Nonparametric Factor Analysis with Beta Process Priors

John Paisley JWP4@EE.DUKE.EDU
Lawrence Carin LCARIN@EE.DUKE.EDU
Department of Electrical & Computer Engineering
Duke University, Durham, NC 27708

Abstract

We propose a nonparametric extension to the factor analysis problem using a beta process prior. This *beta process factor analysis* (BP-FA) model allows for a dataset to be decomposed into a linear combination of a sparse set of factors, providing information on the underlying structure of the observations. As with the Dirichlet process, the beta process is a fully Bayesian conjugate prior, which allows for analytical posterior calculation and straightforward inference. We derive a variational Bayes inference algorithm and demonstrate the model on the MNIST digits and HGDP-CEPH cell line panel datasets.

1. Introduction

Latent membership models provide a useful means for discovering underlying structure in a dataset by elucidating the relationships between observed data. For example, in *latent class models*, observations are assumed to be generated from one of K classes, with mixture models constituting a classic example. When a single class indicator is considered too restrictive, *latent feature models* can be employed, allowing for an observation to possess combinations of up to K latent features.

As K is typically unknown, Bayesian nonparametric models seek to remove the need to set this value by defining robust, but sparse priors on infinite spaces. For example, the Dirichlet process (Ferguson, 1973) allows for nonparametric mixture modeling in the latent class scenario. In the latent feature paradigm, the beta process (Hjort, 1990) has been defined and can be used toward the same objective, which, when marginalized, is closely related to the Indian buffet process (Griffiths & Ghahramani, 2005; Thibaux & Jordan, 2007).

An example of a latent feature model is the factor analysis model (West, 2003), where a data matrix is decomposed into the product of two matrices plus noise,

$$X = \Phi Z + E \tag{1}$$

In this model, the columns of the $D \times K$ matrix of factor loadings, Φ, can be modeled as latent features and the elements in each of N columns of Z can be modeled as indicators of the possession of a feature for the corresponding column of X (which can be given an associated weight). It therefore seems natural to seek a nonparametric model for this problem.

To this end, several models have been proposed that use the Indian buffet process (IBP) (Knowles & Ghahramani, 2007; Rai & Daumé, 2008; Meeds et al., 2007). However, these models require MCMC inference, which can be slow to converge. In this paper, we propose a *beta process factor analysis* (BP-FA) model that is fully conjugate and therefore has a fast variational solution; this is an intended contribution of this paper. Starting from first principles, we show how the beta process can be formulated to solve the nonparametric factor analysis problem, as the Dirichlet process has been previously shown to solve the nonparametric mixture modeling problem; we intend for this to be a second contribution of this paper.

The remainder of the paper is organized as follows. In Section 2 we review the beta process in detail. We introduce the BP-FA model in Section 3, and discuss some of its theoretical properties. In Section 4 we derive a variational Bayes inference algorithm for fast inference, exploiting full conjugacy within the model. Experimental results are presented in Section 5 on synthetic data, and on the MNIST digits and HGDP-CEPH cell line panel (Rosenberg et al., 2002) datasets. We conclude and discuss future work in Section 6.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

2. The Beta Process

The beta process, first introduced by Hjort for survival analysis (Hjort, 1990), is an independent increments, or Lévy process and can be defined as follows:

Definition: Let Ω be a measurable space and $\mathcal{B}$ its σ-algebra. Let H_0 be a continuous probability measure on $(\Omega, \mathcal{B})$ and α a positive scalar. Then for all disjoint, infinitesimal partitions, $\{B_1, \ldots, B_K\}$, of Ω the beta process is generated as follows,

$$H(B_k) \sim Beta(\alpha H_0(B_k), \alpha(1 - H_0(B_k))) \quad (2)$$

with $K \to \infty$ and $H_0(B_k) \to 0$ for $k = 1, \ldots, K$. This process is denoted $H \sim BP(\alpha H_0)$.

Because of the convolution properties of beta random variables, the beta process does not satisfy the Kolmogorov consistency condition, and is therefore defined in the infinite limit (Billingsley, 1995). Hjort extends this definition to include functions, $\alpha(B_k)$, which for simplicity is here set to a constant.

Like the Dirichlet process, the beta process can be written in set function form,

$$H(\omega) = \sum_{k=1}^{\infty} \pi_k \delta_{\omega_k}(\omega) \quad (3)$$

with $H(\omega_i) = \pi_i$. Also like the Dirichlet process, means for drawing H are not obvious. We briefly discuss this issue in Section 2.1. In the case of the beta process, π does not serve as a probability mass function on Ω, but rather as part of a new measure on Ω that parameterizes a Bernoulli process defined as follows:

Definition: Let the column vector, z_i, be infinite and binary with the k^{th} value, z_{ik}, generated by

$$z_{ik} \sim Bernoulli(\pi_k) \quad (4)$$

The new measure, $X_i(\omega) = \sum_k z_{ik}\delta_{\omega_k}(\omega)$, is then drawn from a Bernoulli process, or $X_i \sim BeP(H)$.

