DENOISING OF IMAGE PATCHES VIA SPARSE REPRESENTATIONS WITH LEARNED STATISTICAL DEPENDENCIES

Tomer Faktor, Yonina C. Eldar, and Michael Elad

Departments of Electrical Engineering and Computer Science
Technion – Israel Institute of Technology, Haifa 32000, Israel
{tomerfa@tx, yonina@ee, elad@cs}.technion.ac.il

ABSTRACT

We address the problem of denoising for image patches. The approach taken is based on Bayesian modeling of sparse representations, which takes into account dependencies between the dictionary atoms. Following recent work, we use a Boltzman machine to model the sparsity pattern. In this work we focus on the special case of a unitary dictionary and obtain the exact MAP estimate for the sparse representation using an efficient message passing algorithm. We present an adaptive model-based scheme for sparse signal recovery, which is based on sparse coding via message passing and on learning the model parameters from the data. This adaptive approach is applied on noisy image patches in order to recover their sparse representations over a fixed unitary dictionary. We compare the denoising performance to that of previous sparse recovery methods, which do not exploit the statistical dependencies, and show the effectiveness of our approach.

Index Terms— Sparse representations, MAP, Boltzmann machine, unitary dictionary, message passing, image denoising.

1. INTRODUCTION

Signal modeling based on sparse representations is used in numerous signal and image processing applications, such as denoising, restoration, source separation, compression and sampling (for a comprehensive review see [1]). The core idea in all of these applications is to recover a sparse representation of the signal over a pre-specified dictionary of atoms.

It is very common to treat these atoms independently from each other. This is an implicit assumption in standard recovery algorithms such as orthogonal matching pursuit (OMP) [2] and it is assumed explicitly in many Bayesian approaches for signal recovery (see for example [3]). Note however that the independence assumption limits the representation power of the signal model and is inaccurate for many types of real-life signals. This motivates the formulation of structured sparsity models. In image processing this structure typically takes the form of wavelet trees (see for example [4]).

A recent work [5] suggested using a Bayesian model for the sparse representations, where the sparsity pattern is modeled by a Boltzmann machine (BM). This is a commonly used Markov random field (MRF) for capturing statistical dependencies in a general and flexible manner. The nonzero representation coefficients are modeled by Gaussian distributions with atom-dependent variances.

The current work relies on the proposed BM generative model. In this context, we address two specific questions: how to perform a pursuit algorithm for finding the sparse representation of a given noisy signal and how to estimate the BM parameters given a set of sparse supports. We suggest using a message passing algorithm for an exact maximum *a posteriori* (MAP) estimation of the sparse representation, when the dictionary is unitary and the interaction matrix is banded. As for learning the BM parameters, we adopt a maximum pseudo-likelihood (MPL) approach, which allows for solving this problem using convex optimization methods. For more details on these topics see [6].

In this work we consider the problem of evaluating both the model parameters and the sparse representations from a given set of signals. The joint estimation problem is handled using a block-coordinate relaxation technique that iteratively updates the representations, and then the parameters. We demonstrate the effectiveness of the BM model and the estimated parameters on denoising of image patches.

2. BM-BASED MAP RECOVERY

The main goal of this paper is exploiting statistical dependencies between the dictionary atoms used for the sparse representation, in order to improve the denoising performance of image patches. In order to set the ground for the image experiments, we first provide a brief review of the proposed generative model, the MAP estimation problem that follows from it and an efficient algorithm that solves this problem under some modeling assumptions.

We consider a signal y which is built as $y = Ax+e$, where A is a unitary dictionary of size m-by-m, x is a sparse representation over this dictionary and e is additive white Gaussian noise with variance σ_e^2. The idea to model the noise-free signal y_0 as Ax with a sparse representation vector x is a very common and long-studied concept in signal and image processing. We denote the sparsity pattern of x by $S \in \{-1, 1\}^m$, where $S_i = 1$ implies that the index i belongs to the support of x, whereas $S_i = -1$ implies that $x_i = 0$. The nonzero coefficients of x are denoted by x_s, where s is the support of x. We assume a Gaussian distribution with zero mean and variance $\sigma_{x,i}^2$ for each nonzero representation coefficient x_i.

