

---

# Submodular Dictionary Selection for Sparse Representation

---

Andreas Krause

KRAUSEA@CALTECH.EDU

California Institute of Technology, Computer Science Department

Volkan Cevher

VOLKAN.CEVHER@{EPFL,IDIAP}.CH

Ecole Polytechnique Federale de Lausanne, STI-IEL-LIONS & Idiap Research Institute

## Abstract

We develop an efficient learning framework to construct signal dictionaries for sparse representation by selecting the dictionary columns from multiple candidate bases. By sparse, we mean that only a few dictionary elements, compared to the ambient signal dimension, can exactly represent or well-approximate the signals of interest. 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 training 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 available dictionary column vectors are incoherent, our objective function satisfies approximate *submodularity*. We exploit this property to develop  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$ , two greedy algorithms with approximation guarantees. We also describe how our learning framework enables dictionary selection for structured sparse representations, e.g., where the sparse coefficients occur in restricted patterns. We evaluate our approach on synthetic signals and natural images for representation and inpainting problems.

## 1. Introduction

An important problem in machine learning, signal processing and computational neuroscience is to determine a *dictionary* of basis functions for sparse representation of signals. A signal  $y \in \mathbb{R}^d$  has a sparse representation with  $y = \mathcal{D}\alpha$  in a dictionary  $\mathcal{D} \in \mathbb{R}^{d \times n}$ , when  $k \ll d$  coefficients of  $\alpha$  can exactly represent or well-approximate  $y$ . Myriad applications in data analysis and processing—from deconvolution to data mining and from compression to compressive sensing—involve such representations. Surprisingly, there are only two main approaches for determining data-sparsifying dictionaries: dictionary design and dictionary learning.

In *dictionary design*, researchers assume an abstract functional space that can concisely capture the underlying characteristics of the signals. A classical example is based on Besov spaces and the set of natural images, for which the Besov norm measures spatial smoothness between edges (c.f., Choi & Baraniuk (2003) and the references therein). 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  $y_k^{\mathcal{D}}$  is quantified by keeping the  $k$  largest dictionary elements via an  $\ell_p$  norm, such as  $\sigma_p(y, y_k^{\mathcal{D}}) = \|y - y_k^{\mathcal{D}}\|_p \equiv \left(\sum_{i=1}^d \|y_i - y_{k,i}^{\mathcal{D}}\|^p\right)^{1/p}$ ; the faster  $\sigma_p(y, y_k^{\mathcal{D}})$  decays with  $k$ , the better the observations can be compressed. While the designed dictionaries have well-characterized rate distortion and approximation performance on signals in the assumed functional space, they are data-independent and hence their empirical performance on the actual observations can greatly vary:  $\sigma_2(y, y_k^{\mathcal{W}}) = \mathcal{O}(k^{-0.1})$  (practice) vs.  $\mathcal{O}(k^{-0.5})$  (theory) for wavelets on natural images (Cevher, 2008).

In *dictionary learning*, researchers develop algorithms to learn a dictionary for sparse representation directly from data using techniques such as regularization, clustering, and nonparametric Bayesian inference. Regularization-based approaches define an objective function that minimize the data error, regularized by the  $\ell_1$  or the total variation (TV) norms to enforce sparsity under the dictionary representation. The proposed objective function is then jointly optimized in the dictionary entries and the sparse coefficients (Olshausen & Field, 1996; Zhang & Chan, 2009; Mairal et al., 2008). Clustering approaches learn dictionaries by sequentially determining clusters where sparse coefficients overlap on the dictionary and then updating the corresponding dictionary elements based on singular value decomposition (Aharon et al., 2006). Bayesian approaches use hierarchical probability models to nonparametrically infer the dictionary size and

---

Appearing in *Proceedings of the 27<sup>th</sup> International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

its composition (Zhou et al., 2009). Although dictionary learning approaches have great empirical performance on many data sets in denoising and inpainting of natural images, they lack theoretical rate distortion characterizations of the dictionary design approaches.

In this paper, we investigate a hybrid approach between dictionary design and learning. We propose a learning framework based on *dictionary selection*: We build a sparsifying dictionary for a set of observations by selecting the dictionary 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 a variance reduction metric over the set of observations.

We then propose  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$ , two computationally efficient, greedy algorithms for dictionary selection. We show that under certain incoherence assumptions on the candidate vectors, the dictionary selection problem is approximately submodular, and we use this insight to derive theoretical performance guarantees for our algorithms. We also demonstrate that our framework naturally extends to dictionary selection with restrictions on the allowed sparsity patterns in signal representation. As a stylized example, we study a dictionary selection problem where the sparse signal coefficients exhibit *block sparsity*, e.g., sparse coefficients appear in pre-specified blocks.

Lastly, we first evaluate the performance of our algorithms in both on synthetic and real data. Our main contributions can be summarized as follows:

1. We introduce the problem of dictionary selection and cast the dictionary learning/design problems in a new, discrete optimization framework.
2. We propose new algorithms and provide their theoretical performance characterizations by exploiting a geometric connection between submodularity and sparsity.
3. We extend our dictionary selection framework to allow structured sparse representations.
4. We evaluate our approach on several real-world sparse representation and image inpainting problems and show that it provides practical insights to existing image coding standards.

