

# Model-Based Compressive Sensing

Richard G. Baraniuk, *Fellow, IEEE*, Volkan Cevher, *Member, IEEE*,  
Marco F. Duarte, *Member, IEEE*, and Chinmay Hegde, *Student Member, IEEE*

**Abstract**—Compressive sensing (CS) is an alternative to Shannon/Nyquist sampling for the acquisition of sparse or compressible signals that can be well approximated by just  $K \ll N$  elements from an  $N$ -dimensional basis. Instead of taking periodic samples, CS measures inner products with  $M < N$  random vectors and then recovers the signal via a sparsity-seeking optimization or greedy algorithm. **Standard CS dictates that robust signal recovery is possible from  $M = \mathcal{O}(K \log(N/K))$  measurements. It is possible to substantially decrease  $M$  without sacrificing robustness by leveraging more realistic signal models that go beyond simple sparsity and compressibility by including structural dependencies between the values and locations of the signal coefficients. This paper introduces a model-based CS theory that parallels the conventional theory and provides concrete guidelines on how to create model-based recovery algorithms with provable performance guarantees. A highlight is the introduction of a new class of structured compressible signals along with a new sufficient condition for robust structured compressible signal recovery that we dub the restricted amplification property, which is the natural counterpart to the restricted isometry property of conventional CS. Two examples integrate two relevant signal models — wavelet trees and block sparsity — into two state-of-the-art CS recovery algorithms and prove that they offer robust recovery from just  $M = \mathcal{O}(K)$  measurements. Extensive numerical simulations demonstrate the validity and applicability of our new theory and algorithms.**

**Index Terms**—Compressive sensing, sparsity, signal model, union of subspaces, wavelet tree, block sparsity

## I. INTRODUCTION

WE ARE in the midst of a digital revolution that is enabling the development and deployment of new sensors and sensing systems with ever increasing fidelity and resolution. The theoretical foundation is the Shannon/Nyquist sampling theorem, which states that a signal's information is preserved if it is uniformly sampled at a rate at least two times faster than its Fourier bandwidth. Unfortunately, in many important and emerging applications, the resulting Nyquist rate

can be so high that we end up with too many samples and must compress in order to store or transmit them. In other applications the cost of signal acquisition is prohibitive, either because of a high cost per sample, or because state-of-the-art samplers cannot achieve the high sampling rates required by Shannon/Nyquist. Examples include radar imaging and exotic imaging modalities outside visible wavelengths.

Transform compression systems reduce the effective dimensionality of an  $N$ -dimensional signal  $x$  by re-representing it in terms of a sparse or compressible set of coefficients  $\alpha$  in a basis expansion  $x = \Psi\alpha$ , with  $\Psi$  an  $N \times N$  basis matrix. By sparse we mean that only  $K \ll N$  of the coefficients  $\alpha$  are nonzero and need to be stored or transmitted. By compressible we mean that the coefficients  $\alpha$ , when sorted, decay rapidly enough to zero that  $\alpha$  can be well-approximated as  $K$ -sparse. The sparsity and compressibility properties are pervasive in many signal classes of interest. For example, smooth signals and images are compressible in the Fourier basis, while piecewise smooth signals and images are compressible in a wavelet basis [1]; the JPEG and JPEG2000 standards are examples of practical transform compression systems based on these bases.

Compressive sensing (CS) provides an alternative to Shannon/Nyquist sampling when the signal under acquisition is known to be sparse or compressible [2–4]. In CS, we measure not periodic signal samples but rather inner products with  $M \ll N$  measurement vectors. In matrix notation, the measurements  $y = \Phi x = \Phi \Psi \alpha$ , where the rows of the  $M \times N$  matrix  $\Phi$  contain the measurement vectors. While the matrix  $\Phi \Psi$  is rank deficient, and hence loses information in general, it can be shown to preserve the information in sparse and compressible signals if it satisfies the so-called *restricted isometry property* (RIP) [3]. Intriguingly, a large class of random matrices have the RIP with high probability. To recover the signal from the compressive measurements  $y$ , we search for the sparsest coefficient vector  $\alpha$  that agrees with the measurements. To date, research in CS has focused primarily on reducing both the number of measurements  $M$  (as a function of  $N$  and  $K$ ) and on increasing the robustness and reducing the computational complexity of the recovery algorithm. Today's state-of-the-art CS systems can robustly recover  $K$ -sparse and compressible signals from just  $M = \mathcal{O}(K \log(N/K))$  noisy measurements using polynomial-time optimization solvers or greedy algorithms.

While this represents significant progress from Nyquist-rate sampling, our contention in this paper is that it is possible to do even better by more fully leveraging concepts from state-of-the-art signal compression and processing algorithms.

Manuscript received August 26, 2008; revised July 2, 2009. The authors are listed alphabetically. This work was supported by grants NSF CCF-0431150, CCF-0728867, CNS-0435425, and CNS-0520280, DARPA/ONR N66001-08-1-2065, ONR N00014-07-1-0936, N00014-08-1-1067, N00014-08-1-1112, and N00014-08-1-1066, AFOSR FA9550-07-1-0301, ARO MURI W911NF-07-1-0185, and the Texas Instruments Leadership University Program, and was completed while M. F. Duarte was a Ph.D. student at Rice University. The material in this paper was presented in part at the SIAM Conference on Imaging Science, San Diego, CA, June 2008, and at the Conference on Information Sciences and Systems (CISS), Baltimore, MD, March 2009.

R. G. Baraniuk, V. Cevher, and C. Hegde are with Rice University, Houston, TX 77005 USA (e-mail: richb@rice.edu; vcevher@rice.edu; chinmay@rice.edu).

V. Cevher is also with Ecole Polytechnique Fédérale de Laussane, Laussane, Switzerland (e-mail: volkan.cevher@epfl.ch).

M. F. Duarte is with Princeton University, Princeton, NJ 08544 USA (e-mail: mduarte@princeton.edu).

In many such algorithms, the key ingredient is a more realistic *structured sparsity model* that goes beyond simple sparsity by codifying the inter-dependency *structure* among the signal coefficients  $\alpha$ .<sup>1</sup> For instance, modern wavelet image coders exploit not only the fact that most of the wavelet coefficients of a natural image are small but also the fact that the values and locations of the large coefficients have a particular structure. Coding the coefficients according to a structured sparsity model enables these algorithms to compress images close to the maximum amount possible – significantly better than a naïve coder that just processes each large coefficient independently. We have previously developed a new CS recovery algorithm that promotes structure in the sparse representation by tailoring the recovered signal according to a sparsity-promoting probabilistic model, such as an Ising graphical model [5]. Such probabilistic models favor certain configurations for the magnitudes and indices of the significant coefficients of the signal.

In this paper, we expand on this concept by introducing a model-based CS theory that parallels the conventional theory and provides concrete guidelines on how to create structured signal recovery algorithms with provable performance guarantees. By reducing the number of degrees of freedom of a sparse/compressible signal by permitting only certain configurations of the large and zero/small coefficients, structured sparsity models provide two immediate benefits to CS. First, they enable us to reduce, in some cases significantly, the number of measurements  $M$  required to stably recover a signal. Second, during signal recovery, they enable us to better differentiate true signal information from recovery artifacts, which leads to a more robust recovery.

To precisely quantify the benefits of model-based CS, we introduce and study several new theoretical concepts that could be of more general interest. We begin with structured sparsity models for  $K$ -sparse signals and make precise how the structure reduces the number of potential sparse signal supports in  $\alpha$ . Then using the *model-based restricted isometry property* from [6,7], we prove that such *structured sparse signals* can be robustly recovered from noisy compressive measurements. Moreover, we quantify the required number of measurements  $M$  and show that for some structured sparsity models  $M$  is independent of  $N$ . These results unify and generalize the limited related work to date on structured sparsity models for strictly sparse signals [6–10]. We then introduce the notion of a *structured compressible signal*, whose coefficients  $\alpha$  are no longer strictly sparse but have a structured power-law decay. To establish that structured compressible signals can be robustly recovered from compressive measurements, we generalize the standard RIP to a new *restricted amplification property* (RAmP). Using the RAmP, we show that the required number of measurements  $M$  for recovery of structured com-

pressible signals is independent of  $N$ .

To take practical advantage of this new theory, we demonstrate how to integrate structured sparsity models into two state-of-the-art CS recovery algorithms, CoSaMP [11] and iterative hard thresholding (IHT) [12–16]. The key modification is surprisingly simple: we merely replace the nonlinear sparse approximation step in these greedy algorithms with a structured sparse approximation. Thanks to our new theory, both new model-based recovery algorithms have provable robustness guarantees for both structured sparse and structured compressible signals.

To validate our theory and algorithms and demonstrate their general applicability and utility, we present two specific instances of model-based CS and conduct a range of simulation experiments. The first structured sparsity model accounts for the fact that the large wavelet coefficients of piecewise smooth signals and images tend to live on a rooted, connected *tree structure* [17]. Using the fact that the number of such trees is much smaller than  $\binom{N}{K}$ , the number of  $K$ -sparse signal supports in  $N$  dimensions, we prove that a tree-based CoSaMP algorithm needs only  $M = \mathcal{O}(K)$  measurements to robustly recover tree-sparse and tree-compressible signals. This provides a significant reduction against the standard CS requirement  $M = \mathcal{O}(K \log(N/K))$  as the signal length  $N$  increases. Figure 1 indicates the potential performance gains on a tree-compressible, piecewise smooth signal.

The second structured sparsity model accounts for the fact that the large coefficients of many *sparse signals cluster together* [8,9]. Such a *so-called block sparse* model is equivalent to a *joint sparsity* model for an ensemble of  $J$ , length- $N$  signals [10], where the supports of the signals' large coefficients are shared across the ensemble. Using the fact that the number of clustered supports is much smaller than  $\binom{JN}{JK}$ , we prove that a block-based CoSaMP algorithm needs only  $M = \mathcal{O}(JK + K \log(\frac{N}{K}))$  measurements to robustly recover block-sparse and block-compressible signals. In contrast, standard CS requires  $M = \mathcal{O}(JK \log(N/K))$ ; block sparsity reduces the dependence of  $M$  on the signal length  $N$ , particularly for large block sizes  $J$ .

Our new theory and methods relate to a small body of previous work aimed at integrating structured sparsity into CS. Several groups have developed structured sparse signal recovery algorithms [6–8, 18–24]; however, their approaches have either been ad hoc or focused on a single structured sparsity model. Most previous work on unions of subspaces [6, 7, 24] has focused exclusively on strictly sparse signals and has considered neither compressibility nor feasible recovery algorithms. A related CS modeling framework for structured sparse and compressible signals [9] collects the  $N$  samples of a signal into  $D$  groups,  $D \leq N$ , and allows signals where  $K$  out of  $D$  groups have nonzero coefficients. This framework is immediately applicable to block-sparse signals and signal ensembles with common sparse supports. While [9] provides recovery algorithms, measurement bounds, and recovery guarantees similar to those provided in Section VI, our proposed framework has the ability to focus on arbitrary

<sup>1</sup>Obviously, sparsity and compressibility correspond to simple signal models where each coefficient is treated independently; for example in a sparse model, the fact that the coefficient  $\alpha_i$  is large has no bearing on the size of any  $\alpha_j$ ,  $j \neq i$ . We will reserve the use of the term “model” for situations where we are enforcing structured dependencies between the values and the locations of the coefficients  $\alpha_i$ .

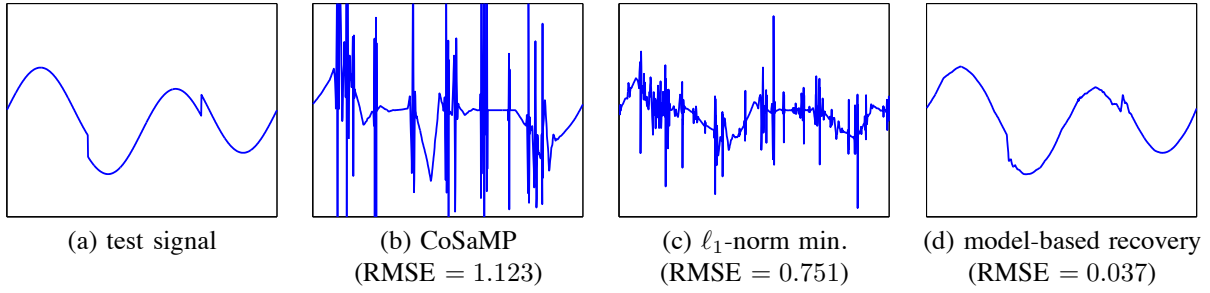


Fig. 1. Example performance of structured signal recovery. (a) Piecewise smooth HeaviSine test signal of length  $N = 1024$ . This signal is compressible under a connected wavelet tree model. Signal recovered from  $M = 80$  random Gaussian measurements using (b) the iterative recovery algorithm CoSaMP, (c) standard  $\ell_1$ -norm minimization via linear programming, and (d) the wavelet tree-based CoSaMP algorithm from Section V. In all figures, root mean-squared error (RMSE) values are normalized with respect to the  $\ell_2$  norm of the signal.

subsets of the  $\binom{D}{K}$  groups that yield more elaborate structures, such as connected subtrees for wavelet coefficients. To the best of our knowledge, our general algorithmic framework for model-based recovery, the concept of a model-compressible signal, and the associated RAMP are new to the literature.

This paper is organized as follows. A review of the CS theory in Section II lays out the foundational concepts that we extend to the model-based case in subsequent sections. Section III develops the concept of structured sparse signals and introduces the concept of structured compressible signals. We also quantify how structured sparsity models improve the measurement and recovery process by exploiting the model-based RIP for structured sparse signals and by introducing the RAMP for structured compressible signals. Section IV indicates how to tune CoSaMP to incorporate structured sparsity models and establishes its robustness properties for structured sparse and structured compressible signals; the modifications to the IHT algorithm are very similar, so we defer them to an appendix to reduce redundancy. Sections V and VI then specialize our theory to the special cases of wavelet tree and block sparse signal models, respectively, and report on a series of numerical experiments that validate our theoretical claims. We conclude with a discussion in Section VII. To make the paper more readable, all proofs are relegated to a series of appendices.

## II. BACKGROUND ON COMPRESSIVE SENSING

### A. Sparse and compressible signals

Given a basis  $\{\psi_i\}_{i=1}^N$ , we can represent every signal  $x \in \mathbb{R}^N$  in terms of  $N$  coefficients  $\{\alpha_i\}_{i=1}^N$  as  $x = \sum_{i=1}^N \alpha_i \psi_i$ ; stacking the  $\psi_i$  as columns into the  $N \times N$  matrix  $\Psi$ , we can write succinctly that  $x = \Psi\alpha$ . In the sequel, we will assume without loss of generality that the signal  $x$  is sparse or compressible in the canonical domain so that the sparsity basis  $\Psi$  is the identity and  $\alpha = x$ .

A signal  $x$  is  $K$ -sparse if only  $K \ll N$  entries of  $x$  are nonzero. We call the set of indices corresponding to the nonzero entries the *support* of  $x$  and denote it by  $\text{supp}(x)$ . The set of all  $K$ -sparse signals is the union of the  $\binom{N}{K}$ ,  $K$ -dimensional subspaces aligned with the coordinate axes in  $\mathbb{R}^N$ . We denote this union of subspaces by  $\Sigma_K$ .

