Distributed Compressed Sensing

Dror Baron, Michael B. Wakin, Marco F. Duarte,
Shriram Sarvotham, and Richard G. Baraniuk[*]

Department of Electrical and Computer Engineering
Rice University
Houston, TX 77005, USA

November 27, 2005

Abstract

Compressed sensing is an emerging field based on the revelation that a small collection of linear projections of a sparse signal contains enough information for reconstruction. In this paper we introduce a new theory for *distributed compressed sensing* (DCS) that enables new distributed coding algorithms for multi-signal ensembles that exploit both intra- and inter-signal correlation structures. The DCS theory rests on a new concept that we term the *joint sparsity* of a signal ensemble. We study in detail three simple models for jointly sparse signals, propose algorithms for joint recovery of multiple signals from incoherent projections, and characterize theoretically and empirically the number of measurements per sensor required for accurate reconstruction. We establish a parallel with the Slepian-Wolf theorem from information theory and establish upper and lower bounds on the measurement rates required for encoding jointly sparse signals. In two of our three models, the results are asymptotically best-possible, meaning that both the upper and lower bounds match the performance of our practical algorithms. Moreover, simulations indicate that the asymptotics take effect with just a moderate number of signals. In some sense DCS is a framework for distributed compression of sources with memory, which has remained a challenging problem for some time. DCS is immediately applicable to a range of problems in sensor networks and arrays.

Keywords: Compressed sensing, distributed source coding, sparsity, incoherent projections, random matrices, linear programming, sensor networks, array processing.

1 Introduction

A core tenet of signal processing and information theory is that signals, images, and other data often contain some type of *structure* that enables intelligent representation and processing. The notion of structure has been characterized and exploited in a variety of ways for a variety of purposes. In this paper, we focus on exploiting signal *correlations* for the purpose of *compression*.

Current state-of-the-art compression algorithms employ a decorrelating transform such as an exact or approximate Karhunen-Loève transform (KLT) to compact a correlated signal's energy into just a few essential coefficients [4–6]. Such *transform coders* exploit the fact that many signals

[*] This work was supported by NSF-CCF, NSF-NeTS, ONR, and AFOSR. Preliminary versions of this work have appeared at the 43rd Allerton Conference on Communication, Control, and Computing [1], the 39th Asilomar Conference on Signals, Systems and Computers [2], and the 19th Conference on Neural Information Processing Systems [3]. Email: {drorb, wakin, duarte, shri, richb}@rice.edu; Web: dsp.rice.edu/cs

have a *sparse* representation in terms of some basis, meaning that a small number K of adaptively chosen transform coefficients can be transmitted or stored rather than $N \gg K$ signal samples. For example, smooth signals are sparse in the Fourier basis, and piecewise smooth signals are sparse in a wavelet basis [7]; the commercial coding standards MP3 [8], JPEG [9], and JPEG2000 [10] directly exploit this sparsity.

1.1 Distributed source coding

While the theory and practice of compression have been well developed for individual signals, many applications involve multiple signals, for which there has been less progress. As a motivating example, consider a *sensor network*, in which a potentially large number of distributed sensor nodes can be programmed to perform a variety of data acquisition tasks as well as to network themselves to communicate their results to a central collection point [11, 12]. In many sensor networks, and in particular battery-powered ones, communication energy and bandwidth are scarce resources; both factors make the reduction of communication critical.

Fortunately, since the sensors presumably observe related phenomena, the ensemble of signals they acquire can be expected to possess some joint structure, or *inter-signal correlation*, in addition to the *intra-signal correlation* in each individual sensor's measurements. For example, imagine a microphone network recording a sound field at several points in space. The time-series acquired at a given sensor might have considerable intra-signal (temporal) correlation and might be sparsely represented in a local Fourier basis. In addition, the ensemble of time-series acquired at all sensors might have considerable inter-signal (spatial) correlation, since all microphones listen to the same sources. In such settings, *distributed source coding* that exploits both intra- and inter-signal correlations might allow the network to save on the communication costs involved in exporting the ensemble of signals to the collection point [13–17].

A number of distributed coding algorithms have been developed that involve collaboration amongst the sensors, including several based on predictive coding [18–20], a distributed KLT [21], and distributed wavelet transforms [22, 23]. Three-dimensional wavelets have been proposed to exploit both inter- and intra-signal correlations [24]. Note, however, that any collaboration involves some amount of inter-sensor communication overhead.

In the *Slepian-Wolf* framework for lossless distributed coding [13–17], the availability of correlated side information at the collection point / decoder enables each sensor node to communicate losslessly at its conditional entropy rate rather than at its individual entropy rate. Slepian-Wolf coding has the distinct advantage that the sensors need not collaborate while encoding their measurements, which saves valuable communication overhead. Unfortunately, however, most existing coding algorithms [15, 16] exploit only inter-signal correlations and not intra-signal correlations. To date there has been only limited progress on distributed coding of so-called "sources with memory." (We briefly mention some limitations here and elaborate in Section 2.1.3.) The direct implementation for such sources would require huge lookup tables [13, 25]. Furthermore, approaches combining pre- or post-processing of the data to remove intra-signal correlations combined with Slepian-Wolf coding for the inter-signal correlations appear to have limited applicability. Finally, although a recent paper by Uyematsu [26] provides compression of spatially correlated sources with memory, the solution is specific to lossless distributed compression and cannot be readily extended to lossy compression setups. We conclude that the design of constructive techniques for distributed coding of sources with both intra- and inter-signal correlation is still an open and challenging problem with many potential applications.

1.2 Compressed sensing (CS)

A new framework for single-signal sensing and compression has developed recently under the rubric of *Compressed Sensing* (CS). CS builds on the ground-breaking work of Candès, Romberg, and Tao [27] and Donoho [28], who showed that if a signal has a sparse representation in one basis then it can be recovered from a small number of projections onto a second basis that is *incoherent* with the first.[1] In fact, for an N-sample signal that is K-sparse,[2] only $K + 1$ projections of the signal onto the incoherent basis are required to reconstruct the signal with high probability (Theorem 2). Unfortunately, this requires a combinatorial search, which is prohibitively complex. Candès et al. [27] and Donoho [28] have recently proposed tractable recovery procedures based on linear programming, demonstrating the remarkable property that such procedures provide the same result as the combinatorial search as long as cK projections are used to reconstruct the signal (typically $c \approx 3$ or 4) [31–33]. Iterative greedy algorithms have also been proposed [34–36], allowing even faster reconstruction at the expense of slightly more measurements.

The implications of CS are promising for many applications, especially sensing signals that have a sparse representation in some basis. Instead of sampling a K-sparse signal N times, only cK incoherent measurements suffice, where K can be orders of magnitude less than N. (For example, Takhar et al. [37] develop a camera that dispenses with the usual N-pixel CCD or CMOS imaging array by computing cK incoherent image projections optically using a digital micromirror device.) Therefore, a sensor can transmit far fewer measurements to a receiver, which can reconstruct the signal and then process it in any manner. Moreover, the cK measurements need not be manipulated in any way before being transmitted, except possibly for some quantization. Finally, independent and identically distributed (i.i.d.) Gaussian or Bernoulli/Rademacher (random ± 1) vectors provide a useful *universal* basis that is incoherent with all others.[3] Hence, when using a random basis, CS is universal in the sense that the sensor can apply the same measurement mechanism no matter what basis the signal is sparse in (and thus the coding algorithm is independent of the sparsity-inducing basis) [28, 29].

While powerful, the CS theory at present is designed mainly to exploit intra-signal structures at a single sensor. To the best of our knowledge, the only work to date that applies CS in a multi-sensor setting is Haupt and Nowak [38] (see Section 2.2.6). However, while their scheme exploits inter-signal correlations, it ignores intra-signal correlations.

1.3 Distributed compressed sensing (DCS)

In this paper we introduce a new theory for *distributed compressed sensing* (DCS) that enables new distributed coding algorithms that exploit both intra- and inter-signal correlation structures. In a typical DCS scenario, a number of sensors measure signals (of any dimension) that are each individually sparse in some basis and also correlated from sensor to sensor. Each sensor *independently* encodes its signal by projecting it onto another, incoherent basis (such as a random one) and then transmits just a few of the resulting coefficients to a single collection point. Under the right conditions, a decoder at the collection point can reconstruct each of the signals precisely.

The DCS theory rests on a concept that we term the *joint sparsity* of a signal ensemble. We will study in detail three simple models for jointly sparse signals, propose tractable algorithms for joint recovery of signal ensembles from incoherent projections, and characterize theoretically and empirically the number of measurements per sensor required for accurate reconstruction. While

[1] Roughly speaking, *incoherence* means that no element of one basis has a sparse representation in terms of the other basis. This notion has a variety of formalizations in the CS literature [27–30].

[2] By K-sparse, we mean that the signal can be written as a sum of K basis functions from some known basis.

[3] Since the "incoherent" measurement vectors must be known for signal recovery, in practice one may use a pseudorandom basis with a known random seed.

the sensors operate entirely without collaboration, we will see in many cases that the measurement rates relate directly to the signals' *conditional sparsities*, in parallel with the Slepian-Wolf theory. The joint sparsity models (JSMs) we study are as follows.

JSM-1: Sparse common component + innovations: In this model each signal consists of a sum of two components: a *common* component that is present in all of the signals and an *innovations* component that is unique to each signal. Both the common and innovations components are sparsely representable in some basis. Such signals may arise in settings where large-scale phenomena affect all sensors and local phenomena affect individual sensors; one example would be a network of temperature sensors in a forest, where the sun has a global effect, and shade, water, and animals have more local effects.

For JSM-1, we will show that there exists a *measurement rate region* analogous to the Slepian-Wolf rate region for distributed coding [14] (see Figure 6). The notion of joint sparsity suggests a joint reconstruction technique based on linear programming. We provide a converse bound (Theorem 6) and an achievable bound (Theorem 7) on the measurement rate region using linear programming techniques.

Our simulations reveal that in practice the savings in the total number of required measurements can be substantial over separate CS encoding/decoding, especially when the common component dominates. In one of our scenarios with just two sensors, the savings in the number of measurements can be as large as 30% (Theorem 7). Detailed numerical results appear in Section 4.7.

JSM-2: Common sparse supports: In this model, all signals are constructed from the same sparse set of basis vectors, but with different coefficient values. Examples of JSM-2 scenarios include MIMO communication [34] and audio signal arrays; the signals may be sparse in the Fourier domain, for example, yet multipath resulting from differing propagation paths causes different attenuations among the frequency components. (Note that while all sensors may be "listening" to the same underlying signal, in applications such as localization and beamforming it can be important to recover all of the individual signals and not just a single composite signal.)

We develop two techniques based on iterative greedy pursuit for signal ensemble reconstruction from independent, incoherent measurements. Our analysis (Theorem 9) and simulations (in Section 5.3 and Figure 10) indicate that as the number of sensors grows, the oversampling factor c required for exact reconstruction of all signals shrinks to $c = 1$. Since an "oracle" system that knows in advance the positions of the sparse basis vectors also requires $c = 1$ (Theorem 9), our DCS encoder/decoder provides the best-possible performance. From an information theoretic perspective, for JSM-2 we have tight converse and achievable measurement rate bounds. Our simulations indicate that the asymptotics take effect with even a moderate number of signals.

JSM-3: Nonsparse common component + sparse innovations: This model extends JSM-1 so that the common component need no longer be sparse in any basis. Since the common component is not sparse, no individual signal contains enough structure to permit efficient compression or CS; in general N measurements would be required for each individual N-sample signal. We demonstrate, however, that the common structure shared by the signals permits a dramatic reduction in the required measurement rates. In fact, asymptotically, the required measurement rates relate simply to the sparsity K of the innovation components; as the number of sensors grows, each sensor may again reduce its oversampling factor to $c = 1$ (Theorem 10). Again, this is best-possible performance that could not be bettered by an oracle that knew the common nonsparse component in advance.

1.4 Advantages of DCS

In addition to offering substantially reduced measurement rates in multi-signal applications, the DCS-based distributed source coding schemes we develop here share many of the attractive and intriguing properties of CS, particularly when we employ random projections at the sensors. As in single-signal CS, random measurement bases are *universal* in the sense that they can be paired with any sparse basis. This allows exactly the same encoding strategy to be applied in a variety of different sensing environments; knowledge of the nuances of the environment are needed only at the decoder. Moreover, random measurements are also *future-proof*: if a better sparsity-inducing basis is found for the signals, then the same random measurements can be used to reconstruct an even more accurate view of the environment. A pseudorandom basis can be generated using a simple algorithm according to a random seed. Such encoding effectively implements a form of *encryption*: the randomized measurements will themselves resemble noise and be meaningless to an observer who does not know the associated seed. Random coding is also *robust*: the randomized measurements coming from each sensor have equal priority, unlike the Fourier or wavelet coefficients in current coders. Thus they allow a *progressively better reconstruction* of the data as more measurements are obtained; one or more measurements can also be lost without corrupting the entire reconstruction.

Two additional properties of DCS make it well-matched to distributed applications such as sensor networks and arrays [11, 12]. First, each sensor encodes its measurements independently, which reduces inter-sensor communication overhead to zero. Second, DCS distributes its computational complexity asymmetrically, placing most of it in the joint decoder, which will often have more substantial computational resources than any individual sensor node. The encoders are very simple; they merely compute incoherent projections with their signals and make no decisions.

This paper focuses primarily on the basic task of reducing the *measurement rate* of a signal ensemble in order to reduce the communication cost of source coding that ensemble. In practical settings (such as sensor networks), additional criteria may be relevant for measuring performance. For example, the measurements will typically be real numbers that must be quantized and encoded, which will gradually degrade the reconstruction quality as the quantization becomes coarser [39] (see also Section 7). Characterizing DCS in light of practical considerations such as rate-distortion tradeoffs, power consumption in sensor networks, etc., are topics of ongoing research [40].

1.5 Paper organization

Section 2 overviews the distributed source coding and single-signal CS theories and provides two new results on CS reconstruction. While readers may be familiar with some of this material, we include it to make the paper self-contained. Section 3 introduces our three models for joint sparsity: JSM-1, 2, and 3. We provide our detailed analysis and simulation results for these models in Sections 4, 5, and 6, respectively. We close the paper with a discussion and conclusions in Section 7. Several appendices contain the proofs and other mathematical details.

2 Background

2.1 Information theory

2.1.1 Lossless source coding

In a typical lossless coding scenario, we have a *sequence* $x = x(1), x(2), \ldots, x(N)$ of N *symbols*, $n \in \{1, 2, \ldots, N\}$, where each symbol $x(n)$ belongs to a finite alphabet $\mathcal{X}$. Our goal is to encode x using bits in such a way that we can reconstruct x perfectly[4] at the decoder. In order to represent

[4] In this paper we use the terms *perfect* and *exact* interchangeably.

the sequence x, a straightforward approach is to represent each symbol $x(n)$ using $\lceil \log_2(|\mathcal{X}|) \rceil$ bits, where $|\cdot|$ denotes the cardinality of a set, $\log_2(\cdot)$ is the base-two logarithm, and $\lceil \cdot \rceil$ rounds up to the nearest integer.

The sequence x is often modeled in a way that enables a more compact representation of the sequence. This is called *compression.* To elaborate further, we now describe a standard setting for lossless source coding. Consider a *source* X that generates such sequences, where the symbols of x are i.i.d., and the probability mass function assigns each symbol $x(n) \in \mathcal{X}$ a probability $p(x(n))$. The key idea in lossless compression is to represent each symbol $x(n)$ that has probability p using $-\log_2(p)$ bits. Using this insight, the *entropy* $H(X)$ is defined as

$$H(X) \triangleq - \sum_{x(n) \in \mathcal{X}} p(x(n)) \log_2(p(x(n))). \tag{1}$$

Not only can sequences be compressed close to their entropy using this insight [13], but it also turns out that the entropy provides the lowest per-symbol rate that enables lossless compression. Various techniques such as arithmetic coding [13] can be used to compress near the entropy rate.

2.1.2 Distributed source coding

Information theory [13, 17] has also provided tools that characterize the performance of distributed source coding. For correlated length-N sequences x_1 and x_2 generated by sources X_1 and X_2 over discrete alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$, we have entropies $H(X_1)$ and $H(X_2)$ as before in (1). The *joint entropy* of X_1 and X_2, which is the lowest rate that enables compression of x_1 and x_2 together, is defined as

$$H(X_1, X_2) \triangleq - \sum_{x_1(n) \in \mathcal{X}_1, x_2(n) \in \mathcal{X}_2} p(x_1(n), x_2(n)) \log_2(p(x_1(n), x_2(n))).$$

The extension to more than two signals is straightforward [13].

The *conditional entropy* is the lowest per-symbol rate that enables lossless compression, conditioned on the side information that is available at both encoder and decoder. More formally,

$$H(X_1|X_2) \triangleq - \sum_{x_1(n) \in \mathcal{X}_1, x_2(n) \in \mathcal{X}_2} p(x_1(n), x_2(n)) \log_2(p(x_1(n)|x_2(n))),$$

and it can be shown that

$$H(X_1) + H(X_2|X_1) = H(X_1, X_2) = H(X_2) + H(X_1|X_2).$$

If the sources X_1 and X_2 are independent, then $H(X_2|X_1) = H(X_2)$ and $H(X_1|X_2) = H(X_1)$, and so the joint entropy is the sum of the individual entropies $H(X_1)$ and $H(X_2)$. In this case, separate (independent) compression of each of the sequences x_1 and x_2 can achieve the optimal compression rate $H(X_1, X_2)$. However, if X_1 and X_2 are *correlated sources*, then the joint entropy satisfies

$$H(X_1, X_2) < H(X_1) + H(X_2),$$

meaning that the separate encoding of each sequence is wasteful. The potential savings in the coding rate in this setup motivated Slepian and Wolf to study the distributed coding of correlated sources [14].

In the Slepian-Wolf framework [13–17, 25], the sequences x_1 and x_2 are encoded separately and decoded jointly. The *rates* R_1 and R_2 at which we encode x_1 and x_2 are the normalized number of bits used per source symbol. For Slepian-Wolf coding, there is an entire *rate region* of rate pairs (R_1, R_2) that enable us to correctly reconstruct x_1 and x_2. This rate region is characterized in the following theorem.

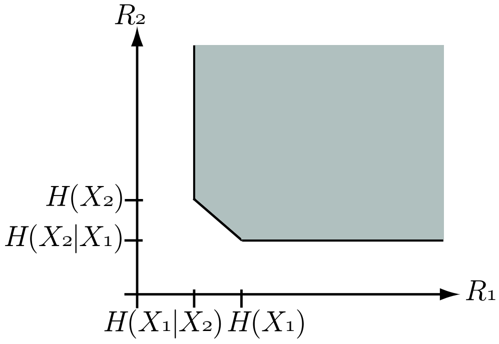


Figure 1: *The shaded area indicates the Slepian-Wolf achievable rate region for distributed source coding (Theorem 1).*

Theorem 1 [14] *Consider sources X_1 and X_2 that generate length-N sequences x_1 and x_2. The sequences are encoded separately using rates R_1 and R_2. As N increases, the sequences can be reconstructed jointly with vanishing probability of error if and only if*

$$\begin{aligned} R_1 &> H(X_1|X_2), && \text{(2a)} \\ R_2 &> H(X_2|X_1), && \text{(2b)} \\ R_1 + R_2 &> H(X_1, X_2). && \text{(2c)} \end{aligned}$$

The surprising result is that it suffices to encode each sequence above its conditional entropy as long as the sum rate exceeds the joint entropy. In contrast, separate encoding must encode each source at its entropy, and the sum rate is often greater than the joint entropy.

The Slepian-Wolf rate region [13, 14, 17, 25] — shown in Figure 1 — has been completely characterized: any rate pair (R_1, R_2) that satisfies the conditions (2a)–(2c) enables decoding of x_1 and x_2 with vanishing probability of error as N increases. This characterization is accomplished by providing converse and achievable results that are tight. The *converse* part of the analysis shows that for any rate pair for which the conditions do not hold, the probability of error does not vanish. The *achievable* part shows that for any rate pair that satisfies these conditions (2a)–(2c), there exist constructive schemes that enable us to reconstruct x_1 and x_2 with vanishing probability of error.

The constructive schemes usually used in these analyses are based on *random binning* [13, 25]. In this technique, every possible sequence x_1 is assigned a bin index

$$i(x_1) \in \left\{1, 2, \ldots, 2^{NR_1}\right\},$$

where the probability of the bin index assigned to any x_1 is uniform over all 2^{NR_1} possible indices. The other sequence x_2 is assigned an index $i(x_2)$ in an analogous manner. The *encoders* for x_1 and x_2 assign these indices and can thus encode x_1 and x_2 using NR_1 and NR_2 bits, respectively. The *decoders* search for a pair of sequences $(\widehat{x}_1, \widehat{x}_2)$ such that $i(x_1) = i(\widehat{x}_1)$, $i(x_2) = i(\widehat{x}_2)$, and the pair is *jointly typical.* Loosely speaking, joint typicality means that the sequences $\widehat{x}_1$ and $\widehat{x}_2$ match the joint statistics well. As long as the conditions (2a)–(2c) hold, the probability of error vanishes as N increases.

2.1.3 Challenges for distributed coding of sources with memory

One approach to distributed compression of data with both inter- and intra-signal correlations ("sources with memory") is to perform Slepian-Wolf coding using source models with temporal memory. Cover [25] showed how random binning can be applied to compress ergodic sources in a distributed manner. Unfortunately, implementing this approach would be challenging, since it requires maintaining lookup tables of size 2^{NR_1} and 2^{NR_2} at the two encoders. Practical Slepian-Wolf encoders are based on dualities to channel coding [15, 16] and hence do not require storing vast lookup tables.

An alternative approach would use a transform to remove intra-signal correlations. For example, the Burrows-Wheeler Transform (BWT) permutes the symbols of a block in a manner that removes correlation between temporal symbols and thus can be viewed as the analogue of the Karhunen-Lòeve transform for sequences over finite alphabets. The BWT handles temporal correlation efficiently in single-source lossless coding [41, 42]. For distributed coding, the BWT could be proposed to remove temporal correlations by pre-processing the sequences prior to Slepian-Wolf coding. Unfortunately, the BWT is input-dependent, and hence temporal correlations would be removed only if all sequences were available at the encoders. Using a transform as a post-processor following Slepian-Wolf coding does not seem promising either, since the distributed encoders' outputs will each be i.i.d.

