^*} \,|\, \lambda)\right] \\ & + \sum_{n=1}^{N}\left(\mathrm{E}_q\left[\log p(Z_n \,|\, \mathbf{V})\right] + \mathrm{E}_q\left[\log p(x_n \,|\, Z_n)\right]\right) \\ & - \mathrm{E}_q\left[\log q(\mathbf{Z}, \mathbf{V}, \boldsymbol{\eta^*})\right].\end{aligned} \tag{23}$$

The issue that we must face to make use of this bound is that of constructing an approximation to the distribution of the infinite-dimensional random measure G, expressed in terms of the infinite sets $\mathbf{V} = \{V_1, V_2, \ldots\}$ and $\boldsymbol{\eta^*} = \{\eta^*_1, \eta^*_2, \ldots\}$. Our approach is based on using truncated stick-breaking representations for the variational distributions. Thus, we fix a value T and let $q(v_T = 1) = 1$. As in the truncated Dirichlet process, under the truncated variational distribution, the mixture proportions θ_t are equal to zero for $t > T$ and we can thus ignore the parameters η^*_t for $t > T$.

The factorized distribution that we propose to use as a basis for mean-field variational inference is thus of the following form:

$$q(\mathbf{v}, \boldsymbol{\eta^*}, \mathbf{z}, T) = \prod_{t=1}^{T-1} q(v_i \,|\, \gamma_i) \prod_{t=1}^{T} q(\eta^*_t \,|\, \tau_t) \prod_{n=1}^{N} q(z_n \,|\, \phi_n), \tag{24}$$

where γ_n are the parameters for a beta distribution, τ_t are natural parameters for the distributions of η^*_t, and ϕ_n are parameters for a multinomial distribution.

[3] This relationship has inspired the software package VIBES (Bishop et al., 2003), which is a variational version of the popular BUGS package (Gilks et al., 1996).

Notice that in the model of Figure 1, the variables $\mathbf{V}$, $\boldsymbol{\eta}^*$, and $\mathbf{Z}$ are each identically distributed, whereas, under the variational distribution, there is a different parameter for each variable. For example, the choice of the mixture component z_n for the nth data point is governed by a multinomial distribution indexed by a variational parameter ϕ_n that depends on n. This reflects the conditional nature of variational inference.

We emphasize the difference between role that truncation plays in our variational method and the role that it plays in the blocked Gibbs sampler of Ishwaran and James (2001) (see Section 3.2). The blocked Gibbs sampler estimates the posterior of a truncated approximation to the DP. In contrast, we use a truncated stick-breaking distribution to approximate the true posterior of a full DP mixture model—the posterior itself is *not* truncated. The truncation level T is a variational parameter which can be freely set; it is not a part of the prior model specification.

4.2 Coordinate-ascent algorithm

We now derive a coordinate-ascent algorithm for optimizing the bound in Equation (23) with respect to the variational parameters. The third term in the bound is the only term that requires attention, as all of the other terms in the bound involve standard computations in the exponential family. We rewrite the third term using indicator random variables:

$$\begin{aligned}
\mathrm{E}_q\left[\log p(Z_n \mid \mathbf{V})\right] &= \mathrm{E}_q\left[\log\left(\textstyle\prod_{i=1}^T (1-V_i)^{\mathbf{1}[Z_n>i]} V_i^{Z_n^i}\right)\right] \\
&= \textstyle\sum_{i=1}^T q(z_n > i)\mathrm{E}\left[\log(1-V_i)\right] + q(z_n = i)\mathrm{E}\left[\log V_i\right],
\end{aligned}$$

where:

$$\begin{aligned}
q(z_n = i) &= \phi_{n,i} \\
q(z_n > i) &= \textstyle\sum_{j=i+1}^K \phi_{n,j} \\
\mathrm{E}\left[\log V_i\right] &= \Psi(\gamma_{i,1}) - \Psi(\gamma_{i,1} + \gamma_{i,2}) \\
\mathrm{E}\left[\log(1-V_i)\right] &= \Psi(\gamma_{i,2}) - \Psi(\gamma_{i,1} + \gamma_{i,2}).
\end{aligned}$$

(Note that Ψ is the digamma function arising from the derivative of the log normalization factor in the beta distribution.)

