Learning Dictionary with Spatial and Inter-dictionary Dependency

Yizhe Zhang, Ricardo Henao, Chunyuan Li, Lawrence Carin
Department of Electrical and Computer Engineering
Duke University Durham, NC 27708
{yz196,rhenao,chunyuan.li,lcarin}@duke.edu

Abstract

In dictionary learning for analysis of images, spatial correlation from extracted patches can be leveraged to improve characterization power. We propose a Bayesian framework for dictionary learning, where spatial location dependencies are captured by imposing a multiplicative Gaussian process prior on the latent units representing binary activations. Data augmentation and Kronecker methods allow for efficient Markov chain Monte Carlo sampling. We further extend our model with a sigmoid belief network, linking Gaussian processes and high-level latent binary units to capture inter-dictionary dependencies, while yielding additional computational savings. Applications to image denoising, inpainting and depth-information restoration demonstrate that the proposed model outperforms traditional Bayesian dictionary learning approaches.

1 INTRODUCTION

Learning overcomplete sparse latent representations for signal restoration and characterization has recently led to state-of-the-art results in tasks such as image denoising, inpainting, super-resolution and compressive sensing (Zhou et al., 2009; Yang et al., 2012). In dictionary learning, signals are represented in a latent factor space, where each signal is encoded as a sparse linear combination of dictionary elements (factors). Non-parametric Bayesian approaches have been successful at tackling these challenges (Zhou et al., 2009, 2012; Polatkan et al., 2015), by employing methodologies like the Indian buffet process (IBP) (Ghahramani and Griffiths, 2005). Recent work has demonstrated that modeling images as a collection of sub-regions or *patches* is important, because leveraging local image structure is instrumental for representational quality (Mairal et al., 2009; Zhou et al., 2009). Furthermore, dictionary learning can be greatly improved by imposing that patches close in space are likely to use the same or similar dictionary elements (Zhou et al., 2011).

In this paper, we propose a framework for dictionary learning where patch-to-patch spatial dependencies are modeled via GP (Rasmussen and Williams, 2006; Wilson et al., 2014) priors linked to binary dictionary element activations. They are appealing because one can use them, in a principled non-parametric manner, to estimate correlation as a function of relative spatial location. Despite great flexibility, GPs are known to be computationally expensive. To address this challenges, we consider an efficient Kronecker inference method, with multiplicative covariance functions (Gilboa et al., 2015). Furthermore, we utilize Sigmoid Belief Networks (SBNs) to impose correlation structure across dictionary elements, which we demonstrate leads to performance improvements in many cases. The GPs link to the binary units at the top layer of a SBN, and the number of these top-layer units may be made small relative to the number of dictionary elements. Therefore, there is a substantial computational savings manifested by placing the (small number of) GPs at the top of an SBN, rather than placing GP for each of the (large number of) dictionary elements.

2 DICTIONARY LEARNING WITH GAUSSIAN PROCESSES

Assume observed data $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \in \mathbb{R}^{J \times N}$, where $\boldsymbol{x}_i$ represents data from one of N patches extracted from a single image. In Bayesian dictionary learning, the goal is to learn dictionary elements, $\boldsymbol{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\} \in \mathbb{R}^{J \times M}$ from data, $\boldsymbol{X}$. The i-th observations, $\boldsymbol{x}_i$, is represented as

$$\begin{aligned} \boldsymbol{x}_i &= \boldsymbol{D}(\boldsymbol{w}_i \odot \boldsymbol{z}_i) + \boldsymbol{\varepsilon}_i\,, \quad & \boldsymbol{\varepsilon}_i &\sim \mathcal{N}(0, \sigma_\varepsilon^2 \boldsymbol{I}_J)\,, \\ \boldsymbol{d}_m &\sim \mathcal{N}(0, \boldsymbol{I}_J)\,, & \boldsymbol{w}_i &\sim \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_M)\,, \end{aligned} \tag{1}$$

where $\odot$ denotes the element-wise (Hadamard) product, and $\boldsymbol{I}_J$ is a $J \times J$ identity matrix. Vectors $\boldsymbol{w}_i = \{w_{i1}, \ldots, w_{iM}\} \in \mathbb{R}^M$ and $\boldsymbol{z}_i = \{z_{i1}, \ldots, z_{iM}\} \in \{0,1\}^M$ represent weights and *binary activations*, respectively. Specifically, $\boldsymbol{z}_i$, encodes the presence or absence of dictionary elements for the i-th sample. $\boldsymbol{\varepsilon}_i$ is *i.i.d.* additive Gaussian noise.

