Nonparametric Learning of Dictionaries for Sparse Representation of Sensor Signals

Mingyuan Zhou, John Paisley and Lawrence Carin
Electrical & Computer Engineering Department
Duke University
Durham, NC 27708–0291
Email: lcarin@ee.duke.edu

***Abstract*—Nonparametric Bayesian techniques are considered for learning dictionaries for sparse data representations, with applications in sparse rendering of sensor data. The beta process is employed as a prior for learning the dictionary, and this nonparametric method naturally infers an appropriate dictionary size. The proposed method can learn a sparse dictionary, and may also be used to denoise a signal under test. The noise variance need not be known, and can be non-stationary. The dictionary coefficients for a given sensor signal may be employed within a classifier. Several example results are presented, using both Gibbs and variational Bayesian inference, with comparisons to other state-of-the-art approaches.**

I. Introduction

There has been significant recent interest in sparse signal expansions, in several settings. For example, such algorithms as the support vector machine (SVM) [3], the relevance vector machine (RVM) [18], Lasso [17] and many others have been developed for sparse regression (and classification). A sparse representation has several advantages, including the fact that it encourages a simple model, and therefore over-training is often avoided. The inferred sparse coefficients also often have biological/physical meaning, of interest for model interpretation [11].

Of relevance for the current paper, there has recently been significant interest in sparse representations in the context of denoising and classification [19]. All of these applications exploit the fact that most images may be sparsely represented in an appropriate dictionary. Most of the CS literature assumes "off-the-shelf" wavelet and DCT bases/dictionaries [6], but recent research has demonstrated the significant advantages of learning an often over-complete dictionary matched to the signals of interest ($e.g.$, images) [1], [5], [10], [8], [9], [15], [4], [14]. The purpose of this paper is to perform dictionary learning using new nonparametric Bayesian technology [16], [12], that offers several advantages not found in earlier approaches, which have generally sought point estimates. We demonstrate how these may be applied to sensor data.

II. Dictionary Learning with a Beta Process

In traditional *sparse coding* tasks, one considers a signal $\boldsymbol{x} \in \Re^n$ and a *fixed* dictionary $\mathbf{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_M)$ where each $\boldsymbol{d}_m \in \Re^n$. We wish to impose that any $\boldsymbol{x} \in \Re^n$ may be represented approximately as $\hat{\boldsymbol{x}} = \mathbf{D}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \Re^M$ is sparse, and our objective is to also minimize the ℓ_2 error $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2$. With a proper dictionary, a sparse $\boldsymbol{\alpha}$ often manifests robustness to noise (the model doesn't fit noise well), and the model also yields effective inference of $\boldsymbol{\alpha}$ even when $\boldsymbol{x}$ is partially or indirectly observed via a small number of measurements (of interest for inpainting, interpolation and compressive sensing [1], [10]). To the authors' knowledge, all previous work in this direction has been performed in the following manner: (i) if $\mathbf{D}$ is given, the sparse vector $\boldsymbol{\alpha}$ is estimated via a point estimate (without a posterior distribution), typically based on orthogonal matching pursuits (OMP), basis pursuits or related methods, for which the stopping criteria is defined by assuming knowledge (or off-line estimation) of the noise variance or the sparsity level of $\boldsymbol{\alpha}$; and (ii) when the dictionary $\mathbf{D}$ is to be learned, the dictionary size M must be set *a priori*, and a point estimate is achieved for $\mathbf{D}$ (in practice one may infer M via cross-validation, with this step avoided in the proposed method). In many applications one may not know the noise variance or an appropriate sparsity level of $\boldsymbol{\alpha}$; further, one may be interested in the confidence of the estimate ($e.g.$, "error bars" on the estimate of $\boldsymbol{\alpha}$). To address these goals, we propose development of a non-parametric Bayesian formulation to this problem, in terms of the beta process, this allowing one to infer the appropriate values of M and $\|\boldsymbol{\alpha}\|_0$ (sparsity level) jointly, also manifesting a full posterior density function on the learned $\mathbf{D}$ and the inferred $\boldsymbol{\alpha}$ (for a particular $\boldsymbol{x}$), yielding a measure of confidence in the inversion. As discussed further below, the non-parametric Bayesian formulation also allows one to relax other assumptions that have been made in the field of learning $\mathbf{D}$ and $\boldsymbol{\alpha}$ for inpainting, denoising and compressive sensing. Further, the addition of other goals are readily addressed within the non-parametric Bayesian paradigm, $e.g.$ designing $\mathbf{D}$ for *joint* compression *and* classification.

