VARIATIONAL INFERENCE FOR NONPARAMETRIC SUBSPACE DICTIONARY LEARNING WITH HIERARCHICAL BETA PROCESS

Shaoyang Li, Xiaoming Tao, Jianhua Lu

Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ABSTRACT

Nonparametric Bayesian models have been implemented in dictionary learning. However, for signal samples from multiple subspaces, existing methods only learn one uniform dictionary and thus are not optimal for representing the subspace structures. To address this issue, we first utilize a combination of Dirichlet process and hierarchical Beta process as priors to infer the latent subspace number and dictionary dimension automatically; second, to derive tractable variational inference, we modify the priors with the Sethuraman's construction and further employ the multinomial approximation. Experimental results indicate that our approach can achieve a set of nonparametric subspace dictionaries, while showing performance enhancements in the tasks of image denoising.

Index Terms— Nonparametric Bayes, subspace dictionary learning, hierarchical Beta process, variational inference, image denoising.

1. INTRODUCTION

Dictionary learning builds a framework of seeking for appropriate atoms to sparsely represent high-dimensional signals (*e.g.*, images). The atoms used for representing the signals of interest are learned from the given data samples [1]–[3]. In contrast to the conventional dictionary learning algorithms which set fixed atom number in advance, recently nonparametric Bayesian methods arise to infer the required dictionary size by employing Beta-Bernoulli process [4]–[6].

Although the existing Bayesian strategies avoid predefining the atom numbers, they only learn one uniform dictionary even for the data samples belonging to different low-dimensional subspaces or manifolds [4]–[6]. For instance, however, if the data samples are small patches extracted from specific images, their patch textures may be grouped into different categories [7]. Thus, by exploring the structure of given data and learning multiple dictionaries simultaneously, it is possible to better depict the characteristics of the latent subspaces. Some researches have utilized a clustering techniques to choose different dictionaries for data points in different subspaces [7], [8]. Unfortunately, the algorithms considering multiple dictionaries have not considered Bayesian methods, and thus have to set fixed dictionary number and dimension. Furthermore, if we would like to learn subspace dictionaries using nonparametric Bayesian methods, the hierarchical Beta process (HBP) [9] is required. Nevertheless, posterior inference for HBP is intractable, hence existing approaches have to resort to the Markov chain Monte Carlo (MCMC) methods which require much time consumption [10], [11].

To tackle the above issues, we present two improvements in this paper. First, by using Dirichlet process (DP) [12], [13] and HBP together, we build a nonparametric Bayesian model to automatically infer the appropriate subspace number and dictionary dimensions. Second, to infer with the time-saving variational methods [14], [15], we modify the HBP prior via the Sethuraman's stick-breaking construction, and further develop a closed-form coordinate updating algorithm via the multinomial approximation. Finally, to evaluate the performance of our model and the inference strategy, we implement the proposed method in the image denoising tasks. Experimental results have indicated that our nonparametric subspace dictionary learning algorithm outperforms other state-of-the-art methods in the application of reconstructing noisy images.

2. NONPARAMETRIC BAYESIAN MODEL FOR SUBSPACE DICTIONARY LEARNING

Within the traditional dictionary learning framework, given a set of signal samples $\{\boldsymbol{x}_i\}_{i=1}^N$, one considers to learn a fixed-sized dictionary $\mathbf{D} \in \mathbb{R}^{P\times K}$ and a sparse weight vector $\boldsymbol{w}_i \in \mathbb{R}^K$ simultaneously to represent the samples. In this manner, the signal vector $\boldsymbol{x}_i \in \mathbb{R}^P$ can be expressed as $\boldsymbol{x}_i = \mathbf{D}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ denotes the residual noise.