In short, approaches based on separating source coding into two components — distributed coding to handle inter-signal correlations and a transform to handle intra-signal correlations — appear to have limited applicability. In contrast, a recent paper by Uyematsu [26] proposed a universal Slepian-Wolf scheme for correlated Markov sources. Uyematsu's approach constructs a sequence of universal codes such that the probability of decoding error vanishes when the coding rates lie within the Slepian-Wolf region. Such codes can be constructed algebraically, and the encoding/decoding complexity is $O(N^3)$. While some of the decoding schemes developed below have similar (or lower) complexity, they have broader applicability. First, we deal with continuous sources, whereas Uyematsu's work considers only finite alphabet sources. Second, quantization of the measurements will enable us to extend our schemes to lossy distributed compression, whereas Uyematsu's work is confined to lossless settings. Third, Uyematsu's work only considers Markov sources. In contrast, the use of different bases enables our approaches to process broader classes of jointly sparse signals.

2.2 Compressed sensing

2.2.1 Transform coding

Consider a length-N, real-valued signal x of any dimension (without loss of generality, we will focus on one dimension for notational simplicity) indexed as $x(n)$, $n \in \{1, 2, \ldots, N\}$. Suppose that the basis $\Psi = [\psi_1, \ldots, \psi_N]$ [7] provides a K-sparse representation of x; that is

$$x = \sum_{n=1}^{N} \theta(n)\, \psi_n = \sum_{\ell=1}^{K} \theta(n_\ell)\, \psi_{n_\ell},$$

where x is a linear combination of K vectors chosen from Ψ, $\{n_\ell\}$ are the indices of those vectors, and $\{\theta(n)\}$ are the coefficients; the concept is extendable to tight frames [7]. Alternatively, we can write in matrix notation

$$x = \Psi\theta,$$

where x is an $N \times 1$ column vector, the *sparse basis* matrix Ψ is $N \times N$ with the basis vectors ψ_n as columns, and θ is an $N \times 1$ column vector with K nonzero elements. Using $\|\cdot\|_p$ to denote the ℓ_p norm,[5] we can write that $\|\theta\|_0 = K$. Various expansions, including wavelets [7], Gabor bases [7],

[5] The ℓ_0 "norm" $\|\theta\|_0$ merely counts the number of nonzero entries in the vector θ.

curvelets [43], etc., are widely used for representation and compression of natural signals, images, and other data.

In this paper, we will focus on exactly K-sparse signals and defer discussion of the more general situation where the coefficients decay rapidly but not to zero (see Section 7 for additional discussion and [40] for DCS simulations on real-world compressible signals). The standard procedure for compressing sparse signals, known as *transform coding*, is to (*i*) acquire the full N-sample signal x; (*ii*) compute the complete set of transform coefficients $\{\theta(n)\}$; (*iii*) locate the K largest, significant coefficients and discard the (many) small coefficients; (*iv*) encode the *values and locations* of the largest coefficients.

This procedure has three inherent inefficiencies: First, for a high-dimensional signal, we must start with a large number of samples N. Second, the encoder must compute *all* of the N transform coefficients $\{\theta(n)\}$, even though it will discard all but K of them. Third, the encoder must encode the locations of the large coefficients, which requires increasing the coding rate since the locations change with each signal.

2.2.2 Incoherent projections

This raises a simple question: For a given signal, is it possible to directly estimate the set of large $\theta(n)$'s that will not be discarded? While this seems improbable, Candès, Romberg, and Tao [27, 29] and Donoho [28] have shown that a reduced set of projections can contain enough information to reconstruct sparse signals. An offshoot of this work, often referred to as *Compressed Sensing* (CS) [28, 29, 44–49], has emerged that builds on this principle.

In CS, we do not measure or encode the K significant $\theta(n)$ directly. Rather, we measure and encode $M < N$ projections $y(m) = \langle x, \phi_m^T \rangle$ of the signal onto a *second set* of basis functions $\{\phi_m\}, m = 1, 2, \ldots, M$, where ϕ_m^T denotes the transpose of ϕ_m and $\langle \cdot, \cdot \rangle$ denotes the inner product. In matrix notation, we measure

$$y = \Phi x,$$

where y is an $M \times 1$ column vector and the *measurement basis* matrix Φ is $M \times N$ with each row a basis vector ϕ_m. Since $M < N$, recovery of the signal x from the measurements y is ill-posed in general; however the additional assumption of signal *sparsity* makes recovery possible and practical.

The CS theory tells us that when certain conditions hold, namely that the basis $\{\phi_m\}$ cannot sparsely represent the elements of the basis $\{\psi_n\}$ (a condition known as *incoherence* of the two bases [27–30]) and the number of measurements M is large enough, then it is indeed possible to recover the set of large $\{\theta(n)\}$ (and thus the signal x) from a similarly sized set of measurements $\{y(m)\}$. This incoherence property holds for many pairs of bases, including for example, delta spikes and the sine waves of a Fourier basis, or the Fourier basis and wavelets. Significantly, this incoherence also holds with high probability between an arbitrary fixed basis and a randomly generated one. Signals that are sparsely represented in frames or unions of bases can be recovered from incoherent measurements in the same fashion.

2.2.3 Signal recovery via ℓ_0 optimization

The recovery of the sparse set of significant coefficients $\{\theta(n)\}$ can be achieved using *optimization* by searching for the signal with ℓ_0-sparsest coefficients $\{\theta(n)\}$ that agrees with the M observed measurements in y (recall that $M < N$). Reconstruction relies on the key observation that, given some technical conditions on Φ and Ψ, the coefficient vector θ is the solution to the ℓ_0 minimization

$$\widehat{\theta} = \arg\min \|\theta\|_0 \quad \text{s.t. } y = \Phi\Psi\theta \tag{3}$$

with overwhelming probability. (Thanks to the incoherence between the two bases, if the original signal is sparse in the θ coefficients, then no other set of sparse signal coefficients θ' can yield the same projections y.) We will call the columns of $\Phi\Psi$ the *holographic basis*.

In principle, remarkably few incoherent measurements are required to recover a K-sparse signal via ℓ_0 minimization. Clearly, more than K measurements must be taken to avoid ambiguity; the following theorem establishes that $K+1$ random measurements will suffice. The proof appears in Appendix A; similar results were established by Venkataramani and Bresler [50].

Theorem 2 *Let Ψ be an orthonormal basis for $\mathbb{R}^N$, and let $1 \leq K < N$. Then the following statements hold:*

1. *Let Φ be an $M \times N$ measurement matrix with i.i.d. Gaussian entries with $M \geq 2K$. Then with probability one the following statement holds: all signals $x = \Psi\theta$ having expansion coefficients $\theta \in \mathbb{R}^N$ that satisfy $\|\theta\|_0 = K$ can be recovered uniquely from the M-dimensional measurement vector $y = \Phi x$ via the ℓ_0 optimization (3).*

2. *Let $x = \Psi\theta$ such that $\|\theta\|_0 = K$. Let Φ be an $M \times N$ measurement matrix with i.i.d. Gaussian entries (notably, independent of x) with $M \geq K+1$. Then with probability one the following statement holds: x can be recovered uniquely from the M-dimensional measurement vector $y = \Phi x$ via the ℓ_0 optimization (3).*

3. *Let Φ be an $M \times N$ measurement matrix, where $M \leq K$. Then, aside from pathological cases (specified in the proof), no signal $x = \Psi\theta$ with $\|\theta\|_0 = K$ can be uniquely recovered from the M-dimensional measurement vector $y = \Phi x$.*

Remark 1 *The second statement of the theorem differs from the first in the following respect: when $K < M < 2K$, there will necessarily exist K-sparse signals x that cannot be uniquely recovered from the M-dimensional measurement vector $y = \Phi x$. However, these signals form a set of measure zero within the set of* all *K-sparse signals and can safely be avoided if Φ is randomly generated independently of x.*

The intriguing conclusion from the second and third statements of Theorem 2 is that one measurement separates the *achievable region*, where perfect reconstruction is possible with probability one, from the *converse region*, where with overwhelming probability reconstruction is impossible. Moreover, Theorem 2 provides a *strong converse measurement region* in a manner analogous to the strong channel coding converse theorems of Wolfowitz [17].

Unfortunately, solving this ℓ_0 optimization problem is prohibitively complex, requiring a combinatorial enumeration of the $\binom{N}{K}$ possible sparse subspaces. In fact, the ℓ_0-recovery problem is known to be NP-complete [31]. Yet another challenge is robustness; in the setting of Theorem 2, the recovery may be very poorly conditioned. In fact, *both* of these considerations (computational complexity and robustness) can be addressed, but at the expense of slightly more measurements.

2.2.4 Signal recovery via ℓ_1 optimization

The practical revelation that supports the new CS theory is that it is not necessary to solve the ℓ_0-minimization problem to recover the set of significant $\{\theta(n)\}$. In fact, a much easier problem yields an equivalent solution (thanks again to the incoherency of the bases); we need only solve for the ℓ_1-sparsest coefficients θ that agree with the measurements y [27–29, 44–48]

$$\widehat{\theta} = \arg\min \|\theta\|_1 \quad \text{s.t. } y = \Phi\Psi\theta. \tag{4}$$

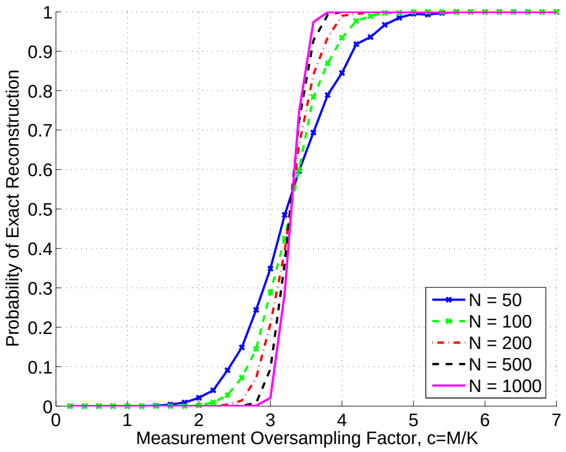


Figure 2: *Performance of Basis Pursuit for single-signal Compressed Sensing (CS) reconstruction. A signal x of normalized sparsity $S = K/N = 0.1$ and various lengths N is encoded in terms of a vector y containing $M = cK$ projections onto i.i.d. Gaussian random basis elements. The vertical axis indicates the probability that the linear program yields the correct answer x as a function of the oversampling factor $c = M/K$.*

This optimization problem, also known as *Basis Pursuit* [51], is significantly more approachable and can be solved with traditional linear programming techniques whose computational complexities are polynomial in N.

There is no free lunch, however; according to the theory, more than $K + 1$ measurements are required in order to recover sparse signals via Basis Pursuit. Instead, one typically requires $M \geq cK$ measurements, where $c > 1$ is an *oversampling factor*. As an example, we quote a result asymptotic in N. For simplicity, we assume that the sparsity scales linearly with N; that is, $K = SN$, where we call S the *sparsity rate*.

Theorem 3 [31–33] *Set $K = SN$ with $0 < S \ll 1$. Then there exists an oversampling factor $c(S) = O(\log(1/S))$, $c(S) > 1$, such that, for a K-sparse signal x in basis Ψ, the following statements hold:*

1. *The probability of recovering x via Basis Pursuit from $(c(S)+\epsilon)K$ random projections, $\epsilon > 0$, converges to one as $N \to \infty$.*

2. *The probability of recovering x via Basis Pursuit from $(c(S)-\epsilon)K$ random projections, $\epsilon > 0$, converges to zero as $N \to \infty$.*

The typical performance of Basis Pursuit-based CS signal recovery is illustrated in Figure 2. Here, the linear program (4) attempts to recover a K-sparse signal x of length N, with the normalized sparsity rate fixed at $S = K/N = 0.1$ (each curve corresponds to a different N). The horizontal axis indicates the oversampling factor c, that is, the ratio between the number of measurements M (length of y) employed in (4) and the signal sparsity K. The vertical axis indicates the probability that the linear program yields the correct answer x. Clearly the probability increases with the number of measurements $M = cK$. Moreover, the curves become closer to a step function as N grows.

In an illuminating series of recent papers, Donoho and Tanner [32, 33] have characterized the oversampling factor $c(S)$ precisely. With appropriate oversampling, reconstruction via Basis Pursuit is also provably robust to measurement noise and quantization error [27]. In our work, we have

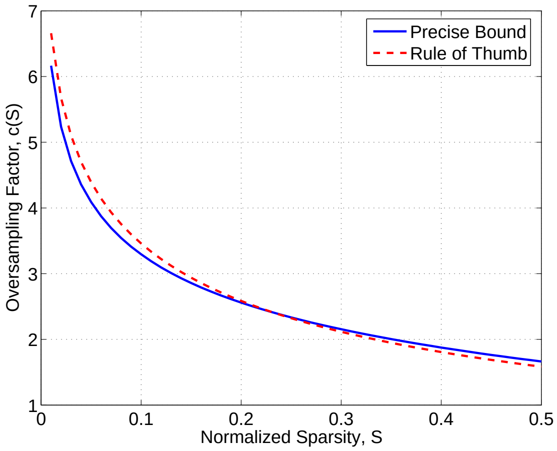


Figure 3: *Oversampling factor for ℓ_1 reconstruction. The solid line indicates the precise oversampling ratio $c(S)$ required for ℓ_1 recovery in CS [32, 33]. The dashed line indicates our proposed rule of thumb $c(S) \approx \log_2(1 + S^{-1})$.*

noticed that the oversampling factor is quite similar to $\log_2(1 + S^{-1})$. We find this expression a useful rule of thumb to approximate the precise oversampling ratio and illustrate the similarity in Figure 3.

Rule of Thumb *The oversampling factor $c(S)$ in Theorem 3 satisfies $c(S) \approx \log_2\left(1 + S^{-1}\right)$.*

In the remainder of the paper, we often use the abbreviated notation c to describe the oversampling factor required in various settings even though $c(S)$ depends on the sparsity K and signal length N.

2.2.5 Signal recovery via greedy pursuit

At the expense of slightly more measurements, iterative greedy algorithms have also been developed to recover the signal x from the measurements y. Examples include the iterative Orthogonal Matching Pursuit (OMP) [30], matching pursuit (MP), and tree matching pursuit (TMP) [35, 36] algorithms. OMP, for example, iteratively selects the vectors from the holographic basis $\Phi\Psi$ that contain most of the energy of the measurement vector y. The selection at each iteration is made based on inner products between the columns of $\Phi\Psi$ and a residual; the residual reflects the component of y that is orthogonal to the previously selected columns.

OMP is guaranteed to converge within a finite number of iterations. In CS applications, OMP requires $c \approx 2\ln(N)$ [30] to succeed with high probability. In the following, we will exploit both Basis Pursuit and greedy algorithms for recovering jointly sparse signals from incoherent measurements. We note that Tropp and Gilbert require the OMP algorithm to succeed in the first K iterations [30]; however, in our simulations, we allow the algorithm to run up to the maximum of M possible iterations. While this introduces a potential vulnerability to noise in the measurements, our focus in this paper is on the noiseless case. The choice of an appropriate practical stopping criterion (likely somewhere between K and M iterations) is a subject of current research in the CS community.

2.2.6 Related work

Recently, Haupt and Nowak [38] formulated a setting for CS in sensor networks that exploits inter-signal correlations. In their approach, each sensor $n \in \{1, 2, \ldots, N\}$ simultaneously records a single reading $x(n)$ of some spatial field (temperature at a certain time, for example).[6] Each of the sensors generates a pseudorandom sequence $r_n(m), m = 1, 2, \ldots, M$, and modulates the reading as $x(n)r_n(m)$. Each sensor n then transmits its M numbers in sequence in an analog and synchronized fashion to the collection point such that it automatically aggregates them, obtaining M measurements $y(m) = \sum_{n=1}^{N} x(n)r_n(m)$. Thus, defining $x = [x(1), x(2), \ldots, x(N)]^T$ and $\phi_m = [r_1(m), r_2(m), \ldots, r_N(m)]$, the collection point automatically receives the measurement vector $y = [y(1), y(2), \ldots, y(M)]^T = \Phi x$ after M transmission steps. The samples $x(n)$ of the spatial field can then be recovered using CS provided that x has a sparse representation in a known basis. The coherent analog transmission scheme also provides a power amplification property, thus reducing the power cost for the data transmission by a factor of N. There are some shortcomings to this approach, however. Sparse representations for x are straightforward when the spatial samples are arranged in a grid, but establishing such a representation becomes much more difficult when the spatial sampling is irregular [22]. Additionally, since this method operates at a single time instant, it exploits only inter-signal and not intra-signal correlations; that is, it essentially assumes that the sensor field is i.i.d. from time instant to time instant. In contrast, we will develop signal models and algorithms that are agnostic to the spatial sampling structure and that exploit both inter- *and* intra-signal correlations.

3 Joint Sparsity Models

In this section, we generalize the notion of a signal being sparse in some basis to the notion of an ensemble of signals being *jointly sparse*. In total, we consider three different *joint sparsity models* (JSMs) that apply in different situations. In the first two models, each signal is itself sparse, and so we could use the CS framework from Section 2.2 to encode and decode each one separately (independently). However, there also exists a framework wherein a *joint representation* for the ensemble uses fewer total vectors. In the third model, no signal is itself sparse, yet there still exists a joint sparsity among the signals that allows recovery from significantly fewer measurements per sensor.

We will use the following notation for signal ensembles and our measurement model. Denote the *signals* in the ensemble by x_j, $j \in \{1, 2, \ldots, J\}$, and assume that each signal $x_j \in \mathbb{R}^N$. We use $x_j(n)$ to denote sample n in signal j, and we assume that there exists a known *sparse basis* Ψ for $\mathbb{R}^N$ in which the x_j can be sparsely represented. The coefficients of this sparse representation can take arbitrary real values (both positive and negative). Denote by Φ_j the *measurement matrix* for signal j; Φ_j is $M_j \times N$ and, in general, the entries of Φ_j are different for each j. Thus, $y_j = \Phi_j x_j$ consists of $M_j < N$ *incoherent measurements* of x_j.[7] We will emphasize random i.i.d. Gaussian matrices Φ_j in the following, but other schemes are possible, including random ±1 Bernoulli/Rademacher matrices, and so on.

In previous sections, we discussed signals with intra-signal correlation (within each x_j) or signals with inter-signal correlation (between x_{j_1} and x_{j_2}). The three following models sport both kinds of correlation simultaneously.

[6] Note that in this section only, N refers to the number of sensors and not the length of the signals.

[7] The measurements at sensor j can be obtained either indirectly by sampling the signal x_j and then computing the matrix-vector product $y_j = \Phi_j x_j$ or directly by special-purpose hardware that computes y_j without first sampling (see [37], for example).

3.1 JSM-1: Sparse common component + innovations

In this model, all signals share a *common* sparse component while each individual signal contains a sparse *innovation* component; that is,

$$x_j = z_C + z_j, \quad j \in \{1, 2, \ldots, J\}$$

with

$$z_C = \Psi\theta_C, \quad \|\theta_C\|_0 = K_C \qquad \text{and} \qquad z_j = \Psi\theta_j, \quad \|\theta_j\|_0 = K_j.$$

Thus, the signal z_C is common to all of the x_j and has sparsity K_C in basis Ψ. The signals z_j are the unique portions of the x_j and have sparsity K_j in the same basis. Denote by Ω_C the support set of the nonzero θ_C values and by Ω_j the support set of θ_j.

A practical situation well-modeled by JSM-1 is a group of sensors measuring temperatures at a number of outdoor locations throughout the day. The temperature readings x_j have both temporal (intra-signal) and spatial (inter-signal) correlations. Global factors, such as the sun and prevailing winds, could have an effect z_C that is both common to all sensors and structured enough to permit sparse representation. More local factors, such as shade, water, or animals, could contribute localized innovations z_j that are also structured (and hence sparse). A similar scenario could be imagined for a network of sensors recording light intensities, air pressure, or other phenomena. All of these scenarios correspond to measuring properties of physical processes that change smoothly in time and in space and thus are highly correlated.

3.2 JSM-2: Common sparse supports

In this model, all signals are constructed from the same sparse set of basis vectors, but with different coefficients; that is,

$$x_j = \Psi\theta_j, \quad j \in \{1, 2, \ldots, J\}, \tag{5}$$

where each θ_j is nonzero only on the common coefficient set $\Omega \subset \{1, 2, \ldots, N\}$ with $|\Omega| = K$. Hence, all signals have ℓ_0 sparsity of K, and all are constructed from the same K basis elements but with arbitrarily different coefficients.

A practical situation well-modeled by JSM-2 is where multiple sensors acquire replicas of the same Fourier-sparse signal but with phase shifts and attenuations caused by signal propagation. In many cases it is critical to recover each one of the sensed signals, such as in many acoustic localization and array processing algorithms. Another useful application for JSM-2 is MIMO communication [34].

Similar signal models have been considered by different authors in the area of *simultaneous sparse approximation* [34, 52, 53]. In this setting, a collection of sparse signals share the same expansion vectors from a redundant dictionary. The sparse approximation can be recovered via greedy algorithms such as *Simultaneous Orthogonal Matching Pursuit* (SOMP) [34, 52] or *MMV Order Recursive Matching Pursuit* (M-ORMP) [53]. We use the SOMP algorithm in our setting (see Section 5) to recover from incoherent measurements an ensemble of signals sharing a common sparse structure.

3.3 JSM-3: Nonsparse common component + sparse innovations

This model extends JSM-1 so that the common component need no longer be sparse in any basis; that is,

$$x_j = z_C + z_j, \quad j \in \{1, 2, \ldots, J\}$$

with

$$z_C = \Psi\theta_C \qquad \text{and} \qquad z_j = \Psi\theta_j, \quad \|\theta_j\|_0 = K_j,$$

but z_C is not necessarily sparse in the basis Ψ. We also consider the case where the supports of the innovations are shared for all signals, which extends JSM-2. Note that separate CS reconstruction cannot be applied under JSM-3, since the common component is not sparse.

A practical situation well-modeled by JSM-3 is where several sources are recorded by different sensors together with a background signal that is not sparse in any basis. Consider, for example, an idealized computer vision-based verification system in a device production plant. Cameras acquire snapshots of components in the production line; a computer system then checks for failures in the devices for quality control purposes. While each image could be extremely complicated, the ensemble of images will be highly correlated, since each camera is observing the same device with minor (sparse) variations.