We desire the model $\boldsymbol{x} = \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where $\boldsymbol{x} \in \Re^n$ and $\mathbf{D} \in \Re^{n \times M}$, and we wish to learn $\mathbf{D}$ and in so doing infer M. Toward this end, we consider a dictionary $\mathbf{D} \in \Re^{n \times K}$, with $K \to \infty$; by inferring the number of columns of $\mathbf{D}$ that are required for accurate representation of $\boldsymbol{x}$, the appropriate value of M is implicitly inferred (work has been considered in [7], [13] for the related but distinct application of factor analysis). We wish to also impose that $\boldsymbol{\alpha} \in \Re^K$ is sparse, and therefore only a small fraction of the columns of $\mathbf{D}$ are used for representation of a given $\boldsymbol{x}$. Specifically, assume that we

978-1-4244-5180-7/09/$26.00 ©2009 IEEE

have a training set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1,N}$, where $\boldsymbol{x}_i \in \Re^n$ and $y_i \in \{1, 2, \dots, N_c\}$, where $N_c \geq 2$ represents the number of classes from which the data arise; when learning the dictionary we ignore the class labels y_i, and later discuss how they may be considered in the learning process.

The two-parameter beta process (BP) was developed in [12], to which the reader is referred for further details; we here only provide those details of relevance for the current application. The BP with parameters $a > 0$ and $b > 0$, and base measure H_0, is represented as BP(a, b, H_0), and a draw $H \sim$ BP(a, b, H_0) may be represented as

$$\begin{aligned} H(\boldsymbol{\psi}) &= \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\psi}_k}(\boldsymbol{\psi}) \\ \pi_k &\sim \text{Beta}(a/K, b(K-1)/K) \\ \boldsymbol{\psi}_k &\sim H_0 \end{aligned} \tag{1}$$

with this a valid measure as $K \to \infty$. The expression $\delta_{\boldsymbol{\psi}_k}(\boldsymbol{\psi})$ equals one if $\boldsymbol{\psi} = \boldsymbol{\psi}_k$ and is zero otherwise. Therefore, $H(\boldsymbol{\psi})$ represents a vector of K probabilities, with each associated with a respective atom $\boldsymbol{\psi}_k$. In the limit $K \to \infty$, $H(\boldsymbol{\psi})$ corresponds to an infinite-dimensional vector of probabilities, and each probability has an associated atom $\boldsymbol{\psi}_k$ drawn i.i.d. from H_0.

Using $H(\boldsymbol{\psi})$, we may now draw N *binary* vectors, the ith of which is denoted $\boldsymbol{z}_i \in \{0, 1\}^K$, and the kth component of $\boldsymbol{z}_i$ is drawn $z_{ik} \sim$ Bernoulli(π_k). These N binary column vectors are used to constitute a matrix $\mathbf{Z} \in \{0, 1\}^{K \times N}$, with ith column corresponding to $\boldsymbol{z}_i$; the kth row of $\mathbf{Z}$ is associated with atom $\boldsymbol{\psi}_k$, drawn as discussed above. For our problem the atoms $\boldsymbol{\psi}_k \in \Re^n$ will correspond to candidate members of our dictionary $\mathbf{D}$, and the binary vector $\boldsymbol{z}_i$ defines which members of the dictionary are used to represent sample $\boldsymbol{x}_i \in \mathcal{D}$.