Many natural and manmade signals are not strictly sparse, but can be approximated as such; we call such signals *compressible*. Consider a signal  $x$  whose coefficients, when sorted in order of decreasing magnitude, decay according to the power law

$$|x_{\mathcal{I}(i)}| \leq G i^{-1/r}, \quad i = 1, \dots, N, \quad (1)$$

where  $\mathcal{I}$  indexes the sorted coefficients. Thanks to the rapid decay of their coefficients, such signals are well-approximated by  $K$ -sparse signals. Let  $x_K \in \Sigma_K$  represent the best  $K$ -term approximation of  $x$ , which is obtained by keeping just the first  $K$  terms in  $x_{\mathcal{I}(i)}$  from (1). Denote the error of this approximation in the  $\ell_p$  norm as

$$\sigma_K(x)_p := \min_{\bar{x} \in \Sigma_K} \|x - \bar{x}\|_p = \|x - x_K\|_p, \quad (2)$$

where the  $\ell_p$  norm of the vector  $x$  is defined as  $\|x\|_p = \left(\sum_{i=1}^N |x_i|^p\right)^{1/p}$  for  $0 < p < \infty$ . Then, for  $r < p$ , we have that

$$\sigma_K(x)_p \leq (rs)^{-1/p} G K^{-s}, \quad (3)$$

with  $s = \frac{1}{r} - \frac{1}{p}$ . That is, when measured in the  $\ell_p$  norm, the signal's best approximation error has a power-law decay with exponent  $s$  as  $K$  increases. In the sequel we let  $p = 2$ , yielding  $s = 1/r - 1/2$ , and we dub a signal that obeys (3) an  $s$ -compressible signal.

The approximation of compressible signals by sparse signals is the basis of *transform coding* as is used in algorithms like JPEG and JPEG2000 [1]. In this framework, we acquire the full  $N$ -sample signal  $x$ ; compute the complete set of transform coefficients  $\alpha$  via  $\alpha = \Psi^{-1}x$ ; locate the  $K$  largest coefficients and discard the  $(N - K)$  smallest coefficients; and encode the  $K$  values and locations of the largest coefficients. While a widely accepted standard, this sample-then-compress framework suffers from three inherent inefficiencies. First, we must start with a potentially large number of samples  $N$  even if the ultimate desired  $K$  is small. Second, the encoder must compute all of the  $N$  transform coefficients  $\alpha$ , even though it will discard all but  $K$  of them. Third, the encoder faces the overhead of encoding the locations of the large coefficients.

### B. Compressive measurements and the restricted isometry property

Compressive sensing (CS) integrates the signal acquisition and compression steps into a single process [2–4]. In CS we do not acquire  $x$  directly but rather acquire  $M < N$  linear measurements  $y = \Phi x$  using an  $M \times N$  measurement matrix  $\Phi$ . We then recover  $x$  by exploiting its sparsity or compressibility. Our goal is to push  $M$  as close as possible to  $K$  in order to perform as much signal “compression” during acquisition as possible.

In order to recover a good estimate of  $x$  (the  $K$  largest  $x_i$ ’s, for example) from the  $M$  compressive measurements, the measurement matrix  $\Phi$  should satisfy the *restricted isometry property* (RIP) [3].

*Definition 1:* An  $M \times N$  matrix  $\Phi$  has the  $K$ -restricted isometry property ( $K$ -RIP) with constant  $\delta_K$  if, for all  $x \in \Sigma_K$ ,

$$(1 - \delta_K)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K)\|x\|_2^2. \quad (4)$$

In words, the  $K$ -RIP ensures that all submatrices of  $\Phi$  of size  $M \times K$  are close to an isometry, and therefore distance (and information) preserving. Practical recovery algorithms typically require that  $\Phi$  have a slightly stronger  $2K$ -RIP,  $3K$ -RIP, or higher-order RIP in order to preserve distances between  $K$ -sparse vectors (which are  $2K$ -sparse in general), three-way sums of  $K$ -sparse vectors (which are  $3K$ -sparse in general), and other higher-order structures.

While checking whether a measurement matrix  $\Phi$  satisfies the  $K$ -RIP is an NP-Complete problem in general [26], random matrices whose entries are independent and identically distributed (i.i.d.) Gaussian, Rademacher ( $\pm 1$ ), or more generally subgaussian<sup>2</sup> work with high probability provided  $M = \mathcal{O}(K \log(N/K))$ . These random matrices also have a so-called *universality* property in that, for any choice of orthonormal basis matrix  $\Psi$ ,  $\Phi\Psi$  has the  $K$ -RIP with high probability. This is useful when the signal is sparse not in the canonical domain but in basis  $\Psi$ . A random  $\Phi$  corresponds to an intriguing data acquisition protocol in which each measurement  $y_j$  is a randomly weighted linear combination of the entries of  $x$ .

### C. Recovery algorithms

Since there are infinitely many signal coefficient vectors  $x'$  that produce the same set of compressive measurements  $y = \Phi x$ , to recover the “right” signal we exploit our a priori knowledge of its sparsity or compressibility. For example, we could seek the sparsest  $x$  that agrees with the measurements  $y$ :

$$\hat{x} = \arg \min_{x'} \|x'\|_0 \text{ s.t. } y = \Phi x', \quad (5)$$

<sup>2</sup>A random variable  $X$  is called subgaussian if there exists  $c > 0$  such that  $\mathbb{E}(e^{Xt}) \leq e^{c^2 t^2/2}$  for all  $t \in \mathbb{R}$ . Examples include the Gaussian, Bernoulli, and Rademacher random variables, as well as any bounded random variable. [25]

where the  $\ell_0$  “norm” of a vector counts its number of nonzero entries. While this optimization can recover a  $K$ -sparse signal from just  $M = 2K$  compressive measurements, it is unfortunately a combinatorial, NP-hard problem [26]; furthermore, the recovery is not stable in the presence of noise [4].

Practical, stable recovery algorithms rely on the RIP (and therefore require at least  $M = \mathcal{O}(K \log(N/K))$  measurements); they can be grouped into two camps. The first approach convexifies the  $\ell_0$  “norm” minimization (5) to the  $\ell_1$ -norm minimization

$$\hat{x} = \arg \min_{x'} \|x'\|_1 \text{ s.t. } y = \Phi x'. \quad (6)$$

This corresponds to a linear program that can be solved in polynomial time [2, 3]. Adaptations to deal with additive noise in  $y$  or  $x$  include basis pursuit with denoising (BPDN) [27], complexity-based regularization [28], and the Dantzig Selector [29].

The second approach finds the sparsest  $x$  agreeing with the measurements  $y$  through an iterative, greedy search. Algorithms such as matching pursuit, orthogonal matching pursuit [30], StOMP [31], iterative hard thresholding (IHT) [12–16], CoSaMP [11], and Subspace Pursuit (SP) [32] all revolve around a best  $L$ -term approximation for the estimated signal, with  $L$  varying for each algorithm; typically  $L$  is  $\mathcal{O}(K)$ .

### D. Performance bounds on signal recovery

Given  $M = \mathcal{O}(K \log(N/K))$  compressive measurements, a number of different CS signal recovery algorithms, including all of the  $\ell_1$ -norm minimization techniques mentioned above and the CoSaMP, SP, and IHT iterative techniques, offer provably stable signal recovery with performance close to optimal  $K$ -term approximation (recall (3)) [2, 3, 11, 16]. For a random  $\Phi$ , all results hold with high probability.

For a noise-free,  $K$ -sparse signal, these algorithms offer perfect recovery, meaning that the signal  $\hat{x}$  recovered from the compressive measurements  $y = \Phi x$  is exactly  $\hat{x} = x$ .

For a  $K$ -sparse signal  $x$  whose measurements are corrupted by noise  $n$  of bounded norm (that is, we measure  $y = \Phi x + n$ ) the mean-squared error of the signal  $\hat{x}$  is

$$\|x - \hat{x}\|_2 \leq C\|n\|_2, \quad (7)$$

with  $C$  a small constant.

For an  $s$ -compressible signal  $x$  whose measurements are corrupted by noise  $n$  of bounded norm, the mean-squared error of the recovered signal  $\hat{x}$  is

$$\|x - \hat{x}\|_2 \leq C_1 \|x - x_K\|_2 + C_2 \frac{1}{\sqrt{K}} \|x - x_K\|_1 + C_3 \|n\|_2. \quad (8)$$

Using (3) we can simplify this expression to

$$\|x - \hat{x}\|_2 \leq \frac{C_1 GK^{-s}}{\sqrt{2s}} + \frac{C_2 GK^{-s}}{s - 1/2} + C_3 \|n\|_2. \quad (9)$$

For the recovery algorithm (6), we obtain a bound very similar to (8), albeit with the  $\ell_2$ -norm error component removed [33].

### III. STRUCTURED SPARSITY AND COMPRESSIBILITY

While many natural and manmade signals and images can be described to first-order as sparse or compressible, the support of their large coefficients often has an underlying interdependency structure. This phenomenon has received only limited attention by the CS community to date [6–9, 19–23]. In this section, we introduce a model-based theory of CS that captures such structure. A model reduces the degrees of freedom of a sparse/compressible signal by permitting only certain configurations of supports for the large coefficient. As we will show, this allows us to reduce, in some cases significantly, the number of compressive measurements  $M$  required to stably recover a signal.

#### A. Structured sparse signals

Recall from Section II-A that a  $K$ -sparse signal vector  $x$  lives in  $\Sigma_K \subset \mathbb{R}^N$ , which is a union of  $\binom{N}{K}$  subspaces of dimension  $K$ . Other than its  $K$ -sparsity, there are no further constraints on the support or values of its coefficients. A *structured sparsity model* endows the  $K$ -sparse signal  $x$  with additional structure that allows certain  $K$ -dimensional subspaces in  $\Sigma_K$  and disallows others [6, 7].

To state a formal definition of a structured sparsity model, let  $x|_\Omega$  represent the entries of  $x$  corresponding to the set of indices  $\Omega \subseteq \{1, \dots, N\}$ , and let  $\Omega^C$  denote the complement of the set  $\Omega$ .

**Definition 2:** A *structured sparsity model*  $\mathcal{M}_K$  is defined as the union of  $m_K$  canonical  $K$ -dimensional subspaces

$$\mathcal{M}_K = \bigcup_{m=1}^{m_K} \mathcal{X}_m, \text{ s.t. } \mathcal{X}_m := \{x : x|_{\Omega_m} \in \mathbb{R}^K, x|_{\Omega_m^C} = 0\},$$

where  $\{\Omega_1, \dots, \Omega_{m_K}\}$  is the set containing all allowed supports, with  $|\Omega_m| = K$  for each  $m = 1, \dots, m_K$ , and each subspace  $\mathcal{X}_m$  contains all signals  $x$  with  $\text{supp}(x) \subseteq \Omega_m$ .

Signals from  $\mathcal{M}_K$  are called  *$K$ -structured sparse*. Clearly,  $\mathcal{M}_K \subseteq \Sigma_K$  and contains  $m_K \leq \binom{N}{K}$  subspaces.

In Sections V and VI below we consider two concrete structured sparsity models. The first model accounts for the fact that the large wavelet coefficients of piecewise smooth signals and images tend to live on a rooted, connected *tree structure* [17]. The second model accounts for the fact that the large coefficients of sparse signals often *cluster together into blocks* [8–10].

#### B. Model-based RIP

If we know that the signal  $x$  being acquired is  $K$ -structured sparse, then we can relax the RIP constraint on the CS measurement matrix  $\Phi$  and still achieve stable recovery from the compressive measurements  $y = \Phi x$  [6, 7].

**Definition 3:** [6, 7] An  $M \times N$  matrix  $\Phi$  has the  $\mathcal{M}_K$ -*restricted isometry property* ( $\mathcal{M}_K$ -RIP) with constant  $\delta_{\mathcal{M}_K}$  if,

for all  $x \in \mathcal{M}_K$ , we have

$$(1 - \delta_{\mathcal{M}_K})\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_{\mathcal{M}_K})\|x\|_2^2. \quad (10)$$

Blumensath and Davies [6] have quantified the number of measurements  $M$  necessary for a random CS matrix to have the  $\mathcal{M}_K$ -RIP with a given probability.

**Theorem 1:** [6] Let  $\mathcal{M}_K$  be the union of  $m_K$  subspaces of  $K$ -dimensions in  $\mathbb{R}^N$ . Then, for any  $t > 0$  and any

$$M \geq \frac{2}{c\delta_{\mathcal{M}_K}^2} \left( \ln(2m_K) + K \ln \frac{12}{\delta_{\mathcal{M}_K}} + t \right),$$

where  $c$  is a positive constant, an  $M \times N$  i.i.d. subgaussian random matrix has the  $\mathcal{M}_K$ -RIP with constant  $\delta_{\mathcal{M}_K}$  with probability at least  $1 - e^{-t}$ .

This bound can be used to recover the conventional CS result by substituting  $m_K = \binom{N}{K} \approx (Ne/K)^K$ . Similarly, as the number of subspaces  $m_K$  that arise from the structure imposed can be significantly smaller than the standard  $\binom{N}{K}$ , the number of rows needed for a random matrix to have the  $\mathcal{M}_K$ -RIP can be significantly lower than the number of rows needed for the standard RIP. The  $\mathcal{M}_K$ -RIP property is sufficient for robust recovery of structured sparse signals, as we show below in Section IV-B.

#### C. Structured compressible signals

Just as compressible signals are “nearly  $K$ -sparse” and thus live close to the union of subspaces  $\Sigma_K$  in  $\mathbb{R}^N$ , structured compressible signals are “nearly  $K$ -structured sparse” and live close to the restricted union of subspaces  $\mathcal{M}_K$ . In this section, we make this new concept rigorous. Recall from (3) that we defined compressible signals in terms of the decay of their  $K$ -term approximation error.

The  $\ell_2$  error incurred by approximating  $x \in \mathbb{R}^N$  by the best structured sparse approximation in  $\mathcal{M}_K$  is given by

$$\sigma_{\mathcal{M}_K}(x) := \inf_{\tilde{x} \in \mathcal{M}_K} \|x - \tilde{x}\|_2.$$

We define  $\mathbb{M}_B(x, K)$  as the algorithm that obtains the best  $K$ -term structured sparse approximation of  $x$  in the union of subspaces  $\mathcal{M}_K$ :

$$\mathbb{M}(x, K) = \arg \min_{\tilde{x} \in \mathcal{M}_K} \|x - \tilde{x}\|_2.$$

This implies that  $\|x - \mathbb{M}(x, K)\|_2 = \sigma_{\mathcal{M}_K}(x)$ . The decay of this approximation error defines the structured compressibility of a signal.

**Definition 4:** The set of *s-structured compressible signals* is defined as

$$\mathfrak{M}_s = \{x \in \mathbb{R}^N : \sigma_{\mathcal{M}_K}(x) \leq GK^{-s}, 1 \leq K \leq N, G < \infty\}.$$

Define  $|x|_{\mathfrak{M}_s}$  as the smallest value of  $G$  for which this condition holds for  $x$  and  $s$ .

We say that  $x \in \mathfrak{M}_s$  is an *s-structured compressible signal* under the structured sparsity model  $\mathcal{M}_K$ . These approximation classes have been characterized for certain structured



sparsity models; see Section V for an example. We will select the value of  $s$  for which the distance between the approximation errors  $\sigma_{\mathcal{M}_K}(x)$  and the corresponding bounds  $GK^{-1/s}$  is minimal.

#### D. Nested model approximations and residual subspaces

In conventional CS, the same requirement (RIP) is a sufficient condition for the stable recovery of both sparse and compressible signals. In model-based recovery, however, the class of structured compressible signals is much larger than that of structured sparse signals, since the union of subspaces defined by structured sparse signals does not contain all canonical  $K$ -dimensional subspaces.