We now use the general expression in Equation (21) to derive a mean-field coordinate ascent algorithm. Computing the derivatives with respect to the variational parameters, the bound in Equation (23) is optimized via the following set of updates, for $t \in \{1, \ldots, T\}$ and $n \in \{1, \ldots, N\}$:

$$\begin{aligned} \gamma_{t,1} &= 1 + \textstyle\sum_n \phi_{n,t} && (25) \\ \gamma_{t,2} &= \alpha + \textstyle\sum_n \sum_{j=t+1}^{T} \phi_{n,j} && (26) \\ \tau_{t,1} &= \lambda_1 + \textstyle\sum_n \phi_{n,t} x_n && (27) \\ \tau_{t,2} &= \lambda_2 + \textstyle\sum_n \phi_{n,t}. && (28) \\ \phi_{n,t} &\propto \exp(S), && (29) \end{aligned}$$

where:

$$S = \mathrm{E}\left[\log V_t \mid \gamma_t\right] + \mathrm{E}\left[\eta_t \mid \tau_t\right]^T X_n - \mathrm{E}\left[a(\eta_t) \mid \tau_t\right] - \sum_{j=t+1}^{T} \mathrm{E}\left[\log(1 - V_j) \mid \gamma_j\right].$$

Iterating these updates optimizes Equation (23) with respect to the variational parameters defined in Equation (24). That is, we find $q(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z})$ which, when plugged in to the factored expression displayed in Equation (24), yield a distribution that is a mean-field approximation to the true posterior.

Practical applications of variational methods must address initialization of the variational distribution. While the algorithm yields a bound for any starting values of the variational parameters, poor choices of initialization can lead to local maxima that yield poor bounds. We initialize the variational distribution by incrementally updating the parameters according to a random permutation of the data points. In a sense, this is a variational version of sequential importance sampling. We run the algorithm multiple times and choose the final parameter settings that give the best bound on the marginal likelihood.

Given a (possibly locally) optimal set of variational parameters, the approximate predictive distribution is:

$$p(x_{N+1} \mid \mathbf{z}, \mathbf{x}, \alpha, \lambda) = \sum_{t=1}^{T} \mathrm{E}_q\left[\theta_t \mid \boldsymbol{\gamma}\right] \mathrm{E}_q\left[p(x_{N+1} \mid \tau_t)\right]. \tag{30}$$

This approximation has a form similar to the approximate predictive distribution under the blocked Gibbs sampler in Equation (10). In the variational case, however, the averaging is done parametrically via the variational distribution rather than by a Monte Carlo integral.

When G_0 is not conjugate, a simple coordinate ascent update for τ_i may not be available if $p(\eta_i^* \mid \mathbf{z}, \mathbf{x}, \lambda)$ is not in the exponential family. However, if G_0 is a mixture of conjugate priors, then a simple coordinate ascent algorithm is available.

Finally, we extend the variational inference algorithm to posterior updates for the scaling parameter α with a Gamma(s_1, s_2) prior. Using the exact posterior of α in Equation (11), the variational posterior Gamma(w_1, w_2) distribution is:

$$\begin{aligned} w_1 &= s_1 + T - 1 \\ w_2 &= s_2 - \sum_{i=1}^{T-1} \mathrm{E}_q\left[\log(1 - V_i)\right]), \end{aligned}$$

and we replace α with its expectation $\mathrm{E}_q\left[\alpha \,|\, w\right] = w_1/w_2$ in the updates for $\gamma_{t,2}$ in Equation (26).

4.3 Discussion

Qualitatively, variational methods offer several potential advantages over Gibbs sampling. They are deterministic, and have an optimization criterion given by Equation (23) that can be used to assess convergence. In contrast, assessing convergence of a Gibbs sampler—namely, determining when the Markov chain has reached its stationary distribution—is an active field of research. Theoretical bounds on the mixing time are of little practical use, and there is no consensus on how to choose among the several empirical methods developed for this purpose (Robert and Casella, 1999).

Furthermore, in this context, the variational technique provides an explicit estimate of the infinite-dimensional parameter G by using the truncated stick-breaking construction. The best Gibbs samplers (e.g., the collapsed Gibbs sampler) marginalize out G and rely on the Pólya urn scheme representation (Neal, 2000). This precludes computation of quantities, such as quantiles, which cannot be expressed as expectations of G. (See Gelfand and Kottas (2002) for a method which combines urn-based sampling and TDP-based blocked sampling to compute such quantities.)

