# Variational Methods for the Dirichlet Process

**David M. Blei** BLEI@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720

**Michael I. Jordan** JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720

## Abstract

Variational inference methods, including mean field methods and loopy belief propagation, have been widely used for approximate probabilistic inference in graphical models. While often less accurate than MCMC, variational methods provide a fast deterministic approximation to marginal and conditional probabilities. Such approximations can be particularly useful in high dimensional problems where sampling methods are too slow to be effective. A limitation of current methods, however, is that they are restricted to parametric probabilistic models. MCMC does not have such a limitation; indeed, MCMC samplers have been developed for the Dirichlet process (DP), a nonparametric distribution on distributions (Ferguson, 1973) that is the cornerstone of Bayesian nonparametric statistics (Escobar & West, 1995; Neal, 2000). In this paper, we develop a mean-field variational approach to approximate inference for the Dirichlet process, where the approximate posterior is based on the truncated stick-breaking construction (Ishwaran & James, 2001). We compare our approach to DP samplers for Gaussian DP mixture models.

## 1. Introduction

The goal of nonparametric statistics is to allow the complexity of a statistical model to grow as a function of the number of data points, in such a way that the eventual model can be as complex as necessary. This requires considering a large family of possible distributions for the data—a family which makes as few assumptions as possible about the distribution actually underlying the data. In the Bayesian approach to nonparametric statistics, it is necessary to place a prior probability distribution on this family of probability distributions. A number of such priors have been described in the literature in recent years; one important example is the Dirichlet process (DP) (Ferguson, 1973).

A DP is thus a distribution on distributions. It is a special distribution on distributions in that (with probability one) the distributions drawn from a DP are discrete. That is, a draw from a DP is a distribution that places its probability mass on a countably infinite subset of the underlying sample space. This discreteness has a number of important consequences; for example, it means that the DP framework can be used to address problems associated with inferring the structure of a model.

Consider in particular the problem of choosing the number of mixture components in a mixture model. This perennial issue can be addressed as a model selection problem using Bayes factors or other methods (Kass & Raftery, 1995). The DP provides an alternative approach via the *Dirichlet process mixture model*.

In a DP mixture, the draw from the Dirichlet process is treated as a latent variable. Thus we consider drawing a distribution from the DP, drawing parameters from this distribution (e.g., means and covariance matrices in the case of Gaussian mixtures), and drawing data conditional on those parameters. Now consider integrating over the latent variable (integrating using the Dirichlet process) and repeatedly drawing parameters from the marginal. The resulting sequence of parameters turns out to have a simple characterization in terms of a Pólya urn model (Blackwell & MacQueen, 1973). In particular, parameters take on identical values with positive probability, and the probability of

Appearing in *Proceedings of the 21$^{st}$ International Conference on Machine Learning*, Banff, Canada, 2004. 

sampling a given value increases as more parameters take on that value. Thus, the DP yields a clustering effect. New clusters continue to emerge, and after $N$ parameters are drawn, there are on average $\log N$ distinct clusters of parameters.

The DP thus provides a nonparametric prior for the parameters of a mixture model that allows the number of mixture components to grow as the size of the training set grows. While classical model selection techniques generally assume that a fixed number of components underly the data, the DP approach makes no such assumption. Moreover, the DP approach allows test data points to induce still more components—there is no assumption that test data must belong to the clusters associated with the training data.

DP mixtures have been used extensively in statistics as an alternative to model selection (Escobar & West, 1995) and have had particular application in fields such as population genetics, where there is good reason to accommodate the possibility that new components may arise with new data (Ewens, 1972). Machine learning researchers have also become aware of the opportunities offered by the DP, and various extensions of DP mixtures have been developed in recent years to place DP-based priors on graphical models such as hidden Markov models (Beal et al., 2002) and topic models (Blei et al., 2004).

A drawback with the DP approach is its dependence on Monte Carlo Markov chain (MCMC) methods for posterior inference. While such methods are accurate, they can be prohibitively slow, especially in the context of large-scale, multivariate, and highly-correlated data. One would like to be able to consider alternative inference algorithms for the DP, in particular variational methods (e.g., mean field methods and loopy belief propagation) which have provided fast deterministic alternatives to MCMC for approximating otherwise intractable posteriors in simpler settings. However, even in their most mature form, the application of variational methods has thus far been limited to finite-dimensional (i.e., parametric) models (Wainwright & Jordan, 2003).

