# A Stick-Breaking Construction of the Beta Process

**John Paisley**[1] JWP4@EE.DUKE.EDU
**Aimee Zaas**[2] AIMEE.ZAAS@DUKE.EDU
**Christopher W. Woods**[2] WOODS004@MC.DUKE.EDU
**Geoffrey S. Ginsburg**[2] GINSB005@DUKE.EDU
**Lawrence Carin**[1] LCARIN@EE.DUKE.EDU
[1]Department of ECE, [2]Duke University Medical Center, Duke University, Durham, NC

## Abstract

We present and derive a new stick-breaking construction of the beta process. The construction is closely related to a special case of the stick-breaking construction of the Dirichlet process (Sethuraman, 1994) applied to the beta distribution. We derive an inference procedure that relies on Monte Carlo integration to reduce the number of parameters to be inferred, and present results on synthetic data, the MNIST handwritten digits data set and a time-evolving gene expression data set.

## 1. Introduction

The Dirichlet process (Ferguson, 1973) is a powerful Bayesian nonparametric prior for mixture models. There are two principle methods for drawing from this infinite-dimensional prior: ($i$) the Chinese restaurant process (Blackwell & MacQueen, 1973), in which samples are drawn from a marginalized Dirichlet process and implicitly construct the prior; and ($ii$) the stick-breaking process (Sethuraman, 1994), which is a fully Bayesian construction of the Dirichlet process.

Similarly, the beta process (Hjort, 1990) is receiving significant use recently as a nonparametric prior for latent factor models (Ghahramani et al., 2007; Thibaux & Jordan, 2007). This infinite-dimensional prior can be drawn via marginalization using the Indian buffet process (Griffiths & Ghahramani, 2005), where samples again construct the prior. However, unlike the Dirichlet process, the fully Bayesian stick-breaking construction of the beta process has yet to be derived (though related methods exist, reviewed in Section 2).

To review, a Dirichlet process, $G$, can be constructed according to the following stick-breaking process (Sethuraman, 1994; Ishwaran & James, 2001),

$$\begin{aligned} G &= \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1-V_j)\delta_{\theta_i} \\ V_i &\overset{iid}{\sim} \text{Beta}(1,\alpha) \\ \theta_i &\overset{iid}{\sim} G_0 \end{aligned} \tag{1}$$

This stick-breaking process is so-called because proportions, $V_i$, are sequentially broken from the remaining length, $\prod_{j=1}^{i-1}(1-V_j)$, of a unit-length stick. This produces a probability (or weight), $V_i\prod_{j=1}^{i-1}(1-V_j)$, that can be visually represented as one of an infinite number of contiguous sections cut out of a unit-length stick. As $i$ increases, these weights stochastically decrease, since smaller and smaller fractions of the stick remain, and so only a small number of the infinite number of weights have appreciable value. By construction, these weights occur first, which allows for practical implementation of this prior.

The contribution of this paper is the derivation of a stick-breaking construction of the beta process. We use a little-known property of the constructive definition in (Sethuraman, 1994), which is equally applicable to the beta distribution – a two-dimensional Dirichlet distribution. The construction presented here will be seen to result from an infinite collection of these stick-breaking constructions of the beta distribution.

The paper is organized as follows. In Section 2, we review the beta process, the stick-breaking construction of the beta distribution, as well as related work in this area. In Section 3, we present the stick-breaking construction of the beta process and its derivation. We derive an inference procedure for the construction in Section 4 and present experimental results on synthetic, MNIST digits and gene expression data in Section 5.

Appearing in *Proceedings of the 26$^{th}$ International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

## 2. The Beta Process

Let $H_0$ be a continuous measure on the space $(\Theta, \mathcal{B})$ and let $H_0(\Theta) = \gamma$. Also, let $\alpha$ be a positive scalar and define the process $H_K$ as follows,

$$\begin{aligned} H_K &= \sum_{k=1}^{K} \pi_k \delta_{\theta_k} \\ \pi_k &\overset{iid}{\sim} \text{Beta}\left(\frac{\alpha\gamma}{K}, \alpha(1-\frac{\gamma}{K})\right) \\ \theta_k &\overset{iid}{\sim} \frac{1}{\gamma} H_0 \end{aligned} \qquad (2)$$

then as $K \to \infty$, $H_K \to H$ and $H$ is a beta process, which we denote $H \sim \text{BP}(\alpha H_0)$.

We avoid a complete measure-theoretic definition, since the stick-breaking construction to be presented is derived in reference to the limit of (2). That $H$ is a beta process can be shown in the following way: Integrating out $\boldsymbol{\pi}^{(K)} = (\pi_1, \ldots, \pi_K)^T \in (0,1)^K$, letting $K \to \infty$ and sampling from this marginal distribution produces the two-parameter extension of the Indian buffet process discussed in (Thibaux & Jordan, 2007), which is shown to have the beta process as its underlying de Finetti mixing distribution.

Before deriving the stick-breaking construction of the beta process, we review a property of the beta *distribution* that will be central to the construction. We also review related work to distinguish the presented construction from other constructions in the literature.

