Greedy Dictionary Selection for Sparse Representation

Volkan Cevher (Senior Member) and Andreas Krause

***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 SDS$_{OMP}$ and 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.**

I. 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 α 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., [1] 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 ℓ_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 [2].

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 ℓ_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 [3], [4], [5]. 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 [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 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

VC is with Swiss Federal Institute of Technology, Lausanne and the Idiap Research Institute. AK is with Swiss Federal Institute of Technology, Zurich. Email: volkan.cevher@epfl.ch, krausea@ethz.ch.
This research is supported by the European Commission under Grant MIRG-268398, DARPA KeCoM program #11-DARPA-1055, ONR N000140811112, ARO MURI W911NF-09-1-0383, AFOSR FA9550-07-1-0301, ONR grants N00014-09-1-1044 and N00014-08-1-1112, NSF grants CNS-0932392 and IIS-0953413, AFOSR FA9550-07-1-0301, ARO W911NF-09-1-0383, DARPA N66001-08-1-2065, ARO MURI W911NF-09-1-0383, and an Okawa Foundation Research Grant. VC also acknowledges Rice University ECE for his Faculty Fellowship position. AK acknowledges the California Institute of Technology.

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 SDS_{OMP} and 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 amounts to optimizing a function that is approximately submodular. We then 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 prespecified 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 show that it provides practical insights to existing image coding standards. We also provide an image inpainting example to understand the limitations of our approach as compared to dictionary learning.

This work extends our earlier work [9]. Compared to [9], we introduce a new structured sparsity model for dictionary selection in this paper to enforce sparsity *on average* for the given collection of training signals. We show that this model leads to a matroid constraint that can be readily handled within our dictionary selection framework. Additional experiments on natural images show that learning with the average sparsity model leads to better dictionaries for sparse representation on test data. Our preliminary results were also presented at [8].

The paper is organized as follows. Section II sets the stage by introducing the dictionary selection for sparse representation and describing its computational challenges. Section III unifies key geometric and combinatorial properties in dictionary selection, which motivate the use of two computationally scalable greedy approximation algorithms. Section IV then describes the algorithms along with their theoretical guarantees. Sections V and VI discuss structured models in dictionary selection for the sparse representation of individual signals as well as the signal ensembles. Section VII provide extensive numerical studies that support the effectiveness of our algorithms. Section VIII presents concluding remarks and discusses promising directions for future research.

II. The dictionary selection problem

In the *dictionary selection problem* (DiSP), we seek a dictionary $\mathcal{D}$ to sparsely represent a given collection of signals $\mathcal{Y} = \{y_1, \ldots, y_m\} \in \mathbb{R}^{d \times m}$. We compose $\mathcal{D}$ using the variance reduction metric, defined below, by selecting a subset out of a candidate set of vectors $\Phi = \{\phi_1, \ldots, \phi_N\}$, indexed by set $\mathcal{V} = \{1, \ldots, N\}$, and where each $\phi_i \in \mathbb{R}^d$. Without loss of generality, we assume $\|y_i\|_2 \leq 1$ and $\|\phi_i\|_2 = 1$, $\forall i$. In the sequel, we define $\Phi_{\mathcal{A}} = [\phi_{i_1}, \ldots, \phi_{i_Q}]$ as a matrix containing the vectors in Φ as indexed by $\mathcal{A} = \{i_1, \ldots, i_Q\}$ 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:

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

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* $\mathcal{Y}$ on the average. 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 $L_s(\emptyset) = \|y_s\|_2^2$ and $F_s(\mathcal{D})$ 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

$$F(\mathcal{D}) = \frac{1}{m}\sum_{s} F_s(\mathcal{D}). \tag{4}$$

The optimal solution to the DiSP is then given by

$$\mathcal{D}^* = \underset{|\mathcal{D}|\leq n}{\operatorname{argmax}}\, F(\mathcal{D}), \tag{5}$$

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 (5) 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 [10]. To circumvent these combinatorial challenges, the existing dictionary learning work relies on continuous relaxations, such as replacing the combinatorial sparsity constraint with the ℓ_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.

III. 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 ε, if for $\mathcal{D} \subseteq \mathcal{D}' \subseteq \mathcal{V}$ and $v \in \mathcal{V} \setminus \mathcal{D}'$ it holds that

$$F(\mathcal{D} \cup \{v\}) - F(\mathcal{D}) \geq F(\mathcal{D}' \cup \{v\}) - F(\mathcal{D}') - \varepsilon. \tag{6}$$

In the context of DiSP, the above definition implies that adding a new column v to a larger dictionary $\mathcal{D}'$ helps at most ε 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 [11] 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

$$v_i = \underset{v\in\mathcal{V}\setminus\mathcal{D}}{\operatorname{argmax}}\, G(\mathcal{D}_{i-1} \cup \{v\}), \tag{7}$$

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

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

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, [12] show that the same greedy algorithm, when applied to approximately submodular functions, instead inherits the following–slightly weaker–guarantee

$$F(\mathcal{D}_n) \geq (1 - 1/e) \max_{|\mathcal{D}|\leq n} F(\mathcal{D}) - n\varepsilon. \tag{9}$$