In this paper, we develop a variational inference algorithm for the DP mixture model. This is a new direction for variational methods, departing from the traditional parametric setting in which they are usually deployed

The paper is organized as follows. In Sections 2 and 3, we review the construction of DP mixture models and MCMC algorithms for posterior inference. In Section 4, we derive a variational approximation to that posterior and describe the corresponding variational inference algorithm. Finally, in Section 5 we compare the two approaches on simulated and real data.

## 2. Dirichlet process mixture models

Let $\eta$ be a continuous random variable, $G_0$ be a non-atomic probability distribution on $\eta$, and $\alpha$ be a scalar. Suppose $G$ is a random probability distribution on $\eta$. The variable $G$ is distributed according to a *Dirichlet process* (DP) (Ferguson, 1973) with scaling parameter $\alpha$ and baseline distribution $G_0$ if, for all natural numbers $k$ and $k$-partitions of the space of $\eta$:

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

Integrating out $G$, the joint distribution on the collection of variables $\eta_{1:N}$ will exhibit a clustering effect; conditioned on $n-1$ draws, the $n$th value will, with some probability, be exactly equal to one of those draws:

$$p(\eta \,|\, \eta_{1:n-1}) \propto \alpha p(\eta \,|\, G_0) + \textstyle\sum_{i=1}^{n-1} \delta(\eta, \eta_i). \quad (1)$$

Thus, the $\eta_{1:N}$ are randomly partitioned into variables taking on the same value, where the partition structure is given by a Pòlya urn scheme (Blackwell & MacQueen, 1973). Denote the $k$ unique values which $\eta_{1:n-1}$ take on by $\eta^*_{1:k}$; the next draw from the Dirichlet process follows the urn distribution:

$$\eta_n = \begin{cases} \eta^*_i & \text{with prob } \frac{n_i}{n-1+\alpha} \\ \eta, \eta \sim G_0 & \text{with prob } \frac{\alpha}{n-1+\alpha}, \end{cases} \quad (2)$$

where $n_i$ are the number of instances of $\eta^*_i$ in $\eta_{1:n-1}$.

The DP can be used as a nonparametric prior in a hierarchical Bayesian model (Antoniak, 1974). Suppose our data are generated as follows:

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

Since the parameters are drawn from $G$, the data will cluster according to those values drawn from the same parameter value; this model is referred to as a *Dirichlet process mixture model*.

The DP mixture is also referred to as an "infinite" mixture model in that the data exhibit a finite number of components but new data can exhibit previously unseen components (Neal, 2000). This view is highlighted in the *stick-breaking construction* of $G$ (Ishwaran & James, 2001). 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:

$$\begin{aligned} \theta_i &= V_i \prod_{j=1}^{i-1}(1 - V_i) \\ G(\eta) &= \sum_{i=1}^{\infty} \theta_i \delta(\eta, \eta_i^*). \end{aligned} \tag{3}$$

(This construction's name comes from imagining successively breaking pieces off a unit-length stick with size proportional to random draws from a Beta distribution. The components of $\theta$ are the proportions of each of the infinite pieces of stick relative to its original size.)

From the perspective of infinite mixture models, $\theta$ comprise the infinite vector of mixing proportions and $\eta_{1:\infty}^*$ are parameters for the infinite set of mixture components. In the development of the subsequent algorithms, it will also be useful to consider an auxiliary variable $Z_n$ which denotes the mixture component associated with $X_n$.[1] We thus consider the following augmented model:

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

Finally, we can truncate this construction at $K$ by setting $V_{K-1} = 1$ (Ishwaran & James, 2001). Note from Eq. (3) that all $\theta_k$ for $k > K$ are zero. The resulting distribution, the *truncated Dirichlet process* (TDP), can be shown to closely approximate a true Dirichlet process for $K$ chosen large enough relative to the number of data. Guidelines for choosing $K$ as a function of $N$ are given in Ishwaran and James (2001).