We now turn to describe the prior distribution on S. MRFs are a useful tool for capturing statistical dependencies between a set of random variables. An MRF provides a full and concise description for the joint distribution of these variables, where the statistical dependencies are conveniently displayed in the form of an undirected graph. We focus on the special case of a BM, which can serve as a useful and powerful prior on S. The BM distribution is given by:

$$\Pr(S) = \frac{1}{Z} \exp\left(b^T S + \frac{1}{2} S^T W S\right), \tag{1}$$

This work was supported by the Israel Science Foundation under Grants 1081/07 and 599/08, by the European Commission's FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (grant agreement no. 216715) and by the EU Framework 7 FET-Open project FP7-ICT-225913-SMALL: Sparse Models, Algorithms and Learning for Large-Scale data.

where W is symmetric with zero entries on the main diagonal and Z is a constant which normalizes the distribution. The BM distribution can be easily represented by an MRF - a bias b_i is associated with a node i and a nonzero entry W_{ij} in the interaction matrix results in an edge connecting nodes i and j with the specified weight. Consequently, the zero entries in W have the simple interpretation of missing edges in the corresponding graph.

In our setup we assume a BM prior on S with a low order banded W matrix. In an Lth order banded matrix only the $2L+1$ principle diagonals consist of nonzero elements. This type of BMs can serve as a useful relaxation for a general dependency model, as they can achieve a substantial decrease in computational complexity, while still capturing the significant dependencies. More specifically, inference tasks like finding the most probable configuration can be computed exactly and efficiently in this case by means of a message passing algorithm.

Next, we turn to describe the design objective for the sparse recovery problem. Our goal is to recover x given y. However, due to algorithmic considerations implied by our Bayesian framework, we suggest to first perform MAP estimation of S given y and then proceed with MAP estimation of x given y and the estimated support $\hat{s}$. The latter takes a simple closed-form formula, as in the oracle estimator (see [3]):

$$\hat{x}_{s_{MAP}} = \underset{x_s \in \mathbb{R}^k}{\text{argmax}} \Pr(x|y,\hat{s}) = \left(A_{\hat{s}}^T A_{\hat{s}} + \sigma_e^2 \Sigma_{\hat{s}}^{-1}\right)^{-1} A_{\hat{s}}^T y. \quad (2)$$

For MAP estimation of S given y we can take advantage of a useful theorem that holds for a unitary dictionary: The BM distribution is a conjugate prior for this problem, namely $\Pr(S|y)$ is a BM with the same interaction matrix W and a modified bias vector q, which depends on the original bias vector b, the variances $\left\{\sigma_{x,i}^2\right\}_{i=1}^m$, the noise variance σ_e^2, the dictionary atoms and the signal y. It follows that finding the MAP estimator for S becomes an inference task on a BM with parameters W, q. Using our modeling assumption on W (banded with a low order L), we have that the MAP problem can be solved exactly by message passing. A proof of the above mentioned theorem and a concrete message passing algorithm for this setup are given in [6].

3. ADAPTIVE SPARSE SIGNAL RECOVERY

In an actual problem suite we are given a set of signals $\left\{y^{(l)}\right\}_{l=1}^N$ from which we would like to estimate both the sparse representations and the model parameters. We address this joint estimation problem in this section. Note that throughout this section we will assume that the unitary dictionary A and the noise variance σ_e^2 are known.

We begin by assuming that the sparse representations are known, namely we are given a data set of independent and identically distributed (i.i.d.) examples $\mathcal{D} = \left\{y^{(l)}, x^{(l)}, S^{(l)}\right\}_{l=1}^N$, from which we would like to learn the model parameters $\Theta = \left[W, b, \left\{\sigma_{x,i}^2\right\}_{i=1}^m\right]$. To estimate Θ we suggest a maximum likelihood (ML) approach, and using the BM generative model we can write:

$$\hat{\Theta}_{ML} = \underset{\Theta}{\text{argmax}} \ \Pr\left(\mathcal{D}|\Theta\right) = \underset{\Theta}{\text{argmax}} \sum_{i=1}^m \mathcal{L}(\sigma_{x,i}^2) + \mathcal{L}(W,b), \quad (3)$$

