Variational Inference for Stick-Breaking Beta Process Priors

John Paisley[1] JPAISLEY@PRINCETON.EDU
Lawrence Carin[2] LCARIN@DUKE.EDU
David Blei[1] BLEI@PRINCETON.EDU
[1] Department of Computer Science, Princeton University, Princeton, NJ, USA
[2] Department of Electrical & Computer Engineering, Duke University, Durham, NC, USA

Abstract

We present a variational Bayesian inference algorithm for the stick-breaking construction of the beta process. We derive an alternate representation of the beta process that is amenable to variational inference, and present a bound relating the truncated beta process to its infinite counterpart. We assess performance on two matrix factorization problems, using a non-negative factorization model and a linear-Gaussian model.

1. Introduction

The beta process (Hjort, 1990) has recently found use in machine learning as a Bayesian nonparametric prior for sparse latent feature models (Ghahramani et al., 2007), for example in latent topic modeling (Williamson et al., 2010) and image reconstruction (Zhou et al., 2009). The beta process is closely related to the Indian buffet process (IBP) (Griffiths & Ghahramani, 2006), which can be linked to the beta process after a slight parameter modification (Thibaux & Jordan, 2007). Teh et al. (2007) presented a fully Bayesian representation of the IBP, for which Doshi-Velez et al. (2009) derived a variational inference algorithm. This representation is of the original one-parameter IBP, and does not extend to the beta process presented in (Hjort, 1990).

Recently, Paisley et al. (2010) derived a stick-breaking construction of the beta process that differs from that in (Teh et al., 2007); we discuss this difference in Section 4. Paisley et al. (2010) presented an inference algorithm that relied heavily on Monte Carlo integration to avoid learning many parameters in the model. This approximate integration gave another inference algorithm for marginalized beta processes.

In this paper, we present an algorithm for performing variational Bayesian inference (Jordan et al., 1999) for the stick-breaking construction of the beta process. Using a simpler representation of the construction given in (Paisley et al., 2010), this inference algorithm does not marginalize any parameters of the model, but rather approximates the full posterior.

We truncate the posterior of the beta process for variational inference. We present a bound on the closeness of this truncation to its infinite counterpart that parallels the bound given for truncated Dirichlet processes (Ishwaran & James, 2001), and is similar to the bound given in (Doshi-Velez et al., 2009) for truncated IBPs. We assess the performance of the variational algorithm on a non-negative matrix factorization model and a linear-Gaussian model.

2. Constructing Beta Processes

The beta process is a Bayesian nonparametric method for generating an infinite collection of atoms with corresponding weights that have degenerate beta distributions. That is, let Ω be a space and $\mathcal{B}$ be the set of all measurable subsets of that space. Let H_0 be a non-atomic measure on $(\Omega, \mathcal{B})$ with $H_0(\Omega) = \gamma$ and γ finite, and let $\alpha > 0$. Then H is a beta process if

$$H(d\omega) \sim \text{Beta}(\alpha H_0(d\omega), \alpha(1 - H_0(d\omega))) \tag{1}$$

for an infinitesimal set $d\omega \in \mathcal{B}$. This definition differs slightly from (Hjort, 1990) in that $\gamma < \infty$, α is constant and the space can be more general than $\mathbb{R}_+$.

We can write H in the form $H = \sum_{k=1}^{\infty} \pi_k \delta_{\omega_k}$, where π_k tends to decrease as k increases (Ghahramani et al., 2007; Paisley et al., 2010; Teh et al., 2007). H parameterizes a Bernoulli process, denoted $X_n \sim \text{BeP}(H)$; see (Thibaux & Jordan, 2007) for more details.

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

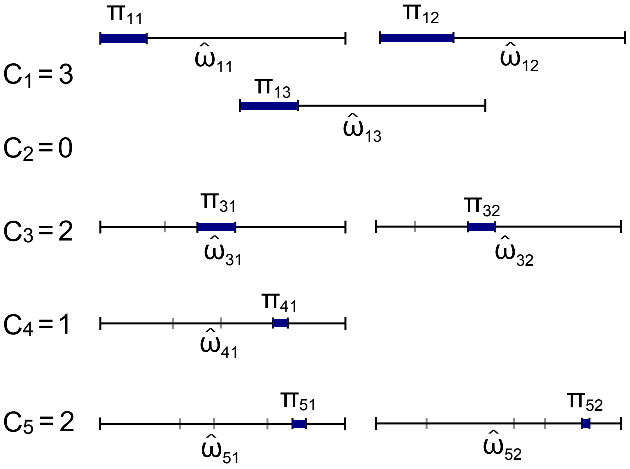

Figure 1. An illustration of the stick-breaking construction of the beta process. Each stick corresponds to an atom drawn i.i.d. from a base H_0. The length of each blue horizontal bar is a specific atom's weight. In round i, this weight equals the ith break from a Beta$(1, \alpha)$ stick-breaking process. The value of C_i denotes the number of atoms contributed in round i. A beta process is $H = \sum_{ij} \pi_{ij} \delta_{\hat{\omega}_{ij}}$.

2.1. A Stick-Breaking Construction of the Beta Process

Paisley et al. (2010) presented a method for constructing H in which each weight and atom is indexed by two values, $(\pi_{ij}, \hat{\omega}_{ij})$. The intuition behind this construction is that atoms are introduced into the model in "rounds" (indexed by i), and an atom in round i is one of a collection of atoms in that round (indexed by j). The number of atoms in round i is a random variable, C_i, with a Poisson distribution. The weight given to an atom in the ith round is the ith break from an *atom-specific* Beta$(1, \alpha)$ stick-breaking process (Sethuraman, 1994). We illustrate this process in Figure 1.