### 2.1. Exponential family mixtures

In this paper, we will consider DP and TDP mixtures for which the data are drawn from an exponential family distribution with natural parameter $\eta$, and the baseline distribution $G_0$ is the conjugate prior with hyperparameter $\lambda$. In this setting, the DP mixture model is illustrated as a graphical model in Figure 1. The distributions on $V_i$ and $Z_n$ are described above. The conditional distribution of $X_n$ given $Z_n$ is:

$$p(x_n \,|\, z_n, \boldsymbol{\eta}^*) = \prod_{i=1}^{K} \left( h(x_n) \exp\{{\eta_i^*}^T x_n - a(\eta_i^*)\} \right)^{z_n^i},$$

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

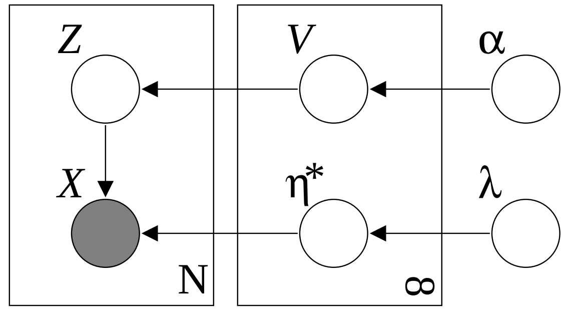


*Figure 1.* Graphical model representation of an exponential family DP mixture model.

where $a(\eta_i^*)$ is the log normalizer for whichever exponential family distribution is being used and we assume, for notational simplicity, that $X$ is its own sufficient statistic. The corresponding conjugate distribution on $\eta_i^*$ given $\lambda$ is:

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

Note that, in the standard conjugate exponential family set-up, $\lambda$ has dimension $\dim(\eta^*) + 1$ and $-a(\eta^*)$ is the last component of the sufficient statistic of $\eta^*$ (Bernardo & Smith, 1994).

## 3. MCMC for DP mixtures

In both the DP and TDP mixture models, the posterior distribution over their respective hidden variables is intractable to compute. However, Markov chain Monte Carlo (MCMC) methods have become increasingly popular for approximating these posteriors (Escobar & West, 1995; Neal, 2000; Ishwaran & James, 2001).

The idea behind MCMC is to construct a Markov chain on the hidden variables for which the stationary distribution is the posterior conditioned on the data. One collects samples from a converged Markov chain to construct an empirical estimate of the posterior, and this estimate can then be used to approximate various posterior expectations. In this work, we are interested in the predictive distribution of the next data point $x_{N+1}$ conditioned on the data $\mathbf{x} = x_{1:N}$.

One of the simplest MCMC algorithm is the Gibbs sampler, in which the Markov chain is obtained by iteratively sampling each hidden random variable conditioned on the data and the previously sampled values of the other hidden random variables. Below, we review the Gibbs sampling algorithms for the DP and TDP mixture models.

### 3.1. Collapsed Gibbs sampling

Gibbs sampling is straightforward in the exponential family DP mixture with a conjugate baseline distribution (Neal, 2000). We integrate out all random variables except $Z_n$, resulting in a *collapsed Gibbs sampler*. The algorithm is to iteratively sample each $Z_n$ from:

$$p(z_n^k = 1 \mid \mathbf{x}, \mathbf{z}_{-n}, \lambda, \alpha) \propto p(x_n \mid \mathbf{x}_{-n}, \mathbf{z}_{-n}, z_n^k = 1, \lambda) p(z_n^k = 1 \mid \mathbf{z}_{-n}, \alpha), \quad (4)$$

where $\mathbf{z}_{-n}$ denotes all the previously sampled cluster variables except for the $n$th one.