where

$$\mathcal{L}(\sigma_{x,i}^2) = \frac{1}{2}\sum_{l=1}^N \left[\frac{1}{\sigma_{x,i}^2}\left(x_i^{(l)}\right)^2 + \ln\left(\sigma_{x,i}^2\right)\right] \mathbf{1}\left[i \in s^{(l)}\right] \quad (4)$$

$$\mathcal{L}(W,b) = \frac{1}{2}\sum_{l=1}^N \ \left[\left(S^{(l)}\right)^T W S^{(l)} + b^T S^{(l)}\right] - N\ln(Z(W,b)) \quad (5)$$

are the log likelihood functions for the model parameters. This decomposition allows separate estimation of the variances $\{\sigma_{x_i}^2\}_{i=1}^m$ and the Boltzmann parameters W, b.

Starting with the variances we have the close-form estimator:

$$\hat{\sigma}_{x,i}^2 = \left(\sum_{l=1}^N \left(x_i^{(l)}\right)^2 \mathbf{1}\left[i \in s^{(l)}\right]\right) / \left(\sum_{l=1}^N \mathbf{1}\left[i \in s^{(l)}\right]\right). \quad (6)$$

Similar estimators for the variances were also used in [5].

ML estimation of W, b is computationally intensive due to the exponential complexity in m associated with the partition function $Z(W,b)$. Therefore, we turn to approximated ML estimators. A widely used approach is applying Gibbs sampling and mean-field techniques, as practiced in [5]. However, these methods are usually computationally demanding. A simpler approach is to replace the likelihood function by a pseudo-likelihood, leading to MPL estimation [7]. The basic idea is to replace the BM prior $\Pr(S|W,b)$ by the product of all the conditional distributions of each node S_i given the rest of the nodes S_{i^c}: $\prod_{i=1}^m \Pr\left(S_i|S_{i^c}, W, b\right)$. This leads to the following log-PL objective (up to an additive constant):

$$\mathcal{L}_p(W,b) = \sum_{l=1}^N \left(S^{(l)}\right)^T \left(WS^{(l)} + b\right) - \mathbf{1}^T \rho\left(WS^{(l)} + b\right) \quad (7)$$

where $\rho(z) = \ln(\cosh(z))$ and the function $\rho(\cdot)$ operates on a vector entry-wise. This function is concave with respect to W, b, implying that maximization of this function can be done efficiently using convex optimization techniques.

We suggest using the sequential subspace optimization (SESOP) method [8] for the MPL estimation problem. This method is very useful for large-scale unconstrained convex problems, as opposed to gradient descent which suffers from a slow convergence rate and Newton iterations which does not scale up well. The core idea is to optimize over a low-dimensional subspace spanned by several recent update directions and the current gradient. In each iteration of SESOP we perform an inner optimization stage to find the steps sizes in each direction, using Newton method.

When W is constrained to be banded, we suggest a post-processing of the MPL estimate. More specifically, we define the energy of W as the l_1 norm for the entries in the banding zone. The basic idea is to perform pairwise permutations in $\hat{W}$, namely switch the roles of pairs of atoms, so that the energy will be maximal. A greedy approach can be used, so that in each iteration we replace the roles of one pair of atoms, where this replacement is optimal in the sense of maximizing the energy. The algorithm converges when we cannot increase the energy anymore. At this point we set all entries located outside the banding zone to zero. The suggested post-processing stage serves as a projection onto the banding constraint. Note that the estimated biases and variances should be modified to account for the changes in the atom roles.