It is reasonable to assume that patches located near each other are likely to be represented in terms of the same or similar dictionary elements, and thus binary activations of nearby patches are likely to be consistent with spatial dependencies (Zhou et al., 2011). In order to incorporate such a prior belief, we employ a GP on a 2-D spatial field. Our approach differs from that by Zhou et al. (2011), in that GPs allow estimation of binary activation dependencies (by connecting the GP output to a logistic link function). We call our method GP-FA, for short. The construction for the m-th binary activation, z_{im}, is given by $z_{im} \sim \text{ Bernoulli}(\sigma(y_{im}))$, where

$$y_{im} = f_m(\boldsymbol{l}_i)\,, \qquad f_m(\cdot) \sim \mathcal{GP}(b_m, k_m(\boldsymbol{l}_i, \cdot))\,, \tag{2}$$

$$b_m \sim \mathcal{N}(\lambda_m, \sigma_b^2)\,, \qquad \lambda_m \sim \mathcal{N}_-(0, \sigma_\lambda^2)\,, \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function, y_{im} is the value of function $f_m(\cdot)$ evaluated at the 2-D spatial coordinates of the i-th patch, $\boldsymbol{l}_i = \{l_i^{(1)}, l_i^{(2)}\}$. The function $f_m(\cdot)$ is drawn from a GP with constant mean function, $\mu_m(\cdot) = b_m$, and multiplicative covariance function, $k_m(\boldsymbol{l}_i, \cdot)$, defined as $k_m(\boldsymbol{l}_i, \cdot) = k_m^{(1)}(l_i^{(1)}, \cdot) \otimes k_m^{(2)}(l_i^{(2)}, \cdot)$.

For the mean, b_m, we specify a Gaussian prior with mean and variance, λ_m and σ_b^2, respectively. To encourage sparsity in the activations, z_{im}, we bias function instances, y_{im}, towards negative values using a zero-mean Gaussian distribution truncated above zero (*i.e.*, negative support) with variance σ_λ^2. Since this prior is shared by all factors, it encourages sparsity globally. Further, the hierarchy in (3) is convenient from a practical stand point, because it yields local conjugacy.

For the covariance function, $k_m(\cdot, \cdot)$, in our implementation we consider the widely used squared exponential (SE) function. Specifically, the covariance function for axis $s = \{1, 2\}$, is defined as

$$k_m^{(s)}(l^{(s)}, l^{(s')}; \boldsymbol{\Theta}_m) = (\sigma_f^2)_m \exp\{-(l^{(s)} - l^{(s')})^2 / \theta_m\}\,,$$

where $\boldsymbol{\Theta}_m = \{(\sigma_f^2)_m, \theta_m\}$ is the set of parameters for the m-th dictionary element, $(\sigma_f^2)_m$ is the signal variance and θ_m is the characteristic length scale (Rasmussen and Williams, 2006).

Pólya-gamma augmentation To perform Gibbs inference for such model, we leverage the Pólya-Gamma (PG) data augmentation scheme of Polson et al. (2013). In contrast to probit-based augmentation, PG augmentation has been shown to be efficient with sophisticated posteriors (Gan et al., 2015), while enjoying theoretical guarantees in terms of unbiased estimates of posterior expectations (Choi et al., 2013). Briefly, if the auxiliary variable γ is draw from Pólya-gamma distribution, *i.e.*, $\gamma \sim \mathcal{PG}(1, 0)$, the conditional distribution of $y_i m$ given γ is $\mathcal{N}(\mu_*, \sigma_*)$, where, ($m$ is omitted for clarity. $\boldsymbol{K}$ represents the Gram matrix, subscript $\backslash i$ indicates $(i' : i' \neq i, i' \in \{1, \cdots, N\})$)

