Non-Parametric Bayesian Dictionary Learning for Sparse Image Representations

Mingyuan Zhou, Haojun Chen, John Paisley, Lu Ren,
[1]Guillermo Sapiro and Lawrence Carin

Department of Electrical and Computer Engineering
Duke University, Durham, NC, USA
[1]Department of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN, USA

NIPS, December 2009

Outline

- Introduction
- Dictionary learning
- Model and inference
- Image denosing
- Image inpainting
- Compressive sensing
- Conclusions

Introduction

- Sparse representations: simple models, interpretable dictionary elements and sparse coefficients.
- Applications: Image denosing, inpainting, and compressive sensing.
- “Off-the-shelf” bases/dictionaries.
- Over-complete dictionary matched to the signals of interest may improve performance.

Introduction: *sparse coding*

DUKE ECE

- Objective function:

  Given $\mathbf{D} \in \Re^{P \times K}$ and $\boldsymbol{x} \in \Re^{P}$

  $$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0 \quad \text{subject to } \|\boldsymbol{x} - \mathbf{D}\boldsymbol{w}\|_2^2 \leq \epsilon$$

- Exact solution: a NP-hard problem
- Approximate solutions:
  - Greedy algorithms (OMP)
  - Convex relaxation approaches (Lars, Lasso, BCS)
- Sparse representation under an appropriate dictionary: data recovery

Introduction: *dictionary learning*

- "Off-the-shelf" bases/dictionaries
  - DFT, DCT, Wavelet
  - Simple, fast computation

- Dictionaries adapted to the data under test
  - Improved performance
  - Better interpretation

Dictionary Learning: *General Approach*

- Global objective function

$$\min_{\mathbf{D},\mathbf{W}}\{\|\mathrm{X}-\mathbf{DW}\|\} \quad \text{subject to } \forall i, \|\boldsymbol{w}_i\|_0 \le T_0$$

- Sparse coding stage (fix the dictionary)

$$\min_{\boldsymbol{w}_i} \|\boldsymbol{w}_i\|_0 \quad \text{subject to } \|\boldsymbol{x}_i - \mathbf{D}\boldsymbol{w}_i\|_2^2 \le C\sigma^2$$

$$\text{or} \quad \min_{\boldsymbol{w}_i} \|\boldsymbol{x}_i - \mathbf{D}\boldsymbol{w}_i\|_2^2 \quad \text{subject to } \|\boldsymbol{w}_i\|_0 \le T_0$$

- Dictionary update stage
  - Method of optimal direction, MOD (fix the sparse codes):

$$\mathbf{D} = \mathbf{X}\mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$$

  - K-SVD (fix the sparsity pattern, rank-1 approximation):

$$\tilde{\boldsymbol{d}}_k\tilde{\boldsymbol{w}}_{k:} \approx \mathrm{X} - \sum_{j\neq k} \boldsymbol{d}_j\boldsymbol{w}_{j:}$$

Dictionary Learning: *Restrictions and solutions*

- Restrictions of previous dictionary learning approaches:
  - The noise variance or sparsity level are assumed to be known.
  - The size of the dictionary need to be set *a priori*.
  - Only point estimates are provided.

- How to relax these restrictions?
  - Introduce a non-parametric Bayesian dictionary learning approach.
  - Use sparsity promoting priors instead of enforcing the sparsity level/noise variance.
  - Preset a large dictionary size and let the data itself infer an appropriate dictionary size.

Dictionary Learning with Beta process Priors

- Representation (naive form):

$$\boldsymbol{x}_i = \mathbf{D}\boldsymbol{z}_i + \boldsymbol{\epsilon}_i \qquad \mathbf{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_K]$$

- Beta process formulation:

$$H \sim \mathrm{BP}(a_0, b_0, H_0) \qquad H(\boldsymbol{d}) = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{d}_k}(\boldsymbol{d}) \quad \boldsymbol{d}_k \sim H_0$$

$$\pi_k \sim \mathrm{Beta}(a_0/K, b_0(K-1)/K)$$

- Binary weights:

$$z_{ik} \sim \mathrm{Bernoulli}(\pi_k)$$

- Representation (with pseudo weights):

$$\boldsymbol{x}_i = \mathbf{D}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i \qquad \boldsymbol{w}_i = \boldsymbol{z}_i \odot \boldsymbol{s}_i \qquad \boldsymbol{s}_i \sim \mathcal{N}(0, \gamma_s^{-1}\mathbf{I}_K)$$

