Bayesian Analysis (2006) **1**, Number 1, pp. 121–144

Variational Inference for Dirichlet Process Mixtures

David M. Blei* Michael I. Jordan†

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 the application of nonparametric Bayesian methods 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.

Keywords: Dirichlet processes, hierarchical models, variational inference, image processing, Bayesian computation

1 Introduction

The methodology of Monte Carlo Markov chain (MCMC) sampling has energized Bayesian statistics for more than a 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 one—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

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, http://www.cs.berkeley.edu/~blei/

†Department of Statistics and Computer Science Division, University of California, Berkeley, CA, http://www.cs.berkeley.edu/~jordan/

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 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 by developing variational methods for Dirichlet process mixtures.

The Dirichlet process (DP), introduced in Ferguson (1973), is a measure on measures. The DP is parameterized by a base distribution G_0 and a positive scaling parameter α.[1] Suppose we draw a random measure G from a Dirichlet process, and independently draw N random variables η_n from G:

$$\begin{aligned} G \,|\, \{G_0, \alpha\} &\sim \mathrm{DP}(G_0, \alpha) \\ \eta_n &\sim G, \quad n \in \{1, \ldots, N\}. \end{aligned}$$

Marginalizing out the random measure G, the joint distribution of $\{\eta_1, \ldots, \eta_N\}$ follows a Pólya urn scheme (Blackwell and MacQueen 1973). Positive probability is assigned to configurations in which different η_n take on identical values; moreover, 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 by 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.

Given a sample $\{x_1, \ldots, x_N\}$ from a DP mixture, our goal is to compute the predictive density:

$$p(x \,|\, x_1, \ldots, x_N, \alpha, G_0) = \int p(x \,|\, \eta) p(\eta \,|\, x_1, \ldots, x_N, \alpha, G_0) d\eta, \tag{1}$$

As in many hierarchical Bayesian models, the posterior distribution $p(\eta \,|\, x_1, \ldots, x_N, G_0, \alpha)$ is complicated and is not available in a closed form. MCMC provides one class of approximations for this posterior and the predictive density (MacEachern 1994; Escobar and West 1995; Neal 2000).

[1] Ferguson (1973) parameterizes the Dirichlet process by a single base measure, which is αG_0 in our notation.

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 truncate only the variational distribution.

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. In Section 3 we present a variational inference algorithms for DP mixtures. Section 4 overviews MCMC algorithms for the DP mixture, discussing algorithms based both on the Pólya urn representation and the stick-breaking representation. Section 5 presents the results of experimental comparisons, Section 6 presents an analysis of natural image data, 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 distribution G_0, if for all natural numbers k and k-partitions $\{B_1, \ldots, B_k\}$,

$$(G(B_1), G(B_2), \ldots, G(B_k)) \sim \mathrm{Dir}(\alpha G_0(B_1), \alpha G_0(B_2), \ldots, \alpha G_0(B_k)). \tag{2}$$

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

$$p(\cdot \mid \eta_1, \ldots, \eta_{n-1}) \propto \alpha G_0(\cdot) + \sum_{i=1}^{n-1} \delta_{\eta_i}(\cdot). \tag{3}$$

Thus, the variables $\{\eta_1, \ldots, \eta_{n-1}\}$ 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 $\mathbf{c} = \{c_1, \ldots, c_{n-1}\}$ be assignment variables such that $\eta_i = \eta_{c_i}^*$, and let $|\mathbf{c}|$ denote the number of cells in the partition. The distribution of η_n follows the urn distribution:

$$\eta_n = \begin{cases} \eta_i^* & \text{with prob. } \frac{|\{j : c_j = i\}|}{n-1+\alpha} \\ \eta, \eta \sim G_0 & \text{with prob. } \frac{\alpha}{n-1+\alpha}, \end{cases} \tag{4}$$

where $|\{j : c_j = i\}|$ 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 specification (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 the distinct values of the parameter. Taking this view, the DP mixture has a natural interpretation as a flexible mixture model in which the number of components (i.e., the number of cells in the partition) is random and grows as new data are observed.

The definition of the DP via its finite dimensional distributions in Equation (2) reposes on the Kolmogorov consistency theorem (Ferguson 1973). Sethuraman (1994) provides a more explicit characterization of the DP in terms of a *stick-breaking construction*. Consider two infinite collections of independent random variables, $V_i \sim \mathrm{Beta}(1, \alpha)$ and $\eta_i^* \sim G_0$, for $i = \{1, 2, \ldots\}$. The stick-breaking representation of G is as follows:

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j) \tag{5}$$

$$G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v}) \delta_{\eta_i^*}. \tag{6}$$

This representation of the DP makes clear that G is discrete (with probability one); the support of G consists of a countably infinite set of atoms, drawn independently from G_0. The mixing proportions $\pi_i(\mathbf{v})$ 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 $\mathrm{Beta}(1, \alpha)$ distribution.

In the DP mixture, the vector $\pi(\mathbf{v})$ comprises the infinite vector of mixing proportions and $\{\eta_1^*, \eta_2^*, \ldots\}$ are the atoms representing the mixture components. Let Z_n be an assignment variable of the mixture component with which the data point x_n is associated. The data can be described as arising from the following process:

1. Draw $V_i \,|\, \alpha \sim \mathrm{Beta}(1, \alpha), \quad i = \{1, 2, \ldots\}$
2. Draw $\eta_i^* \,|\, G_0 \sim G_0, \quad i = \{1, 2, \ldots\}$
3. For the nth data point:
   (a) Draw $Z_n \,|\, \{v_1, v_2, \ldots\} \sim \mathrm{Mult}(\pi(\mathbf{v}))$.
   (b) Draw $X_n \,|\, z_n \sim p(x_n \,|\, \eta_{z_n}^*)$.