### 2.1. A Construction of the Beta Distribution

The constructive definition of a Dirichlet prior derived in (Sethuraman, 1994) applies to more than the infinite-dimensional Dirichlet process. In fact, it is applicable to Dirichlet priors of any dimension, of which the beta distribution can be viewed as a special, two-dimensional case.[1] Focusing on this special case, Sethuraman showed that one can sample

$$\pi \sim \text{Beta}(a, b) \qquad (3)$$

according to the following stick-breaking construction,

$$\begin{aligned} \pi &= \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1} (1 - V_j) \mathbb{I}(Y_i = 1) \\ V_i &\overset{iid}{\sim} \text{Beta}(1, a+b) \\ Y_i &\overset{iid}{\sim} \text{Bernoulli}\left(\frac{a}{a+b}\right) \end{aligned} \qquad (4)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

In this construction, weights are drawn according to the standard stick-breaking construction of the DP (Ishwaran & James, 2001), as well as their respective locations, which are independent of the weights and iid among themselves. The major difference is that the set of locations is finite, 0 or 1, which results in more than one term being active in the summation.

Space restrictions prohibit an explicit proof of this construction here, but we note that Sethuraman implicitly proves this in the following way: Using notation from (Sethuraman, 1994), let the space, $\mathcal{X} = \{0, 1\}$, and the prior measure, $\alpha$, be $\alpha(1) = a$, $\alpha(0) = b$, and therefore $\alpha(\mathcal{X}) = a + b$. Carrying out the proof in (Sethuraman, 1994) for this particular space and measure yields (4). We note that this $\alpha$ is different from that in (2).

### 2.2. Related Work

There are three related constructions in the machine learning literature, each of which differs significantly from that presented here. The first construction, proposed by (Teh et al., 2007), is presented specifically for the Indian buffet process (IBP) prior. The generative process from which the IBP and this construction are derived replaces the beta distribution in (2) with Beta$(\frac{\alpha}{K}, 1)$. This small change greatly facilitates this construction, since the parameter 1 in Beta$(\frac{\alpha}{K}, 1)$ allows for a necessary simplification of the beta distribution. This construction does not extend to the two-parameter generalization of the IBP (Ghahramani et al., 2007), which is equivalent in the infinite limit to the marginalized representation in (2).

A second method for drawing directly from the beta process prior has been presented in (Thibaux & Jordan, 2007), and more recently in (Teh & Görür, 2009) as a special case of a more general power-law representation of the IBP. In this representation, no stick-breaking takes place of the form in (1), but rather the weight for each location is simply beta-distributed, as opposed to the usual function of multiple beta-distributed random variables. The derivation relies heavily upon connecting the marginalized process to the fully Bayesian representation, which does not factor into the similar derivation for the DP (Sethuraman, 1994). This of course does not detract from the result, which appears to have a simpler inference procedure than that presented here.

A third representation presented in (Teh & Görür, 2009) based on the inverse Lévy method (Wolpert & Ickstadt, 1998) exists in theory only and does not simplify to an analytic stick-breaking form. See (Damien et al., 1996; Lee & Kim, 2004) for two approximate methods for sampling from the beta process.

[1] We thank Jayaram Sethuraman for his valuable correspondence regarding his constructive definition.

## 3. A Stick-Breaking Construction of the Beta Process

We now define and briefly discuss the stick-breaking construction of the beta process, followed by its derivation. Let $\alpha$ and $H_0$ be defined as in (2). If $H$ is constructed according to the following process,

$$\begin{aligned} H &= \sum_{i=1}^{\infty}\sum_{j=1}^{C_i} V_{ij}^{(i)} \prod_{\ell=1}^{i-1}(1-V_{ij}^{(\ell)})\delta_{\theta_{ij}} \\ C_i &\overset{iid}{\sim} \text{Poisson}(\gamma) \\ V_{ij}^{(\ell)} &\overset{iid}{\sim} \text{Beta}(1,\alpha) \\ \theta_{ij} &\overset{iid}{\sim} \frac{1}{\gamma}H_0 \end{aligned} \tag{5}$$

then $H \sim \text{BP}(\alpha H_0)$.

Since the first row of (5) may be unclear at first sight, we expand it for the first few values of $i$ below,

$$\begin{aligned} H = & \sum_{j=1}^{C_1} V_{1,j}^{(1)}\delta_{\theta_{1,j}} + \\ & \sum_{j=1}^{C_2} V_{2,j}^{(2)}(1-V_{2,j}^{(1)})\delta_{\theta_{2,j}} + \\ & \sum_{j=1}^{C_3} V_{3,j}^{(3)}(1-V_{3,j}^{(2)})(1-V_{3,j}^{(1)})\delta_{\theta_{3,j}} + \cdots \end{aligned} \tag{6}$$

For each value of $i$, which we refer to as a "round," there are $C_i$ atoms, where $C_i$ is itself random and drawn from Poisson($\gamma$). Therefore, every atom is defined by two subscripts, $(i,j)$. The mass associated with each atom in round $i$ is equal to the $i^{\text{th}}$ break from an *atom-specific* stick, where the stick-breaking weights follow a Beta$(1,\alpha)$ stick-breaking process (as in (1)). Superscripts are used to index the $i$ random variables that construct the weight on atom $\theta_{ij}$. Since the number of breaks from the unit-length stick prior to obtaining a weight increases with each level in (6), the weights stochastically decrease as $i$ increases, in a similar manner as in the stick-breaking construction of the Dirichlet process (1).