With this intuition, Paisley et al. (2010) showed that H is a beta process if

$$\begin{aligned} H &= \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} \hat{V}_{ij}^{(i)} \prod_{l=1}^{i-1} (1 - \hat{V}_{ij}^{(l)}) \delta_{\hat{\omega}_{ij}}, \\ \hat{V}_{ij}^{(l)} &\overset{iid}{\sim} \text{Beta}(1, \alpha), \\ C_i &\overset{iid}{\sim} \text{Poisson}(\gamma), \\ \hat{\omega}_{ij} &\overset{iid}{\sim} \frac{1}{\gamma} H_0, \end{aligned} \tag{2}$$

which we denote $H \sim \text{BP}(\alpha, H_0)$. Hats are used over variables that will be re-indexed in the alternate representation given in Section 2.2. Beta processes are useful as sparse priors for matrix factorization models, which we consider in Section 5.

2.2. A Simpler Representation

We derive a simpler representation of H that will allow for easier approximate posterior inference with variational Bayesian methods. The first step is to re-index H in (2) into a single summation. We define a new latent indicator d_k which marks the round in which the kth atom *overall* appears. That is,

$$d_k = 1 + \sum_{i=1}^{\infty} \mathbb{I}\left(\sum_{j=1}^{i} C_j < k \right). \tag{3}$$

For example, if $d_5 = 3$, then the fifth atom in H occurs in round three. Therefore, $C_1 < 5$ and $C_1 + C_2 < 5$, but $C_1 + C_2 + C_3 \geq 5$, and equation (3) returns the correct round.[1] If $C_1 = 3, C_2 = 0, C_3 = 2, C_4 = 1, \ldots$, then we encode this in the vector $\boldsymbol{d} = (1, 1, 1, 3, 3, 4, \ldots)$, with d_k being the kth element of $\boldsymbol{d}$. Using this latent indicator, we rewrite H in (2) as

$$H = \sum_{k=1}^{\infty} V_{k,d_k} \prod_{l=1}^{d_k - 1} (1 - V_{k,l}) \delta_{\omega_k}, \tag{4}$$

with ω and V drawn as before.

We next represent $\prod_{l<d_k}(1 - V_{k,l})$ using a single random variable. Let $T_k := -\sum_{l<d_k} \ln(1 - V_{k,l})$. Since each individual $-\ln(1 - V_{k,l}) \overset{iid}{\sim}$ Exponential(α), it follows that $T_k \sim \text{Gamma}(d_k - 1, \alpha)$. Therefore, the random variable $\exp\{-T_k\} \overset{d}{=} \prod_{l<d_k}(1 - V_{k,l})$.

Finally, we relax the strict ordering of $\boldsymbol{d}$; we no longer require that $d_k \leq d_{k+1}$, but only that the total number of any given integer appearing in $\boldsymbol{d}$ has a Poisson(γ) distribution. This gives the following representation of the beta process,

$$\begin{aligned} H &= \sum_{k=1}^{\infty} V_k \mathrm{e}^{-T_k} \delta_{\omega_k}, \\ V_k &\overset{iid}{\sim} \text{Beta}(1, \alpha), \\ T_k &\sim \text{Gamma}(d_k - 1, \alpha), \\ \sum_{k=1}^{\infty} \mathbf{1}_{d_k}(r) &\overset{iid}{\sim} \text{Poisson}(\gamma), \quad r \in \mathbb{N}_+, \\ \omega_k &\overset{iid}{\sim} \frac{1}{\gamma} H_0. \end{aligned} \tag{5}$$

Note that each d_k does not have a distribution, but instead the cardinality of $\{k : d_k = r\}$ is drawn. Also, although we indicate in (5) that T_k is drawn when $d_k = 1$ (with $T_k = 0$ with probability one), in the next section we parameterize the gamma prior such that this distribution only applies when $d_k > 1$.

[1] We use the notation $\mathbb{I}(a > b)$ to indicate $a > b$, and the more compact $\mathbf{1}_a(b)$ to indicate $a = b$.

3. Variational Inference for the SB-BP

We derive a mean-field variational inference algorithm (Jordan et al., 1999) for approximate posterior inference of H as represented in Section 2.2. Variational inference methods approximate the true, full posterior of a set of latent variables with a simpler, factorized distribution $Q(\cdot) = \prod q(\cdot)$; the form of Q is defined in advance. Each individual q distribution is defined on a single (or subset) of parameters and is given a specific functional form with variational parameters. In variational inference, we fit these parameters to minimize the KL divergence between Q and the full posterior.

3.1. Joint Likelihood of the SB-BP Model

Let $\mathcal{D}$ represent the data, Θ be the set of all latent variables in the model and Υ all the hyperparameters. Let $\theta_k = \{V_k, T_k, d_k, z_{1:N,k}\}$ be the variables for each atom, $\Theta_H = \{\theta_k\}$ and Θ_{-H} be all other model-specific variables. We include gamma priors on α and γ

$$\alpha \sim \mathrm{Gamma}(a_1, a_2), \qquad \gamma \sim \mathrm{Gamma}(b_1, b_2). \tag{6}$$

The joint likelihood of the model is $p(\mathcal{D}, \Theta|\Upsilon) = p(\mathcal{D}, \Theta_{-H}|\Theta_H, \Upsilon)p(\Theta_H|\Upsilon)$. We focus on the hidden variables for the beta-Bernoulli process prior,

$$p(\Theta_H|\Upsilon) = p(\alpha)p(\gamma)p(\boldsymbol{d}|\gamma) \times \prod_{k=1}^{\infty} p(V_k|\alpha)p(T_k|d_k, \alpha) \textstyle\prod_{n=1}^{N} p(z_{nk}|V_k, T_k, d_k). \tag{7}$$

