Greedy Dictionary Selection for Sparse Representation

Volkan Cevher
Rice University
volkan@rice.edu

Andreas Krause
Caltech
krausea@caltech.edu

Abstract

We discuss how to construct a dictionary by selecting its columns from multiple candidate bases to allow sparse representation of signals. By sparse representation, we mean that only a few dictionary elements, compared to the ambient signal dimension, can be used to well-approximate the signals. We formulate both the selection of the dictionary columns and the sparse representation of signals as a joint *combinatorial* optimization problem. The proposed combinatorial objective maximizes variance reduction over the set of signals by constraining the size of the dictionary as well as the number of dictionary columns that can be used to represent each signal. We show that if the columns of the candidate bases are incoherent, our objective function satisfies approximate submodularity. We exploit this property to develop efficient greedy algorithms with well-characterized theoretical performance. Applications of dictionary selection include denoising, inpainting, and compressive sensing. We evaluate our approach to reconstruct dictionaries from sparse samples, and also apply it to an image inpainting problem.

1 Introduction

Data compression is a well-established paradigm for alleviating storage, communication, and processing bottlenecks in information systems. Typically, a compression algorithm selects a *dictionary* $\mathcal{D} \in \mathbb{R}^{d \times n}$ where the discrete signal $y \in \mathbb{R}^d$ to be compressed can be well approximated by a sparse set of nonzero coefficients. By sparse, we mean that only k dictionary elements are nonzero where $k \ll d$. As a result, compared to the signal dimensions, only a small number of coefficients are codified and processed, reducing space, bandwidth, and computational requirements.

The utility of sparse data representation is not specific to the signal compression problem. Sparse representations with appropriate dictionaries underlie the minimax optimality of denoising via thresholding in signal processing, the stable embedding and decoding properties of codes in information theory, tractability and generalization of learning algorithms in machine learning, effectiveness of data streaming algorithms in theoretical computer science, and neuronal information processing and interactions in computational neuroscience. Unfortunately, the appropriate sparsifying dictionary is oft-times unknown and must be determined for each individual problem. Interestingly, research to date on determining data sparsifying dictionaries has paved only two distinct paths, with one following dictionary design and the other pursuing dictionary learning.

In *dictionary design*, researchers assume an abstract functional space that can concisely capture the underlying characteristics of the observations. A stylized example is Besov spaces and the set of natural images, for which the Besov norm measures spatial smoothness between edges (cf. [1] and references within). Along with the functional space, a matching dictionary is naturally introduced, e.g., wavelets ($\mathcal{W}$) for Besov spaces, to efficiently calculate the induced norm. Then, the rate distortion of the partial signal reconstructions $y_k^{\mathcal{D}}$ on the dictionary is quantified by keeping the k largest dictionary elements via an ℓ_p metric, such as $\sigma_p(y, y_k^{\mathcal{D}}) = \|y - y_k^{\mathcal{D}}\|_p^p = \left(\sum_{i=1}^d \|y_i - y_{k,i}^{\mathcal{D}}\|^p\right)^{1/p}$;

the faster $\sigma_p(y, y_k^{\mathcal{D}})$ decays with k, the better the observations can be compressed. Although the designed dictionaries have well-characterized approximation performance on signals in the assumed functional space, their empirical performance on the actual observations can greatly vary: $\sigma_2(y, y_k^{\mathcal{W}}) = O(k^{-0.2})$ (practice) vs. $O(k^{-1})$ (theory) for wavelets on natural images [2].

In *dictionary learning*, researchers develop algorithms to learn a sparsifying dictionary directly from data using techniques such as optimization, clustering, and nonparametric Bayesian inference. Optimization-based approaches define an objective function that minimize the data error, regularized by the ℓ_1 or total variation (TV) norms to enforce sparsity in the dictionary representation. The proposed objective function is then jointly solved in the dictionary entries and the sparse coefficients [3, 4, 5]. Clustering approaches learn dictionaries by sequentially determining clusters where sparse coefficients overlap on the dictionary and updating the corresponding dictionary elements based on singular value decomposition [6]. Bayesian approaches use hierarchical probability models to nonparametrically infer the dictionary size and its composition [7]. Although dictionary learning approaches have great empirical performance on many data sets of great interest, e.g., in denoising and inpainting of natural images, they lack theoretical rate distortion characterizations and have computational disadvantages compared to the dictionary design approaches.