Let $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_K)$, and we may consider the limit $K \to \infty$. A naive form of our model, for representation of sample $\boldsymbol{x}_i \in \mathcal{D}$, is $\boldsymbol{x}_i = \boldsymbol{\Psi}\boldsymbol{z}_i + \boldsymbol{\epsilon}_i$. However, this is highly restrictive, as it imposes that the coefficients of the dictionary expansion must be binary. To address this, we draw weights $\boldsymbol{w}_i \sim \mathcal{N}(0, \gamma_w^{-1}\mathbf{I}_K)$, where γ_w is the precision or inverse variance; the dictionary weights are now $\boldsymbol{\alpha}_i = \boldsymbol{z}_i \circ \boldsymbol{w}_i$, and $\boldsymbol{x}_i = \boldsymbol{\Psi}\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i$, where $\circ$ represents the Hadamard (element-wise) multiplication of two vectors. Note that, by construction, $\boldsymbol{\alpha}$ is sparse; this imposition of sparseness is distinct from the widely used Laplace shrinkage prior [17], which imposes that many coefficients are small but not necessarily exactly zero.

For simplicity we assume that the dictionary elements, defined by the atoms $\boldsymbol{\psi}_k$, are drawn from a multivariate Gaussian base H_0, and the components of the error vectors $\boldsymbol{\epsilon}_i$ are drawn i.i.d. from a zero-mean Gaussian. The hierarchical form of the model may now be expressed as

$$\begin{aligned} \boldsymbol{x}_i &= \boldsymbol{\Psi}\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i \ , \qquad \boldsymbol{\alpha}_i = \boldsymbol{z}_i \circ \boldsymbol{w}_i \\ \boldsymbol{\Psi} &= (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_K) \ , \quad \boldsymbol{\psi}_k \sim \mathcal{N}(0, n^{-1}\mathbf{I}_n) \\ \boldsymbol{w}_i &\sim \mathcal{N}(0, \gamma_w^{-1}\mathbf{I}_K) \ , \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \gamma_\epsilon^{-1}\mathbf{I}_n) \\ \boldsymbol{z}_i &\sim \prod_{k=1}^{K} \text{Bernoulli}(\pi_k), \ \pi_k \sim \text{Beta}(a/K, b(K-1)/K) \end{aligned} \tag{2}$$

Non-informative gamma hyper-priors are typically placed on γ_w and γ_ϵ. Consecutive elements in the above hierarchical model are in the conjugate exponential family, and therefore inference may be implemented via a variational Bayesian [2] or Gibbs-sampling analysis, with analytic update equations (all inference update equations, and the software, will be referenced in a technical report, if the paper is accepted). After performing such inference, we retain those columns of $\boldsymbol{\Psi}$ that are used in the representation of the data in $\mathcal{D}$, thereby inferring $\mathbf{D}$ and hence M.

To impose our desire that the vector of dictionary weights $\boldsymbol{\alpha}$ is sparse, one may adjust the parameters a and b. Particularly, as discussed in [12], in the limit $K \to \infty$, the number of elements of $\boldsymbol{z}_i$ that are non-zero is a random variable drawn from Poisson(a/b). In Section 3.1 we discuss the fact that these parameters are in general non-informative and the sparsity is intrinsic to the data.

III. Example Results

For the denoising results considered below, we observed that the Gibbs sampler provided better performance than associated variational Bayesian inference. For denoising we may exploit shifted versions of the data, which accelerates convergence substantially (discussed in detail below). Therefore, all denoising and inpainting results are based on efficient Gibbs sampling. The same set of model hyper-parameters are used across all our denoising examples (no tuning was performed): all gamma priors are set as Gamma$(10^{-6}, 10^{-6})$, along the lines suggested in [18], and the beta distribution parameters are set with $a = K$ and $b = N/8$ (many other settings of a and b yield similar results).

