Dependent Hierarchical Beta Process for Image Interpolation and Denoising

[1]Mingyaun Zhou [2]Hongxia Yang [3]Guillermo Sapiro [2]David Dunson [1]Lawrence Carin
[1]Department of ECE, [2]Department of Statistical Science, Duke University, Durham, NC 27708
[3]Department of ECE, University of Minnesota, Minneapolis, MN 55455

Abstract

A dependent hierarchical beta process (dHBP) is developed as a prior for data that may be represented in terms of a sparse set of latent features, with covariate-dependent feature usage. The dHBP is applicable to general covariates and data models, imposing that signals with similar covariates are likely to be manifested in terms of similar features. Coupling the dHBP with the Bernoulli process, and upon marginalizing out the dHBP, the model may be interpreted as a covariate-dependent hierarchical Indian buffet process. As applications, we consider interpolation and denoising of an image, with covariates defined by the location of image patches within an image. Two types of noise models are considered: (*i*) typical white Gaussian noise; and (*ii*) spiky noise of arbitrary amplitude, distributed uniformly at random. In these examples, the features correspond to the atoms of a dictionary, learned based upon the data under test (without *a priori* training data). State-of-the-art performance is demonstrated, with efficient inference using hybrid Gibbs, Metropolis-Hastings and slice sampling.

1 INTRODUCTION

There has been significant recent interest in the Indian buffet process (IBP) (Griffiths and Ghahramani, 2005; Knowles and Ghahramani, 2007; Miller et al., 2008; Rai and Daumé, 2008; Williamson et al., 2010) and in the related beta process (BP) (Paisley and Carin, 2009; Teh and Gorur, 2009; Thibaux and Jordan, 2007; Zhou et al., 2009). These models have been applied to factor analysis to infer a set of factors (features/dictionary atoms) with which data may be sparsely represented. In many applications the signal (and hence features) are dependent on observable covariates. For example, in image-processing applications (Mairal et al., 2009, 2008; Zhou et al., 2009) one often represents an image in terms of a set of local patches (each composed of a contiguous subset of pixels), and the objective is to represent each patch as a sparse linear combination of dictionary atoms (features). All patches are processed jointly, and it is desirable to account for their spatial locations (the covariates in this application) when learning the underlying dictionary.

For sparse image analysis, $\boldsymbol{x}_i \in \mathbb{R}^P$ represents the ith image and $\{\boldsymbol{x}_i\}_{i=1,N}$ represents the set of images under analysis. The $\boldsymbol{x}_i$ may represent the ith of N patches from a single image, or it may represent the ith entire image in a set of N; in both cases, all N images $\{\boldsymbol{x}_i\}_{i=1,N}$ are analyzed jointly ("collaboratively"). Each $\boldsymbol{x}_i$ is assumed to be represented as a linear combination of a sparse set of atoms from a dictionary $\mathbf{D} \in \mathbb{R}^{P\times K}$, where the columns of $\mathbf{D}$ represent dictionary atoms. A prior is placed on $\mathbf{D}$, and a posterior density function on $\mathbf{D}$ is to be learned *in situ*, based on $\{\boldsymbol{x}_i\}_{i=1,N}$ (no additional training data). Further, the size of the dictionary (total number of active atoms across all $\boldsymbol{x}_i$) is unknown, and to be inferred. Specifically, $\boldsymbol{x}_i = \mathbf{D}\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\alpha}_i$ is sparse and $\boldsymbol{\epsilon}_i$ represents noise or residual. A prior is placed on $\{\boldsymbol{\epsilon}_i\}_{i=1,N}$, and the statistics of the residual/noise are also to be inferred. In recent research, it has been demonstrated that the beta process (BP) and Bernoulli process (BeP) may be coupled to constitute a prior on $\{\boldsymbol{\alpha}_i\}_{i=1,N}$ and $\mathbf{D}$ (Zhou et al., 2009), to impose the desired sparseness and to infer the dictionary size and composition.

One may wish to impose a prior belief that samples (here $\{\boldsymbol{x}_i\}_{i=1,N}$) with similar covariates are likely to employ the same or similar factors; this is related to

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

previous work on joint sparse analysis of multiple data vectors, but in that work (Chen and Huo, 2006; Mishali and Eldar, 2008; Tropp, 2006) covariates were not explicitly employed. To address this challenge, we develop a new model, termed the dependent hierarchical beta process (dHBP), and relate it (via the Bernoulli process) to a dependent hierarchical IBP.

Using the IBP metaphor, the proposed model imposes that data ("customers") that are far away from each other in covariate space interact and possibly share atoms/parameters ("dishes") via a "global buffet"; this global buffet is manifested as in a conventional IBP, and reflects dish popularity across all data. However, customers that are closer to each other in covariate space tend to have more sharing, manifested by "local buffets" that reflect popularity of dishes in local neighborhoods of covariate space.

1.1 Related Models

The phylogenetic IBP removes the assumption of sample exchangeability by imposing prior knowledge on inter-sample relationships via a tree structure (Miller et al., 2008). The form of the tree may be constituted as a result of covariates that are available with the samples, but the tree is not necessarily unique, and therefore it may be desirable in some applications to design a model based on the covariates directly. Toward this end, a dependent IBP (dIBP) model has been introduced recently, with a hierarchical Gaussian process (GP) used to account for covariate dependence (Williamson et al., 2010), with the covariates embedded into the covariance matrix through a kernel. For the problem of interest here, when considering a potential set of K features and N samples, one need perform K N-dimensional GP draws, which may be expensive for the large N and K of interest in large-scale applications. The proposed model is most related to Bayesian density regression (BDR) (Dunson et al., 2007). The original BDR construction was developed for models based on the Dirichlet process (DP), and here we extend it to a beta-Bernoulli process construction. Additionally, we here develop a covariate-dependent *hierarchical* beta process, related to the work in (Thibaux and Jordan, 2007), while (Dunson et al., 2007) did not consider a hierarchical DP.

1.2 Contributions

In this paper a new dependent hierarchical beta process is developed, its properties analyzed, and an efficient framework is constituted for analysis of large-scale problems. Efficient inference is performed by integrating Gibbs, Metropolis-Hastings and slice sampling. The model is applied to image interpolation and denoising, where in this case the samples correspond to patches of pixels within the image, and the covariates are linked to the position of the patches within the overall image. For the denoising applications we jointly consider two types of noise: (i) traditional white Gaussian noise; and (ii) sparse spiky noise of arbitrary amplitude, situated uniformly at random within the image. The noise considered in (ii) generalizes ideas from robust principal components analysis (PCA) (Candès et al., 2011; Chandrasekaran et al., 2009; Wright et al., 2009) to a new class of problems.