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

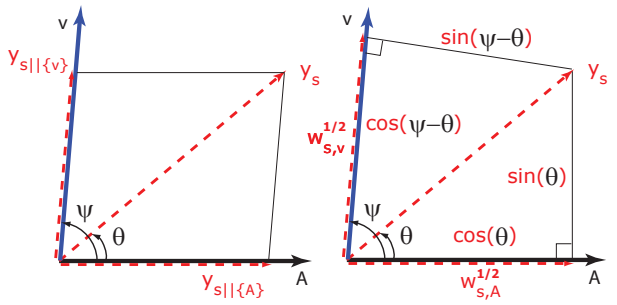


Fig. 1. *Example geometry in DiSP. (Left) Minimum error decomposition. (Right) Modular decomposition.*

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

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

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

The following theorem establishes a key relationship between ε and μ for DiSP.

Theorem 1: If $\Phi_{\mathcal{V}}$ has incoherence μ, then the variance reduction objective F in DiSP is ε-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

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

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 μ. 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., $||y_s||^2 = 1$; 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 θ and ψ are defined in Fig. 1. We thus can bound $\varepsilon_s \equiv |R_s(\mathcal{A}\cup\{v\}) - R_s(\mathcal{A}) - w_{v,s}|$ by

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

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

$$\begin{aligned}&F(\mathcal{D}\cup\{v\}) - F(\mathcal{D}) - F(\mathcal{D}'\cup\{v\}) + F(\mathcal{D}') \\ \geq&\widehat{F}(\mathcal{D}\cup\{v\}) - \widehat{F}(\mathcal{D}) - \widehat{F}(\mathcal{D}'\cup\{v\}) + \widehat{F}(\mathcal{D}') - 4k\mu \\ \geq& -4k\mu,\end{aligned}$$

which proves the claim. ∎

When the incoherency μ is small, the approximation guarantee in (9) 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 μ. For example, it is possible to achieve incoherence $\mu \approx d^{-1/2}$ with the union of $d/2$ orthonormal bases (c.f. Theorem 2 of [13]). In general settings, the Welch bound can be used to obtain a lower-bound on the value of μ.

Unfortunately, when $n = \Omega(d)$ and $\varepsilon = 4k\mu$, the guarantee (9) is vacuous since the maximum value of F for DiSP is 1. In Section IV, 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 $w_{s,\mathcal{A}} = \Phi_\mathcal{A}^T y_s$:

$$\widehat{F}_s(\mathcal{D}) = \max_{\mathcal{A}\subseteq\mathcal{D},|\mathcal{A}|\leq k} \sum_{v\in\mathcal{A}} w_{s,v}, \text{ and } \widehat{F}(\mathcal{D}) = \frac{1}{m}\sum_s \widehat{F}_s(\mathcal{D}), \tag{10}$$

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 (10) 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 corollary below follows directly from the proof of Theorem 1 and summarizes the essential properties of $\widehat{F}$:

Corollary 1: Suppose $\Phi_\mathcal{V}$ is incoherent with constant μ. 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.

Proof: Using the same arguments in Theorem 1, we first note that $|F_s(\mathcal{A}\cup\{v\}) - F_s(\mathcal{A}) - w_{s,v}| \leq \mu$. By concatenation, we then have $|F(\mathcal{D}) - \sum_{v\in\mathcal{A},\mathcal{A}\subseteq\mathcal{D},|\mathcal{A}|\leq k} w_{s,v}| \leq k\mu$, proving the desired result. ∎

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 IV.

IV. Sparsifying dictionary selection

In this section, we describe two sparsifying dictionary selection (SDS) algorithms with theoretical performance guarantees: SDS$_{OMP}$ and SDS$_{MA}$. Both algorithms make locally greedy choices to handle the combinatorial challenges **C1** and **C2**, defined in Section II. Pseudocode for the algorithms is presented in Algorithm 1. 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 (7). That is, both algorithms start with the empty set and greedily

Input: Collection of N candidate column vectors Φ; collection of m signals $\mathcal{Y}$; desired sparsity level k; bound on number n of columns selected; approximation method $M \in \{OMP, MA\}$
Output: Dictionary $\mathcal{D}$
begin
  $\mathcal{D} \leftarrow \emptyset$;
  for $\ell = 1$ **to** n **do**
    if *M=OMP* **then**
    $i^* \leftarrow \text{argmax}_{i \in \mathcal{V} \setminus \mathcal{D}} F_{OMP}(\mathcal{D} \cup \{i\}; \Phi, \mathcal{Y}, k)$;
    else if *M=MA* **then**
    $i^* \leftarrow \text{argmax}_{i \in \mathcal{V} \setminus \mathcal{D}} \hat{F}(\mathcal{D} \cup \{i\}; \Phi, \mathcal{Y}, k)$;
    Set $\mathcal{D} \leftarrow \mathcal{D} \cup \{i^*\}$;
  end
end