But there are several potential disadvantages of variational methods as well. First, variational methods are deterministic optimization procedures that can fall prey to local minima. Local minima can be mitigated with restarts, or removed via the incorporation of additional variational parameters, but these strategies may slow the overall convergence of the procedure and nullify the advantage over MCMC. Second, any given fixed variational representation yields only an approximation to the posterior. There are methods for considering hierarchies of variational representations that approach the posterior in the limit, but these methods may again incur serious computational costs. Lacking a theory by which these issues can be evaluated in the general setting of DP mixtures, we turn to experimental evaluation.

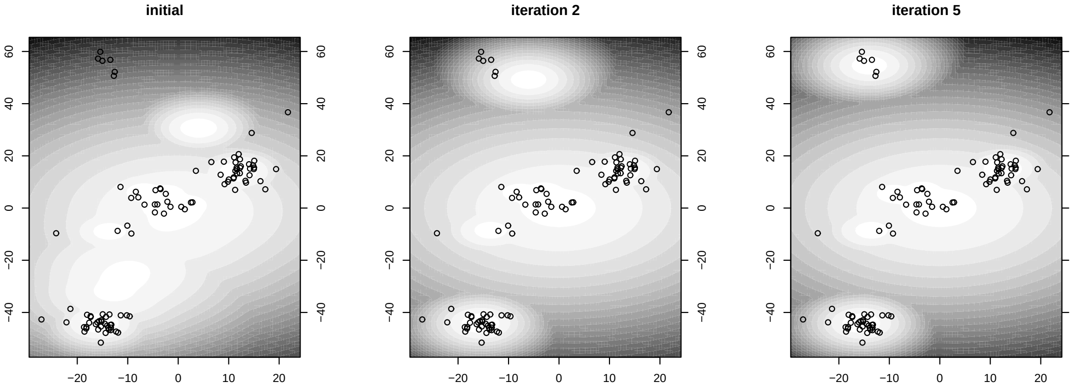

Figure 2: The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.

5 Empirical comparison

We studied the performance of the variational algorithm of Section 4 and the Gibbs samplers of Section 3 in the setting of Gaussian DP mixtures. Thus, likelihood is Gaussian with fixed covariance matrix Λ and the Dirichlet process mixes over the mean of the Gaussian. The base measure for the DP is Gaussian, with covariance given by Λ/λ_2, which is conjugate to the likelihood.

Figure 2 provides an illustrative example of the variational inference algorithm on a small problem involving 100 data points sampled from a two-dimensional Gaussian DP mixture with diagonal covariance. Each panel in the figure illustrates the data and the predictive distribution given by the variational inference algorithm, with truncation level 20. As seen in the first panel, the initialization of the variational parameters yields a largely flat distribution on the data. After one iteration, the algorithm has found the modes of the predictive distribution and, after convergence, it has further refined those modes. Even though 20 mixture components are represented in the variational distribution, the fitted approximate posterior only uses five of them.

To compare the variational inference algorithm to the Gibbs sampling algorithms, we conducted a systematic set of experiments in which the dimensionality of the data was varied from 5 to 50. In each case, we generated 100 data from a Gaussian DP mixture model of the chosen dimensionality and generated 100 additional points as held-out data. In testing on the held-out data, each point is treated as the 101st

