LEARNING SPARSE REPRESENTATIONS FOR ADAPTIVE COMPRESSIVE SENSING

Akshay Soni and Jarvis Haupt

University of Minnesota, Twin Cities
Department of Electrical and Computer Engineering
Minneapolis, Minnesota USA 55455
e-mail: {sonix022,jdhaupt}@umn.edu

ABSTRACT

Breakthrough results in compressive sensing (CS) have shown that high dimensional signals (vectors) can often be accurately recovered from a relatively small number of non-adaptive linear projection observations, provided that they possess a sparse representation in some basis. Subsequent efforts have established that the reconstruction performance of CS can be improved by employing additional prior signal knowledge, such as dependency in the location of the non-zero signal coefficients (structured sparsity) or by collecting measurements sequentially and adaptively, in order to focus measurements into the proper subspace where the unknown signal resides. In this paper, we examine a powerful hybrid of adaptivity and structure. We identify a particular form of structured sparsity that is amenable to adaptive sensing, and using concepts from sparse hierarchical dictionary learning we demonstrate that sparsifying dictionaries exhibiting the appropriate form of structured sparsity can be learned from a collection of training data. The combination of these techniques (structured dictionary learning and adaptive sensing) results in an effective and efficient adaptive compressive acquisition approach which we refer to as LASeR (Learning Adaptive Sensing Representations.)

Index Terms— Compressive sensing, adaptive sensing, structured sparsity, principal component analysis

1. INTRODUCTION

Motivated in large part by the surge of research activity in compressive sensing [1–3], the identification of efficient sensing and reconstruction procedures for high dimensional inference problems remains an extremely active research area. The basic problem can be explained as follows. Let $x \in \mathbb{R}^n$ represent some unknown signal, and suppose that x can be accurately represented using only a small number of *atoms* $d_i \in \mathbb{R}^n$ from some *dictionary* D, so that

$$x = \sum_{i \in \mathcal{S}} a_i d_i + \epsilon, \tag{1}$$

where $|\mathcal{S}|$ is small (relative to the ambient dimension n), the a_i are the coefficients corresponding to the relative weight of the contribution of each of the d_i that contribute to the approximation, and the vector ϵ represents a (nominally small) modeling error. The dictionary D may, for example, consist of all of the columns of an orthonormal matrix (eg., a discrete wavelet or Fourier transform matrix), though other representations may be possible (eg., D may be a frame). In any case, when $|\mathcal{S}|$ is small relative to the ambient dimension n, we say that the signal x is *sparse*, or that it possesses a sparse representation in the dictionary D.

Initial results in compressive sensing (CS) established sparse vectors can often be recovered from $m \ll n$ measurements, each in the form of a *randomized* linear combination of the entries of x. The weights associated with each linear combination may, for example, be selected as i.i.d. realizations of zero-mean random variables such as Gaussian or symmetric Bernoulli, and these random measurements can be modeled as inner products between the signal vector and a sequence of randomly generated "test" vectors. Suppose, for the sake of illustration, that D is an orthonormal matrix. Then, the main result of CS is that signals that possess a sparse representation with no more than s nonzero coefficients in this (known) dictionary D can, with high probability, be *exactly* recovered from $m \leq Cs \log n$ so-called *randomized projection* measurements, where $C \geq 0$ is a constant independent of s and n (see, eg., [2, 3]). The salient point is that the number of measurements required for exact reconstruction is on the order of the sparsity s, not the ambient dimension n, and when $s \ll n$ the savings in the number of measurements required for recovery can be quite significant.

A number of subsequent efforts in CS have examined settings where, in addition to being sparse, the representation of x in terms of the dictionary D possesses some additional structure (see, for example, the tutorial article [4] and the references therein). The nonzero coefficients may, for example, occur in clusters, or it may be the case that the presence of a particular coefficient in the representation guarantees the presence of other coefficients, according to some a priori known dependency structure. This latter case of coefficient dependency occurs, for example, in the wavelet coefficients of piecewise smooth signals and many natural images, where the nonzero coefficients cluster across levels of a rooted connected tree. In any case, taking advantage of this so-called *structured sparsity* has been shown to result in further reductions in the number of measurements required for recovery of s-sparse n-dimensional vectors from randomized projections. For example, it was shown in [5, 6] that in these cases $m \leq C'k$ randomized measurements suffice for exact recovery, where $C' \geq 0$ is another constant. The savings in this case amount to a reduction in the scaling behavior by a factor of $\log n$, which can itself be a significant savings when n is large.