To address this difference, we introduce some additional tools to develop a *sufficient* condition for the stable recovery of structured compressible signals. We will pay particular attention to structured sparsity models  $\mathcal{M}_K$  that generate *nested approximations*, since they are more amenable to analysis and computation.

**Definition 5:** A structured sparsity model  $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \dots\}$  has the *nested approximation property* (NAP) if  $\text{supp}(\mathbb{M}(x, K)) \subset \text{supp}(\mathbb{M}(x, K'))$  for all  $K < K'$  and for all  $x \in \mathbb{R}^N$ .

In words, a structured sparsity model generates nested approximations if the support of the best  $K'$ -term structured sparse approximation contains the support of the best  $K$ -term structured sparse approximation for all  $K < K'$ . An important example of a NAP-generating structured sparse model is the standard compressible signal model of (3).

When a structured sparsity model obeys the NAP, the support of the difference between the best  $jK$ -term structured sparse approximation and the best  $(j+1)K$ -term structured sparse approximation of a signal can be shown to lie in a small union of subspaces, thanks to the structure enforced by the model. This structure is captured by the set of subspaces that are included in each subsequent approximation, as defined below.

**Definition 6:** The  $j^{\text{th}}$  set of *residual subspaces* of size  $K$  is defined as  $\mathcal{R}_{j,K}(\mathcal{M}) = \{u \in \mathbb{R}^N : u = \mathbb{M}(x, jK) - \mathbb{M}(x, (j-1)K) \text{ for some } x \in \mathbb{R}^N\}$ , for  $j = 1, \dots, \lceil N/K \rceil$ .

Under the NAP, each structured compressible signal  $x$  can be partitioned into its best  $K$ -term structured sparse approximation  $x_{T_1}$ , the additional components present in the best  $2K$ -term structured sparse approximation  $x_{T_2}$ , and so on, with  $x = \sum_{j=1}^{\lceil N/K \rceil} x_{T_j}$  and  $x_{T_j} \in \mathcal{R}_{j,K}(\mathcal{M})$  for each  $j$ . Each signal partition  $x_{T_j}$  is a  $K$ -sparse signal, and thus  $\mathcal{R}_{j,K}(\mathcal{M})$  is a union of subspaces of dimension  $K$ . We will denote by  $R_j$  the number of subspaces that compose  $\mathcal{R}_{j,K}(\mathcal{M})$  and omit the dependence on  $\mathcal{M}$  in the sequel for brevity.

Intuitively, the norms of the partitions  $\|x_{T_j}\|_2$  decay as  $j$  increases for signals that are structured compressible. As the next subsection shows, this observation is instrumental in relaxing the isometry restrictions on the measurement matrix  $\Phi$

and bounding the recovery error for  $s$ -structured compressible signals when the model obeys the NAP.

#### E. The restricted amplification property (RAmP)

For exactly  $K$ -structured sparse signals, we discussed in Section III-B that the number of compressive measurements  $M$  required for a random matrix to have the  $\mathcal{M}_K$ -RIP is determined by the number of canonical subspaces  $m_K$  via (11). Unfortunately, such structured sparse concepts and results do not immediately extend to structured compressible signals. Thus, we develop a generalization of the  $\mathcal{M}_K$ -RIP that we will use to quantify the stability of recovery for structured compressible signals.

One way to analyze the robustness of compressible signal recovery in conventional CS is to consider the tail of the signal outside its  $K$ -term approximation as contributing additional “noise” to the measurements of size  $\|\Phi(x - x_K)\|_2$  [11, 16, 33]. Consequently, the conventional  $K$ -sparse recovery performance result can be applied with the augmented noise  $n + \Phi(x - x_K)$ .

This technique can also be used to quantify the robustness of structured compressible signal recovery. The key quantity we must control is the amplification of the structured sparse approximation residual through  $\Phi$ . The following property is a new generalization of the RIP and model-based RIP.

**Definition 7:** A matrix  $\Phi$  has the  $(\epsilon_K, r)$ -*restricted amplification property* (RAmP) for the residual subspaces  $\mathcal{R}_{j,K}$  of model  $\mathcal{M}$  if

$$\|\Phi u\|_2^2 \leq (1 + \epsilon_K) j^{2r} \|u\|_2^2 \quad (11)$$

for any  $u \in \mathcal{R}_{j,K}$  for each  $1 \leq j \leq \lceil N/K \rceil$ .

The regularity parameter  $r > 0$  caps the growth rate of the amplification of  $u \in \mathcal{R}_{j,K}$  as a function of  $j$ . Its value can be chosen so that the growth in amplification with  $j$  balances the decay of the norm in each residual subspace  $\mathcal{R}_{j,K}$  with  $j$ .

We can quantify the number of compressive measurements  $M$  required for a random measurement matrix  $\Phi$  to have the RAmP with high probability; we prove the following in Appendix A.

**Theorem 2:** Let  $\Phi$  be an  $M \times N$  matrix with i.i.d. subgaussian entries and let the set of residual subspaces  $\mathcal{R}_{j,K}$  of the structured sparsity model  $\mathcal{M}$  contain  $R_j$  subspaces of dimension  $K$  for each  $1 \leq j \leq \lceil N/K \rceil$ . If

$$M \geq \max_{1 \leq j \leq \lceil N/K \rceil} \frac{2K + 4 \ln \frac{R_j N}{K} + 2t}{(j^r \sqrt{1 + \epsilon_K} - 1)^2}, \quad (12)$$

then the matrix  $\Phi$  has the  $(\epsilon_K, r)$ -RAmP with probability  $1 - e^{-t}$ .

The order of the bound of Theorem 2 is lower than  $\mathcal{O}(K \log(N/K))$  as long as the number of subspaces  $R_j$  grows slower than  $N^K$ .

Armed with the RaMP, we can state the following result, which will provide robustness for the recovery of structured compressible signals; see Appendix B for the proof.

*Theorem 3:* Let  $x \in \mathfrak{M}_s$  be an  $s$ -structured compressible signal under a structured sparsity model  $\mathcal{M}$  that obeys the NAP. If  $\Phi$  has the  $(\epsilon_K, r)$ -RaMP and  $r = s - 1$ , then we have

$$\|\Phi(x - \mathbb{M}(x, K))\|_2 \leq C_s \sqrt{1 + \epsilon_K} K^{-s} \ln \left\lceil \frac{N}{K} \right\rceil \|x\|_{\mathfrak{M}_s},$$

where  $C_s$  is a constant that depends only on  $s$ .

#### IV. MODEL-BASED SIGNAL RECOVERY ALGORITHMS

To take practical advantage of our new theory for model-based CS, we demonstrate how to integrate structured sparsity models into two state-of-the-art CS recovery algorithms, CoSaMP [11] (in this section) and iterative hard thresholding (IHT) [12–16] (in Appendix C to avoid repetition). The key modification is simple: we merely replace the best  $K$ -term sparse approximation step in these greedy algorithms with a best  $K$ -term structured sparse approximation. Since at each iteration we need only search over the  $m_K$  subspaces of  $\mathcal{M}_K$  rather than  $\binom{N}{K}$  subspaces of  $\Sigma_K$ , fewer measurements will be required for the same degree of robust signal recovery. Or, alternatively, using the same number of measurements, more accurate recovery can be achieved.

After presenting the modified CoSaMP algorithm, we prove robustness guarantees for both structured sparse and structured compressible signals. To this end, we must define an enlarged union of subspaces that includes sums of elements in the structured sparsity model.

*Definition 8:* The  $B$ -order sum for the set  $\mathcal{M}_K$ , with  $B > 1$  an integer, is defined as

$$\mathcal{M}_K^B = \left\{ x = \sum_{r=1}^B x^{(r)}, \text{ with } x^{(r)} \in \mathcal{M}_K \right\}.$$

Define  $\mathbb{M}_B(x, K)$  as the algorithm that obtains the best approximation of  $x$  in the enlarged union of subspaces  $\mathcal{M}_K^B$ :

$$\mathbb{M}_B(x, K) = \arg \min_{\tilde{x} \in \mathcal{M}_K^B} \|x - \tilde{x}\|_2.$$

We note that  $\mathbb{M}(x, K) = \mathbb{M}_1(x, K)$ . Note also that for many structured sparsity models, we will have  $\mathcal{M}_K^B \subset \mathcal{M}_{BK}$ , and so the algorithm  $\mathbb{M}(x, BK)$  will provide a strictly better approximation than  $\mathbb{M}_B(x, K)$ .

##### A. Model-based CoSaMP

We choose to modify the CoSaMP algorithm [11] for two reasons. First, it has robust recovery guarantees that are on par with the best convex optimization-based approaches. Second, it has a simple iterative, greedy structure based on a best  $BK$ -term approximation (with  $B$  a small integer) that is easily modified to incorporate a best  $BK$ -term structured sparse approximation  $\mathbb{M}_B(K, x)$ . These properties also make

the IHT and SP algorithms amenable to modification; see Appendix C for details on IHT. Pseudocode for the modified CoSaMP algorithm is given in Algorithm 1, where  $A^\dagger$  denotes the Moore-Penrose pseudoinverse of  $A$ .

##### B. Performance of structured sparse signal recovery

We now study the performance of model-based CoSaMP signal recovery on structured sparse and structured compressible signals. A robustness guarantee for noisy measurements of structured sparse signals can be obtained using the model-based RIP (10). Our performance guarantee for structured sparse signal recovery will require that the measurement matrix  $\Phi$  be a near-isometry for all subspaces in  $\mathcal{M}_K^B$  for some  $B > 1$ . This requirement is a direct generalization of the  $2K$ -RIP,  $3K$ -RIP, and higher-order RIPs from the conventional CS theory. The following theorem is proven in Appendix D.

*Theorem 4:* Let  $x \in \mathcal{M}_K$  and let  $y = \Phi x + n$  be a set of noisy CS measurements. If  $\Phi$  has an  $\mathcal{M}_K^4$ -RIP constant of  $\delta_{\mathcal{M}_K^4} \leq 0.1$ , then the signal estimate  $\hat{x}_i$  obtained from iteration  $i$  of the model-based CoSaMP algorithm satisfies

$$\|x - \hat{x}_i\|_2 \leq 2^{-i} \|x\|_2 + 15 \|n\|_2. \quad (13)$$

This guarantee matches that of the CoSaMP algorithm [11, Theorem 4.1]; however, our guarantee is only for structured sparse signals rather than for all sparse signals.

##### C. Performance of structured compressible signal recovery

Using the new tools introduced in Section III, we can provide a robustness guarantee for noisy measurements of structured compressible signals, using the RaMP as a condition on the measurement matrix  $\Phi$ .

*Theorem 5:* Let  $x \in \mathfrak{M}_s$  be an  $s$ -structured compressible signal from a structured sparsity model  $\mathcal{M}$  that obeys the NAP, and let  $y = \Phi x + n$  be a set of noisy CS measurements. If  $\Phi$  has the  $\mathcal{M}_K^4$ -RIP with  $\delta_{\mathcal{M}_K^4} \leq 0.1$  and the  $(\epsilon_K, r)$ -RaMP with  $\epsilon_K \leq 0.1$  and  $r = s - 1$ , then the signal estimate  $\hat{x}_i$  obtained from iteration  $i$  of the model-based CoSaMP algorithm satisfies

$$\begin{aligned} \|x - \hat{x}_i\|_2 &\leq 2^{-i} \|x\|_2 + 35 \|n\|_2 \\ &\quad + 35 C_s \|x\|_{\mathfrak{M}_s} K^{-s} (1 + \ln \lceil N/K \rceil). \end{aligned} \quad (14)$$

*Proof sketch.* To prove the theorem, we first bound the optimal structured sparse recovery error for an  $s$ -structured compressible signal  $x \in \mathfrak{M}_s$  when the matrix  $\Phi$  has the  $(\epsilon_K, r)$ -RaMP with  $r \leq s - 1$  (see Theorem 3). Then, using Theorem 4, we can easily prove the result by following the analogous proof in [11].  $\square$

The standard CoSaMP algorithm also features a similar guarantee for structured compressible signals, with the constant changing from 35 to 20.

**Algorithm 1** Model-based CoSaMP

---

Inputs: CS matrix  $\Phi$ , measurements  $y$ , structured sparse approximation algorithm  $\mathbb{M}$ 
Output:  $K$ -sparse approximation  $\hat{x}$  to true signal  $x$  $\hat{x}_0 = 0$ ,  $d = y$ ;  $i = 0$  {initialize}**while** halting criterion false **do**1.  $i \leftarrow i + 1$ 2.  $e \leftarrow \Phi^T d$  {form signal residual estimate}3.  $\Omega \leftarrow \text{supp}(\mathbb{M}_2(e, K))$  {prune residual estimate according to structure}4.  $T \leftarrow \Omega \cup \text{supp}(\hat{x}_{i-1})$  {merge supports}5.  $b|_T \leftarrow \Phi_T^\dagger y$ ,  $b|_{T^c}$  {form signal estimate}6.  $\hat{x}_i \leftarrow \mathbb{M}(b, K)$  {prune signal estimate according to structure}7.  $d \leftarrow y - \Phi \hat{x}_i$  {update measurement residual}**end while**return  $\hat{x} \leftarrow \hat{x}_i$ *D. Robustness to model mismatch*

We now analyze the robustness of model-based CS recovery to *model mismatch*, which occurs when the signal being recovered from compressive measurements does not conform exactly to the structured sparsity model used in the recovery algorithm.

We begin with optimistic results for signals that are “close” to matching the recovery structured sparsity model. First consider a signal  $x$  that is not  $K$ -structured sparse as the recovery algorithm assumes but rather  $(K + \kappa)$ -structured sparse for some small integer  $\kappa$ . This signal can be decomposed into  $x_K$ , the signal’s  $K$ -term structured sparse approximation, and  $x - x_K$ , the error of this approximation. For  $\kappa \leq K$ , we have that  $x - x_K \in \mathcal{R}_{2,K}$ . If the matrix  $\Phi$  has the  $(\epsilon_K, r)$ -RAmP, then it follows that

$$\|\Phi(x - x_K)\|_2 \leq 2^r \sqrt{1 + \epsilon_K} \|x - x_K\|_2. \quad (15)$$

Using equations (13) and (15), we obtain the following guarantee for the  $i^{\text{th}}$  iteration of model-based CoSaMP:

$$\|x - \hat{x}_i\|_2 \leq 2^{-i} \|x\|_2 + 16 \cdot 2^r \sqrt{1 + \epsilon_K} \|x - x_K\|_2 + 15 \|n\|_2.$$

By noting that  $\|x - x_K\|_2$  is small, we obtain a guarantee that is close to (13).

Second, consider a signal  $x$  that is not  $s$ -structured compressible as the recovery algorithm assumes but rather  $(s - \epsilon)$ -structured compressible. The following bound can be obtained under the conditions of Theorem 5 by modifying the argument in Appendix B:

$$\begin{aligned} \|x - \hat{x}_i\|_2 \leq & 2^{-i} \|x\|_2 + 35 \left( \|n\|_2 \right. \\ & \left. + C_s |x|_{\mathfrak{M}_s} K^{-s} \left( 1 + \frac{\lceil \frac{N}{K} \rceil^\epsilon - 1}{\epsilon} \right) \right). \end{aligned}$$

As  $\epsilon$  becomes smaller, the factor  $\frac{\lceil N/K \rceil^\epsilon - 1}{\epsilon}$  approaches  $\log \lceil N/K \rceil$ , matching (14). In summary, as long as the deviations from the structured sparse and structured compressible classes are small, our model-based recovery guarantees still apply within a small bounded constant factor.