JSM-3 could also be useful in some non-distributed scenarios. For example, it motivates the compression of data such as video, where the innovations or differences between video frames may be sparse, even though a single frame may not be very sparse. In this case, JSM-3 suggests that we encode each video frame independently using CS and then decode all frames of the video sequence jointly. This has the advantage of moving the bulk of the computational complexity to the video decoder. Puri and Ramchandran have proposed a similar scheme based on Wyner-Ziv distributed encoding in their PRISM system [54]. In general, JSM-3 may be invoked for ensembles with significant inter-signal correlations but insignificant intra-signal correlations.

3.4 Refinements and extensions

Each of the JSMs proposes a basic framework for joint sparsity among an ensemble of signals. These models are intentionally *generic*; we have not, for example, mentioned the processes by which the index sets and coefficients are assigned. In subsequent sections, to give ourselves a firm footing for analysis, we will often consider specific *stochastic* generative models, in which (for example) the nonzero indices are distributed uniformly at random and the nonzero coefficients are drawn from a random Gaussian distribution. While some of our specific analytical results rely on these assumptions, the basic algorithms we propose should generalize to a wide variety of settings that resemble the JSM-1, 2, and 3 models.

It should also be clear that there are many possible joint sparsity models beyond the three we have introduced. One immediate extension is a combination of JSM-1 and JSM-2, where the signals share a common set of sparse basis vectors but with different expansion coefficients (as in JSM-2) plus additional innovation components (as in JSM-1). For example, consider a number of sensors acquiring different delayed versions of a signal that has a sparse representation in a multiscale basis such as a wavelet basis. The acquired signals will share the same wavelet coefficient support at coarse scales with different values, while the supports at each sensor will be different for coefficients at finer scales. Thus, the coarse scale coefficients can be modeled as the common support component, and the fine scale coefficients can be modeled as the innovation components.

Further work in this area will yield new JSMs suitable for other application scenarios. Applications that could benefit include multiple cameras taking digital photos of a common scene from various angles [55]. Additional extensions are discussed in Section 7.

4 Recovery Strategies for Sparse Common Component + Innovations (JSM-1)

In Section 2.1.2, Theorem 1 specified an entire region of rate pairs where distributed source coding is feasible (recall Figure 1). Our goal is to provide a similar characterization for measurement rates in DCS. In this section, we characterize the sparse common signal and innovations model (JSM-1);

we study JSMs 2 and 3 in Sections 5 and 6, respectively.

We begin this section by presenting a stochastic model for signals in JSM-1 in Section 4.1, and then present an information-theoretic framework where we scale the size of our reconstruction problem in Section 4.2. We study the set of viable representations for JSM-1 signals in Section 4.3. After defining our notion of a measurement rate region in Section 4.4, we present bounds on the measurement rate region using ℓ_0 and ℓ_1 reconstructions in Sections 4.5 and 4.6, respectively. We conclude with numerical examples in Section 4.7.

4.1 Stochastic signal model for JSM-1

To give ourselves a firm footing for analysis, we consider in this section a specific *stochastic* generative model for jointly sparse signals in JSM-1. Though our theorems and experiments are specific to this context, the basic ideas, algorithms, and results can be expected to generalize to other, similar scenarios.

For our model, we assume without loss of generality that $\Psi = I_N$, where I_N is the $N \times N$ identity matrix.[8] Although the extension to arbitrary bases is straightforward, this assumption simplifies the presentation because we have $x_1(n) = z_C(n) + z_1(n) = \theta_C(n) + \theta_1(n)$ and $x_2(n) = z_C(n)+z_2(n) = \theta_C(n)+\theta_2(n)$. We generate the common and innovation components in the following manner. For $n \in \{1, 2, ..., N\}$ the decision whether $z_C(n)$ is zero or not is an i.i.d. process, where the probability of a nonzero value is given by a parameter denoted S_C. The values of the nonzero coefficients are then generated from an i.i.d. Gaussian distribution. In a similar fashion we pick the K_j indices that correspond to the nonzero indices of z_j independently, where the probability of a nonzero value is given by a parameter S_j. The values of the nonzero innovation coefficients are then generated from an i.i.d. Gaussian distribution.

The outcome of this process is that each component z_j has an *operational sparsity* of K_j, where K_j has a Binomial distribution with mean NS_j, that is, $K_j \sim \text{Binomial}(N, S_j)$. A similar statement holds for z_C, K_C, and S_C. Thus, the parameters S_j and S_C can be thought of as *sparsity rates* controlling the random generation of each signal.

4.2 Information theoretic framework and notion of sparsity rate

In order to glean some theoretic insights, consider the simplest case where a *single joint encoder* processes $J = 2$ signals. By employing the CS machinery, we might expect that (i) $(K_C + K_1)c$ measurements suffice to reconstruct x_1, (ii) $(K_C + K_2)c$ measurements suffice to reconstruct x_2, and (iii) $(K_C + K_1 + K_2)c$ measurements suffice to reconstruct both x_1 and x_2, because we have $K_C + K_1 + K_2$ nonzero elements in x_1 and x_2.[9] Next, consider the case where the two signals are processed by *separate encoders*. Given the $(K_C + K_1)c$ measurements for x_1 as side information and assuming that the partitioning of x_1 into z_C and z_1 is known, cK_2 measurements that describe z_2 should allow reconstruction of x_2. Similarly, conditioned on x_2, we should need only cK_1 measurements to reconstruct x_1.

These observations seem related to various types of entropy from information theory; we thus expand our notions of sparsity to draw such analogies. As a motivating example, suppose that the signals $x_j,\ j \in \{1, 2, \ldots, J\}$ are generated by sources $X_j,\ j \in \{1, 2, \ldots, J\}$ using our stochastic model. As the signal length N is incremented one by one, the sources provide new values for $z_C(N)$ and $z_j(N)$, and the operational sparsity levels increase roughly linearly in the signal length. We thus define the *sparsity rate of* X_j as the limit of the proportion of coefficients that need to be

[8] If the measurement basis Φ is i.i.d. random Gaussian, then the matrix $\Phi\Psi$ remains i.i.d. Gaussian no matter what (orthonormal) sparse basis Ψ we choose.

[9] With a slight abuse of notation, we denote the oversampling factors for coding x_1, x_2, or both signals by c.

specified in order to reconstruct the signal x_j given its support set Ω_j; that is,

$$S(X_j) \triangleq \lim_{N\to\infty} \frac{K_C + K_j}{N}, \quad j \in \{1, 2, \ldots, J\}.$$

We also define the *joint sparsity* $S(X_{j_1}, X_{j_2})$ of x_{j_1} and x_{j_2} as the proportion of coefficients that need to be specified in order to reconstruct both signals given the support sets Ω_{j_1}, Ω_{j_2} of both signals. More formally,

$$S(X_{j_1}, X_{j_2}) \triangleq \lim_{N\to\infty} \frac{K_C + K_{j_1} + K_{j_2}}{N}, \quad j_1, j_2 \in \{1, 2, \ldots, J\}.$$

Finally, the *conditional sparsity* of x_{j_1} given x_{j_2} is the proportion of coefficients that need to be specified in order to reconstruct x_{j_1}, where x_{j_2} and Ω_{j_1} are available

$$S(X_{j_1}|X_{j_2}) \triangleq \lim_{N\to\infty} \frac{K_{j_1}}{N}, \quad j_1, j_2 \in \{1, 2, \ldots, J\}.$$

The joint and conditional sparsities extend naturally to groups of more than two signals.

The sparsity rate of the common source Z_C can be analyzed in a manner analogous to the *mutual information* [13] of traditional information theory; that is, $S_C = I(X_1; X_2) = S(X_1) + S(X_2) - S(X_1, X_2)$. While our specific stochastic model is somewhat simple, we emphasize that these notions can be extended to additional models in the class of stationary ergodic sources. These definitions offer a framework for joint sparsity with notions similar to the entropy, conditional entropy, and joint entropy of information theory.

4.3 Ambiguous representations for signal ensembles

As one might expect, the basic quantities that determine the measurement rates for a JSM-1 ensemble will be the sparsities K_C and K_j of the components z_C and z_j, $j = 1, 2, \ldots, J$. However we must account for an interesting side effect of our generative model. The representation $(z_C, z_1, \ldots, z_J)$ for a given signal ensemble $\{x_j\}$ is not unique; in fact many sets of components $(z_C, z_1, \ldots, z_J)$ (with different sparsities K_C and K_j) could give rise to the same signals $\{x_j\}$. We refer to any representation $(\overline{z_C}, \overline{z_1}, \ldots, \overline{z_J})$ for which $x_j = \overline{z_C} + \overline{z_j}$ for all j as a *viable representation* for the signals $\{x_j\}$. The sparsities of these viable representations will play a significant role in our analysis.

To study JSM-1 viability, we confine our attention to $J = 2$ signals. Consider the n-th coefficient $z_C(n)$ of the common component z_C and the corresponding innovation coefficients $z_1(n)$ and $z_2(n)$. Suppose that these three coefficients are all nonzero. Clearly, the same signals x_1 and x_2 could have been generated using at most two nonzero values among the three, for example by adding the value $z_C(n)$ to $z_1(n)$ and $z_2(n)$ (and then setting $z_C(n)$ to zero). Indeed, when all three coefficients are nonzero, we can represent them equivalently by any subset of two coefficients. Thus, there exists a sparser representation than we might expect given K_C, K_1, and K_2. We call this process *sparsity reduction*.

Likelihood of sparsity reduction: Having realized that *sparsity reduction* might be possible, we now characterize when it can happen and how likely it is. Consider the modification of $z_C(n)$ to some fixed $\overline{z_C(n)}$. If $z_1(n)$ and $z_2(n)$ are modified to

$$\overline{z_1(n)} \triangleq z_1(n) + z_C(n) - \overline{z_C(n)} \quad \text{and} \quad \overline{z_2(n)} \triangleq z_2(n) + z_C(n) - \overline{z_C(n)},$$

then $\overline{z_C(n)}$, $\overline{z_1(n)}$, and $\overline{z_2(n)}$ form a viable representation for $x_1(n)$ and $x_2(n)$. For example, if $z_C(n)$, $z_1(n)$, and $z_2(n)$ are nonzero, then

$$\overline{z_C(n)} = 0, \quad \overline{z_1(n)} = z_1(n) + z_C(n) \quad \text{and} \quad \overline{z_2(n)} = z_2(n) + z_C(n)$$

form a viable representation with reduced sparsity. Certainly, if all three original coefficients $z_C(n)$, $z_1(n)$, and $z_2(n)$ are nonzero, then the ℓ_0 sparsity of the n-th component can be reduced to two. However, once the sparsity has been reduced to two, it can only be reduced further if multiple original nonzero coefficient values were equal. Since we have assumed independent Gaussian coefficient amplitudes (see Section 4.1), further sparsity reduction is possible only with probability zero. Similarly, if two or fewer original coefficients are nonzero, then the probability that the sparsity can be reduced is zero. We conclude that *sparsity reduction is possible with positive probability only in the case where three original nonzero coefficients have an equivalent representation using two nonzero coefficients.*

Since the locations of the nonzero coefficients are uniform (Section 4.1), the probability that for one index n all three coefficients are nonzero is

$$\Pr(\text{sparsity reduction}) = \frac{K_C}{N}\frac{K_1}{N}\frac{K_2}{N}. \tag{6}$$

We denote the number of indices n for which $z_C(n)$, $z_1(n)$, and $z_2(n)$ are all nonzero by K_{C12}. Similarly, we denote the number of indices n for which both $z_C(n)$ and $z_1(n)$ are nonzero by K_{C1}, and so on. Asymptotically, the probability that all three elements are nonzero is

$$S_{C12} \triangleq \Pr(\text{sparsity reduction}) = S_C S_1 S_2.$$

Similarly, we denote the probability that both $z_C(n)$ and $z_1(n)$ are nonzero by $S_{C1} = S_C S_1$, and so on.

The previous arguments indicate that with probability one the total number of nonzero coefficients $K_C + K_1 + K_2$ can be reduced by K_{C12} but not more.[10] Consider a viable representation with minimal number of nonzero coefficients. We call this a *minimal sparsity representation.* Let the sparsity of the viable common component $\overline{z_C}$ be $\overline{K_C}$, and similarly let the number of nonzero coefficients of the viable j-th innovation component $\overline{z_j}$ be $\overline{K_j}$. The previous arguments indicate that with probability one a minimal sparsity representation satisfies

$$\overline{K_C} + \overline{K_1} + \overline{K_2} = K_C + K_1 + K_2 - K_{C12}. \tag{7}$$

One can view $\left(\overline{K_C}, \overline{K_1}, \overline{K_2}\right)$ as *operational* sparsities that represent the sparsest way to express the signals at hand.

Sparsity swapping: When the three signal coefficients $z_C(n)$, $z_1(n)$, $z_2(n)$ are nonzero, an alternative viable representation exists in which any one of them is zeroed out through sparsity reduction. Similarly, if any two of the coefficients are nonzero, then with probability one the corresponding signal values $x_1(n)$ and $x_2(n)$ are nonzero and differ. Again, any two of the three coefficients suffice to represent both values, and we can "zero out" any of the coefficients that are currently nonzero at the expense of the third coefficient, which is currently zero. This *sparsity swapping* provides numerous equivalent representations for the signals x_1 and x_2. To characterize sparsity swapping, we denote the number of indices for which at least two original coefficients are nonzero by

$$K_\cap \triangleq K_{C1} + K_{C2} + K_{12} - 2K_{C12};$$

this definition is easily extendable to $J > 2$ signals. As before, we use the generic notation Ω to denote the coefficient support set. Since $\Psi = I_N$, the coefficient vectors θ_C, θ_1, and θ_2 correspond to the signal components z_C, z_1, and z_2, which have support sets Ω_C, Ω_1, and Ω_2, respectively.

[10] Since N is finite, the expected number of indices n for which further sparsity reduction is possible is zero.

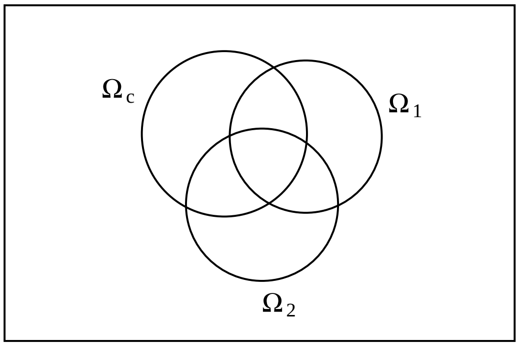


Figure 4: *Minimal sparsity representation region illustrating the overlap between the supports (denoted by Ω) of z_C, z_1, and z_2. The outer rectangle represents the set $\{1, 2, ..., N\}$ of which Ω_C, Ω_1, and Ω_2 are subsets. Due to independence, the sparsity of the overlap between multiple sets can be computed as the product of the individual sparsities.*

The intersections between the different support sets are illustrated in Figure 4. In an asymptotic setting, the probability of intersection satisfies

$$S_\cap \triangleq S_{C1} + S_{C2} + S_{12} - 2S_{C12}. \tag{8}$$

We call $K_\cap$ the *intersection sparsity* and $S_\cap$ the *intersection sparsity rate*. In addition to satisfying (7), a minimal sparsity representation must also obey

$$\overline{K_\cap} = K_\cap, \tag{9}$$

since for every index n where two or more coefficients intersect, $x_1(n)$ and $x_2(n)$ will differ and be nonzero with probability one, and so will be represented by two nonzero coefficients in any minimal sparsity representation. Furthermore, for any index n where two nonzero coefficients intersect, any of the three coefficients $\overline{z_C(n)}$, $\overline{z_1(n)}$, and $\overline{z_2(n)}$ can be "zeroed out." Therefore, the set of minimal representations lies in a cube with sidelength $K_\cap$.

We now ask where this cube lies. Clearly, no matter what sparsity reduction and swapping we perform, the potential for reducing K_C is no greater than $K_{C1} + K_{C2} - K_{C12}$. (Again, Figure 4 illustrates these concepts.) We denote the minimal sparsity that $\overline{z_C}$, $\overline{z_1}$, and $\overline{z_2}$ may obtain by K_C', K_1', and K_2', respectively. We have

$$\begin{aligned}
\overline{K_C} &\geq K_C' \triangleq K_C - K_{C1} - K_{C2} + K_{C12}, &(10a)\\
\overline{K_1} &\geq K_1' \triangleq K_1 - K_{C1} - K_{12} + K_{C12}, &(10b)\\
\overline{K_2} &\geq K_2' \triangleq K_2 - K_{C2} - K_{12} + K_{C12}. &(10c)
\end{aligned}$$

Therefore, the minimal sparsity representations lie in the cube $[K_C', K_C' + K_\cap] \times [K_1', K_1' + K_\cap] \times [K_2', K_2' + K_\cap]$. We now summarize the discussion with a result on sparsity levels of minimal sparsity representations.

Lemma 1 *With probability one, the sparsity levels $\overline{K_C}$, $\overline{K_1}$, and $\overline{K_2}$ of a minimal sparsity representation satisfy*

$$\begin{aligned}
K_C' \leq \overline{K_C} &\leq K_C' + K_\cap, &(11a)\\
K_1' \leq \overline{K_1} &\leq K_1' + K_\cap, &(11b)\\
K_2' \leq \overline{K_2} &\leq K_2' + K_\cap, &(11c)\\
\overline{K_C} + \overline{K_1} + \overline{K_2} &= K_C' + K_1' + K_2' + 2K_\cap. &(11d)
\end{aligned}$$

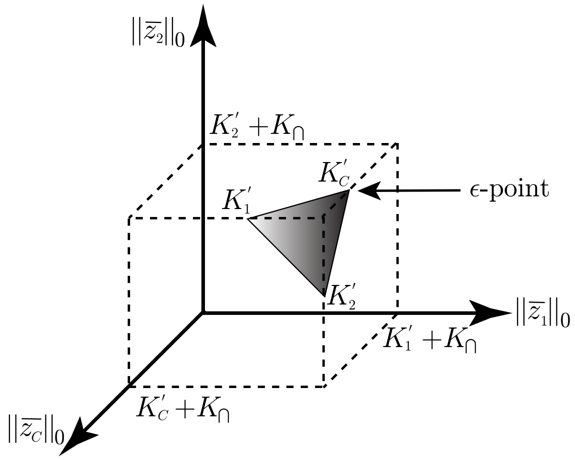


Figure 5: *Sparsity reduction and swapping. For $J = 2$ and a given K_C, K_1, and K_2, the possibility of overlap between signal components allows us to find a minimal sparsity representation with sparsity $\overline{K_C}$, $\overline{K_1}$ and $\overline{K_2}$. The shaded section of the triangle gives the set of minimal sparsity representations. The triangle lies on a hyperplane described by (7). The cube is described in Lemma 1. The ϵ-point, which essentially describes the measurement rates required for joint ℓ_0 reconstruction, lies on a corner of the triangle.*

Remark 2 *Equation (11d) is obtained by combining (7) with the useful identity*

$$K_C + K_1 + K_2 - K_{C12} = K_C' + K_1' + K_2' + 2K_\cap.$$

Combining these observations, among minimal sparsity representations, the values $\left(\overline{K_C}, \overline{K_1}, \overline{K_2}\right)$ lie on the intersection of a plane (7) with a cube. This intersection forms a triangle, as illustrated in Figure 5.

ϵ-point: Among all minimal sparsity representations $(\overline{z_C}, \overline{z_1}, \overline{z_2})$ there is one of particular interest because it determines the minimal measurement rates necessary to recover the signal ensemble $\{x_j\}$. The fact is that one cannot exploit *any* minimal sparsity representation for reconstruction. Consider, for example, the situation where the supports of z_C, z_1, and z_2 are identical. Using sparsity swapping and reduction, one might conclude that a representation where $\overline{z_C} = z_2$, $\overline{z_1} = z_1 - z_2$, and $\overline{z_2} = 0$ could be used to reconstruct the signal, in which case there is no apparent need to measure x_2 at all. Of course, since x_1 and x_2 differ and are both nonzero, it seems implausible that one could reconstruct x_2 without measuring it at all.

Theorems 4 and 5 suggest that the representation of particular interest is the one that places as few entries in the common component $\overline{z_C}$ as possible. As shown in Figure 5, there is a unique minimal sparsity representation that satisfies this condition. We call this representation the ϵ-point (for reasons that will be more clear in Section 4.5.2), and we denote its components by z_C^ϵ, z_1^ϵ, and z_2^ϵ. The sparsities of these components satisfy

$$\begin{aligned} K_C^\epsilon &= K_C', && \text{(12a)} \\ K_1^\epsilon &= K_1' + K_\cap, && \text{(12b)} \\ K_2^\epsilon &= K_2' + K_\cap. && \text{(12c)} \end{aligned}$$

We also define the sparsity rates S_C^ϵ, S_1^ϵ, and S_2^ϵ in an analogous manner.

4.4 Measurement rate region

To characterize DCS performance, we introduce a *measurement rate region.* Let M_1 and M_2 be the number of measurements taken of x_1 and x_2, respectively. We define the measurement rates R_1 and R_2 in an asymptotic manner as

$$R_1 \triangleq \lim_{N\to\infty} \frac{M_1}{N} \quad \text{and} \quad R_2 \triangleq \lim_{N\to\infty} \frac{M_2}{N}.$$

For a measurement rate pair (R_1, R_2) and sources X_1 and X_2, we wish to see whether we can reconstruct the signals with vanishing probability as N increases. In this case, we say that the measurement rate pair is *achievable.*

For signals that are jointly sparse under JSM-1, the individual sparsity rate of signal x_j is $S(X_j) = S_C + S_j - S_C S_j$. Separate recovery via ℓ_0 minimization would require a measurement rate $R_j = S(X_j)$. Separate recovery via ℓ_1 minimization would require an oversampling factor $c(S(X_j))$, and thus the measurement rate would become $S(X_j) \cdot c(S(X_j))$. To improve upon these figures, we adapt the standard machinery of CS to the joint recovery problem.

4.5 Joint recovery via ℓ_0 minimization

In this section, we begin to characterize the theoretical measurement rates required for joint reconstruction. We provide a lower bound for *all* joint reconstruction techniques, and we propose a reconstruction scheme based on ℓ_0 minimization that approaches this bound but has high complexity. In the Section 4.6 we pursue more efficient approaches.