2 PRELIMINARIES

We review the BP, BeP and IBP, to set notation and to motivate the need to account for covariates; we follow (Thibaux and Jordan, 2007). A beta process $B \sim \mathrm{BP}(c, B_0)$ is a positive random measure on a space Ω, where c is a positive function over Ω, and B_0 is a fixed measure on Ω, called the base measure; we assume c is a constant. If B_0 is non-atomic, then a draw B may be represented as $B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}$, where ω_k are i.i.d. draws from $B/B(\Omega)$ and p_k are i.i.d. draws from a degenerate beta distribution with parameter c. If B_0 is discrete and of the form $B_0 = \sum_k q_k \delta_{\omega_k}$, then $B = \sum_k p_k \delta_{\omega_k}$, with $p_k \sim \mathrm{Beta}(cq_k, c(1-q_k))$. If B_0 is mixed discrete-continuous, B is the sum of the two independent contributions. We now consider a draw $X \sim \mathrm{BeP}(B)$ from a Bernoulli process, for measure B on Ω. If B is continuous, then $X = \sum_{k=1}^{K} \delta_{\omega_k}$, where $K \sim \mathrm{Poisson}(B(\Omega))$, and ω_k are i.i.d. draws from $B_0/B_0(\Omega)$. If B is discrete and of the form $B = \sum_k p_k \delta_{\omega_k}$, then $X = \sum_k b_k \delta_{\omega_k}$, where the $b_k \sim \mathrm{Bernoulli}(p_k)$ independently.

If we consider $B \sim \mathrm{BP}(c, B_0)$ and $X_i \sim \mathrm{BeP}(B)$, for $i = 1, \ldots, n$, then the posterior distribution of B is

$$B|\{X_i\}_{i=1,n} \sim \mathrm{BP}\bigg(c+n, \frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} X_i\bigg)$$

Hence, the BP is the conjugate prior for the BeP. Further, by integrating out B

$$X_{n+1}|\{X_i\}_{i=1,n} \sim \mathrm{BeP}\bigg(\frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} X_i\bigg).$$

Note that $\frac{1}{c+n}\sum_{i=1}^{n} X_i = \sum_k \frac{m_{n,k}}{c+n}\delta_{\omega_k}$, where ω_k represent the unique dishes/atoms selected by the first n "customers" and $m_{n,k}$ represents the number of customers selecting the kth dish ω_k. Additionally, $X_{n+1}|\{X_i\}_{i=1,n} = U + V$, where $U \sim \mathrm{BeP}(\frac{c}{c+n}B_0)$ and $V \sim \mathrm{BeP}(\frac{1}{c+n}\sum_{i=1}^{n} X_i)$. Thibaux and Jordan (Thibaux and Jordan, 2007) explicitly relate the above construction to the Indian buffet process (Griffiths and Ghahramani, 2005).

Note that the posteriors $p(B|\{X_i\}_{i=1,n})$ and $p(X_{n+1}|\{X_i\}_{i=1,n})$ are only dependent on the

count of atom usage among $\{X_i\}_{i=1,n}$, and therefore the order of the $\{X_i\}_{i=1,n}$ may be interchanged with no change in the posterior probabilities. This exchangeability assumption is inappropriate in many applications, and motivates the proposed model.

3 DEPENDENT HIERARCHICAL BETA PROCESS

3.1 Model Constructon

We consider data $\{\boldsymbol{x}_i\}_{i=1,N}$ with $\boldsymbol{x}_i \in \mathbb{R}^P$, and our objective is again to associate a set of features represented by $X_i = \sum_k \delta_{\omega_k}$ with sample $\boldsymbol{x}_i$. We also now introduce a corresponding set of covariates $\{\boldsymbol{\ell}_i\}_{i=1,N}$. The form of the covariates is general, but to be explicit, we here assume $\boldsymbol{\ell}_i \in \mathbb{R}^{\mathcal{L}}$. The covariates are used to impose relationships between the N samples, summarized by a matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, where each row of $\mathbf{A}$ sums to one, and its (i,j)th component is

$$a_{ij} = \mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_j) \Big/ \sum_{j'=1}^{N} \mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_{j'}) \tag{1}$$

where $\mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_j)$ is a kernel which diminishes with increasing distance between $\boldsymbol{\ell}_i$ and $\boldsymbol{\ell}_j$ and has the properties $0 \le \mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_j) \le 1$ and $\mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_i) = 1$. The matrix $\mathbf{A}$ is analogous to the random-walk matrix associated with diffusion analysis methods (Nadler et al., 2005).

A measure B_i is associated with each sample i:

$$B_i = \sum_{j=1}^{N} a_{ij} B_j^* \ , \quad B_j^* \sim \mathrm{BP}(c_1, B) \ , \quad B \sim \mathrm{BP}(c_0, B_0) \tag{2}$$

where c_0 and c_1 are positive constants, and B_0 is a base measure on Ω. The latent feature vector $X_i \sim \mathrm{BeP}(B_i)$ is associated with each sample i. The relationship between the B_j^* and B is analogous to the hierarchical BP in (Thibaux and Jordan, 2007), but now the covariate-dependent sample inter-relationships are accounted for via the a_{ij}. To help elucidate its properties, we express the dHBP in an alternative manner, introducing the latent indicator Z_i, as

$$\begin{aligned} X_i \sim \mathrm{BeP}(B_{Z_i}^*) \ , \quad & Z_i \sim \textstyle\sum_{j=1}^{N} a_{ij}\delta_j \ , \\ B_j^* \sim \mathrm{BP}(c_1, B) \ , \quad & B \sim \mathrm{BP}(c_0, B_0). \end{aligned} \tag{3}$$

If we marginalize out $\{B_i^*\}_{i=1,N}$ and B, the $\{X_i\}_{i=1,N}$ may be generated as follows, as a covariate-dependent generalization of the Indian buffet process (IBP).