The data and model-specific variables are contained in $p(\mathcal{D}, \Theta_{-H}|\Theta_H, \Upsilon)$, which is left undefined. Later, we will consider matrix factorization problems, in which case these are the relevant terms. We expand

$$p(z_{nk}|V_k, T_k, d_k) = p(z_{nk}|V_k)^{\mathbf{1}_{d_k}(1)} p(z_{nk}|V_k, T_k)^{\mathbb{I}(d_k>1)}$$

to account for the round in which an atom appears. This representation activates T_k when it becomes part of the model, i.e. when $d_k > 1$. We also focus on two terms in (7):

$$p(T_k|d_k, \alpha) = \frac{\alpha^{v_k(1)}}{\prod\limits_{r\geq 2} \Gamma(r-1)^{\mathbf{1}_{d_k}(r)}} T_k^{v_k(2)}\, \mathrm{e}^{-\alpha T_k \mathbb{I}(d_k>1)},$$

where $v_k(s) := \Sigma_{r\geq 2}(r-s)\mathbf{1}_{d_k}(r)$, and

$$p(\boldsymbol{d}|\gamma) = \prod_{r=1}^{\infty} \frac{\gamma^{\sum_k \mathbf{1}_{d_k}(r)}}{\{\sum_k \mathbf{1}_{d_k}(r)\}!} \mathrm{e}^{-\gamma \mathbb{I}\left(\sum\limits_{r'=r}^{\infty} \sum\limits_{k=1}^{\infty} \mathbf{1}_{d_k}(r')>0\right)}.$$

We use several indicator functions in these probability distributions to obtain the proper form of the joint likelihood. In $p(T_k|d_k, \alpha)$, indicators are used to select the parameters for the gamma prior distribution on T_k, or remove this term if $d_k = 1$. The distribution $p(\boldsymbol{d}|\gamma)$ is a product of Poisson distributions with C_r replaced by $\sum_{k=1}^{\infty} \mathbf{1}_{d_k}(r)$. The term $\mathbb{I}(\sum_{r'=r}^{\infty} \sum_{k=1}^{\infty} \mathbf{1}_{d_k}(r') > 0)$ is introduced in the exponential for inference purposes, which we discuss in detail in Section 3.3.3. Under the infinite beta process prior, this indicator is equal to one with probability one, and therefore does not change the form of the Poisson distribution.

3.2. Variational Posterior and Lower Bound

Since the model evidence $p(\mathcal{D}|\Upsilon)$ is intractable, the posterior $p(\Theta|\mathcal{D}, \Upsilon)$ cannot be found by normalizing the joint likelihood. We therefore approximate the true posterior with a factorized variational distribution. We use two truncations for inference; we truncate the number of factors at K, and the number of rounds at R. The variational distribution is

$$Q = q(\alpha)q(\gamma) \prod_{k=1}^{K} q(d_k)q(V_k)q(T_k) \textstyle\prod_{n=1}^{N} q(z_{nk}), \tag{8}$$

and we select q distributions as follows,

$$\begin{aligned}
q(d_k) &= \mathrm{Multinomial}(d_k|\varphi_k), \\
q(z_{nk}) &= \mathrm{Bernoulli}(z_{nk}|\phi_{nk}), \\
q(V_k) &= \mathrm{Beta}(V_k|a'_k, b'_k), \\
q(T_k) &= \mathrm{Gamma}(T_k|u'_k, v'_k), \\
q(\alpha) &= \mathrm{Gamma}(\alpha|\kappa_1, \kappa_2), \\
q(\gamma) &= \mathrm{Gamma}(\gamma|\tau_1, \tau_2).
\end{aligned} \tag{9}$$

Let Ψ be the set of variational parameters. We expand the lower bound $\mathcal{L}(\mathcal{D}, \Psi) = \mathbb{E}_Q[\ln p(\mathcal{D}, \Theta|\Upsilon)] - \mathbb{E}_Q[\ln Q]$ for the SB-BP prior terms below,

$$\begin{aligned}
\mathcal{L}(\mathcal{D}, \Psi) &= \mathbb{E}_Q\left[\ln p(\mathcal{D}, \Theta_{-H}|\Theta_H, \Upsilon)\right] \;\ldots \\
&+ \sum_{n,k} \varphi_k(1)\mathbb{E}_q[\ln p(z_{nk}|V_k)] + \sum_{k=1}^{K} \mathbb{E}_q[\ln p(T_k|\alpha, d_k)] \\
&+ \sum_{n,k} \varphi_k(r>1)\mathbb{E}_q[\ln p(z_{nk}|V_k, T_k)] + \sum_{k=1}^{K} \mathbb{E}_q[\ln p(V_k|\alpha)] \\
&+ \sum_{r=1}^{R} \mathbb{E}_q[\ln p(\Sigma_k \mathbf{1}_{d_k}(r)|\gamma)] + \mathbb{E}_q[\ln p(\alpha)] + \mathbb{E}_q[\ln p(\gamma)] \\
&- \mathbb{E}_Q[\ln Q_{-T}] - \sum_{k=1}^{K} \varphi_k(r>1)\mathbb{E}_{q(T_k)}[\ln q(T_k)],
\end{aligned} \tag{10}$$

where $\varphi_k(r > 1) := \sum_{r>1} \varphi_k(r)$. We multiply the entropy of T_k by the variational probability that atom ω_k is not in the first round. This keeps the entropy of T_k from blowing up when $\varphi_k(1) \to 1$, since it is the only term that remains involving T_k in this case.

This variational objective function requires two approximations to achieve an analytical form, which we discuss in the next section. We optimize the variational objective using a coordinate ascent algorithm, where each parameter is updated to approximately maximize $\mathcal{L}(\mathcal{D}, \Psi)$ conditioned on the current values of all other parameters. For the variational SB-BP model, variational parameters for $q(d_k)$, $q(z_{nk})$, $q(\alpha)$ and $q(\gamma)$ are updated analytically, while those for $q(V_k)$ and $q(T_k)$ require gradient methods. We give the variational inference algorithm in the appendix.

3.3. A Discussion on Variational Inference for the SB-BP Model

We discuss three terms in the expansion of (10). Two terms require approximations before the variational objective can be put into a tractable form, and the third term relates to the modified Poisson prior on the number of atoms in round r.

