

Compressed Sensing: Theory and Applications

Gitta Kutyniok

March 12, 2012

Abstract

Compressed sensing is a novel research area, which was introduced in 2006, and since then has already become a key concept in various areas of applied mathematics, computer science, and electrical engineering. It surprisingly predicts that high-dimensional signals, which allow a sparse representation by a suitable basis or, more generally, a frame, can be recovered from what was previously considered highly incomplete linear measurements by using efficient algorithms. This article shall serve as an introduction to and a survey about compressed sensing.

Key Words. Dimension reduction. Frames. Greedy algorithms. Ill-posed inverse problems. ℓ_1 minimization. Random matrices. Sparse approximation. Sparse recovery.

Acknowledgements. The author would like to thank Emmanuel Candès, David Donoho, Michael Elad, and Yonina Eldar for various discussions on related topics. The author acknowledges partial support by Deutsche Forschungsgemeinschaft (DFG) Grants SPP-1324 KU 1446/13 and KU 1446/14, and is also grateful for the support of the Einstein Foundation.

1 Introduction

The area of compressed sensing was initiated in 2006 by two ground breaking papers, namely [18] by Donoho and [11] by Candès, Romberg, and Tao. Nowadays, after only 6 years, an abundance of theoretical aspects of compressed sensing are already explored in more than 1000 articles. Moreover, this methodology is to date extensively utilized by applied mathematicians, computer scientists, and engineers for a variety of applications in astronomy, biology, medicine, radar, and seismology, to name a few.

The key idea of compressed sensing is to recover a sparse signal from very few non-adaptive, linear measurements by convex optimization. Taking a different viewpoint, it concerns the exact recovery of a high-dimensional sparse vector after a dimension reduction step. From a yet again different standpoint, we can regard the problem as computing a sparse coefficient vector for a signal with respect to an overcomplete system. The theoretical foundation of compressed sensing has links to and also explores methodologies from various other fields such as, for example, applied harmonic analysis, frame theory, geometric functional analysis, numerical linear algebra, optimization theory, and random matrix theory.

It is interesting to notice that this development – the problem of sparse recovery – can in fact be traced back to earlier papers from the 90s such as [24] and later the prominent papers by Donoho and Huo [21] and Donoho and Elad [19]. When the previously mentioned two fundamental papers introducing compressed sensing were published, the term ‘compressed sensing’ was initially utilized for random sensing matrices, since those allow for a minimal number of non-adaptive, linear measurements. Nowadays, the terminology ‘compressed sensing’ is more and more often used interchangeably with ‘sparse recovery’ in general, which is a viewpoint we will also take in this survey paper.

1.1 The Compressed Sensing Problem

To state the problem mathematically precisely, let now $x = (x_i)_{i=1}^n \in \mathbb{R}^n$ be our signal of interest. As prior information, we either assume that x itself is *sparse*, i.e., it has very few non-zero coefficients in the sense that

$$\|x\|_0 := \#\{i : x_i \neq 0\}$$

is small, or that there exists an orthonormal basis or a frame¹ Φ such that $x = \Phi c$ with c being sparse. For this, we let Φ be the matrix with the elements of the orthonormal basis or the frame as column vectors. In fact, a frame provides more flexibility than an orthonormal basis due to its redundancy and hence leads to improved sparsifying properties, wherefore in this setting customarily frames are more often employed than orthonormal bases. Sometimes the notion of sparsity is weakened, which we for now – before we will make this precise in Section 2 – will refer to as *approximately sparse*. Further, let A be an $m \times n$ matrix, which is typically called *sensing matrix* or *measurement matrix*. Throughout we will always assume that $m < n$ and that A does not possess any zero columns, even if not explicitly mentioned.

Then the *Compressed Sensing Problem* can be formulated as follows: Recover x from knowledge of

$$y = Ax,$$

or recover c from knowledge of

$$y = A\Phi c.$$

In both cases, we face an underdetermined linear system of equations with sparsity as prior information about the vector to be recovered. This leads us to the following questions:

- What are suitable signal and sparsity models?
- What are suitable sensing matrices?
- How can the signal be algorithmically recovered?
- When and with which accuracy can the signal be recovered?