4.5.1 Lower bound

For simplicity but without loss of generality we again consider the case of $J = 2$ received signals and sparsity basis $\Psi = I_N$. We can formulate the recovery problem using matrices and vectors as

$$z \triangleq \begin{bmatrix} z_C \\ z_1 \\ z_2 \end{bmatrix}, \quad x \triangleq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad y \triangleq \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \Phi \triangleq \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix}. \tag{13}$$

Since $\Psi = I_N$, we can define

$$\widetilde{\Psi} \triangleq \begin{bmatrix} \Psi & \Psi & 0 \\ \Psi & 0 & \Psi \end{bmatrix} \tag{14}$$

and write $x = \widetilde{\Psi} z$. We measure the sparsity of a representation z by its total ℓ_0 sparsity

$$\|z\|_0 = \|z_C\|_0 + \|z_1\|_0 + \|z_2\|_0.$$

We assume that any two representations z and $\widehat{z}$ for which $y = \Phi\widetilde{\Psi}z = \Phi\widetilde{\Psi}\widehat{z}$ and $\|z\|_0 = \|\widehat{z}\|_0$ are indistinguishable to any recovery algorithm.

The following theorem is proved in Appendix B. It essentially incorporates the lower bound of Theorem 2 for single signal CS into every measurement component of the representation region described in Lemma 1.

Theorem 4 *Assume the measurement matrices Φ_j contain i.i.d. Gaussian entries. The following conditions are necessary to enable recovery of all signals in the ensemble $\{x_j\}$:*

$$M_j \geq K'_j + K_\cap + 1, \quad j = 1, 2, \ldots, J, \tag{15a}$$

$$\sum_j M_j \geq K'_C + \sum_j K'_j + J \cdot K_\cap + 1. \tag{15b}$$

The measurement rates required in Theorem 4 are somewhat similar to those in the Slepian-Wolf theorem [14], where each signal must be encoded above its conditional entropy rate, and the entire collection must be coded above the joint entropy rate. In particular, we see that the measurement rate bounds reflect the sparsities of the ϵ-point defined in (12a)–(12c). We note also that this theorem is easily generalized beyond the stochastic model of Section 4.1 to other JSM-1 scenarios.

4.5.2 Constructive algorithm

We now demonstrate an achievable result, tight with the converse bounds, by considering a specific algorithm for signal recovery. As suggested by Theorem 2, to approach the theoretical bounds we must employ ℓ_0 minimization. We solve

$$\widehat{z} = \arg\min \|z_C\|_0 + \|z_1\|_0 + \|z_2\|_0 \quad \text{s.t. } y = \Phi\widetilde{\Psi}z. \tag{16}$$

The following theorem is proved in Appendix C.

Theorem 5 *Assume the measurement matrices Φ_j contain i.i.d. Gaussian entries. Then the ℓ_0 optimization program (16) recovers all signals in the ensemble $\{x_j\}$ almost surely if the following conditions hold:*

$$\begin{aligned} M_j &\geq K'_j + K_\cap + 1, \quad j = 1, 2, \ldots, J, & \text{(17a)} \\ \sum_j M_j &\geq K'_C + \sum_j K'_j + J \cdot K_\cap + 1. & \text{(17b)} \end{aligned}$$

As before, one measurement separates the achievable region of Theorem 5, where perfect reconstruction is possible with probability one, from the converse region of Theorem 4. These results again provide a strong converse measurement rate region in a manner analogous to the results by Wolfowitz [17]. Our joint recovery scheme provides a significant savings in measurements, because the common component can be measured as part of all J signals.

We note that when it succeeds, the ℓ_0 optimization program (16) could recover *any* of the minimal sparsity representations (each has the same sparsity $\|z\|_0$ and each provides a valid reconstruction of x). If one were so inclined, this program could be modified to provide a unique solution (the ϵ-point) by replacing the optimization program (16) with

$$\widehat{z} = \arg\min\,(1+\epsilon)\|z_C\|_0 + \|z_1\|_0 + \|z_2\|_0 \quad \text{s.t. } y = \Phi\widetilde{\Psi}z, \tag{18}$$

for small $\epsilon > 0$. This slight ϵ-modification to a minimization problem of the form $\arg\min \|z\|_0$ (16) prioritizes the innovations components in cases where sparsity swapping is possible. It is from this formulation that the ϵ-point draws its name.

Despite the elegance of Theorem 5, it is of limited utility, since in practice we do not know how much sparsity reduction and swapping can be performed. However, if we fix the common sparsity rate S_C and innovation sparsity rates $S_1, S_2, \ldots, S_J$ and increase N, then

$$\lim_{N\to\infty} \frac{K_{C12}}{N} = S_{C12}.$$

Using (7), the minimal sparsity representation satisfies

$$\lim_{N\to\infty} \frac{\overline{K} + \sum_j \overline{K_j}}{N} = S_C + \sum_j S_j - S_{C12} = S'_C + \sum_j S'_j + J \cdot S_\cap, \tag{19}$$

and the sparsity rates of the ϵ-point satisfy

$$S_C^\epsilon \triangleq \lim_{N\to\infty} \frac{\overline{K_C^\epsilon}}{N} = S_C',$$

$$S_j^\epsilon \triangleq \lim_{N\to\infty} \frac{\overline{K_j^\epsilon}}{N} = S_j' + S_\cap,$$

where the minimal sparsity rates S_C', S_1', and S_2' are derived from (10a)–(10c):

$$S_C' \triangleq S_C - S_{C1} - S_{C2} + S_{C12}, \tag{20a}$$

$$S_1' \triangleq S_1 - S_{C1} - S_{12} + S_{C12}, \tag{20b}$$

$$S_2' \triangleq S_2 - S_{C2} - S_{12} + S_{C12}. \tag{20c}$$

We incorporate these results to characterize the measurement rate region in the following corollary.

Corollary 1 *Assume the measurement matrices Φ_j contain i.i.d. Gaussian entries. Then as N increases, the ℓ_0 optimization program (16) recovers all signals in the ensemble $\{x_j\}$ almost surely if the following conditions hold:*

$$\begin{aligned} R_j &> S_j' + S_\cap, \quad j = 1, 2, \ldots, J, \\ \sum_j R_j &> S_C' + \sum_j S_j' + J \cdot S_\cap. \end{aligned}$$

4.6 Joint recovery via ℓ_1 minimization

We again confine our attention to $J = 2$ signals with $\Psi = I_N$. We also assume that the innovation sparsity rates are equal and dub them $S_I \triangleq S(Z_1) = S(Z_2)$.

4.6.1 Formulation

As discussed in Section 2.2.3, solving an ℓ_0 optimization problem is NP-complete, and so in practice we must relax our ℓ_0 criterion in order to make the solution tractable. In regular (non-distributed) CS (Section 2.2.3), ℓ_1 minimization can be implemented via linear programming but requires an oversampling factor of $c(S)$ (Theorem 3). In contrast, ℓ_0 reconstruction only requires one measurement above the sparsity level K, both for regular and distributed compressed sensing (Theorems 2, 4, and 5). We now wish to understand what penalty must be paid for ℓ_1 reconstruction of jointly sparse signals.

Using the frame $\widetilde{\Psi}$, as shown in (14), we can represent the data vector x sparsely using the coefficient vector z, which contains $K_C + K_1 + K_2$ nonzero coefficients, to obtain $x = \widetilde{\Psi} z$. The concatenated measurement vector y is computed from separate measurements of the signals x_j, where the joint measurement basis is Φ and the joint holographic basis is then $V = \Phi\widetilde{\Psi}$. With sufficient oversampling, we can recover a vector $\widehat{z}$, which is a viable representation for x, by solving the linear program

$$\widehat{z} = \arg\min \|z\|_1 \quad \text{s.t. } y = \Phi\widetilde{\Psi} z. \tag{21}$$

The vector z enables the reconstruction of the original signals x_1 and x_2.

We find it helpful to modify the Basis Pursuit algorithm to account for the special structure of JSM-1 recovery. In the linear program (21), we replace the ℓ_1 performance metric

$$||z||_1 = ||z_C||_1 + ||z_1||_1 + ||z_2||_1$$

with the modified ℓ_1 metric

$$\gamma_C||z_C||_1 + \gamma_1||z_1||_1 + \gamma_2||z_2||_1, \tag{22}$$

where $\gamma_C, \gamma_1, \gamma_2 \geq 0$. We call this the *$\gamma$-weighted ℓ_1 formulation.* If $K_1 = K_2$ and $M_1 = M_2$, then we set $\gamma_1 = \gamma_2$. In this scenario, without loss of generality, we set $\gamma_1 = \gamma_2 = 1$ and optimize γ_C. We discuss the asymmetric case with $K_1 = K_2$ and $M_1 \neq M_2$ below in Section 4.6.3.

In Section 4.6.2 we study the γ-weighted ℓ_1 formulation (22) and provide converse bounds on its reconstruction performance. This technique relies on the γ values; we discuss these effects in Section 4.6.3. While we have not as yet been able to provide theoretical achievable bounds for this method, Section 4.6.4 describes another ℓ_1-based reconstruction algorithm whose performance is more amenable to analysis. Numerical results in Section 4.7 indicate that the γ-weighted ℓ_1 formulation can offer favorable performance.

4.6.2 Converse bounds on performance of γ-weighted ℓ_1 signal recovery

We now provide several converse bounds that describe what measurement rate pairs *cannot* be achieved via ℓ_1 recovery. Before proceeding, we shed some light on the notion of a converse region in this computational scenario. We focus on the setup where each signal x_j is measured via multiplication by the M_j by N matrix Φ_j and joint reconstruction of the J signals is performed via our γ-weighted ℓ_1 formulation (22). Within this setup, a converse region is a set of measurement rates for which the reconstruction techniques fail with overwhelming probability as N increases.

We now present our bounds, assuming $J = 2$ sources with innovation sparsity rates satisfying $S_1 = S_2 = S_I$. For brevity we define the measurement function

$$c'(S) \triangleq S \cdot c(S)$$

based on Donoho and Tanner's oversampling factor c (Theorem 3 [31–33]). We begin with a result, proved in Appendix D, that provides necessary conditions to reconstruct the viable components $\overline{z_C}$, $\overline{z_1}$, and $\overline{z_2}$ for x_1 and x_2.

Lemma 2 *Consider any γ_C, γ_1, and γ_2 in the γ-weighted ℓ_1 formulation (22). The components $\overline{z_C}$, $\overline{z_1}$, and $\overline{z_2}$ can be recovered using measurement matrices Φ_1 and Φ_2 only if (i) $\overline{z_1}$ can be recovered via ℓ_1 CS reconstruction (4) using Φ_1 and measurements $\Phi_1\overline{z_1}$; (ii) $\overline{z_2}$ can be recovered via ℓ_1 CS reconstruction using Φ_2 and measurements $\Phi_2\overline{z_2}$; and (iii) $\overline{z_C}$ can be recovered via ℓ_1 CS reconstruction using the joint matrix $[\Phi_1^T \;\; \Phi_2^T]^T$ and measurements $[\Phi_1^T \;\; \Phi_2^T]^T\overline{z_C}$.*

Remark 3 *This result provides deterministic necessary conditions for correct reconstruction using the γ-weighted ℓ_1 formulation (22).*

Lemma 2 can be interpreted as follows. If M_1 and M_2 are not large enough individually, then the innovation components $\overline{z_1}$ and $\overline{z_2}$ cannot be reconstructed. This implies converse bounds on the individual measurement rates R_1 and R_2. Similarly, combining Lemma 2 with the converse bound of Theorem 3 for standard ℓ_1 reconstruction of the common component $\overline{z_C}$ yields a lower bound on the sum measurement rate $R_1 + R_2$. We have incorporated these insights to prove the following result in Appendix E.

Theorem 6 *Let $J = 2$ and fix the sparsity rate of the common part to $S(Z_C) = S_C$ and the innovation sparsity rates to $S(Z_1) = S(Z_2) = S_I$. Then the following conditions on the measurement* *rates are necessary to enable reconstruction using the γ-weighted ℓ_1 formulation (22) with vanishing probability of error:*

$$\begin{aligned} R_1 &\geq c'(S'_I), \\ R_2 &\geq c'(S'_I), \\ R_1 + R_2 &\geq c'(S'_C). \end{aligned}$$

The theorem provides a converse region such that, if (R_1, R_2) violate these conditions and we perform $M_1 = \lceil (R_1 - \epsilon)N \rceil$ measurements for x_1 or $M_2 = \lceil (R_2 - \epsilon)N \rceil$ measurements for x_2, then the probability of incorrect reconstruction will converge to one as N increases.

Anticipated converse: Recall the ϵ-point from the ℓ_0 formulation (12a)–(12c). As mentioned earlier, we speculate that for indices n such that $x_1(n)$ and $x_2(n)$ differ and are nonzero, each sensor must take measurements to account for one of the two coefficients. The sum sparsity rate is $S'_C + S'_1 + S'_2 + 2S_\cap$ and, in the simplified case where $S_1 = S_2 = S_I$, the sum sparsity becomes $S'_C + 2S'_I + 2S_\cap$. It can be shown that the oversampling factor $c'(\cdot)$ is concave, and so it is best to "explain" as many of the sparse coefficients in one of the signals and as few as possible in the other. For the ϵ-point, we have

$$S_1^\epsilon = S_2^\epsilon = S'_I + S_\cap.$$

Consequently, one of the signals must "explain" this sparsity rate, whereas the other signal must explain the rest

$$[S'_C + 2S'_I + 2S_\cap] - [S'_I + S_\cap] = S'_C + S'_I + S_\cap.$$

We conclude with the following conjecture.

Conjecture 1 *Let $J = 2$ and fix the sparsity rate of the common part $S(Z_C) = S_C$ and the innovation sparsity rates $S(Z_1) = S(Z_2) = S_I$. Then the following conditions on the measurement rates are necessary to enable reconstruction with vanishing probability of error*

$$\begin{aligned} R_1 &\geq c'\left(S'_I + S_\cap\right), \\ R_2 &\geq c'\left(S'_I + S_\cap\right), \\ R_1 + R_2 &\geq c'\left(S'_I + S_\cap\right) + c'\left(S'_C + S'_I + S_\cap\right). \end{aligned}$$

Remark 4 *In the rigorous converse Theorem 6, the individual and sum rate bounds are $c'(S'_I)$ and $c'(S'_C)$, whereas in Conjecture 1 the bounds are $c'(S'_I + S_\cap)$ and $c'(S'_I + S_\cap) + c'(S'_C + S'_I + S_\cap)$, respectively. Therefore, the differences in the bounds are that (i) the sparsity rate of the ϵ-point adds $S_\cap$ terms and (ii) the sum rate bound of Lemma 2 proves that the sum rate must suffice to describe the common component, whereas in our conjecture we speculate that a total sparsity rate of $S'_C + 2S'_I + 2S_\cap$ must somehow be accounted for. Note also that the terms in our bounds are analogous to notions of joint entropy, conditional entropy, and mutual information.*

4.6.3 Optimal γ values

In our ℓ_1 reconstruction (22), the optimal choice of γ_C, γ_1, and γ_2 depends on the relative sparsities K_C, K_1, and K_2. At this stage we have not been able to determine the optimal values analytically. Instead, we rely on a numerical optimization, which is computationally intense. In this section we offer our intuition behind the choice of the optimal γ (confirmed with numerical simulations).

When the number of signals J is small and the measurement matrices Φ_j are different, and in any case when $\sum_j M_j < N$, it is possible to construct a common signal z_C that explains all the measurements without the need for any innovation signals. (This is accomplished by concatenating all the measurements and using a pseudoinverse.) However, such a z_C will presumably not be sparse. Therefore, when using different Φ_j matrices and jointly reconstructing, it may be important to penalize the sparsity of z_C, and the tradeoff is biased in favor of larger γ_C. This is especially important when $\sum_j K_j \gg K_C$.

An entirely different behavior occurs if identical measurement matrices Φ_j are used. In this case, we cannot "hide" all the measurements in z_C, and so it may be less important to penalize the sparsity of z_C via γ_C, and the bias to increase γ_C is reduced. However, in the setup where we try to recover $(z_C, z_1, \ldots, z_J)$ jointly, the measurement matrix Φ has worse incoherency with the sparsity matrix $\widetilde{\Psi}$ when all Φ_j are the same. The biggest problem comes in the first N columns of $\widetilde{\Psi}$ — those that are measuring z_C. Hence the incoherency is most challenging when $K_C \gg \sum_j K_j$.

When J is large, we have abundant information for recovering the common component. Using identical Φ_j matrices, we can average our (many) observations to obtain a good approximation of Φz_C from which we can recover z_C via single-signal CS. Using different Φ_j, we could use a pseudoinverse to recover z_C, completely ignoring the fact that it may be sparse (a similar procedure is applied to recover z_C in JSM-3; see Section 6). Both methods may provide somewhat noisy reconstructions, but that noise should decrease as J becomes larger. In any case, as J increases the bias is to increase γ_C, since the abundant information to reconstruct the common component must be offset by a penalty that increases the ℓ_1 term.

Finally, γ_C must be modified when asymmetric measurement rates are used. Consider as a simple example the case where $J = 2$ and $K_1 = K_2$. Suppose also that we use the convention where a single γ_C is used for the common component (instead of weighting z_1 and z_2 differently in the reconstruction), and $M_1 + M_2 = M$ is fixed. If M_1 is increased, then fewer measurements are available to reconstruct z_2; hence γ_C must be increased. Unfortunately, this causes a degradation in performance, as illustrated in Figure 6, where M must be increased to provide the same probability of correct reconstruction. We also evaluated the case where z_1 and z_2 are weighted differently by choosing $\gamma_C = 1$ and optimizing γ_1 and γ_2 numerically. Our preliminary results indicate an insignificant performance enhancement.

4.6.4 Achievable bounds on performance of ℓ_1 signal recovery

Now that we have ruled out part of the measurement region, we wish to specify regions where joint reconstruction can succeed. Unfortunately, we have not been able to characterize the performance of our γ-weighted ℓ_1 formulation (22) analytically. Instead, Theorem 7 below, proved in Appendix F, uses an alternative ℓ_1-based reconstruction technique. The proof describes a constructive reconstruction algorithm that is very insightful. We construct measurement matrices Φ_1 and Φ_2 which each consist of two parts. The first parts of the matrices are identical and reconstructs $x_1 - x_2$. The second parts of the matrices are different and enable the reconstruction of $\frac{1}{2}x_1 + \frac{1}{2}x_2$. Once these two components have been reconstructed, the computation of x_1 and x_2 is straightforward. The measurement rate can be computed by considering both common and different parts of the measurement matrices.

Theorem 7 *Let $J = 2$ and fix the sparsity rate of the common part $S(Z_C) = S_C$ and the innovation sparsity rates $S(Z_1) = S(Z_2) = S_I$. Then there exists an ℓ_1 reconstruction technique (along with a*

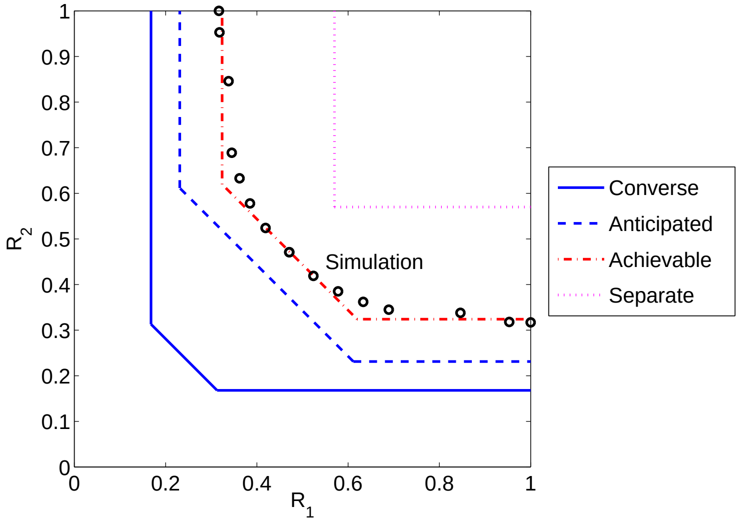


Figure 6: *Rate region for Joint Sparsity Model 1. We chose a common sparsity rate $S_C = 0.2$ and innovation sparsity rates $S_I = S_1 = S_2 = 0.05$. Our simulation results use the γ-weighted ℓ_1-formulation on signals of length $N = 1000$. Note that the anticipated converse of Conjecture 1 is significantly closer to our achievable from Theorem 7 than our converse from Theorem 6.*

measurement strategy) if the measurement rates satisfy the following conditions

$$\begin{aligned} R_1 &\geq c'(2S_I - (S_I)^2), && \text{(23a)} \\ R_2 &\geq c'(2S_I - (S_I)^2), && \text{(23b)} \\ R_1 + R_2 &\geq c'(2S_I - (S_I)^2) + c'(S_C + 2S_I - 2S_C S_I - (S_I)^2 + S_C(S_I)^2). && \text{(23c)} \end{aligned}$$

Furthermore, as $S_I \to 0$ the sum measurement rate approaches $c'(S_C)$.

This reconstruction is based on linear programming. It can be extended from $J = 2$ to an arbitrary number of signals by reconstructing all signal differences of the form $x_{j_1} - x_{j_2}$ in the first stage of the algorithm and then reconstructing $\frac{1}{J}\sum_j x_j$ in the second stage. Despite these advantages, the achievable measurement rate region of Theorem 7 is loose with respect to the region of the converse Theorem 6, as shown in Figure 6. Note, however, that the achievable region is significantly closer to the anticipated converse bound of Conjecture 1. Ultimately, we aim to provide a tight measurement rate region for reconstruction techniques with moderate (polynomial) computational requirements; we leave this for future work.

Comparison to γ-weighted ℓ_1 formulation (22): The achievable approach of Theorem 7 offers a computational advantage with respect to our γ-weighted ℓ_1 formulation (22). In our previous reconstruction approach (22), the linear program must reconstruct the $J+1$ vectors $z_C, z_1, \ldots, z_J$. Since the complexity of linear programming is roughly cubic, the computational burden scales with J^3. In contrast, the achievable approach of Theorem 7 reconstructs $J(J-1)/2$ pairs of the form $x_{j_1} - x_{j_2}$ and one additional average part, but each such reconstruction is only for a length-N signal. Therefore the computational load is lighter by an $O(J)$ factor. However, our γ-weighted ℓ_1 formulation also offers some advantages. Although we have not been able to characterize its achievable performance theoretically, our simulation tests indicate that it can reconstruct using fewer measurements than the approach of Theorem 7.