3.3.1. The Term $\mathbb{E}_Q[\ln(1 - V_k \mathrm{e}^{-T_k})]$

The first term we discuss is $\mathbb{E}_Q[\ln(1 - V_k \mathrm{e}^{-T_k})]$, which appears in $\mathbb{E}_q[\ln p(z_{nk}|V_k, T_k)]$. As written, this term is intractable. We use a Taylor expansion of the natural logarithm about the point 1,

$$\ln(1 - V_k \mathrm{e}^{-T_k}) = -\sum_{m=1}^{\infty} \frac{1}{m}\left(V_k \mathrm{e}^{-T_k}\right)^m, \tag{11}$$

which converges since $|V_k \exp\{-T_k\}| < 1$. The expectation of this sum becomes the sum of the expectations by monotone convergence, after which each expectation can be calculated analytically. We truncate the summation at a large number, M. For example, we set $M = 1000$ in our experiments. Since $V_k \mathrm{e}^{-T_k} \in (0, 1)$, the error in the approximation decreases rapidly as M increases for values of $V_k \mathrm{e}^{-T_k}$ that are not very close to one, and likewise for their expectations.

3.3.2. The Term $\mathbb{E}_Q[\ln(\{\sum_k \mathbf{1}_{d_k}(r)\}!)]$

The second term of interest is $\mathbb{E}_Q[\ln(\{\sum_k \mathbf{1}_{d_k}(r)\}!)]$, which is the expectation of the log of the denominator of the Poisson distribution with respect to Q. We write this term in the following, more tractable form

$$\begin{aligned} \ln\left(\sum_k \mathbf{1}_{d_k}(r)\right)! &= \sum_{\ell=1}^{\sum_k \mathbf{1}_{d_k}(r)} \ln \ell \qquad (12) \\ &= \sum_{\ell=1}^{\infty} \mathbb{I}\left(\sum_k \mathbf{1}_{d_k}(r) \geq \ell\right) \ln \ell. \end{aligned}$$

Once $\ell > \sum_k \mathbf{1}_{d_k}(r)$, the natural logarithm is multiplied by zero for the remainder of the summation. The expectation of this outer indicator is the probability of the event $\sum_k \mathbf{1}_{d_k}(r) \geq \ell$ with respect to the variational distributions $\varphi_{1:K}$. Since this term is combinatorially intractable, we use Markov's inequality to lower bound the negative of this value (and therefore $\mathcal{L}$),

$$\mathbb{P}_Q\left(\sum_{k=1}^{\infty} \mathbf{1}_{d_k}(r) \geq \ell\right) \leq \frac{1}{\ell^2}\mathbb{E}_Q\left[\left(\sum_{k=1}^{\infty} \mathbf{1}_{d_k}(r)\right)^2\right], \tag{13}$$

and we replace $-\sum_{r,\ell} \mathbb{P}_Q(\sum_k \mathbf{1}_{d_k}(r) \geq \ell) \ln \ell$ with

$$-\sum_{r=1}^{R}\left\{\sum_{k=1}^{K} \varphi_k(r) + \sum_{i \neq j} \varphi_i(r)\varphi_j(r)\right\}\sum_{\ell=1}^{K} \frac{\ln \ell}{\ell^2} \tag{14}$$

in the truncated model. Though the value of $\sum_{\ell=1}^{K} \ell^{-2} \ln \ell$ can be calculated for any K, we use the value of the original infinite summation, which to four significant digits is $\xi := \sum_{\ell=1}^{\infty} \ell^{-2} \ln \ell \approx 0.9375$. Markov's inequality holds for all powers, p, but we select $p = 2$ because $\xi = \infty$ for $p = 1$, and for the computational ease relative to $p > 2$.

The impact of this approximation is to change the penalty for increasing the number of atoms in a round. Previously, the factorial term in the denominator of the Poisson distribution penalized additional atoms with a penalty that grows in the number of atoms, which is easily seen in (12). The derivative of (14) with respect to $\varphi_k(r)$ is $-\xi - \xi \sum_{i \neq k} \varphi_i(r)$. We see that our approximation replaces the increasing per-atom penalty of the Poisson distribution with a constant penalty ξ, and therefore the overall penalty is linear in the number of atoms in a round. This overall penalty is larger than the original penalty for the first four atoms, and smaller afterwards.

3.3.3. The Term $\mathbb{P}_Q(\sum_{r'=r}^{R} \sum_{k=1}^{K} \mathbf{1}_{d_k}(r') > 0)$

Finally, we discuss the term $\mathbb{P}_Q(\sum_{r' \geq r, k} \mathbf{1}_{d_k}(r') > 0)$, which is the expectation of the indicator in the exponent of the Poisson distribution. This is the probability with respect to $\varphi_{1:K}$ that at least one of the K indexed atoms occurs in round r or higher. Under the infinite beta process prior, there will always be an atom that occurs in round r or higher, and so this will always equal one. However, this is not the case for the truncated model, where there are only K atoms.

The probability that $\sum_{r' \geq r, k} \mathbf{1}_{d_k}(r') > 0$ given $\varphi_{1:K}$ is

$$\mathbb{P}_Q\left(\sum_{r'=r}^{R}\sum_{k=1}^{K} \mathbf{1}_{d_k}(r') > 0\right) = 1 - \prod_{k=1}^{K}\sum_{r'=1}^{r-1} \varphi_k(r'). \tag{15}$$

We include this term because is keeps the parameter γ relevant to the truncated model. When not included in the exponential of the Poisson distribution during inference, the truncated model reads all rounds that do not contain atoms as having *zero* atoms, which is a valid draw from the underlying Poisson distribution. In this case, the value of $E_{q(\gamma)}[\gamma]$ always equals K/R, since the model considers there to be a total of K atoms in R rounds. Therefore, removing this term forces γ to be fixed and depend entirely on K and R, which makes the selection of these values critically important. Instead, we wish to say that, if there is an r' for which there are no atoms in round $r' \leq r \leq R$, then those rounds *do not exist*. This term achieves this by down-weighting these rounds. The number of used rounds can still increase, but does so gradually.