Algorithm 1: The SDS_{OMP} and SDS_{MA} algorithms

Input: Collection of N candidate column vectors Φ; collection of m signals $\mathcal{Y}$; desired sparsity level k; candidate dictionary $\mathcal{D}$
Output: Value $F_{OMP}(\mathcal{D}) = F_{OMP}(\mathcal{D}; \Phi, \mathcal{Y}, k)$
begin
  for $s = 1$ **to** m **do**
    Use OMP to approximately solve $\mathcal{A}_s = \text{argmin}_{\mathcal{A} \subseteq \mathcal{D}: ||\mathcal{A}|| \leq k} L_s(\mathcal{A})$;
    $r_s \leftarrow L_s(\emptyset) - L_s(\mathcal{A}_s)$;
  end
  return $\frac{1}{m} \sum_{s=1}^{m} r_s$
end

Algorithm 2: Algorithm for computing F_{OMP}

add dictionary columns to solve DiSP. Interestingly, while SDS_{MA} has better theoretical guarantees and is much faster than SDS_{OMP}, Section VII empirically shows that SDS_{OMP} often performs better.

SDS_{OMP}: SDS_{OMP} employs the orthogonal matching pursuit (OMP) [14] to approximately solve the sparse representation problem in (2). It greedily maximizes F_{OMP} (pseudo code for evaluating F_{OMP} is given in Algorithm 2), and has the following theoretical guarantee:

Theorem 2: SDS_{OMP} uses the scheme in (7) to build a dictionary $\mathcal{D}_{\text{OMP}}$ one column at a time such that

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

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

Input: Collection of N candidate column vectors Φ; collection of m signals $\mathcal{Y}$; desired sparsity level k; candidate dictionary $\mathcal{D}$
Output: Value $\widehat{F}(\mathcal{D}) = \widehat{F}(\mathcal{D}; \Phi, \mathcal{Y}, k)$
begin
  for $s = 1$ **to** m **do**
    for $v = 1$ **to** N **do** $\hat{w}_{s,v} \leftarrow \phi_v^T y_s$;
    Sort $\hat{w}_{s,1}, \ldots, \hat{w}_{s,N}$, and let $i_1 \neq \cdots \neq i_N$ s.t. $\hat{w}_{s,i_1} \geq \cdots \geq \hat{w}_{s,i_N}$;
    $r_s \leftarrow \sum_{\ell=1}^{k} \hat{w}_{s,i_\ell}$;
  end
  return $\frac{1}{m} \sum_{s=1}^{m} r_s$
end

Algorithm 3: Algorithm for computing $\widehat{F}$

Proposition 1: At each iteration, 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 of Theorem 2: From Theorem 1 and Proposition 1 we can see that F_{OMP} is $6kn\mu$-approximately submodular. Thus, according to [12]:

$$F_{OMP}(\mathcal{D}_{\text{OMP}}) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F_{OMP}(\mathcal{D}) - 6kn\mu. \tag{11}$$

Using Proposition 1, we substitute $F(\mathcal{D}_{\text{OMP}}) + k\mu \geq F_{OMP}(\mathcal{D}_{\text{OMP}})$ and $\max_{|\mathcal{D}| \leq n} F_{OMP}(\mathcal{D}) \geq \max_{|\mathcal{D}| \leq n} F(\mathcal{D}) - k\mu$ into (11) to prove the claim. ■

SDS_{MA}: SDS_{MA} greedily (according to (7)) optimizes the modular approximation (MA) $\widehat{F}$ of the DiSP objective F (pseudo code for evaluating $\widehat{F}$ is given in Algorithm 3) and has the following guarantee:

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

$$F(\mathcal{D}_{MA}) \geq (1 - 1/e) \max_{|\mathcal{D}| \leq n} F(\mathcal{D}) - (2 - 1/e)k\mu. \tag{12}$$

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 (12) of SDS_{MA} is negligible. Note that the approximation guarantee of Theorems 2 and 3 is stated in terms of the variance reduction F, instead of the residual reconstruction error. We leave the derivation of approximation guarantees for the reconstruction error as an open problem for future work.

At the time of this publication, [15] improved our additive bounds on dictionary selection with multiplicative bounds by using a new concept called submodularity ratio.

V. 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 [16].

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 [17]. 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

$$F_i(\mathcal{D}) = \sum_{s \in B_i} L_s(\emptyset) - \min_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \leq k} \sum_{s \in B_i} L_s(\mathcal{A}), \qquad (13)$$

where $B_i \subseteq \{1, \ldots, m\}$ 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 VII further elaborates. This change preserves (approximate) submodularity.

VI. Dictionary selection for average sparsity

When facing a large collection of natural signals, it is only expected that some signals carry a lot of information (e.g., faces in natural images), whereas other signals can be compressed using a only few non-zero coefficients (e.g., flat background). In such settings, it may be advantageous to use different amounts of compression for different signals. Thus, a valid question is whether sparsifying dictionaries can be selected for which signals can be represented using a small number of columns *on average*.