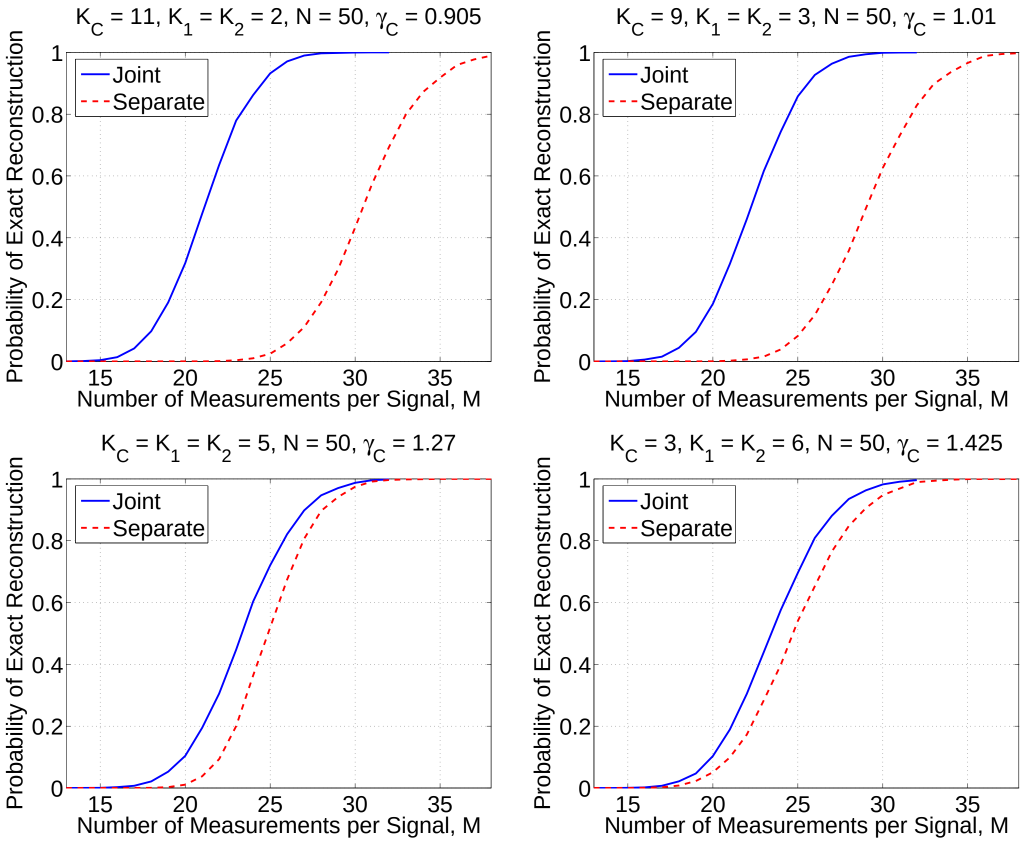


Figure 7: *Comparison of joint decoding and separate decoding for JSM-1. The advantage of joint over separate decoding depends on the common sparsity.*

4.7 Simulations for JSM-1

We now present simulation results for several different JSM-1 setups. We used the γ-weighted ℓ_1 formulation (22) throughout.

Reconstructing two signals with symmetric measurement rates: Our simulation setup is as follows. We consider two correlated signals x_1 and x_2 that satisfy our conditions for joint sparsity (Section 3.1). The signal components z_C, z_1, and z_2 are assumed (without loss of generality) to be sparse in $\Psi = I_N$ with sparsities K_C, K_1, and K_2, respectively. We assign random Gaussian values to the nonzero coefficients. We restrict our attention to the symmetric setup in which $K_1 = K_2$ and $M_1 = M_2$, and consider signals of length $N = 50$ and sparsity parameters chosen such that $K_C + K_1 + K_2 = 15$ and $K_1 = K_2$.

In our joint decoding simulations, we consider values of M_1 and M_2 in the range between 10 and 40. We find the optimal γ_C in the γ-weighted ℓ_1 formulation (22) using a line search optimization, where simulation indicates the "goodness" of specific γ_C values in terms of the likelihood of reconstruction. With the optimal γ_C, for each set of values we run several thousand trials to determine the empirical probability of success in decoding z_1 and z_2. The results of the simulation are summarized in Figure 7. The results reveal that the degree to which joint decoding outperforms separate decoding is directly related to the amount of shared information K_C. The savings in the number of required measurements M can be substantial, especially when the common component

K_C is large (Figure 7). For $K_C = 11$, $K_1 = K_2 = 2$, M is reduced by approximately 30%. For smaller K_C, joint decoding barely outperforms separate decoding, since most of the measurements are expended on innovation components.

Reconstructing two signals with asymmetric measurement rates: In Figure 6, we compare separate CS reconstruction with the converse bound of Theorem 6, the anticipated converse bound of Conjecture 1, the achievable bound of Theorem 7, and numerical results.

We use $J = 2$ signals and choose a common sparsity rate $S_C = 0.2$ and innovation sparsity rates $S_I = S_1 = S_2 = 0.05$. Several different asymmetric measurement rates are considered. In each such setup, we constrain M_2 to have the form $M_2 = \alpha M_1$ for some α. In the simulation itself, we first find the optimal γ_C using a line search optimization as described above. In order to accelerate this intense optimization, we use relatively short signals of length $N = 40$. Once the optimal gammas have been determined, we simulate larger problems of size $N = 1000$. The results plotted indicate the smallest pairs (M_1, M_2) for which we always succeeded reconstructing the signal over 100 simulation runs. The figure shows that in some areas of the measurement rate region our γ-weighted ℓ_1 formulation (22) requires less measurements than the achievable approach of Theorem 7.

Reconstructing multiple signals with symmetric measurement rates: The reconstruction techniques of this section are especially promising when more than $J = 2$ sensors are used, since the innovation sparsity rates may become smaller as additional side information from other signals becomes available, thus enabling even greater savings in the measurement rates. These savings may be especially valuable in applications such as sensor networks, where data may contain strong spatial (inter-source) correlations.

We use $J \in \{2, \ldots, 10\}$ signals and choose the same sparsity rates $S_C = 0.2$ and $S_I = 0.05$ as the asymmetric rate simulations; here we use symmetric measurement rates. We first find the optimal γ_C using a line search optimization as described above; during this procedure we use relatively short signals of length $N = 40$ to accelerate the computation. Once the optimal gammas are determined, we simulate larger problems of size $N = 500$ (since the computation scales with $(J+1)^3$, as mentioned in Section 4.6.3, we used shorter signals than in the asymmetric rate $J = 2$ signal simulations described above). The results of Figure 8 describe the smallest symmetric measurement rates for which we always succeeded reconstructing the signal over 100 simulation runs. Clearly, as J increases, lower measurement rates can be used.

5 Recovery Strategies for Common Sparse Supports (JSM-2)

Under the JSM-2 signal ensemble model from Section 3.2, separate recovery of each signal via ℓ_0 minimization would require $K + 1$ measurements per signal, while separate recovery via ℓ_1 minimization would require cK measurements per signal. As we now demonstrate, the total number of measurements can be reduced substantially by employing specially tailored joint reconstruction algorithms that exploit the common structure among the signals, in particular the common coefficient support set Ω.

The algorithms we propose are inspired by conventional greedy pursuit algorithms for CS (such as OMP [30]). In the single-signal case, OMP iteratively constructs the sparse support set Ω; decisions are based on inner products between the columns of $\Phi\Psi$ and a residual. In the multi-signal case, there are more clues available for determining the elements of Ω.

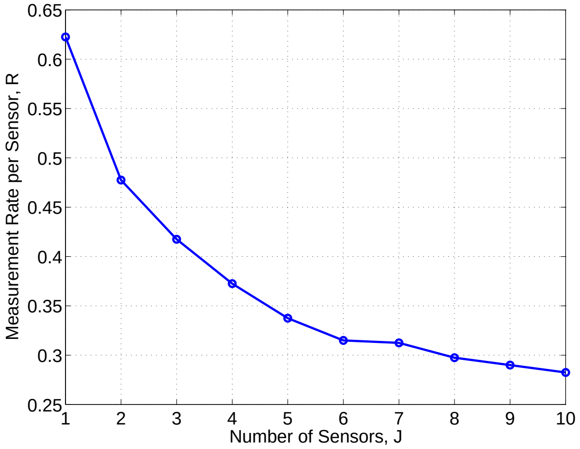


Figure 8: *Multi-sensor measurement results for JSM-1. We choose a common sparsity rate* $S_C = 0.2$, *innovation sparsity rates* $S_I = 0.05$, *and signals of length* $N = 500$.

5.1 Recovery via One-Step Greedy Algorithm (OSGA)

When there are many correlated signals in the ensemble, a simple non-iterative greedy algorithm based on inner products will suffice to recover the signals jointly. For simplicity but without loss of generality, we again assume that $\Psi = I_N$ and that an equal number of measurements $M_j = M$ are taken of each signal. We write Φ_j in terms of its columns, with $\Phi_j = [\phi_{j,1}, \phi_{j,2}, \ldots, \phi_{j,N}]$.

One-Step Greedy Algorithm (OSGA) for JSM-2

1. **Get greedy:** Given all of the measurements, compute the test statistics

$$\xi_n = \frac{1}{J}\sum_{j=1}^{J}\langle y_j, \phi_{j,n}\rangle^2, \qquad n \in \{1, 2, \ldots, N\} \tag{24}$$

and estimate the elements of the common coefficient support set by

$$\widehat{\Omega} = \{n \text{ having one of the } K \text{ largest } \xi_n\}.$$

When the sparse, nonzero coefficients are sufficiently generic (as defined below), we have the following surprising result, which is proved in Appendix G.

Theorem 8 *Let* Ψ *be an orthonormal basis for* $\mathbb{R}^N$, *let the measurement matrices* Φ_j *contain i.i.d. Gaussian entries, and assume that the nonzero coefficients in the* θ_j *are i.i.d. Gaussian random variables. Then with* $M \geq 1$ *measurements per signal, OSGA recovers* Ω *with probability approaching one as* $J \to \infty$.

In words, with *fewer* than K measurements per sensor, it is possible to recover the sparse support set Ω under the JSM-2 model.[11] Of course, this approach does not recover the K coefficient values for each signal; K measurements per sensor are required for this.

[11] One can also show the somewhat stronger result that, as long as $\sum_j M_j \gg N$, OSGA recovers Ω with probability approaching one. We have omitted this additional result for brevity.

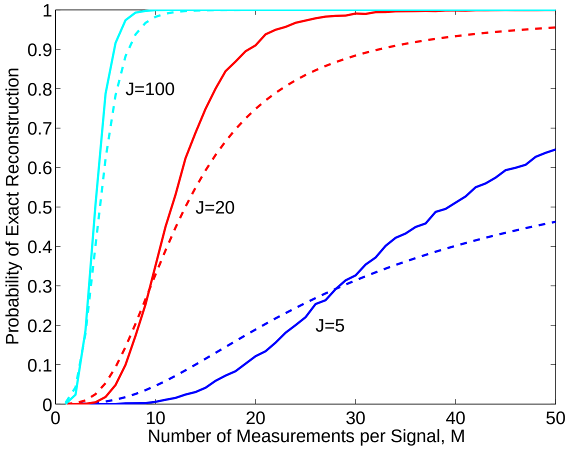


Figure 9: *Reconstruction using OSGA for JSM-2. Approximate formula (dashed lines) for the probability of error in recovering the support set Ω in JSM-2 using OSGA given J, N, K, and M [56] compared against simulation results (solid) for fixed $N = 50$, $K = 5$ and varying number of measurements M and number of signals $J = 5$, $J = 20$, and $J = 100$.*

Theorem 9 *Assume that the nonzero coefficients in the θ_j are i.i.d. Gaussian random variables. Then the following statements hold:*

1. *Let the measurement matrices Φ_j contain i.i.d. Gaussian entries, with each matrix having an oversampling factor of $c = 1$ (that is, $M_j = K$ for each measurement matrix Φ_j). Then OSGA recovers all signals from the ensemble $\{x_j\}$ with probability approaching one as $J \to \infty$.*

2. *Let Φ_j be a measurement matrix with oversampling factor $c < 1$ (that is, $M_j < K$), for some $j \in \{1, 2, \ldots, J\}$. Then with probability one, the signal x_j cannot be uniquely recovered by any algorithm for any value of J.*

The first statement is an immediate corollary of Theorem 8; the second statement follows because each equation $y_j = \Phi_j x_j$ would be underdetermined even if the nonzero indices were known. Thus, under the JSM-2 model, the one-step greedy algorithm asymptotically performs as well as an oracle decoder that has prior knowledge of the locations of the sparse coefficients. From an information theoretic perspective, Theorem 9 provides tight achievable and converse bounds for JSM-2 signals.

In a technical report [56], we derive an approximate formula for the probability of error in recovering the common support set Ω given J, N, K, and M. Figure 9 depicts the performance of the formula in comparison to simulation results. While theoretically interesting and potentially practically useful, these results require J to be large. Our numerical experiments show that OSGA works well even when M is small, as long as J is sufficiently large. However, in the case of fewer signals (small J), OSGA performs poorly. We propose next an alternative recovery technique based on simultaneous greedy pursuit that performs well for small J.

5.2 Recovery via iterative greedy pursuit

In practice, the common sparse support among the J signals enables a fast iterative algorithm to recover all of the signals jointly. Tropp and Gilbert have proposed one such algorithm, called *Simultaneous Orthogonal Matching Pursuit* (SOMP) [34], which can be readily applied in our DCS framework. SOMP is a variant of OMP that seeks to identify Ω one element at a time. (A similar simultaneous sparse approximation algorithm has been proposed using convex optimization; see [57] for details.) We dub the DCS-tailored SOMP algorithm DCS-SOMP.

To adapt the original SOMP algorithm to our setting, we first extend it to cover a different measurement basis Φ_j for each signal x_j. Then, in each DCS-SOMP iteration, we select the column index $n \in \{1, 2, \ldots, N\}$ that accounts for the greatest amount of residual energy across *all* signals. As in SOMP, we orthogonalize the remaining columns (in each measurement basis) after each step; after convergence we obtain an expansion of the measurement vector on an orthogonalized subset of the holographic basis vectors. To obtain the expansion coefficients in the sparse basis, we then reverse the orthogonalization process using the QR matrix factorization. We assume without loss of generality that $\Psi = I_N$.

DCS-SOMP Algorithm for JSM-2

1. **Initialize:** Set the iteration counter $\ell = 1$. For each signal index $j \in \{1, 2, \ldots, J\}$, initialize the orthogonalized coefficient vectors $\widehat{\beta}_j = 0$, $\widehat{\beta}_j \in \mathbb{R}^M$; also initialize the set of selected indices $\widehat{\Omega} = \emptyset$. Let $r_{j,\ell}$ denote the residual of the measurement y_j remaining after the first ℓ iterations, and initialize $r_{j,0} = y_j$.

2. **Select** the dictionary vector that maximizes the value of the sum of the magnitudes of the projections of the residual, and add its index to the set of selected indices

$$\begin{aligned} n_\ell &= \arg\max_{n=1,2,\ldots,N} \sum_{j=1}^{J} \frac{|\langle r_{j,\ell-1}, \phi_{j,n}\rangle|}{\|\phi_{j,n}\|_2}, \\ \widehat{\Omega} &= [\widehat{\Omega} \ \ n_\ell]. \end{aligned}$$

3. **Orthogonalize** the selected basis vector against the orthogonalized set of previously selected dictionary vectors

$$\gamma_{j,\ell} = \phi_{j,n_\ell} - \sum_{t=0}^{\ell-1} \frac{\langle \phi_{j,n_\ell}, \gamma_{j,t}\rangle}{\|\gamma_{j,t}\|_2^2} \gamma_{j,t}.$$

4. **Iterate:** Update the estimate of the coefficients for the selected vector and residuals

$$\begin{aligned} \widehat{\beta}_j(\ell) &= \frac{\langle r_{j,\ell-1}, \gamma_{j,\ell}\rangle}{\|\gamma_{j,\ell}\|_2^2}, \\ r_{j,\ell} &= r_{j,\ell-1} - \frac{\langle r_{j,\ell-1}, \gamma_{j,\ell}\rangle}{\|\gamma_{j,\ell}\|_2^2} \gamma_{j,\ell}. \end{aligned}$$

5. **Check for convergence:** If $\|r_{j,\ell}\|_2 > \epsilon \|y_j\|_2$ for all j, then increment ℓ and go to Step 2; otherwise, continue to Step 6. The parameter ϵ determines the target error power level allowed for algorithm convergence. Note that due to Step 3 the algorithm can only run for up to M iterations.

6. **De-orthogonalize:** Consider the relationship between $\Gamma_j = [\gamma_{j,1}, \gamma_{j,2}, \ldots, \gamma_{j,M}]$ and the Φ_j given by the QR factorization

$$\Phi_{j,\widehat{\Omega}} = \Gamma_j R_j,$$

where $\Phi_{j,\widehat{\Omega}} = [\phi_{j,n_1}, \phi_{j,n_2}, \ldots, \phi_{j,n_M}]$ is the so-called *mutilated basis*.[12] Since $y_j = \Gamma_j \beta_j = \Phi_{j,\widehat{\Omega}} x_{j,\widehat{\Omega}} = \Gamma_j R_j x_{j,\widehat{\Omega}}$, where $x_{j,\widehat{\Omega}}$ is the mutilated coefficient vector, we can compute the

[12] We define a *mutilated basis* Φ_Ω as a subset of the basis vectors from $\Phi = [\phi_1, \phi_2, \ldots, \phi_N]$ corresponding to the indices given by the set $\Omega = \{n_1, n_2, \ldots, n_M\}$, that is, $\Phi_\Omega = [\phi_{n_1}, \phi_{n_2}, \ldots, \phi_{n_M}]$. This concept can be extended to vectors in the same manner.

signal estimates $\{\widehat{x}_j\}$ as

$$\begin{aligned} \widehat{\theta}_{j,\widehat{\Omega}} &= R_j^{-1}\widehat{\beta}_j, \\ \widehat{x}_j &= \Psi\widehat{\theta}_j, \end{aligned}$$

where $\widehat{\theta}_{j,\widehat{\Omega}}$ is the mutilated version of the sparse coefficient vector $\widehat{\theta}_j$.

In practice, each sensor projects its signal x_j via $\Phi_j x_j$ to produce $\widehat{c}K$ measurements for some $\widehat{c}$. The decoder then applies DCS-SOMP to reconstruct the J signals jointly. We orthogonalize because as the number of iterations approaches M the norms of the residues of an orthogonal pursuit decrease faster than for a non-orthogonal pursuit.

Thanks to the common sparsity structure among the signals, we believe (but have not proved) that DCS-SOMP will succeed with $\widehat{c} < c(S)$. Empirically, we have observed that a small number of measurements proportional to K suffices for a moderate number of sensors J. We conjecture that $K+1$ measurements per sensor suffice as $J \to \infty$; numerical experiments are presented in Section 5.3. Thus, in practice, this efficient greedy algorithm enables an oversampling factor $\widehat{c} = (K+1)/K$ that approaches 1 as J, K, and N increase.

5.3 Simulations for JSM-2

We now present a simulation comparing separate CS reconstruction versus joint DCS-SOMP reconstruction for a JSM-2 signal ensemble. Figure 10 plots the probability of perfect reconstruction corresponding to various numbers of measurements M as the number of sensors varies from $J = 1$ to 32. We fix the signal lengths at $N = 50$ and the sparsity of each signal to $K = 5$.

With DCS-SOMP, for perfect reconstruction of all signals the average number of measurements per signal decreases as a function of J. The trend suggests that, for very large J, close to K measurements per signal should suffice. On the contrary, with separate CS reconstruction, for perfect reconstruction of all signals the number of measurements per sensor *increases* as a function of J. This surprise is due to the fact that each signal will experience an independent probability $p \leq 1$ of successful reconstruction; therefore the overall probability of complete success is p^J. Consequently, each sensor must compensate by making additional measurements. This phenomenon further motivates joint reconstruction under JSM-2.

Finally, we note that we can use algorithms other than DCS-SOMP to recover the signals under the JSM-2 model. Cotter et al. [53] have proposed additional algorithms (such as the M-FOCUSS algorithm) that iteratively eliminate basis vectors from the dictionary and converge to the set of sparse basis vectors over which the signals are supported. We hope to extend such algorithms to JSM-2 in future work.

6 Recovery Strategies for Nonsparse Common Component + Sparse Innovations (JSM-3)

The JSM-3 signal ensemble model from Section 3.3 provides a particularly compelling motivation for joint recovery. Under this model, no individual signal x_j is sparse, and so recovery of each signal separately would require fully N measurements per signal. As in the other JSMs, however, the commonality among the signals makes it possible to substantially reduce this number.

6.1 Recovery via Transpose Estimation of Common Component (TECC)

Successful recovery of the signal ensemble $\{x_j\}$ requires recovery of both the nonsparse common component z_C and the sparse innovations $\{z_j\}$. To illustrate the potential for signal recovery using

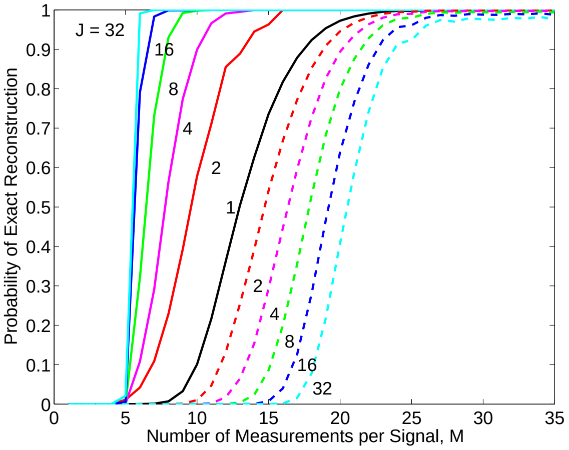


Figure 10: *Reconstructing a signal ensemble with common sparse supports (JSM-2). We plot the probability of perfect reconstruction via DCS-SOMP (solid lines) and independent CS reconstruction (dashed lines) as a function of the number of measurements per signal M and the number of signals J. We fix the signal length to $N = 50$, the sparsity to $K = 5$, and average over 1000 simulation runs. An oracle encoder that knows the positions of the large signal expansion coefficients would use 5 measurements per signal.*

far fewer than N measurements per sensor, consider the following gedankenexperiment. Again, for simplicity but without loss of generality, we assume $\Psi = I_N$.