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 distribution for the DP is the corresponding conjugate prior.

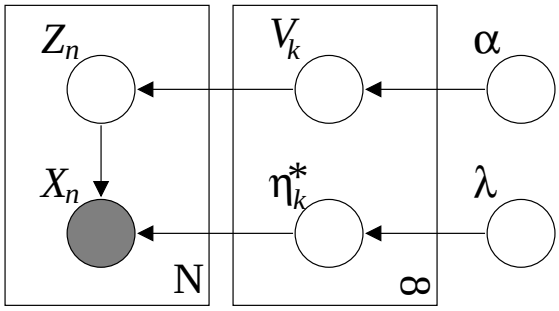

Figure 1: Graphical model representation of an exponential family DP mixture. Nodes denote random variables, edges denote possible dependence, and plates denote replication.

The stick-breaking construction for the DP mixture is depicted as a graphical model in Figure 1. The conditional 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

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

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

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

$$p(\eta^* \mid \lambda) = h(\eta^*) \exp\{\lambda_1^T \eta^* + \lambda_2(-a(\eta^*)) - a(\lambda)\}, \tag{7}$$

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

3 Variational inference for DP mixtures

There is no direct way to compute the posterior distribution under a DP mixture prior. Approximate inference methods are required for DP mixtures and Markov chain Monte Carlo (MCMC) sampling methods have become the methodology of choice (MacEachern 1994; Escobar and West 1995; MacEachern 1998; Neal 2000; Ishwaran and James 2001).

Variational inference provides an alternative, deterministic methodology for approximating likelihoods and posteriors (Wainwright and Jordan 2003). Consider a model with hyperparameters θ, latent variables $\mathbf{W} = \{W_1, \ldots, W_M\}$, and observations $\mathbf{x} = \{x_1, \ldots, x_N\}$. The posterior distribution of the latent variables is:

$$p(\mathbf{w} \mid \mathbf{x}, \theta) = \exp\{\log p(\mathbf{x}, \mathbf{w} \mid \theta) - \log p(\mathbf{x} \mid \theta)\}. \tag{8}$$

Working directly with this posterior is typically precluded by the need to compute the normalizing constant. The log marginal probability of the observations is:

$$\log p(\mathbf{x} \mid \theta) = \log \int p(\mathbf{w}, \mathbf{x} \mid \theta) d\mathbf{w}, \tag{9}$$

which may be difficult to compute given that the latent variables become dependent when conditioning on observed data.

MCMC algorithms circumvent this computation by constructing an approximate posterior based on samples from a Markov chain whose stationary distribution is the posterior of interest. Gibbs sampling is the simplest MCMC algorithm; one iteratively samples each latent variable conditioned on the previously sampled values of the other latent variables:

$$p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta) = \exp\{\log p(\mathbf{w}, \mathbf{x} \mid \theta) - \log p(\mathbf{w}_{-i}, \mathbf{x} \mid \theta)\}. \tag{10}$$

The normalizing constants for these conditional distributions are assumed to be available analytically for settings in which Gibbs sampling is appropriate.

Variational inference is based on reformulating the problem of computing the posterior distribution as an optimization problem, perturbing (or, "relaxing") that problem, and finding solutions to the perturbed problem (Wainwright and Jordan 2003). In this paper, we work with a particular class of variational methods known as *mean-field* methods. These are based on optimizing Kullback-Leibler (KL) divergence with respect to a so-called *variational distribution*. In particular, let $q_\nu(\mathbf{w})$ be a family of distributions indexed by a *variational parameter* ν. We aim to minimize the KL divergence between $q_\nu(\mathbf{w})$ and $p(\mathbf{w} \mid \mathbf{x}, \theta)$:

$$\mathrm{D}(q_\nu(\mathbf{w}) || p(\mathbf{w} \mid \mathbf{x}, \theta)) = \mathrm{E}_q \left[\log q_\nu(\mathbf{W})\right] - \mathrm{E}_q \left[\log p(\mathbf{W}, \mathbf{x} \mid \theta)\right] + \log p(\mathbf{x} \mid \theta), \tag{11}$$

where here and elsewhere in the paper we omit the variational parameters ν when using q as a subscript of an expectation. Notice that the problematic marginal probability does not depend on the variational parameters; it can be ignored in the optimization.

The minimization in Equation (11) can be cast alternatively as the maximization of a lower bound on the log marginal likelihood:

$$\log p(\mathbf{x} \mid \theta) \geq \mathrm{E}_q \left[\log p(\mathbf{W}, \mathbf{x} \mid \theta)\right] - \mathrm{E}_q \left[\log q_\nu(\mathbf{W})\right]. \tag{12}$$

The gap in this bound is the divergence between $q_\nu(\mathbf{w})$ and the true posterior.

For the mean-field framework to yield a computationally effective inference method, it is necessary to choose a family of distributions $q_\nu(\mathbf{w})$ such that we can tractably optimize Equation (11). In constructing that family, one typically breaks some of the dependencies between latent variables that make the true posterior difficult to compute. In the next sections, we consider fully-factorized variational distributions which break all of the dependencies.

3.1 Mean field variational inference in exponential families

For each latent variable, let us assume that the conditional distribution $p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta)$ is a member of the exponential family[2]:

$$p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta) = h(w_i) \exp\{g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta)^T w_i - a(g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta))\}, \tag{13}$$

where $g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta)$ is the natural parameter for w_i when conditioning on the remaining latent variables and the observations.