Several techniques for implementing some form of feedback in the compressive measurement process have also been examined recently in the CS literature. These so-called sequential *adaptive* measurement procedures attempt to glean some information about the unknown signal from initial compressive measurements, which is then used to *shape* subsequent test vectors in order to *focus* more directly on the subspace in which the signal resides. Adaptive CS

This work was supported by grant DARPA/ONR N66001-10-1-4090.

procedures have been shown to provide an improved resilience (relative to traditional CS) in the presence of additive measurement noise (see, for example, [7–9], as well as the summary article [10] and the references therein).

In this paper, we examine a powerful hybrid of the notions of structured sparsity and adaptivity. Our approach, which we refer to as **L**earning **A**daptive **Se**nsing **R**epresentations, or **LASeR**, entails the *identification* of dictionaries in which each element in a given collection of training data exhibits a special form of structured sparsity that is amenable to a particular adaptive compressive measurement strategy. This approach is described and evaluated in the remainder of this paper, which is organized as follows. Our dictionary identification procedure which comprises an extension of techniques recently proposed in the literature on *dictionary learning*, is described in Section 2, along with our proposed adaptive sensing procedure. The performance of the LASeR procedure is evaluated in Section 3, and conclusions and directions for future work are discussed in Section 4.

2. LEARNING ADAPTIVE SENSING REPRESENTATIONS

2.1. Structured Dictionary Learning

Consider a matrix $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n\times p}$, whose p columns of ambient dimension n each represent a training vector from a collection of training data. *Dictionary learning* describes a general matrix factorization problem, the goal of which is to to identify matrices $D \in \mathbb{R}^{n\times q}$ and $A \in \mathbb{R}^{q\times p}$ such that $X \approx DA$. Such factorizations are generally non-unique, so it is common to impose additional constraints on the coefficient matrix A — for example, requiring its columns $a_i \in \mathbb{R}^q$, $i = 1, 2, \ldots, p$, be sparse [11, 12]. Such conditions result in learned representations having the property that each column of the training data matrix may be expressed as a sparse linear combination of dictionary *atoms* denoted by columns of the matrix D. Overall, this type of factorization may be accomplished by obtaining a (local) solution to an optimization of the form

$$\{D, A\} = \arg \min_{D\in\mathbb{R}^{n\times q}, \{a_i\}\in\mathbb{R}^q} \sum_{i=1}^{p} \|x_i - Da_i\|_2^2 + \lambda\|a_i\|_1, \quad (2)$$

where λ is a (non-negative) regularization parameter.

Techniques for enforcing that each column of A additionally exhibit some form of pre-defined dependency structure have recently been examined in the literature [13]. Suppose that the set $\{1, 2, \ldots, q\}$ can be put into one-to-one association with q nodes in a known rooted binary tree $\mathcal{T}$. We say that a coefficient vector $a_i \in \mathbb{R}^q$ exhibits tree sparsity if the set of nonzero coefficients of a_i exist on a rooted connected subtree of $\mathcal{T}$. Efficient software packages have been developed (eg., [14]) for solving the dictionary learning problem while enforcing tree sparsity on the columns of the learned coefficient matrix. In this case, the factorization can be obtained by solving an optimization of the form

$$\{D, A\} = \arg \min_{D\in\mathbb{R}^{n\times q}, \{a_i\}\in\mathbb{R}^q} \sum_{i=1}^{p} \|x_i - Da_i\|_2^2 + \lambda\Omega(a_i). \quad (3)$$

The (convex) regularization term is given by

$$\Omega(a_i) = \sum_{g\in\mathcal{G}} \omega_g \|(a_i)_g\|, \quad (4)$$

where $\mathcal{G}$ is the set of p groups, each comprised of a node with all of its descendants in the tree $\mathcal{T}$, the notation $(a_i)_g$ refers to the subvector of a_i restricted to the indices in the set $g \in \mathcal{G}$, the ω_g are non-negative weights, and the norm can be either the ℓ_2 or ℓ_∞ norm.