We now turn to the joint estimation problem, where both the sparse representations and the model parameters are unknown. We suggest using a block-coordinate optimization approach for approximating the solution of the joint estimation problem, which results in an iterative scheme for adaptive sparse signal recovery. Each iteration in this scheme consists of two stages. The first is sparse coding where we apply the proposed message passing algorithm and then (2) to obtain MAP estimates for the sparse representations based on the most recent estimate for the model parameters. This is followed by model update where we re-estimate all the model parame-

σ_e	Lena		Barbara		Boats		House		Peppers		Average	
2	2.47	2.43	2.63	2.46	2.7	2.59	2.23	2.19	2.81	2.55	2.58	2.45
	2.39	**2.24**	2.33	**2.22**	2.43	**2.31**	2.17	**2.01**	2.42	**2.3**	2.35	**2.22**
5	4.3	4.29	4.83	4.78	5.09	5.05	4.15	3.97	4.86	4.85	4.66	4.61
	4.18	**3.92**	4.61	**4.31**	4.91	**4.64**	4.03	**3.83**	4.7	**4.32**	4.5	**4.21**
10	6.43	6.26	7.91	7.56	7.8	7.59	6.28	5.9	7.63	7.45	7.24	6.99
	6.1	**5.98**	7.51	**7.22**	7.37	**7.09**	6.02	**5.74**	7.29	**6.92**	6.89	**6.62**
15	8.14	7.82	10.47	**9.62**	9.89	9.48	7.8	7.15	9.92	9.41	9.31	8.75
	7.58	**7.27**	9.9	9.67	9.25	**8.86**	7.4	**7.1**	9.3	**8.83**	8.74	**8.41**
20	9.6	9.21	12.61	**11.34**	11.61	11.05	9.21	**8.21**	11.83	11.08	11.05	10.25
	8.85	**8.59**	11.87	12.04	10.8	**10.57**	8.68	8.43	10.88	**10.53**	10.29	**10.12**
25	10.89	10.36	14.43	**12.89**	13.12	12.5	10.63	**9.35**	13.57	12.52	12.62	11.61
	9.95	**9.84**	13.5	13.72	12.12	**11.92**	9.86	9.57	12.34	**12.13**	11.64	**11.54**

Table 1. Summary of average denoising results (Root-MSE per pixel). In each cell four denoising results are reported. Top left: Unitary OMP. Top right: K-SVD. Bottom left: Independent adaptive recovery. Bottom right: BM adaptive recovery.

ters given the current estimate of the sparse representations. We use (6) for the variances and MPL estimation via SESOP for the Boltzmann parameters. Then the post-processing stage mentioned above is applied on these estimates.

4. SIMULATIONS ON IMAGE PATCHES

In this section we address real-life signals - patches of size 8-by-8 that are extracted out of natural images. The adaptive BM-based sparse recovery scheme that was suggested in the previous section is applied on noisy image patches using a unitary discrete cosine transform (DCT) dictionary, in order to demonstrate the effectiveness of our approach. Note that we are not suggesting here an improved image denoising algorithm, and in contrast to common denoising methods, we do not exploit self-similarities in the image (see for example [9]). Therefore our comparison is limited to denoising schemes that recover each patch separately.

We compare our algorithm to three denoising schemes: OMP with a unitary DCT dictionary, K-SVD which uses an adaptive dictionary with a redundancy factor of 4 as in [10], and an adaptive sparse recovery approach which is based on a unitary DCT dictionary and an independent-based prior as in [3]. In the latter each atom is assigned a different prior probability p_i to be turned "on", which is estimated from the data using $p_i \approx \frac{1}{N}\sum_{l=1}^{N} \mathbf{1}\left[S_i^{(l)}=1\right]$ for all i, where $\mathbf{1}[\cdot]$ is the indicator function.

To initialize the parameters of the two adaptive model-based approaches, we set all the variances to 50^2 and use an i.i.d. prior on the support, namely $\Pr(S_i=1)=p$ for all i. This prior is obtained by the Boltzmann parameters $\hat{W}=\mathbf{0}^{m\times m}$ and $\hat{b}_i=0.5\ln({}^{p}/_{(1-p)})$. Note that p has the intuitive meaning of the ratio ${}^{k}/_{m}$ where k is our prior belief on the mean cardinality of the support. We use a prior belief that the average cardinality for image patches is $k=10$, which leads to initial biases $b_i=-0.84$ for all i. We then perform two iterations for each of the adaptive schemes. In the BM-based scheme W is constrained to be a 9th order banded matrix. As for denoising via OMP, it consists of one iteration where we apply the OMP algorithm using the unitary DCT dictionary. Finally, for K-SVD denoising we set an overcomplete DCT dictionary for the initialization, following the suggestion in [10], and apply 10 iterations with OMP for the sparse coding stage. Throughout this section we use the abbreviations "unitary OMP", "K-SVD", "independent adaptive recovery" and "BM adaptive recovery" to denote the four methods.