However, for the signal samples $\{\boldsymbol{x}_i\}_{i=1}^N$ existing in multiple subspaces with distinct features, one fixed dictionary is not flexible enough to express the diversity, thus different dictionaries may be required. Assuming that the corresponding dictionary for $\boldsymbol{x}_i$ is denoted as $\mathbf{D}_{c(i)}$, the signal vector can be expressed in the form of

$$\boldsymbol{x}_i = \mathbf{D}_{c(i)}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $c(i)$ denotes the subspace index of $\boldsymbol{x}_i$. Based on this point, we build a Bayesian model and define prior distributions for the model parameters. As $\boldsymbol{x}_i$ is sparsely represented by the atoms of $\mathbf{D}_{c(i)}$, we introduce a binary indicator $\boldsymbol{z}_i \in$

This work was supported by the National Basic Research Project of China (973) (2013CB329006), and National Natural Science Foundation of China (NSFC, 61622110, 61471220, 91538107).

$\{0,1\}^K$, which determines the atoms of $\mathbf{D}_{c(i)}$ to be active or not, to induce the sparseness structure of $\boldsymbol{w}_i$ by

$$\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{D}_{c(i)}\boldsymbol{w}_i, \alpha_{c(i)}^{-1}\mathbf{I}_P), \tag{2}$$

$$\mathbf{D}_{c(i)} \sim \prod_{k=1}^{K} \mathcal{N}(\mathbf{0}, \tfrac{1}{P}\mathbf{I}_P), \tag{3}$$

$$\boldsymbol{w}_i = \boldsymbol{z}_i \odot \boldsymbol{s}_i,\ \boldsymbol{s}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \tag{4}$$

$$\boldsymbol{z}_i|\boldsymbol{\pi}_c \sim \prod_{k=1}^{K} \text{Bernoulli}(\pi_{ck}),\ \forall\, i : c(i) = c, \tag{5}$$

where $\boldsymbol{s}_i$ is the weight vector, π_{ck} is the probability of choosing the kth element of $\mathbf{D}_c$, and $\odot$ denotes the Hadamard product. We can observe that all the samples $\boldsymbol{x}_i$ in a specific subspace share one common probability vector $\boldsymbol{\pi}_c$.

Besides, in our nonparametric Bayesian model, the number of subspaces is not restricted to be finite, but is inferred automatically. Toward this end, we place a DP prior on the multiple dictionaries to be learned. Such a strategy can be described in a statistical measure-based style

$$G = \sum_{c=1}^{\infty} \xi_c \delta_{G_c} \sim \mathcal{DP}(\eta, \{G_c\}_{c=1}^{\infty}), \tag{6}$$

$$c(i) \sim \text{Multinomial}(\boldsymbol{\xi}), \tag{7}$$

$$\boldsymbol{\xi} \sim \text{GEM}(\eta),\ \eta \sim \text{Gamma}(a, b), \tag{8}$$

where $\mathcal{DP}(\eta, \{G_c\}_{c=1}^{\infty})$ denotes a DP with concentration parameters η and subspace-specified base measure $\{G_c\}_{c=1}^{\infty}$. With such a DP prior, $\boldsymbol{x}_i$ is assigned into one subspace via the multinomial distribution in (7) where $p(c(i) = c) = \xi_c$. The $\text{GEM}(\eta)$ in (8) represents the stick-breaking construction for the sequence $\boldsymbol{\xi}$ and can be written as

$$\xi_c = \rho_c \prod_{l=1}^{c-1}(1-\rho_l),\ \rho_c \sim \text{Beta}(1, \eta).$$

We set a truncation level for the DP to C in practice and let $\rho_C = 1$ which guarantees that $\sum_{c=1}^{C} \xi_c = 1$.