If z_C were known, then each innovation z_j could be estimated using the standard single-signal CS machinery on the adjusted measurements

$$y_j - \Phi_j z_C = \Phi_j z_j.$$

While z_C is not known in advance, it can be *estimated* from the measurements. In fact, across all J sensors, a total of $\sum_j M_j$ random projections of z_C are observed (each corrupted by a contribution from one of the z_j). Since z_C is not sparse, it cannot be recovered via CS techniques, but when the number of measurements is sufficiently large ($\sum_j M_j \gg N$), z_C can be estimated using standard tools from linear algebra. A key requirement for such a method to succeed in recovering z_C is that each Φ_j be different, so that their rows combine to span all of $\mathbb{R}^N$. In the limit (again, assuming the sparse innovation coefficients are well-behaved), the common component z_C can be recovered while still allowing each sensor to operate at the minimum measurement rate dictated by the $\{z_j\}$. A prototype algorithm is listed below, where we assume that each measurement matrix Φ_j has i.i.d. $\mathcal{N}(0, \sigma_j^2)$ entries.

TECC Algorithm for JSM-3

1. **Estimate common component:** Define the matrix $\widehat{\Phi}$ as the concatenation of the regularized individual measurement matrices $\widehat{\Phi}_j = \frac{1}{M_j \sigma_j^2} \Phi_j$, that is, $\widehat{\Phi} = [\widehat{\Phi}_1, \widehat{\Phi}_2, \ldots, \widehat{\Phi}_J]$. Calculate the estimate of the common component as $\widehat{z_C} = \frac{1}{J} \widehat{\Phi}^T y$.

2. **Estimate measurements generated by innovations:** Using the previous estimate, subtract the contribution of the common part on the measurements and generate estimates for the measurements caused by the innovations for each signal: $\widehat{y}_j = y_j - \Phi_j \widehat{z_C}$.

3. **Reconstruct innovations:** Using a standard single-signal CS reconstruction algorithm, obtain estimates of the innovations $\widehat{z}_j$ from the estimated innovation measurements $\widehat{y}_j$.

4. **Obtain signal estimates:** Estimate each signal as the sum of the common and innovations estimates; that is, $\widehat{x}_j = \widehat{z_C} + \widehat{z}_j$.

The following theorem, proved in Appendix H, shows that asymptotically, by using the TECC algorithm, each sensor need only measure at the rate dictated by the sparsity K_j.

Theorem 10 *Assume that the nonzero expansion coefficients of the sparse innovations z_j are i.i.d. Gaussian random variables and that their locations are uniformly distributed on $\{1, 2, ..., N\}$. Then the following statements hold:*

1. *Let the measurement matrices Φ_j contain i.i.d. $\mathcal{N}(0, \sigma_j^2)$ entries with $M_j \geq K_j + 1$. Then each signal x_j can be recovered using the TECC algorithm with probability approaching one as $J \to \infty$.*

2. *Let Φ_j be a measurement matrix with $M_j \leq K_j$ for some $j \in \{1, 2, ..., J\}$. Then with probability one, the signal x_j cannot be uniquely recovered by any algorithm for any value of J.*

For large J, the measurement rates permitted by Statement 1 are the lowest possible for *any* reconstruction strategy on JSM-3 signals, even neglecting the presence of the nonsparse component. Thus, Theorem 10 provides a tight achievable and converse for JSM-3 signals. The CS technique employed in Theorem 10 involves combinatorial searches for estimating the innovation components. More efficient techniques could also be employed (including several proposed for CS in the presence of noise [38, 39, 45, 48, 51]). It is reasonable to expect similar behavior; as the error in estimating the common component diminishes, these techniques should perform similarly to their noiseless analogues (Basis Pursuit [45, 48], for example).

6.2 Recovery via Alternating Common and Innovation Estimation (ACIE)

The preceding analysis demonstrates that the number of required measurements in JSM-3 can be substantially reduced through joint recovery. While Theorem 10 suggests the theoretical gains as $J \to \infty$, practical gains can also be realized with a moderate number of sensors. For example, suppose in the TECC algorithm that the initial estimate $\widehat{z_C}$ is not accurate enough to enable correct identification of the sparse innovation supports $\{\Omega_j\}$. In such a case, it may still be possible for a rough approximation of the innovations $\{z_j\}$ to help refine the estimate $\widehat{z_C}$. This in turn could help to refine the estimates of the innovations. Since each component helps to estimate the other components, we propose an iterative algorithm for JSM-3 recovery.

The Alternating Common and Innovation Estimation (ACIE) algorithm exploits the observation that once the basis vectors comprising the innovation z_j have been identified in the index set Ω_j, their effect on the measurements y_j can be removed to aid in estimating z_C. Suppose that we have an estimate for these innovation basis vectors in $\widehat{\Omega}_j$. We can then partition the measurements into two parts: the projection into $\text{span}(\{\phi_{j,n}\}_{n \in \widehat{\Omega}_j})$ and the component orthogonal to that span. We build a basis for the $\mathbb{R}^{M_j}$ where y_j lives:

$$\mathbf{B}_j = [\Phi_{j,\widehat{\Omega}_j} \ Q_j],$$

where $\Phi_{j,\widehat{\Omega}_j}$ is the mutilated holographic basis corresponding to the indices in $\widehat{\Omega}_j$, and the $M_j \times (M_j - |\widehat{\Omega}_j|)$ matrix $Q_j = [q_{j,1} \ \ldots \ q_{j,M_j - |\widehat{\Omega}_j|}]$ has orthonormal columns that span the orthogonal complement of $\Phi_{j,\widehat{\Omega}_j}$.

This construction allows us to remove the projection of the measurements into the aforementioned span to obtain measurements caused exclusively by vectors not in $\widehat{\Omega}_j$

$$\begin{aligned} \widetilde{y}_j &= Q_j^T y_j, && (25) \\ \widetilde{\Phi}_j &= Q_j^T \Phi_j. && (26) \end{aligned}$$

These modifications enable the sparse decomposition of the measurement, which now lives in $\mathbb{R}^{M_j - |\widehat{\Omega}_j|}$, to remain unchanged

$$\widetilde{y}_j = \sum_{n=1}^{N} \alpha_j \widetilde{\phi}_{j,n}.$$

Thus, the modified measurements $\widetilde{Y} = \left[\widetilde{y}_1^T \ \widetilde{y}_2^T \ \ldots \ \widetilde{y}_J^T\right]^T$ and modified holographic basis $\widetilde{\Phi} = \left[\widetilde{\Phi}_1^T \ \widetilde{\Phi}_2^T \ \ldots \ \widetilde{\Phi}_J^T\right]^T$ can be used to refine the estimate of the measurements caused by the common part of the signal

$$\widetilde{z_C} = \widetilde{\Phi}^\dagger \widetilde{Y}, \tag{27}$$

where $A^\dagger = (A^T A)^{-1} A^T$ denotes the pseudoinverse of matrix A.

In the case where the innovation support estimate is correct ($\widehat{\Omega}_j = \Omega_j$), the measurements $\widetilde{y}_j$ will describe only the common component z_C. If this is true for every signal j and the number of remaining measurements $\sum_j M_j - KJ \geq N$, then z_C can be perfectly recovered via (27). However, it may be difficult to obtain correct estimates for all signal supports in the first iteration of the algorithm, and so we find it preferable to refine the estimate of the support by executing several iterations.

ACIE Algorithm for JSM-3

1. **Initialize:** Set $\widehat{\Omega}_j = \emptyset$ for each j. Set the iteration counter $\ell = 1$.

2. **Estimate common component:** Update estimate $\widetilde{z_C}$ according to (25)–(27).

3. **Estimate innovation supports:** For each sensor j, after subtracting the contribution $\widetilde{z_C}$ from the measurements, $\widehat{y}_j = y_j - \Phi_j \widetilde{z_C}$, estimate the sparse support of each signal innovation $\widehat{\Omega}_j$.

4. **Iterate:** If $\ell < L$, a preset number of iterations, then increment ℓ and return to Step 2. Otherwise proceed to Step 5.

5. **Estimate innovation coefficients:** For each signal j, estimate the coefficients for the indices in $\widehat{\Omega}_j$

   $$\widehat{\theta}_{j,\widehat{\Omega}_j} = \Phi^\dagger_{j,\widehat{\Omega}_j}(y_j - \Phi_j \widetilde{z_C}),$$

   where $\widehat{\theta}_{j,\widehat{\Omega}_j}$ is a mutilated version of the innovation's sparse coefficient vector estimate $\widehat{\theta}_j$.

6. **Reconstruct signals:** Compute the estimate of each signal as $\widehat{x}_j = \widetilde{z_C} + \widehat{z}_j = \widetilde{z_C} + \Phi_j \widehat{\theta}_j$.

Estimation of the sparse supports in Step 3 can be accomplished using a variety of techniques. We propose to run ℓ iterations of OMP; if the supports of the innovations are known to match across signals — as in the JSM-2 scenario — then more powerful algorithms like SOMP can be used.

6.3 Simulations for JSM-3

We now present simulations of JSM-3 reconstruction in the following scenario. Consider J signals of length $N = 50$ containing a common white noise component $z_C(n) \sim \mathcal{N}(0,1)$ for $n \in \{1, 2, \dots, N\}$ that, by definition, is not sparse in any fixed basis. Each innovations component z_j has sparsity $K = 5$ (once again in the time domain), resulting in $x_j = z_C + z_j$. The support for each innovations component is randomly selected with uniform probability from all possible supports for K-sparse, length-N signals. We draw the values of the innovation coefficients from a unit-variance Gaussian distribution.

We study two different cases. The first is an extension of JSM-1: we select the supports for the various innovations independently and then apply OMP independently to each signal in Step 3 of the ACIE algorithm in order to estimate its innovations component. The second case is an extension of JSM-2: we select one common support for all of the innovations across the signals and then apply the DCS-SOMP algorithm from Section 5.2 to estimate the innovations in Step 3. In both cases we set $L = 10$. We test the algorithms for different numbers of signals J and calculate the probability of correct reconstruction as a function of the (same) number of measurements per signal M.

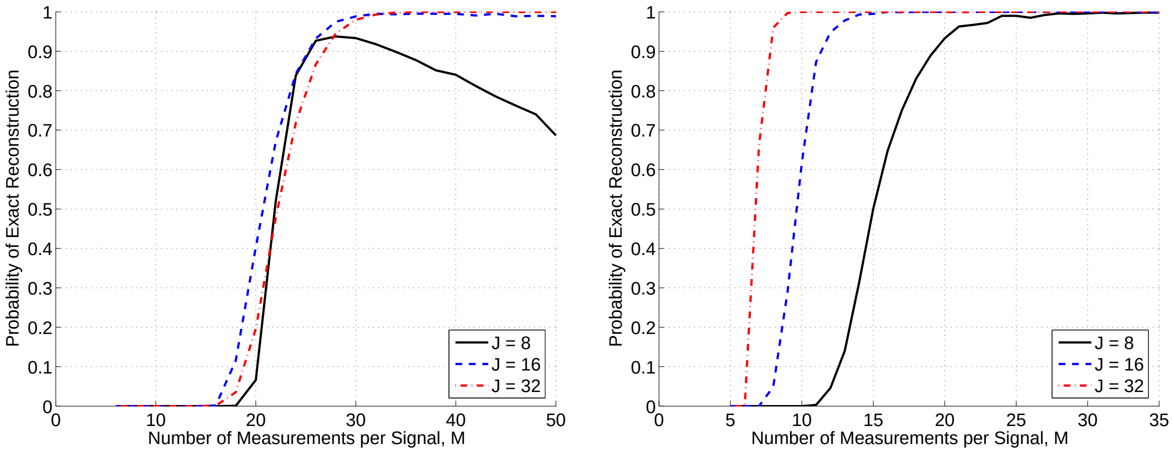
Figure 11(a) shows that, for sufficiently large J, we can recover all of the signals with significantly fewer than N measurements per signal. We note the following behavior in the graph. First, as J grows, it becomes more difficult to perfectly reconstruct all J signals. We believe this is inevitable, because even if z_C were known without error, then perfect ensemble recovery would require the successful execution of J *independent* runs of OMP. Second, for small J, the probability of success can decrease at high values of M. We believe this behavior is due to the fact that initial errors in estimating z_C may tend to be somewhat sparse (since $\widehat{z_C}$ roughly becomes an average of the signals $\{x_j\}$), and these sparse errors can mislead the subsequent OMP processes. For more moderate M, it seems that the errors in estimating z_C (though greater) tend to be less sparse. We expect that a more sophisticated algorithm could alleviate such a problem, and we note that the problem is also mitigated at higher J.

Figure 11(b) shows that when the sparse innovations share common supports we see an even greater savings. As a point of reference, a traditional approach to signal encoding would require 1600 total measurements to reconstruct these $J = 32$ nonsparse signals of length $N = 50$. Our approach requires only approximately 10 random measurements per sensor for a total of 320 measurements.

7 Discussion and Conclusions

In this paper we have taken the first steps towards extending the theory and practice of Compressed Sensing (CS) to multi-signal, distributed settings. Our three simple joint sparsity models (JSMs) for signal ensembles with both intra- and inter-signal correlations capture the essence of real physical scenarios, illustrate the basic analysis and algorithmic techniques, and indicate the significant gains to be realized from joint recovery. In some sense Distributed Compressed Sensing (DCS) is a framework for distributed compression of sources with memory, which has remained a challenging problem for some time.

For JSM-1, we have established a measurement rate region analogous to the Slepian-Wolf theorem [14], with converse and achievable bounds on performance. This required a careful analysis of equivalent viable representations that can explain the measurements. Simulations with our γ-weighted ℓ_1 signal recovery algorithm revealed that in practice the savings in the total number of required measurements can be substantial over separate CS encoding/decoding, especially when the common component dominates. In one of our scenarios with just two sensors, the savings in



(a) (b)

Figure 11: *Reconstructing a signal ensemble with nonsparse common component and sparse innovations (JSM-3) using ACIE. (a) Reconstruction using OMP independently on each signal in Step 3 of the ACIE algorithm (innovations have arbitrary supports). (b) Reconstruction using DCS-SOMP jointly on all signals in Step 3 of the ACIE algorithm (innovations have identical supports). Signal length $N = 50$, sparsity $K = 5$. The common structure exploited by DCS-SOMP enables dramatic savings in the number of measurements. We average over 1000 simulation runs.*

the number of measurements was as large as 30%.

For JSM-2, we have demonstrated using OSGA that important collective signal properties (the sparse support of the ensemble, for example) can be learned from as few as one measurement per signal. We also introduced DCS-SOMP, an efficient greedy algorithm for joint signal recovery based on the SOMP algorithm for simultaneous sparse approximation. DCS-SOMP features asymptotically best-possible performance that cannot be improved upon by any compression scheme: only roughly K measurements per signal are required in order to reconstruct an ensemble of K-sparse signals as their number grows. Simulations indicate that the asymptotics take effect with just a moderate number of signals.

For JSM-3, we have developed TECC and ACIE, procedures that recover a special class of nonsparse signal ensembles from few measurements, despite the fact that no individual signal in the ensemble could be recovered without the full number of measurements. These algorithms offer a surprising gain in performance: only roughly cK measurements per signal are required in order to reconstruct the ensemble of signals as their number grows (and only K measurements per signal are required if the innovations share common supports), essentially overcoming the complexity of sampling a nonsparse component. Finally, while we have used sparsity basis $\Psi = I_N$ in our theoretical development and simulations (which implies our signals are spikes in the time domain), we emphasize that our results hold for signals sparse in any basis.

There are many opportunities for applications and extensions of these ideas.

Application to sensor networks: The area that appears most likely to benefit immediately from the new DCS theory is low-powered sensor networks, where energy and communication bandwidth limitations require that we perform data compression while minimizing inter-sensor communications [11, 12]. DCS encoders work completely independently; therefore inter-sensor communication is required in a DCS-enabled sensor network only to support multi-hop networking to the data collection point. Moreover, the fact that typical sensor networks are designed to measure physical phenomena suggests that their data will possess considerable joint structure in the form of inter- and intra-signal correlations [40].

Compressible signals: In practice natural signals are not exactly ℓ_0 sparse but rather can be better modeled as ℓ_p sparse with $0 < p \leq 1$. Roughly speaking, a signal in a *weak*-ℓ_p ball has coefficients that decay as $n^{-1/p}$ once sorted according to magnitude [28]. The key concept is that the ordering of these coefficients is important. For JSM-2, we can extend the notion of simultaneous sparsity for ℓ_p-sparse signals whose sorted coefficients obey roughly the same ordering. This condition could perhaps be enforced as an ℓ_p constraint on the composite signal

$$\left\{ \sum_{j=1}^{J} |x_j(1)|, \ \sum_{j=1}^{J} |x_j(2)|, \ \ldots, \ \sum_{j=1}^{J} |x_j(N)| \right\}.$$

Quantized and noisy measurements: In general, (random) measurements will be real numbers; quantization will gradually degrade the reconstruction quality as the quantization becomes coarser [39]. Moreover, in many practical situations some amount of measurement noise will corrupt the $\{x_j\}$, making them not exactly sparse in any basis. While characterizing these effects and the resulting rate-distortion consequences in the DCS setting are topics for future work, there has been work in the single-signal CS literature that we should be able to leverage, including Basis Pursuit with Denoising [28, 45, 51, 58], robust iterative reconstruction algorithms [38], CS noise sensitivity analysis [27], and the Dantzig Selector [39].

Fast algorithms: In some applications, the linear program associated with some DCS decoders (in JSM-1 and JSM-3) could prove too computationally intense. As we saw in JSM-2, efficient iterative and greedy algorithms could come to the rescue, but these need to be extended to the multi-signal case. SOMP is a solid first step, and some progress has been made on fast algorithms for certain sparse signal classes, such as piecewise smooth signals in wavelet bases [35, 36].

Sparse signals with positive expansion coefficients: Tanner and Donoho have shown that the oversampling factor $c(S)$ required for perfect reconstruction drops dramatically when the sparse expansion coefficients are positive in some basis, that is, when $\theta_j(n) \geq 0$. While we cannot improve upon the (best-possible) theoretical performance of our algorithms for JSM-2 and JSM-3, the measurement rates in JSM-1 could benefit from this additional knowledge of the signal structure.

Acknowledgments

Thanks to Emmanuel Candès, Hyeokho Choi, Albert Cohen, Ron DeVore, Anna Gilbert, Rob Nowak, Jared Tanner, and Joel Tropp for informative and inspiring conversations. Special thanks to Jared Tanner for his code to compute the sparsity phase transition boundary in Figure 3 and to Mark Davenport for a thorough critique of our manuscript. Final thanks to Ryan King for supercharging our computational capabilities.

A Proof of Theorem 2

We first prove Statement 2, followed by Statements 1 and 3.

Statement 2 (Achievable, $M \geq K + 1$): Since Ψ is an orthonormal basis, it follows that entries of the $M \times N$ matrix $\Phi\Psi$ will be i.i.d. Gaussian. Thus without loss of generality, we assume Ψ to be the identity, $\Psi = I_N$, and so $y = \Phi\theta$. We concentrate on the "most difficult" case where $M = K + 1$; other cases follow similarly.

Let Ω be the index set corresponding to the nonzero entries of θ; we have $|\Omega| = K$. Also let Φ_Ω be the $M \times K$ mutilated matrix obtained by selecting the columns of Φ corresponding to the indices

Ω. The measurement y is then a linear combination of the K columns of Φ_Ω. With probability one, the columns of Φ_Ω are linearly independent. Thus, Φ_Ω will have rank K and can be used to recover the K nonzero entries of θ.

The coefficient vector θ can be uniquely determined if no other index set $\widehat{\Omega}$ can be used to explain the measurements y. Let $\widehat{\Omega} \neq \Omega$ be a different set of K indices (possibly with up to $K-1$ indices in common with Ω). We will show that (with probability one) y is not in the column span of $\Phi_{\widehat{\Omega}}$, where the column span of the matrix A is defined as the vector space spanned by the columns of A and denoted by colspan(A).

First, we note that with probability one, the columns of $\Phi_{\widehat{\Omega}}$ are linearly independent and so $\Phi_{\widehat{\Omega}}$ will have rank K. Now we examine the concatenation of these matrices $\left[\Phi_\Omega \;\; \Phi_{\widehat{\Omega}}\right]$. The matrix $\left[\Phi_\Omega \;\; \Phi_{\widehat{\Omega}}\right]$ cannot have rank K unless colspan$(\Phi_\Omega) =$ colspan$(\Phi_{\widehat{\Omega}})$, a situation that occurs with probability zero. Since these matrices have $M = K+1$ rows, it follows that $\left[\Phi_\Omega \;\; \Phi_{\widehat{\Omega}}\right]$ will have rank $K+1$; hence the column span is $\mathbb{R}^{K+1}$.

Since the combined column span of Φ_Ω and $\Phi_{\widehat{\Omega}}$ is $\mathbb{R}^{K+1}$ and since each matrix has rank K, it follows that colspan$(\Phi_\Omega) \cap$ colspan$(\Phi_{\widehat{\Omega}})$ is a $(K-1)$-dimensional linear subspace of $\mathbb{R}^{K+1}$. (Each matrix contributes one additional dimension to the column span.) This intersection is the set of measurements in the column span of Φ_Ω that could be confused with signals generated from the vectors $\widehat{\Omega}$. Based on its dimensionality, this set has measure zero in the column span of Φ_Ω; hence the probability that θ can be recovered using $\widehat{\Omega}$ is zero. Since the number of sets of K indices is finite, the probability that there exists $\widehat{\Omega} \neq \Omega$ that enables recovery of θ is zero.

Statement 1 (Achievable, $M \geq 2K$): We first note that, if $K \geq N/2$, then with probability one, the matrix Φ has rank N, and there is a unique (correct) reconstruction. Thus we assume that $K < N/2$. The proof of Statement 1 follows similarly to the proof of Statement 2. The key fact is that with probability one, all subsets of up to $2K$ columns drawn from Φ are linearly independent. Assuming this holds, then for two index sets $\Omega \neq \widehat{\Omega}$ such that $|\Omega| = |\widehat{\Omega}| = K$, colspan$(\Phi_\Omega) \cap$colspan$(\Phi_{\widehat{\Omega}})$ has dimension equal to the number of indices common to both Ω and $\widehat{\Omega}$. A signal projects to this common space only if its coefficients are nonzero on exactly these (fewer than K) common indices; since $\|\theta\|_0 = K$, this does not occur. Thus every K-sparse signal projects to a unique point in $\mathbb{R}^M$.