4. Truncated Beta Processes

By truncating Q for variational inference at K, we are assuming that at most K atoms appear in the posterior of H. To give a sense of how reasonable this assumption is, we give a truncation bound for the beta process. This bound is a measure of closeness between the truncated and infinite beta processes, and provides information when selecting K and R.

In the spirit of (Doshi-Velez et al., 2009; Ishwaran & James, 2001), we derive a bound on the L_1 distance between the marginal distributions of data that are drawn from a beta process prior, $m_\infty(X)$, and a BP truncated after the Rth round, $m_R(X)$. When truncating K, this bound is valid when K contains the first R rounds, or $K \geq \sum_{r=1}^R C_r$. From Doshi-Velez et al. (2009), we have

$$\frac{1}{4}\int |m_R(X) - m_\infty(X)|\, dX \qquad (16)$$
$$\leq \mathbb{P}\left(\exists k > \Sigma_{r=1}^R C_r \text{ and } 1 \leq n \leq N : z_{nk} = 1\right).$$

The value of $\mathbb{P}$ is the probability that, in N binary vectors sampled from $\mathrm{BeP}(H)$ with $H \sim \mathrm{BP}(\alpha, H_0)$, there exists a $k > \sum_{r=1}^R C_r$ and n for which $z_{nk} = 1$. Intuitively this means that, from the perspective of the data, a truncation of H at round R will be noticed.

Theorem 1 *Let N samples be drawn from* $\mathrm{BeP}(H)$*, where $H \sim \mathrm{BP}(\alpha, H_0)$ is constructed according to (2) and truncated after the first K atoms. Then the bound*

$$\frac{1}{4}\int |m_R(X) - m_\infty(X)|\, dX \leq 1 - \exp\left\{-2\gamma N \left(\frac{\alpha}{1+\alpha}\right)^R\right\}$$

is valid with probability $1 - \frac{\gamma^K}{\Gamma(k)}\int_0^R u^{K-1}\mathrm{e}^{-\gamma u}\, du$.

We sketch a proof of Theorem 1 in the appendix. The integral is of a $\mathrm{Gamma}(K, \gamma)$ distribution, and is the probability that the truncation $K \geq \sum_{r=1}^R C_r$.

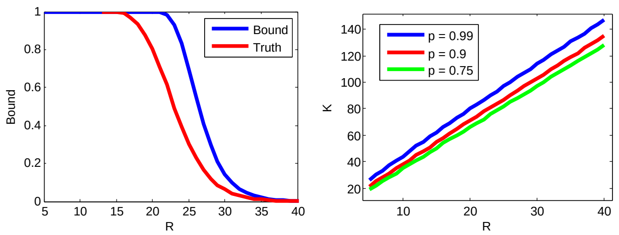

Figure 2. Let $N = 5000$, $\gamma = 3$ and $\alpha = 2$. (left) The bound as a function of R. (right) R vs. K for different probability thresholds, p, on the bound being valid.[2]

Figure 2 contains an example of this bound, where we use 5000 samples of H for each value of R to approximate the ground truth.

Though Theorem 1 can help in selecting K and R, we note that the variational inference algorithm here effectively learns R. We recall the discussion in Section 3.3.3, where a penalty is given to rounds that are far away from those currently occupied by the atoms. Even when R is set to a large number, this penalty adds resistance to exploring higher indices, which must be overcome by the data.

A key difference between the truncated BP and the truncated IBP presented in (Doshi-Velez et al., 2009) is that the prior in Section 2.2 does not require a strict ordering of the atom weights. Recall that for the stick-breaking construction of the IBP, the weight given to the kth atom is $\prod_{i=1}^k V_i$, where $V_i \overset{iid}{\sim} \mathrm{Beta}(\alpha, 1)$ (Teh et al., 2007). In contrast, in Section 2.2 the strict ordering of $d_1, d_2, \ldots$ was relaxed, which allows atoms with high and low probability to take any index value without being penalized by the prior. Also, through the variational posteriors $\varphi_{1:K}$, we don't enforce a hard assignment, but learn distributions on these values.

5. Experiments

We evaluate performance of the variational SB-BP prior in the matrix factorization setting. We consider a $V \times N$ data matrix $D \sim f(\Lambda)$, where $f(\cdot)$ is some distribution, Λ is a latent matrix and the above notation indicates that $d_{vn} \sim f(\lambda_{vn})$. We model $\Lambda = \Phi(W \circ Z)$, where $\circ$ indicates element-wise multiplication and Z has a beta-Bernoulli process prior. Therefore, though the initial rank of the factorization may be large, the posterior will place a high probability on a low rank representation. We consider two cases: (i) non-negative matrix factorization, in which Φ and W contain *iid* gamma

[2] As another example, let $N = 1000$, $\gamma = 2$ and $\alpha = 3$. If $K = 180$ and $R = 75$, then with probability greater than 0.99 the upper bound is $\approx 1.7 \times 10^{-6}$.

Table 1. The average number of factors used per-document and corpus-wide for *The New York Times* and *Science*. These results are for $\beta = 0.5$.

	New York Times		Science	
	per-doc	corpus	per-doc	corpus
SB-BP	12.9	68.4	14.7	125.4
VB-IBP	15.4	94.8	20.0	164

random variables and $f(\cdot)$ is a Poisson distribution; and (ii) a linear-Gaussian model, in which Φ and W contain *iid* Gaussian random variables and $f(\cdot)$ is a normal distribution. We set $K = 200$ and $R = 50$ in both models. We place vague gamma priors on α and γ. We terminate each algorithm when the fractional change in the lower bound falls below 10^{-3}.