Moreover, the multiple subspaces for signals samples may overlap to some extent, the generative processes of $\{\boldsymbol{x}_i\}_{i=1}^{N}$ in different subspaces should not be entirely independent. We depict the correlations among G_c via a soft manner, where a HBP is further employed

$$H = \sum_{t=1}^{\infty} \upsilon_t \delta_{\phi_t} \sim \mathcal{BP}(\lambda, H_0),\ \phi_t \overset{\text{iid}}{\sim} H_0, \tag{9}$$

$$G_c = \sum_{k=1}^{\infty} \pi_{ck} \delta_{\varphi_{ck}} \overset{\text{iid}}{\sim} \mathcal{BP}(\gamma_c, H), \tag{10}$$

$$\varphi_{ck} = \phi_{u_{ck}},\ u_{ck} \sim \text{Multinomial}(\boldsymbol{\upsilon}), \tag{11}$$

where $\mathcal{BP}(\cdot,\cdot)$ denotes the Beta process. In contract to the conventional HBP priors defining the distribution of π_{ck} with respect to υ_k, we utilize a Sethuraman's stick-breaking construction which can viewed as a two-dimensional case of hierarchical Dirichlet process [16]. The reason for using Sethuraman's construction is that the conventional strategy is unable to achieve closed-form updates in variational inference. Instead, we introduce an auxiliary indicator $\boldsymbol{u}_c$ in (11) to build the connections between G_c and the global measure H.

To achieve analytical variational inference for the hierarchical model, the sequences of $\boldsymbol{\pi}_c$ and $\boldsymbol{\upsilon}$ are similarly generated from Indian buffet process (IBP) [17] with stick-breaking constructions which can be denoted as $\boldsymbol{\pi}_c \sim \mathcal{IBP}(\gamma_c)$ and $\boldsymbol{\upsilon} \sim \mathcal{IBP}(\lambda)$. Specifically, they are constructed as

$$\pi_{ck} = \prod_{m=1}^{k} \varpi_{cm},\ \varpi_{cm} \sim \text{Beta}(\gamma_c, 1),$$
$$\upsilon_t = \prod_{j=1}^{t} \beta_j,\ \beta_j \sim \text{Beta}(\lambda, 1).$$

Similar to the stick-breaking construction for DP, we set the truncations for the global and local buffets to T and K, and here K should be much smaller than T.

Finally, to infer all the hyper-parameters involved in a Bayesian framework, except for η in (8), we further set their priors as $\gamma_c \sim \text{Gamma}(c, d)$, $\lambda \sim \text{Gamma}(e, f)$, $\alpha_c \sim \text{Gamma}(g, h)$ to accomplish a full hierarchical Bayesian model with DP and HBP.

3. VARIATIONAL INFERENCE FOR THE DP-HBP

We focus on variational inference procedures for the proposed DP-HBP-based model. Mean-field variational strategies approximate the true posteriors by updating the factorized variational distributions $\mathcal{Q}$ to minimize their KL divergence [18]. Let $\mathcal{X}$ represent the data samples, $\mathcal{M}$ be the set of all the latent model parameters, and $\mathcal{H}$ be all the hyper-parameters. The variational objective function arises by maximizing the marginal likelihood lower bound $\mathcal{L}$ which is shown as

$$\mathcal{L} = \mathbb{E}_{\mathcal{Q}}[p(\mathcal{X}, \mathcal{M}|\mathcal{H})] + \mathbb{H}[\mathcal{Q}], \tag{12}$$

where $\mathbb{E}_{\mathcal{Q}}[\cdot]$ is the expectation operation w.r.t $\mathcal{Q}$, and $\mathbb{H}[\cdot]$ denotes the entropy function. Based on the joint distribution of our model, the variational distribution is factorized as

$$\begin{aligned}\mathcal{Q} =& q(\boldsymbol{\rho})q(\boldsymbol{\beta})q(\eta)q(\lambda) \times \prod_{i=1}^{N} q(\boldsymbol{z}_i)q(\boldsymbol{s}_i)q(c(i)) \\ &\times \prod_{c=1}^{C} q(\mathbf{D}_c)q(\boldsymbol{\varpi}_c)q(\boldsymbol{u}_c)q(\alpha_c)q(\gamma_c),\end{aligned}$$

and we let the q distributions subject to the same types of distributions with their priors.