2.2. Structured Sparsity and Adaptive Sensing

Suppose that we wish to efficiently sense and acquire a signal $x \in \mathbb{R}^n$. If we know a priori that x possesses a sparse representation (1) in some dictionary D having orthonormal columns, and that the coefficients a_i exhibit tree sparsity (as described above), then we may acquire x using an efficient sequential adaptive sensing procedure, as follows. Without loss of generality, let the index 1 correspond to the root of the tree. Begin by initializing a stack (or queue) with the index 1, and collect a measurement by projecting x onto the dictionary element d_1; that is, obtain the measurement

$$y_1 = d_1^T x = \sum_{i\in\mathcal{S}} a_i d_1^T d_i + d_1^T \epsilon \quad (5)$$

$$= a_1 + d_1^T \epsilon. \quad (6)$$

Notice that the last equality follows from the fact that we assumed D to have orthonormal columns, so that the sum reduces to the single coefficient a_i. Now, perform a significance test on the measured value y_1 (eg., compare its amplitude to a threshold τ). If the measurement is deemed significant, then add the locations of its immediate descendants to the stack (or queue). If the measurement is not deemed significant, then proceed by processing the stack (or queue) to obtain the index of the next coefficient to observe (ie., the index of the next vector d_i to project x onto). Notice that using a stack in the aforementioned process results in a *depth-first* traversal of the tree, while using a queue results in *breadth-first* traversal. Similar tree-based sequential sensing procedures have been proposed in the context of rapid MRI imaging using non-Fourier encoding [15] and more recently in the context of adaptive compressive imaging via so-called Direct Wavelet Sensing [16].

For **LASeR**, our goal is to learn tree-sparse approximations of each element in a collection of training data, and then to apply the aforementioned adaptive (compressive) sensing procedure to efficiently obtain signals that are similar to the training data. Here, we learn the sparsifying representation for the training data by solving an optimization of the form presented in (3), with the additional constraint that the learned dictionary have orthonormal columns (ie, we solve (3) subject to $D^T D = I$). A local solution to this optimization can be obtained by alternating minimization over the dictionary D, and the coefficient matrix $A = [a_1, a_2, \ldots a_q]$. Here, we used the SPAMS software [17] to solve the optimization over A (keeping D fixed), while the optimization over D keeping A fixed has the following closed-form expression

$$D = XA^T[AX^TXA^T]^{-1/2}. \quad (7)$$

The next section describes empirical results for LASeR.

3. EXPERIMENTAL RESULTS

We performed separate experiments on two databases. The first is comprised of 72 man-made and 91 natural images from the Psychological Image Collection at Stirling [18] (some example images are shown in the top row of Fig.1). Each image in the original database is of size 256×256, but here, we rescaled each to 128×128 to reduce computational demands on the dictionary learning procedure. The training data were then each reshaped to a 16384×1 vector and stacked together to form the training matrix $X \in \mathbb{R}^{16384\times 163}$. The

Fig. 1: Sample images from the PICS database (top row), and reconstructions obtained from different compressive sensing approaches (bottom row) for the first image in the top row. From left to right, the reconstructions depicted in the second row correspond to estimates obtained from 40 measurements of the first image in the top row, using LASeR (with thresholds $\tau = 0, 0.4, 0.8$, and 2), PCA, direct wavelet sensing ($\tau = 0$), and CS/LASSO. In each case, the measurements were subject to additive noise with $\sigma = 0.3$.

second database we used was comprised of a total of 4500 synthetically generated training vectors of length 1023, where each vector possessed a tree-sparse representation with 150 nonzeros (the support of each corresponding to a rooted connected subtree of a tree of degree two) in a randomly generated orthonormal basis. We learn balanced binary tree-structured orthonormal dictionaries with 7 levels (comprising 127 orthogonal dictionary elements) for the data from the PICS database and 10 levels (comprising 1023 orthogonal dictionary elements) for the database of synthetically generated data.