In this setting it is natural to consider the following family of distributions as mean-field variational approximations (Ghahramani and Beal 2001):

$$q_{\boldsymbol{\nu}}(\mathbf{w}) = \prod_{i=1}^{M} \exp\{\nu_i^T w_i - a(w_i)\}, \tag{14}$$

where $\boldsymbol{\nu} = \{\nu_1, \nu_2, \ldots, \nu_M\}$ are variational parameters. Indeed, it turns out that the variational algorithm that we obtain using this fully-factorized family is reminiscent of Gibbs sampling. In particular, as we show in Appendix 7, the optimization of KL divergence with respect to a single variational parameter ν_i is achieved by computing the following expectation:

$$\nu_i = \mathrm{E}_q \left[g_i(\mathbf{W}_{-i}, \mathbf{x}, \theta) \right]. \tag{15}$$

Repeatedly updating each parameter in turn by computing this expectation amounts to performing coordinate ascent in the KL divergence.

Notice the interesting relationship of this algorithm to the Gibbs sampler. In Gibbs sampling, we iteratively draw the latent variables w_i from the distribution $p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta)$. In mean-field variational inference, we iteratively update the variational parameter ν_i by setting it equal to the expected value of $g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta)$. This expectation is computed under the variational distribution.

3.2 DP mixtures

In this section we develop a mean-field variational algorithm for the DP mixture. Our algorithm is based on the stick-breaking representation of the DP mixture (see Figure 1). In this representation the latent variables are the stick lengths, the atoms, and the cluster assignments: $\mathbf{W} = \{\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z}\}$. The hyperparameters are the scaling parameter and the parameter of the conjugate base distribution: $\theta = \{\alpha, \lambda\}$.

Following the general recipe in Equation (12), we write the variational bound on the

[2]Examples of models in which $p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta)$ is an exponential family distribution include hidden Markov models, mixture models, state space models, and hierarchical Bayesian models with conjugate and mixture of conjugate priors.

log marginal probability of the data:

$$\begin{aligned}\log p(\mathbf{x} \mid \alpha, \lambda) \geq\ & \mathrm{E}_q\left[\log p(\mathbf{V} \mid \alpha)\right] + \mathrm{E}_q\left[\log p(\boldsymbol{\eta}^* \mid \lambda)\right] \\ & + \sum_{n=1}^{N} \left(\mathrm{E}_q\left[\log p(Z_n \mid \mathbf{V})\right] + \mathrm{E}_q\left[\log p(x_n \mid Z_n)\right]\right) \\ & - \mathrm{E}_q\left[\log q(\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z})\right].\end{aligned} \tag{16}$$

To exploit this bound, we must find a family of variational distributions that approximates the distribution of the infinite-dimensional random measure G, where the random measure is expressed in terms of the infinite sets $\mathbf{V} = \{V_1, V_2, \ldots\}$ and $\boldsymbol{\eta}^* = \{\eta_1^*, \eta_2^*, \ldots\}$. We do this by considering truncated stick-breaking representations. Thus, we fix a value T and let $q(v_T = 1) = 1$; this implies that the mixture proportions $\pi_t(\mathbf{v})$ are equal to zero for $t > T$ (see Equation 5).

Truncated stick-breaking representations have been considered previously by Ishwaran and James (2001) in the context of sampling-based inference for an approximation to the DP mixture model. Note that our use of truncation is rather different. In our case, the model is a full Dirichlet process and is not truncated; only the variational distribution is truncated. The truncation level T is a variational parameter which can be freely set; it is not a part of the prior model specification (see Section 5).

We thus propose the following factorized family of variational distributions for mean-field variational inference:

$$q(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z}) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\eta_t^*) \prod_{n=1}^{N} q_{\phi_n}(z_n) \tag{17}$$

where $q_{\gamma_t}(v_t)$ are beta distributions, $q_{\tau_t}(\eta_t^*)$ are exponential family distributions with natural parameters τ_t, and $q_{\phi_n}(z_n)$ are multinomial distributions. In the notation of Section 3.1, the free variational parameters are

$$\boldsymbol{\nu} = \{\gamma_1, \ldots, \gamma_{T-1}, \tau_1, \ldots, \tau_T, \phi_1, \ldots, \phi_N\}.$$

It is important to note that there is a different variational parameter for each latent variable under the variational distribution. 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. This reflects the conditional nature of variational inference.

Coordinate ascent algorithm

In this section we present an explicit coordinate ascent algorithm for optimizing the bound in Equation (16) with respect to the variational parameters.

All of the terms in the bound involve standard computations in the exponential family, except for the third term. We rewrite the third term using indicator random

variables:

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

Recall that $\mathrm{E}_q\left[\log(1-V_T)\right] = 0$ and $q(z_n > T) = 0$. Consequently, we can truncate this summation at $t = T$:

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

where

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

The digamma function, denoted by Ψ, arises from the derivative of the log normalization factor in the beta distribution.

We now use the general expression in Equation (15) to derive a mean-field coordinate ascent algorithm. This yields:

$$\gamma_{t,1} = 1 + \textstyle\sum_n \phi_{n,t} \tag{18}$$

$$\gamma_{t,2} = \alpha + \textstyle\sum_n \sum_{j=t+1}^{T} \phi_{n,j} \tag{19}$$

$$\tau_{t,1} = \lambda_1 + \textstyle\sum_n \phi_{n,t} x_n \tag{20}$$

$$\tau_{t,2} = \lambda_2 + \textstyle\sum_n \phi_{n,t}. \tag{21}$$

$$\phi_{n,t} \propto \exp(S_t), \tag{22}$$

for $t \in \{1,\ldots,T\}$ and $n \in \{1,\ldots,N\}$, where

$$S_t = \mathrm{E}_q\left[\log V_t\right] + \textstyle\sum_{i=1}^{t-1}\mathrm{E}_q\left[\log(1-V_i)\right] + \mathrm{E}_q\left[\eta_t^*\right]^T X_n - \mathrm{E}_q\left[a(\eta_t^*)\right].$$

Iterating these updates optimizes Equation (16) with respect to the variational parameters defined in Equation (17).