Let

$$\begin{aligned} \tau_{k,1} &= \lambda_1 + \textstyle\sum_{i \neq n} z_i^k x_i \\ \tau_{k,2} &= \lambda_2 + \textstyle\sum_{i \neq n} z_i^k. \end{aligned} \quad (5)$$

The first term of Eq. (4) is:

$$p(x_n \mid \mathbf{x}_{-n}, \mathbf{z}_{-n}, z_n^k = 1, \lambda) = \exp\left\{a(\tau_1 + X_n, \tau_2 + 1) - a(\tau_1, \tau_2)\right\}, \quad (6)$$

which is simply a ratio of normalizing constants. The second term of Eq. (4) comes from the partition structure of the Dirichlet process. When $k$ is a previously seen component:

$$p(z_n^k = 1 \mid \mathbf{z}_{-n}, \alpha) = \frac{n_k}{\alpha + N - 1}, \quad (7)$$

where $n_k$ are the number of instances of $z_n^k = 1$ in $\mathbf{z}_{-n}$. When $k$ is an unvisited component:

$$p(z_n^k = 1 \mid \mathbf{z}_{-n}, \alpha) = \frac{\alpha}{\alpha + N - 1}. \quad (8)$$

Once this chain has run for long enough, we will have in hand samples from $p(\mathbf{z} \mid \mathbf{x}, \alpha, \lambda)$ and can construct an empirical approximation to this posterior. The approximate predictive distribution for $x_{N+1}$ will be an average of the predictive distributions for each of the collected samples. For a particular sample, that distribution is:

$$p(x_{N+1} \mid \mathbf{z}, \mathbf{x}, \alpha, \lambda) = \sum_{i=1}^{K+1} p(z_{N+1}^i = 1 \mid \mathbf{z}) p(x_{N+1} \mid \mathbf{z}, \mathbf{x}, z_{N+1}^i = 1), \quad (9)$$

where $K$ are the number of components exhibited in the sample $\mathbf{z}$, and note that the next component $z_{N+1}$ can take on $K+1$ possible values.

### 3.2. Gibbs sampling for a TDP mixture

In the collapsed Gibbs sampler, the distribution of each cluster variable $Z_n$ depends on the previously sampled values of every other cluster variable. Thus, the $Z_n$ variables must be updated one at a time, which can slow down the algorithm compared with a blocking strategy. To this end, Ishwaran and James (2001) developed the TDP mixture model, which is equivalent to the model in Figure 1 except that there are only $K$ Beta variables $V_i$ and mixture component parameters $\eta_i^*$. We review its corresponding Gibbs sampler here.

Rather than integrating out $G$, the random distribution drawn from the DP, the TDP explicitly represents its approximation via the truncated stick-breaking construction. The Beta variables $V_{1:K}$ determine the mixing proportions using Eq. (3) truncated at $K$, and the parameters $\eta_{1:K}^*$ are associated with the $K$ different mixture components.

Given data $\mathbf{x}$, the Gibbs sampler for the TDP mixture model iterates between the following steps:

1. Conditioned on $\mathbf{v}$, $\boldsymbol{\eta}^*$, and $\mathbf{x}$, the variables $\mathbf{Z}$ are sampled independently:

   $$p(z_n^k = 1 \mid \mathbf{v}, \boldsymbol{\eta}^*, \mathbf{x}) = \theta_k p(x_n \mid \eta_k^*),$$

   where $\theta_k$ is a function of $\mathbf{V}$ as given in Eq. (3).

2. Conditioned on $\mathbf{z}$ and $\mathbf{x}$, the $V_k$ variables are independently sampled from $\text{Beta}(\gamma_{k,1}, \gamma_{k,2})$, where:

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

3. Conditioned on $\mathbf{z}$ and $\mathbf{x}$, the mixture component parameters $\eta_k^*$ are sampled from the appropriate posterior distribution $p(\eta_k^* \mid \tau_k)$, where $\tau_k$ is defined in Eq. (5).

After a sufficient burn-in period, we can collect samples and construct an approximate predictive distribution of the next data point. Again, this distribution will be an average of the predictive distributions for each of the collected states. The state-specific distribution is:

$$p(x_{N+1} \mid \mathbf{z}, \alpha, \lambda) = \textstyle\sum_{i=1}^K \mathrm{E}\left[\theta_i \mid \boldsymbol{\gamma}\right] p(x_{N+1} \mid \tau_i), \quad (10)$$

where $\mathrm{E}\left[\theta_i \mid \boldsymbol{\gamma}\right]$ is the expectation of the product of independent Beta variables given in Eq. (3). Note that this distribution only depends on $\mathbf{z}$. The other variables are needed in the Gibbs sampling procedure, but can be integrated out here.