$$\mu_* = \left(\frac{\boldsymbol{k}_{i,\backslash i} \boldsymbol{K}_{\backslash i,\backslash i}^{-1} \boldsymbol{y}_{\backslash i}^T}{k_{i,i} - \boldsymbol{k}_{i,\backslash i} \boldsymbol{K}_{\backslash i,\backslash i}^{-1} \boldsymbol{k}_{i,\backslash i}^T} + z_i - \frac{1}{2} - \gamma_i b \right) \sigma_*^2\,, \quad \sigma_*^2 = \left(\frac{1}{k_{i,i} - \boldsymbol{k}_{i,\backslash i} \boldsymbol{K}_{\backslash i,\backslash i}^{-1} \boldsymbol{K}_{i,\backslash i}^T} + \gamma_i \right)^{-1}$$

The conditional posterior for binary activations, z_{im}, is dependent on both dictionary factorization and Gaussian process prior, thus we can write $z_{im}|- \sim \text{ Bernoulli}(p_{im}^*/(1 + p_{im}^*))$, where

$$p_{im}^* = \exp\left\{ \frac{1}{\sigma_\varepsilon^2} \sum_{j=1}^{J} \left(x_{ij} - \sum_{m' \neq m} d_{jm'} s_{im'} \right) d_{jm} w_{im} - \frac{1}{2\sigma_\varepsilon^2} (d_{jm} w_{im})^2 + y_{im} + b_m \right\}.$$

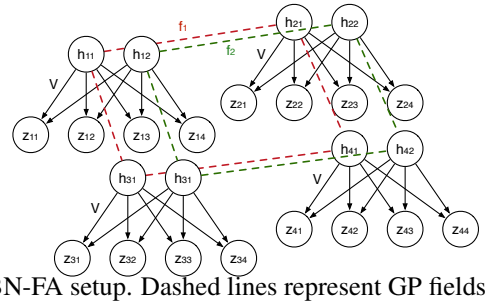


Figure 1: GP-SBN-FA setup. Dashed lines represent GP fields ($f_1(\cdot)$ and $f_2(\cdot)$).

Kronecker method Unfortunately, such naive implementation scales as $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ memory per patch, due to matrix inversion. In this paper, we adopt the fast inference method of Gilboa et al. (2015), where the computational cost can be effectively reduced to $\mathcal{O}(N^{3/2})$ time and $\mathcal{O}(N)$ memory per patch. Specifically, by defining $\boldsymbol{\Gamma} = \begin{bmatrix} \gamma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, from the block matrix inversion lemma (Petersen and Pedersen, 2012) we can write $(\boldsymbol{K} + \boldsymbol{\Gamma})^{-1} \overset{\gamma \to 0}{\approx} \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_{\setminus i, \setminus i}^{-1} \end{bmatrix}$. Such approximation allows us to perform a single inversion on the full Gram matrix, $\boldsymbol{K}$, instead of $\boldsymbol{K}_{\setminus i, \setminus i}$. This is desirable because $\boldsymbol{K}$ can be represented as $\boldsymbol{K} = \boldsymbol{K}_1 \otimes \boldsymbol{K}_2$. Existing Kronecker methods can be applied via Preconditioned Conjugate Gradient (PCG) (Shewchuk, 1994)based on the fast computation of $\boldsymbol{\alpha} = (\otimes_{d=1}^{D} \boldsymbol{A}_d)\boldsymbol{b}$.

Automatic relevance determination To estimate the parameters of the covariance functions, for $(\sigma_f^2)_m$ and θ_m, we use *maximum a posterior* (MAP) estimation, corresponding to dictionary element m. This is done by maximizing the conditional log-posterior function. Again, the Kronecker product trick can be employed for fast inference via Cholesky decompositions.