A. Denoising

We consider denoising a 256×256 image, with comparison of the proposed approach to K-SVD [5] (for which the noise variance is assumed known and fixed); the *true* noise standard deviation is set at 15, 25 and 50 in the examples below. We show results for three algorithms: (i) mismatched K-SVD (with noise standard deviation of 30), (ii) K-SVD when the standard deviation is properly matched, and (iii) the proposed BP approach. For (iii) a non-informative prior is placed on the noise precision, and the same BP model is run for all three noise levels (with the underlying noise levels inferred). The BP and K-SVD employed no *a priori* training data. In Figure 1 are shown the noisy images at the three different noise levels, as well as the reconstructions via BP and K-SVD. A preset large dictionary size $K = 256$ is used for both algorithms, and for the BP results we inferred that approximately M = 219, 143, and 28 dictionary elements were important for noise standard deviations 15, 25, and 50, respectively; the remaining elements of the dictionary were used less than 0.1% of the time. As seen within the bottom portion of the right part of Figure 1, the unused dictionary elements appear as random draws from the prior, since they are not used and hence influenced by the data.

Note that K-SVD works well when the set noise variance is at or near truth, but the method is undermined by mismatch. The proposed BP approach is robust to changing noise levels. Quantitative performance is summarized in Table I. The BP denoiser estimates a full posterior density function on the noise standard deviation; for the examples considered here, the modes of the inferred standard-deviation posteriors were 15.52, 25.33, and 48.13, for true standard deviations 15, 25, and 50, respectively.

To achieve these BP results, we employ a sequential implementation of the Gibbs sampler (a batch implementation converges to the same results but with higher computational cost); this discussed in further detail below, when presenting inpainting results.

TABLE I
Peak signal-to-reconstructed image measure (PSNR) for the data in Figure 1, for K-SVD [5] and the proposed BP method. The true standard deviation was 15, 25 and 50, respectively, from the top to the bottom row. For the mismatched K-SVD results, the noise stand deviation was fixed at 30.

Original Noisy Image (dB)	K-SVD mismatched (dB)	K-SVD matched (dB)	Beta Process Denoising (dB)
24.58	30.67	34.22	34.19
20.19	31.52	32.08	31.89
14.56	19.60	27.07	27.85

B. Learning dictionaries for UXO sensing

As a practical sensing example, we consider magnetometer data collected at a real former bombing site in the United States (used previously for soldier training). Our objective is to learn a dictionary in which the data may be compactly rendered. Details of this problem will be presented in the talk. The measured data are shown in Figure 2, where each small image chip represents a given item that passed a pre-screener, based on signal energy (a given item may be unexploded ordnance (UXO) or clutter (non-UXO). The learned dictionary is depicted in Figure 3. In the talk we will demonstrate how this dictionary may be used to perform UXO classification effectively.

IV. Conclusion

The nonparametric beta process has been presented for dictionary learning with the goal of image denoising and learning of dictionaries matched to sensing data. In the context of noisy underlying data, the noise variance need not be known in advance, and it need not be spatially uniform. The algorithm may be used to learn dictionaries that are matched to the sensing signals of interest, with sensing of unexploded ordnance (UXO) a particular example considered here.

References

[1] M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Trans. Signal Processing*, 54, 2006.

[2] M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[4] J.M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IMA Preprint Series 2211*, 2008.

[5] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Processing*, 15, 2006.

[6] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56, 2008.

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

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the International Conference on Machine Learning*, 2009.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Proceedings of Neural and Information Processing Systems*, 2008.

[10] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17, 2008.

[11] B.A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37, 1997.

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

[13] P. Rai and H. Daume III. Nonparametric bayesian sparse hierarchical factor modeling and regression. In *Advances in Neural Information Processing Systems*, 2008.

[14] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proc. International Conference on Machine Learning*, 2007.

[15] M. Ranzato, C. Poultney, S. Chopra, and Y. Lecun. Efficient learning of sparse representations with an energy-based model. In *Proc. Neural Information Processing Systems*, 2006.

[16] R. Thibaux and M.I. Jordan. Hierarchical beta processes and the indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 2007.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58:267–288, 1996.

[18] M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, June 2001.

[19] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis Machine Intelligence*, 2009.