In this section, we explain how our dictionary selection framework allows to handle an average sparsity structure for signal ensembles for the dictionary selection problem. To define our model, we reformulate the variance reduction objective $F(\mathcal{D})$ from (4) as

$$F_{avg}(\mathcal{D}) = \max_{\substack{\mathcal{A}_1, \ldots, \mathcal{A}_m \subseteq \mathcal{D} \\ |\mathcal{A}_s| \leq k', \sum_s |\mathcal{A}_s| \leq mk}} \frac{1}{m} \sum_s \Big(L_s(\emptyset) - L_s(\mathcal{A}_s) \Big). \qquad (14)$$

Thus, the value of a dictionary $\mathcal{D}$ is the average variance reduction across all signals, where each signal s is represented using a set $\mathcal{A}_s$ of at most k' columns from $\mathcal{D}$. For generality, we also impose an additional constraint that at most mk columns are selected overall, where k' is given as a parameter.

At first glance, the problem

$$\max_{|\mathcal{D}| \leq n} F_{avg}(\mathcal{D})$$

appears to be more challenging: Previously, in order to evaluate $F(\mathcal{D})$, we had to solve m sparse reconstruction problems with *fixed* sparsity budget for each signal. Now, in addition, we have to optimize over the number of columns $|\mathcal{A}_s|$ selected for each signal s.

Fortunately, we can still resort to our modular approximation technique: Reusing the notation from Section III, we define the modular approximation

$$\widehat{F}_{avg}(\mathcal{D}) = \max_{\substack{\mathcal{A}_1, \ldots, \mathcal{A}_m \subseteq \mathcal{D} \\ |\mathcal{A}_s| \leq k', \sum_s |\mathcal{A}_s| \leq mk}} \frac{1}{m} \sum_s \widehat{R}_s(\mathcal{A}_s),$$

where $\widehat{R}_s(\mathcal{A}) = \sum_{v \in \mathcal{A}} w_{s,v}$ is the modular approximation to the variance reduction for signal s using columns $\mathcal{A}$.

We have the following result, which strictly generalizes Corollary 1:

Theorem 4: Suppose $\Phi_{\mathcal{V}}$ is incoherent with constant μ. Then, for any $\mathcal{D} \subseteq \mathcal{V}$, we have $|\widehat{F}_{avg}(\mathcal{D}) - F_{avg}(\mathcal{D})| \leq k'\mu$. Furthermore, $\widehat{F}_{avg}$ is monotonic and submodular.

Proof: Similar arguments as used in the proof of Theorem 1 show that $|\widehat{R}_s(\mathcal{A}) - R_s(\mathcal{A})| \leq k'\mu$ for all

signals s and sets $\mathcal{A}$ s.t. $|\mathcal{A}| \leq k'$. This immediately proves the first claim. Monotonicity of $\widehat{F}_{avg}$ is immediate as well. It remains to prove the submodularity of $\widehat{F}_{avg}$. Define the set $\mathcal{V}' = \{1, \ldots, m\} \times \mathcal{V}$ of all pairs of signals and columns of Φ. Any subset $\mathcal{A}' \subseteq \Phi'$ can thus be interpreted as a "joint support set", indicating which columns of $\mathcal{V}$ are used as support in each of the m signals. Define set function $G : 2^{\mathcal{V}'} \to \mathbb{R}$ as

$$G(\mathcal{A}') = \sum_{(s,v) \in \mathcal{A}'} w_{s,v},$$

i.e., is the total weight of the joint support in the modular approximation. G is a modular function. Call a subset $\mathcal{A}' \subseteq \mathcal{V}'$ *independent* if $|\mathcal{A}'| \leq mk$ and, for all s, $|(\{s\} \times \mathcal{V}) \cap \mathcal{A}'| \leq k'$. Thus, a candidate joint support set $\mathcal{A}'$ is independent if it respects the constraints dictated by average sparsity. Now fix a subset $\mathcal{W}' \subseteq \mathcal{V}'$, and let $\mathcal{I}(\mathcal{W}') \subseteq 2^{\mathcal{V}'}$ be the collection of all subsets $\mathcal{A}' \subseteq \mathcal{W}'$ that are independent. It can be seen that for any $\mathcal{W}'$, the pair $(\mathcal{W}', \mathcal{I}(\mathcal{W}'))$ forms a matroid, and $\mathcal{I}(\mathcal{W}_1) \subseteq \mathcal{I}(\mathcal{W}_2)$ whenever $\mathcal{W}_1 \subseteq \mathcal{W}_2$. Further,

$$\max_{\substack{\mathcal{A}_1, \ldots, \mathcal{A}_m \subseteq \mathcal{D} \\ |\mathcal{A}_s| \leq k', \sum_s |\mathcal{A}_s| \leq mk}} \sum_s \widehat{R}_s(\mathcal{A}_s) = \max_{\mathcal{A}' \in \mathcal{I}(\{1,\ldots,m\} \times \mathcal{D})} G(\mathcal{A}').$$