3 DICTIONARY LEARNING WITH GP-SBN

We leverage sigmoid belief networks (SBNs) as an alternative way of linking binary activations in (1) with the binary output from the GP. As shown in Figure 1, instead of placing a GP prior on each of the M dictionary elements as in GP-FA, we use GPs to impose spatial dependency on the hidden units of an SBN. We denote this model as GP-SBN-FA, for short. Building upon recent work on SBNs (Gan et al., 2015), we consider an SBN with L binary units, which can be written as

$$\boldsymbol{z}_i \sim \text{Bernoulli}(\sigma(\boldsymbol{V}\boldsymbol{h}_i + \boldsymbol{b}))\,, \boldsymbol{h}_i \sim \quad \text{Bernoulli}(\sigma(\boldsymbol{y}_i))\,, \tag{4}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iM})^T$ has a GP prior, as in (2), and $\boldsymbol{h}_i \in \{0,1\}^L$ is a vector of L binary units. The weight matrix, $\boldsymbol{V} \in \mathbb{R}^{M \times L}$, contains L features encoding M dictionary elements correlations. We have observed that $L = M/2$ works well in practice. We place a three-parameter beta normal prior on the weight matrix, $\boldsymbol{V}$, which has demonstrated good mixing performance (Gan et al., 2015). Further, we let $\boldsymbol{b} \sim \mathcal{N}(0, \boldsymbol{I}_M)$, for simplicity. Closed-form conditional posteriors for $\{\boldsymbol{V}, \boldsymbol{b}\}$ via Gibbs sampling are available via Pólya-gamma data augmentation (Gan et al., 2015). In this work we only consider one-layer SBNs as in (4). However, adding layers to form deep architectures is straightforward, as previously described by Gan et al. (2015). Use of deep SBNs may result in more computational savings, as the number of top-layer deep SBN units can be small, reducing the number of needed GPs. With this said, the single-layer SBN considered here yields excellent results, and computational savings.

4 EXPERIMENTS

4.1 2-D Grayscale Images

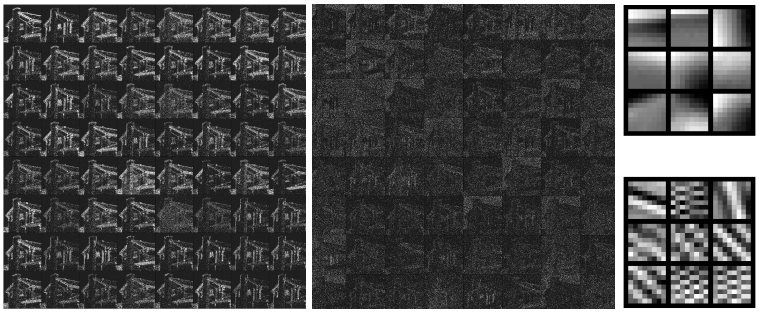
We analyzed 10 gray-scale images for image denoising and inpainting. For denoising, we added isotropic *i.i.d.* Gaussian noise, $\mathcal{N}(0, \sigma)$, to each pixel with $\sigma = 25$. For inpainting, we consider 50% observation ratio (observed pixels selected uniformly at random). Each image was partitioned into 8×8 patches with sliding distance of one pixel, *i.e.*, the distance between centers of neighbor

Method	C.man	House	Pepper	Lena	Barbara	Method	C.man	House	Pepper	Lena	Barbara
BP	28.41	31.92	29.36	31.25	28.83	BP	28.90	38.02	32.58	36.94	33.17
GP-FA	28.70	32.22	29.65	31.42	29.11	GP-FA	**29.03**	38.53	32.84	**37.18**	33.18
GP-SBN-FA	**28.99**	**32.23**	**29.78**	**31.51**	**29.18**	GP-SBN-FA	28.98	**38.89**	**33.04**	37.01	**33.33**
Method	Boats	F.print	Man	Couple	Hill	Method	Boats	F.print	Man	Couple	Hill
BP	29.25	27.44	29.06	28.89	29.29	BP	33.78	33.53	33.29	**35.56**	34.23
GP-FA	29.49	**27.55**	**29.27**	29.04	29.49	GP-FA	**34.16**	**34.08**	**33.83**	34.63	**34.46**
GP-SBN-FA	**29.56**	27.54	29.23	**29.15**	**29.52**	GP-SBN-FA	33.98	33.89	33.54	33.60	34.31

Table 1: Denoising (left) and inpainting (right) results. Performance is measured as PSNR in dB.

patches is one pixel. We ran 500 MCMC iterations with random initialization and kept the last 50 samples for image reconstruction (averaging over these collection samples). The hyper-parameters controlling Gaussian distribution variances, *i.e.*, σ_w, σ_ε, σ_b and σ_λ, were all set to 0.1, the hyper-parameters for the Gamma distributions were set to 10^{-6}. Dictionary sizes in both GP-FA and GP-SBN-FA are initially set to 128. We use Peak Signal-to-Noise Ratio (PSNR) to measure the recovery performance of original images. Compared with BPFA (Zhou et al., 2009), as shown in Table 1, GP-SBN-FA yields the best results for most images under different regimes.