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. (This can be viewed as 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.

To compute the predictive distribution, we use the variational posterior in a manner analogous to the way that the empirical approximation is used by an MCMC sampling algorithm. The predictive distribution is:

$$p(x_{N+1} \,|\, \mathbf{x}, \alpha, \lambda) = \int \left(\sum_{t=1}^{\infty} \pi_t(\mathbf{v}) p(x_{N+1} \,|\, \eta_t^*) \right) dP(\mathbf{v}, \boldsymbol{\eta}^* \,|\, \mathbf{x}, \lambda, \alpha).$$

Under the factorized variational approximation to the posterior, the distribution of the atoms and the stick lengths are decoupled and the infinite sum is truncated. Consequently, we can approximate the predictive distribution with a product of expectations which are straightforward to compute under the variational approximation,

$$p(x_{N+1} \,|\, \mathbf{x}, \alpha, \lambda) \approx \sum_{t=1}^{T} \mathrm{E}_q \left[\pi_t(\mathbf{V}) \right] \mathrm{E}_q \left[p(x_{N+1} \,|\, \eta_t^*) \right], \tag{23}$$

where q depends implicitly on $\mathbf{x}$, α, and λ.

Finally, we remark on two possible extensions. First, when G_0 is not conjugate, a simple coordinate ascent update for τ_i may not be available, particularly when $p(\eta_i^* \,|\, \mathbf{z}, \mathbf{x}, \lambda)$ is not in the exponential family. However, such an update is available for the special case of G_0 being a mixture of conjugate distributions. Second, it is often important in applications to integrate over a diffuse prior on the scaling parameter α. As we show in Appendix 7, it is straightforward to extend the variational algorithm to include a gamma prior on α.

4 Gibbs sampling

For comparison to variational inference, we review the collapsed Gibbs sampler and blocked Gibbs sampler for DP mixtures.

4.1 Collapsed Gibbs sampling

The *collapsed Gibbs sampler* for a DP mixture with conjugate base distribution (MacEachern 1994) integrates 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 $\mathbf{c} = \{c_1, \ldots, c_N\}$. (Recall that $|\mathbf{c}|$ denotes the number of cells in the partition.)

The algorithm iteratively samples each assignment variable C_n, for $n \in \{1, \ldots, N\}$, conditional on the other cells in the partition, $\mathbf{c}_{-n}$. The assignment C_n can be one of $|\mathbf{c}_{-n}| + 1$ values: either the nth data point is in a cell with other data points, or in a cell by itself.

Exchangeability implies that C_n has the following multinomial distribution:

$$p(c_n = k \,|\, \mathbf{x}, \mathbf{c}_{-n}, \lambda, \alpha) \propto p(x_n \,|\, \mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n = k, \lambda) p(c_n = k \,|\, \mathbf{c}_{-n}, \alpha). \tag{24}$$

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:

$$p(x_n \mid \mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n = k, \lambda) = \frac{\exp\left\{a(\lambda_1 + \sum_{m \neq n} \mathbf{1}\left[c_m = k\right] x_m + x_n, \lambda_2 + \sum_{m \neq n} \mathbf{1}\left[c_m = k\right] + 1)\right\}}{\exp\left\{a(\lambda_1 + \sum_{m \neq n} \mathbf{1}\left[c_m = k\right] x_m, \lambda_2 + \sum_{m \neq n} \mathbf{1}\left[c_m = k\right])\right\}}. \tag{25}$$

The second term is given by the Pólya urn scheme:

$$p(c_n = k \mid \mathbf{c}_{-n}) \propto \begin{cases} |\{j : c_{-n,j} = k\}| & \text{if } k \text{ is an existing cell in the partition} \\ \alpha & \text{if } k \text{ is a new cell in the partition,} \end{cases} \tag{26}$$

where $|\{j : c_{-n,j} = k\}|$ denotes the number of data points in the kth cell of the partition $\mathbf{c}_{-n}$.

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 across the Monte Carlo samples:

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

For a given sample, that distribution is

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

When G_0 is not conjugate, the distribution in Equation (25) does not have a simple closed form. Effective algorithms for handling this case are given in Neal (2000).

4.2 Blocked Gibbs sampling

In the collapsed Gibbs sampler, the assignment variable C_n is drawn from a distribution that depends on the most recently sampled values of the other assignment variables. Consequently, these variables must be updated one at a time which can potentially slow down the algorithm when compared to a blocking strategy. To this end, Ishwaran and James (2001) developed a blocked Gibbs sampling algorithm based on the stick-breaking representation of Figure 1.

The main issue to face in developing a blocked Gibbs sampler for the stick-breaking DP mixture is that one needs to sample the infinite collection of stick lengths $\mathbf{V}$ before sampling the finite collection of cluster assignments $\mathbf{Z}$. Ishwaran and James (2001) face this issue by defining a truncated Dirichlet process (TDP) in which V_{K-1} is set equal to one for some fixed value K. This yields $\pi_i(\mathbf{V}) = 0$ for $i \geq K$, and converts the infinite sum in Equation (5) into a finite sum. Ishwaran and James (2001) justify substituting a

TDP mixture model for a full DP mixture model by showing that the truncated process closely approximates a true Dirichlet process when the truncation level is chosen large relative to the number of data points.