Proposition 3.2 of [11] now proves that the function

$$\widehat{F}_{avg}(\mathcal{D}) = \max_{\mathcal{A}' \in \mathcal{I}(\bigcup_{v \in \mathcal{D}} \{1,\ldots,m\} \times \{v\})} G(\mathcal{A}')$$

is submodular. ∎

Thus, the results of Theorems 2 and 3 generalize, with k replaced by k'. In addition, the proof of Theorem 4 suggests an efficient algorithm for evaluating $\widehat{F}_{avg}(\mathcal{D})$. The problem

$$\max_{\substack{\mathcal{A}_1, \ldots, \mathcal{A}_m \subseteq \mathcal{D} \\ |\mathcal{A}_s| \leq k', \sum_s |\mathcal{A}_s| \leq mk}} \sum_s \widehat{R}_s(\mathcal{A}_s)$$

requires maximizing a modular function subject to a matroid constraint, which is optimally solved using a greedy algorithm: We start with $\mathcal{A}_s = \emptyset$ for all s, and then greedily choose the pair (s, v) such that all constraints remain satisfied, and w_{s^*,v^*} is maximized. We then add column v^* to set $\mathcal{A}_{s^*}$. We continue until no more elements can be added. The resultant collection of support sets $\mathcal{A}_1, \ldots, \mathcal{A}_m$ satisfies $\widehat{F}_{avg}(\mathcal{D}) = \frac{1}{m} \sum_s \widehat{R}_s(\mathcal{A}_s)$.

Note that if we set $k' = \Omega(\sqrt{d})$, then even for the case of incoherent ($\mu = \Omega(\frac{1}{\sqrt{d}})$) collections $\mathcal{V}$ of columns, the guarantees of Theorem 4 can be rather weak. However, in practice, one likely intends to limit the maximum number of coefficients used to represent each signal k', for example, to counter overfitting. In such cases, where k' is a small constant, the guarantees of Theorem 4 are quite useful.

VII. Experiments

We evaluate our SDS$_{OMP}$ and SDS$_{MA}$ algorithms on several sparse representation problems both on synthetic and real data. In our implementation, we use lazy evaluations [18] to speed up the SDS$_{OMP}$ and SDS$_{MA}$ algorithms.

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.

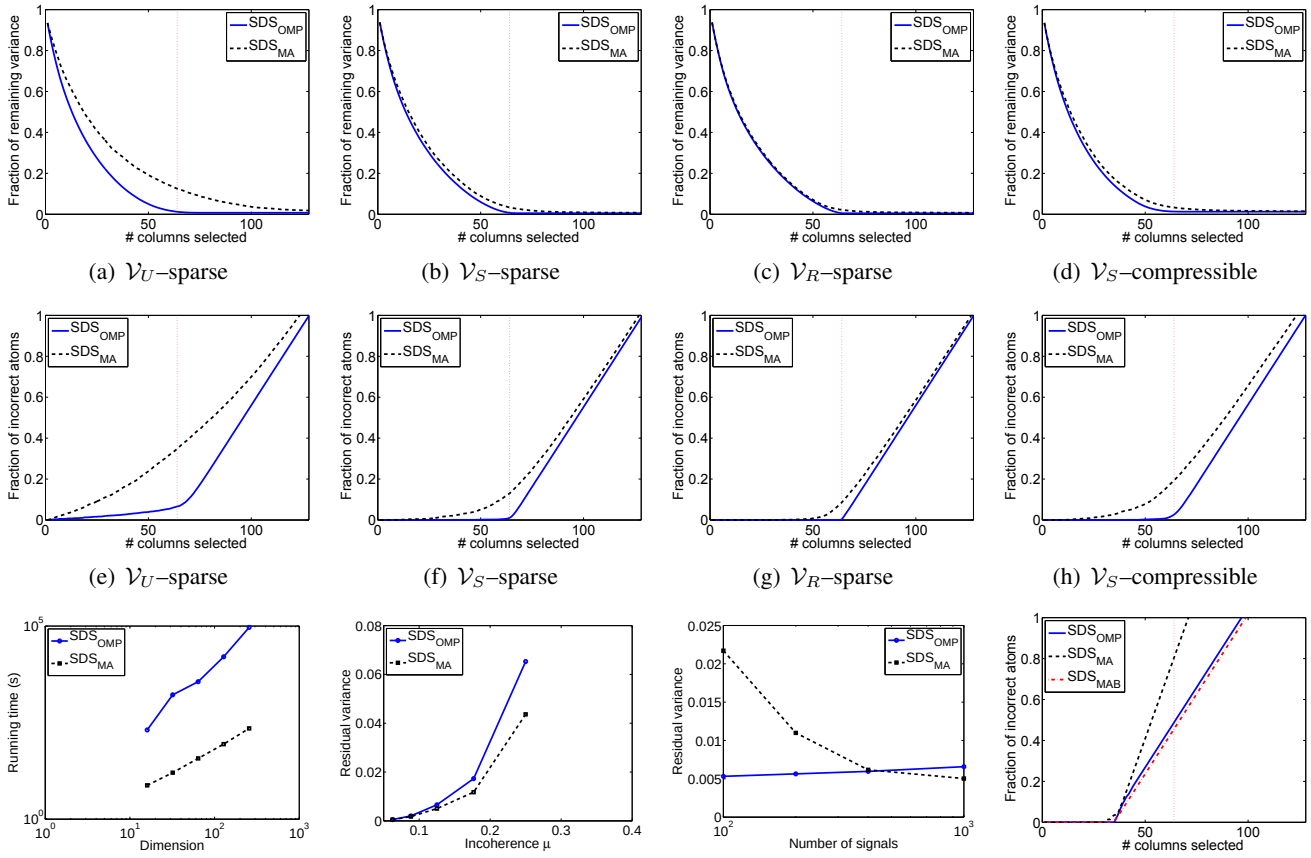
We generate a collection Φ_U with 400 columns by forming a union of five orthonormal bases and a normalized tight frame with $d = 64$, including the discrete cosine transform (DCT), different wavelet bases (Haar, Daub4, Coiflets), noiselets, and the Gabor frame. This collection Φ_U is not incoherent—in fact, the various bases contain perfectly coherent columns. As alternatives, we first create a separate collection Φ_S from Φ_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 Φ_R with 150 random columns, which results in $\mu_R = 0.2$.