For X_1, which is associated with $\boldsymbol{x}_1$, we first draw $Z_1 \sim \sum_{j=1}^{N} a_{1j}\delta_j$; X_1 is drawn from $\mathrm{BeP}(B_{Z_1}^*)$, meaning $M_1 \sim \mathrm{Poisson}(B_0(\Omega))$ atoms $\{\omega_k\}_{k=1,M_1}$ are drawn from the base measure B_0, and $X_1 = \sum_{k=1}^{M_1} \delta_{\omega_k}$. We now have

$$\begin{aligned} B_i^* | i \neq Z_1 &\sim \mathrm{BP}(c_1, B) \\ B_{Z_1}^* | X_1 &\sim \mathrm{BP}\Big(c_1 + 1, \frac{c_1}{c_1+1} B + \frac{1}{c_1+1} X_1\Big) \\ B | X_{g1} &\sim \mathrm{BP}\Big(c_0 + 1, \frac{c_0}{c_0+1} B_0 + \frac{1}{c_0+1} X_{g1}\Big), \end{aligned}$$

where $X_1 = X_{g1}$. For X_2, we draw $Z_2 \sim \sum_{j=1}^{N} a_{2j}\delta_j$. If $Z_2 \neq Z_1$, then $X_2 \sim \mathrm{BeP}(\frac{c_0}{c_0+1}B_0 + \frac{1}{c_0+1}X_{g1})$. In this case the M_1 atoms in X_{g1} are selected for inclusion in X_2 via i.i.d. sampling from Bernoulli$(\frac{1}{c_0+1})$. Further, $M_2 \sim \mathrm{Poisson}(\frac{c_0}{c_0+1}B_0(\Omega))$ new atoms are drawn i.i.d. from B_0; X_2 is a sum of the selected atoms from X_{g1} as well as the new draw of M_2 atoms, and in this case again $X_2 = X_{g2}$. On the other hand, if $Z_2 = Z_1$, we need to distinguish dishes selected off the "local" buffet from those selected off the "global" buffet. Since $X_2 \sim \mathrm{BeP}(\frac{c_1}{c_1+1}B + \frac{1}{c_1+1}X_1)$, in this case X_2 selects from among the M_1 atoms of X_1 on the "local" buffet, these drawn i.i.d. as Bernoulli$(\frac{1}{c_1+1})$; it also selects $M_2 \sim \mathrm{Poisson}(\frac{c_1}{c_1+1}\frac{c_0}{c_0+1}B_0(\Omega))$ new atoms drawn i.i.d. from B_0; finally, it selects from among the atoms in X_{g1} on the "global" buffet, drawn i.i.d. Bernoulli$(\frac{c_1}{c_1+1}\frac{1}{c_0+1})$. The vector X_2 is represented as a sum of *all* selected atoms, as well as the new set of M_2 atoms; X_{g2} corresponds only to the new atoms and the ones selected from among the atoms in X_{g1}. After doing this N times, we have

$$B_i^* | \mathbf{X}_N, \mathbf{X}_{gN}, \mathbf{Z}_N \sim$$
$$\mathrm{BP}\Big(c_1 + n_i, \tfrac{c_1}{c_1+n_i} B + \tfrac{1}{c_1+n_i} \textstyle\sum_{j=1}^{N} 1(Z_j = i) X_j\Big)$$
$$B | \mathbf{X}_{gN} \sim \mathrm{BP}\Big(c_0 + N, \frac{c_0}{c_0+N} B_0 + \frac{1}{c_0+N} \sum_{j=1}^{N} X_{gj}\Big)$$

where $\mathbf{X}_N = \{X_1, \ldots, X_N\}$, $\mathbf{X}_{gN} = \{X_{g1}, \ldots, X_{gN}\}$, $\mathbf{Z}_N = \{Z_1, \ldots, Z_N\}$, $n_i = \sum_{j=1}^{N} 1(Z_j = i)$, and $1(\cdot)$ is equal to one if the expression inside the brackets is true, and it is zero otherwise. The $\frac{1}{c_1+n_i}\sum_{j=1}^{N} 1(Z_j = i)X_j$ defines the "local" buffet of atoms at node i, and the probability of selecting each atom; $\frac{1}{c_0+N}\sum_{j=1}^{N} X_{gj}$ similarly defines the "global" buffet, constituted as all N local buffets are formed. The global buffet is still exchangeable, as in the original IBP, but the local buffets change with index i, removing exchangeability.

3.2 Handling New Samples

Assume we employ the dHBP to analyze data $\{\boldsymbol{x}_i\}_{i=1,N}$, and we now wish to use this model to infer the appropriate set of features/atoms for new sample $\boldsymbol{x}_{N+1}$, with associated covariate $\boldsymbol{\ell}_{N+1}$. We wish to do this without having to go back and analyze $\{\boldsymbol{x}_i\}_{i=1,N+1}$, as before, but rather we wish to directly use the result from the previous analysis of $\{\boldsymbol{x}_i\}_{i=1,N}$. For the specific examples we consider below all the $\{\boldsymbol{x}_i\}_{i=1,N}$ are known *a priori* and therefore this issue of handling new samples does not occur. However, the

subsequent analysis nevertheless provides insight into the model, and there are other applications for which one may wish to add a new sample $\boldsymbol{x}_{N+1}$.

Based upon the previous analysis of $\{\boldsymbol{x}_i\}_{i=1,N}$, assume access to the atoms used by each sample, this denoted $\mathbf{X}_N = \{X_i\}_{i=1,N}$, as well as the set of atoms $\mathbf{X}_{gN} = \{X_{gi}\}_{i=1,N}$ associated with the "global" buffet. Further, we *initially* assume access to $\mathbf{Z}_{N+1} = \{Z_i\}_{i=1,N+1}$. Then X_{N+1}, associated with $\boldsymbol{x}_{N+1}$, is drawn

$$X_{N+1}|\mathbf{X}_N, \mathbf{X}_{gN}, \mathbf{Z}_{N+1} \sim \text{BeP}\left(\frac{c_1 B_N}{c_1+\sum_{i=1}^N 1(Z_i=Z_{N+1})} + \frac{\sum_{i=1}^N 1(Z_i=Z_{N+1})X_i}{c_1+\sum_{i=1}^N 1(Z_i=Z_{N+1})}\right) \quad (4)$$

where $B_N = \frac{c_0}{c_0+N}B_0 + \frac{1}{c_0+N}\sum_{i=1}^N X_{gi}$.
Assume $N \gg c_0$, such that $\frac{c_0}{c_0+N}B_0(\Omega) \approx 0$, and therefore the probability of drawing new atoms for representation of $\boldsymbol{x}_{N+1}$ is negligibly small. Assuming we know $\mathbf{Z}_{N+1}$ and assuming $\frac{c_0}{c_0+N}B_0(\Omega) = 0$, we therefore have

$$\begin{aligned} X_{N+1}|\mathbf{X}_N, \mathbf{X}_{gN}, \mathbf{Z}_{N+1} &= U + V \\ U \sim \text{BeP}&\left(\frac{c_1}{c_1+\sum_{i=1}^N 1(Z_i=Z_{N+1})}\frac{1}{c_0+N}\sum_{i=1}^N X_{gi}\right) \\ V \sim \text{BeP}&\left(\frac{\sum_{i=1}^N 1(Z_i=Z_{N+1})X_i}{c_1+\sum_{i=1}^N 1(Z_i=Z_{N+1})}\right) \end{aligned} \quad (5)$$

Equation (5) provides further insight into the model. Specifically, if $\boldsymbol{x}_{N+1}$ is in a neighborhood of many members of $\{\boldsymbol{x}_i\}_{i=1,N}$ (*i.e.*, if c_1 is small relative to $\sum_{i=1}^N 1(Z_i = Z_{N+1})$), then $\frac{c_1}{c_1+\sum_{i=1}^N 1(Z_i=Z_{N+1})}$ is likely to be small, and therefore U is unlikely to contribute atoms for X_{N+1}. In this case V will dominate, it corresponding to a buffet of atoms that are popular within a neighborhood of $\boldsymbol{x}_{N+1}$, and these atoms are more probable to be selected by $\boldsymbol{x}_{N+1}$ if they are popular within the neighborhood. By contrast, if $\boldsymbol{x}_{N+1}$ is isolated from the samples in $\{\boldsymbol{x}_i\}_{i=1,N}$, then it is expected that $\sum_{i=1}^N 1(Z_i = Z_{N+1})$ will be small or zero, in which case the atoms associated with $\boldsymbol{x}_{N+1}$ will be constituted primarily from U, which corresponds to the global buffet.