Statement 3 (Converse, $M \leq K$): If $M < K$, then there is insufficient information in the vector y to recover the K nonzero coefficients of θ; thus we assume $M = K$. In this case, there is a single explanation for the measurements only if there is a single set Ω of K linearly independent columns *and* the nonzero indices of θ are the elements of Ω. Aside from this pathological case, the rank of subsets $\Phi_{\widehat{\Omega}}$ will generally be less than K (which would prevent robust recovery of signals supported on $\widehat{\Omega}$) or will be equal to K (which would give ambiguous solutions among all such sets $\widehat{\Omega}$). □

B Proof of Theorem 4

Denote by $V = \Phi\widetilde{\Psi}$ the joint matrix whose columns should offer a sparse decomposition of y; we assume without loss of generality that $\Psi = I_N$. This proof essentially requires us to provide an extension of Theorem 2 to the multi-signal case; for generality we state the proof for an arbitrary number of signals $J \geq 2$. Recall from (7) that all minimal sparsity representations $(\overline{z_C}, \overline{z_1}, \overline{z_2}, \dots)$ have total sparsity $K'_C + \sum_j K'_j + J \cdot K_\cap$.

For this proof we will use the index 0 to denote the common component; that is, $z_0 \triangleq z_C$. For a particular minimal sparsity representation $(\overline{z_0}, \overline{z_1}, \overline{z_2}, \dots)$, we will also let $\Omega_j \subset \{1, 2, \dots, N\}$ be the set of nonzero indices corresponding to $\overline{z_j}$; note that $|\Omega_j| = \overline{K_j}$. More formally, we say that $n \in \Omega_j$

if and only if $\overline{z_j(n)} \neq 0$. We write $\Omega = (\Omega_0, \Omega_1, \dots, \Omega_J)$, where we define the tuple cardinality as the sum of the cardinalities of the elements: $|\Omega| = K'_C + \sum_j K'_j + J \cdot K_\cap$. We let V_Ω be the $(K'_C + \sum_j K'_j + J \cdot K_\cap) \times (K'_C + \sum_j K'_j + J \cdot K_\cap)$ matrix obtained by sampling the columns Ω from V, and for brevity we denote $V_{\Omega_j} = V_{(\emptyset, \dots, \emptyset, \Omega_j, \emptyset, \dots, \emptyset)}$, where Ω_j is the j-th element in the tuple.

Case $M_j = K'_j + K_\cap$ for some j: We will show that, if the first condition (15a) is violated (in particular by one measurement, though other cases follow similarly), then there exists a representation $(\widetilde{z_C}, \widetilde{z}_1, \widetilde{z}_2, \dots)$, that has total sparsity $K'_C + \sum_j K'_j + J \cdot K_\cap$ and explains the measurements y (and is therefore indistinguishable during reconstruction from any minimal sparsity representation) but does not generate the correct signals x.

We can construct such a representation as follows. Let $(\overline{z_C}, \overline{z_1}, \overline{z_2}, \dots)$ be a minimal sparsity representation such that $\overline{K_j} = K'_j + K_\cap$. To construct the representation $(\widetilde{z_C}, \widetilde{z}_1, \widetilde{z}_2, \dots)$, we modify only $\overline{z_j}$ by choosing an arbitrary $\widehat{\Omega}_j \subset \{1, 2, \dots, N\}$ such that $|\widehat{\Omega}_j| = |\Omega_j| = K'_j + K_\cap$ but $\widehat{\Omega}_j \neq \Omega_j$ and then assigning coefficients to $\widetilde{z}_j$ to ensure that $\Phi_j \widetilde{z}_j = \Phi_j \overline{z_j}$. As in Theorem 2, this is possible with probability one because the rank of Φ_j is $K'_j + K_\cap$. The new representation $(\widetilde{z_C}, \widetilde{z}_1, \widetilde{z}_2, \dots)$ will have total sparsity $K'_C + \sum_j K'_j + J \cdot K_\cap$; however since $\widetilde{z}_j \neq \overline{z_j}$ the reconstruction $\widetilde{x}_j = \widetilde{z_C} + \widetilde{z}_j$ will be incorrect. Thus x_j cannot be uniquely recovered if the first condition (15a) is violated.

Case $\sum_j M_j = K'_C + \sum_j K'_j + J \cdot K_\cap$ and $M_j \geq K'_j + K_\cap + 1 \ \forall j$: If the second condition (15b) is violated (again by just one measurement), we will again show the existence of an unfavorable representation $(\widetilde{z_C}, \widetilde{z}_1, \widetilde{z}_2, \dots)$.

We recall at this point that any minimal sparsity representation must have, for each $n \in \{1, \dots, N\}$, either $\overline{z(n)} = 0$ or have at least one j such that $\overline{z_j(n)} = 0$; otherwise a sparser representation exists. In other words, we have $\bigcap_{j=0}^{J} \Omega_j = \emptyset$. This also implies that, when Ω is the support of a minimum sparsity representation, V_Ω has full rank $K'_C + \sum_j K'_j + J \cdot K_\cap$. (Sets of columns of V are linearly independent with probability one, unless a column index n is included from each Ω_j, $j = 0, 1, \dots, J$.)

Let h be a column index such that $h \notin \Omega_0 \cup \Omega_j$ for some $j \in \{1, 2, \dots, J\}$. (If no such j exists, then all signals x_j are nonsparse and each requires a full N measurements.) Without loss of generality, we assume that $j = 1$. Now construct a new index set $\widehat{\Omega} = (\widehat{\Omega}_0, \widehat{\Omega}_1, \dots, \widehat{\Omega}_J)$ such that *(i)* $\widehat{\Omega}_0 = \Omega_0 \cup \{h\} \backslash \{n\}$ for some $n \in \Omega_0$ and *(ii)* $\widehat{\Omega}_j = \Omega_j$, for $j = 1, 2, \dots, J$. It follows that $|\widehat{\Omega}| = K'_C + \sum_j K'_j + J \cdot K_\cap$. The matrix $V_{\widehat{\Omega}}$ should also have full rank $K'_C + \sum_j K'_j + J \cdot K_\cap$ (since $\bigcap_{j=0}^{J} \widehat{\Omega}_j = \emptyset$), and so the indices $\widehat{\Omega}$ yield a feasible signal set $(\widehat{z_C}, \widehat{z}_1, \dots, \widehat{z_J})$ that explains the measurements y. We know that $\widehat{z_C}(h) \neq 0$, since otherwise the solution $(\widehat{z_C}, \widehat{z}_1, \dots, \widehat{z_J})$ would have sparsity less than $K'_C + \sum_j K'_j + J \cdot K_\cap$. From this it follows that $\widehat{x_1}(h) = \widehat{z_C}(h) \neq 0 = x_1(h)$, and so this solution will be incorrect. Since this solution produces the same measurement vector y and has the same sparsity $K'_C + \sum_j K'_j + J \cdot K_\cap$, no algorithm exists that could distinguish it from the correct solution. Thus both conditions (15a), (15b) must be satisfied to permit signal recovery. □

C Proof of Theorem 5

Suppose $M_j \geq K'_j + K_\cap + 1 \ \forall j$ and $\sum_j M_j \geq K'_C + \sum_j K'_j + J \cdot K_\cap + 1$. For this proof we consider the specific minimal sparsity representation $(z^\epsilon_C, z^\epsilon_1, z^\epsilon_2, \dots)$ known as the ϵ-point (defined in Section 4.3). We recall from (12a)–(12c) that

$$K^\epsilon_C = K'_C \quad \text{and} \quad K^\epsilon_j = K'_j + K_\cap \quad \forall j.$$

As in the proof of Theorem 4, we let Ω denote the support of this representation. As a minimal sparsity representation, the ϵ-point has total sparsity $K'_C + \sum_j K'_j + J \cdot K_\cap$. We aim to show that with probability one, any other representation that explains the measurements y and has the same (or lower) total sparsity will also yield the same signals $\{x_j\}$, and therefore solving the ℓ_0 minimization problem will yield the correct signals $\{x_j\}$. For this proof, we denote by $V_{j,n}$ the column of V corresponding to element n of component j; that is, $V_{j,n} \triangleq V_{(\emptyset,\ldots,n,\ldots,\emptyset)}$ with the j-th index set being nonempty.

As before, we select an index set $\widehat{\Omega}$, with $|\widehat{\Omega}| = K'_C + \sum_j K'_j + J \cdot K_\cap$, that could support another representation (that is, that explains the measurements y and has the same total sparsity). Without loss of generality, we can assume that $\widehat{\Omega}$ (like the ϵ-point) is concentrated into the innovation components. That is, for each n such that $\#\{j : n \in \widehat{\Omega}_j\} = J$, we have $n \notin \widehat{\Omega}_0$ and $n \in \widehat{\Omega}_j$, $j \in \{1, 2, \ldots, J\}$ (the common component is neglected in favor of the J innovations). We see that for any $\widehat{\Omega}$ that does not feature such concentration, there exists another $\widehat{\Omega}'$ that is concentrated in the innovation components for which $\text{colspan}(V_{\widehat{\Omega}}) = \text{colspan}(V_{\widehat{\Omega}'})$ and $|\widehat{\Omega}| = |\widehat{\Omega}'|$, and so any feasible solution with support $\widehat{\Omega}$ can also be described using support $\widehat{\Omega}'$.

We must consider several situations involving the measurement rates and the column spans.

Case 1: $|\widehat{\Omega}_j| < M_j$ **for all** j**:** For Ω and $\widehat{\Omega}$ to be feasible supports for sparsest solutions, it is required that V_Ω and $V_{\widehat{\Omega}}$ have rank at least $K'_C + \sum_j K'_j + J \cdot K_\cap$, as argued previously. The joint matrix $[V_\Omega \; V_{\widehat{\Omega}}]$ has a rank greater than $K'_C + \sum_j K'_j + J \cdot K_\cap$, unless $\text{colspan}(V_\Omega) = \text{colspan}(V_{\widehat{\Omega}})$.

Case 1(a): $\text{colspan}(V_\Omega) = \text{colspan}(V_{\widehat{\Omega}})$**:** In this case we argue that $\Omega = \widehat{\Omega}$. To see this, consider a vector $V_{j,n} \in \text{colspan}(V_\Omega)$ for $j \geq 1$. This can occur only if $n \in \Omega_j$; if $n \notin \Omega_j$ then because $|\Omega_j| = K'_j + K_\cap < M_j$, $\text{colspan}(V_{\Omega_j})$ contains $V_{j,n}$ only with probability zero. Similarly, with probability one, a column $V_{0,n} \in \text{colspan}(V_\Omega)$ only if $n \in \Omega_0$ or $\#\{j : n \in \Omega_j\} = J$.

Repeating these arguments for $\text{colspan}(V_{\widehat{\Omega}})$ (and exploiting the assumption that $|\widehat{\Omega}_j| < M_j$ for all j), we conclude that for $j \geq 1$, $V_{j,n} \in \text{colspan}(V_{\widehat{\Omega}})$ only if $n \in \widehat{\Omega}_j$, and that $V_{0,n} \in \text{colspan}(V_{\widehat{\Omega}})$ only if $n \in \widehat{\Omega}_0$ or $\#\{j : n \in \widehat{\Omega}_j\} = J$. Since $\text{colspan}(V_\Omega) = \text{colspan}(V_{\widehat{\Omega}})$, we conclude that $\Omega = \widehat{\Omega}$, and so trivially we see that the reconstructions x and $\widehat{x}$ must be equal.

Case 1(b): $\text{colspan}(V_\Omega) \neq \text{colspan}(V_{\widehat{\Omega}})$**:** For this case we mimic the arguments from the achievable proof of Theorem 2. Here $[V_\Omega \; V_{\widehat{\Omega}}]$ has rank at least $K'_C + \sum_j K'_j + J \cdot K_\cap + 1$, which in turn implies that $\text{colspan}(V_\Omega) \cap \text{colspan}(V_{\widehat{\Omega}})$ is a $\left(K'_C + \sum_j K'_j + J \cdot K_\cap - 1\right)$-dimensional subspace of $\mathbb{R}^{K'_C + \sum_j K'_j + J \cdot K_\cap + 1}$ (each matrix contributes one additional dimension, giving $[V_\Omega \; V_{\widehat{\Omega}}]$ a rank of $K'_C + \sum_j K'_j + J \cdot K_\cap + 1$). Once again, $\text{colspan}(V_\Omega) \cap \text{colspan}(V_{\widehat{\Omega}})$ is the subspace containing all measurements that can be explained by two distinct signals embedded in a space with dimension at least $K'_C + \sum_j K'_j + J \cdot K_\cap + 1$. Based on its dimensionality, this set has measure zero in the column span of V_Ω, and so an arbitrary set of signals with support given by Ω can be recovered with probability one.

Case 2: $|\widehat{\Omega}_j| \geq M_j$ **for some** j**:** In this case we conclude that $\text{colspan}(V_\Omega) \neq \text{colspan}(V_{\widehat{\Omega}})$. To see this we consider the signal j for which $|\widehat{\Omega}_j| \geq M_j$. Then with probability one every column $V_{j,n}$, $n \in \{1, 2, \ldots, N\}$ is contained in $\text{colspan}(V_{\widehat{\Omega}})$ because the $|\widehat{\Omega}_j|$ columns in $V_{\widehat{\Omega}_j}$ span the M_j-dimensional measurement space. However these columns $V_{j,n}$ cannot *all* be in $\text{colspan}(V_\Omega)$ because $|\Omega_j| = K'_j + K_\cap < M_j$ and so V_{Ω_j} has insufficient dimension to span this same space. Thus, $\text{colspan}(V_\Omega)$ and $\text{colspan}(V_{\widehat{\Omega}})$ must differ.

Since $\text{colspan}(V_\Omega) \neq \text{colspan}(V_{\widehat{\Omega}})$, the arguments from Case 1(b) above apply, and so we conclude that an arbitrary set of signals with support given by Ω can be recovered with probability one. □

D Proof of Lemma 2

Necessary conditions on innovation components: We begin by proving that in order to reconstruct $\overline{z_C}$, $\overline{z_1}$, and $\overline{z_2}$ via the γ-weighted ℓ_1 formulation it is necessary that $\overline{z_1}$ can be recovered via single-signal ℓ_1 CS reconstruction using Φ_1 and measurements $\widetilde{y_1} = \Phi_1 \overline{z_1}$.

Consider the single-signal ℓ_1 reconstruction problem

$$\widetilde{z_1} = \arg\min \|z_1\|_1 \quad \text{s.t. } \widetilde{y_1} = \Phi_1 z_1.$$

Suppose that this ℓ_1 reconstruction for $\overline{z_1}$ fails; that is, there exists $\widetilde{z_1} \neq \overline{z_1}$ such that $\widetilde{y_1} = \Phi_1 \widetilde{z_1}$ and $\|\widetilde{z_1}\|_1 \leq \|\overline{z_1}\|_1$. Therefore, substituting $\widetilde{z_1}$ instead of $\overline{z_1}$ in the γ-weighted ℓ_1 formulation (22) provides an alternate explanation for the measurements with a smaller or equal modified ℓ_1 penalty. Consequently, reconstruction of $\overline{z_1}$ using (22) will fail, and thus we will reconstruct x_1 incorrectly. We conclude that the single-signal ℓ_1 CS reconstruction of $\overline{z_1}$ using Φ_1 is necessary for successful reconstruction using the γ-weighted ℓ_1 formulation. A similar necessary condition for ℓ_1 CS reconstruction of $\overline{z_2}$ using Φ_2 and measurements $\Phi_2 \overline{z_2}$ can be proved in an analogous manner.

Necessary condition on common component: We now prove that in order to reconstruct $\overline{z_C}$, $\overline{z_1}$, and $\overline{z_2}$ via the γ-weighted ℓ_1 formulation it is necessary that $\overline{z_C}$ can be recovered via single-signal ℓ_1 CS reconstruction using the joint matrix $[\Phi_1^T \;\; \Phi_2^T]^T$ and measurements $[\Phi_1^T \;\; \Phi_2^T]^T \overline{z_C}$.

The proof is very similar to the previous proof for the innovation component $\overline{z_1}$. Consider the single-signal ℓ_1 reconstruction problem

$$\widetilde{z_C} = \arg\min \|z_C\|_1 \quad \text{s.t. } \widetilde{y_C} = [\Phi_1^T \;\; \Phi_2^T]^T \overline{z_C}.$$

Suppose that this ℓ_1 reconstruction for $\overline{z_C}$ fails; that is, there exists $\widetilde{z_C} \neq \overline{z_C}$ such that $\widetilde{y_C} = [\Phi_1^T \;\; \Phi_2^T]^T \widetilde{z_C}$ and $\|\widetilde{z_C}\|_1 \leq \|\overline{z_C}\|_1$. Therefore, substituting $\widetilde{z_C}$ instead of $\overline{z_C}$ in the γ-weighted ℓ_1 formulation (22) provides an alternate explanation for the measurements with a smaller modified ℓ_1 penalty. Consequently, the reconstruction of $\overline{z_C}$ using the γ-weighted ℓ_1 formulation (22) will fail, and thus we will reconstruct x_1 and x_2 incorrectly. We conclude that the single-signal ℓ_1 reconstruction of $\overline{z_C}$ using $[\Phi_1^T \;\; \Phi_2^T]^T$ is necessary for successful reconstruction using the γ-weighted ℓ_1 formulation. □

E Proof of Theorem 6

Consider our γ-weighted ℓ_1 formulation (22). Among minimal sparsity representations, there is one that has smallest γ-weighted ℓ_1 norm. We call this the *star representation* and denote its components by z_C^*, z_1^*, and z_2^*. If the components z_C^*, z_1^*, and z_2^* cannot be recovered, then the solution of the γ-weighted ℓ_1 minimization has a smaller ℓ_1 norm. But the star representation has the smallest γ-weighted ℓ_1 norm among minimal sparsity representations and thus among viable representations. Therefore, correct reconstruction of the components z_C^*, z_1^*, and z_2^* is a necessary condition for reconstruction of x_1 and x_2.

Conditions on individual rates: Using Lemma 1, the innovation components z_1^* and z_2^* must have sparsity rate no smaller than S_I'. Lemma 2 requires single-signal ℓ_1 reconstruction of z_1^* and z_2^* using Φ_1 and Φ_2, respectively. The converse bound of Theorem 3 indicates that $R_1 \geq c'(S_I')$ and $R_2 \geq c'(S_I')$ are necessary conditions for single-signal ℓ_1 reconstruction of z_1^* and z_2^*, respectively. Combining these observations, these conditions on R_1 and R_2 are necessary for the γ-weighted ℓ_1 formulation (22) to reconstruct x_1 and x_2 correctly as N increases.

Condition on sum rate: Using Lemma 1, z_C^* must have sparsity rate no smaller than S_C'. Lemma 2 requires single-signal ℓ_1 reconstruction of z_C^* using $[\Phi_1^T \;\; \Phi_2^T]^T$. The converse bound of

Theorem 3 indicates that $R_1+R_2 \geq c'(S'_C)$ is a necessary condition for single-signal ℓ_1 reconstruction of z^*_C. Combining these observations, the condition $R_1+R_2 \geq c'(S'_C)$ is necessary for the γ-weighted ℓ_1 formulation (22) to reconstruct x_1 and x_2 correctly as N increases. □

F Proof of Theorem 7

We construct measurement matrices Φ_1 and Φ_2 that consist of two sets of rows. The first set of rows is common to both and reconstructs the signal *difference* $x_1 - x_2$. The second set is different and reconstructs the signal *average* $\frac{1}{2}x_1 + \frac{1}{2}x_2$. Let the submatrix formed by the common rows for the signal difference be Φ_D, and let the submatrices formed by unique rows for the signal average be $\Phi_{A,1}$ and $\Phi_{A,2}$. In other words, the measurement matrices Φ_1 and Φ_2 are of the following form:

$$\Phi_1 = \left[\begin{array}{c} \Phi_D \\ -- \\ \Phi_{A,1} \end{array} \right] \text{ and } \Phi_2 = \left[\begin{array}{c} \Phi_D \\ -- \\ \Phi_{A,2} \end{array} \right].$$

The submatrices Φ_D, $\Phi_{A,1}$, and $\Phi_{A,2}$ contain i.i.d. Gaussian entries. Once the difference $x_1 - x_2$ and average $\frac{1}{2}x_1 + \frac{1}{2}x_2$ have been reconstructed using the above technique, the computation of x_1 and x_2 is straightforward. The measurement rate can be computed by considering both parts of the measurement matrices.

Reconstruction of signal difference: The submatrix Φ_D is used to reconstruct the signal difference. By subtracting the product of Φ_D with the signals x_1 and x_2, we have

$$\Phi_D x_1 - \Phi_D x_2 = \Phi_D(x_1 - x_2).$$

In the original representation we have $x_1 - x_2 = z_1 - z_2$ with sparsity rate $2S_I$. But $z_1(n) - z_2(n)$ is nonzero only if $z_1(n)$ is nonzero or $z_2(n)$ is nonzero. Therefore, the sparsity rate of $x_1 - x_2$ is equal to the sum of the individual sparsities reduced by the sparsity rate of the overlap, and so we have $S(X_1 - X_2) = 2S_I - (S_I)^2$. Therefore, any measurement rate greater than $c'(2S_I - (S_I)^2)$ for each Φ_D permits reconstruction of the length N signal $x_1 - x_2$. (As always, the probability of correct reconstruction approaches one as N increases.)