For each of Φ_U, Φ_S and Φ_R with respective index sets $\mathcal{V}_U$, $\mathcal{V}_S$ and $\mathcal{V}_R$, we repeatedly (50 trials) pick at random a dictionary $\mathcal{D}^* \subseteq \mathcal{V}$ (where $\mathcal{V} \in \{\mathcal{V}_U, \mathcal{V}_S, \mathcal{V}_R\}$) 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 SDS$_{OMP}$ and 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 SDS$_{OMP}$ and SDS$_{MA}$ in terms of their variance reduction as a function of the selected number of columns. Interestingly, in all 50 trials, SDS$_{OMP}$ perfectly reconstructs the hidden basis $\mathcal{D}^*$ when selecting 64 columns for Φ_S and Φ_R. SDS$_{MA}$ performs slightly worse than 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 Φ_U, even SDS$_{OMP}$ does not achieve perfect recovery. However, even with high coherence, $\mu = 0.5$ for Φ_S, SDS$_{OMP}$ exactly identifies $\mathcal{D}^*$. 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



(a) $\mathcal{V}_U$–sparse (b) $\mathcal{V}_S$–sparse (c) $\mathcal{V}_R$–sparse (d) $\mathcal{V}_S$–compressible

(e) $\mathcal{V}_U$–sparse (f) $\mathcal{V}_S$–sparse (g) $\mathcal{V}_R$–sparse (h) $\mathcal{V}_S$–compressible

(i) Dimension vs. runtime (j) Incoherence vs. res. var. (k) # Signals vs. res. var. (l) Block-sparse

Fig. 2. *Results of 50 trials: (a-c) Variance reduction achieved by* SDS_{OMP} *and* SDS_{MA} *on the collections* $\mathcal{V}_U$, $\mathcal{V}_S$ *and* $\mathcal{V}_R$ *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)* SDS_{MA} *is orders of magnitude faster than* 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* SDS_{MA} *improves with the number of training samples. (l) Exploiting block-sparse structure in signals leads to improved dictionary selection performance.*

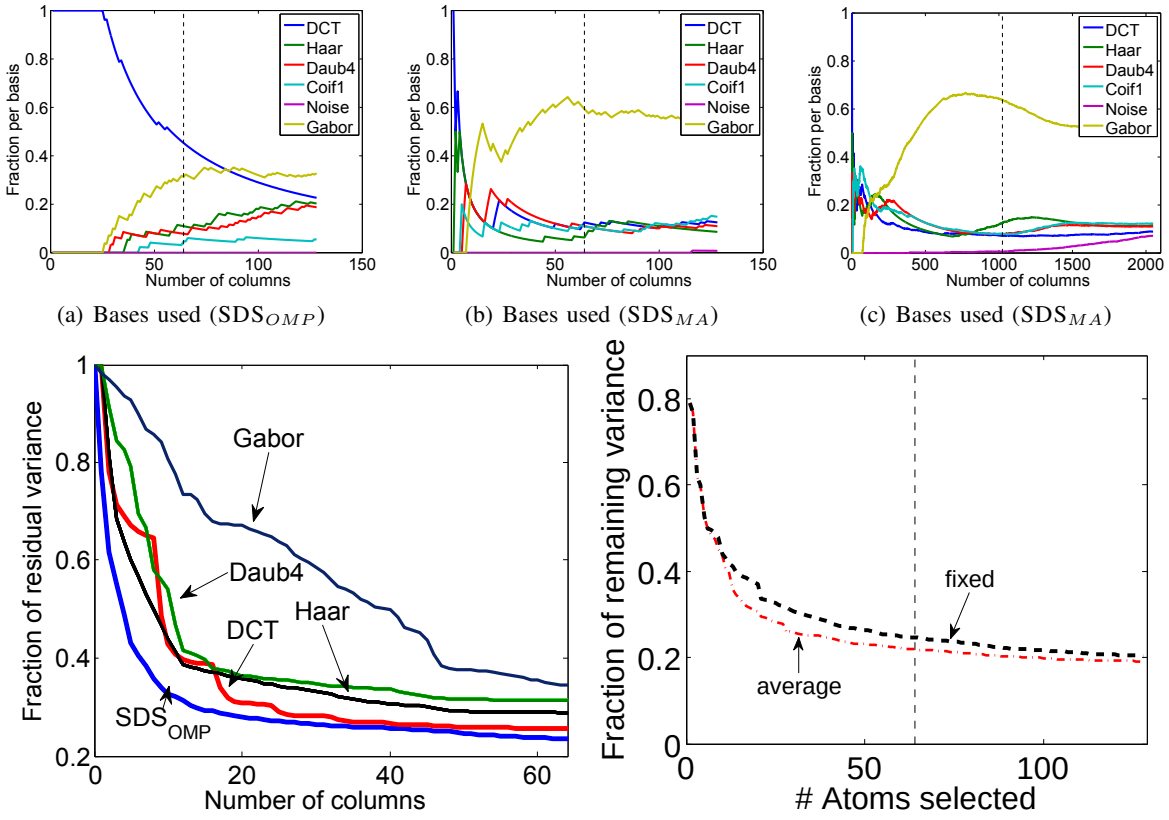
power-law with decay rate of 2. These signals can be well-approximated as sparse; however, the residual error in sparse representation creates discrepancies in measurements which can be modeled as noise in DiSP. Figures 2(d) and 2(h) repeat the above experiments for Φ_S; both SDS_{OMP} and SDS_{MA} perform quite well.