For GP-SBN-FA, the learned dictionary elements, binary activations for dictionary elements and the binary units of SBN are also shown in Figure 2. The imposed SBN architecture encourages blocks of dictionary elements to simultaneously turn on or turn off. Such an inter-dictionary dependency assumption is useful if the dictionary elements are heavily correlated.



(a) (b) (c)

Figure 2: GP-SBN-FA. a) Binary activations of dictionary elements learned. b) Binary hidden units of SBN. c) Inter-dictionary dependency captured by hidden units.

4.2 Depth Restoration

We applied our methods to the 30 images of the Middlebury stereo dataset (Scharstein and Szeliski, 2002; Lu et al., 2014). The task is to jointly recover the corrupted pixels in the depth map and to denoise RGB-D channels. We compared our methods with BPFA (Zhou et al., 2009) and dHBP (Zhou et al., 2011). We used 500 burn-in samples for our methods, and kept 50 MCMC collection samples for image reconstruction. For BPFA and dHBP, we use default settings for the hyper-parameters, and perform 64 sequential MCMC iterations with incomplete data (Zhou et al., 2009), followed by 300 MCMC iterations. An overall comparison of depth channel interpolation task is shown in Figure 3. In general, GP-FA is marginally better than GP-SBN-FA, as in about 75% images it performs better than GP-SBN-FA. However, GP-SBN-FA is approximately 25% faster than GP-FA. GP-FA and GP-SBN-FA are consistently better than dHBP and BPFA in all images.

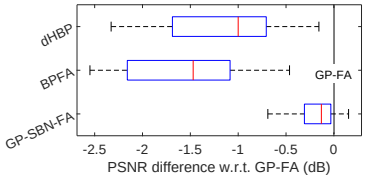


Figure 3: Results of depth-information restoration, each method was compared with GP-FA.

References

H. M. Choi, J. P. Hobert, et al. The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013.

Z. Gan, R. Henao, D. Carlson, and L. Carin. Learning Deep Sigmoid Belief Networks with Data Augmentation. 2015.

Z. Ghahramani and T. L. Griffiths. Infinite latent feature models and the indian buffet process. In *NIPS*, pages 475–482, 2005.

E. Gilboa, Y. Saatci, and J. P. Cunningham. Scaling Multidimensional Inference for Structured Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 424–436, 2015.

S. Lu, X. Ren, and F. Liu. Depth Enhancement via Low-Rank Matrix Completion. pages 3390–3397, 2014.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, pages 689–696, 2009.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012.

G. Polatkan, M. Zhou, L. Carin, D. Blei, and I. Daubechies. A Bayesian Nonparametric Approach to Image Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37 (2):346–358, Feb. 2015.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *JASA*, 108(504):1339–1349, Aug. 2013.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Jan. 2006.

D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 2002.

J. R. Shewchuk. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, 1994.

A. Wilson, E. Gilboa, J. P. Cunningham, and A. Nehorai. Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems*, pages 3626–3634, 2014.

J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *Image Processing, IEEE Transactions on*, 21(8):3467–3478, 2012.

M. Zhou, H. Chen, L. Ren, G. Sapiro, L. Carin, and J. W. Paisley. Non-Parametric Bayesian Dictionary Learning for Sparse Image Representations. pages 2295–2303, 2009.

M. Zhou, H. Yang, and G. Sapiro. Dependent hierarchical beta process for image interpolation and denoising. *AISTATS*, 2011.

M. Zhou, H. Chen, J. W. Paisley, L. Ren, L. Li, Z. Xing, D. B. Dunson, G. Sapiro, and L. Carin. Nonparametric Bayesian Dictionary Learning for Analysis of Noisy and Incomplete Images. *IEEE Transactions on Image Processing*, 21(1):130–144, 2012.