In this paper, we investigate a mixture of dictionary design and learning approaches to determine sparsifying dictionaries. We focus on *dictionary selection*, where we build a dictionary by selecting its columns from multiple candidate bases, typically designed for the observations of interest. We constrain the size of the dictionary as well as the number of dictionary columns that can be used to represent each signal with user defined parameters n and k, respectively. We formulate both the selection of basis functions and the sparse reconstruction as a joint *combinatorial* optimization problem. Our objective function maximizes an approximation error reduction metric (defined in Section 2) over the set of observations.

We show that the dictionary selection problem is approximately submodular where the approximation depends on the coherence of the candidate bases. We then propose two computationally efficient algorithms for dictionary selection and characterize their theoretical performances in Section 4. Compared to the continuous nature of the dictionary learning approaches, our approach is discrete and provides new theoretical characterizations. Compared to the dictionary design approaches, our approach is data adaptive and has better empirical performance on data sets. Moreover, it can automatically exploit the existing dictionary designs for the observations. In Section 5, we evaluate approach to recover a correct dictionary via synthetic examples. We also apply our algorithms to an image inpainting task, where missing image pixels are determined based on the dictionary, learned in-situ from the observed pixels.

2 The dictionary selection problem

In the dictionary selection problem, we are interested in determining a dictionary $\mathcal{D}$ to sparsely represent a given collection of signals, $\mathcal{Y} = y_1, \ldots, y_m \in \mathbb{R}^d$. We compose the dictionary using a variance reduction metric defined below by selecting a subset out of a set of candidate vectors $\mathcal{V} = \{\phi_1, \ldots, \phi_N\}$, where $\phi_i \in \mathbb{R}^d$. Without loss of generality, we assume $\|y_i\|_2 \leq 1$ and $\|\phi_i\|_2 = 1$ for all i. In the sequel, we define $\Phi_{\mathcal{A}} = [\phi_{i_1}, \ldots, \phi_{i_k}]$ as a matrix containing the vectors in $\mathcal{V}$ as indexed by $\mathcal{A} = \{i_1, \ldots, i_k\}$ where $\mathcal{A} \subseteq \mathcal{V}$. Moreover, we do not assume any particular ordering of $\mathcal{V}$.

For a fixed signal y_s and a set of vectors $\mathcal{A}$, we define the *reconstruction* accuracy as

$$L_s(\mathcal{A}) = \sigma_2(y_s, y^{\mathcal{A}}) = \min_w ||y_s - \Phi_{\mathcal{A}} w||_2^2.$$

The problem of optimal k-sparse reconstruction with respect to a fixed dictionary $\mathcal{D}$ then requires solving the following discrete optimization problem:

$$\mathcal{A}_s^* = \min_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} L_s(\mathcal{A}), \tag{1}$$

where $|\cdot|$ counts the set cardinality and k is the sparsity constraint on the number of atoms used in the reconstruction.

Formally, we now seek a dictionary $\mathcal{D} \subseteq \mathcal{V}$ that allows to obtain the best reconstruction accuracy (cf. (1)) possible not just for a single signal, but for *all* signals $y_1, \ldots, y_m$. Note that each signal y_s

can potentially use a different set of atoms $\mathcal{A}_s \subseteq \mathcal{D}$. Thus, we define, for each signal y_s a separate utility function

$$F_s(\mathcal{D}) = L_s(\emptyset) - \min_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} L_s(\mathcal{A}), \tag{2}$$

i.e., $F_s(\mathcal{D})$ measures the improvement in reconstruction accuracy (also known as *variance reduction*) for signal y_s in case dictionary $\mathcal{D}$ gets selected.

Based on the notation above, we define the average improvement for all signals as

$$F(\mathcal{D}) = \frac{1}{m} \sum_s F_s(\mathcal{D}),$$

and define the *dictionary selection problem* as seeking the solution to

$$\mathcal{D}^* = \operatorname*{argmax}_{|\mathcal{D}| \leq n} F(\mathcal{D}), \tag{3}$$

where n is a constraint on the number of atoms that the dictionary can be composed of.

The optimization problem in (3) presents combinatorial challenges. In fact, even evaluating $F_s(\mathcal{D})$ requires finding the set $\mathcal{A}_s$ of k basis functions–out of exponentially many options–for best reconstruction of y_s. Furthermore, even if we could evaluate F_s, we would have to search over an exponential number of possible dictionaries $\mathcal{D}$.