We end with an intuitive worst-case result for signals that are arbitrarily far away from structured sparse or structured compressible. Consider such an arbitrary  $x \in \mathbb{R}^N$  and compute its nested structured sparse approximations  $x_{jK} = \mathbb{M}(x, jK)$ ,  $j = 1, \dots, \lceil N/K \rceil$ . If  $x$  is not structured compressible, then the structured sparse approximation error  $\sigma_{jK}(x)$  is not guaranteed to decay as  $j$  decreases. Additionally, the number of residual subspaces  $\mathcal{R}_{j,K}$  could be as large as  $\binom{N}{K}$ ; that is, the  $j^{\text{th}}$  difference between subsequent structured sparse approximations  $x_{T_j} = x_{jK} - x_{(j-1)K}$  might lie in any arbitrary  $K$ -dimensional subspace. This worst case is equivalent to setting  $r = 0$  and  $R_j = \binom{N}{K}$  in Theorem 2. It is easy to see that the resulting condition on the number of measurements  $M$  matches that of the standard RIP for CS. Hence, if we inflate the number of measurements to  $M = \mathcal{O}(K \log(N/K))$  (the usual number for conventional CS), the performance of model-based CoSaMP recovery on an arbitrary signal  $x$  follows the distortion of the best  $K$ -term *structured sparse approximation* error of  $x$  within a bounded constant factor.

*E. Computational complexity of model-based recovery*

The computational complexity of a structured signal recovery algorithm differs from that of a standard algorithm by two factors. The first factor is the reduction in the number of measurements  $M$  necessary for recovery: since most current recovery algorithms have a computational complexity that is linear in the number of measurements, any reduction in  $M$  reduces the total complexity. The second factor is the cost of the structured sparse approximation. The  $K$ -term approximation used in most current recovery algorithms can be implemented with a simple sorting operation ( $\mathcal{O}(N \log N)$  complexity, in general). Ideally, the structured sparsity model should support a similarly efficient approximation algorithm.

To validate our theory and algorithms and demonstrate their general applicability and utility, we now present two specific instances of model-based CS and conduct a range of simulation experiments.



## V. EXAMPLE: WAVELET TREE MODEL

Wavelet decompositions have found wide application in the analysis, processing, and compression of smooth and piecewise smooth signals because these signals are  $K$ -sparse and compressible, respectively [1]. Moreover, the wavelet coefficients can be naturally organized into a tree structure, and for many kinds of natural and manmade signals the largest coefficients cluster along the branches of this tree. This motivates a connected tree model for the wavelet coefficients [34–36].

While CS recovery for wavelet-sparse signals has been considered previously [19–23], the resulting algorithms integrated the tree constraint in an ad-hoc fashion. Furthermore, the algorithms provide no recovery guarantees or bounds on the necessary number of compressive measurements.

### A. Tree-sparse signals

We first describe tree sparsity in the context of sparse wavelet decompositions. We focus on one-dimensional signals and binary wavelet trees, but all of our results extend directly to  $d$ -dimensional signals and  $2^d$ -ary wavelet trees.

Consider a signal  $x$  of length  $N = 2^I$ , for an integer value of  $I$ . The wavelet representation of  $x$  is given by

$$x = v_0\nu + \sum_{i=0}^{I-1} \sum_{j=0}^{2^i-1} w_{i,j} \psi_{i,j},$$

where  $\nu$  is the scaling function and  $\psi_{i,j}$  is the wavelet function at scale  $i$  and offset  $j$ . The wavelet transform consists of the scaling coefficient  $v_0$  and wavelet coefficients  $w_{i,j}$  at scale  $i$ ,  $0 \leq i \leq I-1$ , and position  $j$ ,  $0 \leq j \leq 2^i-1$ . In terms of our earlier matrix notation,  $x$  has the representation  $x = \Psi\alpha$ , where  $\Psi$  is a matrix containing the scaling and wavelet functions as columns, and  $\alpha = [v_0 \ w_{0,0} \ w_{1,0} \ w_{1,1} \ w_{2,0} \ \dots]^T$  is the vector of scaling and wavelet coefficients. We are, of course, interested in sparse and compressible  $\alpha$ .

The nested supports of the wavelets at different scales create a parent/child relationship between wavelet coefficients at different scales. We say that  $w_{i-1,\lfloor j/2 \rfloor}$  is the *parent* of  $w_{i,j}$  and that  $w_{i+1,2j}$  and  $w_{i+1,2j+1}$  are the *children* of  $w_{i,j}$ . These relationships can be expressed graphically by the wavelet coefficient tree in Figure 2.

Wavelet functions act as local discontinuity detectors, and using the nested support property of wavelets at different scales, it is straightforward to see that a signal discontinuity will give rise to a chain of large wavelet coefficients along a branch of the wavelet tree from a leaf to the root. Moreover, smooth signal regions will give rise to regions of small wavelet coefficients. This “connected tree” property has been well-exploited in a number of wavelet-based processing [17, 37, 38] and compression [39, 40] algorithms. In this section, we will specialize the theory developed in Sections III and IV to a connected tree model  $\mathcal{T}$ .

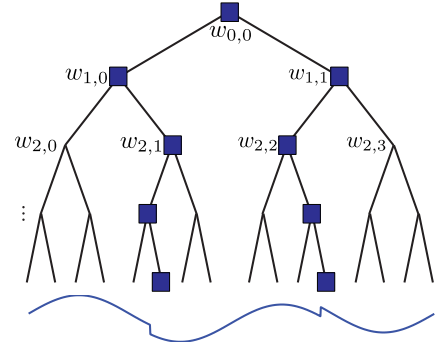


Fig. 2. Binary wavelet tree for a one-dimensional signal. The squares denote the large wavelet coefficients that arise from the discontinuities in the piecewise smooth signal drawn below; the support of the large coefficients forms a rooted, connected tree.

A set of wavelet coefficients  $\Omega$  forms a *connected subtree* if, whenever a coefficient  $w_{i,j} \in \Omega$ , then its parent  $w_{i-1,\lfloor j/2 \rfloor} \in \Omega$  as well. Each such set  $\Omega$  defines a subspace of signals whose support is contained in  $\Omega$ ; that is, all wavelet coefficients outside  $\Omega$  are zero. In this way, we define the structured sparsity model  $\mathcal{T}_K$  as the union of all  $K$ -dimensional subspaces corresponding to supports  $\Omega$  that form connected subtrees.

*Definition 9:* Define the set of  $K$ -tree sparse signals as

$$\mathcal{T}_K = \left\{ x = v_0\nu + \sum_{i=0}^{I-1} \sum_{j=1}^{2^i} w_{i,j} \psi_{i,j} : w|_{\Omega^c} = 0, \right. \\ \left. |\Omega| = K, \Omega \text{ forms a connected subtree} \right\}.$$

To quantify the number of subspaces in  $\mathcal{T}_K$ , it suffices to count the number of distinct connected subtrees of size  $K$  in a binary tree of size  $N$ . We prove the following result in Appendix E.

*Proposition 1:* The number of subspaces in  $\mathcal{T}_K$  obeys  $T_K \leq \frac{4^{K+4}}{Ke^2}$  for  $K \geq \log_2 N$  and  $T_K \leq \frac{(2e)^K}{K+1}$  for  $K < \log_2 N$ .

To simplify the presentation in the sequel, we will simply use the weaker bound  $T_K \leq \frac{(2e)^K}{K+1}$  for all values of  $K$  and  $N$ .

### B. Tree-based approximation

To implement tree-based signal recovery, we seek an efficient algorithm  $\mathbb{T}(x, K)$  to solve the optimal approximation

$$x_K^{\mathcal{T}} = \arg \min_{\bar{x} \in \mathcal{T}_K} \|x - \bar{x}\|_2. \quad (16)$$

Fortuitously, an efficient solver exists, called the *condensing sort and select algorithm* (CSSA) [34–36]. Recall that subtree approximation coincides with standard  $K$ -term approximation (and hence can be solved by simply sorting the wavelet coefficients) when the wavelet coefficients are monotonically nonincreasing along the tree branches out from the root. The CSSA solves (16) in the case of general wavelet coefficient

values by *condensing* the nonmonotonic segments of the tree branches using an iterative sort-and-average routine during a greedy search through the nodes. For each node in the tree, the algorithm calculates the average wavelet coefficient magnitude for each subtree rooted at that node, and records the largest average among all the subtrees as the energy for that node. The CSSA then searches for the unselected node with the largest energy and adds the subtree corresponding to the node's energy to the estimated support as a *supernode*: a single node that provides a condensed representation of the corresponding subtree [36]. Condensing a large coefficient far down the tree accounts for the potentially large cost (in terms of the total budget of tree nodes  $K$ ) of growing the tree to that point.

Since the first step of the CSSA involves sorting all of the wavelet coefficients, overall it requires  $\mathcal{O}(N \log N)$  computations. However, once the CSSA grows the optimal tree of size  $K$ , it is trivial to determine the optimal trees of size  $< K$  and computationally efficient to grow the optimal trees of size  $> K$  [34].

The constrained optimization (16) can be rewritten as an unconstrained problem by introducing the Lagrange multiplier  $\lambda$  [41]:

$$\min_{\bar{x} \in \bar{\mathcal{T}}} \|x - \bar{x}\|_2^2 + \lambda(\|\bar{\alpha}\|_0 - K),$$

where  $\bar{\mathcal{T}} = \cup_{n=1}^N \mathcal{T}_n$  and  $\bar{\alpha}$  are the wavelet coefficients of  $\bar{x}$ . Except for the inconsequential  $\lambda K$  term, this optimization coincides with Donoho's *complexity penalized sum of squares* [41], which can be solved in only  $\mathcal{O}(N)$  computations using coarse-to-fine dynamic programming on the tree. Its primary shortcoming is the nonobvious relationship between the tuning parameter  $\lambda$  and the resulting size  $K$  of the optimal connected subtree.

### C. Tree-compressible signals

Specializing Definition 2 from Section III-C to  $\mathcal{T}$ , we make the following definition.

**Definition 10:** Define the set of *s-tree compressible signals* as

$$\mathcal{T}_s = \{x \in \mathbb{R}^N : \|x - \mathbb{T}(x, K)\|_2 \leq GK^{-s}, \\ 1 \leq K \leq N, G < \infty\}.$$

Furthermore, define  $|x|_{\mathcal{T}_s}$  as the smallest value of  $G$  for which this condition holds for  $x$  and  $s$ .

Tree approximation classes contain signals whose wavelet coefficients have a loose (and possibly interrupted) decay from coarse to fine scales. These classes have been well-characterized for wavelet-sparse signals [35, 36, 40] and are intrinsically linked with the Besov spaces  $B_q^s(L_p([0, 1]))$ . Besov spaces contain functions of one or more continuous variables that have (roughly speaking)  $s$  derivatives in  $L_p([0, 1])$ ; the parameter  $q$  provides finer distinctions of smoothness. When a Besov space signal  $x_a \in B_q^s(L_p([0, 1]))$  with  $s > 1/p - 1/2$  is sampled uniformly and converted to a length- $N$  vector  $x$ ,

its wavelet coefficients belong to the tree approximation space  $\mathcal{T}_s$ , with

$$|x_N|_{\mathcal{T}_s} \asymp \|x_a\|_{L_p([0, 1])} + \|x_a\|_{B_q^s(L_p([0, 1]))},$$

where " $\asymp$ " denotes an equivalent norm. The same result holds if  $s = 1/p - 1/2$  and  $q \leq p$ .

### D. Stable tree-based recovery from compressive measurements

For tree-sparse signals, by applying Theorem 1 and Proposition 1, we find that a subgaussian random matrix has the  $\mathcal{T}_K$ -RIP property with constant  $\delta_{\mathcal{T}_K}$  and probability  $1 - e^{-t}$  if the number of measurements obeys

$$M \geq \frac{2}{c\delta_{\mathcal{T}_K}^2} \left( K \ln \frac{48}{\delta_{\mathcal{T}_K}} + \ln \frac{512}{Ke^2} + t \right).$$

Thus, the number of measurements necessary for stable recovery of tree-sparse signals is linear in  $K$ , without the dependence on  $N$  present in conventional non-model-based CS recovery.

For tree-compressible signals, we must quantify the number of subspaces  $R_j$  in each residual set  $\mathcal{R}_{j,K}$  for the approximation class. We can then apply the theory of Section IV-C with Proposition 1 to calculate the smallest allowable  $M$  via Theorem 5.

**Proposition 2:** The number of  $K$ -dimensional subspaces that comprise  $\mathcal{R}_{j,K}$  obeys

$$R_j \leq \frac{(2e)^{K(2j+1)}}{(Kj + K + 1)(Kj + 1)}. \quad (17)$$

Using Proposition 2 and Theorem 5, we obtain the following condition for the matrix  $\Phi$  to have the RAmP, which is proved in Appendix F.

**Proposition 3:** Let  $\Phi$  be an  $M \times N$  matrix with i.i.d. subgaussian entries. If

$$M \geq \frac{2 \left( 10K + 2 \ln \frac{N}{K(K+1)(2K+1)} + t \right)}{(\sqrt{1 + \epsilon_K} - 1)^2},$$

then the matrix  $\Phi$  has the  $(\epsilon_K, s)$ -RAmP for the structured sparsity model  $\mathcal{T}$  and all  $s > 0.5$  with probability  $1 - e^{-t}$ .

Both cases give a simplified bound on the number of measurements required as  $M = \mathcal{O}(K)$ , which is a substantial improvement over the  $M = \mathcal{O}(K \log(N/K))$  required by conventional CS recovery methods. Thus, when  $\Phi$  satisfies Proposition 3, we have the guarantee (14) for sampled Besov space signals from  $B_q^s(L_p([0, 1]))$ .

### E. Experiments

We now present the results of a number of numerical experiments that illustrate the effectiveness of a tree-based recovery algorithm. Our consistent observation is that explicit incorporation of the structured sparsity model in the recovery

process significantly improves the quality of recovery for a given number of measurements. In addition, model-based recovery remains stable when the inputs are no longer tree-sparse, but rather are tree-compressible and/or corrupted with differing levels of noise. We employ the Daubechies-6 wavelet basis for sparsity, and recover the signal using model-based CoSaMP (Algorithm 1) with a CSSA-based structured sparse approximation step in all experiments.

We first study one-dimensional signals that match the connected wavelet-tree model described above. Among such signals is the class of piecewise smooth functions, which are commonly encountered in analysis and practice.

Figure 1 illustrates the results of recovering the tree-compressible *HeaviSine* signal of length  $N = 1024$  from  $M = 80$  noise-free random Gaussian measurements using CoSaMP,  $\ell_1$ -norm minimization using the `l1_eq` solver from the  $\ell_1$ -Magic toolbox,<sup>3</sup> and our tree-based recovery algorithm. It is clear that the number of measurements ( $M = 80$ ) is far fewer than the minimum number required by CoSaMP and  $\ell_1$ -norm minimization to accurately recover the signal. In contrast, tree-based recovery using  $K = 26$  is accurate and uses fewer iterations to converge than conventional CoSaMP. Moreover, the normalized magnitude of the squared error for tree-based recovery is equal to 0.037, which is remarkably close to the error between the noise-free signal and its *best*  $K$ -term tree-approximation (0.036).

Figure 3(a) illustrates the results of a Monte Carlo simulation study on the impact of the number of measurements  $M$  on the performance of model-based and conventional recovery for a class of tree-sparse piecewise polynomial signals. Each data point was obtained by measuring the normalized recovery error of 500 sample trials. Each sample trial was conducted by generating a new piecewise polynomial signal of length  $N = 1024$  with five polynomial pieces of cubic degree and randomly placed discontinuities, computing its best  $K$ -term tree-approximation using the CSSA, and then measuring the resulting signal using a matrix with i.i.d. Gaussian entries. Model-based recovery attains near-perfect recovery at  $M = 3K$  measurements, while CoSaMP only matches this performance at  $M = 5K$ .