In the TDP mixture, the state of the Markov chain consists of the beta variables $\mathbf{V} = \{V_1, \ldots, V_{K-1}\}$, the mixture component parameters $\boldsymbol{\eta}^* = \{\eta_1^*, \ldots, \eta_K^*\}$, and the indicator 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

$$p(z_n = k \,|\, \mathbf{v}, \boldsymbol{\eta}^*, \mathbf{x}) = \pi_k(\mathbf{v}) p(x_n \,|\, \eta_k^*),$$

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

$$\begin{aligned} \gamma_{k,1} &= 1 + \textstyle\sum_{n=1}^{N} \mathbf{1}\left[z_n = k\right] \\ \gamma_{k,2} &= \alpha + \textstyle\sum_{i=k+1}^{K} \sum_{n=1}^{N} \mathbf{1}\left[z_n = i\right]. \end{aligned}$$

   This step follows from the conjugacy between the multinomial distribution 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 $p(\eta_k^* \,|\, \tau_k)$. This distribution is in the same family as the base distribution, with parameters

$$\begin{aligned} \tau_{k,1} &= \lambda_1 + \textstyle\sum_{i \neq n} \mathbf{1}\left[z_i = k\right] x_i \\ \tau_{k,2} &= \lambda_2 + \textstyle\sum_{i \neq n} \mathbf{1}\left[z_i = k\right]. \end{aligned} \tag{27}$$

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

$$p(x_{N+1} \,|\, \mathbf{z}, \mathbf{x}, \alpha, \lambda) = \sum_{k=1}^{K} \mathrm{E}\left[\pi_i(\mathbf{V}) \,|\, \gamma_1, \ldots, \gamma_k\right] p(x_{N+1} \,|\, \tau_k), \tag{28}$$

where $\mathrm{E}\left[\pi_i \,|\, \gamma_1, \ldots, \gamma_k\right]$ is the expectation of the product of independent beta variables given in Equation (5). This distribution only depends on $\mathbf{z}$; the other variables are needed in the Gibbs sampling procedure, but can be integrated out here. Note that this approximation has a form similar to the approximate predictive distribution under the variational distribution in Equation (23). In the variational case, however, the averaging is done parametrically via the variational distribution rather than by a Monte Carlo integral.

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

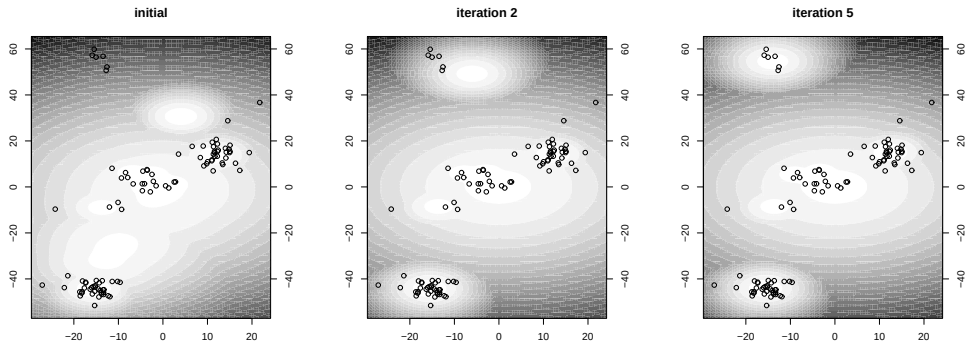

Figure 2: The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.

5 Empirical comparison

Qualitatively, variational methods offer several potential advantages over Gibbs sampling. They are deterministic, and have an optimization criterion given by Equation (16) 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 2004).

But there are several potential disadvantages of variational methods as well. First, the optimization procedure can fall prey to local maxima in the variational parameter space. Local maxima 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. 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.

We studied the performance of the variational algorithm of Section 3 and the Gibbs samplers of Section 4 in the setting of DP mixtures of Gaussians with fixed inverse covariance matrix Λ (i.e., the DP mixes over the mean of the Gaussian). The natural conjugate base distribution for the DP is Gaussian, with covariance given by Λ/λ_2 (see Equation 7).

Figure 2 provides an illustrative example of variational inference on a small problem involving 100 data points sampled from a two-dimensional DP mixture of Gaussians with diagonal covariance. Each panel in the figure plots the data and presents the

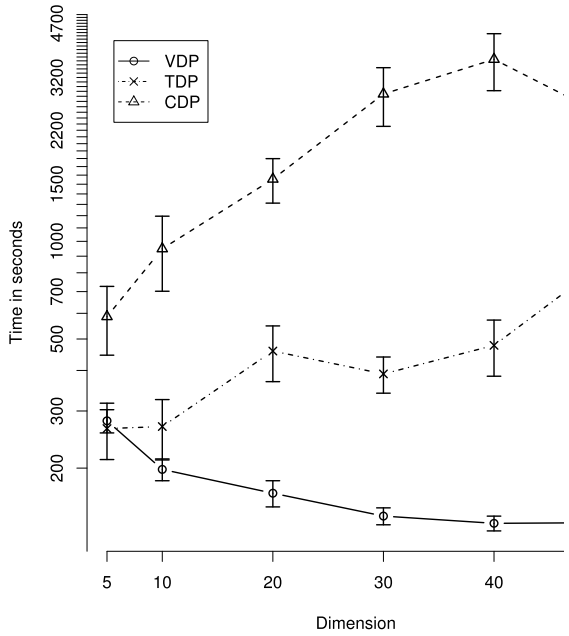

Figure 3: Mean convergence time and standard error across ten data sets per dimension for variational inference, TDP Gibbs sampling, and the collapsed Gibbs sampler.

predictive distribution given by the variational inference algorithm at a given iteration (see Equation (23)). The truncation level was set to 20. As seen in the first panel, the initialization of the variational parameters yields a largely flat distribution. 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 simulation experiments in which the dimensionality of the data was varied from 5 to 50. The covariance matrix was given by the autocorrelation matrix for a first-order autoregressive process, chosen so that the components are highly dependent ($\rho = 0.9$). The base distribution was a zero-mean Gaussian with covariance appropriately scaled for comparison across dimensions. The scaling parameter α was set equal to one.