We derive a coordinate ascent algorithm for achieving a local maximum of $\mathcal{L}$ in (12). Since many priors defined in our model are conjugate to the likelihood function, we discuss more about the non-conjugate terms.

3.1. Coordinate Update for DP

In the procedures of updating the variational distributions in DP, we focus on the subspace index (i), stick weight $\boldsymbol{\rho}$ and the concentration parameter η. We assume that $q(c(i) = c) = \bar{\xi}_c$ and $q(\rho_c) = \text{Beta}(a_c, b_c)$, then the corresponding q distributions are alternately updated as follows:

$$\begin{aligned}\bar{\xi}_c \propto \exp\big\{&\mathbb{E}\left[-\tfrac{P}{2}\ln(2\pi\alpha_c) - \alpha_c\|\boldsymbol{x}_i - \mathbf{D}_c\boldsymbol{w}_i\|_2^2/2\right] \\ &+\mathbb{E}[\ln\rho_c + \textstyle\sum_{l=1}^{c-1}\ln(1-\rho_l)]\big\},\end{aligned} \tag{13}$$

$$q(\rho_c) = \text{Beta}\left(1 + \textstyle\sum_{i=1}^{N}\bar{\xi}_c, \eta + \sum_{i=1}^{N}\sum_{l=c+1}^{C}\bar{\xi}_l\right), \tag{14}$$

$$q(\eta) = \text{Gamma}\left(a + C - 1, b - \textstyle\sum_{c=1}^{C-1}\mathbb{E}[\ln(1-\rho_c)]\right), \tag{15}$$

where $\mathbb{E}[\ln \rho_c] = \psi(a_c) - \psi(a_c + b_c)$, $\mathbb{E}[\ln 1 - \rho_c)] = \psi(b_c) - \psi(a_c + b_c)$, and $\psi(\cdot)$ is the digamma function. Besides, for the first term in (13), if $q(\alpha_c) = \mathrm{Gamma}(g_c, h_c)$, then we have $\mathbb{E}[\ln \alpha_c] = \psi(g_c) - \ln(h_c)$ based on the property of Gamma distribution. For the second term in (13), we require $q(\mathbf{D}_c)$, $q(\boldsymbol{w}_i)$ and $q(\alpha_c)$ to compute the norm expectation which is relatively trivial here.

3.2. Coordinate Update for HBP

To obtain the q distributions of the parameters in HBP, we first need to evaluate an expectation term $\mathbb{E}\left[\ln\left(1 - \prod_{m=1}^{k} \varpi_{cm}\right)\right]$, which is a byproduct of $\mathbb{E}[\ln p(z_i|\boldsymbol{\varpi}_c)]$ in the lower bound $\mathcal{L}$. However, since such a term is intractable, we resort to the multinomial approximation to lower bound it tightly instead. Via introducing an auxiliary multinomial distribution $q_k(y)$ and employing Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varpi}_c}\left[\ln\left(1 - \prod_{m=1}^{k} \varpi_{cm}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\varpi}_c}\left[\ln\left(\sum_{y=1}^{k} q_k(y) \frac{(1-\varpi_{cy})\prod_{m=1}^{y-1}\varpi_{cm}}{q_k(y)}\right)\right] \qquad (16) \\
&\geq \mathbb{E}_{\boldsymbol{\varpi}_c}\mathbb{E}_y\left[\ln(1-\varpi_{cy}) + \sum_{m=1}^{y-1}\ln\varpi_{cm}\right] + \mathbb{H}(q_k). \qquad (17)
\end{aligned}$$

To maximize the lower bound in (17), we take derivatives to find $q_k(y)$ with $q(\varpi_{cm}) = \mathrm{Beta}(\kappa_{cm,1}, \kappa_{cm,2})$, then $q_k(y) \propto$

$$\exp\left[\psi(\kappa_{cy,2}) + \sum_{m=1}^{y-1}\psi(\kappa_{cm,1}) - \sum_{m=1}^{y}\psi(\kappa_{cm,1}+\kappa_{cm,2})\right].$$