Using (1), we may compute $a_{N+1,j}$ for all $j \in \{1,\ldots,N+1\}$, with $\sum_{j=1}^{N+1} a_{N+1,j} = 1$. Doing this, we may now marginalize out $\mathbf{Z}_{N+1}$ in (5). However, we are not interested explicitly in what the values of $\{Z_i\}_{i=1,N}$ are, only in whether $Z_i = Z_{N+1}$. We therefore marginalize out Z_i, and consider all possible outcomes for $1(Z_i = Z_{N+1})$. Toward this end, let $m_i = 1$ if $Z_i = Z_{N+1}$, while $m_i = 0$ otherwise. Then upon marginalizing out $\mathbf{Z}_{N+1}$, we have

$$X_{N+1}|\mathbf{X}_N, \mathbf{X}_{gN} \sim \sum_{j=1}^{N+1}\sum_{m_1=0}^1 \cdots \sum_{m_N=0}^1 p(j, m_1, \ldots, m_N) \text{BeP}\left(\frac{c_1}{c_1+\sum_{i=1}^N m_i}\frac{1}{c_0+N}\sum_{i=1}^N X_{gi} + \frac{\sum_{i=1}^N m_i X_i}{c_1+\sum_{i=1}^N m_i}\right)$$

where

$$p(j, m_1, \ldots, m_N) = a_{N+1,j} a_{1j}^{m_1}(1-a_{1j})^{1-m_1}\cdots a_{Nj}^{m_N}(1-a_{Nj})^{1-m_N} \quad (6)$$

Note that the expression for $X_{N+1}|\mathbf{X}_N, \mathbf{X}_{gN}$ involves 2^N summation terms for each $j \in \{1,\ldots,N+1\}$. However, typically only a small number of terms need actually be computed, as $a_{N+1,j}$ will only be non-negligible for j for which the associated $\boldsymbol{x}_j$ are within a neighborhood of $\boldsymbol{x}_{N+1}$. Further, it is only probable that $m_i \neq 0$ when $\boldsymbol{x}_i$ is within a neighborhood of $\boldsymbol{x}_{N+1}$, and therefore one only need consider sums over m_i for i associated with $\boldsymbol{x}_i$ in a neighborhood of $\boldsymbol{x}_{N+1}$. Therefore, in practice only a small set of the sums need actually be computed, this constituting an efficient prior for the atoms needed to model the new sample $\boldsymbol{x}_{N+1}$.

3.3 Covariate-Dependent Correlations

Theorem 1: For any measurable subset S and $\boldsymbol{x}_i, \boldsymbol{x}_{i'} \in \mathbb{R}^P$, B_i and $B_{i'}$ are dependent random probability measures, with

$$\text{corr}\{B_i(S), B_{i'}(S)\} = \frac{<\boldsymbol{a}_i, \boldsymbol{a}_{i'}>}{\|\boldsymbol{a}_i\| \cdot \|\boldsymbol{a}_{i'}\|} \quad (7)$$

where $\boldsymbol{a}_i = [a_{i1}, \ldots, a_{iN}]^T$. □

Proof: The correlation between $B_i(S)$ and $B_{i'}(S)$ can be expressed as

$$\text{corr}\{B_i(S), B_{i'}(S)\} = \frac{E\{B_i(S)B_{i'}(S)\} - E\{B_i(S)\}E\{B_{i'}(S)\}}{[V\{B_i(S)\}V\{B_{i'}(S)\}]^{1/2}}.$$

Following (Dunson et al., 2007), the numerator can be simplified as $\sum_{j=1}^N a_{ij}a_{i'j}V\{B_j^*(S)\}$. Since $V\{B_i^*(S)\} = V\{B_j^*(S)\}$ for $i \neq j$, $V\{B_i(S)\}$ can be expressed as $V\{B_i(S)\} = \|\boldsymbol{a}_i\|^2 V\{B_j^*(S)\}$. Hence, $\text{corr}\{B_i(S), B_{i'}(S)\}$ may be expressed as

$$\frac{\sum_{j=1}^N a_{ij}a_{i'j}V\{B_j^*(S)\}}{\left[\|\boldsymbol{a}_i\|^2 V\{B_i^*(S)\}\|\boldsymbol{a}_{i'}\|^2 V\{B_{i'}^*(S)\}\right]^{1/2}} = \frac{<\boldsymbol{a}_i, \boldsymbol{a}_{i'}>}{\|\boldsymbol{a}_i\| \cdot \|\boldsymbol{a}_{i'}\|}. \quad \square$$

4 DICTIONARY LEARNING FOR INTERPOLATION & DENOISING

4.1 Dictionary Learning with Beta Process

When *not* considering the covariate dependence (assuming $\{\boldsymbol{x}_i\}_{i=1,N}$ are exchangeable), we draw the dictionary atoms and the associated atom usage probabilities from a (truncated) beta process (BP) and link them to a sample's atom usage via a Bernoulli process (BeP). The data $\boldsymbol{x}_i$ is represented as

$$\boldsymbol{x}_i = \mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i) + \boldsymbol{\epsilon}_i \tag{8}$$

where $\odot$ represents the Hadamard product, $\mathbf{D} = [\boldsymbol{d}_1, \cdots, \boldsymbol{d}_K]$, $\boldsymbol{s}_i = [s_{i1}, \cdots, s_{iK}]^T$, $\boldsymbol{z}_i = [z_{i1}, \cdots, z_{iK}]^T$, $s_{ik} \in \mathbb{R}$, $z_{ik} \in \{0, 1\}$ indicates whether the kth atom is *active* within sample i, and $\boldsymbol{\epsilon}_i$ is the residual error; the truncation level K is set large enough such that not all K atoms are actually used in the representation of the data $\{\boldsymbol{x}_i\}_{i=1,N}$, and in this sense the size of active dictionary elements is inferred. The binary indicator $z_{ik} = X_i(\boldsymbol{d}_k)$ is drawn as $z_{ik} \sim \text{Bernoulli}(\pi_k)$ with $\pi_k = B(\boldsymbol{d}_k)$.