For evaluation, the LASeR sensing procedure was applied to acquire a signal from each database. We evaluated the performance of the procedure for various values of τ (the threshold for determining significance of a measured coefficient) in a noise free setting as well as when measurements are corrupted by zero-mean additive white Gaussian measurement noise. In either case, the reconstruction from the LASeR procedure is obtained as the weighted sum of the atoms of the dictionary used to obtain the projections, where the weights are taken to be the actual observation values obtained by projecting onto the corresponding atom. When assessing the performance of the procedure in noisy settings, we averaged over a total of 500 trials corresponding to different realizations of the random noise.

Reconstruction performance is quantified by the reconstruction signal to noise ratio (SNR), given by

$$\text{SNR} = 10 \log_{10} \left(\frac{\|\mathbf{x}\|_2^2}{\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2} \right). \tag{8}$$

To provide a performance comparison for LASeR, we also evaluate the reconstruction performance of the direct wavelet sensing algorithm described in [16], as well as principal component analysis (PCA) based reconstruction. For PCA, the reconstruction is obtained by taking projections of the test signal onto the principal components (obtained from the mean-centered training data) along with one additional projection onto mean of the training data, and then performing a least squares fit to get the final reconstruction. We also compare with "traditional" compressed sensing, where measurements are obtained by projecting onto random vectors (in this case, vectors whose entries are i.i.d. zero-mean Gaussian distributed) and reconstruction is obtained via the LASSO by enforcing sparsity in the learned dictionary (the regularization parameter was chosen clairvoyantly to give best performance). In order to make a fair comparison among all of the different measurement strategies, we normalize so that each measurement is obtained by projecting onto a vectors of unit norm.

Plots of reconstruction SNR values vs. number of measurements for two test signals are shown in Fig. 2. The results in the top row (for a test signal from PICS databse) show that a range of choices of threshold value τ results in good reconstruction SNR from only $40-65$ measurements using LASeR. Similar behavior is observed in the noisy case, where a good reconstruction SNR can be obtained by around 65 measurements with any threshold value. Actual estimates obtained via the various procedures for this test image are shown in the second row of Fig.1. We can see that LASeR gives a very good reconstruction with only few measurements for some (larger) choices of the threshold, whereas the other approaches perform comparably poorly. The results in the bottom row of Fig. 2 (corresponding to estimates of a test signal from the synthetic database) demonstrate further the performance gain of LASeR over the other schemes compared.

4. CONCLUSIONS

In this paper, we presented a novel sensing and reconstruction procedure called LASeR, which uses dictionaries learned from training data, in conjunction with adaptive sensing, to perform compressive sensing. Simulations demonstrate that the proposed procedure can provide significant improvements over traditional compressive sensing (based on random projection measurements), as well as other established methods such as PCA. The proposed procedure was also shown to be robust to additive noise. Future work in this direction will entail obtaining a complete characterization of the performance of the LASeR procedure for different dictionaries, and for different learned tree structures (we restricted attention here to binary trees, though higher degrees can also be obtained via the same procedure). We also note that in a related effort, we have recently obtained precise performance guarantees (for support recovery and estimation error) to quantify the performance of the top-down adaptive compressive sensing procedures (as employed in LASeR, and in the procedure proposed in [16]) in the presence of measurement noise [19].