## 2. The dictionary selection problem

In the *dictionary selection problem* (DiSP), we seek a dictionary  $\mathcal{D}$  to sparsely represent a given collection

of signals  $\mathcal{Y} = \{y_1, \dots, y_m\} \in \mathbb{R}^{d \times m}$ . We compose  $\mathcal{D}$  using the variance reduction metric, defined below, by selecting a subset of a candidate vector set  $\mathcal{V} = \{\phi_1, \dots, \phi_N\} \in \mathbb{R}^{d \times N}$ . Without loss of generality, we assume  $\|y_i\|_2 \leq 1$  and  $\|\phi_i\|_2 = 1, \forall i$ . In the sequel, we define  $\Phi_{\mathcal{A}} = [\phi_{i_1}, \dots, \phi_{i_Q}]$  as a matrix containing the vectors in  $\mathcal{V}$  as indexed by  $\mathcal{A} = \{i_1, \dots, i_Q\}$  where  $\mathcal{A} \subseteq \mathcal{V}$  and  $Q = |\mathcal{A}|$  is the cardinality of the set  $\mathcal{A}$ . We do not assume any particular ordering of  $\mathcal{V}$ .

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

$$L_s(\mathcal{A}) = \sigma_2^2(y_s, y^{\mathcal{A}}) = \min_w \|y_s - \Phi_{\mathcal{A}} w\|_2^2. \quad (1)$$

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

$$\mathcal{A}_s = \underset{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k}{\operatorname{argmin}} L_s(\mathcal{A}), \quad (2)$$

where  $k$  is the user-defined sparsity constraint on the number of columns in the reconstruction.

In DiSP, we are interested in determining a dictionary  $\mathcal{D} \subseteq \mathcal{V}$  that obtains the best possible reconstruction accuracy for not only a single signal but *all signals*  $\mathcal{Y}$ . Each signal  $y_s$  can potentially use different columns  $\mathcal{A}_s \subseteq \mathcal{D}$  for representation; we thus define

$$F_s(\mathcal{D}) = L_s(\emptyset) - \min_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} L_s(\mathcal{A}), \quad (3)$$

where  $F_s(\mathcal{D})$  measures the improvement in reconstruction accuracy, also known as *variance reduction*, for the signal  $y_s$  and the dictionary  $\mathcal{D}$ . Moreover, we define the average improvement for all signals as

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

The optimal solution to the DiSP is then given by

$$\mathcal{D}^* = \underset{|\mathcal{D}| \leq n}{\operatorname{argmax}} F(\mathcal{D}), \quad (4)$$

where  $n$  is a user-defined constraint on the number of dictionary columns. For instance, if we are interested in selecting a basis, we have  $n = d$ .

**DiSP challenges:** The optimization problem in (4) presents two combinatorial challenges. **(C1)** Evaluating  $F_s(\mathcal{D})$  requires finding the set  $\mathcal{A}_s$  of  $k$  basis functions—out of exponentially many options—for the best reconstruction accuracy of  $y_s$ . **(C2)** Even if we could evaluate  $F_s$ , we would have to search over an exponential number of possible dictionaries to determine  $\mathcal{D}^*$  for all signals. Even the special case of  $k = n$  is NP-hard (Davis et al., 1997). To circumvent these

combinatorial challenges, the existing dictionary learning work relies on continuous relaxations, such as replacing the combinatorial sparsity constraint with the  $\ell_1$ -norm of the dictionary representation of the signal. However, these approaches result in non-convex objectives, and the performance of such relaxations is typically not well-characterized for dictionary learning.

### 3. Submodularity in sparse representation

In this section, we first describe a key structure in the DiSP objective function: *approximate submodularity*. We then relate this structure to a geometric property of the candidate vector set, called *incoherence*. We use these two concepts to develop efficient algorithms with provable guarantees in the next section.

**Approximate submodularity in DiSP:** To define this concept, we first note that  $F(\emptyset) = 0$  and whenever  $\mathcal{D} \subseteq \mathcal{D}'$  then  $F(\mathcal{D}) \leq F(\mathcal{D}')$ , i.e.,  $F$  increases monotonically with  $\mathcal{D}$ . In the sequel, we will show that  $F$  is approximately submodular: A set function  $F$  is called *approximately submodular* with constant  $\varepsilon$ , if for  $\mathcal{D} \subseteq \mathcal{D}' \subseteq \mathcal{V}$  and  $v \in \mathcal{V} \setminus \mathcal{D}'$  it holds that

$$F(\mathcal{D} \cup \{v\}) - F(\mathcal{D}) \geq F(\mathcal{D}' \cup \{v\}) - F(\mathcal{D}') - \varepsilon. \quad (5)$$

In the context of DiSP, the above definition implies that adding a new column  $v$  to a larger dictionary  $\mathcal{D}'$  helps at most  $\varepsilon$  more than adding  $v$  to a subset  $\mathcal{D} \subseteq \mathcal{D}'$ . When  $\varepsilon = 0$ , the set function is called *submodular*.

A fundamental result by Nemhauser et al. (1978) proves that for monotonic submodular functions  $G$  with  $G(\emptyset) = 0$ , a simple greedy algorithm that starts with the empty set  $\mathcal{D}_0 = \emptyset$ , and at every iteration  $i$  adds a new element via

$$v_i = \operatorname{argmax}_{v \in \mathcal{V} \setminus \mathcal{D}} G(\mathcal{D}_{i-1} \cup \{v\}), \quad (6)$$

where  $\mathcal{D}_i = \{v_1, \dots, v_i\}$ , obtains a near-optimal solution. That is, for the solution  $\mathcal{D}_n$  returned by the greedy algorithm, we have the following guarantee:

$$G(\mathcal{D}_n) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} G(\mathcal{D}). \quad (7)$$

The solution  $\mathcal{D}_n$  hence obtains at least a constant fraction of  $(1 - 1/e) \approx 63\%$  of the optimal value.

Using similar arguments, Krause et al. (2008) show that the same greedy algorithm, when applied to approximately submodular functions, instead inherits the following—slightly weaker—guarantee

$$F(\mathcal{D}_n) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F(\mathcal{D}) - n\varepsilon. \quad (8)$$

In Section 4, we explain how this greedy algorithm can be adapted to DiSP. But first, we elaborate on how  $\varepsilon$  depends on the candidate vector set  $\Phi_{\mathcal{V}}$ .