In this section, we will debate these questions briefly to build up intuition for the subsequent sections.

¹Recall that a *frame* for a Hilbert space \mathcal{H} is a system $(\varphi_i)_{i \in I}$ in \mathcal{H} , for which there exist *frame bounds* $0 < A \leq B < \infty$ such that $A\|x\|_2^2 \leq \sum_{i \in I} |\langle x, \varphi_i \rangle|^2 \leq B\|x\|_2^2$ for all $x \in \mathcal{H}$. A *tight frame* allows $A = B$. If $A = B = 1$ can be chosen, $(\varphi_i)_{i \in I}$ forms a *Parseval frame*. For further information, we refer to [12].

1.2 Sparsity: A Reasonable Assumption?

As a first consideration, one might question whether sparsity is indeed a reasonable assumption. Due to the complexity of real data certainly only a heuristic answer is possible.

If a natural image is taken, it is well known that wavelets typically provide sparse approximations. This is illustrated in Figure 1, which shows a wavelet decomposition [45] of an exemplary image. It can clearly be seen that most coefficients are small in absolute value, indicated by a darker color.

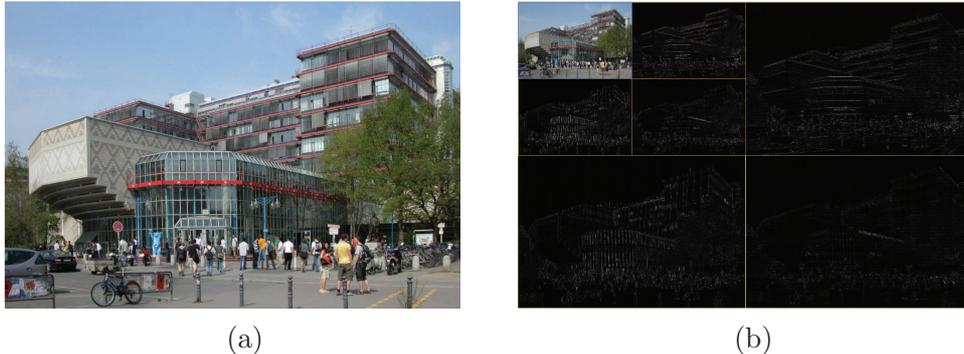


Figure 1: (a) Mathematics building of TU Berlin (Photo by TU-Pressestelle); (b) Wavelet decomposition

Depending on the signal, a variety of representation systems which can be used to provide sparse approximations is available and is constantly expanded. In fact, it was recently shown that wavelet systems do not provide optimally sparse approximations of most natural images, but the novel system of shearlets does [42, 41]. Hence, assuming some foreknowledge of the signal to be sensed or compressed, typically suitable, well analyzed representation systems are already at hand. If this is not the case, more data sensitive methods such as dictionary learning algorithms, in which a suitable representation system is computed for a given set of test signals, are available.

Depending on the application at hand, often x is already sparse itself. Think, for instance, of digital communication, when a cell phone network with n antennas and m users needs to be modelled. Or in genomics, when in a test study m genes shall be analyzed with n patients taking part in the study. In the first scenario, very few of the users have an ongoing call at a specific time; in the second scenario, very few of the genes are actually active. Thus, x being sparse itself is also a very natural assumption.

In the compressed sensing literature, most results indeed assume that x itself is sparse, and the problem $y = Ax$ is considered. Very few articles study the problem of incorporating a sparsifying orthonormal basis or frame, and we mention [55, 9]. In this paper, we will also assume throughout that x is already a sparse vector. It should be emphasized that ‘exact’ sparsity is often too restricting or unnatural, and weakened sparsity notions need to be taken into account. On the other hand, sometimes – such as with the tree structure of wavelet coefficients – some structural information of the non-zero coefficients is known, which leads to diverse structured sparsity models. Section 2 provides an overview of such models.