By arranging samples of the infinite-dimensional vector, z_i, in matrix form, $Z = [z_1, \ldots, z_N]$, the beta process is seen to be a prior over infinite binary matrices, with each row in the matrix Z corresponding to a location, δ_ω.

2.1. The Marginalized Beta Process and the Indian Buffet Process

As previously mentioned, sampling H directly from the infinite beta process is difficult, but a marginalized approach can be derived in the same manner as the corresponding Chinese restaurant process (Aldous, 1985), used for sampling from the Dirichlet process.

We briefly review this marginalization, discussing the link to the Indian buffet process (Griffiths & Ghahramani, 2005) as well as other theoretical properties of the beta process that arise as a result.

We first extend the beta process to take two scalar parameters, a, b, and partition Ω into K regions of equal measure, or $H_0(B_k) = 1/K$ for $k = 1, \ldots, K$. We can then write the generative process in the form of (3) as

$$\begin{aligned} H(B) &= \sum_{k=1}^{K} \pi_k \delta_{B_k}(B) \\ \pi_k &\sim Beta\left(a/K, b(K-1)/K\right) \end{aligned} \quad (5)$$

where $B \in \{B_1, \ldots, B_K\}$. Marginalizing the vector π and letting $K \to \infty$, the matrix, Z, can be generated directly from the beta process prior as follows:

1. For an infinite matrix, Z, initialized to all zeros, set the first $c_1 \sim Po(a/b)$ rows of z_1 to 1. Sample the associated locations, ω_i, $i = 1, \ldots, c_1$, independently from H_0.
2. For observation N, sample $c_N \sim Po\left(\frac{a}{b+N-1}\right)$ and define $C_N \equiv \sum_{i=1}^{N} c_i$. For rows $k = 1, \ldots, C_{N-1}$ of z_N, sample

$$z_{Nk} \sim Bernoulli\left(\frac{n_{Nk}}{b+N-1}\right) \quad (6)$$

where $n_{Nk} \equiv \sum_{i=1}^{N-1} z_{ik}$, the number of previous observations with a 1 at location k. Set indices $C_{N-1}+1$ to C_N equal to 1 and sample associated locations independently from H_0.

If we define

$$H(\omega) \equiv \sum_{k=1}^{\infty} \frac{n_{Nk}}{b+N-1}\delta_{\omega_k}(\omega) \quad (7)$$

then $H \sim BP(a, b, H_0)$ in the limit as $N \to \infty$, and the exchangeable columns of Z are drawn iid from a beta process. As can be seen, in the case where $b = 1$, the marginalized beta process is equivalent to the Indian buffet process (Thibaux & Jordan, 2007).

This representation can be used to derive some interesting properties of the beta process. We observe that the random variable, C_N, has a Poisson distribution, $C_N \sim Po\left(\sum_{i=1}^{N} \frac{a}{b+i-1}\right)$, which provides a sense of how the matrix Z grows with sample size. Furthermore, since $\sum_{i=1}^{N} \frac{a}{b+i-1} \to \infty$ as $N \to \infty$, we can deduce that the entire space of Ω will be explored as the number of samples grows to infinity.

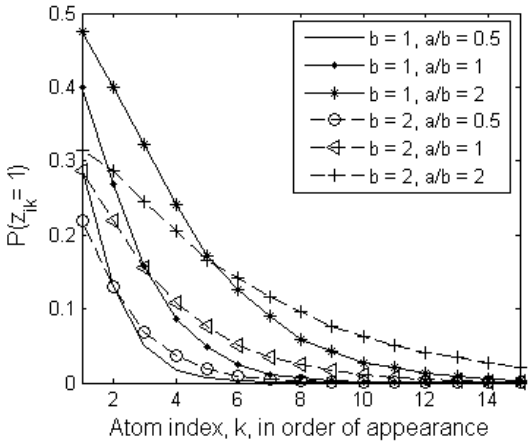

Figure 1. Estimation of π from 5000 marginal beta process runs of 500 samples each, with various a, b initializations.

We show in Figure 1 the expectation of π calculated empirically by drawing from the marginalized beta process. As can be seen, the a, b parameters offer flexibility in both the magnitude and shape of π and can be tuned.