Since the expectation of the mass on the $k^{\text{th}}$ atom drawn overall does not simplify to a compact and transparent form, we omit its presentation here. However, we note the following relationship between $\alpha$ and $\gamma$ in the construction. As $\alpha$ decreases, weights decay more rapidly as $i$ increases, since smaller fractions of each unit-length stick remains prior to obtaining a weight. As $\alpha$ increases, the weights decay more gradually over several rounds. The expected weight on an atom in round $i$ is equal to $\alpha^{(i-1)}/(1+\alpha)^i$. The number of atoms in each round is controlled by $\gamma$.

### 3.1. Derivation of the Construction

Starting with (2), we now show how Sethuraman's constructive definition of the beta distribution can be used to derive that the infinite limit of (2) has (5) as an alternate representation that is equal in distribution. We begin by observing that, according to (4), each $\pi_k$ value can be drawn as follows,

$$\begin{aligned} \pi_k &= \sum_{l=1}^{\infty} \hat{V}_{kl} \prod_{m=1}^{l-1}(1-\hat{V}_{km})\mathbb{I}(\hat{Y}_{kl}=1) \\ \hat{V}_{kl} &\overset{iid}{\sim} \text{Beta}(1,\alpha) \\ \hat{Y}_{kl} &\overset{iid}{\sim} \text{Bernoulli}\left(\frac{\gamma}{K}\right) \end{aligned} \tag{7}$$

where the marker ^ is introduced because $V$ will later be re-indexed values of $\hat{V}$. We also make the observation that, if the sum is instead taken to $K'$, and we then let $K' \to \infty$, then this truncated representation converges to (7).

This suggests the following procedure for constructing the limit of the vector $\boldsymbol{\pi}^{(K)}$ in (2). We define the matrices $\hat{\boldsymbol{V}} \in (0,1)^{K\times K}$ and $\hat{\boldsymbol{Y}} \in \{0,1\}^{K\times K}$, where

$$\begin{aligned} \hat{V}_{kl} &\overset{iid}{\sim} \text{Beta}(1,\alpha) \\ \hat{Y}_{kl} &\overset{iid}{\sim} \text{Bernoulli}\left(\frac{\gamma}{K}\right) \end{aligned} \tag{8}$$

for $k = 1,\ldots,K$ and $l = 1,\ldots,K$. The $K$-truncated weight, $\pi_k$, is then constructed "horizontally" by looking at the $k^{\text{th}}$ row of $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{Y}}$, and where we define that the error of the truncation is assigned to $1-\pi_k$ (i.e., $Y_{k,l'} := 0$ for the extension $l' > K$.)

It can be seen from the matrix definitions in (8) and the underlying function of these two matrices, defined for each row as a $K$-truncated version of (7), that in the limit as $K \to \infty$, this representation converges to the infinite beta process when viewed vertically, and to a construction of the individual beta-distributed random variables when viewed horizontally, each of which occur simultaneously.

Before using these two matrices to derive (5), we derive a probability that will be used in the infinite limit. For a given column, $i$, of (8), we calculate the probability that, for a particular row, $k$, there is at least one $\hat{Y} = 1$ in the set $\{\hat{Y}_{k,1},\ldots,\hat{Y}_{k,i-1}\}$, in other words, the probability that $\sum_{i'=1}^{i-1}\hat{Y}_{ki'} > 0$. This value is

$$\mathbb{P}\left(\sum_{i'=1}^{i-1}\hat{Y}_{ki'} > 0|\gamma, K\right) = 1-(1-\frac{\gamma}{K})^{i-1} \tag{9}$$

In the limit as $K \to \infty$, this can be shown to converge to zero for all fixed values of $i$.

As with the Dirichlet process, the problem with drawing each $\pi_k$ explicitly in the limit of (2) is that there are an infinite number of them, and any given $\pi_k$ is equal to zero with probability one. With the representation in (8), this problem appears to have doubled, since there are now an infinite number of random variables to sample in two dimensions, rather than one. However, this is only true when viewed *horizontally.* When viewed vertically, drawing the values of interest becomes manageable.

First, we observe that, in (8), we only care about the set of indices $\{(k,l) : \hat{Y}_{kl} = 1\}$, since these are the locations which indicate that mass is to be added to their respective $\pi_k$ values. Therefore, we seek to bypass the drawing of all indices for which $\hat{Y} = 0$, and directly draw those indices for which $\hat{Y} = 1$.

To do this, we use a property of the binomial distribution. For any column, $i$, of $\hat{\boldsymbol{Y}}$, the number of nonzero locations, $\sum_{k=1}^{K} \hat{Y}_{ki}$, has the Binomial$(K, \frac{\gamma}{K})$ distribution. Also, it is well-known that

$$\text{Poisson}(\gamma) = \lim_{K\to\infty} \text{Binomial}\left(K, \frac{\gamma}{K}\right) \tag{10}$$

Therefore, in the limit as $K \to \infty$, the sum of each column (as well as row) of $\hat{\boldsymbol{Y}}$ produces a random variable with a Poisson$(\gamma)$ distribution. This suggests the procedure of first drawing the number of nonzero locations for each column, followed by their corresponding indices.