Hierarchical model

- Data are fully observed

$$
\begin{aligned}
\boldsymbol{x}_i &= \mathbf{D}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i & \pi_k &\sim \mathrm{Beta}(a_0/K, b_0(K-1)/K) \\
\boldsymbol{w}_i &= \boldsymbol{z}_i \odot \boldsymbol{s}_i & \boldsymbol{s}_i &\sim \mathcal{N}(0, \gamma_s^{-1}\mathbf{I}_K) \\
\boldsymbol{d}_k &\sim \mathcal{N}(0, P^{-1}\mathbf{I}_P) & \boldsymbol{\epsilon}_i &\sim \mathcal{N}(0, \gamma_\epsilon^{-1}\mathbf{I}_P) \\
\boldsymbol{z}_i &\sim \prod_{k=1}^{K} \mathrm{Bernoulli}(\pi_k) & \gamma_s &\sim \Gamma(c_0, d_0) \\
& & \gamma_\epsilon &\sim \Gamma(e_0, f_0)
\end{aligned}
$$

- Data are partially observed

$$\boldsymbol{y}_i = \boldsymbol{\Sigma}_i \boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_i^T = \mathbf{I}_{||\boldsymbol{\Sigma}_i||_0} \quad ||\boldsymbol{\Sigma}_i^T \boldsymbol{\Sigma}_i||_0 = ||\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_i^T||_0 = ||\boldsymbol{\Sigma}_i||_0$$

Hierarchical model

- Full likelihood

$$
\begin{aligned}
&P(\mathbf{Y}, \boldsymbol{\Sigma}, \mathbf{D}, \mathbf{Z}, \mathbf{S}, \boldsymbol{\pi}, \gamma_s, \gamma_\epsilon) \\
&= \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\Sigma}_i \mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i), \gamma_\epsilon^{-1} \mathbf{I}_{\|\boldsymbol{\Sigma}_i\|_0}) \mathcal{N}(\mathbf{s}_i; 0, \gamma_s^{-1} \mathbf{I}_K) \\
&\prod_{k=1}^{K} \mathcal{N}(\boldsymbol{d}_k; 0, P^{-1} \mathbf{I}_P) Beta(\pi_k; a_0, b_0) \\
&\prod_{i=1}^{N} \prod_{k=1}^{K} Bernoulli(z_{ik}; \pi_k) \\
&\Gamma(\gamma_s; c_0, d_0) \Gamma(\gamma_\epsilon; e_0, f_0)
\end{aligned}
$$

- Gibbs Sampling Inference

Model comparison

- MOD

  two stages: dictionary learning, sparse coding.
- K-SVD

  two stages: dictionary learning (enforced sparsity pattern), sparse coding.
- Dictionary learning with beta process priors

  three stages: dictionary learning (enforced sparsity pattern), sparsity pattern update, pseudo weights update.
- The three models have apparent differences in the level of exploiting previous obtained information.

Sequential learning for large data sets

- Partitioning the whole date set to be

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots \mathcal{D}_{J-1} \cup \mathcal{D}_J$$

Instead of directly calculating

$$p(\mathbf{D}|\mathcal{D}, \mathbf{\Theta})$$

we first calculate

$$p(\mathbf{D}|\mathcal{D}_1, \mathbf{\Theta})$$

The posterior is then used as prior for **D** for calculating

$$p(\mathbf{D}|\mathcal{D}_1 \cup \mathcal{D}_2, \mathbf{\Theta})$$

Non-Parametric Bayesian Dictionary Learning

- The noise variance/sparsity level need not be known.
- The dictionary size is automatically inferred.
- Training data are not required.
- The average sparsity level of the representation is inferred from the data itself, and based on the posterior, each sample x_i has its own unique sparse representation.
- A single model applicable for gray-scale, RGB, and hyperspectral image denoising & inpainting.