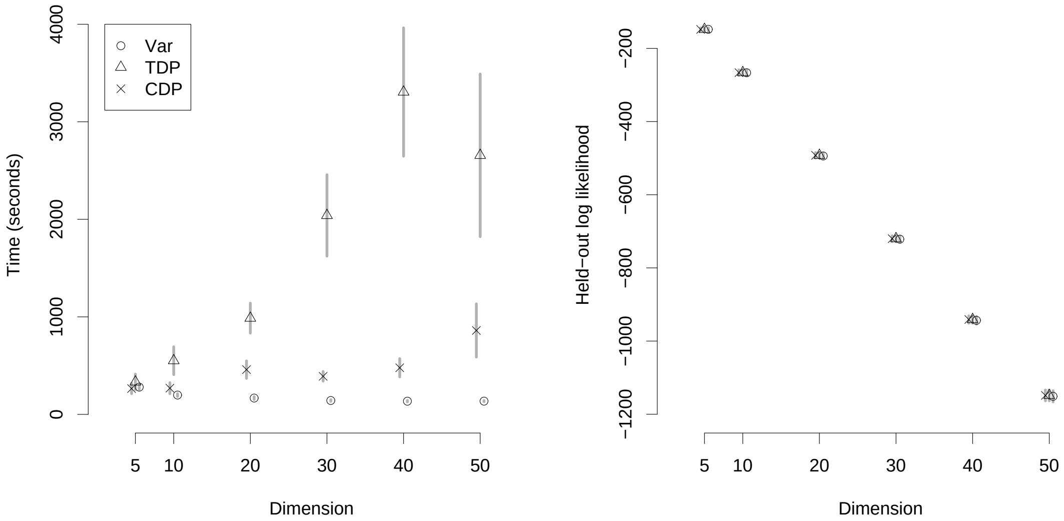

Figure 3: (Left) Convergence time across ten datasets per dimension for variational inference, TDP Gibbs sampling, and the collapsed Gibbs sampler (grey bars are standard error). (Right) Average held-out log likelihood for the corresponding predictive distributions.

data point in the collection.

The covariance matrix was given by the autocorrelation matrix for a first-order autogressive process, chosen so that the components are highly dependent ($\rho = 0.9$). The base measure was a zero-mean Gaussian with covariance appropriately scaled for comparison across dimensions. The scaling parameter α was set equal to one.

We ran all algorithms to convergence and measure the computation time.[4] Convergence was assessed in the following way. For the Gibbs samplers, we assess convergence to the stationary distribution with the diagnostic given by Raftery and Lewis (1992), and collect 25 additional samples to estimate the predictive distribution (the same diagnostic provides an appropriate lag at which to collect uncorrelated samples). The TDP approximation and variational posterior approximation are both truncated at 20 components. For the variational inference algorithm we measure convergence by the relative change in the likelihood bound, stopping the algorithm when it is less than $1e^{-10}$. Note that there is a certain inevitable arbitrariness in these choices; in general it is difficult to envisage measures of computation time that allow stochastic MCMC algorithms and deterministic variational algorithms to be

[4] All timing computations were made on a Pentium III 1GHZ desktop machine.

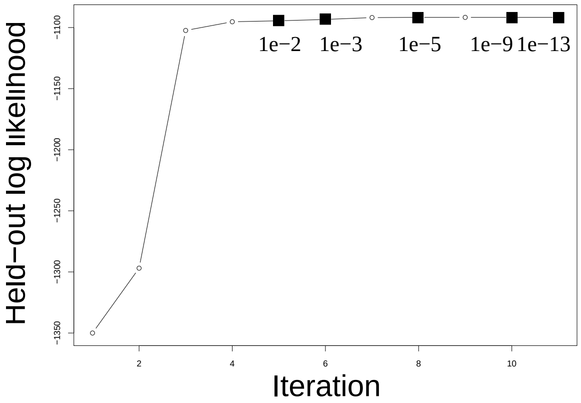

Figure 4: Held-out likelihood as a function of iteration of the variational inference algorithm for a 50-dimensional simulated dataset. The relative change in likelihood bound is labeled at selected iterations.

compared in a standardized way. Nonetheless, we have made what we consider to be reasonable, pragmatic choices. Note in particular that the choice of stopping time for the variational algorithm is quite conservative, as illustrated in Figure 4.

Figure 3 (left) illustrates the average convergence time across ten datasets per dimension, for dimensions ranging from 5 to 50. With the caveats in mind regarding convergence time measurement, it appears that the variational algorithm is quite competitive with the MCMC algorithms. The variational algorithm was faster and exhibited significantly less variance in its convergence time. Moreover, there is little evidence of an increase in convergence time across dimensionality for the variational algorithm.

Note that the collapsed Gibbs sampler converged faster than the TDP Gibbs sampler. Though an iteration of collapsed Gibbs is slower than an iteration of TDP Gibbs, the TDP Gibbs sampler required a longer burn-in and greater lag to obtain uncorrelated samples. This is illustrated in the example autocorrelation plots of Figure 5. Comparing the two MCMC algorithms, we find no advantage to the truncated approximation.