## 4. Variational inference

Mean-field variational inference provides an alternative, deterministic method for approximating likelihoods and posteriors in an intractable probabilistic model (Jordan et al., 1999). Given a model with observed variables $\mathbf{x}$ and hidden variables $\mathbf{H}$, we can lower bound the log likelihood using Jensen's inequality:

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int_{\mathbf{h}} p(\mathbf{x}, \mathbf{h}) d\mathbf{h} \\
&= \log \int_{\mathbf{h}} \frac{q(\mathbf{h}) p(\mathbf{x}, \mathbf{h})}{q(\mathbf{h})} d\mathbf{h} \\
&\geq \int_{\mathbf{h}} q(\mathbf{h}) \log p(\mathbf{x}, \mathbf{h}) - \int_{\mathbf{h}} q(\mathbf{h}) \log q(\mathbf{h}) \\
&= \mathrm{E}\left[\log p(\mathbf{x}, \mathbf{H})\right] - \mathrm{E}\left[\log q(\mathbf{H})\right], \qquad (11)
\end{aligned}$$

for an arbitrary density $q(\mathbf{h})$.

The idea behind variational methods is to restrict $q(\mathbf{h})$ to a parametric family such that optimizing the bound in Eq. (11) is tractable. The solution to this optimization problem provides a lower bound on the log probability of the observed variables. Furthermore, the optimal $q$ is the distribution closest in KL to the true posterior within the chosen parametric family.

### 4.1. Variational inference for a DP mixture

We can apply the mean-field variational approach to the stick-breaking construction of the DP mixture in Figure 1. The hidden variables of the model are $\mathbf{V}$, $\boldsymbol{\eta}^*$, and $\mathbf{Z}$; the variables $\boldsymbol{\eta}^*$ and $\mathbf{Z}$ are coupled in the likelihood, making it intractable to compute. Thus, we will introduce a variational distribution $q(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z})$ in which all the hidden variables are independent. The bound on the likelihood of the data given by Eq. (11) is:

$$\begin{aligned}
\log p(\mathbf{x} \mid \alpha, \lambda) \geq\ & \mathrm{E}\left[\log p(\mathbf{V} \mid \alpha)\right] \\
&+ \mathrm{E}\left[\log p(\boldsymbol{\eta}^* \mid \lambda)\right] \\
&+ \textstyle\sum_{n=1}^{N} \mathrm{E}\left[\log p(Z_n \mid \mathbf{V})\right] \qquad (12) \\
&+ \textstyle\sum_{n=1}^{N} \mathrm{E}\left[\log p(x_n \mid Z_n)\right] \\
&- \mathrm{E}\left[\log q(\mathbf{Z}, \mathbf{V}, \boldsymbol{\eta}^*)\right].
\end{aligned}$$

To define $q$, we need to construct a distribution on an infinite set of $V_k$ and $\eta_k^*$ random variables. For this approach be tractable, we truncate the variational distribution at some value $K$ by setting $q(V_K = 1) = 1$. As in the truncated Dirichlet process, the mixture proportions $\theta_k$ for $k > K$ will be zero, and we can ignore $\eta_k^*$ for $k > K$. Note that the blocked Gibbs sampler of Section 3.2 estimates the posterior of a model which is a truncated approximation to the DP. In contrast, we are using a truncated process to approximate the posterior of the full DP mixture model. The truncation level $K$ is a variational parameter, and is not a part of the model specification. The factorized variational distribution is thus:

$$q(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z}, K) = \textstyle\prod_{i=1}^{K-1} q(v_i \mid \gamma_i) \prod_{i=1}^{K} q(\eta_i^* \mid \tau_i) \prod_{n=1}^{N} q(z_n \mid \phi_n), \qquad (13)$$

where $\gamma_n$ are the Beta parameters for the distributions on $V_i$, $\tau_i$ are natural parameters for the distributions on $\eta_i^*$, and $\phi_n$ are multinomial parameters for the distributions on $Z_n$.

In the model of Figure 1, all the $\mathbf{V}$, $\boldsymbol{\eta}^*$, and $\mathbf{Z}$ variables are governed by the same distributions. It is important to emphasize that, under the variational approximation, there is a different distribution for each variable. For example, each data point $x_n$ is associated with a different distribution over its corresponding hidden factor $Z_n$.