With such an evaluation in hand, we can further derive the expression of $q(z_i k) = \mathrm{Bernoulli}(\nu_{ik})$. Since

$$\ln q(z_{ik}) = \mathbb{E}[\ln p(\{z_{ik}\}|\boldsymbol{\varpi}_{c(i)}) + \ln\mathcal{N}(\boldsymbol{x}_i|\mathbf{D}_{c(i)}\boldsymbol{w}_i, \alpha_{c(i)}^{-1}\mathbf{I}_P)],$$

then we can obtain that $\nu_{ik} = \frac{1}{1+e^{-\vartheta}}$ with

$$\begin{aligned}
\vartheta = \sum_{c=1}^{C}\bar{\xi}_c\Bigg\{ & z_{ik}\sum_{m=1}^{k}\mathbb{E}[\ln\varpi_{cm}] + (1-z_{ik})\mathbb{E}[\ln(1-\prod_{m=1}^{k}\varpi_{cm})] \\
& + z_{ik}\left[-\tfrac{1}{2}\mathbb{E}[s_{ik}^2\mathbf{D}_{ck}^{\mathsf{T}}\mathbf{D}_{ck}] + \alpha_c\mathbb{E}[(\boldsymbol{x}_i^{-ck})^{\mathsf{T}}\mathbf{D}_{ck}s_{ik}]\right]\Bigg\},
\end{aligned}$$

where $\boldsymbol{x}_i^{-ck} \triangleq \boldsymbol{x}_i - \sum_{k'\neq k}\mathbf{D}_{ck'}z_{ik'}s_{ik'}$.

Furthermore, based on the lower bound in (17), we can update $q(\varpi_{cm}) = \mathrm{Beta}(\kappa_{cm,1}, \kappa_{cm,2})$ by

$$\begin{aligned}
\kappa_{ck,1} &= \gamma_c + \sum_{m=k}^{K} N_{cm} + \sum_{m=k+1}^{K}\overline{N}_{cm}\left(\sum_{j=k+1}^{m} q_{mj}\right) \\
\kappa_{ck,2} &= 1 + \sum_{m=k}^{K}\overline{N}_{cm}q_{mk}
\end{aligned}$$

where $N_{cm} = \bar{\xi}_c\sum_{i=1}^{N}(\nu_{im})$, and $\overline{N}_{cm} = \bar{\xi}_c\sum_{i=1}^{N}(1-\nu_{im})$.

In addition, to capture the correlations among the local buffets in HBP, the variational distribution of the mapping indicator $\boldsymbol{u}_c$ should also be updated. We assume that $q(u_{ck} = t) = \zeta_{ckt}$, then

$$\zeta_{ckt} \propto \exp\left\{\bar{\xi}_c\sum_{i=i}^{N}\mathbb{E}\left[\ln p(z_{ik}|\upsilon_t) + \mathbb{E}[\ln\upsilon_t]\right]\right\},$$

and here we also require similar techniques to evaluate the lower bound of $\mathbb{E}[\ln(1-\prod_{j=1}^{t}\beta_j)]$. For the hyper-parameters γ_c and λ in HBP, it is simple to derive their corresponding q distributions.

3.3. Coordinate Update for Dictionary Atoms

Let $\mathbf{D}_{ck}$ denote the kth atom of $\mathbf{D}_c$, and then for $q(\mathbf{D}_{ck}) = \mathcal{N}(\boldsymbol{d}_{ck}, \boldsymbol{\Omega}_{ck})$, we can update it via

$$\begin{aligned}
\boldsymbol{\Omega}_{ck} &= \left(P + \bar{\xi}_c\mathbb{E}[\alpha_c]\sum_{i=i}^{N}\nu_{ik}\mathbb{E}[s_{ik}^2]\right)^{-1}\mathbf{I}_P, \\
\boldsymbol{d}_{ck} &= \bar{\xi}_c\mathbb{E}[\alpha_c]\boldsymbol{\Omega}_{ck}\left(\sum_{i=i}^{N}\nu_{ik}\mathbb{E}[s_{ik}\boldsymbol{x}_i^{-ck}]\right),
\end{aligned}$$