Figure 3 (right) illustrates the average log likelihood assigned to the held-out data by the approximate predictive distributions. First, notice that the collapsed DP Gibbs sampler assigned the same likelihood as the posterior from the TDP Gibbs sampler—an indication of the quality of a TDP for approximating a DP. More importantly, however, the predictive distribution based on the variational posterior assigned a similar score as those based on samples from the true posterior. Though

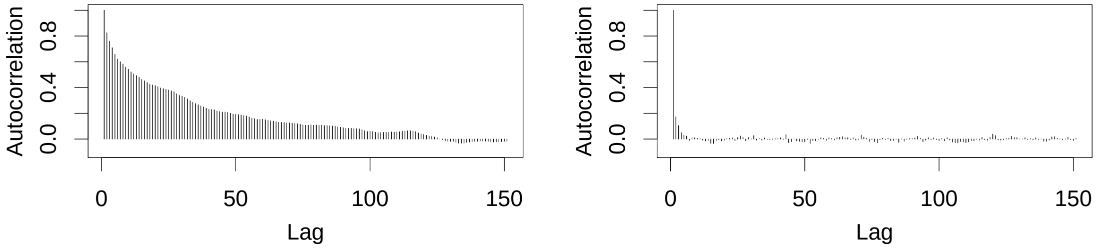

Figure 5: Autocorrelation plots on the size of the largest component for the truncated DP Gibbs sampler (left) and collapsed Gibbs sampler (right) in an example dataset of 50-dimensional Gaussian data.

it is based on an approximation to the true posterior, the resulting predictive distributions are very accurate for this class of DP mixtures.

6 Image analysis

Finite Gaussian mixture models are widely used in computer vision to model natural images for the purposes of automatic clustering, retrieval, and classification (Barnard et al., 2003; Jeon et al., 2003). These applications are often large-scale data analysis problems, involving thousands of data points (images) in hundreds of dimensions (pixels). The appropriate number of mixture components to use in these problems is generally unknown, and DP mixtures would seem to provide an attractive extension of current methods. This deployment of DP mixtures in such problems requires, however, inferential methods that are computationally efficient. To demonstrate the applicability of our variational approach to DP mixtures in the setting of large datasets, we analyzed a collection of 5000 images from the Associated Press under the assumptions of a Gaussian DP mixture model.

Each image is reduced to a 192-dimensional real-valued vector given by an 8×8 grid of average red, green, and blue values. The overall mean is subtracted to yield a dataset with mean zero. We fit a model which is a DP mixture in which the mixture components are Gaussian with mean μ and covariance matrix $\sigma^2 I$. The base measure G_0 is a product measure—Gamma(4,2) for σ^2 and $\mathcal{N}(0, 5\sigma^2)$ for μ. Furthermore, we place a Gamma(1,1) prior on the DP scaling parameter α, as described in Section 4.1. We use a truncation level of 150 for the variational distribution.

The variational algorithm required approximately 4 hours to converge. The resulting approximate posterior uses 79 mixture components to describe the collec-