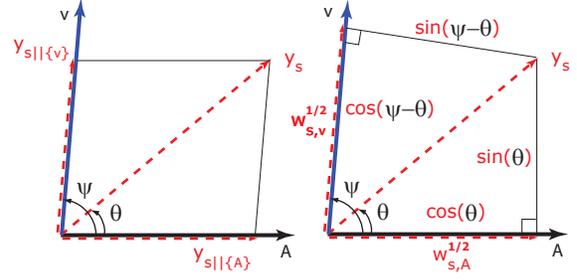


Figure 1. Example geometry in DiSP. (Left) Minimum error decomposition. (Right) Modular decomposition.

**Geometry in DiSP (incoherence):** The approximate submodularity of  $F$  explicitly depends on the maximum *incoherency*  $\mu$  of  $\Phi_{\mathcal{V}} = [\phi_1, \dots, \phi_N]$ :

$$\mu = \max_{\forall (i,j), i \neq j} |\langle \phi_i, \phi_j \rangle| = \max_{\forall (i,j), i \neq j} |\cos \psi_{i,j}|,$$

where  $\psi_{i,j}$  is the angle between the vectors  $\phi_i$  and  $\phi_j$ .

The following lemma establishes a key relationship between  $\varepsilon$  and  $\mu$  for DiSP.

**Theorem 1** *If  $\Phi_{\mathcal{V}}$  has incoherence  $\mu$ , then the variance reduction objective  $F$  in DiSP is  $\varepsilon$ -approximately submodular with  $\varepsilon \leq 4k\mu$ .*

**Proof** Let  $w_{s,v} = \langle \phi_v, y_s \rangle^2$ . When  $\Phi_{\mathcal{V}}$  is an orthonormal basis, the reconstruction accuracy in (1) can be written as follows

$$L_s(\mathcal{A}) = \left\| y_s - \sum_{q=1}^Q \phi_{i_q} \langle y_s, \phi_{i_q} \rangle \right\|_2^2 = \|y_s\|_2^2 - \sum_{v \in \mathcal{A}} w_{s,v}.$$

Hence the function  $R_s(\mathcal{A}) \equiv L_s(\emptyset) - L_s(\mathcal{A}) = \sum_{v \in \mathcal{A}} w_{s,v}$  is additive (modular). It can be seen that then  $F_s(\mathcal{D}) = \max_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} R_s(\mathcal{A})$  is submodular.

Now suppose  $\Phi_{\mathcal{V}}$  is incoherent with constant  $\mu$ . Let  $\mathcal{A} \subseteq \mathcal{V}$  and  $v \in \mathcal{V} \setminus \mathcal{A}$ . Then we claim that  $|R_s(\mathcal{A} \cup \{v\}) - R_s(\mathcal{A}) - w_{s,v}| \leq \mu$ . Consider the special case where  $y_s$  is in the span of two subspaces  $\mathcal{A}$  and  $v$ , and w.l.o.g.,  $\|y_s\|_2^2 = 1$ ; refer to Fig. 1 for an illustration. The reconstruction accuracy as defined in (1) has a well-known closed form solution:  $L_s(\mathcal{A}) = \min_w \|y_s - \Phi_{\mathcal{A}} w\|_2^2 = \|y_s - \Phi_{\mathcal{A}} \Phi_{\mathcal{A}}^\dagger y_s\|_2^2$ , where  $\dagger$  denotes the pseudoinverse; the matrix product  $P = \Phi_{\mathcal{A}} \Phi_{\mathcal{A}}^\dagger$  is simply the projection of the signal  $y_s$  onto the subspace of  $\mathcal{A}$ . We therefore have  $R_s(\mathcal{A}) = 1 - \sin^2(\theta)$ ,  $R_s(\mathcal{A} \cup \{v\}) = 1$ , and  $R_s(\{v\}) = 1 - \sin^2(\psi - \theta)$ , where  $\theta$  and  $\psi$  are defined in Fig. 1. We thus can bound  $\varepsilon_s \equiv |R_s(\mathcal{A} \cup \{v\}) - R_s(\mathcal{A}) - w_{s,v}|$  by

$$\begin{aligned} \varepsilon_s &\leq \max_{\theta} |\sin^2(\psi - \theta) + \sin^2(\theta) - 1| \\ &= |\cos \psi| \max_{\theta} |\cos(\psi - 2\theta)| = \mu. \end{aligned}$$

If  $y_s$  is not in the span of  $\mathcal{A} \cup \{v\}$ , we apply above reasoning to the projection of  $y_s$  onto their span.

Define  $\widehat{R}_s(\mathcal{A}) = \sum_{v \in \mathcal{A}} w_{s,v}$ . Then, by induction, we have  $|\widehat{R}_s(\mathcal{A}) - R_s(\mathcal{A})| \leq k\mu$ . Note that the function  $\widehat{F}_s(\mathcal{D}) = \max_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} \widehat{R}_s(\mathcal{A})$  is submodular. Let  $\mathcal{A}_s = \operatorname{argmax}_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} R_s(\mathcal{A})$  and  $\widehat{\mathcal{A}}_s = \operatorname{argmax}_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} \widehat{R}_s(\mathcal{A})$ . Therefore, it holds that

$$F_s(\mathcal{D}) = R_s(\mathcal{A}_s) \leq \widehat{R}(\mathcal{A}_s) + k\mu \leq \widehat{R}(\widehat{\mathcal{A}}_s) + k\mu = \widehat{F}_s(\mathcal{D}) + k\mu.$$