3 Submodularity in sparse representation

Structure in the Dictionary Selection problem. The key insight that allows us to address problem (3) is that the objective function F exhibits the following properties. By definition, we have $F(\emptyset) = 0$. Whenever $\mathcal{D} \subseteq \mathcal{D}'$ then $F(\mathcal{D}) \leq F(\mathcal{D}')$, i.e., F increases monotonically with $\mathcal{D}$. Moreover, F is *approximately submodular*, i.e., there exists an ε that depends on properties of $\Phi_{\mathcal{V}}$, such that whenever $\mathcal{D} \subseteq \mathcal{D}' \subseteq \mathcal{V}$ and $v \in \mathcal{V} \setminus \mathcal{D}'$ it holds that

$$F(\mathcal{D} \cup \{v\}) - F(\mathcal{D}) \geq F(\mathcal{D}' \cup \{v\}) - F(\mathcal{D}') - \varepsilon.$$

The last property implies that adding a new atom v to a larger dictionary $\mathcal{D}'$ helps at most ε more than adding v to a subset $\mathcal{D} \subseteq \mathcal{D}'$. A fundamental result by Nemhauser et al. [8] proves that for monotonic submodular functions G, a simple greedy algorithm that starts with the empty set $\mathcal{D}_0 = \emptyset$, and at every iteration i adds the element

$$v_i = \operatorname*{argmax}_{v \in \mathcal{V} \setminus \mathcal{D}} G(\mathcal{D}_{i-1} \cup \{v\}),$$

where $\mathcal{D}_i = \{v_1, \ldots, v_i\}$, obtains a near-optimal solution: For the solution $\mathcal{D}_n$ returned by the greedy algorithm it holds that

$$G(\mathcal{D}_n) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} G(\mathcal{D}),$$

i.e., obtains at least of constant fraction of $(1 - 1/e) \approx 63\%$ of the optimal value. We prove in the sequel that the approximate submodularity of F is sufficient to obtain a near-optimal solution to (3), using a similar greedy algorithm.

Approximate submodularity of F. To prove approximate submodularity, we assume that the matrix $\Phi_{\mathcal{V}}$ of basis functions (each normalized to unit norm) is *incoherent* with parameter μ: We assume that for each $\phi_i, \phi_j \in \mathcal{V}$, $i \neq j$ it holds that

$$|\langle \phi_i, \phi_j \rangle| \leq \mu.$$

There has been a significant body of work establishing the existence and construction of dictionaries with low coherence μ. For example, it is possible to achieve incoherence $\mu = d^{-1/2}$ with the union of $d/2$ orthonormal bases (cf. Theorem 2 in [9]).

For a given signal $y_s \in \mathbb{R}^d$ and a candidate vector $\phi_v \in \Phi_{\mathcal{V}}$, we define $w_{s,v} = \langle \phi_v, y_s \rangle^2$. Furthermore, we introduce the following set functions, which will be useful in the analysis below:

$$\widehat{F}_s(\mathcal{D}) = \max_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} \sum_{v \in \mathcal{A}} w_{s,v}, \text{ and } \widehat{F}(\mathcal{D}) = \frac{1}{m} \sum_s \widehat{F}_s(\mathcal{D}).$$

Theorem 1 *Suppose $\Phi_{\mathcal{V}}$ is incoherent with constant μ and let $y_s \in \mathbb{R}^d$. Then, for any $\mathcal{D} \subseteq \mathcal{V}$, it holds that*

$$|\widehat{F}(\mathcal{D}) - F(\mathcal{D})| \leq k\mu.$$

Furthermore, the function $\widehat{F}(\mathcal{D})$ is monotonic and submodular.

Proof [Sketch] The proof of the theorem is geometric; if $\Phi_{\mathcal{V}}$ is in fact an orthonormal basis, the improvement in reconstruction accuracy $G_s(\mathcal{A}) = L_s(\emptyset) - L_s(\mathcal{A}) = \sum_{v \in \mathcal{A}} w_{s,v}$ is additive (modular) due to the Pythagorean theorem. Thus, $\widehat{F}_s(\mathcal{D}) = F_s(\mathcal{D}) = \max_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} \sum_{v \in \mathcal{A}} w_{s,v}$.