Finally, for the atom weight vector $\boldsymbol{s}_i$, we have its q distribution $q(s_{ik}) = \mathcal{N}(\bar{s}_{ik}, \varepsilon^2)$ with

$$\varepsilon^2 = [\nu_{ik}\mathbb{E}(\alpha_c\mathbf{D}_{ck}^{\mathsf{T}}\mathbf{D}_{ck}) + 1]^{-1}, \bar{s}_{ik} = \varepsilon^2\nu_{ik}\mathbb{E}[\alpha_c\boldsymbol{x}_i^{-ck}]^{\mathsf{T}}\boldsymbol{d}_{ck}.$$

For brevity, the expression of $q(\alpha_c)$ is omitted here.

4. EXPERIMENTAL RESULTS

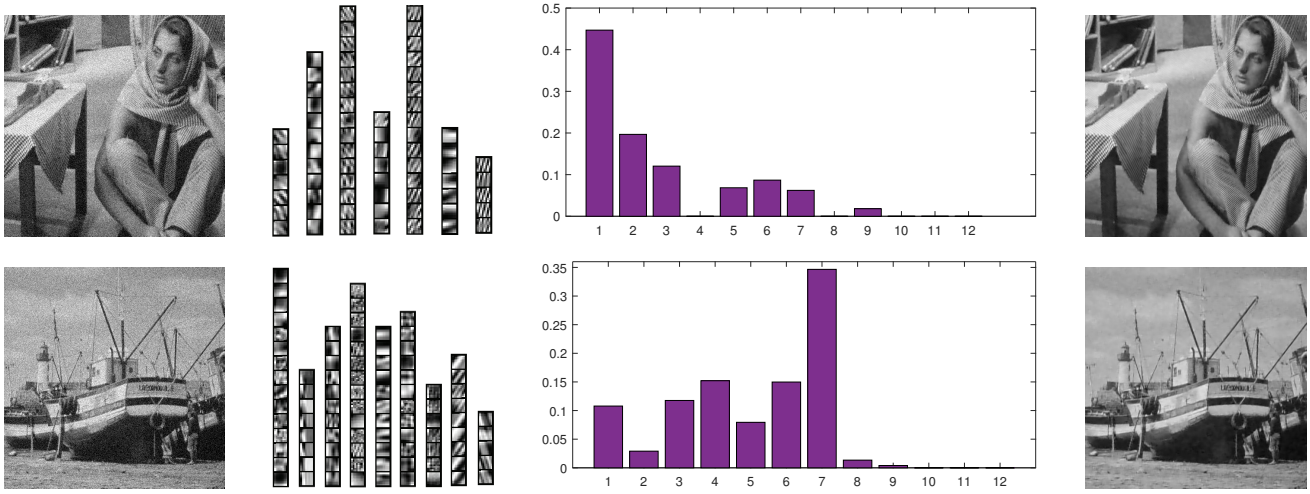
In this section, we evaluate the performance of the proposed variational inference for nonparametric subspace dictionary learning (NSDL) in image denoising tasks. To this end, we compare our strategy with state-of-the-art Bayesian dictionary learning algorithm BPFA [6], the classical K-SVD algorithm [3], and total variation (TV) denoising method [19].