Similarly,  $\widehat{F}_s(\mathcal{D}) \leq F_s(\mathcal{D}) + k\mu$ . Thus,  $|\widehat{F}_s(\mathcal{D}) - F_s(\mathcal{D})| \leq k\mu$ , and hence  $|\widehat{F}(\mathcal{D}) - F(\mathcal{D})| \leq k\mu$  holds for all candidate dictionaries  $\mathcal{D}$ . Therefore, whenever  $\mathcal{D} \subseteq \mathcal{D}'$  and  $v \notin \mathcal{D}'$ , we can obtain the following

$$\begin{aligned} & F(\mathcal{D} \cup \{v\}) - F(\mathcal{D}) - F(\mathcal{D}' \cup \{v\}) + F(\mathcal{D}') \\ & \geq \widehat{F}(\mathcal{D} \cup \{v\}) - \widehat{F}(\mathcal{D}) - \widehat{F}(\mathcal{D}' \cup \{v\}) + \widehat{F}(\mathcal{D}') - 4k\mu \\ & \geq -4k\mu, \text{ which proves the claim.} \quad \blacksquare \end{aligned}$$

When the incoherency  $\mu$  is small, the approximation guarantee in (8) is quite useful. There has been a significant body of work establishing the existence and construction of collections  $\mathcal{V}$  of columns with low coherence  $\mu$ . For example, it is possible to achieve incoherence  $\mu \leq d^{-1/2}$  with the union of  $d/2$  orthonormal bases (c.f. Theorem 2 of Gribonval & Nielsen (2002)).

Unfortunately, when  $n = \Omega(d)$  and  $\varepsilon = 4k\mu$ , the guarantee (8) is vacuous since the maximum value of  $F$  for DiSP is 1. In Section 4, we will show that if, instead of greedily optimizing  $F$ , we optimize a *modular approximation*  $\widehat{F}_s$  of  $F_s$  (as defined below), we can improve the approximation error from  $O(nk\mu)$  to  $O(k\mu)$ .

**A modular approximation to DiSP:** The key idea behind the proof of Theorem 1 is that for incoherent dictionaries the variance reduction  $R_s(\mathcal{A}) = L_s(\emptyset) - L_s(\mathcal{A})$  is approximately additive (modular). We exploit this observation by optimizing a new objective  $\widehat{F}$  that approximates  $F$  by disregarding the non-orthogonality of  $\Phi_{\mathcal{V}}$  in sparse representation. We do this by replacing the weight calculation  $w_{s,\mathcal{A}} = \Phi_{\mathcal{A}}^\dagger y_s$  in  $F$  with  $w_{s,\mathcal{A}} = \Phi_{\mathcal{A}}^T y_s$ :

$$\widehat{F}_s(\mathcal{D}) = \max_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} \sum_{v \in \mathcal{A}} w_{s,v}, \text{ and } \widehat{F}(\mathcal{D}) = \frac{1}{m} \sum_s \widehat{F}_s(\mathcal{D}), \quad (9)$$

where  $w_{s,v} = \langle \phi_v, y_s \rangle^2$  for each  $y_s \in \mathbb{R}^d$  and  $\phi_v \in \Phi_{\mathcal{V}}$ . We call  $\widehat{F}$  a modular approximation of  $F$  as it relies on the approximate modularity of the variance reduction  $R_s$ . Note that in contrast to (3),  $\widehat{F}_s(\mathcal{D})$  in (9) can be exactly evaluated by a greedy algorithm that simply picks the  $k$  largest weights  $w_{s,v}$ . Moreover, the weights must be calculated *only once* during algorithm execution, thereby significantly increasing its efficiency.

The following immediate Corollary to Theorem 1 summarizes the essential properties of  $\widehat{F}$ :

**Corollary 1** *Suppose  $\Phi_{\mathcal{V}}$  is incoherent with constant  $\mu$ . Then, for any  $\mathcal{D} \subseteq \mathcal{V}$ , we have  $|\widehat{F}(\mathcal{D}) - F(\mathcal{D})| \leq k\mu$ . Furthermore,  $\widehat{F}$  is monotonic and submodular.*

Corollary 1 shows that  $\widehat{F}$  is a close approximation of the DiSP set function  $F$ . We exploit this modular approximation to motivate a new algorithm for DiSP and provide better performance bounds in Section 4.

#### 4. Sparsifying dictionary selection

In this section, we describe two sparsifying dictionary selection (SDS) algorithms with theoretical performance guarantees:  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$ . Both algorithms make locally greedy choices to handle the combinatorial challenges **C1** and **C2**, defined in Section 2. The algorithms differ only in the way they address **C1**, which we further describe below. Both algorithms tackle **C2** by the same greedy scheme in (6). That is, both algorithms start with the empty set and greedily add dictionary columns to solve DiSP. Interestingly, while  $\text{SDS}_{MA}$  has better theoretical guarantees and is much faster than  $\text{SDS}_{OMP}$ , Section 6 empirically shows that  $\text{SDS}_{OMP}$  often performs better.

**SDS<sub>OMP</sub>:**  $\text{SDS}_{OMP}$  employs the orthogonal matching pursuit (OMP) (Gilbert & Tropp, 2005) to approximately solve the sparse representation problem in (2) and has the following theoretical guarantee:

**Theorem 2**  *$\text{SDS}_{OMP}$  uses the scheme in (6) to build a dictionary  $\mathcal{D}_{OMP}$  one column at a time such that*

$$F(\mathcal{D}_{OMP}) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F(\mathcal{D}) - k(6n + 2 - 1/e)\mu.$$

Before we prove Theorem 2, we state the following result whose proof directly follows from Theorem 1.