The hierarchical form of the model is

$$\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i), \gamma_\epsilon^{-1}\mathbf{I}_P) \tag{9}$$

$$\boldsymbol{d}_k \sim \mathcal{N}(0, P^{-1}\mathbf{I}_P), \ \boldsymbol{s}_i \sim \mathcal{N}(0, \gamma_s^{-1}\mathbf{I}_K) \tag{10}$$

$$z_{ik} \sim \text{Bernoulli}(\pi_k), \ \pi_k \sim \text{Beta}(c\eta, c(1-\eta)) \tag{11}$$

where gamma hyper-priors are placed on both γ_ϵ and γ_s. The probability distribution $\mathcal{N}(0, P^{-1}I_P)$ plays the role of the base measure B_0; we set $\eta = 1/K$, and for large K $\text{Beta}(c/K, c(1-1/K))$ approximates the improper beta distribution with parameter c.

4.2 Dictionary Learning with dHBP

When employing the hierarchical construction in (2), equation (11) generalizes as

$$\begin{aligned} z_{ik} &\sim \text{Bernoulli}(\pi_{ik}), \quad \pi_{ik} = \sum_{j \in \mathcal{Q}_i} a_{ij}\pi^*_{jk} \\ \pi^*_{jk} &\sim \text{Beta}(c_1\eta_k, c_1(1-\eta_k)) \\ \eta_k &\sim \text{Beta}(c_0\eta_0, c_0(1-\eta_0)). \end{aligned} \tag{12}$$

where $z_{ik} = X_i(\boldsymbol{d}_k)$, $\pi_{ik} = B_i(\boldsymbol{d}_k)$, $\pi^*_{jk} = B^*_j(\boldsymbol{d}_k)$ and $\eta_k = B(\boldsymbol{d}_k)$.

4.3 Implementation

In the application of image interpolation and denoising, we are given an incomplete and noisy image of size $M_x \times M_y$, which is partitioned into N overlapping $W \times W$ patches $\{\boldsymbol{x}_i\}_{i=1,N}$, with $\boldsymbol{x}_i \in \mathbb{R}^P$, $P = W^2$ and $N = (M_x - W + 1) \times (M_y - W + 1)$; we typically set $W = 8$. In this case the ith patch is assumed to have an associated covariate vector $\boldsymbol{\ell}_i = [i_x, i_y]^T$, representing its spatial location in the original image. We consider the kernel function

$$\mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_j) = \delta(j \in \mathcal{Q}_i) \exp(-\|\boldsymbol{\ell}_i - \boldsymbol{\ell}_j\|_2/\sigma) \tag{13}$$

where $\mathcal{Q}_i = \{j : \|\boldsymbol{\ell}_i - \boldsymbol{\ell}_j\|_2 \leq L\}$ is a pre-defined spatial neighborhood of i, $\delta(\cdot) = 1$ if the argument is true and it is zero elsewise, and σ is the kernel width. We can calculate a_{ij} by (1) and we have $a_{ij} \neq 0$ if and only if $j \in \mathcal{Q}_i$. This way of defining neighborhoods is similar to that used in Isomap (Tenenbaum et al., 2000).

4.3.1 Missing Pixels

With missing pixels, we observe $\boldsymbol{y}_i = \boldsymbol{\Sigma}_i\boldsymbol{x}_i$, where $\boldsymbol{\Sigma}_i$ is the sampling matrix, constructed by selecting a subset of rows from the identity matrix $\mathbf{I}_P$. The matrix $\boldsymbol{\Sigma}_i$ is a function of the patch index i, and $\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_i^T = \mathbf{I}_{\|\boldsymbol{\Sigma}_i\|_0}$. With $\boldsymbol{\Sigma}_i^T\boldsymbol{\Sigma}_i\bar{\boldsymbol{\Sigma}}_i^T\bar{\boldsymbol{\Sigma}}_i = \mathbf{0}_P$, where $\bar{\boldsymbol{\Sigma}}_i$ is the sampling matrix for missing pixels in patch i (it identifies the missing-pixel locations), the likelihood term can be expressed as

$$\begin{aligned} &\mathcal{N}(\boldsymbol{x}_i; \mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i), \gamma_\epsilon^{-1}\mathbf{I}_P) = \\ &\mathcal{N}(\boldsymbol{\Sigma}_i^T\boldsymbol{y}_i; \boldsymbol{\Sigma}_i^T\boldsymbol{\Sigma}_i\mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i), \boldsymbol{\Sigma}_i^T\boldsymbol{\Sigma}_i\gamma_\epsilon^{-1}\mathbf{I}_P) \\ &\mathcal{N}(\bar{\boldsymbol{\Sigma}}_i^T\bar{\boldsymbol{\Sigma}}_i\boldsymbol{x}_i; \bar{\boldsymbol{\Sigma}}_i^T\bar{\boldsymbol{\Sigma}}_i\mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i), \bar{\boldsymbol{\Sigma}}_i^T\bar{\boldsymbol{\Sigma}}_i\gamma_\epsilon^{-1}\mathbf{I}_P). \end{aligned} \tag{14}$$

Thus the missing-pixel values can be integrated out analytically, and one may only consider the observed $\boldsymbol{y}_i$ when performing inference.

4.3.2 Sparse Spiky Noise

The sparse spiky noise may exist at any pixel, uniformly at random, and the spike amplitude may be large. The presence of a spike at a given pixel, particularly when the spike amplitude is large, is similar to the case of missing pixels discussed above, as the original pixel value is essentially lost. The complexity of this problem is that the location of the spike is assumed unknown, and must be inferred by the model. We add a sparse noise term $\boldsymbol{v}_i \odot \boldsymbol{m}_i$ to the model as

$$\boldsymbol{x}_i = \mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i) + \boldsymbol{\epsilon}_i + \boldsymbol{v}_i \odot \boldsymbol{m}_i \tag{15}$$

where $\boldsymbol{v}_i = [v_{i1}, \cdots, v_{iP}]^T$ and $\boldsymbol{m}_i = [m_{i1}, \cdots, m_{iP}]^T$, $v_{ip} \in \mathbb{R}$ and $m_{ip} \in \{0, 1\}$ is the binary indicator. A beta-Bernoulli prior is constituted on $\boldsymbol{m}_i$ as

$$m_{ip} \sim \text{Bernoulli}(\pi'_{ip}), \quad \pi'_{ip} \sim \text{Beta}(a_0, b_0). \tag{16}$$

Additionally, we impose $\boldsymbol{v}_i \sim \mathcal{N}(0, \gamma_v^{-1}\mathbf{I}_P)$ with a gamma hyper-prior on γ_v, allowing inference of the spike level. After performing analysis with this model, the noise free data is estimated as $\hat{\boldsymbol{x}}_i = \mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i)$. In the experiments $a_0 = 1$ and $b_0 = 100$.