2.2. Finite Approximation to the Beta Process

As hinted in (5), a finite approximation to the beta process can be made by simply setting K to a large, but finite number. This approximation can be viewed as serving a function similar to the finite Dirichlet distribution in its approximation of the infinite Dirichlet process for mixture modeling. The finite representation is written as

$$\begin{aligned} H(\omega) &= \sum_{k=1}^{K} \pi_k \delta_{\omega_k}(\omega) \\ \pi_k &\sim Beta(a/K, b(K-1)/K) \\ \omega_k &\overset{iid}{\sim} H_0 \end{aligned} \tag{8}$$

with the K-dimensional vector, z_i, drawn from a finite Bernoulli process parameterized by H. The full conjugacy of this representation means posterior computation is analytical, which will allow for variational inference to be performed on the BP-FA model.

We briefly mention that a stick-breaking construction of the beta process has recently been derived (Paisley & Carin, 2009), allowing for exact Bayesian inference. A construction for the Indian buffet process has also been presented (Teh et al., 2007), though this method does not extend to the more general beta process. We will use the finite approximation presented here in the following sections.

3. Beta Process Factor Analysis

Factor analysis can be viewed as the process of modeling a data matrix, $X \in \mathbb{R}^{D\times N}$, as the multiplication of two matrices, $\Phi \in \mathbb{R}^{D\times K}$ and $Z \in \mathbb{R}^{K\times N}$, plus an error matrix, E.

$$X = \Phi Z + E \tag{9}$$

Often, prior knowledge about the structure of the data is used, for example, the desired sparseness properties of the Φ or Z matrices (West, 2003; Rai & Daumé, 2008; Knowles & Ghahramani, 2007). The beta process is another such prior that achieves this sparseness, allowing for K to tend to infinity while only focusing on a small subset of the columns of Φ via the sparse matrix Z.

In *beta process factor analysis* (BP-FA), we model the matrices Φ and Z as N draws from a Bernoulli process parameterized by a beta process, H. First, we recall that draws from the BeP-BP approximation can be generated as

$$\begin{aligned} z_{ik} &\sim Bernoulli(\pi_k) \\ \pi_k &\sim Beta(a/K, b(K-1)/K) \\ \phi_k &\overset{iid}{\sim} H_0 \end{aligned} \tag{10}$$

for observation $i = 1, \ldots, N$ and latent feature (or factor) $k = 1, \ldots, K$. In the general definition, H_0 was unspecified, as was the use of the latent membership vector, z_i. For BP-FA, we let H_0 be multivariate normal and the latent factors be indicators of linear combinations of these locations, which can be written in matrix notation as Φz_i, where $\Phi = [\phi_1, \ldots, \phi_K]$. Adding the noise vector, ϵ_i, we obtain observation x_i. The beta process can thus be seen as a prior on the parameters, $\{\pi, \Phi\}$, with iid Bernoulli process samples composing the expectation matrix, $\mathbb{E}[X] = \Phi Z$ for the factor analysis problem.

As an unweighted linear combination might be too restrictive, we include a weight vector, w_i, which results in the following generative process for observation $i = 1, \ldots, N$,

$$\begin{aligned} x_i &= \Phi(z_i \circ w_i) + \epsilon_i \\ w_i &\sim \mathcal{N}(0, \sigma_w^2 I) \\ z_{ik} &\sim Bernoulli(\pi_k) \\ \pi_k &\sim Beta(a/K, b(K-1)/K) \\ \phi_k &\sim \mathcal{N}(0, \Sigma) \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_n^2 I) \end{aligned} \tag{11}$$

for $k = 1, \ldots, K$ and all values drawn independently. The symbol $\circ$ represents the Hadamard, or elementwise multiplication of two vectors. We show a graphical representation of the BP-FA model in Figure 2.

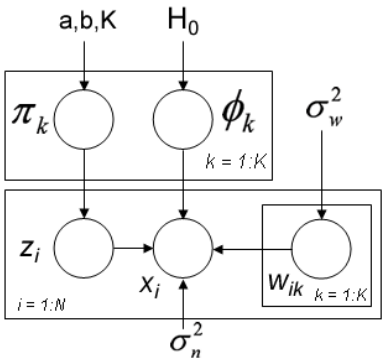

Figure 2. A graphical representation of the BP-FA model.

Written in matrix notation, the weighted BP-FA model of (11) is thus a prior on

$$X = \Phi(Z \circ W) + E \tag{12}$$

Under this prior, the mean and covariance of a given vector, x, can be calculated,

$$\begin{aligned} \mathbb{E}[x] &= 0 \\ \mathbb{E}[xx^T] &= \frac{aK}{a+b(K-1)}\sigma_w^2\Sigma + \sigma_n^2 I \end{aligned} \tag{13}$$

Letting $K \to \infty$, we see that $\mathbb{E}[xx^T] \to \frac{a}{b}\sigma_w^2\Sigma + \sigma_n^2 I$. Therefore, the BP-FA model remains well-defined in the infinite limit. To emphasize, compare this value with z removed, where $\mathbb{E}[xx^T] = K\sigma_w^2\Sigma + \sigma_n^2 I$. The coefficient $\frac{a}{b}$ is significant in that it represents the expected number of factors present in an observation as $K \to \infty$. That is, if we define $m_i \equiv \sum_{k=1}^{\infty} z_{ik}$, where $z_i \sim BeP(H)$ and $H \sim BP(a, b, H_0)$, then by marginalizing H we find that $\mathbb{E}[m_i] = \frac{a}{b}$.