In each case, we generated 100 data points from a DP mixture of Gaussians model of the chosen dimensionality and generated 100 additional points as held-out data. In testing on the held-out data, we treated each point as the 101st data point in the collection and computed its conditional probability using each algorithm's approximate predictive distribution.

Dim	**Mean held out log probability (Std err)**		
	Variational	Collapsed Gibbs	Truncated Gibbs
5	-147.96 (4.12)	-148.08 (3.93)	-147.93 (3.88)
10	-266.59 (7.69)	-266.29 (7.64)	-265.89 (7.66)
20	-494.12 (7.31)	-492.32 (7.54)	-491.96 (7.59)
30	-721.55 (8.18)	-720.05 (7.92)	-720.02 (7.96)
40	-943.39 (10.65)	-941.04 (10.15)	-940.71 (10.23)
50	-1151.01 (15.23)	-1148.51 (14.78)	-1147.48 (14.55)

Table 1: Average held-out log probability for the predictive distributions given by variational inference, TDP Gibbs sampling, and the collapsed Gibbs sampler.

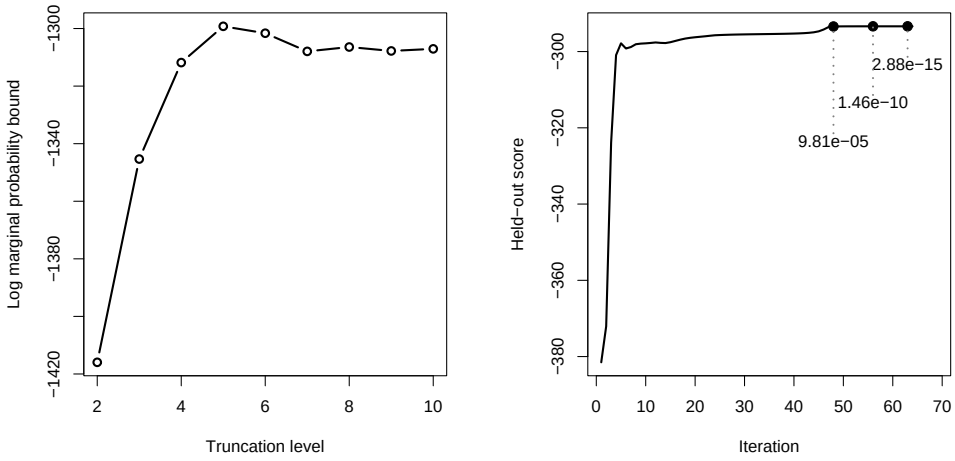

Figure 4: The optimal bound on the log probability as a function of the truncation level (left). There are five clusters in the simulated 20-dimensional DP mixture of Gaussians data set which was used. Held-out probability as a function of iteration of variational inference for the same simulated data set (right). The relative change in the log probability bound of the observations is labeled at selected iterations.

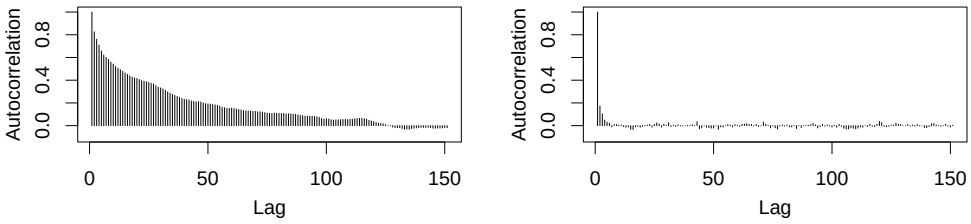

Figure 5: Autocorrelation plots on the size of the largest component for the truncated DP Gibbs sampler (left) and collapsed Gibbs sampler (right) in an example dataset of 50-dimensional Gaussian data.

The TDP approximation was truncated at $K = 20$ components. For the variational algorithm, the truncation level was also $T = 20$ components. Note that in the latter case, the truncation level is simply another variational parameter. While we held T fixed in our simulations, it is also possible to optimize T with respect to the KL divergence. Indeed, Figure 4 (left) shows how the optimal KL divergence changes as a function of the truncation level for one of the simulated data sets.

We ran all algorithms to convergence and measured the computation time.[3] For the collapsed Gibbs sampler, we assessed convergence to the stationary distribution with the diagnostic given by Raftery and Lewis (1992), and collected 25 additional samples to estimate the predictive distribution (the same diagnostic provides an appropriate lag at which to collect uncorrelated samples). We assessed convergence of the blocked Gibbs sampler using the same statistic as for the collapsed Gibbs sampler and used the same number of samples to form the approximate predictive distribution.[4]

Finally, for variational inference, we measured convergence using the relative change in the log marginal probability bound (Equation 16), stopping the algorithm when it was less than $1e^{-10}$.

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

Figure 3 illustrates the average convergence time across ten datasets per dimension. With the caveats in mind regarding convergence time measurement, it appears that the variational algorithm is quite competitive with the MCMC algorithms. The variational

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

[4] Typically, hundreds or thousands of samples are used in MCMC algorithms to form the approximate posterior. However, we found that such approximations did not offer any additional predictive performance in the simulated data. To be fair to MCMC in the timing comparisons, we used a small number of samples to estimate the predictive distributions.