Using a geometric argument, for incoherent dictionaries, G_s can be shown to be *approximately* additive (modular):

$$|G_s(\mathcal{A} \cup \{v\}) - G_s(\mathcal{A}) - w_{s,v}| \leq \mu,$$

and thus $|G_s(\mathcal{A}) - \sum_{v \in \mathcal{A}} w_{s,v}| \leq k\mu$. Note that $F(\mathcal{D}) = \max_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} G(\mathcal{A})$. Thus $|\widehat{F}_s(\mathcal{D}) - F_s(\mathcal{D})| \leq k\mu$ and thus $|\widehat{F}_s(\mathcal{D}) - F_s(\mathcal{D})| \leq k\mu$.

$\widehat{F}_s$ is observed to be a monotonic submodular function. Since the average of monotonic submodular functions remains monotonic submodular, the function $\widehat{F}(\mathcal{D})$ is submodular as well. ■

4 The Submodular Dictionary Selection (SDS) algorithm

Thus far, we have not yet addressed the question on how to evaluate the function $F(\mathcal{D})$ efficiently. Evaluating $F_s(\mathcal{D})$ requires maximizing the variance reduction over exponentially many subsets $\mathcal{A} \subseteq \mathcal{D}$ of size k. However, the result of Nemhauser et al. suggests that greedy strategy for dictionary selection might also be useful due to the approximate submodularity of F. As it turns out, the natural greedy algorithm for the sparse reconstruction problem 1, Orthogonal Matching Pursuit (OMP) [10], can be used to obtain a near-exact evaluation of the function F:

Proposition 1 *Using Orthogonal Matching Pursuit to evaluate* (2) *produces a function F_{OMP} such that $|F_{OMP}(\mathcal{D}) - F(\mathcal{D})| \leq k\mu$ over all dictionaries $\mathcal{D}$.*

The proof of the proposition follows from the approximate modularity of $L_s(\emptyset) - L_s(\mathcal{A})$. Taken together, Theorem 1 and Proposition 1 prove that the objective function $F_{OMP}(\mathcal{D})$ is approximately submodular: For any sets $\mathcal{D} \subseteq \mathcal{D}' \subseteq \mathcal{V}$ and $v \in \mathcal{V} \setminus \mathcal{D}'$, it holds that

$$F_{OMP}(\mathcal{D} \cup \{v\}) - F_{OMP}(\mathcal{D}) \geq F_{OMP}(\mathcal{D}' \cup \{v\}) - F_{OMP}(\mathcal{D}') - k\mu.$$

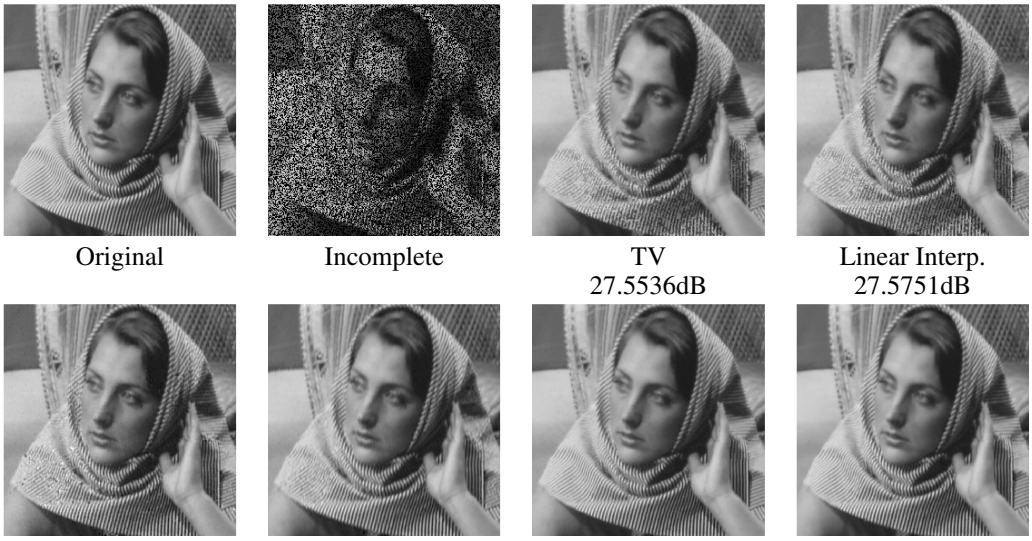
While our objective F is only approximately submodular, the result of Nemhauser et al. [8] can be extended to prove that in this setting for the greedy solution it holds that $F_{OMP}(\mathcal{D}_n) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F_{OMP}(\mathcal{D}) - nk\mu$ [11]. We call the greedy algorithm applied to the function F_{OMP} the SDS_{OMP} algorithm (for submodular dictionary selection using OMP), which inherits the following guarantee:

Theorem 2 SDS_{OMP} *selects a dictionary $\mathcal{D}'$ such that*

$$F(\mathcal{D}') \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F(\mathcal{D}) - 2kn\mu.$$

The solution returned by the SDS_{OMP} algorithm is guaranteed to obtain at least a constant fraction of the optimal variance reduction, minus an absolute error $2kn\mu$. For the setting of high-dimensional signals and small ($n \ll \sqrt{d}$) incoherent dictionaries ($\mu = O(1/\sqrt{d})$) this error is negligible.

Improved guarantees for large dictionaries. If we are interested in selecting a large dictionary $\mathcal{D}$ of size $n = \Omega(\sqrt{d})$, then the guarantee of Theorem 2 becomes vacuous since $F(\mathcal{D}) \leq 1$. Fortunately, instead of running the greedy algorithm on the objective $F_{OMP}(\mathcal{D})$, we can run the greedy algorithm on the approximation $\widehat{F}(\mathcal{D})$ instead. We call the greedy algorithm applied to objective $\widehat{F}$ the SDS_{MA} algorithm (for submodular dictionary selection using modular approximation). We then have the following guarantee:



Original | Incomplete | TV 27.5536dB | Linear Interp. 27.5751dB

Greedy (MA) 30.3195dB | Greedy (OMP) 31.9071dB | Nonlocal TV 32.8615dB | Bayesian 35.1789dB

Figure 1: Comparison of inpainting algorithms.

Theorem 3 *The* SDS_{MA} *algorithm produces a dictionary* $\mathcal{D}'$ *such that*

$$F(\mathcal{D}') \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F(\mathcal{D}) - 2k\mu.$$

In most realistic settings with high-dimensional signals and incoherent dictionaries, this error is negligible. While algorithm SDS_{MA} has better guarantees and is much faster than SDS_{OMP}, as we show in the next section, SDS_{OMP} empirically performs better.

5 Experiments

Finding a basis in a haystack. To understand how the theoretical performance characterizations reflect on the actual performance of the algorithms, we perform experiments on synthetic data. We generate a collection $\mathcal{V}$ of atoms by forming the union of eight orthonormal bases, including the discrete cosine transform (DCT) and different $d = 64$ dimensional wavelet bases (Haar, Daub4 and 8, Coiflets 1, 3 and 5, and Discrete Meyr). We then repeatedly pick a dictionary $\mathcal{D}^* \subseteq \mathcal{V}$ of size $n = 20$ at random, and generated a collection of $m = 100$ random 5-sparse signals according to the dictionary $\mathcal{D}^*$. Our goal is to recover the true dictionary $\mathcal{D}^*$ using our SDS algorithm. For each random trial, we run SDS_{OMP} to select a dictionary $\mathcal{D}$ of size 20. We then look at the overlap $|\mathcal{D} \cap \mathcal{D}^*|$ to measure the performance. Note that the collection $\mathcal{V}$ of atoms given by the union of 8 bases is not incoherent. Nevertheless, on average, $\mathcal{D}$ and $\mathcal{D}^*$ overlap by approximately 75%, i.e., SDS correctly identifies a majority of the randomly chosen atoms in $\mathcal{D}^*$. Furthermore, if we generate the sparse signals by using atoms from only one out of the 8 bases, SDS obtains near-perfect (more than 95% overlap) reconstruction. These results suggest that SDS is able to correctly identify a dictionary "hidden" in a large number of atoms, even when our incoherence assumption is violated.

Dictionary selection from dimensionality reduced data. Natural images provide a scaffold for framing numerous dictionary learning problems encountered in different data modalities. In this subsection, we focus on a specific image processing problem, inpainting, to motivate a dictionary selection problem from dimensionality reduced data. Suppose that instead of observing $\mathcal{Y}$ as assumed in Section 2, we observe $\mathcal{Y}' = \mathcal{P}_1 y_1, \ldots, \mathcal{P}_m y_m \in \mathbb{R}^b$, where $\mathcal{P}_i \in \mathbb{R}^{b \times d} \forall i$ are known linear projection matrices. In the inpainting setting, $\mathcal{P}_i$'s are binary matrices which pass or delete pixels. From a theoretical perspective, dictionary selection from dimensionality reduced data is ill-posed. For the purposes of this demonstration, we will assume that $\mathcal{P}_i$'s are information preserving.