5.1. A Non-negative Factorization Model

We first evaluate performance on a non-negative matrix factorization problem. We consider a word count matrix D, where d_{vn} is the number of times word v appears in document n. We perform experiments on *The New York Times* and *Science* corpora. In each case, we partition the data into five training/testing sets, with 5000 training documents and 3000 testing documents.

For testing, we randomly partition the words in each document into two groups, d_{T_1} and d_{T_2}. We then learn document-specific parameters for the first half of each test document, $d_{T_1} \sim f(\lambda_{T_1})$, where $\lambda_{T_1} = \Phi(w_{T_1} \circ z_{T_1})$. The q distributions from training are used for Φ and as the prior for z_{T_1}. We measure performance using held-out perplexity on the second half of each document. We normalize λ_{T_1} to calculate the word distribution for each test document.[2]

We compare the SB-BP prior with three other models: (i) the VB-IBP prior (Doshi-Velez et al., 2009); (ii) the model with the beta-Bernoulli process removed, called VB-NMF; and (iii) the NMF algorithm of Lee & Seung (2001). The first two models use variational inference. The last model is the maximum-likelihood version of (ii). Models (ii) and (iii) require a pre-set factorization rank; we perform experiments on ranks $K = 10, 25, 50, 75, 100$. The Bayesian models have Gamma$(\beta, \beta V)$ priors on the values in Φ, where V is vocabulary size. We place Gamma$(.25, .25)$ priors on the elements of W. The maximum-likelihood NMF has no parameters (other than the factorization rank). For the SB-BP, we take five steps for each gradient update.

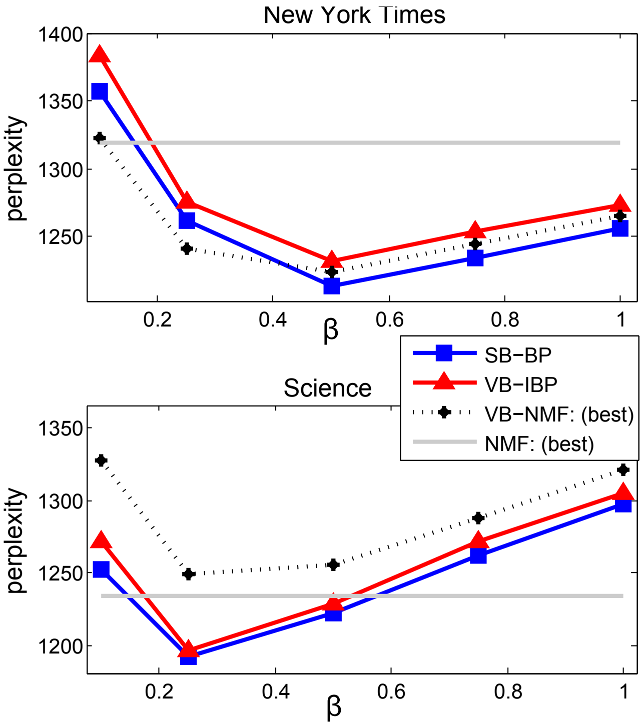

Figure 3. Perplexity results for *The New York Times* and *Science*. The best factorization rank for *The New York Times* was VB-NMF: $K = 50$, NMF: $K = 25$ and for *Science* was VB-NMF: $K = 75$, NMF: $K = 75$.

We show results in Figure 3. For algorithms (ii) and (iii) we show results for the best factorization ranks. Among the variational algorithms, the SB-BP prior performs the best for almost all values of β. The decrease in performance of the VB-IBP is likely due to the strict ordering of the probabilities in the prior. Because gradient methods are used, SB-BP is slower than VB-IBP, taking 30-45 seconds per-iteration to update the variational posterior of the beta process. Figure 4 contains the variational posteriors of the round indicators for one run of *The New York Times* and *Science*. Though we truncate the rounds at 50, the variational distributions φ_k do not explore beyond the 35th round—a result of the penalty discussed in Section 3.3.3. In Table 1 we show latent factor statistics for both corpora. We observe similar results on data from *Huffington Post* and *Wikipedia* (not plotted for space).

5.2. A Linear-Gaussian Model

We also studied the performance of the SB-BP using the linear-Gaussian model. We consider the HGDP-CEPH human genome diversity cell line panel (Conrad et al., 2006), which is a collection of DNA samples from subjects around the world. After whitening, the data matrix is $D \in \mathbb{R}^{377 \times 1056}$. We place $N(0, 1)$ priors on W and $N(0, 10^6)$ (i.e. vague) priors on Φ, and we set the

[2] We recall that, when generating data (e.g. word counts) from a collection of Poisson distributions, the distribution on the specific words *conditioned on* the total number of words is the normalized Poisson parameters.

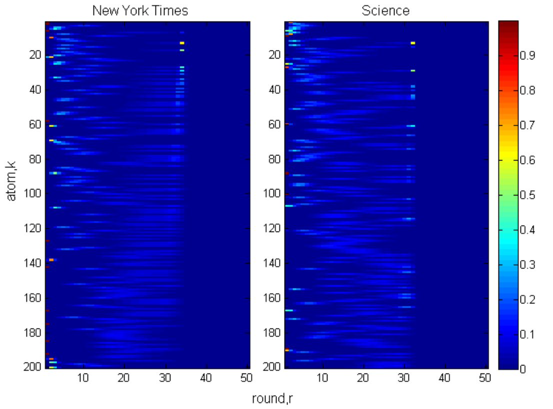

Figure 4. Variational distributions $\varphi_k(r)$ on the latent round indicators for (left) *The New York Times* and (right) *Science*. Not all rounds are used, and atoms of varying probability can take any index value.