**Proposition 1** *At each iteration,  $\text{SDS}_{OMP}$  approximates  $F$  with a value  $F_{OMP}$  such that  $|F_{OMP}(\mathcal{D}) - F(\mathcal{D})| \leq k\mu$  over all dictionaries  $\mathcal{D}$ .*

**Proof** [Theorem 2] From Theorem 1 and Proposition 1 we can see that  $F_{OMP}$  is  $6kn\mu$ -approximately submodular. Thus, according to Krause et al. (2008):

$$F_{OMP}(\mathcal{D}_{OMP}) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F_{OMP}(\mathcal{D}) - 6kn\mu. \quad (10)$$

Using Proposition 1, we substitute  $F(\mathcal{D}_{OMP}) + k\mu \geq F_{OMP}(\mathcal{D}_{OMP})$  and  $\max_{|\mathcal{D}| \leq n} F_{OMP}(\mathcal{D}) \geq \max_{|\mathcal{D}| \leq n} F(\mathcal{D}) - k\mu$  into (10) to prove the claim.  $\blacksquare$

**SDS<sub>MA</sub>:**  $\text{SDS}_{MA}$  greedily (according to (6)) optimizes the modular approximation (MA)  $\widehat{F}$  of the DiSP objective  $F$  and has the following guarantee:

**Theorem 3**  *$\text{SDS}_{MA}$  builds a dictionary  $\mathcal{D}_{MA}$  s.t.*

$$F(\mathcal{D}_{MA}) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F(\mathcal{D}) - (2 - 1/e)k\mu. \quad (11)$$

Corollary 1 and Theorem 2 directly imply Theorem 3.

In most realistic settings with high-dimensional signals and incoherent dictionaries, the term  $(2-1/e)k\mu$  in the approximation guarantee (11) of  $\text{SDS}_{MA}$  is negligible.

## 5. Sparsifying dictionary selection for block sparse representation

**Structured sparsity:** While many man-made and natural signals can be described as sparse in simple terms, their sparse coefficients often have an underlying, problem dependent order. For instance, modern image compression algorithms, such as JPEG, not only exploit the fact that most of the DCT coefficients of a natural image are small. Rather, they also exploit the fact that the large coefficients have a particular structure characteristic of images containing edges. Coding this structure using an appropriate model enables transform coding algorithms to compress images close to the maximum amount possible and significantly better than a naive coder that just assigns bits to each large coefficient independently (Mallat, 1999).

We can enforce structured sparsity for sparse coefficients over the learned dictionaries in DiSP, corresponding to a *restricted union-of-subspaces* (RUS) sparse model by imposing the constraint that the feasible sparsity patterns are a strict subset of all  $k$ -dimensional subspaces (Baraniuk et al., 2008). To facilitate such RUS sparse models in DiSP, we must not only determine the constituent dictionary columns, but also their arrangement within the dictionary. While analyzing the RUS model in general is challenging, we here describe below a special RUS model of broad interest to explain the general ideas.

**Block-sparsity:** Block-sparsity is abundant in many applications. In sensor networks, multiple sensors simultaneously observe a sparse signal over a noisy channel. While recovering the sparse signal *jointly* from the sensors, we can use the fact that the support of the significant coefficients of the signal are common across all the sensors. In DNA microarray applications, specific combinations of genes are also known a priori to cluster over tree structures, called dendrograms. In computational neuroscience problems, decoding of natural images in the primary visual cortex (V1) and statistical behavior of neurons in the retina exhibit clustered sparse responses.

To address block-sparsity in DiSP, we replace (3) by

$$F_i(\mathcal{D}) = \sum_{s \in B_i} L_s(\emptyset) - \min_{A \subseteq \mathcal{D}, |A| \leq k} \sum_{s \in B_i} L_s(A), \quad (12)$$

where  $B_i$  is the  $i$ -th block of signals (e.g., simultaneous recordings by multiple sensors) that must share the same sparsity pattern. Accordingly, we redefine  $F(\mathcal{D}) = \sum_i F_i(\mathcal{D})$  as the sum across blocks, rather than individual signals, as Section 6 further elaborates.

This change preserves (approximate) submodularity.

## 6. Experiments

**Finding a dictionary in a haystack:** To understand how the theoretical performance reflects on the actual performance of the proposed algorithms, we first perform experiments on synthetic data.

We generate a collection  $\mathcal{V}_U$  with 400 columns by forming a union of six orthonormal bases with  $d = 64$ , including the discrete cosine transform (DCT), different wavelet bases (Haar, Daub4, Coiflets), noiselets, and the Gabor frame. This collection  $\mathcal{V}_U$  is not incoherent—in fact, the various bases contain perfectly coherent columns. As alternatives, we first create a separate collection  $\mathcal{V}_S$  from  $\mathcal{V}_U$ , where we greedily removed columns based on their incoherence, until the remaining collection had incoherence of  $\mu_S = 0.5$ . The resulting collection contains 245 columns. We also create a collection  $\mathcal{V}_R$  with 150 random columns of  $\mathcal{V}_U$ , which results in  $\mu_R = 0.23$ .

For  $\mathcal{V}_{U,S,R}$ , we repeatedly (50 trials) pick at random a dictionary  $\mathcal{D}^* \subseteq \mathcal{V}$  of size  $n = 64$  and generate a collection of  $m = 100$  random 5-sparse signals with respect to the dictionary  $\mathcal{D}^*$ . Our goal is to recover the true dictionary  $\mathcal{D}^*$  using our SDS algorithms. For each random trial, we run  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$  to select a dictionary  $\mathcal{D}$  of size 64. We then look at the overlap  $|\mathcal{D} \cap \mathcal{D}^*|$  to measure the performance of selecting the “hidden” basis  $\mathcal{D}^*$ . We also report the fraction of remaining variance after sparse reconstruction.