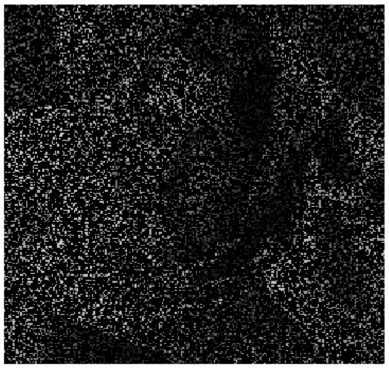
Image denoising

Original Noisy Image (dB)	K-SVD Denoising mismatched variance (dB)	K-SVD Denoising matched variance (dB)	Beta Process Denoising (dB)
24.58	30.67	34.32	34.52
20.19	31.52	32.15	32.19
14.56	19.60	27.95	27.95

Image inpainting

80% Pixels Missing

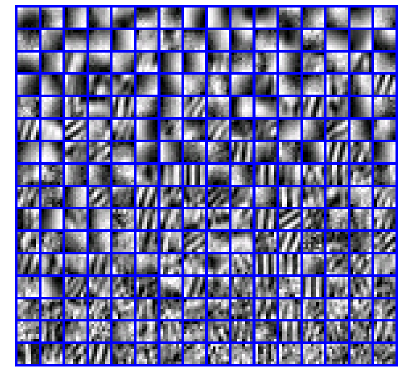
Corrupted image



Original image

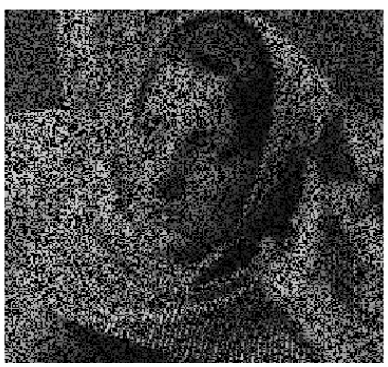
Restored image

Dictionary



50% Pixels Missing

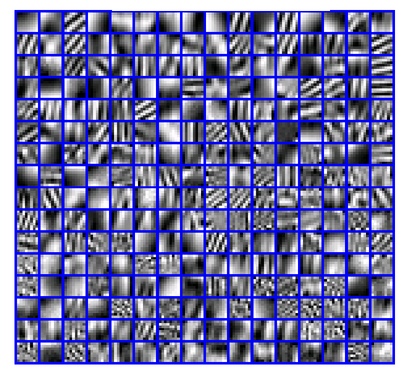
Corrupted image



Original image

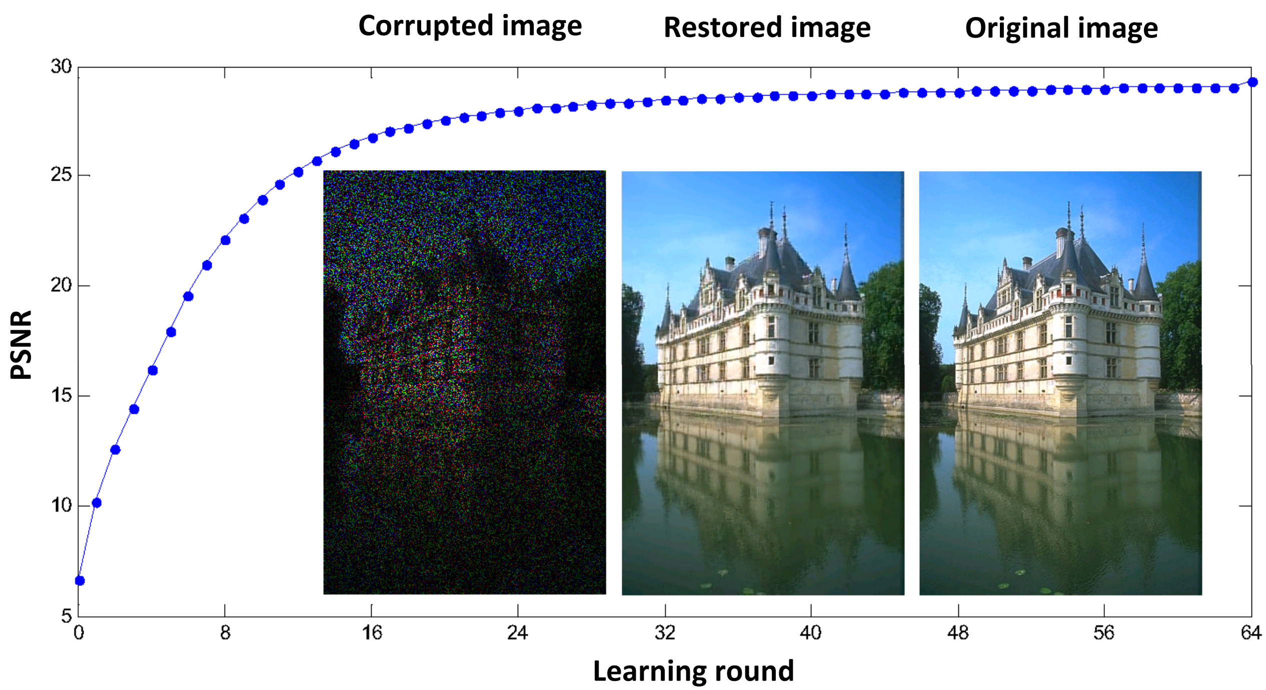
Restored image

Dictionary

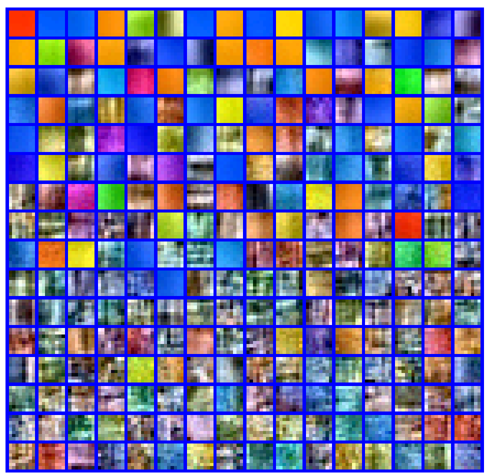


RGB image inpainting

- 480*321 RGB image, 80% missing