1.3 Recovery Algorithms: Optimization Theory and More

Let x now be a sparse vector. It is quite intuitive to recover x from knowledge of y by solving

$$(P_0) \quad \min_x \|x\|_0 \text{ subject to } y = Ax.$$

Due to the unavoidable combinatorial search, this algorithm is however NP-hard [48]. The main idea of Chen, Donoho, and Saunders in the fundamental paper [14] was to substitute the ℓ_0 ‘norm’ by the closest convex norm, which is the ℓ_1 norm. This leads to the following minimization problem, which they coined *Basis Pursuit*:

$$(P_1) \quad \min_x \|x\|_1 \text{ subject to } y = Ax.$$

Due to the shape of the ℓ_1 ball, ℓ_1 minimization indeed promotes sparsity. The general question of when ‘ $\ell_0 = \ell_1$ ’ holds is key to compressed sensing. Both necessary and sufficient conditions have been provided, which not only depend on the sparsity of the original vector x , but also on the incoherence of the sensing matrix A , which will be made precise in Section 3.

Since for very large data sets ℓ_1 minimization is often not feasible even when the solvers are adapted to the particular structure of compressed sensing problems, various other types of recovery algorithms were suggested. These can be roughly separated into convex optimization, greedy, and combinatorial algorithms (cf. Section 5), each one having its own advantages and disadvantages.

1.4 Sensing Matrices: How Much Freedom is Allowed?

As already mentioned, sensing matrices are required to satisfy certain incoherence conditions such as, for instance, a small so-called mutual coherence. If we are allowed to choose the sensing matrix freely, the best choice are random matrices such as Gaussian iid matrices, uniform random ortho-projectors, or Bernoulli matrices, see for instance [11].

It is still an open question (cf. Section 4 for more details) whether deterministic matrices can be carefully constructed to have similar properties with respect to compressed sensing problems. At the moment, different approaches towards this problem are being taken such as structured random matrices by, for instance, Rauhut et al. in [53] or [54]. Moreover, most applications do not allow a free choice of the sensing matrix and enforce a particularly structured matrix. Exemplary situations are the application of data separation, in which the sensing matrix has to consist of two or more orthonormal bases or frames [32, Chapter 11], or for high resolution radar, for which the sensing matrix has to bear a particular time-frequency structure [36].

1.5 Compressed Sensing: Quo Vadis?

At present, a comprising core theory seems established except for some few deep questions such as the construction of deterministic sensing matrices exhibiting properties similar to random matrices.

One current main direction of research which can be identified with already various existing results is the incorporation of additional sparsity properties typically coined *structured sparsity*, see Section 2 for references. Another main direction is the extension or transfer of the Compressed Sensing Problem to other settings such as *matrix completion*, see for instance [10]. Moreover, we are currently witnessing the diffusion of compressed sensing ideas to various *application areas* such as radar analysis, medical imaging, distributed signal processing, and data quantization, to name a few; see [32] for an overview. These applications pose intriguing challenges to the area due to the constraints they require, which in turn initiates novel theoretical problems. Finally, we observe that due to the need of, in particular, fast sparse recovery algorithms, there is a trend to more closely cooperate with *mathematicians from other research areas* such as optimization theorists, numerical linear algebraists, or random matrix theorists.

As three examples of recently initiated research directions, we would like to mention the following. First, while the theory of compressed sensing focusses on digital data, it is desirable to develop a similar theory for the *continuum setting*. Two promising approaches were so far suggested by Eldar et al. (cf. [47]) and Hansen et al. (cf. [1]). Second, in contrast to Basis Pursuit, which minimizes the ℓ_1 norm of the synthesis coefficients, several approaches such as recovery of missing data minimize the ℓ_1 norm of the analysis coefficients – as opposed to minimizing the ℓ_1 norm of the synthesis coefficients –, see Subsections 6.1.2 and 6.2.2. The relation between these two minimization problems is far from being clear, and the recently introduced notion of *co-sparsity* [49] is an interesting approach to shed light onto this problem. Third, the utilization of *frames as a sparsifying system* in the context of compressed sensing has become a topic of increased interest, and we refer to the initial paper [9].