Another important aspect of the BP-FA model is that the π vector enforces sparseness on the *same* subset of factors. In comparison, consider the model where z_i is removed and sparseness is enforced by sampling the elements of w_i iid from a sparseness inducing normal-gamma prior. This is equivalent to learning multiple relevance vector machines (Tipping, 2001) with a jointly learned and shared Φ matrix. A theoretical issue with this model is that the prior does not induce sparseness on the *same* subset of latent factors. As $K \to \infty$, all factors will be used sparsely with equal probability and, therefore, no factors will be shared. This is conceptually similar to the problem of drawing multiple times from a Dirichlet process prior, where individual draws are sparse, but no two draws are sparse on the same subset of atoms. We note that the hierarchical Dirichlet process has been introduced to resolve this particular issue (Teh et al., 2006).

4. Variational Bayesian Inference

In this section, we derive a variational Bayesian algorithm (Beal, 2003) to perform fast inference for the weighted BP-FA model of (11). This is aided by the conjugacy of the beta to the Bernoulli process, where the posterior for the single parameter beta process is

$$H|X_1, \ldots, X_N \sim BP\left(\alpha H_0 + \sum_{i=1}^{N} X_i\right) \tag{14}$$

with $X_i \sim BeP(H)$ being the i^{th} sample from a Bernoulli process parameterized by H. The two-parameter extension has a similar posterior update, though not as compact a written form.

In the following, we define $x_i^{-k} \equiv x_i - \Phi_{-k}(z_i^{-k} \circ w_i^{-k})$, where Φ_{-k}, z^{-k} and w^{-k} are the matrix/vectors with the k^{th} column/element removed; this is simply the portion of x_i remaining considering all but the k^{th} factor. Also, for clarity, we have suppressed certain equation numbers and conditional variables.

4.1. The VB-E Step

Update for Z:
$p(z_{ik}|x_i, \Phi, w_i, z_i^{-k}) \propto p(x_i|z_{ik}, \Phi, w_i, z_i^{-k})p(z_{ik}|\pi)$

The probability that $z_{ik} = 1$ is proportional to

$$\exp[\langle \ln(\pi_k) \rangle] \times$$

$$\exp\left[-\frac{1}{2\sigma_n^2}\left(\langle w_{ik}^2\rangle\langle\phi_k^T\phi_k\rangle - 2\langle w_{ik}\rangle\langle\phi_k\rangle^T\langle x_i^{-k}\rangle\right)\right]$$

where $\langle\cdot\rangle$ indicates the expectation. The probability that $z_{ik} = 0$ is proportional to $\exp[\langle \ln(1-\pi_k) \rangle]$. The expectations can be calculated as

$$\langle \ln(\pi_k) \rangle = \psi\left(\frac{a}{K} + \langle n_k \rangle\right) - \psi\left(\frac{a+b(K-1)}{K} + N\right)$$

$$\langle \ln(1-\pi_k) \rangle =$$

$$\psi\left(\frac{b(K-1)}{K} + N - \langle n_k \rangle\right) - \psi\left(\frac{a+b(K-1)}{K} + N\right)$$

where $\psi(\cdot)$ represents the digamma function and

$$\langle w_{ik}^2 \rangle = \langle w_{ik} \rangle^2 + \Delta_i'^{(k)} \tag{15}$$

$$\langle \phi_k^T \phi_k \rangle = \langle \phi_k \rangle^T \langle \phi_k \rangle + \mathrm{trace}(\Sigma_k') \tag{16}$$

where $\langle n_k \rangle$ is defined in the update for π, Σ_k' in the update for Φ, and $\Delta_i'^{(k)}$ is the k^{th} diagonal element of Δ_i' defined in the update for W.

4.2. The VB-M Step

Update for π:
$p(\pi_k|Z) \propto p(Z|\pi_k)p(\pi_k|a,b,K)$

The posterior of π_k can be shown to be

$$\pi_k \sim Beta\left(\frac{a}{K} + \langle n_k \rangle, \frac{b(K-1)}{K} + N - \langle n_k \rangle\right)$$

where $\langle n_k \rangle = \sum_{i=1}^{N} \langle z_{ik} \rangle$ can be calculated from the VB-E step. The priors a, b can be tuned according to the discussion in Section 2.1. We recall that $\sum_{i=1}^{N} \frac{a}{b+i-1}$ is the expected total number of factors, while a/b is the expected number of factors used by a single observation in the limiting case.