The model developed here, with a usually overcomplete dictionary, generalizes the assumptions of robust PCA (Wright et al., 2009) (which assumes low-rank), widening the class of problems for which sparse spiky noise models are appropriate.

4.4 Inference

The inference is performed using MCMC analysis. Update equations not associated with the BP model discussed in Zhou et al. (2009), are summarized below.

Sample π^*_{jk}. Denote $\pi^{-j}_{ik} = \sum_{\ell \neq j} a_{i\ell}\pi^*_{\ell k} = \pi_{ik} - a_{ij}\pi^*_{jk}$, the posterior of π^*_{jk} can be expressed as

$$\begin{aligned} p(\pi^*_{jk}|-) \propto \text{Beta}\big(\pi^*_{jk}; c_1\eta_k, c_1(1-\eta_k)\big) \\ \prod_{i:\{j \in \mathcal{Q}_i\}} \text{Bernoulli}(z_{ik}; a_{ij}\pi^*_{jk} + \pi^{-j}_{ik}) \end{aligned} \tag{17}$$

which cannot be directly sampled from. However, we

can connect π_{jk}^* with the popularity of the kth atom $\boldsymbol{d}_k$ in the neighborhood of patch j by assuming

$$\pi_{jk}^* \sim \text{Beta}\Big(c_1\eta_k + \sum_{i:\{j\in\mathcal{Q}_i\}} z_{ik}, c_1(1-\eta_k) + \sum_{i:\{j\in\mathcal{Q}_i\}}(1-z_{ik})\Big). \tag{18}$$

We use this as the proposal distribution Q in a Metropolis-Hastings (M-H) independence chain (Hastings, 1970) and accept ${\pi_{jk}^*}^{t+1} = {\pi_{jk}^*}'$ with probability $\min\{p_{jk}, 1\}$, where

$$\begin{aligned} p_{jk} &= \frac{p({\pi_{jk}^*}')}{p({\pi_{jk}^*}^t)}\frac{Q({\pi_{jk}^*}^t)}{Q({\pi_{jk}^*}')} \\ &= \Big(\frac{{\pi_{jk}^*}^t}{{\pi_{jk}^*}'}\Big)^{\sum_{i:\{j\in\mathcal{Q}_i\}} z_{ik}}\Big(\frac{1-{\pi_{jk}^*}^t}{1-{\pi_{jk}^*}'}\Big)^{\sum_{i:\{j\in\mathcal{Q}_i\}}(1-z_{ik})} \\ &\quad \prod_{i:\{j\in\mathcal{Q}_i\}}\Big(1+\frac{\Delta_{jk}}{\pi_{ik}{}^t}\Big)^{z_{ik}}\Big(1-\frac{\Delta_{jk}}{1-\pi_{ik}{}^t}\Big)^{1-z_{ik}} \end{aligned} \tag{19}$$

where $\Delta_{jk} = a_{ij}({\pi_{jk}^*}' - {\pi_{jk}^*}^t)$.

Sample η_k. The posterior of η_k can be expressed as

$$\begin{aligned} p(\eta_k|-) \;\propto\; & \text{Beta}\big(\eta_k; c_0\eta_0, c_0(1-\eta_0)\big) \\ & \prod_{j=1}^{N}\text{Beta}\big(\pi_{jk}^*; c_1\eta_k, c_1(1-\eta_k)\big). \end{aligned} \tag{20}$$

Considering the special case of $c_1 = 1$ and with the Euler's reflection formula $\Gamma(1-x)\Gamma(x) = \pi/\sin(\pi x)$, we have

$$\begin{aligned} p(\eta_k|-) \;\propto\; & \eta_k^{c_0\eta_0-1}(1-\eta_k)^{c_0(1-\eta_0)-1}\sin^N(\pi\eta_k) \\ & \exp\Big(c_1\eta_k\sum_{j=1}^{N}\log\Big(\frac{\pi_{jk}^*}{1-\pi_{jk}^*}\Big)\Big). \end{aligned} \tag{21}$$

With slice sampling (Damlen et al., 1999), we let

$$\begin{aligned} u_k \sim \text{Unif}\big(0, \eta_k^{c_0\eta_0-1}\big), \quad w_k \sim \text{Unif}\big(0, \sin^N(\pi\eta_k)\big) \\ v_k \sim \text{Unif}\big(0, (1-\eta_k)^{c_0(1-\eta_0)-1}\big) \end{aligned} \tag{22}$$

and then draw η_k from the truncated exponential distribution as

$$\eta_k \sim \text{Exp}\Big(-c_1\sum_{j=1}^{N}\log\Big(\frac{\pi_{jk}^*}{1-\pi_{jk}^*}\Big)\Big)\text{I}(\eta_k). \tag{23}$$

where $\text{I}(\eta_k)$ represents the range of η_k derived from (22). We use a default value $c_1 = 1$ in this paper which gives a flexible prior for π_{jk}^* and also obtains efficient computation. Other values of c_1 can also be used and the random walk M-H can be applied.

Sample v_{ip} ***and*** m_{ip}. It is easy to show that v_{ip} and m_{ip} can be sampled as

$$\begin{aligned} v_{ip} &\sim \mathcal{N}\big(\gamma_\epsilon m_{ip}\bar{x}_{ip}(\gamma_v+\gamma_\epsilon m_{ip}^2)^{-1}, (\gamma_v+\gamma_\epsilon m_{ip}^2)^{-1}\big) \\ m_{ip} &\sim \text{Bernoulli}\Big(\frac{\pi'_{ip}\exp[-\frac{\gamma_\epsilon}{2}(-2\bar{x}_{ip}v_{ip}+v_{ip}^2)]}{1-\pi'_{ip}+\pi'_{ip}\exp[-\frac{\gamma_\epsilon}{2}(-2\bar{x}_{ip}v_{ip}+v_{ip}^2)]}\Big) \end{aligned}$$

where $\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i - \mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i)$ and $\bar{x}_{ip}$ is its pth element.

5 EXPERIMENTAL RESULTS

In the analysis that follows, 2500 MCMC iterations are used (2000 burn-in and 500 collection, from a random start). Each MCMC sample produces an image estimate. The results reported here are based on an average of the collection samples. Additionally, the pixel value for each MCMC sample is the average across 64 patches, from the 64 overlapping patches in which each pixel resides (except for near the image edges). Similar image-analysis results were found using as few as 250 MCMC iterations, and using just the last MCMC sample (again averaging patches to constitute the value of each pixel). A quantitative comparison between using 2500 and 250 MCMC iterations is discussed at the end of this section. For dHBP, the M-H acceptance rates were found to be greater than 90%. Finally, in all examples no parameter tuning has been performed; the gamma hyper-priors placed on all precision terms throughout the models were set as $\text{Gamma}(10^{-6}, 10^{-6})$.