As opposed to observing a series of signal vectors, we start with a single image in Fig. 1, albeit missing 50% of its pixels. We break the noisy image into non-overlapping 8 by 8 patches, and train a dictionary for sparse reconstruction of those patches to minimize the average approximation error

on the observed pixels. As candidate bases, we use DCT, wavelets (Haar and Daub4), Coiflets (1 and 3), and Gabor. We test our SDS_{OMP} and SDS_{MA} algorithms, approaches based on total-variation (TV), linear interpolation, nonlocal TV and the nonparametric Bayesian dictionary learning (based on Indian buffet processes) algorithms [4, 5, 7]. The TV and nonlocal TV algorithms use the linear interpolation result as their initial estimates. Figure 1 illustrates the inpainting results for each algorithm sorted in terms of increasing peak signal to noise ratio (PSNR). We do not report the reconstruction results using individual candidate bases here since they are significantly worse than the baseline linear interpolation.

The test image exhibits significant self similarities, decreasing the actual union of subspaces of the sparse coefficients. Hence, for our modular and OMP-based greedy algorithms, we ask the algorithms to select 64×32 dimensional dictionaries. While the modular algorithm SDS_{MA} selects the desired dimensions, the OMP-based greedy algorithm SDS_{OMP} terminates when the dictionary dimensions reach 64×19. Given the selected dictionaries, we determine the sparse coefficients that best explain the observed pixels in a given patch and reconstruct the full patch using the same coefficients. We repeat this process for all the patches in the image that differ by a single pixel. In our final reconstruction, we take the pixel median of all the reconstructed patches. Our OMP algorithm perform on par with nonlocal TV while taking a fraction of its computational time. While the Bayesian approach takes significantly more time (a few order of magnitudes slower), it best exploits the self similarities in the observed image to result in the best reconstruction.

6 Conclusions

We study the problem of selecting a dictionary out of a large number of atoms for the purpose of sparsely representing a collection of signals. We prove that if the original collection of atoms is sufficiently incoherent, the global objective in the dictionary selection problem is approximately submodular. We present two algorithms, SDS_{OMP} (based on Orthogonal Matching Pursuit) and SDS_{MA} (using a simple modular approximation). By exploiting the approximate submodularity property of our objective, we provide theoretical approximation guarantees for the performance of our algorithms. Our empirical results demonstrate the ability of our algorithm to correctly reconstruct a sparsifying dictionary, and to provide fresh insights for image reconstruction. We believe that our results are a promising direction for studying possible connections between sparse reconstruction (ℓ_1-minimization etc.) and combinatorial optimization, in particular submodularity.

Acknowledgments. This work was supported in part by a gift from Microsoft Corporation, by NSF CNS 0932392, by ONR grants N00014-09-1-1044 and N00014-08-1-1112, by AFOSR FA9550-07-1-0301, ARO W911NF-09-1-0383 and DARPA N66001-08-1-2065.

References

[1] H. Choi and R. G. Baraniuk. Wavelet statistical models and Besov spaces. *Lecture Notes in Statistics*, pages 9–30, 2003.

[2] V. Cevher. Learning with compressible priors. In *NIPS*, Vancouver, B.C., Canada, 7–12 December 2008.

[3] B. A. Olshausen and Field D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[4] X. Zhang and T. F. Chan. Wavelet Inpainting by Nonlocal Total Variation. CAM Report (09-64), July 2009.

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

[6] M. Aharon, M. Elad, and A. Bruckstein. The k-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, 54(11):4311–4322, 2006.

[7] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric bayesian dictionary learning for sparse image representations. In *Neural Information Processing Systems (NIPS)*, 2009.

[8] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[9] R. Gribonval and M. Nielsen. Sparse decompositions in "incoherent" dictionaries. In *ICIP*, 2002.

[10] A. C. Gilbert and J. A. Tropp. Signal recovery from random measurements via orthogonal matching pursuit. Technical report, University of Michigan, 2005.

[11] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. In *Journal of Machine Learning Research*, volume 9, 2008.