The first, second, fourth, and fifth terms in Eq. (12) correspond to standard computations in an exponential family distribution. We rewrite the third term with indicator random variables:

$$\mathrm{E}\left[\log p(Z_n \mid \mathbf{V})\right] = \mathrm{E}\left[\log\left(\prod_{i=1}^{K} (1 - V_i)^{\mathbf{1}[Z_n > i]} V_i^{Z_n^i}\right)\right].$$

This expectation simplifies to:

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

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 $\Psi$ is the digamma function, which arises from the derivative of the log normalization factor in the beta distribution.)

Optimization of Eq. (12) is a coordinate ascent algorithm in the variational parameters. The updates for $\tau_n$ and $\gamma_n$ follow the standard recipe for variational inference with exponential family distributions in a conjugate setting (Ghahramani & Beal, 2001):

$$\begin{aligned}
\gamma_{i,1} &= 1 + \textstyle\sum_n \phi_{n,i} \\
\gamma_{i,2} &= \alpha + \textstyle\sum_n \sum_{j=i+1}^{K} \phi_{n,j} \\
\tau_{i,1} &= \lambda_1 + \textstyle\sum_n \phi_{n,i} x_n \\
\tau_{i,2} &= \lambda_2 + \textstyle\sum_n \phi_{n,i}.
\end{aligned}$$

We have intentionally overloaded notation. The variational updates are similar to Gibbs updates in Section 3.2, with the values of the random variables replaced by their means under the variational distribution.

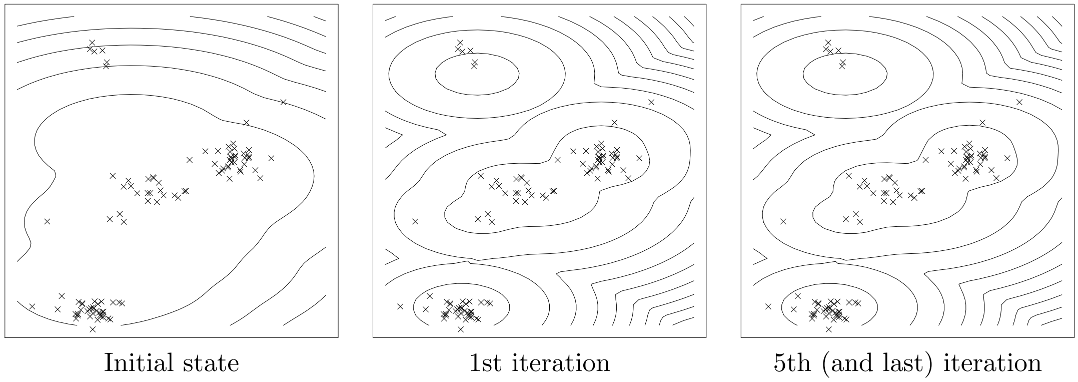


*Figure 2.* The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.