For the same class of signals, we empirically compared the recovery times of our proposed algorithm with those of the standard approach (CoSaMP). Experiments were conducted on a Sun workstation with a 1.8GHz AMD Opteron dual-core processor and 2GB memory running UNIX, using non-optimized Matlab code and a function-handle based implementation of the random projection operator  $\Phi$ . As is evident from Figure 3(b), wavelet tree-based recovery is in general slower than CoSaMP. This is due to the fact that the CSSA step in the iterative procedure is more computationally demanding than simple  $K$ -term approximation. Nevertheless, the highest benefits of model-based CS recovery are obtained around  $M = 3K$ ; in this regime, the runtimes of the two approaches are comparable, with tree-based recovery requiring

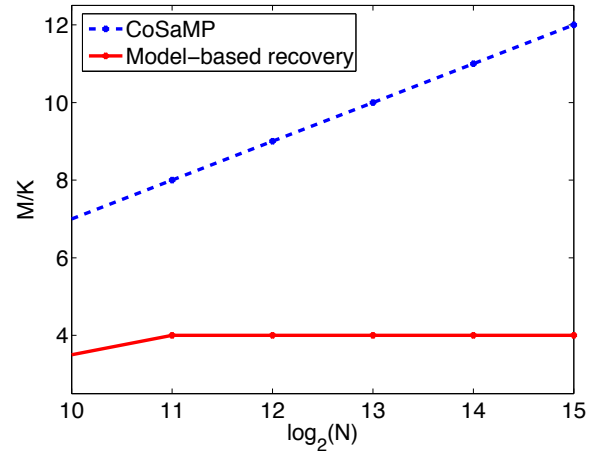


Fig. 4. Required overmeasuring factor  $M/K$  to achieve a target recovery error  $\|x - \hat{x}\|_2 \leq 2.5\sigma_{T_K}(x)$  as a function of the signal length  $N$  for standard and model-based recovery of piecewise smooth signals. While standard recovery requires  $M$  to increase logarithmically with  $N$ , the required  $M$  is essentially constant for model-based recovery.

fewer iterations and yielding much smaller recovery error than standard recovery.

Figure 4 shows the growth of the overmeasuring factor  $M/K$  with the signal length  $N$  for conventional CS and model-based recovery. We generated 50 sample piecewise cubic signals and numerically computed the minimum number of measurements  $M$  required for the recovery error  $\|x - \hat{x}\|_2 \leq 2.5\sigma_{T_K}(x)$ , the *best* tree-approximation error, for every sample signal. The figure shows that while doubling the signal length increases the number of measurements required by standard recovery by  $K$ , the number of measurements required by model-based recovery is constant for all  $N$ . These experimental results verify the theoretical performance described in Proposition 3.

Further, we demonstrate that model-based recovery performs stably in the presence of measurement noise. We generated sample piecewise polynomial signals as above, computed their best  $K$ -term tree-approximations, computed  $M$  measurements of each approximation, and finally added Gaussian noise of variance  $\sigma$  to each measurement, so that the expected variance  $E[\|n\|_2] = \sigma\sqrt{M}$ . We emphasize that this noise model implies that the energy of the noise added will be larger as  $M$  increases. Then, we recovered the signals using CoSaMP and model-based recovery and measured the recovery error in each case. For comparison purposes, we also tested the recovery performance of a  $\ell_1$ -norm minimization algorithm that accounts for the presence of noise, which has been implemented as the `l1_qc` solver in the  $\ell_1$ -Magic toolbox. First, we determined the lowest value of  $M$  for which the respective algorithms provided near-perfect recovery in the absence of noise in the measurements. This corresponds to  $M = 3.5K$  for model-based recovery,  $M = 5K$  for CoSaMP, and  $M = 4.5K$  for  $\ell_1$ -norm minimization. Next, we generated 200 sample tree-structured signals, computed  $M$  *noisy* measurements, recovered the signal using the given algorithm

<sup>3</sup><http://www.acm.caltech.edu/l1magic>.

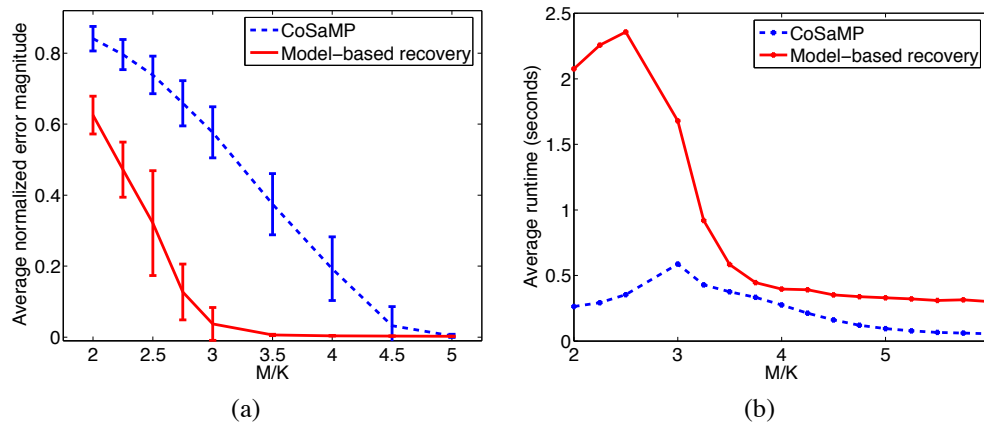


Fig. 3. Performance of CoSaMP vs. wavelet tree-based recovery on a class of tree-sparse signals. (a) Average normalized recovery error and (b) average runtime for each recovery algorithm as a function of the overmeasuring factor  $M/K$ . The number of measurements  $M$  for which the wavelet tree-based algorithm obtains near-perfect recovery is much smaller than that required by CoSaMP. The penalty paid for this improvement is a modest increase in the runtime.

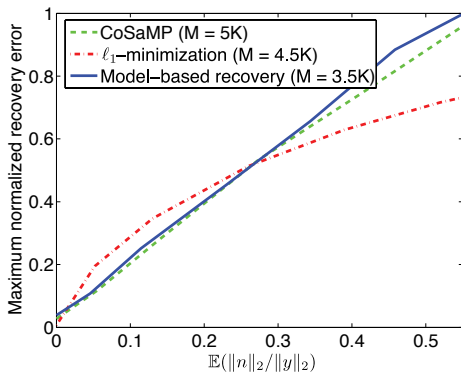


Fig. 5. Robustness to measurement noise for standard and wavelet tree-based CS recovery algorithms. We plot the maximum normalized recovery error over 200 sample trials as a function of the expected signal-to-noise ratio. The linear growth demonstrates that model-based recovery possesses the same robustness to noise as CoSaMP and  $\ell_1$ -norm minimization.

and recorded the recovery error. Figure 5 illustrates the growth in maximum normalized recovery error (over the 200 sample trials) as a function of the expected measurement signal-to-noise ratio for the tree algorithms. We observe similar stability curves for all three algorithms, while noting that model-based recovery offers this kind of stability using significantly fewer measurements.

Finally, we turn to two-dimensional images and a wavelet quadtree model. The connected wavelet-tree model has proven useful for compressing natural images [35]; thus, our algorithm provides a simple and provably efficient method for recovering a wide variety of natural images from compressive measurements. An example of recovery performance is given in Figure 6. The test image (*Peppers*) is of size  $N = 128 \times 128 = 16384$  pixels, and we computed  $M = 5000$  random Gaussian measurements. Model-based recovery again offers higher performance than standard signal recovery algorithms like CoSaMP, both in terms of recovery mean-squared error and visual quality.

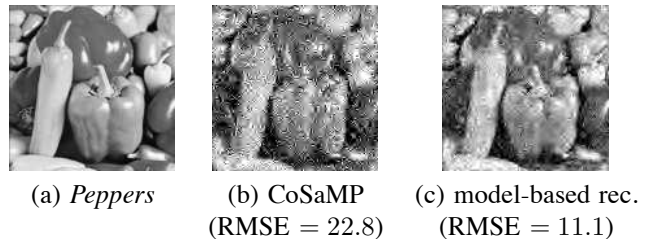


Fig. 6. Example performance of standard and model-based recovery on images. (a)  $N = 128 \times 128 = 16384$ -pixel *Peppers* test image. Image recovery from  $M = 5000$  compressive measurements using (b) conventional CoSaMP and (c) our wavelet tree-based algorithm.

## VI. EXAMPLE: BLOCK-SPARSE SIGNALS AND SIGNAL ENSEMBLES

In a *block-sparse* signal, the locations of the significant coefficients cluster in blocks under a specific sorting order. Block-sparse signals have been previously studied in CS applications, including DNA microarrays and magnetoencephalography [8,9]. An equivalent problem arises in CS for signal ensembles, such as sensor networks and MIMO communication [9,10,42]. In this case, several signals share a common coefficient support set. For example, when a frequency-sparse acoustic signal is recorded by an array of microphones, then all of the recorded signals contain the same Fourier frequencies but with different amplitudes and delays. Such a signal ensemble can be re-shaped as a single vector by concatenation, and then the coefficients can be rearranged so that the concatenated vector exhibits block sparsity.

It has been shown that the block-sparse structure enables signal recovery from a reduced number of CS measurements, both for the single signal case [8,9,43] and the signal ensemble case [10], through the use of specially tailored recovery algorithms. However, the robustness guarantees for the algorithms [8,43] either are restricted to exactly sparse signals and noiseless measurements, do not have explicit bounds on the number of necessary measurements, or are asymptotic in nature. An optimization-based algorithm introduced in [9]

provides similar recovery guarantees to those obtained by the algorithm we present in this chapter; thus, our method can be interpreted as a greedy-based counterpart to that provided in [9].

In this section, we formulate the block sparsity model as a union of subspaces and pose an approximation algorithm on this union of subspaces. The approximation algorithm is used to implement block-based signal recovery. We also define the corresponding class of block-compressible signals and quantify the number of measurements necessary for robust recovery.

#### A. Block-sparse signals

Consider a class  $\mathcal{S}$  of signal vectors  $x \in \mathbb{R}^{JN}$ , with  $J$  and  $N$  integers. This signal can be reshaped into a  $J \times N$  matrix  $X$ , and we use both notations interchangeably in the sequel. We will restrict entire columns of  $X$  to be part of the support of the signal as a group. That is, signals  $X$  in a block-sparse model have entire columns as zeros or nonzeros. The measure of sparsity for  $X$  is its number of nonzero columns. More formally, we make the following definition.

*Definition 11:* [8,9] Define the set of  $K$ -block sparse signals as

$$\mathcal{S}_K = \{X = [x_1 \dots x_N] \in \mathbb{R}^{J \times N} \text{ such that } x_n = 0 \text{ for } n \notin \Omega, \Omega \subseteq \{1, \dots, N\}, |\Omega| = K\}.$$

It is important to note that a  $K$ -block sparse signal has sparsity  $KJ$ , which is dependent on the size of the block  $J$ . We can extend this formulation to ensembles of  $J$ , length- $N$  signals with common support. Denote this signal ensemble by  $\{\tilde{x}_1, \dots, \tilde{x}_J\}$ , with  $\tilde{x}_j \in \mathbb{R}^N$ ,  $1 \leq j \leq J$ . We formulate a matrix representation  $\tilde{X}$  of the ensemble that features the signal  $\tilde{x}_j$  in its  $j^{\text{th}}$  row:  $\tilde{X} = [\tilde{x}_1 \dots \tilde{x}_J]^T$ . The matrix  $\tilde{X}$  features the same structure as the matrix  $X$  obtained from a block-sparse signal; thus, the matrix  $\tilde{X}$  can be converted into a block-sparse vector  $\tilde{x}$  that represents the signal ensemble.

#### B. Block-based approximation

To pose the block-based approximation algorithm, we need to define the mixed norm of a matrix.

*Definition 12:* The  $(p, q)$  mixed norm of the matrix  $X = [x_1 \ x_2 \ \dots \ x_N]$  is defined as

$$\|X\|_{(p,q)} = \left( \sum_{n=1}^N \|x_n\|_p^q \right)^{1/q}.$$

When  $q = 0$ ,  $\|X\|_{(p,0)}$  simply counts the number of nonzero columns in  $X$ .

We immediately find that  $\|X\|_{(p,p)} = \|x\|_p$ , with  $x$  the vectorization of  $X$ . Intuitively, we pose the algorithm  $\mathbb{S}(X, K)$

to obtain the best block-based approximation of the signal  $X$  as follows:

$$X_K^S = \arg \min_{\tilde{X} \in \mathbb{R}^{J \times N}} \|X - \tilde{X}\|_{(2,2)} \text{ s.t. } \|\tilde{X}\|_{(2,0)} \leq K. \quad (18)$$

It is easy to show that to obtain the approximation, it suffices to perform column-wise hard thresholding: let  $\rho$  be the  $K^{\text{th}}$  largest  $\ell_2$ -norm among the columns of  $X$ . Then our approximation algorithm is  $\mathbb{S}(X, K) = X_K^S = [x_{K,1}^S \ \dots \ x_{K,N}^S]$ , where

$$x_{K,n}^S = \begin{cases} x_n & \|x_n\|_2 \geq \rho, \\ 0 & \|x_n\|_2 < \rho, \end{cases}$$

for each  $1 \leq n \leq N$ . Alternatively, a recursive approximation algorithm can be obtained by sorting the columns of  $X$  by their  $\ell_2$  norms, and then selecting the columns with largest norms. The complexity of this sorting process is  $\mathcal{O}(NJ + N \log N)$ .

#### C. Block-compressible signals

The approximation class under the block-compressible model corresponds to signals with blocks whose  $\ell_2$  norm has a power-law decay rate.

*Definition 13:* We define the set of  $s$ -block compressible signals as

$$\mathfrak{S}_s = \{X = [x_1 \dots x_N] \in \mathbb{R}^{J \times N} \text{ such that } \|\mathcal{I}(i)\|_2 \leq Gi^{-s-1/2}, 1 \leq i \leq N, S < \infty\},$$

where  $\mathcal{I}$  indexes the sorted column norms.

We say that  $X$  is an  $s$ -block compressible signal if  $X \in \mathfrak{S}_s$ . For such signals, we have  $\|X - X_K\|_{(2,2)} = \sigma_{S_K}(x) \leq G_1 K^{-s}$ , and  $\|X - X_K\|_{(2,1)} \leq G_2 K^{1/2-s}$ . Note that the block-compressible model does not impart a structure to the decay of the signal coefficients, so that the sets  $\mathcal{R}_{j,K}$  are equal for all values of  $j$ ; due to this property, the  $(\delta_{S_K}, s)$ -RAM-P is implied by the  $\mathcal{S}_K$ -RIP. Taking this into account, we can derive the following result from [11], which is proven similarly to Theorem 4.

*Theorem 6:* Let  $x$  be a signal from the structured sparsity model  $\mathcal{S}$ , and let  $y = \Phi x + n$  be a set of noisy CS measurements. If  $\Phi$  has the  $\mathcal{S}_K^4$ -RIP with  $\delta_{S_K^4} \leq 0.1$ , then the estimate obtained from iteration  $i$  of block-based CoSaMP, using the approximation algorithm (18), satisfies

$$\|x - \hat{x}_i\|_2 \leq 2^{-i} \|x\|_2 + 20 \left( \|X - X_K^S\|_{(2,2)} + \frac{1}{\sqrt{K}} \|X - X_K^S\|_{(2,1)} + \|n\|_2 \right).$$

Thus, the algorithm provides a recovered signal of similar quality to approximations of  $X$  by a small number of nonzero columns. When the signal  $x$  is  $K$ -block sparse, we have that  $\|X - X_K^S\|_{(2,2)} = \|X - X_K^S\|_{(2,1)} = 0$ , obtaining the same result as Theorem 4, save for a constant factor.