Returning to (8), given the number of nonzero locations in column $i$, $\sum_{k=1}^{K} \hat{Y}_{ki} \sim \text{Binomial}(K, \frac{\gamma}{K})$, finding the indices of these locations then becomes a process of sampling uniformly from $\{1, \ldots, K\}$ without replacement. Moreover, since there is a one-to-one correspondence between these indices and the atoms, $\theta_1, \ldots, \theta_K \overset{iid}{\sim} \frac{1}{\gamma} H_0$, which they index, this is equivalent to selecting from the set of atoms, $\{\theta_1, \ldots, \theta_K\}$, uniformly without replacement.

A third more conceptual process, which will aid the derivation, is as follows: Sample the $\sum_{k=1}^{K} \hat{Y}_{ki}$ nonzero indices for column $i$ one at a time. After an index, $k'$, is obtained, check $\{\hat{Y}_{k',1}, \ldots, \hat{Y}_{k',i-1}\}$ to see whether this index has already been drawn. If it has, add the corresponding mass, $V_{k'i} \prod_{l=1}^{i-1}(1 - V_{k'l})$, to the tally for $\pi_{k'}$. If it has not, draw a new atom, $\theta_{k'} \sim \frac{1}{\gamma} H_0$, and associate the mass with this atom.

The derivation concludes by observing the behavior of this last process as $K \to \infty$. We first reiterate that, in the limit as $K \to \infty$, the count of nonzero locations for each column is independent and identically distributed as Poisson$(\gamma)$. Therefore, for $i = 1, 2, \ldots$, we can draw these numbers, $C_i := \sum_{k=1}^{\infty} \hat{Y}_{ki}$, as

$$C_i \overset{iid}{\sim} \text{Poisson}(\gamma) \tag{11}$$

We next need to sample index values uniformly from the positive integers, $\mathbb{N}$. However, we recall from (9) that for all fixed values of $i$, the probability that the drawn index will have previously seen a one is equal to zero. Therefore, using the conceptual process defined above, we can bypass sampling the index value and directly sample the atom which it indexes. Also, we note that the "without replacement" constraint no longer factors.

The final step is simply a matter of re-indexing. Let the function $\sigma_i(j)$ map the input $j \in \{1, \ldots, C_i\}$ to the index of the $j^{\text{th}}$ nonzero element drawn in column $i$, as discussed above. Then the re-indexed random variables $V_{ij}^{(i)} := \hat{V}_{\sigma_i(j),i}$ and $V_{ij}^{(\ell)} := \hat{V}_{\sigma_i(j),\ell}$, where $\ell < i$. We similarly re-index $\theta_{\sigma_i(j)}$ as $\theta_{ij} := \theta_{\sigma_i(j)}$, letting the double and single subscripts remove ambiguity, and hence no $\hat{}$ marker is used. The addition of a subscript/superscript in the two cases above arises from ordering the nonzero locations for each column of (8), i.e., the original index values for the selected rows of each column are being mapped to $1, 2, \ldots$ separately for each column in a many-to-one manner. The result of this re-indexing is the process given in (5).

## 4. Inference for the Stick-Breaking Construction

For inference, we integrate out all stick-breaking random variables, $V$, using Monte Carlo integration (Gamerman & Lopes, 2006), which significantly reduces the number of random variables to be learned. As a second aid for inference, we introduce the latent round-indicator variable,

$$d_k := 1 + \sum_{i=1}^{\infty} \mathbb{I}\left(\sum_{j=1}^{i} C_j < k\right) \tag{12}$$

The equality $d_k = i$ indicates that the $k^{\text{th}}$ atom drawn overall occurred in round $i$. Note that, given $\{d_k\}_{k=1}^{\infty}$, we can reconstruct $\{C_i\}_{i=1}^{\infty}$. Given these latent indicators, the generative process is rewritten as,

$$\begin{aligned} H \mid \{d_k\}_{k=1}^{\infty} &= \sum_{k=1}^{\infty} V_{k,d_k} \prod_{j=1}^{d_k - 1}(1 - V_{kj})\delta_{\theta_k} \\ V_{kj} &\overset{iid}{\sim} \text{Beta}(1, \alpha) \\ \theta_k &\overset{iid}{\sim} \frac{1}{\gamma} H_0 \end{aligned} \tag{13}$$

where, for clarity in what follows, we've avoided introducing a third marker (e.g., $\tilde{V}$) after this re-indexing.