Reconstruction of average: Once $x_1 - x_2$ has been reconstructed, we have

$$x_1 - \frac{1}{2}(x_1 - x_2) = \frac{1}{2}x_1 + \frac{1}{2}x_2 = x_2 + \frac{1}{2}(x_1 - x_2).$$

At this stage, we know $x_1 - x_2$, $\Phi_D x_1$, $\Phi_D x_2$, $\Phi_{A,1}x_1$, and $\Phi_{A,2}x_2$. We have

$$\begin{aligned} \Phi_D x_1 - \frac{1}{2}\Phi_D(x_1 - x_2) &= \Phi_D \left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right), \\ \Phi_{A,1}x_1 - \frac{1}{2}\Phi_{A,1}(x_1 - x_2) &= \Phi_{A,1} \left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right), \\ \Phi_{A,2}x_2 + \frac{1}{2}\Phi_{A,2}(x_1 - x_2) &= \Phi_{A,2} \left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right), \end{aligned}$$

where $\Phi_D(x_1 - x_2)$, $\Phi_{A,1}(x_1 - x_2)$, and $\Phi_{A,2}(x_1 - x_2)$ are easily computable because $(x_1 - x_2)$ has been reconstructed. The signal $\frac{1}{2}x_1 + \frac{1}{2}x_2$ is of length N, and its sparsity rate is clearly upper bounded by $S_C + 2S_I$. The exact sparsity rate is in fact less, since the supports of the original components z_C, z_1, and z_2 overlap. In fact, $\frac{1}{2}x_1 + \frac{1}{2}x_2$ is nonzero only if at least one of $z_C(n)$, $z_1(n)$, and $z_2(n)$ are nonzero. Therefore, the sparsity rate of $\frac{1}{2}x_1 + \frac{1}{2}x_2$ is equal to the sum of the individual sparsities $S_C + 2S_I$ reduced by the sparsity rate of the overlaps, and so we have $S(\frac{1}{2}X_1 + \frac{1}{2}X_2) = S_C + 2S_I - 2S_C S_I - (S_I)^2 + S_C(S_I)^2$. Therefore, any measurement rate greater than $c'(S_C + 2S_I - 2S_C S_I - (S_I)^2 + S_C(S_I)^2)$ aggregated over the matrices Φ_D, $\Phi_{A,1}$, and $\Phi_{A,2}$ enables reconstruction of $\frac{1}{2}x_1 + \frac{1}{2}x_2$.

Computation of measurement rate: By considering the requirements on Φ_D, the individual measurement rates R_1 and R_2 must satisfy (23a) and (23b), respectively. Combining the measurement rates required for $\Phi_{A,1}$ and $\Phi_{A,2}$, the sum measurement rate satisfies (23c). We complete the proof by noting that $c'(\cdot)$ is continuous and that $\lim_{S \to 0} c'(S) = 0$, and so the limit of the sum measurement rate as S_I goes to zero is $c'(S)$. □

G Proof of Theorem 8

We again assume that Ψ is an orthonormal matrix. Like Φ_j itself, the matrix $\Phi_j \Psi$ also has i.i.d. $\mathcal{N}(0,1)$ entries, since Ψ is orthonormal. For convenience, we assume $\Psi = I_N$. The results presented can be easily extended to a more general orthonormal matrix Ψ by replacing Φ_j with $\Phi_j \Psi$.

Assume without loss of generality that $\Omega = \{1, 2, \ldots, K\}$ for convenience of notation. Thus, the correct estimates are $n \le K$, and the incorrect estimates are $n \ge K+1$. Now consider the statistic ξ_n in (24). This is the sample mean of J i.i.d. variables. The variables $\langle y_j, \phi_{j,n} \rangle^2$ are i.i.d. since each $y_j = \Phi_j x_j$, and Φ_j and x_j are i.i.d. Furthermore, these variables have a finite variance.[13] Therefore, we invoke the Law of Large Numbers (LLN) to argue that ξ_n, which is a sample mean of $\langle y_j, \phi_{j,n} \rangle^2$, converges to $E[\langle y_j, \phi_{j,n} \rangle^2]$ as J grows large. We now compute $E[\langle y_j, \phi_{j,n} \rangle^2]$ under two cases. In the first case, we consider $n \ge K+1$ (we call this the "bad statistics case"), and in the second case, we consider $n \le K$ ("good statistics case").

Bad statistics: Consider one of the bad statistics by choosing $n = K+1$ without loss of generality. We have

$$\begin{aligned}
E[\langle y_j, \phi_{j,K+1} \rangle^2] &= E\left[\sum_{n=1}^{K} x_j(n) \langle \phi_{j,n}, \phi_{j,K+1} \rangle\right]^2 \\
&= E\left[\sum_{n=1}^{K} x_j(n)^2 \langle \phi_{j,n}, \phi_{j,K+1} \rangle^2\right] \\
&\quad + E\left[\sum_{n=1}^{K} \sum_{\ell=1, \ell \ne n}^{K} x_j(\ell) x_j(n) \langle \phi_{j,\ell}, \phi_{j,K+1} \rangle \langle \phi_{j,n}, \phi_{j,K+1} \rangle\right] \\
&= \sum_{n=1}^{K} E\left[x_j(n)^2\right] E\left[\langle \phi_{j,n}, \phi_{j,K+1} \rangle^2\right] \\
&\quad + \sum_{n=1}^{K} \sum_{\ell=1, \ell \ne n}^{K} E[x_j(\ell)] E[x_j(n)] E\left[\langle \phi_{j,\ell}, \phi_{j,K+1} \rangle \langle \phi_{j,n}, \phi_{j,K+1} \rangle\right]
\end{aligned}$$

[13] In [56], we evaluate the variance of $\langle y_j, \phi_{j,n} \rangle^2$ as

$$\mathrm{Var}[\langle y_j, \phi_{j,n} \rangle^2] = \begin{cases} M\sigma^4(34MK + 6K^2 + 28M^2 + 92M + 48K + 90 + 2M^3 + 2MK^2 + 4M^2K), & n \in \Omega \\ 2MK\sigma^4(MK + 3K + 3M + 6), & n \notin \Omega. \end{cases}$$

For finite M, K and σ, the above variance is finite.

since the terms are independent. We also have $E[x_j(n)] = E[x_j(\ell)] = 0$, and so

$$\begin{aligned} E[\langle y_j, \phi_{j,K+1}\rangle^2] &= \sum_{n=1}^{K} E\left[x_j(n)^2\right] E\left[\langle \phi_{j,n}, \phi_{j,K+1}\rangle^2\right] \\ &= \sum_{n=1}^{K} \sigma^2 E\left[\langle \phi_{j,n}, \phi_{j,K+1}\rangle^2\right]. \end{aligned} \tag{28}$$

To compute $E\left[\langle \phi_{j,n}, \phi_{j,K+1}\rangle^2\right]$, let $\phi_{j,n}$ be the column vector $[a_1, a_2, ..., a_M]^T$, where each element in the vector is i.i.d. $\mathcal{N}(0,1)$. Likewise, let $\phi_{j,K+1}$ be the column vector $[b_1, b_2, ..., b_M]^T$ where the elements are i.i.d. $\mathcal{N}(0,1)$. We have

$$\begin{aligned} \langle \phi_{j,n}, \phi_{j,K+1}\rangle^2 &= (a_1b_1 + a_2b_2 + ... + a_Mb_M)^2 \\ &= \sum_{m=1}^{M} a_m^2 b_m^2 + 2\sum_{m=1}^{M-1}\sum_{r=m+1}^{M} a_m a_r b_m b_r. \end{aligned}$$

Taking the expected value, we have

$$\begin{aligned} E\left[\langle \phi_{j,n}, \phi_{j,K+1}\rangle^2\right] &= E\left[\sum_{m=1}^{M} a_m^2 b_m^2\right] + 2E\left[\sum_{m=1}^{M-1}\sum_{r=m+1}^{M} a_m a_r b_m b_r\right] \\ &= \sum_{m=1}^{M} E\left[a_m^2 b_m^2\right] + 2\sum_{m=1}^{M-1}\sum_{r=q+1}^{M} E\left[a_m a_r b_m b_r\right] \\ &= \sum_{m=1}^{M} E\left[a_m^2\right] E\left[b_m^2\right] + 2\sum_{m=1}^{M-1}\sum_{r=m+1}^{M} E\left[a_m\right] E\left[a_r\right] E\left[b_m\right] E\left[b_r\right] \\ &\qquad \text{(since the random variables are independent)} \\ &= \sum_{m=1}^{M} (1) + 0 \\ &\qquad \text{(since } E\left[a_m 2\right] = E\left[b_m 2\right] = 1 \text{ and } E\left[a_m\right] = E\left[b_m\right] = 0\text{)} \\ &= M \end{aligned}$$

and thus

$$E\left[\langle \phi_{j,n}, \phi_{j,K+1}\rangle^2\right] = M. \tag{29}$$

Combining this result with (28), we find that

$$E[\langle y_j, \phi_{j,K+1}\rangle^2] = \sum_{n=1}^{K} \sigma^2 M = MK\sigma^2.$$

Thus we have computed $E[\langle y_j, \phi_{j,K+1}\rangle^2]$ and can conclude that as J grows large, the statistic ξ_{K+1} converges to

$$E[\langle y_j, \phi_{j,K+1}\rangle^2] = MK\sigma^2. \tag{30}$$

Good statistics: Consider one of the good statistics, and choose $n = 1$ without loss of generality. Then, we have

$$\begin{aligned}
E[\langle y_j, \phi_{j,1}\rangle^2] &= E\left[\left(x_j(1)\|\phi_{j,1}\|^2 + \sum_{n=2}^{K} x_j(n)\langle\phi_{j,n}, \phi_{j,1}\rangle\right)^2\right] \\
&= E\left[(x_j(1))^2 \|\phi_{j,1}\|^4\right] + E\left[\sum_{n=2}^{K} x_j(n)^2\langle\phi_{j,n}, \phi_{j,1}\rangle^2\right] \\
&\qquad\qquad\qquad \text{(all other cross terms have zero expectation)} \\
&= E\left[x_j(1)^2\right] E\left[\|\phi_{j,1}\|^4\right] + \sum_{n=2}^{K} E\left[x_j(n)^2\right] E\left[\langle\phi_{j,n}, \phi_{j,1}\rangle^2\right] \qquad \text{(by independence)} \\
&= \sigma^2 E\left[\|\phi_{j,1}\|^4\right] + \sum_{n=2}^{K} \sigma^2 E\left[\langle\phi_{j,n}, \phi_{j,1}\rangle^2\right]. \qquad (31)
\end{aligned}$$

Extending the result from (29), we can show that $E\langle\phi_{j,n}, \phi_{j,1}\rangle^2 = M$. Using this result in (31), we find that

$$E[\langle y_j, \phi_{j,1}\rangle^2] = \sigma^2 E\|\phi_{j,1}\|^4 + \sum_{n=2}^{K} \sigma^2 M. \qquad (32)$$

To evaluate $E\left[\|\phi_{j,1}\|^4\right]$, let $\phi_{j,1}$ be the column vector $[c_1, c_2, ..., c_M]^T$, where the elements of the vector are random $\mathcal{N}(0,1)$. Define the random variable $Z = \|\phi_{j,1}\|^2 = \sum_{m=1}^{M} c_m^2$. Note that $E\left[\|\phi_{j,1}\|^4\right] = E\left[Z^2\right]$. From the theory of random variables, we know that Z is chi-squared distributed with M degrees of freedom. Thus, $E\left[\|\phi_{j,1}\|^4\right] = E\left[Z^2\right] = M(M+2)$. Using this result in (32), we have

$$\begin{aligned}
E[\langle y_j, \phi_{j,1}\rangle^2] &= \sigma^2 M(M+2) + (K-1)\sigma^2 M \\
&= M(M+K+1)\sigma^2.
\end{aligned}$$

We have computed the variance of $\langle y_j, \phi_{j,1}\rangle$ and can conclude that as J grows large, the statistic ξ_1 converges to

$$E[\langle y_j, \phi_{j,1}\rangle^2] = (M+K+1)M\sigma^2. \qquad (33)$$

Conclusion: From (30) and (33) we conclude that

$$\lim_{J\to\infty} \xi_n = E[\langle y_j, \phi_{j,n}\rangle^2] = \begin{cases} (M+K+1)M\sigma^2, & n \in \Omega \\ KM\sigma^2, & n \notin \Omega. \end{cases}$$

For any $M \geq 1$, these values are distinct, with a ratio of $\frac{M+K+1}{K}$ between them. Therefore, as J increases we can distinguish between the two expected values of ξ_n with overwhelming probability. □

H Proof of Theorem 10

Statement 2 follows trivially from Theorem 2 (simply assume that z_C is known a priori). The proof of Statement 1 has two parts. First we argue that $\lim_{J\to\infty} \widehat{z_C} = z_C$. Second we show that this implies vanishing probability of error in recovering each innovation z_j.

Part 1: We can write our estimate as

$$\begin{aligned}\widehat{z_C} &= \frac{1}{J}\widehat{\Phi}^T y = \frac{1}{J}\widehat{\Phi}^T \Phi x = \frac{1}{J}\sum_{j=1}^{J} \frac{1}{M_j\sigma_j^2}\Phi_j^T\Phi_j x_j \\ &= \frac{1}{J}\sum_{j=1}^{J}\frac{1}{M_j\sigma_j^2}\sum_{m=1}^{M_j}(\phi_{j,m}^R)^T\phi_{j,m}^R x_i,\end{aligned}$$

where Φ is a diagonal concatenation of the Φ_j's as defined in (13), and $\phi_{j,m}^R$ denotes the m-th row of Φ_j, that is, the m-th measurement vector for node j. Since the elements of each Φ_j are Gaussians with variance σ_j^2, the product $(\phi_{j,m}^R)^T\phi_{j,m}^R$ has the property

$$E[(\phi_{j,m}^R)^T\phi_{j,m}^R] = \sigma_j^2 I_N.$$

It follows that

$$E[(\phi_{j,m}^R)^T\phi_{j,m}^R x_j] = \sigma_j^2 E[x_j] = \sigma_j^2 E[z_C + z_j] = \sigma_j^2 z_C$$

and, similarly, that

$$E\left[\frac{1}{M_j\sigma_j^2}\sum_{m=1}^{M_j}(\phi_{j,m}^R)^T\phi_{j,m}^R x_j\right] = z_C.$$

Thus, $\widehat{z_C}$ is a sample mean of J independent random variables with mean z_C. From the LLN, we conclude that

$$\lim_{J\to\infty}\widehat{z_C} = z_C.$$

Part 2: Consider recovery of the innovation z_j from the adjusted measurement vector $\widehat{y_j} = y_j - \Phi_j\widehat{z_C}$. As a recovery scheme, we consider a combinatorial search over all K-sparse index sets drawn from $\{1, 2, \ldots, N\}$. For each such index set Ω', we compute the distance from $\widehat{y}$ to the column span of $\Phi_{j,\Omega'}$, denoted by $d(\widehat{y}, \text{colspan}(\Phi_{j,\Omega'}))$, where $\Phi_{j,\Omega'}$ is the matrix obtained by sampling the columns Ω' from Φ_j. (This distance can be measured using the pseudoinverse of $\Phi_{j,\Omega'}$.)

For the correct index set Ω, we know that $d(\widehat{y_j}, \text{colspan}(\Phi_{j,\Omega})) \to 0$ as $J \to \infty$. For any other index set Ω', we know from the proof of Theorem 2 that $d(\widehat{y_j}, \text{colspan}(\Phi_{j,\Omega'})) > 0$. Let

$$\zeta \triangleq \min_{\Omega'\neq\Omega} d(\widehat{y_j}, \text{colspan}(\Phi_{i,\Omega'})).$$

With probability one, $\zeta > 0$. Thus for sufficiently large J, we will have $d(\widehat{y_j}, \text{colspan}(\Phi_{j,\Omega})) < \zeta/2$, and so the correct index set Ω can be correctly identified. □

References

[1] D. Baron, M. F. Duarte, S. Sarvotham, M. B. Wakin, and R. G. Baraniuk, "An information-theoretic approach to distributed compressed sensing," in *Proc. 43rd Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sept. 2005.

[2] D. Baron, M. F. Duarte, S. Sarvotham, M. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proc. 39th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2005.

[3] 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.

[4] R. A. DeVore, B. Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 719–746, Mar. 1992.

[5] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.

[6] Z. Xiong, K. Ramchandran, and M. T. Orchard, "Space-frequency quantization for wavelet image coding," *IEEE Trans. Image Processing*, vol. 6, no. 5, pp. 677–693, 1997.

[7] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1999.

[8] K. Brandenburg, "MP3 and AAC explained," in *AES 17th International Conference on High-Quality Audio Coding*, Sept. 1999.

[9] W. Pennebaker and J. Mitchell, "JPEG: Still image data compression standard," *Van Nostrand Reinhold*, 1993.

[10] D. S. Taubman and M. W. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards and Practice*, Kluwer, 2001.

[11] D. Estrin, D. Culler, K. Pister, and G. Sukhatme, "Connecting the physical world with pervasive networks," *IEEE Pervasive Computing*, vol. 1, no. 1, pp. 59–69, 2002.

[12] G. J. Pottie and W. J. Kaiser, "Wireless integrated network sensors," *Comm. ACM*, vol. 43, no. 5, pp. 51–58, 2000.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[14] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, July 1973.

[15] S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *IEEE Trans. Inform. Theory*, vol. 49, pp. 626–643, Mar. 2003.

[16] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Mag.*, vol. 21, pp. 80–94, Sept. 2004.

[17] J. Wolfowitz, *Coding Theorems of Information Theory*, Springer-Verlag, Berlin, 1978.

[18] H. Luo and G. Pottie, "Routing explicit side information for data compression in wireless sensor networks," in *Int. Conf. on Distirbuted Computing in Sensor Systems (DCOSS)*, Marina Del Rey, CA, June 2005.

[19] B. Krishnamachari, D. Estrin, and S. Wicker, "Modelling data-centric routing in wireless sensor networks," *USC Computer Engineering Technical Report CENG 02-14*, 2002.

[20] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "On network correlated data gathering," in *Proc. INFOCOM 2004.*, HongKong, Mar. 2004.

[21] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed Karhunen-Loeve transform," *IEEE Trans. Inform. Theory*, Nov. 2004, Submitted.

[22] R. Wagner, V. Delouille, H. Choi, and R. G. Baraniuk, "Distributed wavelet transform for irregular sensor network grids," in *IEEE Statistical Signal Processing (SSP) Workshop*, Bordeaux, France, July 2005.

[23] A. Ciancio and A. Ortega, "A distributed wavelet compression algorithm for wireless multihop sensor networks using lifting," in *IEEE 2005 Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[24] D. Ganesan, B. Greenstein, D. Perelyubskiy, D. Estrin, and J. Heidemann, "An evaluation of multi-resolution storage for sensor networks," in *Proc. ACM SenSys Conference*, Los Angeles, Nov. 2003, pp. 89–102.

[25] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inform. Theory*, vol. 21, pp. 226–228, Mar. 1975.

[26] T. Uyematsu, "Universal coding for correlated sources with memory," in *Canadian Workshop Inform. Theory*, Vancouver, June 2001.

[27] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, 2004, Submitted.

[28] D. Donoho, "Compressed sensing," 2004, Preprint.

[29] E. Candès and T. Tao, "Near optimal signal recovery from random projections and universal encoding strategies," *IEEE Trans. Inform. Theory*, 2004, Submitted.

[30] J. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," Apr. 2005, Preprint.

[31] E. Candès and T. Tao, "Error correction via linear programming," *Found. of Comp. Math.*, 2005, Submitted.

[32] D. Donoho and J. Tanner, "Neighborliness of randomly projected simplices in high dimensions," Mar. 2005, Preprint.

[33] D. Donoho, "High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension," Jan. 2005, Preprint.

[34] J. Tropp, A. C. Gilbert, and M. J. Strauss, "Simulataneous sparse approximation via greedy pursuit," in *IEEE 2005 Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[35] 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.

[36] C. La and M. N. Do, "Signal reconstruction using sparse tree representation," in *Proc. Wavelets XI at SPIE Optics and Photonics*, San Diego, Aug. 2005.

[37] D. Takhar, V. Bansal, M. Wakin, M. Duarte, D. Baron, J. Laska, K. F. Kelly, and R. G. Baraniuk, "A compressed sensing camera: New theory and an implementation using digital micromirrors," in *Proc. Computational Imaging IV at SPIE Electronic Imaging*, San Jose, Jan. 2006.

[38] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, 2005, Submitted.

[39] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, 2005, Submitted.

[40] M. F. Duarte, M. B. Wakin, D. Baron, and R. G. Baraniuk, "Universal distributed sensing via random projections," in *5th International Workshop on Inf. Processing in Sensor Networks (IPSN '06)*, 2006, Submitted.

[41] D. Baron and Y. Bresler, "An $O(N)$ semi-predictive universal encoder via the BWT," *IEEE Trans. Inform. Theory*, vol. 50, no. 5, pp. 928–937, 2004.

[42] M. Effros, K. Visweswariah, S.R. Kulkarni, and S. Verdu, "Universal lossless source coding with the Burrows Wheeler transform," *IEEE Trans. Inform. Theory*, vol. 48, no. 5, pp. 1061–1081, 2002.

[43] E. Candès and D. Donoho, "Curvelets — A surprisingly effective nonadaptive representation for objects with edges," *Curves and Surfaces*, 1999.

[44] E. Candès and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Found. of Comp. Math.*, 2004, Submitted.

[45] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. on Pure and Applied Math.*, 2005, Submitted.

[46] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, Dec. 2005.

[47] E. Candès and J. Romberg, "Practical signal recovery from random projections," *IEEE Trans. Signal Processing*, 2005, Submitted.

[48] D. Donoho and Y. Tsaig, "Extensions of compressed sensing," 2004, Preprint.

[49] "Compressed sensing website," `http://www.dsp.rice.edu/cs/`.

[50] R. Venkataramani and Y. Bresler, "Further results on spectrum blind sampling of 2D signals," in *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, Chicago, Oct. 1998, vol. 2.

[51] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. on Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[52] V. N. Temlyakov, "A remark on simultaneous sparse approximation," *East J. Approx.*, vol. 100, pp. 17–25, 2004.

[53] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, vol. 51, pp. 2477–2488, July 2005.

[54] R. Puri and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles," in *Proc. 40th Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2002.

[55] R. Wagner, R. G. Baraniuk, and R. D. Nowak, "Distributed image compression for sensor networks using correspondence analysis and super-resolution," in *Proc. Data Comp. Conf. (DCC)*, Mar. 2000.

[56] S. Sarvotham, M. B. Wakin, D. Baron, M. F. Duarte, and R. G. Baraniuk, "Analysis of the DCS one-stage greedy algoritm for common sparse supports," Tech. Rep., Rice University ECE Department, Oct. 2005, available at http://www.ece.rice.edu/~shri/docs/TR0503.pdf.

[57] J. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *EURASIP J. App. Signal Processing*, 2005, To appear.

[58] J. J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. Inform. Theory*, vol. 51, pp. 3601–3608, Oct. 2005.