Figure 2(i) compares SDS_{OMP} and SDS_{MA} in running time. As we increase the dimensionality of the problem, SDS_{MA} is several orders of magnitude faster than 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 μ 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 SDS_{MA} improves, and even exceeds that of SDS_{OMP}.

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

We sample image patches from natural images, and apply our SDS_{OMP} and SDS_{MA} algorithms to select dictionaries from the collection Φ_U, as defined above.



(a) Bases used (SDS_{OMP}) (b) Bases used (SDS_{MA}) (c) Bases used (SDS_{MA})

(d) Test set var. reduction (e)

Fig. 3. *Experiments on natural image patches. (a,b) Fractions of bases selected for* SDS_{OMP} *and* SDS_{MA} *with* $d = 64$. *(c) Fractions of bases selected for* SDS_{MA} *with* $d = 1024$. *(d) The variance reduction on test patches with* $d = 64$. *(e) The variance reduction on test patches with* $d = 64$, *when average sparsity criterion is used.*

Figures 3(a) (for SDS_{OMP}) and 3(b) (for SDS_{MA}) show the fractions of selected columns allocated to the different bases constituting Φ_U for 4000 image patches of size 8×8. 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), SDS_{OMP} prefer 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(d) demonstrates the performance of the learned dictionary against the various bases that comprise Φ_U on a held-out test set of 500 additional image patches. The variance reduction of the dictionary learned by SDS_{OMP} is 8% lower than the variance reduction achieved by the best basis, which, in this case, is DCT.

Moreover, SDS_{MA}, 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_{OMP} and SDS_{MA}; however, the variance reduction performances are comparable. Finally, Figure 3(c) presents the fraction of selected bases for 32×32 sized patches with $k = 102$, which matches well with the 8×8 DiSP problem above.

Figure 3(e) illustrates that the average sparsity assumption can significantly improve the variance reduction objective, when applied to natural images. In this example, we train the dictionary using the average variance reduction criterion; however, we enforce hard sparsity during representation. It is then surprising that only 32 columns are selected with the average sparsity criterion is able to achieve the same amount of the variance reduction when trained with the hard sparsity constraint. We believe that this formulation circumvents the bias caused by the self similar patches, alleviating the column selection process to explore a better column range of the data.

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 Sec-

tion II, 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. 4, albeit missing 50% of its pixels. We break the noisy image into non-overlapping 8×8 patches, and train a dictionary for sparse reconstruction of those patches to minimize the average approximation error on the observed pixels. To form the candidate vectors, we use DCT, Haar and Daub4 wavelets, Coiflets, and Gabor frame. 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. We set $k = 6$ (10%). Figure 4 illustrates the inpainting results for each algorithm sorted in increasing peak signal to noise ratio (PSNR). Figure 4 also shows the PSNR value of the DCT basis alone 29.47dB. The other bases by themselves obtain 26.87dB (Haar), 27.18dB (Daub4), 26.16 (Coiflet), and 11.64dB (noiselet). Gabor frame obtains a denoising performance of 12.36dB by itself with the same sparsity constraint.

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 64×32 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. SDS_{OMP} performs 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.

VIII. 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 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.

References

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

[2] V. Cevher, "Learning with compressible priors," in *NIPS*, (Vancouver, B.C., Canada), 2008.

[3] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 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, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM New York, NY, USA, 2009.