Data is generated iid from $H$ via a Bernoulli process and take the form of infinite-dimensional binary vectors, $z_n \in \{0,1\}^\infty$, where

$$z_{nk} \sim \text{Bernoulli}\left(V_{k,d_k} \prod\nolimits_{j=1}^{d_k-1}(1-V_{kj})\right) \qquad (14)$$

The sufficient statistics calculated from $\{z_n\}_{n=1}^N$ are the counts along each dimension, $k$,

$$m_{1k} = \sum_{n=1}^N \mathbb{I}(z_{nk}=1),\ m_{0k} = \sum_{n=1}^N \mathbb{I}(z_{nk}=0) \qquad (15)$$

### 4.1. Inference for $d_k$

With each iteration, we sample the sequence $\{d_k\}_{k=1}^K$ without using future values from the previous iteration; the value of $K$ is random and equals the number of nonzero $m_{1k}$. The probability that the $k^{\text{th}}$ atom was observed in round $i$ is proportional to

$$p\left(d_k = i | \{d_l\}_{l=1}^{k-1}, \{z_{nk}\}_{n=1}^N, \alpha, \gamma\right) \propto \qquad (16)$$
$$p(\{z_{nk}\}_{n=1}^N | d_k = i, \alpha) p(d_k = i | \{d_l\}_{l=1}^{k-1}, \gamma)$$

Below, we discuss the likelihood and prior terms, followed by an approximation to the posterior.

#### 4.1.1. Likelihood Term

The integral to be solved for integrating out the random variables $\{V_{kj}\}_{j=1}^i$ is

$$p(\{z_{nk}\}_{n=1}^N | d_k = i, \alpha) = \qquad (17)$$
$$\int_{(0,1)^i} f(\{V_{kj}\}_1^i)^{m_{1k}} \{1 - f(\{V_{kj}\}_1^i)\}^{m_{0k}} p(\{V_{kj}\}_1^i | \alpha)\, d\vec{V}$$

where $f(\cdot)$ is the stick-breaking function used in (14). Though this integral can be analytically solved for integer values of $m_{0k}$ via the binomial expansion, we have found that the resulting sum of terms leads to computational precision issues for even small sample sizes. Therefore, we use Monte Carlo methods to approximate this integral.

For $s = 1, \dots, S$ samples, $\{V_{kj}^{(s)}\}_{j=1}^i$, drawn iid from $\text{Beta}(1,\alpha)$, we calculate

$$p(\{z_{nk}\}_{n=1}^N | d_k = i, \alpha) \approx \qquad (18)$$
$$\frac{1}{S}\sum_{s=1}^S f(\{V_{kj}^{(s)}\}_{j=1}^i)^{m_{1k}} \{1 - f(\{V_{kj}^{(s)}\}_{j=1}^i)\}^{m_{0k}}$$

This approximation allows for the use of natural logarithms in calculating the posterior, which was not possible with the analytic solution. Also, to reduce computations, we note that at most two random variables need to be drawn to perform the above stick-breaking, one random variable for the proportion and one for the error; this is detailed in the appendix.

#### 4.1.2. Prior Term

The prior for the sequence of indicators $d_1, d_2, \dots$ is the equivalent sequential process for sampling $C_1, C_2, \dots$, where $C_i = \sum_{k=1}^\infty \mathbb{I}(d_k = i) \sim \text{Poisson}(\gamma)$. Let $\#_{d_{k-1}} = \sum_{j=1}^{k-1} \mathbb{I}(d_j = d_{k-1})$ and let $\mathbb{P}_\gamma(\cdot)$ denote the Poisson distribution with parameter $\gamma$. Then it can be shown that

$$p(d_k = d_{k-1} | \gamma, \#_{d_{k-1}}) = \frac{\mathbb{P}_\gamma(C > \#_{d_{k-1}})}{\mathbb{P}_\gamma(C \geq \#_{d_{k-1}})} \qquad (19)$$

Also, for $h = 1, 2, \dots$, the probability

$$p(d_k = d_{k-1} + h | \gamma, \#_{d_{k-1}}) = \qquad (20)$$
$$\left(1 - \frac{\mathbb{P}_\gamma(C > \#_{d_{k-1}})}{\mathbb{P}_\gamma(C \geq \#_{d_{k-1}})}\right) \mathbb{P}_\gamma(C > 0) \mathbb{P}_\gamma(C = 0)^{h-1}$$

Since $d_k \not< d_{k-1}$, these two terms complete the prior.

#### 4.1.3. Posterior of $d_k$

For the posterior, the normalizing constant requires integration over $h = 0, 1, 2, \dots$, which is not possible given the proposed sampling method. We therefore propose incrementing $h$ until the resulting truncated probability of the largest value of $h$ falls below a threshold (e.g., $10^{-6}$). We have found that the probabilities tend to decrease rapidly for $h > 1$.

### 4.2. Inference for $\gamma$

Given $d_1, d_2, \dots$, the values $C_1, C_2, \dots$ can be reconstructed and a posterior for $\gamma$ can be obtained using a conjugate gamma prior. Since the value of $d_K$ may not be the last in the sequence composing $C_{d_K}$, this value can be "completed" by sampling from the prior, which can additionally serve as proposal factors.

### 4.3. Inference for $\alpha$

Using (18), we again integrate out all stick-breaking random variables to calculate the posterior of $\alpha$,

$$p(\alpha | \{z_n\}_1^N, \{d_k\}_1^K) \propto \prod_{k=1}^K p(\{z_{nk}\}_1^N | \alpha, \{d_k\}_1^K) p(\alpha)$$

Since this is not possible for the positive, real-valued $\alpha$, we approximate this posterior by discretizing the space. Specifically, using the value of $\alpha$ from the previous iteration, $\alpha_{\text{prev}}$, we perform Monte Carlo integration at the points $\{\alpha_{\text{prev}} + t\Delta\alpha\}_{t=-T}^T$, ensuring that $\alpha_{\text{prev}} - T\Delta\alpha > 0$. We use an improper, uniform prior for $\alpha$, with the resulting probability therefore being the normalized likelihood over the discrete set of selected points. As with sampling $d_k$, we again extend the limits beyond $\alpha_{\text{prev}} \pm T\Delta\alpha$, checking that the tails of the resulting probability fall below a threshold.