Average denoising errors per pixel and per patch are evaluated for all the four methods on 5 widely used test images and 6 noise levels: $\sigma_e \in \{2,5,10,15,20,25\}$. For each test image we extract overlapping patches and for each patch we apply a pre-processing stage of DC removal by subtracting the average value of the noisy patch, so that a smooth patch corresponds to a zero representation. A summary of the denoising results is given in Table 1.

We begin by focusing on the three methods which are based on a fixed unitary DCT dictionary. The experiments show that the quality of the denoising improves as we use a more elaborate prior on the support. For the unitary OMP which makes use of the sparsity assumption alone, the denoising performance is the worst. Independent adaptive recovery achieves a significant improvement with respect to unitary OMP: the average gain varies from $0.3[dB]$ to $0.8[dB]$ for the different noise levels. When we turn to BM adaptive recovery, we get that this method outperforms the independent-based one for all tested images and noise levels, apart from the image 'Barbara' with noise levels $\sigma_e \geq 20$. The performance gaps vary from $0.1[dB]$ to $0.6[dB]$ for the different noise levels.

Next, we compare the denoising performance of the two adaptive model-based schemes with that of K-SVD. We can see that BM adaptive recovery succeeds in outperforming K-SVD for most tested images and noise levels (25 out of 30 experiments), despite the fact that K-SVD has 4 times more atoms. Note that for low noise levels ($\sigma_e \leq 10$) there are significant performance gaps - the average gain is $0.6[dB]$ [1]. In fact, even independent adaptive recovery is superior to K-SVD for most experiments (22 out of 30), but here the performance gaps are much less significant.

To give a visual flavor to the numerical observations that were provided above, we show in Fig. 1 a result on one noisy image patch that demonstrates a typical scenario where our approach outperforms others methods. Note that the scale is adjusted to the dynamic range of the given image patch, in order to make the visual differences more coherent. We can see that unitary OMP uses 6 atoms for the recovery: 3 correspond to low frequencies and the other 3 describe complex textures, which are associated more with the noise than the signal. For K-SVD which learns the dictionary from the data 3 atoms are sufficient and the recovery is improved to some extent.

[1] For image denoising the recovered overlapping patches should be averaged, as practiced in [10], in order to prevent artifacts on block boundaries and boost the global denoising performance on the entire image. The performance gaps observed in the denoising of image patches vanish when we apply the averaging process described above and examine the performance on the resulting image. This is a known phenomena, which has also been observed in previous works on the MMSE estimator [3]. Further work is required to maintain the performance gap after the averaging process.

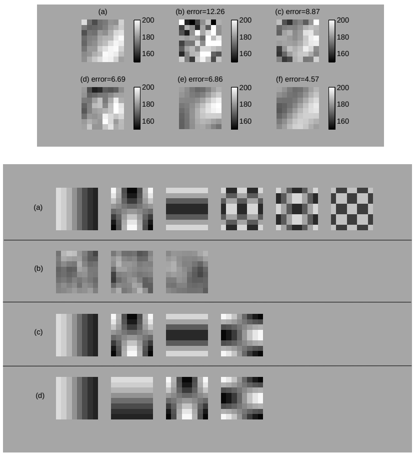

Fig. 1. Results for one patch of the image 'Lena' with a noise level $\sigma_e = 10$. Top - image patches: (a) Noise-free, (b) Noisy, (c) Recovery with unitary OMP, (d) Recovery with K-SVD, (e) Independent adaptive recovery, (f) BM adaptive recovery. Bottom - the recovered atoms for: (a) Unitary OMP, (b) K-SVD, (c) Independent adaptive recovery, (d) BM adaptive recovery.