The reader might also want to consult the extensive webpage dsp.rice.edu/cs containing most published papers in the area of compressed sensing subdivided into different topics. We would also like to draw the reader's attention to the recent books [29] and [32] as well as the survey article [7].

1.6 Outline

In Section 2, we start by discussing different sparsity models including structured sparsity and sparsifying dictionaries. The next section, Section 3, is concerned with presenting both necessary and sufficient conditions for exact recovery using ℓ_1 minimization as a recovery strategy. The delicateness of designing sensing matrices is the focus of Section 4. In Section 5, other algorithmic approaches to sparse recovery are introduced. Finally, applications such as data separation are discussed in Section 6.

2 Signal Models

Sparsity is the prior information assumed of the vector we intend to efficiently sense or whose dimension we intend to reduce, depending on which viewpoint we take. We will start by recalling some classical notions of sparsity. Since applications typically impose a certain structure on the significant coefficients, various structured sparsity models were introduced

which we will subsequently discuss. Finally, we will discuss about imposing sparsity through an appropriate orthonormal basis or frame.

2.1 Sparsity

The most basic notion of sparsity states that a vector has at most k non-zero coefficients. This is measured by the ℓ_0 ‘norm’, which for simplicity we will throughout refer to as a norm although it is certainly not.

Definition 2.1 *A vector $x = (x_i)_{i=1}^n \in \mathbb{R}^n$ is called k -sparse, if*

$$\|x\|_0 = \#\{i : x_i \neq 0\} \leq k.$$

The set of all k -sparse vectors is denoted by Σ_k .

From an application point of view, this situation is unrealistic, wherefore various weaker versions were suggested. In the following definition we present one possibility but do by no means claim this to be the most appropriate one. It might though be very natural, since it analyzes the decay rate of the ℓ_p error of the best k -term approximation of a vector.

Definition 2.2 *Let $1 \leq p < \infty$ and $r > 0$. A vector $x = (x_i)_{i=1}^n \in \mathbb{R}^n$ is called p -compressible with constant C and rate r , if*

$$\sigma_k(x)_p := \min_{\tilde{x} \in \Sigma_k} \|x - \tilde{x}\|_p \leq C \cdot k^{-r} \quad \text{for any } k \in \{1, \dots, n\}.$$

2.2 Structured Sparsity

Typically, the non-zero or significant coefficients do not arise in arbitrary patterns but are rather highly structured. Think of the coefficients of a wavelet decomposition which exhibit a tree structure, see also Figure 1. To take these considerations into account, structured sparsity models were introduced. A first idea might be to identify the clustered set of significant coefficients [22]. An application of this notion will be discussed in Section 6.

Definition 2.3 *Let $\Lambda \subset \{1, \dots, n\}$ and $\delta > 0$. A vector $x = (x_i)_{i=1}^n \in \mathbb{R}^n$ is then called δ -relatively sparse with respect to Λ , if*

$$\|1_{\Lambda^c} x\|_1 \leq \delta.$$

The notion of k -sparsity can also be regarded from a more general viewpoint, which simultaneously imposes additional structure. Let $x \in \mathbb{R}^n$ be a k -sparse signal. Then it belongs to the linear subspace consisting of all vectors with the same support set. Hence the set Σ_k is the union of all subspaces of vectors with support Λ satisfying $|\Lambda| \leq k$. Thus, a natural extension of this concept is the following definition, initially introduced in [44].