Figures 2(a), 2(b), and 2(c) compare  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$  in terms of their variance reduction as a function of the selected number of columns. Interestingly, in all 50 trials,  $\text{SDS}_{OMP}$  perfectly reconstructs the hidden basis  $\mathcal{D}^*$  when selecting 64 columns for  $\mathcal{V}_{S,R}$ .  $\text{SDS}_{MA}$  performs slightly worse than  $\text{SDS}_{OMP}$ .

Figures 2(e), 2(f), and 2(g) compare the performance in terms of the fraction of incorrectly selected basis functions. Note that, as can be expected, in case of the perfectly coherent  $\mathcal{V}_U$ , even  $\text{SDS}_{OMP}$  does not achieve perfect recovery. However, even with high coherence,  $\mu = 0.5$  for  $\mathcal{V}_S$ ,  $\text{SDS}_{OMP}$  exactly identifies  $\mathcal{D}^*$ .  $\text{SDS}_{MA}$  performs a slightly worse but nevertheless correctly identifies a high fraction of  $\mathcal{D}^*$ .

In addition to exact sparse signals, we also generate compressible signals, where the coefficients have power-law with decay rate of 2. These signals can be well-approximated as sparse; however, the residual error in sparse representation creates discrepancies in measurements which can be modeled as noise in DiSP. Figures 2(d) and 2(h) repeat the above experiments for  $\mathcal{V}_S$ ; both  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$  perform quite well.

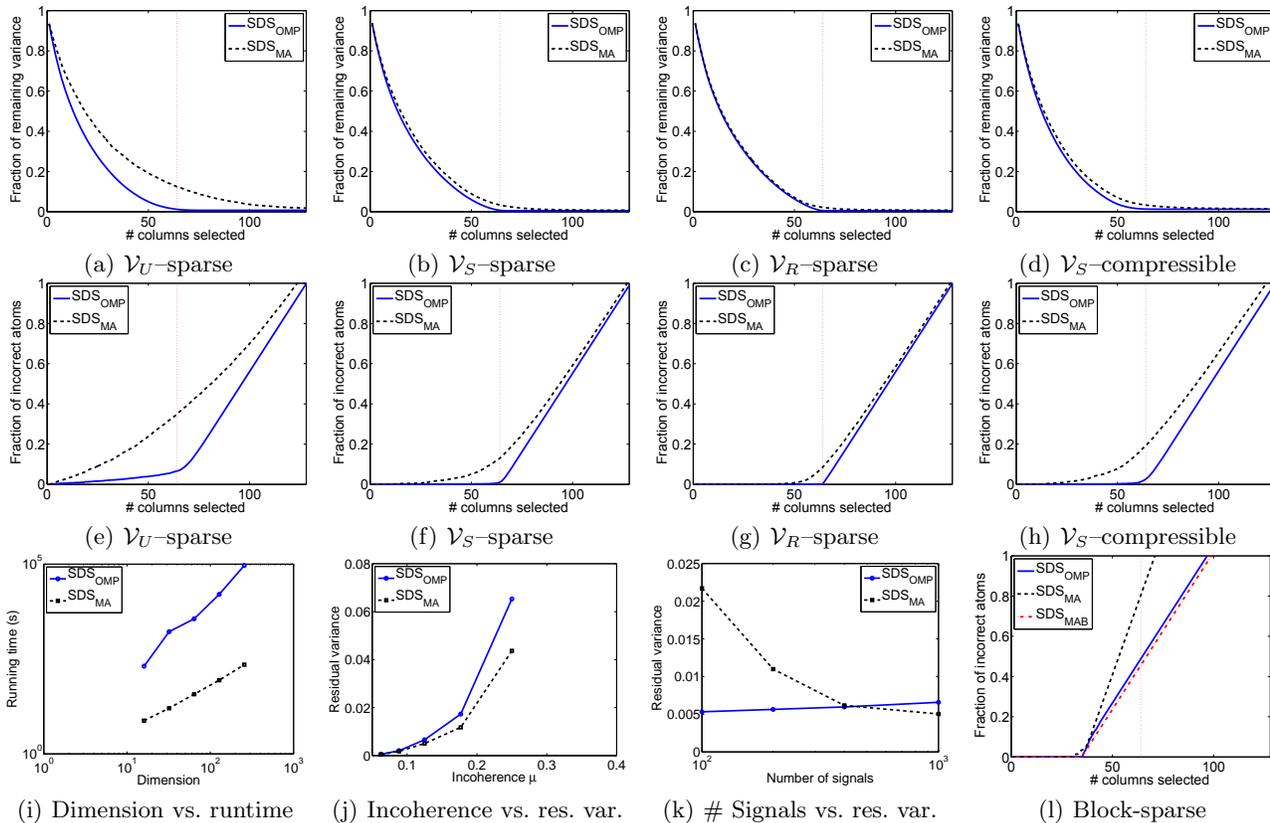


Figure 2. Results of 50 trials: (a-c) Variance reduction achieved by  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$  on the collections  $\mathcal{V}_{U,S,R}$  for 5-sparse signals in 64 dimensions. (e-g) Percentage of incorrectly selected columns on the same collections. (d) Variance reduction for compressible signals in 64 dimensions for  $\mathcal{V}_S$ . (h) Corresponding column selection performance. (i)  $\text{SDS}_{MA}$  is orders of magnitude faster than  $\text{SDS}_{OMP}$  over a broad range of dimensions. (j) As incoherence decreases, the algorithm effectiveness in variance reduction improve. (k) The variance reduction performance of  $\text{SDS}_{MA}$  improves with the number of training samples. (l) Exploiting block-sparse structure in signals leads to improved dictionary selection performance.