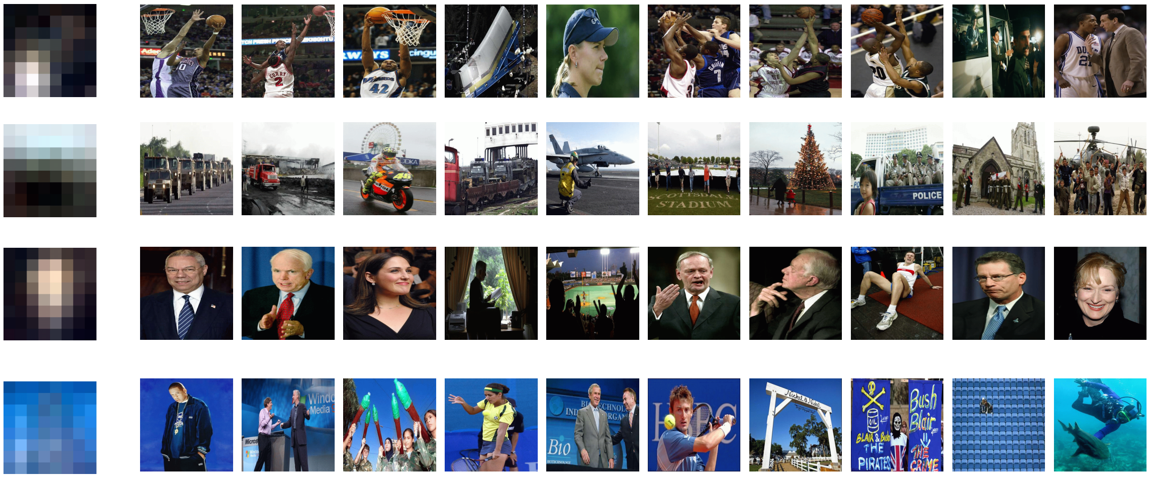

Figure 6: Four sample clusters from a DP mixture analysis of 5000 images from the Associated Press. The left-most column is the posterior mean of each cluster followed by the top ten images associated with it. These clusters capture patterns in the data, such as basketball shots, outdoor scenes on gray days, faces, and pictures with blue backgrounds.

tion. Figure 7 (Left) illustrates the expected number of images allocated to each component. Figure 6 illustrates the ten pictures with highest approximate posterior probability associated with each of four of the components. These clusters appear to capture pictures with basketball shots, outdoor scenes on gray days, faces, and pictures with blue backgrounds.

Figure 7 (Right) illustrates the prior for the scaling parameter α as well as the approximate posterior given by the fitted variational distribution. We see that the approximate posterior is peaked and rather different from the prior, indicating that the data have provided information regarding α. Moreover, the peak is centered around a large value of α, suggesting that the parametric model is inadequate in this case.

7 Conclusions

Bayesian nonparametric models based on the Dirichlet process are powerful tools for flexible data analysis. They offer the inferential strengths of the Bayesian approach together with a degree of robustness that is not always associated with the Bayesian

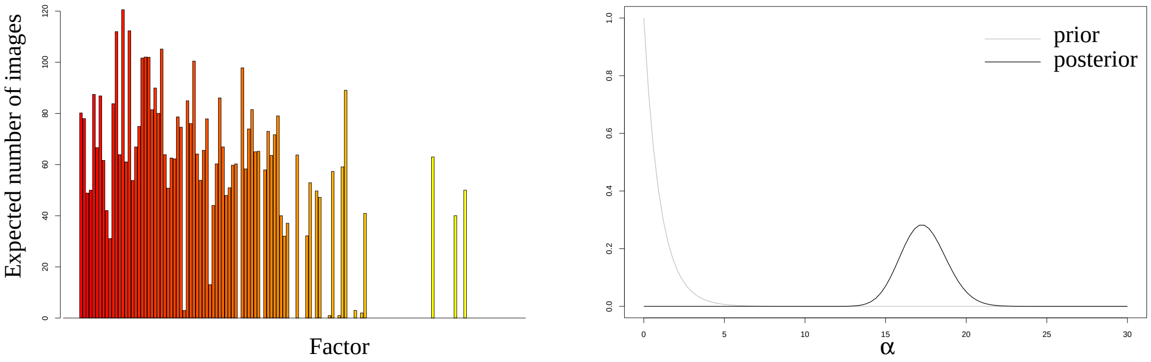

Figure 7: (Left) The expected number of images allocated to each component in the variational posterior. The posterior uses 79 components to describe the data. (Right) The prior for the scaling parameter α and the approximate posterior given by its variational distribution.

approach. For these benefits to be realized, however, the computational issues associated with Bayesian inference must remain a significant part of the research agenda. MCMC methods have become the leading paradigm for computational Bayesian inference, but long convergence times can hinder their usefulness, particularly in the context of large collections of multivariate and highly-correlated data. It would be wise to explore a variety of methods for fitting Bayesian nonparametric models.

We have developed a mean-field variational inference algorithm for the Dirichlet process mixture model and demonstrated its applicability to the kinds of multivariate data for which Gibbs sampling algorithms can exhibit slow convergence. Variational inference was faster than Gibbs sampling in our simulations, and its convergence time was independent of dimensionality for the range which we tested.