#### D. Stable block-based recovery from compressive measurements

Since Theorem 6 poses the same requirement on the measurement matrix  $\Phi$  for sparse and compressible signals, the same number of measurements  $M$  is required to provide performance guarantees for block-sparse and block-compressible signals. The class  $\mathcal{S}_K$  contains  $S = \binom{N}{K}$  subspaces of dimension  $JK$ . Thus, a subgaussian random matrix has the  $\mathcal{S}_K$ -RIP property with constant  $\delta_{\mathcal{S}_K}$  and probability  $1 - e^{-t}$  if the number of measurements obeys

$$M \geq \frac{2}{c\delta_{\mathcal{S}_K}^2} \left( K \left( \ln \frac{2N}{K} + J \ln \frac{12}{\delta_{\mathcal{S}_K}} \right) + t \right). \quad (19)$$

To compare with the standard CS measurement bound, the number of measurements required for robust recovery scales as  $M = \mathcal{O}(JK + K \log(N/K))$ , which is a substantial improvement over the  $M = \mathcal{O}(JK \log(N/K))$  that would be required by conventional CS recovery methods. When the size of the block  $J$  is larger than  $\log(N/K)$ , then this term becomes  $\mathcal{O}(KJ)$ ; that is, it is linear on the total sparsity of the block-sparse signal.

We note in passing that the bound on the number of measurements (19) assumes a dense subgaussian measurement matrix, while the measurement matrices used in [10] have a block-diagonal structure. To obtain measurements from an  $M \times JN$  dense matrix in a distributed setting, it suffices to partition the matrix into  $J$  pieces of size  $M \times N$  and calculate the CS measurements at each sensor with the corresponding matrix; these individual measurements are then summed to obtain the complete measurement vector. For large  $J$ , (19) implies that the total number of measurements required for recovery of the signal ensemble is lower than the bound for the case where each signal recovery is performed independently for each signal ( $M = \mathcal{O}(JK \log(N/K))$ ).

#### E. Experiments

We conducted several numerical experiments comparing model-based recovery to CoSaMP in the context of block-sparse signals. We employ the model-based CoSaMP recovery of Algorithm 1 with the block-based approximation algorithm (18) in all cases. For brevity, we exclude a thorough comparison of our model-based algorithm with  $\ell_1$ -norm minimization and defer it to future work. In practice, we observed that our algorithm performs several times faster than convex optimization-based procedures.

Figure 7 illustrates an  $N = 4096$  signal that exhibits block sparsity, and its recovered version from  $M = 960$  measurements using CoSaMP and model-based recovery. The block size  $J = 64$  and there were  $K = 6$  active blocks in the signal. We observe the clear advantage of using the block-sparsity model in signal recovery.

We now consider block-compressible signals. An example recovery is illustrated in Figure 8. In this case, the  $\ell_2$ -norms of the blocks decay according to a power law, as

described above. Again, the number of measurements is far below the minimum number required to guarantee stable recovery through conventional CS recovery. However, enforcing the structured sparsity model in the approximation process results in a solution that is very close to the best 5-block approximation of the signal.

Figure 9(a) indicates the decay in recovery error as a function of the numbers of measurements for CoSaMP and model-based recovery. We generated sample block-sparse signals as follows: we randomly selected a set of  $K$  blocks, each of size  $J$ , and endow them with coefficients that follow an i.i.d. Gaussian distribution. Each sample point in the curves is generated by performing 200 trials of the corresponding algorithm. As in the connected wavelet-tree case, we observe clear gains using model-based recovery, particularly for low-measurement regimes; CoSaMP matches model-based recovery only for  $M \geq 5K$ .

Figure 9(b) compares the recovery times of the two approaches. For this particular model, we observe that our proposed approach is in general much faster than CoSaMP. This is because of two reasons: a) the block-based approximation step involves sorting fewer coefficients, and thus is faster than  $K$ -term approximation; b) block-based recovery requires fewer iterations to converge to the true solution.

## VII. CONCLUSIONS

In this paper, we have aimed to demonstrate that there are significant performance gains to be made by exploiting more realistic and richer signal models beyond the simplistic sparse and compressible models that dominate the CS literature. Building on the unions of subspaces results of [6] and the proof machinery of [11], we have taken some first steps towards what promises to be a general theory for model-based CS by introducing the notion of a structured compressible signal and the associated restricted amplification **property** (RAmP) condition it imposes on the measurement matrix  $\Phi$ . Our analysis poses the nested approximation property (NAP) as a sufficient condition that is **satisfied** by many structured sparsity models.

For the volumes of natural and manmade signals and images that are wavelet-sparse or compressible, our **tree-based** CoSaMP and IHT algorithms offer performance that **significantly exceeds** today's state-of-the-art while requiring only  $M = \mathcal{O}(K)$  rather than  $M = \mathcal{O}(K \log(N/K))$  random measurements. For block-sparse signals and signal ensembles with common sparse support, our block-based CoSaMP and IHT algorithms offer not only excellent performance but also require just  $M = \mathcal{O}(JK)$  measurements, where  $JK$  is the signal sparsity. Furthermore, block-based recovery can recover signal ensembles using fewer measurements than the number required when each signal is recovered independently; we have shown such advantages using real-world data from environmental sensor networks [44]. Additional structured sparsity models have been developed using our general framework in [45] and [46]; we have also released a Matlab toolbox



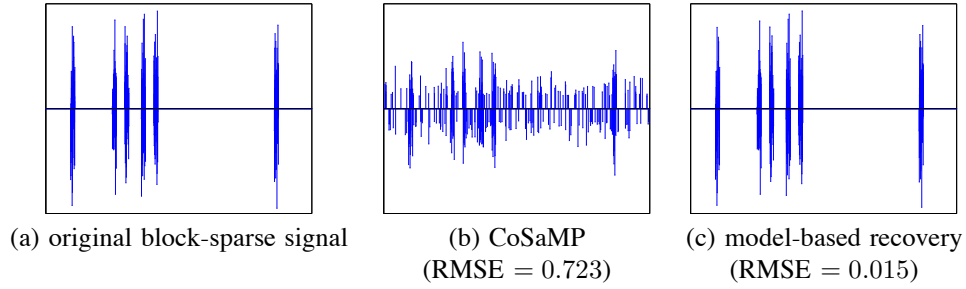


Fig. 7. Example performance of structured signal recovery for a block-sparse signal. (a) Example block-sparse signal of length  $N = 4096$  with  $K = 6$  nonzero blocks of size  $J = 64$ . Recovered signal from  $M = 960$  measurements using (b) conventional CoSaMP recovery and (c) block-based recovery. Standard recovery not only recovers spurious nonzeros, but also attenuates the magnitude of the actual nonzero entries.

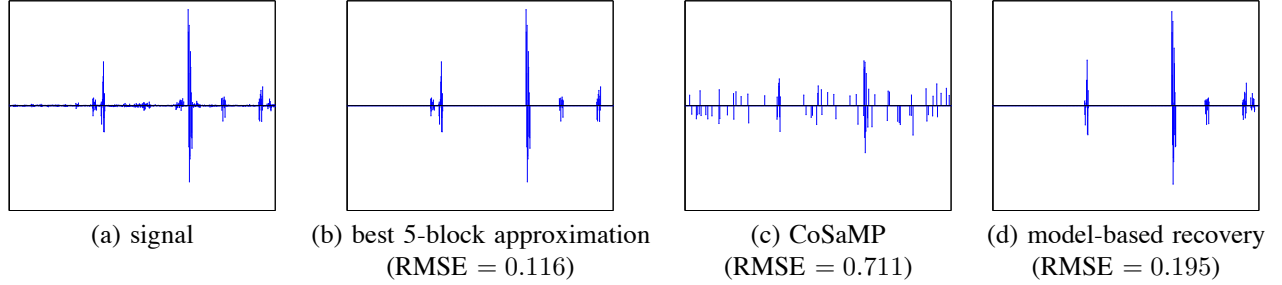


Fig. 8. Example performance of structured signal recovery for block-compressible signals. (a) Example block-compressible signal, length  $N = 1024$ . (b) Best block-based approximation with  $K = 5$  blocks. Recovered signal from  $M = 200$  measurements using both (c) conventional CoSaMP recovery and (d) block-based recovery. Standard recovery not only recovers spurious significant entries, but also attenuates the magnitude of the actual significant entries

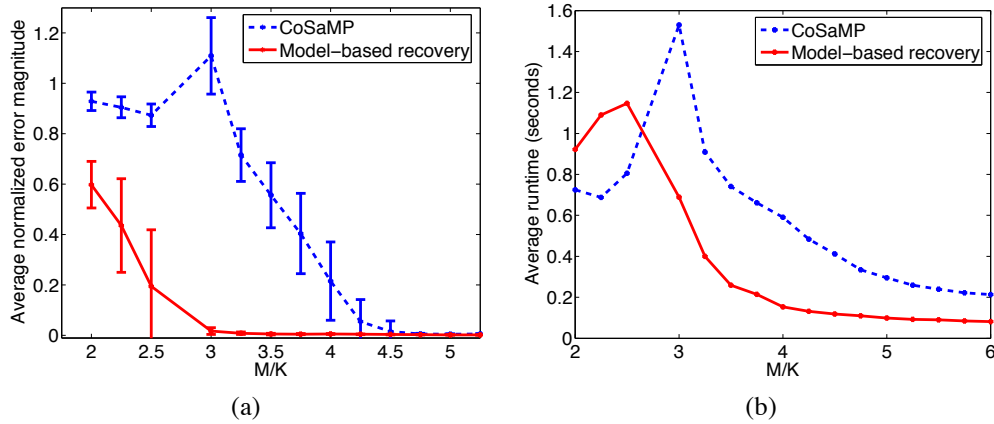


Fig. 9. Performance of CoSaMP vs. block-based recovery on a class of block-sparse signals. (a) Average normalized recovery error and (b) average runtime for each recovery algorithm as a function of the overmeasuring factor  $M/K$ . CoSaMP does not match the performance of the block-based algorithm until  $M = 5K$ . Furthermore, the block-based algorithm has faster convergence time than CoSaMP.

containing the corresponding model-based CS recovery algorithms, available at <http://dsp.rice.edu/software>.

There are many avenues for future work on model-based CS. We have only considered the recovery of signals from models that can be geometrically described as a union of subspaces; possible extensions include other, more complex geometries such as high-dimensional polytopes and nonlinear manifolds. We also expect that the core of our proposed algorithms — a structured sparse approximation step — can be integrated into other iterative algorithms, such as relaxed  $\ell_1$ -norm minimization methods. Furthermore, our framework will benefit from the formulation of new structured sparsity

models that are endowed with efficient structured sparse approximation algorithms.

## APPENDIX A PROOF OF THEOREM 2

To prove this theorem, we will study the distribution of the maximum singular value of a submatrix  $\Phi_T$  of a matrix with i.i.d. Gaussian entries  $\Phi$  corresponding to the columns indexed by  $T$ . From this we obtain the probability that RAmP does not hold for a fixed support  $T$ . We will then evaluate the same probability for all supports  $T$  of elements of  $\mathcal{R}_{j,K}$ , where the desired bound on the amplification is dependent on

the value of  $j$ . This gives us the probability that the RAmP does not hold for a given residual subspace set  $\mathcal{R}_{j,K}$ . We fix the probability of failure on each of these sets; we then obtain probability that the matrix  $\Phi$  does not have the RAmP using a union bound. We end by obtaining conditions on the number of rows  $M$  of  $\Phi$  to obtain a desired probability of failure.

We begin from the following concentration of measure for the largest singular value of a  $M \times K$  submatrix  $\Phi_T$ ,  $|T| = K$ , of an  $M \times N$  matrix  $\Phi$  with i.i.d. subgaussian entries that are properly normalized [25,47,48]:

$$P\left(\sigma_{\max}(\Phi_T) > 1 + \sqrt{\frac{K}{M}} + \tau + \beta\right) \leq e^{-M\tau^2/2}.$$

For large enough  $M$ ,  $\beta \ll 1$ ; thus we ignore this small constant in the sequel. By letting  $\tau = j^r \sqrt{1 + \epsilon_K} - 1 - \sqrt{\frac{K}{M}}$  (with the appropriate value of  $j$  for  $T$ ), we obtain

$$P\left(\sigma_{\max}(\Phi_T) > j^r \sqrt{1 + \epsilon_K}\right) \leq e^{-\frac{M}{2} \left(j^r \sqrt{1 + \epsilon_K} - 1 - \sqrt{\frac{K}{M}}\right)^2}.$$

We use a union bound over all possible  $R_j$  supports for  $u \in \mathcal{R}_{j,K}$  to obtain the probability that  $\Phi$  amplifies the norm of some  $u$  by more than  $j^r \sqrt{1 + \epsilon_K}$ :

$$\begin{aligned} P\left(\|\Phi u\|_2 > (j^r \sqrt{1 + \epsilon_K}) \|u\|_2 \text{ for some } u \in \mathcal{R}_{j,K}\right) \\ \leq R_j e^{-\frac{1}{2}(\sqrt{M}(j^r \sqrt{1 + \epsilon_K} - 1) - \sqrt{K})^2}. \end{aligned}$$

Bound the right hand side by a constant  $\mu$ ; this requires

$$R_j \leq e^{\frac{1}{2}(\sqrt{M}(j^r \sqrt{1 + \epsilon_K} - 1) - \sqrt{K})^2} \mu \quad (20)$$

for each  $j$ . We use another union bound among the residual subspaces  $\mathcal{R}_{j,K}$  to measure the probability that the RAmP does not hold:

$$\begin{aligned} P\left(\frac{\|\Phi u\|_2}{\|u\|_2} > j^r \sqrt{1 + \epsilon_K}, u \in \mathcal{R}_{j,K}, 1 \leq j \leq \lceil N/K \rceil\right) \\ \leq \left\lceil \frac{N}{K} \right\rceil \mu. \end{aligned}$$

To bound this probability by  $e^{-t}$ , we need  $\mu = \frac{K}{N} e^{-t}$ ; plugging this into (20), we obtain

$$R_j \leq e^{\frac{1}{2}(\sqrt{M}(j^r \sqrt{1 + \epsilon_K} - 1) - \sqrt{K})^2} \frac{K}{N} e^{-t}$$

for each  $j$ . Simplifying, we obtain that for  $\Phi$  to possess the RAmP with probability  $1 - e^{-t}$ , the following must hold for all  $j$ :

$$M \geq \left( \frac{\sqrt{2 \left( \ln \frac{R_j N}{K} + t \right) + \sqrt{K}}}{j^r \sqrt{1 + \epsilon_K} - 1} \right)^2. \quad (21)$$

Since  $(\sqrt{a} + \sqrt{b})^2 \leq 2a + 2b$  for  $a, b > 0$ , then the hypothesis (12) implies (21), proving the theorem.  $\square$