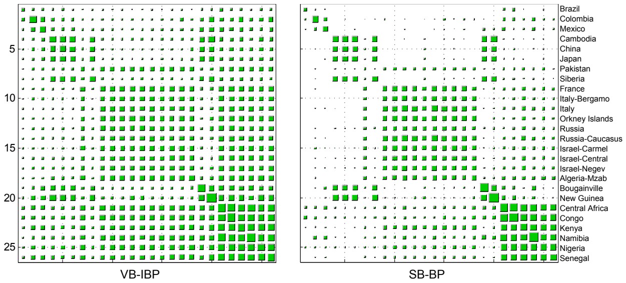

Figure 5. The expected number of shared factors between two samples for (left) VB-IBP and (right) SB-BP. Sharing among regions appears less ambiguous for SB-BP.

noise variance equal to one. For the linear-Gaussian model, we use indicator variational distributions on z_{nk}, which we discuss in the appendix.

We ran the model using the SB-BP and VB-IBP priors, each with the same initialization. We sorted the initialization so that highly probable factors have high index values. We found that SB-BP used fewer total ones (i.e. $\Sigma_{nk} z_{nk}$) than VB-IBP, and had slightly worse reconstruction error in that it learned a larger noise variance. However, the sharing of factors by people of similar geographic regions was less ambiguous for the SB-BP, as shown in Figure 5, indicating that a clearer underlying gene structure may have been found. The number of factors having at least a single one was approximately 130 for each model, so the additional ones for VB-IBP occurred within a set of factors the same size as used by the SB-BP.

6. Conclusion

We have derived a variational inference algorithm for stick-breaking beta process priors using a simpler representation of the construction given in (Paisley et al., 2010). We derived a bound relating truncated and non-truncated BPs to aid in selecting truncation levels. We demonstrated competitive performance on two matrix factorization paradigms.

Acknowledgements David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520, AFOSR 09NL202, the Alfred P. Sloan foundation, and a grant from Google.

References

Conrad, DF, Jakobsson, M, Coop, G, Wen, X, Wall, JD, Rosenberg, NA, and Pritchard, JK. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, 38:1251–1260, 2006.

Doshi-Velez, F., Miller, K.T., Van Gael, J., and Teh, Y.W. Variational inference for the Indian buffet process. In *AISTATS*, 2009.

Ghahramani, Z., Griffiths, T.L., and Sollich, P. Bayesian nonparametric latent feature models. *Bayesian Statistics*, 2007.

Griffiths, T.L. and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2006.

Hjort, N.L. Nonparametric bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294, 1990.

Ishwaran, H. and James, L.F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Jordan, M.I., Ghahramani, Z., Jaakkola, T., and Saul, L.K. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Lee, D.D. and Seung, H.S. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

Paisley, J., Zaas, A., Ginsburg, G., Woods, C., and Carin, L. A stick-breaking construction of the beta process. In *ICML*, 2010.

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

Teh, Y., Gorur, D., and Ghahramani, Z. Stick-breaking construction for the Indian buffet process. In *AISTATS*, 2007.

Thibaux, R. and Jordan, M.I. Hierarchical beta processes and the Indian buffet process. In *AISTATS*, 2007.

Williamson, C., Wang, C., Heller, K., and Blei, D. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.

Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., and Carin, L. Non-parametric Bayesian dictionary learning for sparse image representations. In *NIPS*, 2009.

Appendix

A1. The Variational Inference Algorithm

We present the coordinate ascent algorithm for finding a local maximum of the lower bound of (10). We first define each $\frac{1}{m}\mathbb{E}[(V_k \mathrm{e}^{-T_k})^m]$, which are the expectations of the terms in (11), using the notation

$$\Delta_k(m) := \frac{1}{m}\frac{\Gamma(a'_k+b'_k)}{\Gamma(a'_k+b'_k+m)}\frac{\Gamma(a'_k+m)}{\Gamma(a'_k)}\left(\frac{v'_k}{v'_k+m}\right)^{u'_k}.$$

The algorithm below also uses the derivatives of $\Delta_k(m)$ with respect to a'_k, b'_k, u'_k and v'_k. We also let $\Delta_k(\cdot) := \sum_{m=1}^{M}\Delta_k(m)$, and $n_{1k} := \sum_{n=1}^{N}\phi_{nk}(1)$, $n_{0k} := N - n_{1k}$ and $\xi := 0.9375$. All expectations involving a specific variable are given in the variational update for that variable. Coordinate ascent variational inference cycles through the following updates.

Coordinate update for $q(z_{nk})$. The variational update for ϕ_{nk} is model-specific. We focus on the BP prior. Let the term corresponding to the likelihood be $f_i = \exp\{\partial\mathbb{E}[\ln p(\mathcal{D}, \Theta_H|Z, \Theta_{-H})]/\partial\phi_{nk}(i)\}$. Then

$$\begin{aligned}\phi_{nk}(1) &\propto f_1\exp\left\{\mathbb{E}[\ln V_k] - \varphi_k(r>1)\mathbb{E}[T_k]\right\}, \qquad (17)\\ \phi_{nk}(0) &\propto f_0\exp\left\{\varphi_k(1)\mathbb{E}[\ln(1-V_k)] - \varphi_k(r>1)\Delta_k(\cdot)\right\}.\end{aligned}$$

For an indicator variational distribution, i.e. when $q(z_{nk}) = \mathbf{1}_{z_{nk}}(1)$, we can use the above equations and set $\mathbf{1}_{z_{nk}}(1) = \mathbb{I}(\phi_{nk}(1) > \phi_{nk}(0))$.