Update for Φ:
$p(\phi_k|X, \Phi_{-k}, Z, W) \propto p(X|\phi_k, \Phi_{-k}, Z, W)p(\phi_k|\Sigma)$

The posterior of ϕ_k can be shown to be normal with mean, μ'_k, and covariance, Σ'_k, equal to

$$\Sigma'_k = \left(\frac{1}{\sigma_n^2}\sum_{i=1}^{N}\langle z_{ik}\rangle\langle w_{ik}^2\rangle I + \Sigma^{-1}\right)^{-1} \quad (17)$$

$$\mu'_k = \Sigma'_k\left(\frac{1}{\sigma_n^2}\sum_{i=1}^{N}\langle z_{ik}\rangle\langle w_{ik}\rangle\langle x_i^{-k}\rangle\right) \quad (18)$$

with $\langle w_{ik}^2 \rangle$ given in (15). The prior Σ can be set to the empirical covariance of the data, X.

Update for W:
$p(w_i|x_i, \Phi, z_i) \propto p(x_i|w_i, \Phi, z_i)p(w_i|\sigma_w^2)$

The posterior of w_i can be shown to be multivariate normal with mean, v'_i, and covariance, Δ'_i, equal to

$$\Delta'_i = \left(\frac{1}{\sigma_n^2}\langle\tilde{\Phi}_i^T\tilde{\Phi}_i\rangle + \frac{1}{\sigma_w^2}I\right)^{-1} \quad (19)$$

$$v'_i = \Delta'_i\left(\frac{1}{\sigma_n^2}\langle\tilde{\Phi}_i\rangle^T x_i\right) \quad (20)$$

where we define $\tilde{\Phi}_i \equiv \Phi \circ \tilde{Z}_i$ and $\tilde{Z}_i \equiv [z_i, \ldots, z_i]^T$, with the K-dimensional vector, z_i, repeated D times. Given that $\langle\tilde{\Phi}_i\rangle = \langle\Phi\rangle \circ \langle\tilde{Z}_i\rangle$, we can then calculate

$$\langle\tilde{\Phi}_i^T\tilde{\Phi}_i\rangle = \left(\langle\Phi\rangle^T\langle\Phi\rangle + A\right) \circ \left(\langle z_i\rangle\langle z_i\rangle^T + B_i\right) \quad (21)$$

where A and B_i are calculated as follows

$$A \equiv \text{diag}\left[\text{trace}(\Sigma'_1), \ldots, \text{trace}(\Sigma'_K)\right]$$
$$B_i \equiv \text{diag}\left[\langle z_{i1}\rangle(1-\langle z_{i1}\rangle), \ldots, \langle z_{iK}\rangle(1-\langle z_{iK}\rangle)\right]$$

A prior, discussed below, can be placed on σ_w^2, removing the need to set this value.

Update for σ_n^2:
$p(\sigma_n^2|X, \Phi, Z, W) \propto p(X|\Phi, Z, W, \sigma_n^2)p(\sigma_n^2)$

We can also infer the noise parameter, σ_n^2, by using an inverse-gamma, $InvGa(c,d)$, prior. The posterior can be shown to be inverse-gamma with

$$c' = c + \frac{ND}{2} \quad (22)$$

$$d' = d + \frac{1}{2}\sum_{i=1}^{N}\left(\|x_i - \langle\Phi\rangle(\langle z_i\rangle \circ \langle w_i\rangle)\|^2 + \xi_i\right)$$

where

$$\xi_i \equiv \sum_{k=1}^{K}\left(\langle z_{ik}\rangle\langle w_{ik}^2\rangle\langle\phi_k^T\phi_k\rangle - \langle z_{ik}\rangle^2\langle w_{ik}\rangle^2\langle\phi_k\rangle^T\langle\phi_k\rangle\right) + \sum_{k\neq l}\langle z_{ik}\rangle\langle z_{il}\rangle\Delta'_{i,kl}\langle\phi_k\rangle^T\langle\phi_l\rangle$$

In the previous equations, σ_n^{-2} can then be replaced by $\langle\sigma_n^{-2}\rangle = c'/d'$.

Update for σ_w^2:
$p(\sigma_w^2|W) \propto p(W|\sigma_w^2)p(\sigma_w^2)$

Given a conjugate, $InvGa(e,f)$ prior, the posterior of σ_w^2 is also inverse-gamma with

$$e' = e + \frac{NK}{2} \quad (23)$$

$$f' = f + \frac{1}{2}\sum_{i=1}^{N}\left(\langle w_i\rangle^T\langle w_i\rangle + \text{trace}(\Delta'_i)\right) \quad (24)$$

4.3. Accelerated VB Inference

As with the Dirichlet process, there is a tradeoff in variational inference for the BP-FA; the larger K is set, the more accurate the model should be, but the slower the model inference. We here briefly mention a simple remedy for this problem.