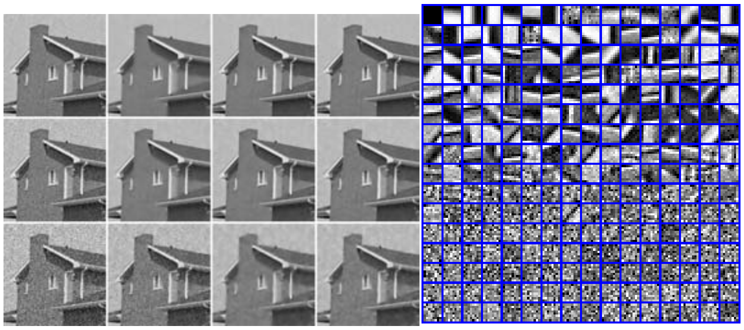

Fig. 1. Left: Representative denoising results, with the top through bottom rows corresponding to noise standard deviations of 15, 25 and 50, respectively. The second and third columns represent K-SVD [5] results with assumed standard deviation equal to 30 and the ground truth, respectively. The fourth column represents the proposed BP reconstructions. The noisy images are in the first column. Right: Inferred BP dictionary elements (mean from the posterior) for noise standard deviation 25, in order of importance from the top-left.

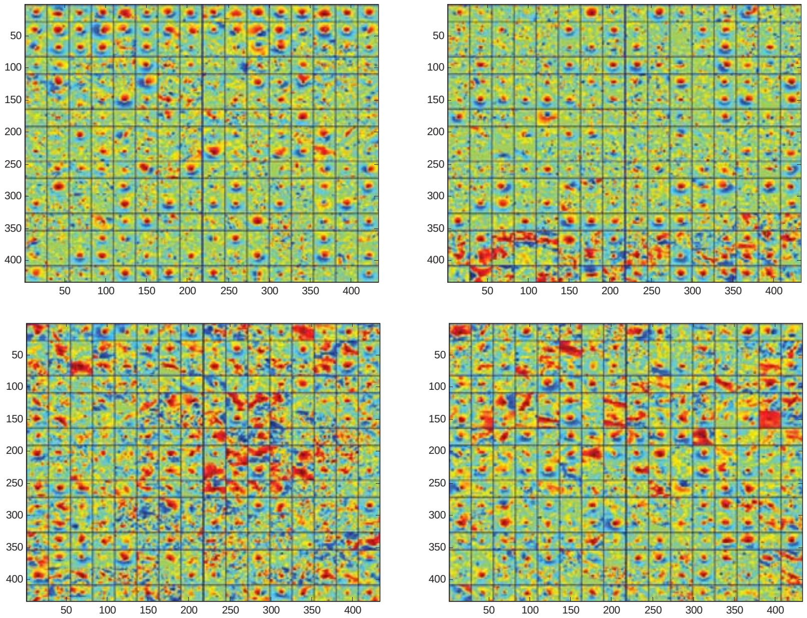

Fig. 2. Measured magnetometer data from buried UXO and clutter.

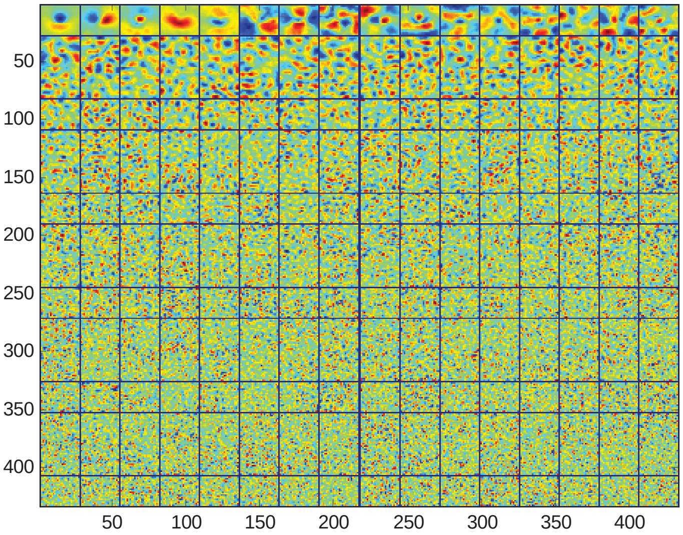

Fig. 3. Learned dictionary for the measured data in Figure 2. Only the top line of dictionary elements are of interest, underscoring the compact representation of the dictionary (the rest are simply draws from the prior, since they are not used to represent any of the data).