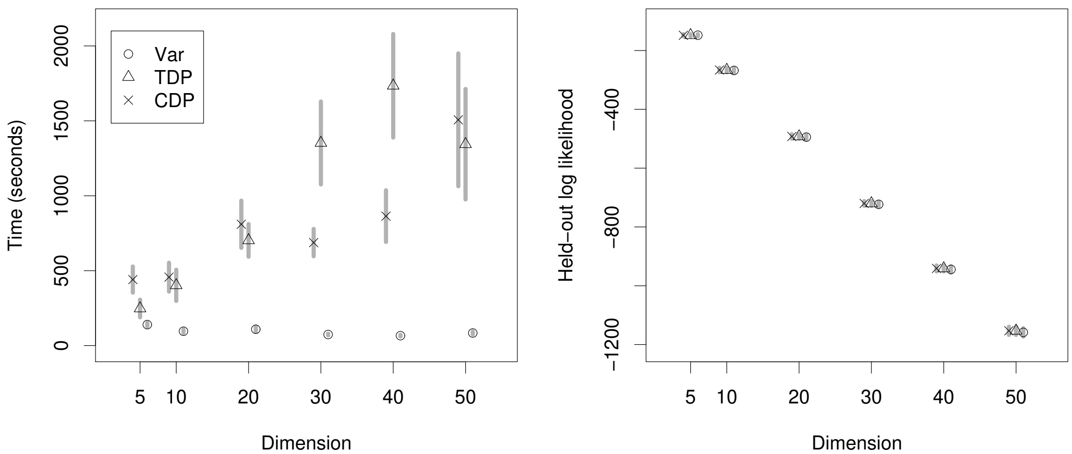


*Figure 3.* (Left) Convergence time per dimension across ten datasets for variational inference (Var), the TDP Gibbs sampler (TDP), and the collapsed Gibbs sampler (CDP). Grey bars are standard error. (Right) Average held-out log likelihood for the corresponding predictive distributions.

The update for the variational multinomial on $Z_n$ is $\phi_{n,i} \propto \exp(E)$ where:

$$\begin{aligned} E = {} & \mathrm{E}\left[\log V_i \,|\, \gamma_i\right] + \mathrm{E}\left[\eta_i \,|\, \tau_i\right]^T X_n \\ & - \mathrm{E}\left[a(\eta_i) \,|\, \tau_i\right] + \textstyle\sum_{j=1}^{i-1} \mathrm{E}\left[\log(1 - V_j) \,|\, \gamma_j\right]. \end{aligned}$$

Iterating between these updates is a coordinate ascent algorithm for optimizing Eq. (12) with respect to the parameters in Eq. (13). We thus find $q(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z})$ which is closest, within the confines of its parameters, to the true posterior. This yields an approximate predictive distribution of the next data point given, as in the TDP Gibbs sampler for a single sample, by Eq. (10).

## 5. Example and Results

We applied the variational algorithm of Section 4 and the two Gibbs samplers of Section 3 to Gaussian-Gaussian DP mixture models. The data are assumed drawn from a multivariate Gaussian with fixed covariance matrix; the mean of each data point is drawn from a DP with a Gaussian baseline distribution (i.e., the conjugate prior).

In Figure 2, we illustrate the variational inference algorithm on a toy problem. We have simulated 100 data points from a two-dimensional Gaussian-Gaussian DP mixture with diagonal covariance. We illustrate the data and the predictive distribution given by the variational inference algorithm of Section 4 with variational truncation level $K$ equal to 20. In the initial setting, the variational approximation places a largely flat distribution on the data. After one iteration, the algorithm has found the various modes of the data and, after convergence, it has further refined those modes. Notice that even though we represent 20 mixture components in the variational distribution, the fitted approximate posterior only uses five of them.

### 5.1. Simulated mixture models

To compare our algorithm to the Gibbs samplers of Section 3, we performed the following simulation. We generated 100 data points from a Gaussian-Gaussian DP mixture model and 100 additional points as held-out data. In the held-out data, each point is treated as the 101st data point in the collection and only depends on the first 100 points. The fixed covariance of the data is given by a first-order autocorrelation matrix such that the components are highly dependent. The baseline distribution for the mean is a zero-mean Gaussian with covariance appropriately scaled for comparison across dimensions.

We run the three algorithms to convergence and measure the computation time.[2] 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 gives an appropriate lag to collect uncorrelated samples). The variational algorithm and TDP Gibbs sampling algorithm are run with truncation level $K$ equal to 20.

In Figure 3 (left), we illustrate the average convergence time for the algorithms as a function of dimensionality, averaging across ten datasets. We see that the variational algorithm is much faster than the sampling methods and exhibits significantly less variance in its convergence time.

It is noteworthy that the collapsed Gibbs sampler usually converges faster than the TDP Gibbs sampler. Though an iteration of collapsed Gibbs is much slower than an iteration of TDP Gibbs, the TDP Gibbs sampler requires a much longer burn-in and greater lag to obtain uncorrelated samples. This is illustrated in the autocorrelation plots in Figure 4.

In Figure 3 (right), we illustrate the average log likelihood assigned to the held-out data by the approximate predictive distributions given by each algorithm. First, notice that the collapsed Gibbs sampler assigned the same likelihood as 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 yields a similar held-out likelihood as that based on samples. Though the fitted variational parameters provide only an approximation to the true posterior, the resulting predictive distributions are very accurate in this setting.