Both variational and MCMC methods have strengths and weaknesses, and it is unlikely that one methodology will dominate the other in general. While MCMC sampling provides theoretical guarantees of accuracy, variational inference provides a fast, deterministic approximation to otherwise unattainable posteriors. Moreover, both MCMC and variational inference are computational paradigms, providing a wide variety of specific algorithmic approaches which trade off speed, accuracy and ease of implementation in different ways. We have investigated the deployment of the simplest form of variational method for DP mixtures—a mean-field variational algorithm—but it worth noting that other variational approaches, such as those described in Wainwright and Jordan (2003), are also worthy of consideration in the nonparametric context.

References

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.

Attias, H. (2000). A variational Bayesian framework for graphical models. In Solla, S., Leen, T., and Muller, K., editors, *Advances in Neural Information Processing Systems 12*, pages 209–215, Cambridge, MA. MIT Press.

Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., and Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference.* PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

Bertsekas, D. (1999). *Nonlinear Programming.* Athena Scientific, Nashua, NH.

Bishop, C., Spiegelhalter, D., and Winn, J. (2003). VIBES: A variational inference engine for Bayesian networks. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 777–784. MIT Press, Cambridge, MA.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355.

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

Brown, L. (1986). *Fundamentals of Statistical Exponential Families.* Institute of Mathematical Statistics, Hayward, CA.

Connor, R. and Mosimann, J. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206.

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

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

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

Gelfand, A. and Smith, A. (1990). Sample based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

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

Ghahramani, Z. and Beal, M. (2001). Propagation algorithms for variational Bayesian learning. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 507–513, Cambridge, MA. MIT Press.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov chain Monte Carlo Methods in Practice*. Chapman and Hall.

Ishwaran, J. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–174.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Informaion Retrieval*, pages 119–126. ACM Press.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Opper, M. and Saad, D. (2001). *Advanced Mean Field Methods: Theory and Practice*. MIT Press, Cambridge, MA.

Raftery, A. and Lewis, S. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7:493–497.

Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer Texts in Statistics. Springer-Verlag, New York, NY.

Rockafellar (1970). *Convex Analysis.* Princeton University Press, Princeton, NJ.

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

Wainwright, M. and Jordan, M. (2003). Graphical models, exponential families, and variational inference. Technical Report 649, U.C. Berkeley, Dept. of Statistics.

Wiegerinck, W. (2000). Variational approximations between mean field theory and the junction tree algorithm. In Boutilier, C. and Goldszmidt, M., editors, *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI-00)*, pages 626–633, San Francisco, CA. Morgan Kaufmann Publishers.

Xing, E., Jordan, M., and Russell, S. (2003). A generalized mean field algorithm for variational inference in exponential families. In Meek, C. and Kjrulff, U., editors, *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 583–591, San Francisco, CA. Morgan Kaufmann Publishers.

A A conjugate prior for the scaling parameter

In this appendix we show that a gamma prior for the scaling parameter α is conjugate to the stick lengths in the stick-breaking representation of the Dirichlet process.

Recall that the V_n are distributed as Beta$(1, \alpha)$:

$$p(v \mid \alpha) = \alpha(1 - v)^{\alpha - 1}.$$

Writing this in the canonical exponential family form:

$$p(v \mid \alpha) = (1/(1 - v)) \exp\{\alpha \log(1 - v) + \log \alpha\},$$

we see that $h(v) = 1/(1 - v)$, $t(v) = \log(1 - v)$, and $a(\alpha) = -\log \alpha$. Thus, we need a distribution in which $t(\alpha) = \langle \alpha, \log \alpha \rangle$.

Consider the gamma distribution for α with shape parameter s_1 and inverse scale parameter s_2:

$$p(\alpha \mid s_1, s_2) = \frac{s_2^{s_1}}{\Gamma(s_1)} \alpha^{s_1 - 1} \exp\{-s_2 \alpha\}.$$

In its canonical form the distribution on α is:

$$p(\alpha \,|\, s_1, s_2) = (1/\alpha) \exp\{-s_2\alpha + s_1 \log \alpha - a(s_1, s_2)\},$$

which is conjugate to $\text{Beta}(1, \alpha)$. The log normalizer is:

$$a(s_1, s_2) = \log \Gamma(s_1) - s_1 \log s_2,$$

and the posterior parameters conditional on data $\{v_1, \ldots, v_K\}$ are:

$$\begin{aligned} \hat{s}_2 &= s_2 - \sum_{i=1}^{K} \log(1 - v_i) \\ \hat{s}_1 &= s_1 + K. \end{aligned}$$