5. REFERENCES

[1] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE*

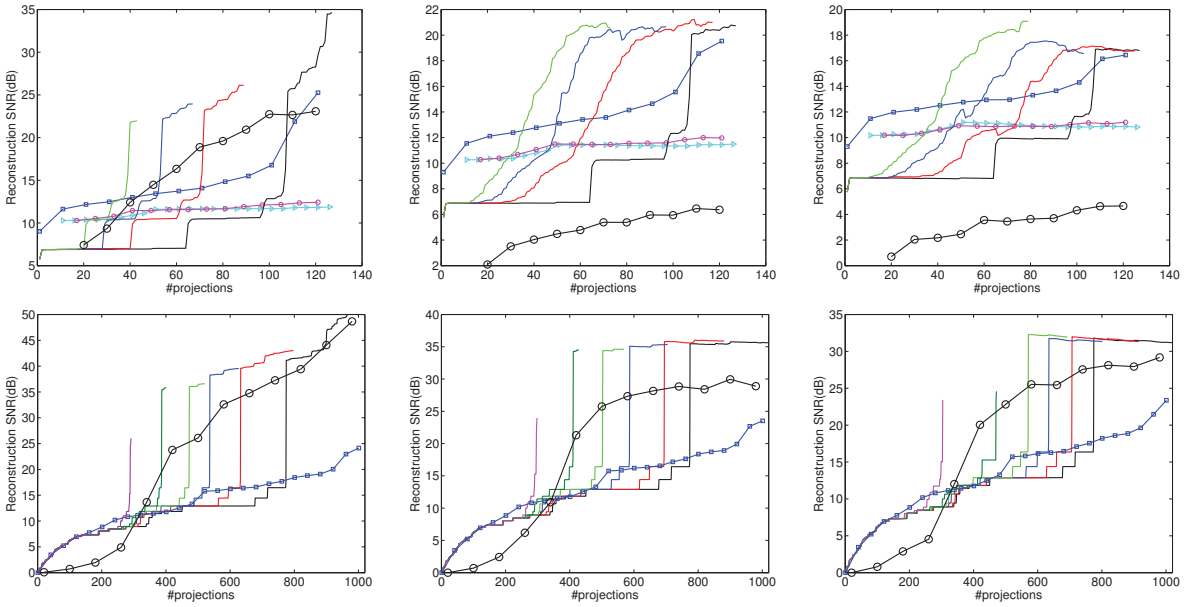

Fig. 2: SNR vs. Number of measurements plots (best viewed in color) with different noise levels and for different schemes (LASeR, PCA, direct wavelet sensing, and CS/LASSO). The top row corresponds to reconstruction SNR achieved when acquiring a test image from the PICS database (column 1: $\sigma = 0$, column 2: $\sigma = 0.3$, column 3: $\sigma = 0.8$). Here, $\square$ is PCA, black $\circ$ is LASSO, sky blue $\triangleright$ and magenta $\circ$ are for direct wavelet sensing with $\tau = 0$ and $\tau = 0.5$ respectively. Colored solid lines (left to right) are for LASeR with $\tau = 0,\ 0.3,\ 0.5$, and 2. The bottom row corresponds to reconstruction SNR achieved when acquiring a test image from the synthetically generated data (column 1: $\sigma = 0$, column 2: $\sigma = 3$, column 3: $\sigma = 5$). Here, $\square$ is PCA, $\circ$ is CS/LASSO. Colored solid lines (left to right) are for LASeR with $\tau = 0,\ 1,\ 2,\ 3,\ 5$ and 8 respectively.

Trans. Inform. Theory, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[4] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *Technical Report*, 2011, Online: `arxiv.org/pdf/1106.6224`.

[5] R. Baraniuk, V. Cevher, M. Duarte, and C Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[6] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *Technical Report*, 2009, Online: `arxiv.org/pdf/0903.3002v2`.

[7] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. on Sig. Proc.*, vol. 56, no. 6, pp. 2346–2356, 2008.

[8] R. Castro, J. Haupt, R. Nowak, and G. Raz, "Finding needles in noisy haystacks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Proc.*, 2008, pp. 5133–5136.

[9] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak, "Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, 2009, pp. 1551–1555.

[10] J. Haupt and R. Nowak, "Adaptive sensing for sparse recovery," To appear in *Compressed Sensing: Theory and applications*, Cambridge University Press, 2011, `www.ece.umn.edu/~jdhaupt/publications/cs10_adaptive_sensing.pdf`.

[11] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37, pp. 3311–3325, 1997.

[12] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Proc.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[13] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *Ann. Statist.*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[14] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for sparse hierarchical dictionary learning," in *Proc. ICML*, 2010.

[15] L. P. Panych and F. A. Jolesz, "A dynamically adaptive imaging algorithm for wavelet-encoded MRI," *Magnetic Resonance in Medicine*, vol. 32, pp. 738–748, 1994.

[16] S. Deutsch, A. Averbuch, and S. Dekel, "Adaptive compressed image sensing based on wavelet modelling and direct sampling," in *8th International Conference on Sampling, Theory and Applications. Marseille, France*, 2009.

[17] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and R. Jenatton, "Sparse modeling software," `www.di.ens.fr/willow/SPAMS/`.

[18] "Psychological image collection at Stirling," `www.pics.stir.ac.uk/`.

[19] A. Soni and J. Haupt, "Efficient adaptive compressive sensing using sparse hierarchical learned dictionaries," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, November 2011.