Following every iteration, the total factor membership expectations, $\{\langle n_k\rangle\}_{k=1}^K$, can be used to assess the relevancy of a particular factor. When this number falls below a small threshold (e.g., 10^{-16}), this factor index can be skipped in following iterations with minimal impact on the convergence of the algorithm. In this way, the algorithm should converge more quickly as the number of iterations increases.

4.4. Prediction for New Observations

Given the outputs, $\{\pi, \Phi\}$, the vectors z^* and w^* can be inferred for a new observation, x^*, using a MAP-EM inference algorithm that iterates between z^* and w^*. The equations are similar to those detailed above, with inference for π and Φ removed.

5. Experiments

Factor analysis models are useful in many applications, for example, for dimensionality reduction in gene expression analysis (West, 2003). In this section, we demonstrate the performance of the BP-FA model on synthetic data, and apply it to the MNIST digits and HGDP-CEPH cell line panel (Rosenberg et al., 2002) datasets.

5.1. A Synthetic Example

For our synthetic example, we generated H from the previously discussed approximation to the Beta process with $a, b = 1$, $K = 100$ and $\phi_k \sim \mathcal{N}(0, I)$ in a $D = 25$ dimensional space. We generated $N = 250$ samples from a Bernoulli process parameterized by H and synthesized X with $W = 1$ and $\sigma_n^2 = 0.0675$. Below, we show results for the model having the highest likelihood selected from five runs, though the results in general were consistent.

In Figure 3, we display the ground truth (top) of Z, rearranged for display purposes. We note that only seven factors were actually used, while several observations contain no factors at all, and thus are pure noise. We initialized our model to $K = 100$ factors, though as the results show (bottom), only a small subset were ultimately used. The inferred $\langle \sigma_n^2 \rangle = 0.0625$ and the elementwise MSE of 0.0186 to the true ΦZ further indicates good performance. For this example, the BP-FA model was able to accurately uncover the underlying latent structure of the dataset.

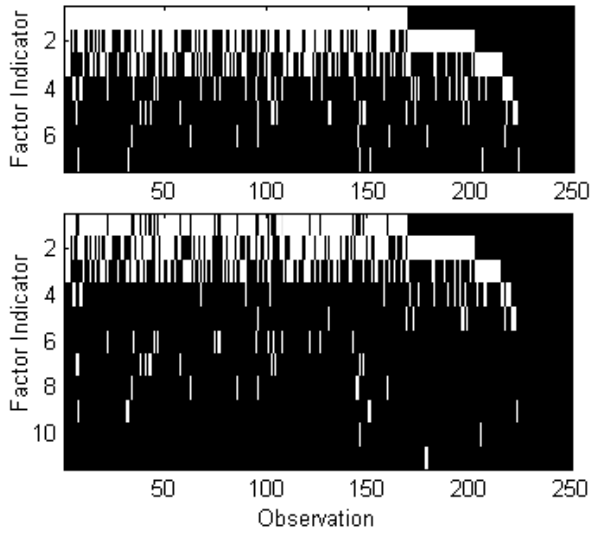

Figure 3. Synthetic Data: Latent factor indicators, Z, for the true (top) and inferred (bottom) models.

5.2. MNIST Handwritten Digits Dataset

We trained our BP-FA model on $N = 2500$ odd digits (500 each) from the MNIST digits dataset. Using PCA, we reduced the dimensionality to $D = 350$, which preserved over 99.5% of the total variance within the data. We truncated the BP-FA model to $K = 100$ factors initialized using the K-means algorithm and ran five times, selecting the run with the highest likelihood, though again the results were consistent.

In Figure 4 below, we show the factor sharing across the digits (left) by calculating the expected number of factors shared between two observations and normalizing by the largest value (0.58); larger boxes indicate more sharing. At right, we show for each of the odd digits the most commonly used factor, followed by the second most used factor *given* the factor to the left. Of particular interest are the digits 3 and 5, where they heavily share the same factor, followed by a factor that differentiates the two numbers.

In Figure 5 (top), we plot the sorted values of $\langle \pi \rangle$ inferred by the algorithm. As can be seen, the algorithm inferred a sparse set of factors, fewer than the 100 initially provided. Also in Figure 5 (bottom), we show an example of a reconstruction of the number 3 that uses four factors. As can be seen, no single factor can individually approximate the truth as well as their weighted linear combination. We note that the BP-FA model was fast, requiring 35 iterations on average to converge and requiring approximately 30 minutes for each run on a 2.66 GHz processor.