## APPENDIX B PROOF OF THEOREM 3

In this proof, we denote  $\mathbb{M}(x, K) = x_K$  for brevity. To bound  $\|\Phi(x - x_K)\|_2$ , we write  $x$  as

$$x = x_K + \sum_{j=2}^{\lceil N/K \rceil} x_{T_j},$$

where

$$x_{T_j} = x_{jK} - x_{(j-1)K}, j = 2, \dots, \lceil N/K \rceil$$

is the difference between the best  $jK$  structured sparse approximation and the best  $(j-1)K$  structured sparse approximation. Additionally, each piece  $x_{T_j} \in \mathcal{R}_{j,K}$ . Therefore, since  $\Phi$  satisfies the  $(\epsilon_K, s-1)$ -RAmP, we obtain

$$\begin{aligned} \|\Phi(x - x_K)\|_2 &= \left\| \Phi \left( \sum_{j=2}^{\lceil N/K \rceil} x_{T_j} \right) \right\|_2 \leq \sum_{j=2}^{\lceil N/K \rceil} \|\Phi x_{T_j}\|_2 \\ &\leq \sum_{j=2}^{\lceil N/K \rceil} \sqrt{1 + \epsilon_K} j^{s-1} \|x_{T_j}\|_2. \end{aligned}$$

Since  $x \in \mathfrak{M}_s$ , the norm of each piece can be bounded as

$$\begin{aligned} \|x_{T_j}\|_2 &= \|x_{jK} - x_{(j-1)K}\|_2 \\ &\leq \|x - x_{(j-1)K}\|_2 + \|x - x_{jK}\|_2 \\ &\leq |x|_{\mathfrak{M}_s} K^{-s} ((j-1)^{-s} + j^{-s}). \end{aligned}$$

Applying this bound in (22), we obtain

$$\begin{aligned} \|\Phi(x - x_K)\|_2 &\leq \sqrt{1 + \epsilon_K} \sum_{j=2}^{\lceil N/K \rceil} j^{s-1} \|x_{T_j}\|_2, \\ &\leq \frac{\sqrt{1 + \epsilon_K}}{K^s} |x|_{\mathfrak{M}_s} \sum_{j=2}^{\lceil N/K \rceil} \frac{j^{s-1}}{(j-1)^s} + \frac{j^{s-1}}{j^s}, \\ &\leq \frac{\sqrt{1 + \epsilon_K}}{K^s} |x|_{\mathfrak{M}_s} \sum_{j=2}^{\lceil N/K \rceil} \frac{1}{j(1 - 1/j)^s} + \frac{1}{j}, \\ &\leq \frac{\sqrt{1 + \epsilon_K}}{K^s} |x|_{\mathfrak{M}_s} \sum_{j=2}^{\lceil N/K \rceil} \frac{2^s}{j} + \frac{1}{j}, \\ &\leq (2^s + 1) \frac{\sqrt{1 + \epsilon_K}}{K^s} |x|_{\mathfrak{M}_s} \sum_{j=2}^{\lceil N/K \rceil} j^{-1}. \end{aligned}$$

It is easy to show, using Euler-Maclaurin summations, that  $\sum_{j=2}^{\lceil N/K \rceil} j^{-1} \leq \ln \lceil N/K \rceil$ ; we then obtain

$$\|\Phi(x - x_K)\|_2 \leq (2^s + 1) \sqrt{1 + \epsilon_K} K^{-s} \ln \left\lceil \frac{N}{K} \right\rceil |x|_{\mathfrak{M}_s},$$

which proves the theorem.  $\square$

## APPENDIX C MODEL-BASED ITERATIVE HARD THRESHOLDING

Our proposed model-based iterative hard thresholding (IHT) is given in Algorithm 2. For this algorithm, Theorems 4,

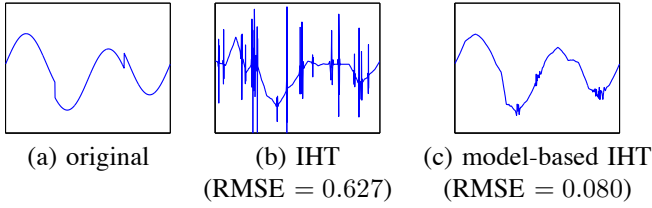


Fig. 10. Example performance of model-based IHT. (a) Piecewise smooth HeaviSine test signal, length  $N = 1024$ . Signal recovered from  $M = 80$  measurements using both (b) standard and (c) model-based IHT recovery. Root mean-squared error (RMSE) values are normalized with respect to the  $\ell_2$  norm of the signal.

5, and 6 can be proven with only a few modifications:  $\Phi$  must have the  $\mathcal{M}_K^3$ -RIP with  $\delta_{\mathcal{M}_K^3} \leq 0.1$ , and the constant factor in the bound changes from 15 to 4 in Theorem 4, from 35 to 10 in Theorem 5, and from 20 to 5 in Theorem 6.

To illustrate the performance of the algorithm, we repeat the *HeaviSine* experiment from Figure 1. Recall that  $N = 1024$ , and  $M = 80$  for this example. The advantages of using our tree-structured sparse approximation step (instead of mere hard thresholding) are evident from Figure 10. In practice, we have observed that our model-based algorithm converges in fewer steps than IHT and yields much more accurate results in terms of recovery error.

#### APPENDIX D PROOF OF THEOREM 4

The proof of this theorem is identical to that of the CoSaMP algorithm in [11, Section 4.6], and requires a set of six lemmas. The sequence of Lemmas 1–6 below are modifications of the lemmas in [11] that are restricted to the structured sparsity model. Lemma 4 does not need any changes from [11], so we state it without proof. The proof of Lemmas 3–6 use the properties in Lemmas 1 and 2, which are simple to prove.

**Lemma 1:** Suppose  $\Phi$  has  $\mathcal{M}$ -RIP with constant  $\delta_{\mathcal{M}}$ . Let  $\Omega$  be a support corresponding to a subspace in  $\mathcal{M}$ . Then we have the following handy bounds.

$$\begin{aligned} \|\Phi_{\Omega}^T u\|_2 &\leq \sqrt{1 + \delta_{\mathcal{M}}} \|u\|_2, \\ \|\Phi_{\Omega}^{\dagger} u\|_2 &\leq \frac{1}{\sqrt{1 - \delta_{\mathcal{M}}}} \|u\|_2, \\ \|\Phi_{\Omega}^T \Phi_{\Omega} u\|_2 &\leq (1 + \delta_{\mathcal{M}}) \|u\|_2, \\ \|\Phi_{\Omega}^T \Phi_{\Omega} u\|_2 &\geq (1 - \delta_{\mathcal{M}}) \|u\|_2, \\ \|(\Phi_{\Omega}^T \Phi_{\Omega})^{-1} u\|_2 &\leq \frac{1}{1 - \delta_{\mathcal{M}}} \|u\|_2, \\ \|(\Phi_{\Omega}^T \Phi_{\Omega})^{-1} u\|_2 &\geq \frac{1}{1 + \delta_{\mathcal{M}}} \|u\|_2. \end{aligned}$$

**Lemma 2:** Suppose  $\Phi$  has  $\mathcal{M}_K^2$ -RIP with constant  $\delta_{\mathcal{M}_K^2}$ . Let  $\Omega$  be a support corresponding to a subspace in  $\mathcal{M}_K$ , and let  $x \in \mathcal{M}_K$ . Then  $\|\Phi_{\Omega}^T \Phi x|_{\Omega^c}\|_2 \leq \delta_{\mathcal{M}_K^2} \|x|_{\Omega^c}\|_2$ .

We begin the proof of Theorem 4 by fixing an iteration  $i \geq 1$  of model-based CoSaMP. We write  $\hat{x} = \hat{x}_{i-1}$  for the

signal estimate at the beginning of the  $i^{\text{th}}$  iteration. Define the signal residual  $s = x - \hat{x}$ , which implies that  $s \in \mathcal{M}_K^2$ . We note that we can write  $r = y - \Phi \hat{x} = \Phi(x - \hat{x}) + n = \Phi s + n$ .

**Lemma 3: (Identification)** The set  $\Omega = \text{supp}(\mathbb{M}_2(e, K))$ , where  $e = \Phi^T r$ , identifies a subspace in  $\mathcal{M}_K^2$ , and obeys

$$\|s|_{\Omega^c}\|_2 \leq 0.2223 \|s\|_2 + 2.34 \|n\|_2.$$

**Proof of Lemma 3:** Define the set  $\Pi = \text{supp}(s)$ . Let  $e_{\Omega} = \mathbb{M}_2(e, K)$  be the structured sparse approximation to  $e$  with support  $\Omega$ , and similarly let  $e_{\Pi}$  be the approximation to  $e$  with support  $\Pi$ . Each approximation is equal to  $e$  for the coefficients in the support, and zero elsewhere. Since  $\Omega$  is the support of the best approximation in  $\mathcal{M}_K^2$ , we must have:

$$\begin{aligned} \|e - e_{\Omega}\|_2^2 &\leq \|e - e_{\Pi}\|_2^2, \\ \sum_{n=1}^N (e[n] - e_{\Omega}[n])^2 &\leq \sum_{n=1}^N (e[n] - e_{\Pi}[n])^2, \\ \sum_{n \notin \Omega} e[n]^2 &\leq \sum_{n \notin \Pi} e[n]^2, \\ \sum_{n=1}^N e[n]^2 - \sum_{n \notin \Omega} e[n]^2 &\geq \sum_{n=1}^N e[n]^2 - \sum_{n \notin \Pi} e[n]^2, \\ \sum_{n \in \Omega} e[n]^2 &\geq \sum_{n \in \Pi} e[n]^2, \\ \sum_{n \in \Omega \setminus \Pi} e[n]^2 &\geq \sum_{n \in \Pi \setminus \Omega} e[n]^2, \\ \|e|_{\Omega \setminus \Pi}\|_2^2 &\geq \|e|_{\Pi \setminus \Omega}\|_2^2, \end{aligned}$$

where  $\Omega \setminus \Pi$  denotes the set difference of  $\Omega$  and  $\Pi$ . These signals are in  $\mathcal{M}_K^4$  (since they arise as the difference of two elements from  $\mathcal{M}_K^2$ ); therefore, we can apply the  $\mathcal{M}_K^4$ -RIP constants and Lemmas 1 and 2 to provide the following bounds on both sides (see [11] for details):

$$\|e|_{\Omega \setminus \Pi}\|_2 \leq \delta_{\mathcal{M}_K^4} \|s\|_2 + \sqrt{1 + \delta_{\mathcal{M}_K^2}} \|n\|_2, \quad (22)$$

$$\begin{aligned} \|e|_{\Pi \setminus \Omega}\|_2 &\geq (1 - \delta_{\mathcal{M}_K^2}) \|s|_{\Omega^c}\|_2 - \delta_{\mathcal{M}_K^2} \|s\|_2 \\ &\quad - \sqrt{1 + \delta_{\mathcal{M}_K^2}} \|n\|_2. \end{aligned} \quad (23)$$

Combining (22) and (23), we obtain

$$\|s|_{\Omega^c}\|_2 \leq \frac{(\delta_{\mathcal{M}_K^2} + \delta_{\mathcal{M}_K^4}) \|s\|_2 + 2\sqrt{1 + \delta_{\mathcal{M}_K^2}} \|n\|_2}{1 - \delta_{\mathcal{M}_K^2}}.$$

The argument is completed by noting that  $\delta_{\mathcal{M}_K^2} \leq \delta_{\mathcal{M}_K^4} \leq 0.1$ .  $\square$

**Lemma 4: (Support Merger)** Let  $\Omega$  be a set of at most  $2K$  indices. Then the set  $\Lambda = \Omega \cup \text{supp}(\hat{x})$  contains at most  $3K$  indices, and  $\|x|_{\Lambda^c}\|_2 \leq \|s|_{\Omega^c}\|_2$ .

**Lemma 5: (Estimation)** Let  $\Lambda$  be a support corresponding to a subspace in  $\mathcal{M}_K^3$ , and define the least squares signal estimate  $b$  by  $b|_T = \Phi_T^{\dagger} y$ ,  $b|_{T^c} = 0$ . Then

$$\|x - b\|_2 \leq 1.112 \|x|_{\Lambda^c}\|_2 + 1.06 \|n\|_2.$$

**Algorithm 2** Model-based Iterative Hard ThresholdingInputs: CS matrix  $\Phi$ , measurements  $y$ , structured sparse approximation algorithm  $\mathbb{M}_K$ Output:  $K$ -sparse approximation  $\hat{x}$  to true signal  $x$  $\hat{x}_0 = 0$ ,  $d = y$ ;  $i = 0$  {initialize}**while** halting criterion false **do**1.  $i \leftarrow i + 1$ 2.  $b \leftarrow \hat{x}_{i-1} + \Phi^T d$  {form signal estimate}3.  $\hat{x}_i \leftarrow \mathbb{M}(b, K)$  {prune residual estimate according to structure}4.  $d \leftarrow y - \Phi \hat{x}_i$  {update measurement residual}**end while**return  $\hat{x} \leftarrow \hat{x}_i$ *Proof of Lemma 5:* It can be shown [11] that

$$\|x - b\|_2 \leq \|x|_{\Lambda^C}\|_2 + \|(\Phi_\Lambda^T \Phi_\Lambda)^{-1} \Phi_\Lambda^T \Phi x|_{\Pi^C}\|_2 + \|\Phi_\Pi^\dagger n\|_2.$$

Since  $\Lambda$  is a support corresponding to a subspace in  $\mathcal{M}_K^3$  and  $x \in \mathcal{M}_K$ , we use Lemmas 1 and 2 to obtain

$$\begin{aligned} \|x - b\|_2 &\leq \|x|_{\Lambda^C}\|_2 + \frac{\|\Phi_\Lambda^T \Phi x|_{\Pi^C}\|_2}{1 - \delta_{\mathcal{M}_K^3}} + \frac{\|n\|_2}{\sqrt{1 - \delta_{\mathcal{M}_K^3}}}, \\ &\leq \left(1 + \frac{\delta_{\mathcal{M}_K^4}}{1 - \delta_{\mathcal{M}_K^3}}\right) \|x|_{\Pi^C}\|_2 + \frac{\|n\|_2}{\sqrt{1 - \delta_{\mathcal{M}_K^3}}}. \end{aligned}$$

Finally, note that  $\delta_{\mathcal{M}_K^3} \leq \delta_{\mathcal{M}_K^4} \leq 0.1$ .  $\square$ *Lemma 6: (Pruning)* The pruned approximation  $\hat{x}_i = \mathbb{M}(b, K)$  is such that

$$\|x - \hat{x}_i\|_2 \leq 2\|x - b\|_2.$$

*Proof of Lemma 6:* Since  $\hat{x}_i$  is the best approximation in  $\mathcal{M}_K$  to  $b$ , and  $x \in \mathcal{M}_K$ , we obtain

$$\|x - \hat{x}_i\|_2 \leq \|x - b\|_2 + \|b - \hat{x}_i\|_2 \leq 2\|x - b\|_2. \quad \square$$

We use these lemmas in reverse sequence for the inequalities below:

$$\begin{aligned} \|x - \hat{x}_i\|_2 &\leq 2\|x - b\|_2, \\ &\leq 2(1.112\|x|_{\Lambda^C}\|_2 + 1.06\|n\|_2), \\ &\leq 2.224\|s|_{\Omega^C}\|_2 + 2.12\|n\|_2, \\ &\leq 2.224(0.2223\|s\|_2 + 2.34\|n\|_2) + 2.12\|n\|_2, \\ &\leq 0.5\|s\|_2 + 7.5\|n\|_2, \\ &\leq 0.5\|x - \hat{x}_{i-1}\|_2 + 7.5\|n\|_2. \end{aligned}$$