Coordinate update for $q(d_k)$. The update for each φ_k is given below for $r = 1, \ldots, R$. For $r \geq 2$, let

$$\rho(r) := (r-1)\mathbb{E}[\ln\alpha] - \ln\Gamma(r-1) + (r-2)\mathbb{E}[\ln T_k],$$

then

$$\varphi_k(1) \propto \exp\left\{n_{0k}\mathbb{E}[\ln(1-V_k)] - \xi\sum_{i\neq k}\varphi_i(1)\right\} \qquad (18)$$

$$\begin{aligned}\varphi_k(r) \propto \exp\Big\{&-(n_{1k}+\mathbb{E}[\alpha])\mathbb{E}[T_k] - n_{0k}\Delta_k(\cdot) + \mathbb{H}[q(T_k)]\\ &+\rho(r) - \xi\sum_{i\neq k}\varphi_i(r) - \mathbb{E}[\gamma]\sum_{j=2}^{r}\prod_{k'\neq k}\sum_{r'=1}^{j-1}\varphi_{k'}(r')\Big\}\end{aligned}$$

We note that the last term in $\varphi_k(r)$ is the penalty for extending into rounds that are unused.

Coordinate update for $q(V_k)$. We use gradient ascent to jointly update (a'_k, b'_k). Let $\lambda_1 := n_{1k} + 1 - a'_k$, $\lambda_2 := n_{0k}\varphi_k(1) + \alpha - b'_k$ and $\lambda_3 := n_{0k}\varphi_k(r > 1)$. The derivatives are

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial a'_k} &= \lambda_1\mathbb{E}[\ln V_k] - \lambda_2\psi'(a'_k+b'_k) - \lambda_3\frac{\partial\Delta_k(\cdot)}{\partial a'_k} \qquad (19)\\ \frac{\partial\mathcal{L}}{\partial b'_k} &= -\lambda_1\psi'(a'_k+b'_k) + \lambda_2\mathbb{E}[\ln(1-V_k)] - \lambda_3\frac{\partial\Delta_k(\cdot)}{\partial b'_k}\end{aligned}$$

with $\mathbb{E}[\ln V_k] = \psi(a'_k) - \psi(a'_k+b'_k)$ and $\mathbb{E}[\ln(1-V_k)] = \psi(b'_k) - \psi(a'_k+b'_k)$.

Coordinate update for $q(T_k)$. We use gradient ascent to jointly update (u'_k, v'_k). The derivatives are

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial u'_k} &= \psi'(u'_k)\sum_{r\geq 2}(r-2)\varphi_k(r) + \varphi_k(r>1) \times \qquad (20)\\ &\left(1 - \frac{n_{1k}+\mathbb{E}[\alpha]}{v'_k} - n_{0k}\frac{\partial\Delta_k(\cdot)}{\partial u'_k} + (1-u'_k)\psi'(u'_k)\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial v'_k} &= -\frac{1}{v'_k}\sum_{r\geq 2}(r-2)\varphi_k(r) + \varphi_k(r>1) \times \qquad (21)\\ &\left(\frac{u'_k}{v'^2_k}(n_{1k}+\mathbb{E}[\alpha]) - n_{0k}\frac{\partial\Delta_k(\cdot)}{\partial v'_k} - \frac{1}{v'_k}\right)\end{aligned}$$

with $\mathbb{E}[T_k] = u'_k/v'_k$ and $\mathbb{E}[\ln T_k] = \psi(u'_k) - \ln v'_k$.

Coordinate update for $q(\alpha)$. The variational parameter updates for $q(\alpha)$ are analytical and are

$$\begin{aligned}\kappa_1 &= K + \sum_{k=1}^{K}\sum_{r\geq 2}(r-1)\varphi_k(r) + a_1 \qquad (22)\\ \kappa_2 &= -\sum_{k=1}^{K}\mathbb{E}[\ln(1-V_k)] + \sum_{k=1}^{K}\mathbb{E}[T_k]\varphi_k(r>1) + a_2\end{aligned}$$

with $\mathbb{E}[\alpha] = \kappa_1/\kappa_2$ and $\mathbb{E}[\ln\alpha] = \psi(\kappa_1) - \ln\kappa_2$.

Coordinate update for $q(\gamma)$. The variational parameter updates for $q(\gamma)$ are analytical and are

$$\begin{aligned}\tau_1 &= K + b_1 \qquad (23)\\ \tau_2 &= \sum_{r=1}^{R}\left\{1 - \prod_{k=1}^{K}\sum_{r'=1}^{r-1}\varphi_k(r')\right\} + b_2\end{aligned}$$

with $\mathbb{E}[\gamma] = \tau_1/\tau_2$ and $\mathbb{E}[\ln\gamma] = \psi(\tau_1) - \ln\tau_2$.

A2. Proof of Theorem 1 (sketch)

Let $\pi_{rj} := V_{rj}^{(r)}\prod_{l<r}(1 - V_{rj}^{(l)})$. An upper bound on (16) can be derived by bounding $\mathbb{P}(\cdot)$ as follows:

$$\begin{aligned}\mathbb{P}(\cdot) &= 1 - \mathbb{E}\left[\mathbb{E}\left[\left(\prod_{r=R+1}^{\infty}\prod_{j=1}^{C_r}(1-\pi_{rj})\right)^N \mid C_r\right]\right] \qquad (24)\\ &\leq 1 - \exp\left\{N\sum_{r=R+1}^{\infty}\mathbb{E}\left[\sum_{j=1}^{C_r}\mathbb{E}\left[\ln(1-\pi_{rj})\right]\right]\right\} \qquad (25)\\ &\leq 1 - \exp\left\{-\frac{\gamma N}{1+\alpha}\sum_{m=1}^{\infty}\frac{1}{m}\sum_{r=R}^{\infty}\left(\frac{\alpha}{m+\alpha}\right)^r\right\} \qquad (26)\end{aligned}$$

We use Jensen's inequality to go from (24) to (25). In (26), we use a Taylor expansion as in (11) and a bound on $\mathbb{E}[V^m]$. To go from (26) to Theorem 1, we refer to Appendix F in the technical report associated with (Doshi-Velez et al., 2009). The corresponding probability follows from constructing a Poisson process on $\mathbb{R}_+$ with rate γ and stopping time R.