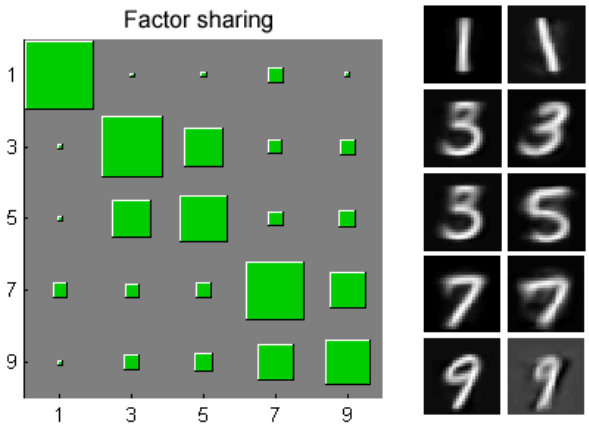

Figure 4. Left: Expected factor sharing between digits. Right: (left) Most frequently used factors for each digit (right) Most used second factor per digit given left factor.

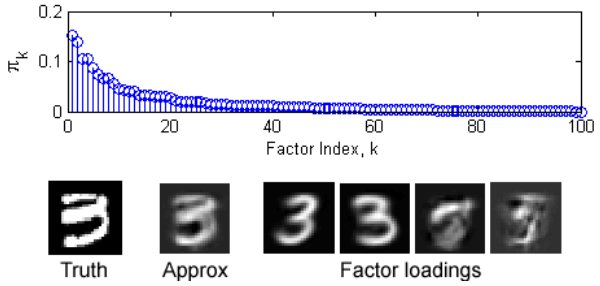

Figure 5. (top) Inferred π indicating sparse factor usage. (bottom) An example reconstruction.

5.3. HGDP-CEPH Cell Line Panel

The HGDP-CEPH Human Genome Diversity Cell Line Panel (Rosenberg et al., 2002) is a dataset comprising genotypes at $D = 377$ autosomal microsatellite loci sampled from $N = 1056$ individuals in 52 populations across the major geographic regions of the world. It is useful for inferring human evolutionary history and migration.

We ran our model on this dataset initializing $K = 100$ factors, though again, only a subset were significantly used. Figure 6 contains the sharing map, as previously calculated for the MNIST dataset, normalized on 0.55. We note the slight differentiation between the Middle East and European regions, a previous issue for this dataset (Rosenberg et al., 2002).

We also highlight the use of BP-FA in denoising. Figure 8 shows the original HGDP-CEPH data, as well as the $\Phi(Z \circ W)$ reconstruction projected onto the first 20 principal components of the raw data. The figure shows how the BP-FA model was able to substantially reduce the noise level within the data, while still retaining the essential structure.

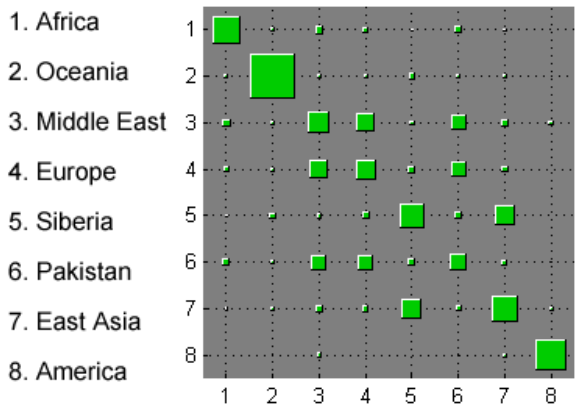

Figure 6. Factor sharing across geographic regions.

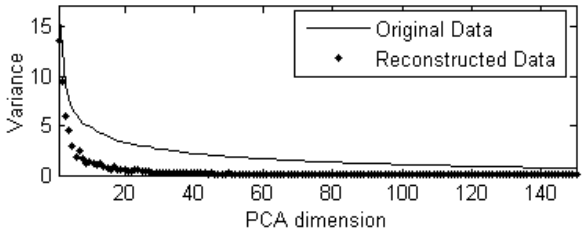

Figure 7. Variance of HGDP-CEPH data along the first 150 principal components of the raw features for original and reconstructed data.

This is also evident in Figure 7, where we plot the variance along these same principal components for the first 150 dimensions. For an apparently noisy dataset such as this, BP-FA can potentially be useful as a preprocessing step in conjunction with other algorithms, in this case, for example, the *Structure* (Rosenberg et al., 2002) or recently proposed *mStruct* (Shringarpure & Xing, 2008) algorithms.