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

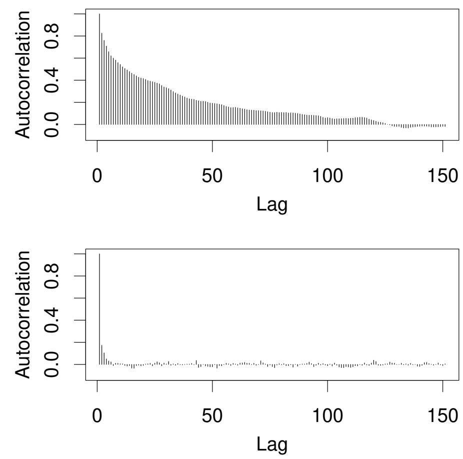


*Figure 4.* Autocorrelation plots on the size of the largest component for the TDP Gibbs sampler (top) and collapsed Gibbs sampler (bottom) in a simulated dataset of 50-dimensional Gaussian data.

| **Alg.** | **Time (sec.)** | **Held-out like.** | **# Comp.** |
|---|---|---|---|
| Var. | 2,819 | -3098.572 | 7 |
| Trunc. | 22,944 | -3097.436 | 8 |
| Coll. | 92,635 | -3097.352 | 5 |

*Table 1.* Variational inference, TDP Gibbs, and collapsed Gibbs inference for the robot data of Section 5.2. Each Markov chain is run to convergence and 25 uncorrelated samples are collected.

### 5.2. Robot data

To assess the quality of our algorithm on a larger dataset, we applied the Gaussian-Gaussian DP mixture to the Pumadyn data from Ueda and Ghahramani (2002). These data are in eight dimensions, which come from realistic simulations of a robot arm. We use 7000 points as training data and keep a held-out set of 250 points. The covariance matrix is the sample covariance, and the mean of the hyperparameter is the sample mean.

Table 1 gives the results of approximate inference under the three algorithms which we are considering. In this case, both the collapsed Gibbs sampler and TDP Gibbs sampler require the same lag (20 iterations) to produce uncorrelated samples. Even though the TDP Gibbs sampler requires more samples to converge, it is faster than collapsed Gibbs in this case. Variational inference is the fastest algorithm to converge, but attains a held-out likelihood which is slightly lower than the likelihoods attained by the Gibbs samplers.

## 6. Summary

Bayesian nonparametric models based on the Dirichlet process are powerful tools for flexible data analysis but have thus far been impractical for large collections of multivariate and highly correlated data. We have developed a fast mean-field variational inference algorithm for the Dirichlet process mixture model and demonstrated its applicability to the kinds of multivariate data in which Gibbs sampling algorithms are slow to converge.

There are three natural next steps in the development of this family of algorithms. First, we would like to test this approach in the setting of hierarchical nonparametric models such as those in Teh et al. (2004) and Blei et al. (2004). Second, we currently fix the variational truncation level to a value known to be beyond the number of components which underlie the data. Optimizing Eq. (12) with respect to this parameter may yield an even faster algorithm for approximate inference.

Finally, the mean-field variational method is only one type of variational algorithm that can be used in this problem. Other approaches, such as those described in Wainwright and Jordan (2003), can also be explored in this context.


## Acknowledgments

David M. Blei is supported by a fellowship from the Microsoft Corporation. We also would like to acknowledge a grant from DARPA for the CALO project.


## References


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

Beal, M., Ghahramani, Z., & Rasmussen, C. (2002). The infinite hidden Markov model. *Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Bernardo, J., & Smith, A. (1994). *Bayesian theory*. Chichester: John Wiley & Sons Ltd.

Blackwell, D., & MacQueen, J. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics*, *1*, 353–355.

Blei, D., Griffiths, T., Jordan, M., & Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

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

Ewens, W. (1972). The sampling theory of selective neutral alleles. *Theor. Popul. Biol.*, *3*, 87–112.

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

Ghahramani, Z., & Beal, M. (2001). Propagation algorithms for variational Bayesian learning. *Neural Information Processing Systems 13*.

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

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

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

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

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

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). *Hierarchical Dirichlet processes* (Technical Report 653). U.C. Berkeley, Dept. of Statistics.

Ueda, N., & Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Netw.*, *15*, 1223–1241.

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