Figure 2(i) compares  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$  in running time. As we increase the dimensionality of the problem,  $\text{SDS}_{MA}$  is several orders of magnitude faster than  $\text{SDS}_{OMP}$  in our MATLAB implementation. Figure 2(j) illustrates the performance of the algorithms as a function of the incoherence. As predicted by Theorems 2 and 3, lower incoherence  $\mu$  leads to improved performance of the algorithms. Lastly, Figure 2(k) compares the residual variance as a function of the training set size (number of signals). Surprisingly, as the number of signals increase, the performance of  $\text{SDS}_{MA}$  improves, and even exceeds that of  $\text{SDS}_{OMP}$ .

We also test the extension of  $\text{SDS}_{MA}$  to block-sparse signals as discussed in Section 5. We generate 200 random signals each with fixed sparsity pattern, comprising 10 blocks, consisting of 20 signals each. We then compare the standard  $\text{SDS}_{MA}$  algorithm with the block-sparse variant  $\text{SDS}_{MAB}$  described in Section 5 in terms of their basis identification performance (see Figure 2(l)).  $\text{SDS}_{MAB}$  drastically outperforms  $\text{SDS}_{MA}$ , and even outperforms the  $\text{SDS}_{OMP}$

algorithm which is computationally far more expensive. Hence, exploiting prior knowledge of the problem structure can significantly aid dictionary selection.

**A battle of bases on image patches:** In this experiment, we try to find the optimal dictionary among an existing set of bases to represent natural images. Since the conventional dictionary learning approaches cannot be applied to this problem, we only present the results of  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$ .

We sample image patches from natural images, and apply our  $\text{SDS}_{OMP}$  and  $\text{SDS}_{MA}$  algorithms to select dictionaries from the collection  $\mathcal{V}_U$ , as defined above. Figures 3(a) (for  $\text{SDS}_{OMP}$ ) and 3(b) (for  $\text{SDS}_{MA}$ ) show the fractions of selected columns allocated to the different bases constituting  $\mathcal{V}_U$  for 4000 image patches of size  $8 \times 8$ . We restrict the maximum number of dictionary coefficients  $k$  for sparse representation to 10% (6). We then observe the following surprising results. While wavelets are considered to be an improvement over the DCT basis for compressing natural images (JPEG2000 vs. JPG),  $\text{SDS}_{OMP}$  prefer

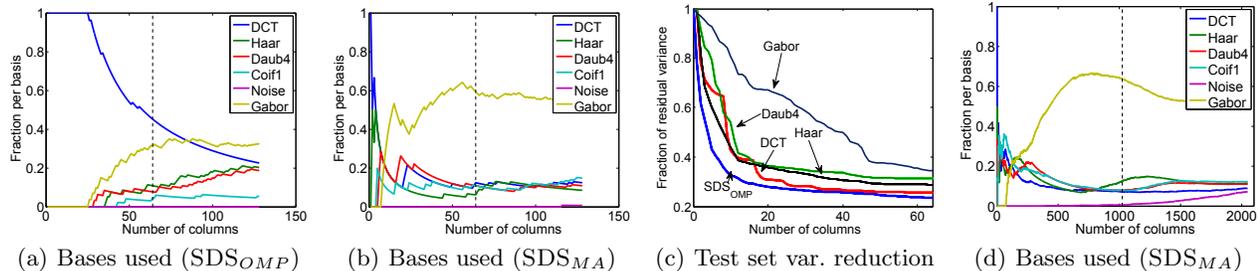


Figure 3. Experiments on natural image patches. (a,b,c) Fractions of bases selected for SDS<sub>OMP</sub> and SDS<sub>MA</sub> with  $d = 64$  (a,b), and the corresponding variance reduction on test patches. (d) Fractions of bases selected for SDS<sub>MA</sub> with  $d = 1024$ .

DCT over wavelets for sparse representation; the cross validation results show that the learned combination of DCT (global) and Gabor functions (local) are better than the wavelets (multiscale) in variance reduction (compression). In particular, Fig. 3(c) demonstrates the performance of the learned dictionary against the various bases that comprise  $\mathcal{V}_U$  on a held-out test set of 500 additional image patches. The variance reduction of the dictionary learned by SDS<sub>OMP</sub> is 8% lower than the variance reduction achieved by the best basis, which, in this case, is DCT.

Moreover, SDS<sub>MA</sub>, which trades off representation accuracy with efficient computation, overwhelmingly prefers Gabor functions that are used to model neuronal coding of natural images. The overall dictionary constituency varies for SDS<sub>OMP</sub> and SDS<sub>MA</sub>; however, the variance reduction performances are comparable. Finally, Figure 3(d) presents the fraction of selected bases for  $32 \times 32$  sized patches with  $k = 102$ , which matches well with the  $8 \times 8$  DiSP problem above.

**Dictionary selection from dimensionality reduced data:** In this experiment, 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  $\mathcal{Y}' = \mathcal{P}_1 y_1, \dots, \mathcal{P}_m y_m \in \mathbb{R}^b$ , 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. 4, albeit missing 50% of its pixels. We break the noisy image into non-overlapping  $8 \times 8$  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<sub>OMP</sub> and SDS<sub>MA</sub> algorithms,

approaches based on total-variation (TV), linear interpolation, nonlocal TV and the nonparametric Bayesian dictionary learning (based on Indian buffet processes) algorithms (Zhang & Chan, 2009; Mairal et al., 2008; Zhou et al., 2009). The TV and nonlocal TV algorithms use the linear interpolation result as their initial estimates. We set  $k = 6$  (10%). Figure 4 illustrates the inpainting results for each algorithm sorted in increasing peak signal to noise ratio (PSNR). We do not report the reconstruction results using individual candidate bases since they are significantly worse than the baseline linear interpolation.