Before proceeding, we note that most nonparametric Bayesian models are under-identified in a frequentist sense, as we have more parameters than data points and certainly cannot obtain unique maximum likelihood estimates of these parameters. Many of the nonparametric and rich parametric models in the literature are over-parameterized in this sense (*e.g.*, (Dunson et al., 2007)), and there are many different regions of the parameter space corresponding to similar likelihoods on observables. For this reason, if we monitor latent quantities not on the observed data level, we may obtain poor mixing. However, this is not a problem in conducting inferences on identifiable quantities (*e.g.*, the observed data density, at the layer of the image in our studies) and indeed for over-parameterized hierarchical models one often obtains excellent mixing for identifiable quantities (see (Bhattacharya and Dunson, 2011) for a discussion); we observed this behavior in the proposed model.

5.1 Images with Missing Pixels and WGN

We assume $\{\boldsymbol{x}_i\}_{i=1,N}$ represent data from (overlapping) patches from a single image, with a subset of pixels missing uniformly at random. The interpolation objective is to recover the missing pixels, with this also to be performed in the presence of additive noise (for this problem the noise is assumed to *not* be spiky). For this application, the model parameters are set as $L = 3$ (*i.e.*, 28 spatial neighbors), $c_0 = 10$, $c_1 = 1$, $\sigma = 5$, and $\eta_0 = 0.5$. The dictionary size K is set as 256 or 512, depending on the size of the image. The dictionary atoms are initialized at random.

We compare dHBP results with those of BP (Zhou et al., 2009) (which assumes exchangeability of the

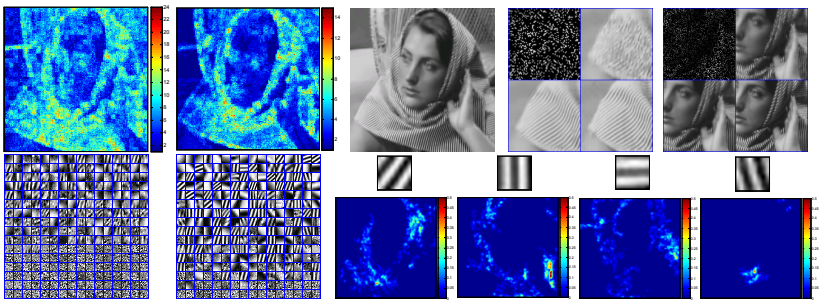

Figure 1: Comparison of interpolation results, considering BP (PSNR 26.90 dB) and dHBP (PSNR 29.92 dB) on the Barbara256 image, with 80% of its pixels missing uniformly at random. In the first row, the left two images show the spatially-dependent number of atoms $\|X_i\|_0$ used for representation of the patches throughout the image, as computed by BP and dHBP, respectively, the third is the dHBP reconstruction, and the fourth and fifth images show two different enlarged regions (top left, top right, bottom left and bottom right quarters corresponding to the original image under test, the BP reconstruction, the dHBP reconstruction, and the original versions, respectively). In the second row, the first two images show the dictionaries (the atoms are ordered based on their probabilities to be selected) inferred by BP and dHBP, respectively, and the third to sixth images show four dictionary atoms (resized from the size of 8×8 to 80×80 for visualization) and the associated atom activation probabilities across the image (each patch has a corresponding π_{ik}).

Figure 2: Comparison of interpolation results, considering BP and dHBP on the 512×512 Boat and Hill images, with 80% of their pixels missing uniformly at random. The left-most and third images show the dHBP reconstructions of Boat and Hill, respectively. The second and fourth images show two enlarged regions as in Figure 1.

Table 1: Gray-scale image interpolation results (PSNR) for BP and dHBP, both using patch size 8×8. The top and bottom rows of each cell show the results of BP and dHBP, respectively.

ratio	C.man	House	Peppers	Lena	Barbara	Boats	F.print	Man	Couple	Hill
20%	24.11	30.12	25.92	31.00	24.80	27.81	26.03	28.24	27.72	29.33
	24.43	**32.23**	**27.06**	**32.00**	**29.51**	**28.66**	**26.80**	**28.86**	**28.55**	**29.94**
30%	25.71	33.14	28.19	33.31	27.52	30.00	29.01	30.06	30.00	31.21
	26.50	**35.64**	**29.30**	**34.23**	**32.29**	**30.90**	**29.23**	**30.76**	**30.65**	**31.67**
50%	28.90	38.02	32.58	36.94	33.17	33.78	**33.53**	33.29	**35.56**	**34.23**
	29.89	**38.83**	**32.90**	**37.14**	**36.03**	**33.92**	32.70	**33.72**	33.54	34.14

Table 2: Joint image interpolation and denoising results (PSNR) for BP and dHBP, considering the Barbara256 image and patch size 8×8. The observed data ratio ranges from 20% to 50% and the noise standard deviation ranges from 0 to 25. The top and bottom rows of each cell show the results of BP and dHBP, respectively.

	0	5	10	15	20	25
20%	26.90	26.81	26.25	25.30	24.44	23.74
	29.92	**29.22**	**27.90**	**26.65**	**25.63**	**24.73**
30%	30.01	29.73	28.38	27.00	25.94	25.00
	32.49	**31.43**	**29.71**	**28.20**	**27.01**	**26.02**
50%	35.41	33.59	31.16	29.31	27.89	26.80
	36.83	**34.42**	**31.94**	**30.20**	**28.77**	**27.77**

patches). The BP results are similar to those produced by KSVD (Aharon et al., 2006; Elad and Aharon, 2006). In Fig. 1 we consider the 256×256 Barbara256 image, with 80% of its pixels missing uniformly at random. In Fig. 2 we show the comparison on the 512×512 Boat and Hill images. The dHBP yields sharp dictionary atoms, and the atom usage frequency map (spatial dependence of $\|X_i\|_0$) reflects the local complexity of the image. Note from Fig. 1 that the dHBP yields substantially more structured dictionary elements than BP, implying that dHBP better tailors dictionary elements to local structure in the image.