[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*, vol. 54, no. 11, pp. 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] V. Cevher and A. Krause, "Greedy dictionary selection for sparse representation." Presented at the Neuronal Information Processing Systems Workshop on Discrete Optimization in Machine Learning, 2009.

[9] A. Krause and V. Cevher, "Submodular dictionary selection for sparse representation," in *Proceedings of the 27th Annual International Conference on Machine Learning*, 2010.

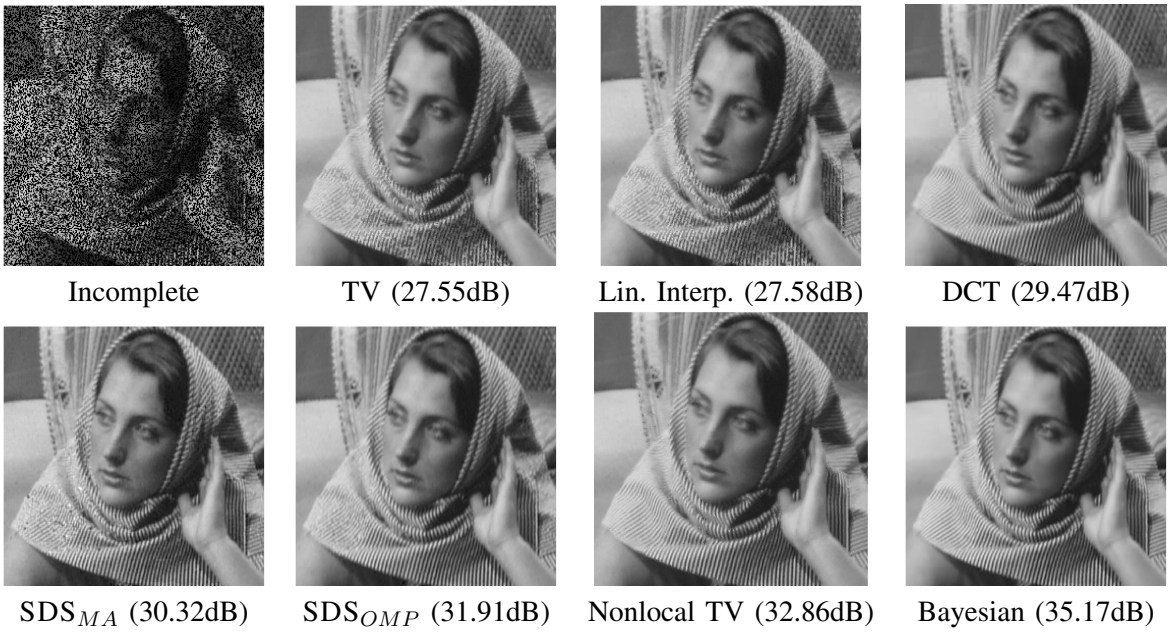


Fig. 4. *Comparison of inpainting algorithms.*

[10] G. Davis, S. Mallat, and M. Avellaneda, "Greedy adaptive approximation," *Journal of Constructive Approximation*, vol. 13, pp. 57–98, 1997.

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

[12] 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*, vol. 9, 2008.

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

[14] A. C. Gilbert and J. A. Tropp, "Signal recovery from random measurements via orthogonal matching pursuit," tech. rep., University of Michigan, 2005.

[15] A. Das and D. Kempe, "Submodular meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection," *Arxiv preprint arXiv:1102.3975*, 2011.

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

[17] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *preprint*, 2008.

[18] M. Minoux, "Accelerated greedy algorithms for maximizing submodular set functions," *Optimization Techniques, LNCS*, pp. 234–243, 1978.

Volkan Cevher received his BSc degree (valedictorian) in Electrical Engineering from Bilkent University in 1999, and his PhD degree in Electrical and Computer Engineering from Georgia Institute of Technology in 2005. He held research scientist positions at University of Maryland, College Park during 2006-2007 and at Rice University during 2008-2009. Currently, he is an assistant professor at Swiss Federal Institute of Technology Lausanne with joint appointment at the Idiap Research Institute and a Faculty Fellow at Rice University. His research interests include signal processing theory, machine learning, graphical models, and information theory. Dr. Cevher received a best paper award at SPARS in 2009 and an ERC StG in 2011.

Andreas Krause received his Diploma in Computer Science and Mathematics from the Technical University of Munich, Germany (2004) and his PhD in Computer Science from Carnegie Mellon University (2008). He joined the California Institute of Technology as an assistant professor of computer science in 2009, and is currently assistant professor in the Department of Computer Science at the Swiss Federal Institute of Technology Zurich. His research is in adaptive systems that actively acquire information, reason and make decisions in large, distributed and uncertain domains, such as sensor networks and the Web. Dr. Krause is a 2010 NAS Kavli Fellow, and received an NSF CAREER award, the Okawa Foundation Research Grant recognizing top young researchers in telecommunications, as well as awards at several premier conferences (KDD, IPSN, ICML, UAI) and the ASCE Journal of Water Resources Planning and Management.