From the recursion on  $\hat{x}_i$ , we obtain  $\|x - \hat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 15\|n\|_2$ . This completes the proof of Theorem 4.  $\square$ APPENDIX E  
PROOF OF PROPOSITION 1When  $K < \log_2 N$ , the number of subtrees of size  $K$  of a binary tree of size  $N$  is the Catalan number [49]

$$T_{K,N} = \frac{1}{K+1} \binom{2K}{K} \leq \frac{(2e)^K}{K+1},$$

using Stirling's approximation. When  $K > \log_2 N$ , we partition this count of subtrees into the numbers of subtrees  $t_{K,h}$  of size  $K$  and height  $h$ , to obtain

$$T_{K,N} = \sum_{h=\lfloor \log_2 K \rfloor + 1}^{\log_2 N} t_{K,h}$$

We obtain the following asymptotic identity from [49, page 51]:

$$\begin{aligned} t_{K,h} &= \frac{4^{K+1.5}}{h^4} \sum_{m \geq 1} \left[ \frac{2K}{h^2} (2\pi m)^4 - 3(2\pi m)^2 \right] e^{-\frac{K(2\pi m)^2}{h^2}} \\ &\quad + 4^K \mathcal{O}\left(e^{-\ln^2 h}\right) + 4^K \mathcal{O}\left(\frac{\ln^8 h}{h^5}\right) + 4^K \mathcal{O}\left(\frac{\ln^8 h}{h^4}\right), \\ &\leq \frac{4^{K+2}}{h^4} \sum_{m \geq 1} \left[ \frac{2K}{h^2} (2\pi m)^4 - 3(2\pi m)^2 \right] e^{-\frac{K(2\pi m)^2}{h^2}}. \quad (24) \end{aligned}$$

We now simplify the formula slightly: we seek a bound for the sum term (which we denote by  $\beta_h$  for brevity):

$$\begin{aligned} \beta_h &= \sum_{m \geq 1} \left[ \frac{2K}{h^2} (2\pi m)^4 - 3(2\pi m)^2 \right] e^{-\frac{K(2\pi m)^2}{h^2}} \\ &\leq \sum_{m \geq 1} \frac{2K}{h^2} (2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}}. \quad (25) \end{aligned}$$

Let  $m_{\max} = \frac{h}{\pi\sqrt{2K}}$ , the value of  $m$  for which the term inside the sum (25) is maximum; this is not necessarily an integer. Then,

$$\begin{aligned} \beta_h &\leq \sum_{m=1}^{\lfloor m_{\max} \rfloor - 1} \frac{2K}{h^2} (2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}} \\ &\quad + \sum_{m=\lfloor m_{\max} \rfloor}^{\lceil m_{\max} \rceil} \frac{2K}{h^2} (2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}} \\ &\quad + \sum_{m \geq \lfloor m_{\max} \rfloor + 1} \frac{2K}{h^2} (2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}}, \\ &\leq \int_1^{\lfloor m_{\max} \rfloor} \frac{2K}{h^2} (2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}} dx \\ &\quad + \sum_{m=\lfloor m_{\max} \rfloor}^{\lceil m_{\max} \rceil} \frac{2K}{h^2} (2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}} \\ &\quad + \int_{\lfloor m_{\max} \rfloor}^{\infty} \frac{2K}{h^2} (2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}} dx, \end{aligned}$$

where the second inequality comes from the fact that the series in the sum is strictly increasing for  $m \leq \lfloor m_{\max} \rfloor$  and strictly decreasing for  $m > \lfloor m_{\max} \rfloor$ . One of the terms in the sum can be added to one of the integrals. If we have that

$$(2\pi \lfloor m_{\max} \rfloor)^4 e^{-\frac{K(2\pi \lfloor m_{\max} \rfloor)^2}{h^2}} < (2\pi \lceil m_{\max} \rceil)^4 e^{-\frac{K(2\pi \lceil m_{\max} \rceil)^2}{h^2}}, \quad \text{we then obtain} \quad (26)$$

then we can obtain

$$\begin{aligned} \beta_h &\leq \int_1^{\lceil m_{\max} \rceil} \frac{2K}{h^2} (2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}} dx \\ &\quad + \frac{2K}{h^2} (2\pi \lceil m_{\max} \rceil)^4 e^{-\frac{K(2\pi \lceil m_{\max} \rceil)^2}{h^2}} \\ &\quad + \int_{\lceil m_{\max} \rceil}^{\infty} \frac{2K}{h^2} (2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}} dx. \end{aligned}$$

When the opposite of (26) is true, we have that

$$\begin{aligned} \beta_h &\leq \int_1^{\lfloor m_{\max} \rfloor} \frac{2K}{h^2} (2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}} dx \\ &\quad + \frac{2K}{h^2} (2\pi \lfloor m_{\max} \rfloor)^4 e^{-\frac{K(2\pi \lfloor m_{\max} \rfloor)^2}{h^2}} \\ &\quad + \int_{\lfloor m_{\max} \rfloor}^{\infty} \frac{2K}{h^2} (2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}} dx. \end{aligned}$$

Since the term in the sum reaches its maximum for  $m_{\max}$ , we will have in all three cases that

$$\beta_h \leq \int_1^{\infty} \frac{2K}{h^2} (2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}} dx + \frac{8h^2}{Ke^2}.$$

We perform a change of variables  $u = 2\pi x$  and define  $\sigma = h/\sqrt{2K}$  to obtain

$$\begin{aligned} \beta_h &\leq \frac{1}{2\pi} \int_0^{\infty} \frac{1}{\sigma^2} u^4 e^{-u^2/2\sigma^2} dx + \frac{8h^2}{Ke^2} \\ &\leq \frac{1}{2\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} u^4 e^{-u^2/2\sigma^2} dx + \frac{8h^2}{Ke^2}. \end{aligned}$$

Using the formula for the fourth central moment of a Gaussian distribution:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} u^4 e^{-u^2/2\sigma^2} dx = 3\sigma^4,$$

we obtain

$$\beta_h \leq \frac{3\sigma^3}{2\sqrt{2\pi}} + \frac{8h^2}{Ke^2} = \frac{3h^3}{8\sqrt{\pi K^3}} + \frac{8h^2}{Ke^2}.$$

Thus, (24) simplifies to

$$t_{K,h} \leq \frac{4^K}{K} \left( \frac{6}{h\sqrt{\pi K}} + \frac{128}{h^2 e^2} \right).$$

Correspondingly,  $T_{K,N}$  becomes

$$\begin{aligned} T_{K,N} &\leq \sum_{h=\lfloor \log_2 K \rfloor + 1}^{\log_2 N} \frac{4^K}{K} \left( \frac{6}{h\sqrt{\pi K}} + \frac{128}{h^2 e^2} \right), \\ &\leq \frac{4^K}{K} \left( \frac{6}{\sqrt{\pi K}} \sum_{h=\lfloor \log_2 K \rfloor + 1}^{\log_2 N} \frac{1}{h} \right. \\ &\quad \left. + \frac{128}{e^2} \sum_{h=\lfloor \log_2 K \rfloor + 1}^{\log_2 N} \frac{1}{h^2} \right). \end{aligned}$$

It is easy to show, using Euler-Maclaurin summations, that

$$\sum_{j=a}^b j^{-1} \leq \ln \frac{b}{a-1} \quad \text{and} \quad \sum_{j=a}^b j^{-2} \leq \frac{1}{a-1};$$

$$\begin{aligned} T_{K,N} &\leq \frac{4^K}{K} \left( \frac{6}{\sqrt{\pi K}} \ln \frac{\log_2 N}{\lfloor \log_2 K \rfloor} + \frac{128}{e^2 \lfloor \log_2 K \rfloor} \right) \\ &\leq \frac{4^{K+4}}{Ke^2 \lfloor \log_2 K \rfloor} \leq \frac{4^{K+4}}{Ke^2}. \end{aligned}$$

This proves the proposition.  $\square$

## APPENDIX F PROOF OF PROPOSITION 3

We wish to find the value of the bound (12) for the subspace count given in (17). We obtain  $M \geq \max_{1 \leq j \leq \lceil N/K \rceil} M_j$ , where

$$\begin{aligned} M_j &= \frac{1}{(j^r \sqrt{1 + \epsilon_K} - 1)^2} \\ &\quad \left( 2K + 4 \ln \frac{(2e)^{K(2j+1)N}}{K(Kj+1)(Kj+K+1)} + 2t \right). \end{aligned}$$

We separate the terms that are linear on  $K$  and  $j$ , and obtain

$$\begin{aligned} M_j &= \frac{1}{(j^r \sqrt{1 + \epsilon_K} - 1)^2} \left( K(3 + 4 \ln 2) + 8Kj(1 + \ln 2) \right. \\ &\quad \left. + 4 \ln \frac{N}{K(Kj+1)(Kj+K+1)} + 2t \right), \\ &= \frac{1}{(j^{s-0.5} \sqrt{1 + \epsilon_K} - j^{-0.5})^2} \\ &\quad \left( 8K(1 + \ln 2) + \frac{K(3 + 4 \ln 2)}{j} \right. \\ &\quad \left. + \frac{4}{j} \ln \frac{N}{K(Kj+1)(Kj+K+1)} + \frac{2t}{j} \right). \end{aligned}$$

The sequence  $\{M_j\}_{j=1}^{\lceil \frac{N}{K} \rceil}$  is a decreasing sequence, since the denominators are decreasing sequences whenever  $s > 0.5$ . We then have

$$\begin{aligned} M &\geq \frac{1}{(\sqrt{1 + \epsilon_K} - 1)^2} \left( K(11 + 12 \ln 2) \right. \\ &\quad \left. + 4 \ln \frac{N}{K(K+1)(2K+1)} + 2t \right). \end{aligned}$$

This completes the proof of Proposition 3.  $\square$

## ACKNOWLEDGEMENTS

We thank Petros Boufounos, Mark Davenport, Yonina Eldar, Moshe Mishali, and Robert Nowak for helpful discussions.

## REFERENCES

- [1] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1999.
- [2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, Sept. 2006.
- [3] E. J. Candès, "Compressive sampling," in *Proc. International Congress of Mathematicians*, Madrid, Spain, 2006, vol. 3, pp. 1433–1452.
- [4] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Mag.*, vol. 24, no. 4, pp. 118–120, 124, July 2007.
- [5] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, "Sparse signal recovery using Markov Random Fields," in *Proc. Workshop on Neural Info. Proc. Sys. (NIPS)*, Vancouver, Canada, Dec. 2008.
- [6] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Info. Theory*, vol. 55, no. 4, pp. 1872–1882, Apr. 2009.
- [7] Y. M. Lu and M. N. Do, "Sampling signals from a union of subspaces," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 41–47, Mar. 2008.
- [8] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Processing*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.
- [9] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Info. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.
- [10] D. Baron, M. F. Duarte, S. Sarvotham, M. B. Wakin, and R. G. Baraniuk, "Distributed compressive sensing," 2005, Preprint.
- [11] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, May 2009.
- [12] M.A.T. Figueiredo and R.D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Processing*, vol. 12, no. 8, pp. 906–916, 2003.
- [13] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.
- [14] Emmanuel J. Candès and Justin K. Romberg, "Signal recovery from random projections," in *Proc. Computational Imaging III at SPIE Electronic Imaging*, San Jose, CA, Jan. 2005, vol. 5674, pp. 76–86.
- [15] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 629–654, Dec. 2008.
- [16] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, Nov. 2009.
- [17] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 886–902, Apr. 1998.
- [18] R. G. Baraniuk, "Fast reconstruction from incoherent projections," Workshop on Sparse Representations in Redundant Systems, May 2005.
- [19] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk, "Fast reconstruction of piecewise smooth signals from random projections," in *Proc. SPARS05*, Rennes, France, Nov. 2005.
- [20] C. La and M. N. Do, "Tree-based orthogonal matching pursuit algorithm for signal reconstruction," in *IEEE International Conference on Image Processing (ICIP)*, Atlanta, GA, Oct. 2006, pp. 1277–1280.
- [21] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk, "Wavelet-domain compressive signal reconstruction using a hidden Markov tree model," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, April 2008, pp. 5137–5140.
- [22] M. N. Do and C. N. H. La, "Tree-based majorize-minimize algorithm for compressed sensing with sparse-tree prior," in *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, Saint Thomas, US Virgin Islands, Dec. 2007, pp. 129–132.
- [23] L. He and L. Carin, "Exploiting structure in wavelet-based Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 57, no. 9, pp. 3488–3497, Sep. 2009.
- [24] K. Lee and Y. Bresler, "Selecting good Fourier measurements for compressed sensing," SIAM Conference on Imaging Science, July 2008.
- [25] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Uniform uncertainty principle for Bernoulli and subgaussian ensembles," *Constructive Approximation*, vol. 28, no. 3, pp. 277–289, Dec. 2008.
- [26] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computation*, vol. 24, no. 2, pp. 227–234, Apr. 1995.
- [27] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33, 1998.
- [28] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Info. Theory*, vol. 52, no. 9, pp. 4036–4048, Sept. 2006.
- [29] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ ," *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.
- [30] J. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Info. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
- [31] D. L. Donoho, I. Drori, Y. Tsaig, and J. L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," 2006, Preprint.
- [32] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing: Closing the gap between performance and complexity," *IEEE Trans. Info. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.
- [33] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences, Series I*, vol. 346, no. 9–10, pp. 589–592, May 2008.
- [34] R. G. Baraniuk and D. L. Jones, "A signal-dependent time-frequency representation: Fast algorithm for optimal kernel design," *IEEE Trans. Signal Processing*, vol. 42, no. 1, pp. 134–146, Jan. 1994.
- [35] R. G. Baraniuk, "Optimal tree approximation with wavelets," in *Wavelet Applications in Signal and Image Processing VII*, Denver, CO, July 1999, vol. 3813 of *Proc. SPIE*, pp. 196–207.
- [36] R. G. Baraniuk, R. A. DeVore, G. Kyriazis, and X. M. Yu, "Near best tree approximation," *Advances in Computational Mathematics*, vol. 16, no. 4, pp. 357–373, May 2002.
- [37] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," *IEEE Trans. Image Processing*, vol. 10, no. 7, pp. 1056–1068, July 2001.
- [38] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using a scale mixture of Gaussians in the wavelet domain," *IEEE Trans. Image Processing*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.
- [39] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.
- [40] A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore, "Tree approximation and optimal encoding," *Applied and Computational Harmonic Analysis*, vol. 11, no. 2, pp. 192–226, Sept. 2001.
- [41] D. Donoho, "CART and best ortho-basis: A connection," *Annals of Statistics*, vol. 25, no. 5, pp. 1870–1911, Oct. 1997.
- [42] M. B. Wakin, S. Sarvotham, M. F. Duarte, D. Baron, and R. G. Baraniuk, "Recovery of jointly sparse signals from few random projections," in *Proc. Workshop on Neural Info. Proc. Sys. (NIPS)*, Vancouver, Nov. 2005.
- [43] J. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, Apr. 2006.
- [44] M. F. Duarte, V. Cevher, and R. G. Baraniuk, "Model-based compressive Sensing for Signal Ensembles," in *Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Oct. 2009.
- [45] C. Hegde, M. F. Duarte, and V. Cevher, "Compressive sensing recovery of spike trains using a structured sparsity model," in *Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, Saint Malo, France, Apr. 2009.
- [46] V. Cevher, P. Indyk, C. Hegde, and R. G. Baraniuk, "Recovery of clustered sparse signals from compressive measurements," in *Int. Conf. on Sampling Theory and Applications (SAMPTA)*, Marseille, France, May 2009.
- [47] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Info. Theory*, vol. 51, pp. 4203–4215, Dec. 2005.
- [48] M. Ledoux, *The Concentration of Measure Phenomenon*, American Mathematical Society, 2001.
- [49] G. G. Brown and B. O. Shubert, "On random binary trees," *Mathematics of Operations Research*, vol. 9, no. 1, pp. 43–65, Feb. 1984.