The test image exhibits significant self similarities, restricting the degrees-of-freedom of the sparse coefficients. Hence, for our modular and OMP-based greedy algorithms, we ask the algorithms to select  $64 \times 32$  dimensional dictionaries. While the modular algorithm SDS<sub>MA</sub> selects the desired dimensions, the OMP-based greedy algorithm SDS<sub>OMP</sub> terminates when the dictionary dimensions reach  $64 \times 19$ . 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. SDS<sub>OMP</sub> performs on par with non-local 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.

## 7. Conclusions

Over the last decade, a great deal of research revolved around recovering, processing, and coding sparse signals. To leverage this experience in new problems, many researchers are now interested in automatically determining data sparsifying dictionaries for their applications. We discussed two alternatives that focus on this problem: dictionary design and dictionary learning. In this paper, we developed a combinatorial theory for dictionary selection that

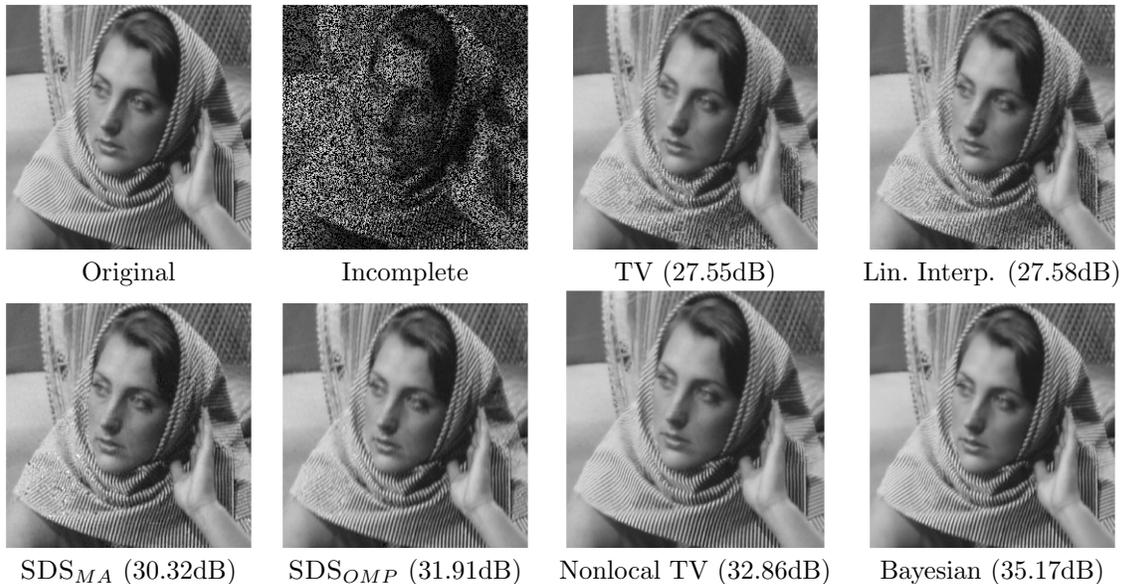


Figure 4. Comparison of inpainting algorithms.

bridges the gap between the two approaches. We explored new connections between the combinatorial structure of submodularity and the geometric concept of incoherence. We presented two computationally efficient algorithms,  $SDS_{OMP}$  based on the OMP algorithm, and  $SDS_{MA}$  using a modular approximation. By exploiting the approximate submodularity property of the DiSP objective, we derived theoretical approximation guarantees for the performance of our algorithms. We also demonstrated the ability of our learning framework to incorporate structured sparsity representations in dictionary learning. Compared to the dictionary design approaches, our approach is data adaptive and has better empirical performance on data sets. Compared to the continuous nature of the dictionary learning approaches, our approach is discrete and provides new theoretical insights to the dictionary learning problem. We believe that our results pave a promising direction for further research, exploiting combinatorial optimization for sparse representations, in particular submodularity.

**Acknowledgements.** This research is supported by ONR grants N00014-09-1-1044 and N00014-08-1-1112, NSF grant CNS-0932392, AFOSR FA9550-07-1-0301, ARO W911NF-09-1-0383, DARPA N66001-08-1-2065, ARO MURI W911NF-09-1-0383.

## References

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

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. Model-based compressive sensing. *IEEE*

*Trans. on Inform. Theory*, 56(11):1982–2001, 2010.

Cevher, V. Learning with compressible priors. In *NIPS*, Vancouver, B.C., Canada, 2008.

Choi, H. and Baraniuk, R. G. Wavelet statistical models and Besov spaces. *Lect. Notes in Statistics*, 2003.

Davis, G., Mallat, S., and Avellaneda, M. Greedy adaptive approximation. *J. Const. Approx.*, 13:57–98, 1997.

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

Gribonval, R. and Nielsen, M. Sparse decompositions in “incoherent” dictionaries. In *ICIP*, 2002.

Krause, A., Singh, A., and Guestrin, C. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. In *JMLR*, vol. 9, 2008.

Mairal, J. and Bach, F. and Ponce, J. and Sapiro, G. Online dictionary learning for sparse coding In *ICML*, 2008.

Mallat, S. G. *A wavelet tour of signal processing*. Academic Press, 1999.

Nemhauser, G., Wolsey, L., and Fisher, M. An analysis of the approximations for maximizing submodular set functions. *Math. Prog.*, 14:265–294, 1978.

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

Zhang, X. and Chan, T. F. Wavelet Inpainting by Non-local Total Variation. CAM Report (09-64), 2009.

Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., and Carin, L. Non-parametric bayesian dictionary learning for sparse image representations. *NIPS*, ’09.