4.1. Experiment Setup

We present image denoising experiments on five 512×512 standard test images – *Barbara, Lena, Boats, House, and Peppers*. The data samples $\{\boldsymbol{x}_i\}_{i=1}^{N}$ are 8×8 patches extracted from the noisy images, thus our target is to reconstruct the clean images using the inferred expectations of $\mathbf{D}_{c(i)}$ and $\boldsymbol{w}_i$. Based on this point, we need to initialize the variational distributions for the model parameters in our Bayesian model. Firstly, we set the truncation level in the DP-HBP priors. For the denoising application, we set $C = 12$, $K = 80$ and $T = 256$, which are guaranteed to be larger than the numbers of latent subspace and dictionary atoms. Besides, the initial values of the concentration parameters for DP and HBP prior are set as $\eta = 1$ and $\lambda = 1$, and the hyper-parameters for the Gamma priors and the Beta priors are set as $\mathrm{Gamma}(10^{-6}, 10^{-6})$ and $\mathrm{Beta}(0.5, 0.5)$, respectively. Moreover, the means of the dictionary atoms are initialized as zero vectors. Finally, total iterations of the variational inference are chosen as 50.

4.2. Subspace Dictionaries learned from Images

Due to the property of DP priors, we can automatically infer the number of the subspace dictionary for the target images. As shown in Fig. 1, taking two images as examples, we present the most frequently used atoms in each subspace dictionary learned from specific noisy images. Since the atom number in one subspace is much smaller than the truncation $K = 80$, the atoms which are not shown here are close to zero vector or Gaussian noise. At the same time, we can also



(a) Noisy image ($\sigma = 25$) (b) Subspace dictionaries (c) Usage probability vs. Subspace index (d) Denoised Image

Fig. 1: Experimental examples for two standard test images "Barbara" and "Boats". Each column in the subspace dictionaries represents the learned atoms of one subspace, while the subspace number, the atom number and their usage probability are inferred adaptively.

obtain the corresponding usage probability of the subspace dictionaries for the two images.

Interestingly, from the results in Fig. 1, we can observe that the patch textures with different characteristics are naturally grouped into the atom sets of different subspace dictionaries. In addition, the atom number of different dictionaries are also adaptively inferred. With such subspace dictionaries in hand, we are capable of better representing one specific image patch using the dictionary atoms in its corresponding subspace. Hence, we can expect the nonparametric subspace dictionaries to show better performance in image denoising or other related tasks.

4.3. Image Denoising

As a matter of fact, our variational inference algorithm can learn the subspace dictionaries and recover the true images simultaneously. Base on the expectations of the variational distributions $q(\mathbf{D}_c)$,$q(\boldsymbol{w}_i)$ and $q(c(i))$ in our Bayesian model, we can rebuild $\boldsymbol{x}_i$ by $\hat{\boldsymbol{x}}_i = \sum_{c=1}^{C} \bar{\xi}_c \sum_{k=1}^{K} \boldsymbol{d}_{ck} \nu_{ik} \bar{s}_{ik}$. In this manner, the denoised images can be achieved with $\{\hat{\boldsymbol{x}}_i\}_{i=1}^{N}$.

As illustrated in Table 1, we compare the reconstruction PSNR results of the proposed NSDL algorithm and three other image denoising methods. In our experiments, the standard deviations of the additive Gaussian noise are set as $\sigma \in \{10, 15, 20, 25\}$ for each image. Since total variation is a fundamental denoising method which does not learn dictionaries for images, it shows relatively poor performance compared to the other three dictionary-based algorithms. In contrast, K-SVD, as a classical dictionary learning method in the framework of optimization, shows better denoising accuracy for all the settings of σ. However, as an Bayesian method, BPFA can be viewed as an extension of K-SVD and introduces nonparametric Beta process to infer the appropriate dictionary atom number, thus it exhibits a further performance shift. Nevertheless, BPFA only learns one uniform dictionary for all the image patches, and hence is a special case of our NSDL when the dictionary number is set as one. Via taking the DP-HBP priors into account in our Bayesian model, the advantages of the proposed NSDL algorithm over BPFA in Table I indicate the effectiveness of learning a set of nonparametric subspace dictionaries.

Table 1: Comparisons of the denoising PSNR for five standard test images as a function of noise standard deviation σ.