### 4.4. Inference for $p(z_{nk} = 1|\alpha, d_k, Z_{\text{prev}})$

In latent factor models, (Griffiths & Ghahramani, 2005), the vectors $\{z_n\}_{n=1}^N$ are to be learned with the rest of the model parameters. To calculate the posterior of a given binary indicator therefore requires a prior, which we calculate as follows

$$\begin{aligned} & p(z_{nk} = 1|\alpha, d_k, Z_{\text{prev}}) \qquad (21) \\ &= \int_{(0,1)^{d_k}} p(z_{nk} = 1|\vec{V}) p(\vec{V}|\alpha, d_k, Z_{\text{prev}})\ d\vec{V} \\ &= \frac{\int_{(0,1)^{d_k}} p(z_{nk} = 1|\vec{V}) p(Z_{\text{prev}}|\vec{V}) p(\vec{V}|\alpha, d_k)\ d\vec{V}}{\int_{(0,1)^{d_k}} p(Z_{\text{prev}}|\vec{V}) p(\vec{V}|\alpha, d_k)\ d\vec{V}} \end{aligned}$$

We again perform Monte Carlo integration (18), where the numerator increments the count $m_{1k}$ of the denominator by one. For computational speed, we treat the previous latent indicators, $Z_{\text{prev}}$, as a block (Ishwaran & James, 2001), allowing this probability to remain fixed when sampling the new matrix, $Z$.

## 5. Experiments

We present experimental results on three data sets: (*i*) A synthetic data set; (*ii*) the MNIST handwritten digits data set (digits 3, 5 and 8); and (*iii*) a time-evolving gene expression data set.

### 5.1. Synthetic Data

For the synthetic problem, we investigate the ability of the inference procedure in Section 4 to learn the underlying $\alpha$ and $\gamma$ used in generating $H$. We use the representation in (2) to generate $\boldsymbol{\pi}^{(K)}$ for $K = 100{,}000$. This provides a sample of $\boldsymbol{\pi}^{(K)}$ that approximates the infinite beta process well for smaller values of $\alpha$ and $\gamma$. We then sample $\{z_n\}_{n=1}^{1000}$ from a Bernoulli process and remove all dimensions, $k$, for which $m_{1k} = 0$. Since the weights in (13) are stochastically decreasing as $k$ increases, while the representation in (2) is exchangeable in $k$, we reorder the dimensions of $\{z_n\}_{n=1}^{1000}$ so that $m_{1,1} \geq m_{1,2} \geq \dots$. The binary vectors are treated as observed for this problem.

We present results in Figure 1 for 5,500 trials, where $\alpha_{\text{true}} \sim \text{Uniform}(1, 10)$ and $\gamma_{\text{true}} \sim \text{Uniform}(1, 10)$. We see that the inferred $\alpha_{\text{out}}$ and $\gamma_{\text{out}}$ values center on the true $\alpha_{\text{true}}$ and $\gamma_{\text{true}}$, but increase in variance as these values increase. We believe that this is due in part to the reordering of the dimensions, which are not strictly decreasing in (5), though some reordering is necessary because of the nature of the two priors. We choose to generate data from (2) rather than (5) because it provides some added empirical evidence as to the correctness of the stick-breaking construction.

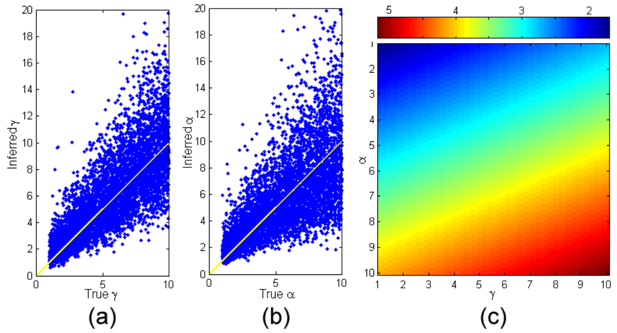

*Figure 1.* Synthetic results for learning $\alpha$ and $\gamma$. For each trial of 150 iterations, 10 samples were collected and averaged over the last 50 iterations. The step size $\Delta\alpha = 0.1$. (a) Inferred $\gamma$ vs true $\gamma$ (b) Inferred $\alpha$ vs true $\alpha$ (c) A plane, shown as an image, fit using least squares that shows the $\ell_1$ distance of the inferred $(\alpha_{\text{out}}, \gamma_{\text{out}})$ to the true $(\alpha_{\text{true}}, \gamma_{\text{true}})$.