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.

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 over the range tested.

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 autocorrelation plots of Figure 5. Comparing the two MCMC algorithms, we found no advantage to the truncated approximation.

Table 1 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 it is based on an approximation to the 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 ap-

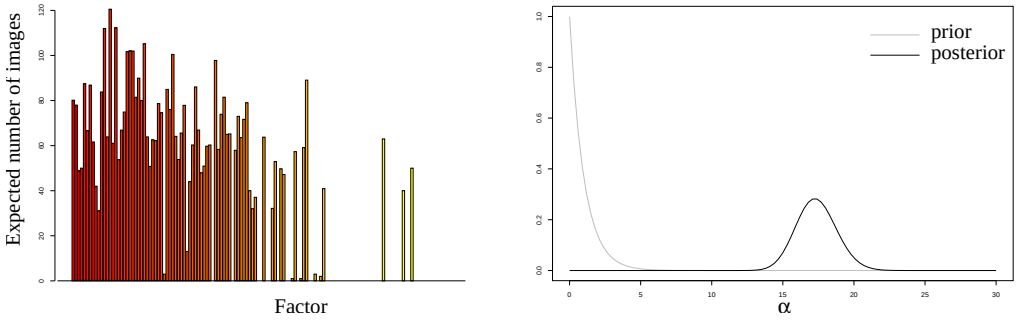

Figure 7: The expected number of images allocated to each component in the variational posterior (left). The posterior uses 79 components to describe the data. The prior for the scaling parameter α and the approximate posterior given by its variational distribution (right).

propriate number of mixture components to use in these problems is generally unknown, and DP mixtures provide an attractive alternative to current methods. However, a deployment of DP mixtures in such problems crucially requires 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 DP mixture of Gaussians model.

Each image was reduced to a 192-dimensional real-valued vector given by an 8×8 grid of average red, green, and blue values. We fit a DP mixture model in which the mixture components are Gaussian with mean μ and covariance matrix $\sigma^2 I$. The base distribution G_0 was a product measure—Gamma(4,2) for $1/\sigma^2$ and $\mathcal{N}(0, 5\sigma^2)$ for μ. Furthermore, we placed a Gamma(1,1) prior on the DP scaling parameter α, as described in Appendix 7. We used a truncation level of 150 for the variational distribution.

The variational algorithm required approximately four hours to converge. The resulting approximate posterior used 79 mixture components to describe the collection. For a rough comparison to Gibbs sampling, an iteration of collapsed Gibbs takes 15 minutes with this data set. In the same four hours, one could perform only 16 iterations. This is not enough for a chain to converge to its stationary distribution, let alone provide a sufficient number of uncorrelated samples to construct an empirical estimate of the posterior.

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

7 Conclusions

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 methods 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.

Appendix-A Variational inference in exponential families

In this appendix, we derive the coordinate ascent algorithm for variational inference described in Section 3.2. Recall that we are considering a latent variable model with hyperparameters θ, observed variables $\mathbf{x} = \{x_1, \ldots, x_N\}$, and latent variables $\mathbf{W} = \{W_1, \ldots, W_M\}$. The posterior can be written as

$$p(\mathbf{w} \,|\, \mathbf{x}, \theta) = \exp\{\log p(\mathbf{w}, \mathbf{x} \,|\, \theta) - \log p(\mathbf{x} \,|\, \theta)\}. \tag{29}$$

The variational bound on the log marginal probability is

$$\log p(\mathbf{x} \,|\, \theta) \geq \mathrm{E}_q \left[\log p(\mathbf{x}, \mathbf{W} \,|\, \theta)\right] - \mathrm{E}_q \left[\log q(\mathbf{W})\right]. \tag{30}$$

This bound holds for any distribution $q(\mathbf{w})$.

For the optimization of this bound to be computationally tractable, we restrict ourselves to fully-factorized variational distributions of the form $q_{\boldsymbol{\nu}}(\mathbf{w}) = \prod_{i=1}^{M} q_{\nu_i}(w_i)$, where $\boldsymbol{\nu} = \{\nu_1, \nu_2, \ldots, \nu_M\}$ are variational parameters and each distribution is in the exponential family (Ghahramani and Beal 2001). We derive a coordinate ascent algorithm in which we iteratively maximize the bound with respect to each ν_i, holding the other variational parameters fixed.

Let us rewrite the bound in Equation (30) using the chain rule:

$$\log p(\mathbf{x} \mid \theta) \geq \log p(\mathbf{x} \mid \theta) + \sum_{m=1}^{M} \mathrm{E}_q \left[\log p(W_m \mid \mathbf{x}, W_1, \ldots, W_{m-1}, \theta)\right] - \sum_{m=1}^{M} \mathrm{E}_q \left[\log q_{\nu_m}(W_m)\right]. \tag{31}$$

To optimize with respect to ν_i, reorder $\mathbf{w}$ such that w_i is last in the list. The portion of Equation (31) depending on ν_i is

$$\ell_i = \mathrm{E}_q \left[\log p(W_i \mid \mathbf{W}_{-i}, \mathbf{x}, \theta)\right] - \mathrm{E}_q \left[\log q_{\nu_i}(W_i)\right]. \tag{32}$$

The variational distribution $q_{\nu_i}(w_i)$ is in the exponential family,

$$q_{\nu_i}(w_i) = h(w_i) \exp\{\nu_i^T w_i - a(\nu_i)\},$$

and Equation (32) simplifies as follows:

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

because $\mathrm{E}_q [W_i] = a'(\nu_i)$.