6. Conclusion and Future Work

We have proposed a *beta process factor analysis* (BP-FA) model for performing nonparametric factor analysis with a potentially infinite number of factors. As with the Dirichlet process prior used for mixture modeling, the beta process is a fully Bayesian prior that assures the sharing of a sparse subset of factors among all observations. Taking advantage of conjugacy within the model, a variational Bayes algorithm was developed for fast model inference requiring an approximation comparable to the finite Dirichlet distribution's approximation to the infinite Dirichlet process. Results were shown on synthetic data, as well as the MNIST handwritten digits and HGDP-CEPH cell line panel datasets.

While several nonparametric factor analysis models have been proposed for applications such as independent components analysis (Knowles & Ghahramani, 2007) and gene expression analysis (Rai & Daumé, 2008; Meeds et al., 2007), these models rely on the Indian buffet process and therefore do not have fast variational solutions - an intended contribution of this paper. Furthermore, while the formal link has been made between the IBF and the beta process (Thibaux & Jordan, 2007), we believe our further development and application to factor analysis to be novel. In future work, the authors plan to develop a stick-breaking process for drawing directly from the beta process similar to (Sethuraman, 1994) for drawing from the Dirichlet process, which will remove the need for finite approximations.

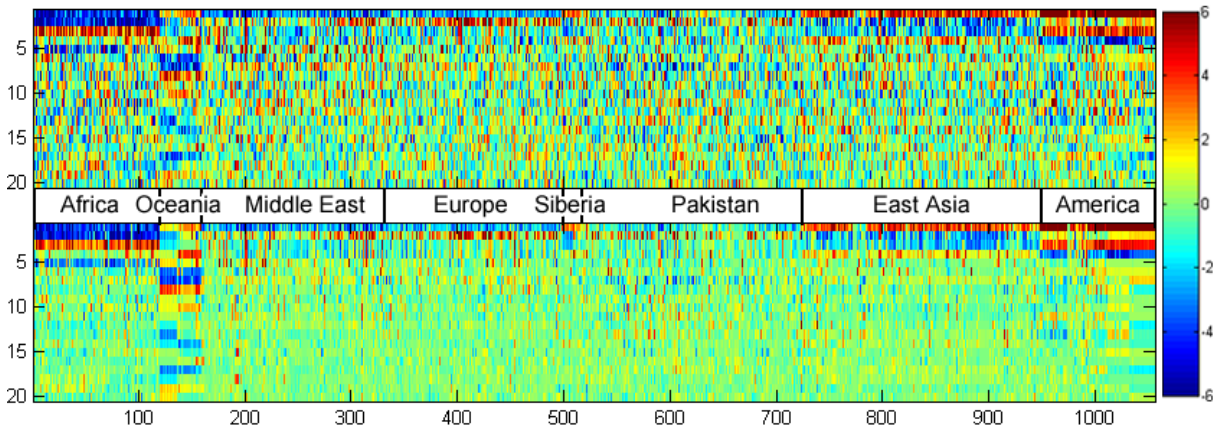

Figure 8. HGDP-CEPH features projected onto the first 20 principal components of the raw features for the (top) original and (bottom) reconstructed data. The broad geographic breakdown is given between the images.

References

Aldous, D. (1985). Exchangeability and related topics. *École d'ete de probabilités de Saint-Flour XIII-1983*, 1–198.

Beal, M. (2003). *Variational algorithms for approximate bayesian inference*. Doctoral dissertation, Gatsby Computational Neuroscience Unit, University College London.

Billingsley, P. (1995). *Probability and measure, 3rd edition*. Wiley Press, New York.

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

Griffiths, T. L., & Ghahramani, Z. (2005). Infinite latent feature models and the indian buffet process. *Advances in Neural Information Processing Systems* (pp. 475–482).

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

Knowles, D., & Ghahramani, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis. *7th International Conference on Independent Component Analysis and Signal Separation*.

Meeds, E., Ghahramani, Z., Neal, R., & Roweis, S. (2007). Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems* (pp. 977–984).

Paisley, J., & Carin, L. (2009). *A stick-breaking construction of the beta process* (Technical Report). Duke University, ee.duke.edu/∼jwp4/StickBP.pdf.

Rai, P., & Daumé, H. (2008). The infinite hierarchical factor regression model. *Advances in Neural Information Processing Systems*.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*, 2381–2385.

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

Shringarpure, S., & Xing, E. P. (2008). mstruct: A new admixture model for inference of population structure in light of both genetic admixing and allele mutation. *Proceedings of the 25th International Conference on Machine Learning* (pp. 952–959).

Teh, Y. W., Görür, D., & Ghahramani, Z. (2007). Stick-breaking construction for the indian buffet process. *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Thibaux, R., & Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. *International Conference on Artificial Intelligence and Statistics*.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, *7*, 723–732.