### 5.2. MNIST Handwritten Digits

We consider the digits 3, 5 and 8 using 1000 observations for each digit and projecting into 50 dimensions using PCA. We model the resulting digits matrix, $X \in \mathbb{R}^{50\times 3000}$, with a latent factor model (Griffiths & Ghahramani, 2005; Paisley & Carin, 2009),

$$X = \Phi(W \circ Z) + E \qquad (22)$$

where the columns of $Z$ are samples from a Bernoulli process, and the elements of $\Phi$ and $W$ are iid Gaussian. The symbol $\circ$ indicates element-wise multiplication. We infer all variance parameters using inverse-gamma priors, and integrate out the weights, $w_n$, when sampling $z_n$. Gibbs sampling is performed for all parameters, except for the variance parameters, where we perform variational inference (Bishop, 2006). We have found that the "inflation" of the variance parameters that results from the variational expectation leads to faster mixing for the latent factor model.

Figure 2 displays the inference results for an initialization of $K = 200$. The top-left figure shows the number of factors as a function of 10,000 Gibbs iterations, and the top-right figure shows the histogram of these values after 1000 burn-in iterations. For Monte Carlo integration, we use $S = 100{,}000$ samples from the stick-breaking prior for sampling $d_k$ and $p(z_{nk} = 1|\alpha, d_k, Z_{\text{prev}})$, and $S = 10{,}000$ samples for sampling $\alpha$, since learning the parameter $\alpha$ requires significantly more overall samples. The average time per iteration was approximately 18 seconds, though this value increases when $K$ increases and vice-versa. In the bottom two rows of Figure 2, we show four example factor loadings (columns of $\Phi$), as well as the probability of its being used by a 3, 5 and 8.

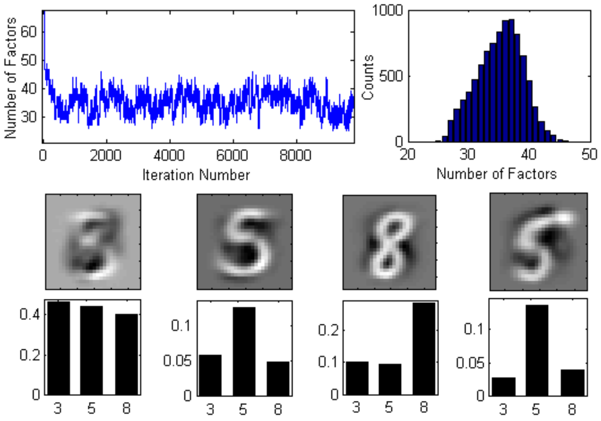

*Figure 2.* Results for MNIST digits 3, 5 and 8. Top left: The number of factors as a function of iteration number. Top right: A histogram of the number of factors after 1000 burn-in iterations. Middle row: Several example learned factors. Bottom row: The probability of a digit possessing the factor directly above.

## 5.3. Time-Evolving Gene Expression Data

We next apply the model discussed in Section 5.2 on data from a viral challenge study (Zaas et al., 2009). In this study, a cohort of 17 healthy volunteers were experimentally infected with the influenza A virus at varying dosages. Blood was taken at intervals between -4 and 120 hours from infection and gene expression values were extracted. Of the 17 patients, 9 ultimately became symptomatic (i.e., became ill), and the goal of the study was to detect this in the gene expression values *prior* to the initial showing of symptoms. There were a total of 16 time points and 267 gene expression extractions, each including expression values for 12,023 genes. Therefore, the data matrix $X \in \mathbb{R}^{267\times 12023}$.

In Figure 3, we show results for 4000 iterations; each iteration took an average of 2.18 minutes. The top row shows the number of factors as a function of iteration, with 100 initial factors, and histograms of the overall number factors, and the number of factors per observation. In the remaining rows, we show four discriminative factor loading vectors, with the statistics from the 267 values displayed as a function of time. We note that the expression values begin to increase for the symptomatic patients prior to the onset of symptoms around the 45th hour. We list the top genes for each factor, as determined by the magnitude of values in $W$ for that factor. In addition, the top three genes in terms of the magnitude of the four-dimensional vector comprising these factors are RSAD2, IFI27 and IFI44L; the genes listed here have a significant overlap with those in the literature (Zaas et al., 2009).