Quantitative comparisons between BP and dHBP on image interpolation are shown in Table 1. Quantitative comparisons for joint image interpolation and denoising on the Barbara256 image are shown in Ta-

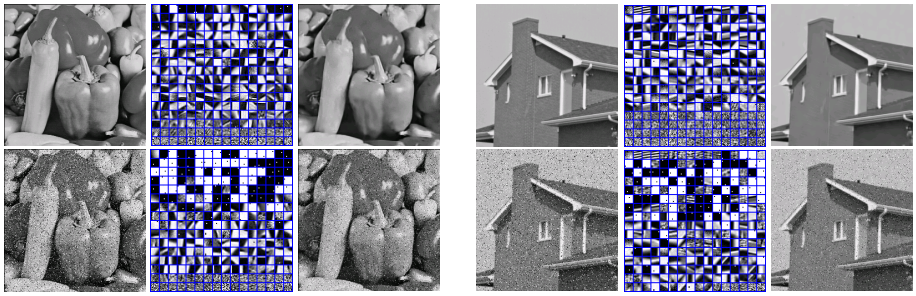

Figure 3: Left: Denoising results for BP (PSNR 18.53 dB) and dHBP (PSNR 29.69 dB) on the Peppers256 image, with 15% of its pixels corrupted by spiky noise situated uniformly at random. The spike amplitudes are uniformly distributed between -255 and 255. WGN with the standard deviation of 15 is also added to the image. Right: Denoising results for BP (PSNR 21.63 dB) and dHBP (PSNR 35.32 dB) on the House image with 10% of its pixels corrupted by spiky noise situated uniformly at random. The spike amplitudes are uniformly distributed between -255 and 255 at random. WGN with the standard deviation of 10 is also added to the image. In both the left and right parts, in the first column, the top and bottom images are the original and corrupted images, respectively. The second and third columns show the learned dictionaries (the atoms are ordered based on their probabilities to be selected) and the recovered images, respectively, with the top and bottom rows showing the results of dBHP and BP, respectively.

ble 2. The noise level is automatically estimated during the learning (in terms of the inverse precision of the noise $\boldsymbol{\epsilon}_i$). For example, the noise standard deviations are estimated by dHBP to be 11.91, 16.47, and 21.25 when the percentage of observed data and and the true noise standard deviation are respectively 20% and 10, 30% and 15, and 50% and 20 (similarly accurate estimates were found in all cases, and were also inferred via the BP model in (Zhou et al., 2009)). In the case for which 80% pixels of the pixels are missing at random, across the 10 test images considered here, the PSNR improvements range from about 0.3 dB to 4.7 dB, with the improvements mainly coming from regions of an image with detailed local structure, such as edges and striped patterns, as demonstrated in Figs. 1 and 2.

5.2 Images with Spiky and WGN Noise

We consider denoising an image corrupted simultaneously by additive noise of two types: (i) sparse spiky noise situated uniformly at random within the image, with amplitude distributed uniformly at random between -255 and 255; and (ii) WGN (with results shown with standard deviation 10 and 15). Comparison between BP and dHBP on the Peppers256 and House images under these two noise settings are shown in Fig. 3. The dHBP successfully separates out the sparse spiky noise, leading to a clean dictionary and an excellent restoration of the original image, while BP fails to do so, with BP yielding dictionary atoms that are severely corrupted by sparse spiky noise. We further tested the dHBP algorithm using joint priors for WGN and spiky noise, considering images only corrupted by sparse spiky noise or only WGN; we found that the algorithm performs well in both of these cases.

We have also considered using a heavy-tailed residual distribution for this spiky noise. Specifically, we considered $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \Gamma_i^{-1}\gamma_s^{-1})$, with $\Gamma_i = \text{diag}(\gamma_{i1}, \ldots, \gamma_{iP})$ and $\gamma_{ij} \sim \text{gamma}(\nu/2, \nu/2)$ with ν the degrees of freedom in the induced t-distribution for the residuals (one may place a prior on ν). The main point of this work is not which of these spiky noise models is selected, since both work well. The key is that we need to use spatial structure (covariates) when learning the model, otherwise it will learn spikes in the dictionary elements.

All algorithms have been implemented in non-optimized Matlab. As an example, when considering denoising a 256×256 image with $62,001$ 8×8 patches, considering joint WGN and spiky noise, BP required 24 seconds per Gibbs iteration, while dHBP required 55 seconds per MCMC iteration using a PC with 2.4 GHz CPU. As a representative example, for 20%, 30% and 50% observations in the image interpolation experiments of Table 1 , the average PSNRs are increased by about 0.4, 0.6 and 1.1 dB when considering 2500 MCMC iterations, as opposed to only 250.

6 CONCLUSIONS

A dependent hierarchical beta process (dHBP) is proposed and its properties are analyzed. Efficient hybrid MCMC inference including Gibbs, Metropolis-Hastings and slice sampling are presented. Encouraging performance is demonstrated on image-processing applications including interpolation, denoising and sparse spiky noise removal.

Acknowledgements

The research reported here was supported by AFOSR, ARO, DARPA, DOE, NGA, ONR and SERDP.

References

M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Trans. Signal Process.*, 2006.

A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 2011.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *accepted for publication in Journal of the ACM*, 2011.

V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A.S. Willsky. Sparse and low-rank matrix decompositions. In *Proc. Allerton conference on Communication, control, and computing*, 2009.

J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Trans. Sigal Process*, 2006.

P. Damlen, J. Wakefield, and S. Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society B Stat. Methodol*, 1999.

D. B. Dunson, N. S. Pillai, and J.-H. Park. Bayesian density regression. *Journal of the Royal Statistical Society B*, 2007.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, 2006.

T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Proc. Neural Information Processing Systems*, 2005.

W. Hastings. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 1970.

D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Proc. International Conference on Independent Component Analysis and Signal Separation*, 2007.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proc. International Conference on Machine Learning*, 2009.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Proc. Neural Information Processing Systems*, 2008.

K. Miller, T. Griffiths, and M. I. Jordan. The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. In *Proc. Conference on Uncertainty in Artificial Intelligence*, 2008.

M. Mishali and Y. C. Eldar. Reduce and boost: Recovering arbitrary sets of jointly sparse vectors. *IEEE Trans. Sigal Process.*, 2008.

B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In *Proc. Neural Information Processing Systems*, 2005.

J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *Proc. Int. Conf. Machine Learning*, 2009.

P. Rai and H. Daumé. The infinite hierarchical factor regression model. In *Proc. Neural Information Processing Systems*, 2008.

Y. W. Teh and D. Gorur. Indian buffet processes with power-law behavior. In *Proc. Neural Information Processing Systems*, 2009.

J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *Proc. International Conference on Artificial Intelligence and Statistics*, 2007.

J. Tropp. Algorithms for simultaneous sparse approximation. part II: Convex relaxation. *Signal Process.*, 2006.

S. Williamson, P. Orbanz, and Z. Ghahramani. Dependent Indian buffet processes. In *Proc. International Conference on Artificial Intelligence and Statistics*, 2010.

J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao. Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization. In *Proc. Neural Information Processing Systems*, 2009.

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