The derivative with respect to ν_i is

$$\frac{\partial}{\partial \nu_i} \ell_i = \frac{\partial}{\partial \nu_i} \left(\mathrm{E}_q \left[\log p(W_i \mid \mathbf{W}_{-i}, \mathbf{x}, \theta)\right] - \mathrm{E}_q \left[\log h(W_i)\right]\right) - \nu_i^T a''(\nu_i). \tag{33}$$

The optimal ν_i satisfies

$$\nu_i = \left[a''(\nu_i)\right]^{-1} \left(\frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log p(W_i \mid \mathbf{W}_{-i}, \mathbf{x}, \theta)\right] - \frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log h(W_i)\right]\right). \tag{34}$$

The result in Equation (34) is general. In many applications of mean field methods, including those in the current paper, a further simplification is achieved. In particular, if the conditional distribution $p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta)$ is an exponential family distribution then

$$p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta) = h(w_i) \exp\{g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta)^T w_i - a(g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta))\},$$

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

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

Using the first derivative in Equation (34), the maximum is attained at

$$\nu_i = \mathrm{E}_q \left[g_i(\mathbf{W}_{-i}, \mathbf{x}, \theta)\right]. \tag{35}$$

We define a coordinate ascent algorithm based on Equation (35) by iteratively updating ν_i for $i \in \{1, \ldots, M\}$. Such an algorithm finds a local maximum of Equation (30) by Proposition 2.7.1 of Bertsekas (1999), under the condition that the right-hand side of Equation (32) is strictly convex.

Relaxing the two assumptions complicates the algorithm, but the basic idea remains the same. If $p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta)$ is not in the exponential family, then there may not be an analytic expression for the update in Equation (34). If $q(\mathbf{w})$ is not a fully factorized distribution, then the second term of the bound in Equation (32) becomes $\mathrm{E}_q[\log q(w_i \mid \mathbf{w}_{-i})]$ and the subsequent simplifications may not be applicable.

Further perspectives on algorithms of this kind can be found in Xing et al. (2003), Ghahramani and Beal (2001), and Wiegerinck (2000). For a more general treatment of variational methods for statistical inference, see Wainwright and Jordan (2003).

Appendix-B Placing a prior on the scaling parameter

The scaling parameter α can have a significant effect on the growth of the number of components grows with the data, and it is generally important to consider extended models which integrate over α. 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 stick-breaking representation, the gamma distribution is convenient because it is conjugate to the stick lengths. We write the gamma distribution in its canonical form:

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

where s_1 is the shape parameter and s_2 is the inverse scale parameter. This distribution is conjugate to 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

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

We extend the variational inference algorithm to include posterior updates for the scaling parameter α. The variational distribution is Gamma(w_1, w_2). The variational parameters are updated as follows:

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

and we replace α with its expectation $\mathrm{E}_q[\alpha] = w_1/w_2$ in the updates for $\gamma_{t,2}$ in Equation (19).

Bibliography

Antoniak, C. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 2(6):1152–1174. 122, 124

Attias, H. (2000). "A variational Bayesian framework for graphical models." In Solla, S., Leen, T., and Muller, K. (eds.), *Advances in Neural Information Processing Systems 12*, 209–215. Cambridge, MA: MIT Press. 121

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. 137

Bertsekas, D. (1999). *Nonlinear Programming*. Nashua, NH: Athena Scientific. 141

Blackwell, D. and MacQueen, J. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1(2):353–355. 122, 123

Blei, D., Ng, A., and Jordan, M. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3:993–1022. 121

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. 132

Escobar, M. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90:577–588. 122, 125, 141

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1:209–230. 122, 123, 124

Ghahramani, Z. and Beal, M. (2001). "Propagation algorithms for variational Bayesian learning." In Leen, T., Dietterich, T., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems 13*, 507–513. Cambridge, MA: MIT Press. 121, 122, 127, 139, 141

Ishwaran, J. and James, L. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96:161–174. 125, 128, 131

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 Information Retrieval*, 119–126. ACM Press. 137

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). "Introduction to variational methods for graphical models." *Machine Learning*, 37:183–233. 121

MacEachern, S. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics B*, 23:727–741. 122, 125, 130

— (1998). "Computational methods for mixture of Dirichlet process models." In Dey, D., Muller, P., and Sinha, D. (eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, 23–44. Springer. 125

Neal, R. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2):249–265. 122, 125, 131

Opper, M. and Saad, D. (2001). *Advanced Mean Field Methods: Theory and Practice*. Cambridge, MA: MIT Press. 121

Raftery, A. and Lewis, S. (1992). "One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo." *Statistical Science*, 7:493–497. 136

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York, NY: Springer-Verlag. 133

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4:639–650. 122, 124

Wainwright, M. and Jordan, M. (2003). "Graphical models, exponential families, and variational inference." Technical Report 649, U.C. Berkeley, Dept. of Statistics. 121, 122, 125, 126, 139, 141

Wiegerinck, W. (2000). "Variational approximations between mean field theory and the junction tree algorithm." In Boutilier, C. and Goldszmidt, M. (eds.), *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI-00)*, 626–633. San Francisco, CA: Morgan Kaufmann Publishers. 121, 141

Xing, E., Jordan, M., and Russell, S. (2003). "A generalized mean field algorithm for variational inference in exponential families." In Meek, C. and Kjærulff, U. (eds.), *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, 583–591. San Francisco, CA: Morgan Kaufmann Publishers. 141

Acknowledgments

We thank Jaety Edwards for providing the AP image data. We want to acknowledge support from Intel Corporation, Microsoft Research, and a grant from DARPA in support of the CALO project.