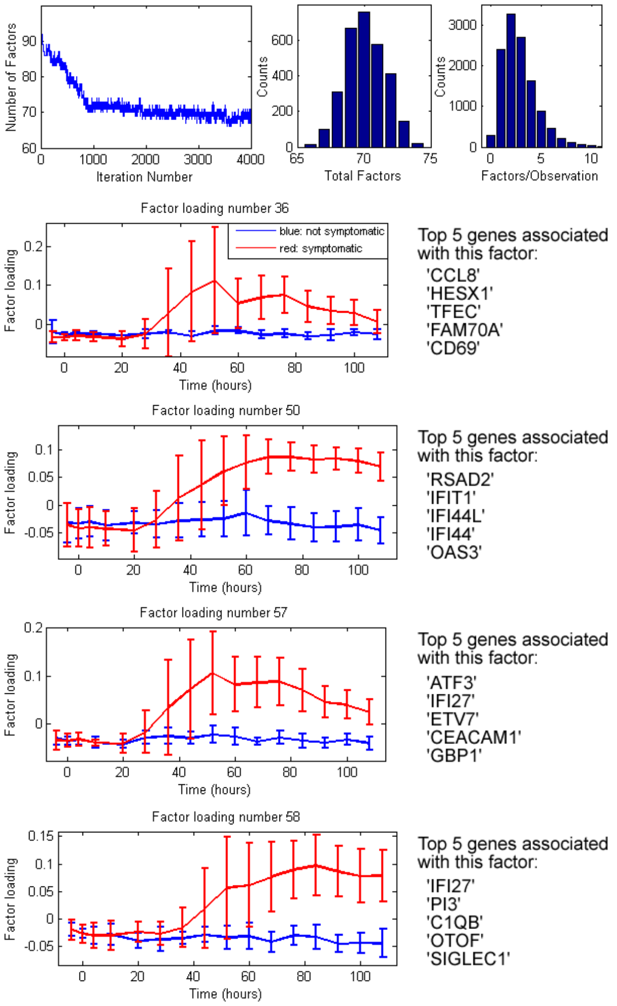

*Figure 3.* Results for time-evolving gene expression data. Top row: (left) Number of factors per iteration (middle) Histogram of the total number of factors after 1000 burn-in iterations (right) Histogram of the number of factors used per observation. Rows 2-5: Discriminative factors and the names of the most important genes associated with each factor (as determined by weight).

As motivated in (Griffiths & Ghahramani, 2005), the values in $Z$ are an alternative to hard clustering, and in this case are useful for group selection. For example, sparse linear classifiers for the model $y = X\beta + \epsilon$, such as the RVM (Bishop, 2006), are prone to select single correlated genes from $X$ for prediction, setting the others to zero. In (West, 2003), latent factor models were motivated as a dimensionality reduction step prior to learning the classifier $y = \Phi\hat{\beta} + \epsilon_2$, where the loading matrix replaces $X$ and unlabeled data are inferred transductively. In this case, discriminative factors selected by the model represent groups of genes associated with that factor, as indicated by $Z$.

## 6. Conclusion

We have presented a new stick-breaking construction of the beta process. The derivation relies heavily upon the constructive definition of the beta distribution, a special case of (Sethuraman, 1994), which has been exclusively used in its infinite form in the machine learning community. We presented an inference algorithm that uses Monte Carlo integration to eliminate several random variables. Results were presented on synthetic data, the MNIST handwritten digits 3, 5 and 8, and time-evolving gene expression data.

As a final comment, we note that the limit of the representation in (2) reduces to the original IBP when $\alpha = 1$. Therefore, the stick-breaking process in (5) should be equal in distribution to the process in (Teh et al., 2007) for this parametrization. The proof of this equality is an interesting question for future work.

**Acknowledgements**
This research was supported by DARPA under the PHD Program, and by the Army Research Office under grant W911NF-08-1-0182.

## References

Bishop, C.M. *Pattern Recognition and Machine Learning.* Springer, New York, 2006.

Blackwell, D. and MacQueen, J.B. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

Damien, Paul, Laud, Purushottam W., and Smith, Adrian F. M. Implementation of bayesian nonparametric inference based on beta processes. *Scandinavian Journal of Statistics*, 23(1):27–36, 1996.

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

Gamerman, D. and Lopes, H.F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition.* Chapman & Hall, 2006.

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*, pp. 475–482, 2005.

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*, 96:161–173, 2001.

Lee, Jaeyong and Kim, Yongdai. A new algorithm to generate beta processes. *Computational Statistics & Data Analysis*, 47(3):441–453, 2004.

Paisley, J. and Carin, L. Nonparametric factor analysis with beta process priors. In *Proc. of the ICML*, pp. 777–784, 2009.

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

Teh, Y.W. and Görür, D. Indian buffet processes with power-law behavior. In *NIPS*, 2009.

Teh, Y.W., Görür, 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.

West, M. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, 2003.

Wolpert, R.L. and Ickstadt, K. Simulations of lévy random fields. *Practical and Semiparametric Bayesian Statistics*, pp. 227–242, 1998.

Zaas, A., Chen, M., Varkey, J., Veldman, T., Hero, A.O., Lucas, J., Huang, Y., Turner, R., Gilbert, A., Lambkin-Williams, R., Oien, N., Nicholson, B., Kingsmore, S., Carin, L., Woods, C., and Ginsburg, G.S. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host & Microbe*, 6:207–217, 2009.

## 7. Appendix

Following $i-1$ breaks from a Beta$(1,\alpha)$ stick-breaking process, the remaining length of the unit-length stick is $\epsilon_i = \prod_{j=1}^{i-1}(1-V_j)$. Let $S_j := -\ln(1-V_j)$. Then, since it can be shown that $S_j \sim \text{Exponential}(\alpha)$, and therefore $\sum_{j=1}^{i-1} S_j \sim \text{Gamma}(i-1,\alpha)$, the value of $\epsilon_i$ can be calculated using only one random variable,

$$\begin{aligned} \epsilon_i &= \mathrm{e}^{-T_i} \\ T_i &\sim \text{Gamma}(i-1,\alpha) \end{aligned}$$

Therefore, to draw $V_i \prod_{j=1}^{i-1}(1-V_j) = \epsilon_i V_i$, one can sample $V_i \sim \text{Beta}(1,\alpha)$ and $\epsilon_i$ as above.