σ \ Image		Barbara	Lena	Boats	House	Peppers
10	TV	29.77	32.71	31.64	33.76	32.40
	K-SVD	33.96	34.87	33.13	35.43	32.99
	BPFA	34.32	35.37	33.54	35.81	34.15
	NSDL	**34.50**	**35.57**	**33.70**	**35.98**	**34.31**
15	TV	27.49	30.96	29.79	31.89	30.44
	K-SVD	31.72	33.01	31.38	33.56	31.25
	BPFA	32.40	33.58	31.71	34.16	32.14
	NSDL	**32.69**	**33.84**	**31.96**	**34.45**	**32.44**
20	TV	26.01	29.84	28.50	30.76	29.25
	K-SVD	30.16	31.53	29.87	32.61	29.53
	BPFA	30.95	32.27	30.39	33.16	30.83
	NSDL	**31.17**	**32.46**	**30.74**	**33.47**	**31.18**
25	TV	25.07	28.87	27.57	29.96	28.26
	K-SVD	28.80	30.48	28.91	31.51	28.35
	BPFA	29.71	31.28	29.36	32.01	29.72
	NSDL	**30.03**	**31.45**	**30.12**	**32.33**	**30.09**

5. CONCLUSIONS

This paper has developed a nonparametric Bayesian approach to learning multiple subspace dictionaries. By combining the DP and the HBP as priors, the proposed model is capable of inferring the dictionary number and size simultaneously. Furthermore, we have also designed a closed-form variational inference engine based on the Sethuraman's stick-breaking construction. Experiments have demonstrated that our approach exhibits significant improvements compared to the existing dictionary learning methods in the image denoising tasks.

6. REFERENCES

[1] I. Tosic, and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, 2011.

[2] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," *Advances in Neural Information Processing Systems*, pp. 1033–1040, 2009.

[3] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *IEEE Trans. Signal Process.*, vol. 61, no. 3, pp. 661–677, 2013.

[4] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 130–144, 2012.

[5] J. Paisley, and L. Carin, "Nonparametric factor analysis with beta process priors," *ACM International Conference on Machine Learning*, pp. 777–784, 2009.

[6] S. Sertoglu and J. Paisley, "Scalable Bayesian nonparametric dictionary learning," *European Signal Processing Conference*, 2015.

[7] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3501–3508, 2010.

[8] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-hilasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4183–4198, 2011.

[9] R. Thibaux, and M. I. Jordan, "Hierarchical beta processes and the Indian buffet process," *International Conference on Artificial Intelligence and Statistics*, pp. 564–571. 2007.

[10] M .Zhou, H. Yang, G. Sapiro, D. B. Dunson, and L. Carin, "Dependent hierarchical beta process for image interpolation and denoising," *International Conference on Artificial Intelligence and Statistics*, pp. 883–891, 2011.

[11] S. K. Gupta, D. Q. Phung, and S. Venkatesh, "A Bayesian Nonparametric Joint Factor Model for Learning Shared and Individual Subspaces from Multiple Data Sources," *SIAM International Conference on Data Mining*, pp. 200–211, 2012.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, 2006.

[13] J. Paisley, C. Wang, D. Blei, and M. I. Jordan, "A Nested HDP for Hierarchical Topic Models," *arXiv preprint arXiv:1301.3570*, 2013.

[14] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," *International Conference on Artificial Intelligence and Statistics*, pp. 752–760, 2011.

[15] M. Hoffman, D. Blei, C. Wang and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, pp. 1005–1031, 2013.

[16] E. Fox, E. Sudderth, and M. Jordan, "An HDP-HMM for systems with state persistence," *ACM International Conference on Machine Learning*, 2008.

[17] T. L. Griffiths, and Z. Ghahramani, "The indian buffet process: An introduction and review," *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, 2011.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[19] T. Goldstein and S. Osher, "The split Bregman method for ℓ_1-regularized problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009.