The model-based adaptive schemes obtain improved recoveries in respect to the unitary OMP by eliminating the 3 atoms which are associated with the noise and replacing them by other atoms which are more adequate. The BM adaptive recovery exploits the dependencies to make a more educated choice for the atom replacement.

To conclude this section we show some results that are related with the Boltzmann parameters that were learned from a corpus of noisy patches. Fig. 2 shows such results for the test image 'peppers' with a noise level $\sigma_e = 10$. On the top we can see the 5 atoms which are characterized by the highest biases $\hat{b}_i$. These atoms correspond to low frequencies. On the middle, the 5 pairs of atoms which correspond to the strongest "excitatory" interactions ($\hat{W}_{ij} > 0$) are shown. We see that for each pair of atoms we typically have a high resemblance between the patterns of the two atoms. This explains their tendency to be used or not used together. Similarly, we can see on the bottom the 5 pairs of atoms which correspond to the strongest "inhibitory" interactions ($\hat{W}_{ij} < 0$). Here we have that the atoms in each pair correspond to patterns with very different natures.

5. CONCLUSIONS AND FUTURE WORK

In this work we have developed a scheme for adaptive model-based recovery of sparse representations that takes into account atom dependencies. We adapted a Bayesian model for these representations, which is based on a Boltzmann machine, and designed efficient estimators for the model parameters. We demonstrated the effectiveness of our proposed approach by real-life experiments on noisy image patches. In these real-life experiments we considered two approaches for adaptive sparse recovery: one is based on dictionary training and the sparsity assumption alone, and the second uses an adaptive and meaningful prior for signal modeling, while the dictionary remains fixed to some reasonable setting. We derived an exciting observation - in terms of denoising performance, we can often benefit more from introducing a well-adjusted prior for the selection of atoms in a fixed dictionary than from dictionary training. A research direction we are considering is merging dictionary training into the adaptive scheme, in order to benefit from both the BM generative model and a dictionary which is better fitted to the data.

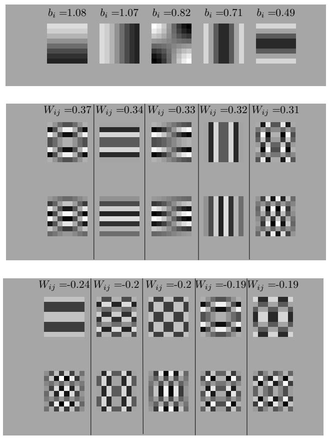

Fig. 2. Results for the Boltzmann parameters that were learned from the patches of the image 'peppers' with a noise level $\sigma_e = 10$. Top: the 5 atoms with the highest biases. Middle: the 5 pairs of atoms with the strongest "excitatory" interactions. Bottom: the 5 pairs of atoms with the strongest "inhibitory" interactions.

6. REFERENCES

[1] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[2] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Systems, and Computers*, 1993, pp. 40–44.

[3] J. S. Turek, I. Yavneh, M. Protter, and M. Elad, "On MMSE and MAP denoising under sparse representation modeling over a unitary dictionary," Tech. Rep., CS Dept., Technion – Israel Institite of Technology, Haifa, Israel, 2010.

[4] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk, "Wavelet-domain compressive signal reconstruction using a hidden Markov tree model," in *ICASSP*, Las Vegas, NV, Apr. 2008.

[5] P. J. Garrigues and B. A. Olshausen, "Learning horizontal connections in a sparse coding model of natural images," in *Advances in Neural Information Processing Systems 20*, 2008, pp. 505–512.

[6] T. Faktor, Y. C. Eldar, and M. Elad, "Exploiting statistical dependencies in sparse representations for signal recovery," *IEEE Trans. Signal Processing*, (submitted).

[7] A. Hyvarinen, "Consistency of pseudolikelihood estimation of fully visible boltzmann machines," *Neural Computation*, vol. 18, no. 10, pp. 2283–2292, 2006.

[8] G. Narkiss and M. Zibulevsky, "Sequential subspace optimization method for large-scale unconstrained optimization," Tech. Rep., EE Dept., Technion – Israel Institite of Technology, Haifa, Israel, 2005.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *ICCV*, Tokyo, Japan, 2009.

[10] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.