Dictionary



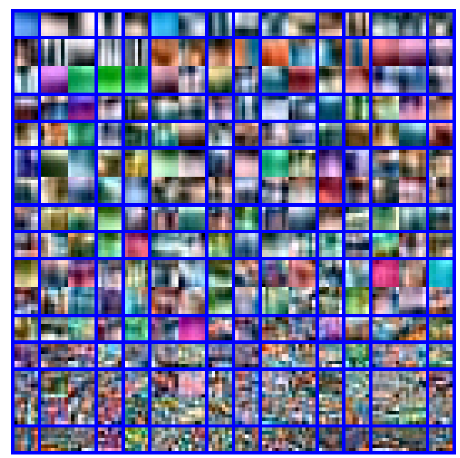
RGB image inpainting

Original image

Corrupted image

Restored image

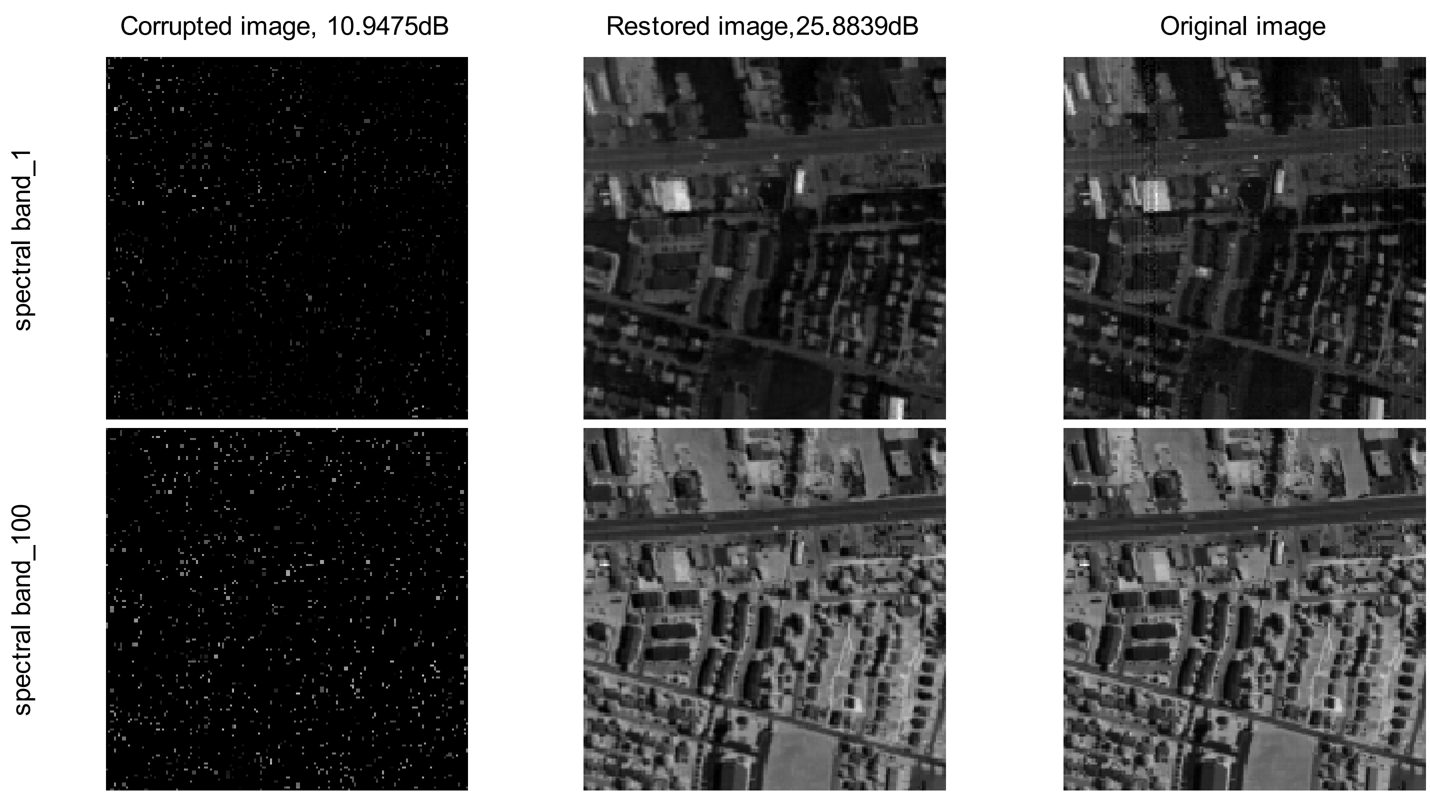
Dictionary



Hyperspectral image inpainting

150*150*210 hyperspectral urban image

95% missing

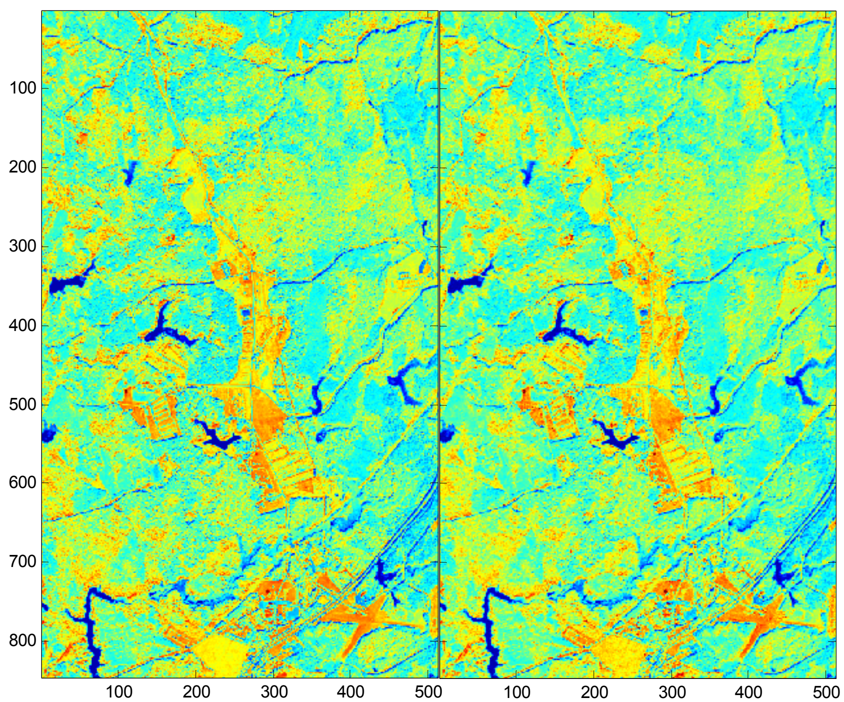


Hyperspectral image inpainting

845*512*106 hyperspectral image

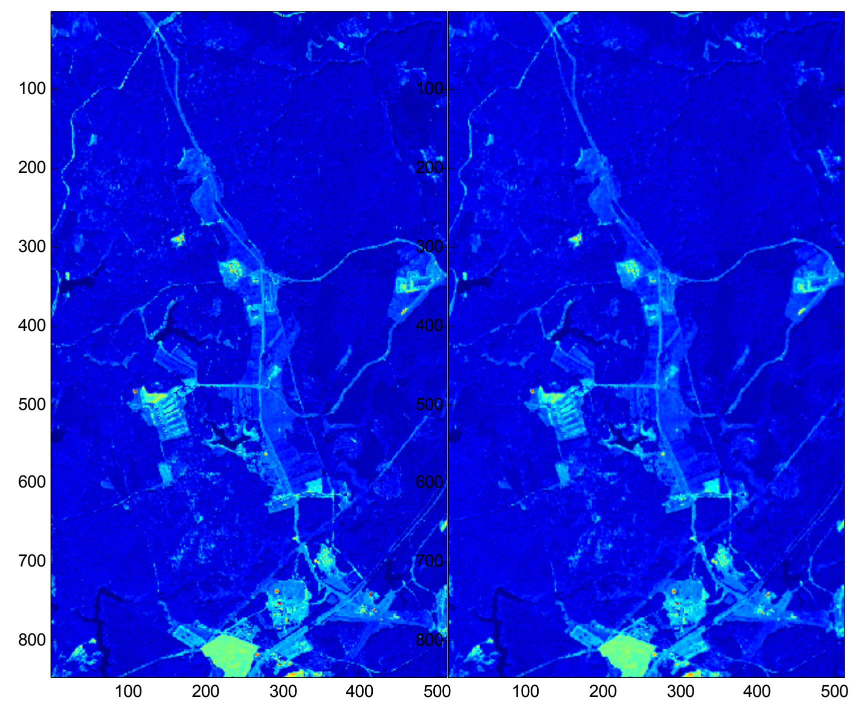
98% missing

Spectral band 50



Original Restored

Spectral band 90

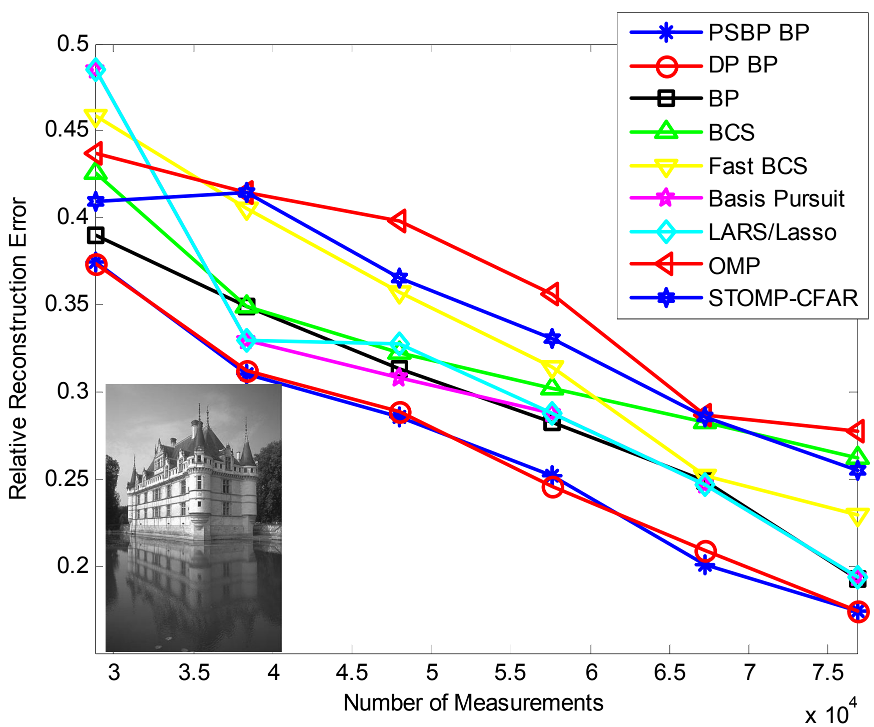


Original Restored

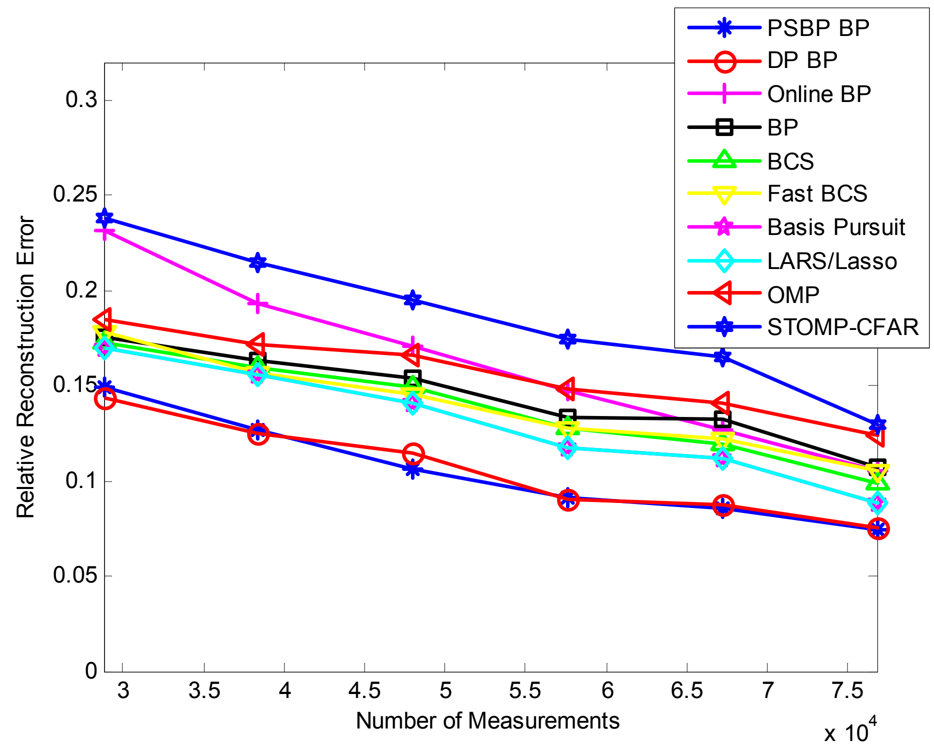
Compressive sensing

Image size: 480 by 320, 2400 8 by 8 patches
153600 coefficients are estimated

DCT



Learned Dictionary



Conclusions

- Non-parametric Bayesian dictionary learning.
- Gray-scale, RGB, and hyperspectral Image denoising, inpainting, and compressive sensing.
- Automatically inferred dictionary size, noise variance and sparsity level.
- Dictionary learning and data reconstruction on the data under test.
- A generative approach for data recovery from redundant noisy and incomplete observations.

References

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

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

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 1994.

[4] B.A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37, 1998.

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

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

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

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

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

[10] 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.

[11] E. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Information Theory*, 52, 2006.

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

[13] 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*, 31, 2009.

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

[15] 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.

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

[17] J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *Proc. International Conference on Machine Learning*, 2009.

[18] T. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 1973.

[19] A. Rodriguez and D.B. Dunson. Nonparametric bayesian models through probit stickbreaking processes. *Univ. California Santa Cruz Technical Report*, 2009.

[20] 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.

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

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

[23] M. Girolami and S. Rogers. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18, 2006.

[24] R.G. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24, 2007.

[25] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 1994.