Definition 2.4 *Let $(\mathcal{W}_j)_{j=1}^N$ be a family of subspaces in \mathbb{R}^n . Then a vector $x \in \mathbb{R}^n$ is k -sparse in the union of subspaces $\bigcup_{j=1}^N \mathcal{W}_j$, if there exists $\Lambda \subset \{1, \dots, N\}$, $|\Lambda| \leq k$, such that*

$$x \in \bigcup_{j \in \Lambda} \mathcal{W}_j.$$

At about the same time, the notion of *fusion frame sparsity* was introduced in [6]. Fusion frames are a set of subspaces having frame-like properties, thereby allowing for stability considerations. A family of subspaces $(\mathcal{W}_j)_{j=1}^N$ in \mathbb{R}^n is a *fusion frame* with bounds A and B , if

$$A\|x\|_2^2 \leq \sum_{j=1}^N \|P_{\mathcal{W}_j}(x)\|_2^2 \leq B\|x\|_2^2 \quad \text{for all } x \in \mathbb{R}^n,$$

where $P_{\mathcal{W}_j}$ denotes the orthogonal projection onto the subspace \mathcal{W}_j , see also [13] and [12, Chapter 13]. Fusion frame theory extends classical frame theory by allowing the analysis of signals through projections onto arbitrary dimensional subspaces as opposed to one-dimensional subspaces in frame theory, hence serving also as a model for distributed processing. Fusion frame sparsity can then be defined in a similar way as for a union of subspaces.

Applications such as manifold learning assume that the signal under consideration lives on a general manifold, thereby forcing us to leave the world of linear subspaces. In such cases, the signal class is often modeled as a non-linear k -dimensional manifold \mathcal{M} in \mathbb{R}^n , i.e.,

$$x \in \mathcal{M} = \{f(\theta) : \theta \in \Theta\}$$

with Θ being a k -dimensional parameter space. Such signals are then considered *k-sparse in the manifold model*, see [58]. For a survey chapter about this topic, the interested reader is referred to [32, Chapter 7].

We wish to finally mention that applications such as matrix completion require generalizations of vector sparsity by considering, for instance, low-rank matrix models. This is however beyond the scope of this survey paper, and we refer to [32] for more details.

2.3 Sparsifying Dictionaries and Dictionary Learning

If the vector itself does not exhibit sparsity, we are required to sparsify it by choosing an appropriate representation system – in this field typically coined *dictionary*. This problem can be attacked in two ways, either non-adaptively or adaptively.

If certain characteristics of the signal are known, a dictionary can be chosen from the vast class of already very well explored representation systems such as the Fourier basis, wavelets, or shearlets, to name a few. The achieved sparsity might not be optimal, but various mathematical properties of these systems are known and fast associated transforms are available.

Improved sparsity can be achieved by choosing the dictionary adaptive to the signals at hand. For this, a test set of signals is required, based on which a dictionary is learnt. This process is customarily termed *dictionary learning*. The most well-known and widely used algorithm is the K-SVD algorithm introduced by Aahron, Elad, and Bruckstein in [2]. However, from a mathematician’s point of view, this approach bears two problems which will however presumably be both solved in the near future. First, almost no convergence results for such algorithms are known. And, second, the learnt dictionaries do not exhibit any mathematically exploitable structure, which makes not only an analysis very hard but also prevents the design of fast associated transforms.

3 Conditions for Sparse Recovery

After having introduced various sparsity notions, in this sense signal models, we next question which conditions we need to impose on the sparsity of the original vector and on the sensing matrix for exact recovery. As sparse recovery method, we will focus on ℓ_1 minimization similar to most published results and refer to Section 5 for further algorithmic approaches. In the sequel of the present section, several incoherence conditions for sensing matrices will be introduced. Section 4 then discusses examples of matrices fulfilling those. Finally, we mention that most results can be slightly modified to incorporate measurements affected by additive noise, i.e., if $y = Ax + \nu$ with $\|\nu\|_2 \leq \varepsilon$.

3.1 Uniqueness Conditions for Minimization Problems

We start by presenting conditions for uniqueness of the minimization problems (P_0) and (P_1) which we introduced in Subsection 1.3.

3.1.1 Uniqueness of (P_0)

The correct condition on the sensing matrix is phrased in terms of the so-called spark, whose definition we first recall. This notion was introduced in [19] and verbally fuses the notions of ‘sparse’ and ‘rank’.

Definition 3.1 *Let A be an $m \times n$ matrix. Then the spark of A denoted by $\text{spark}(A)$ is the minimal number of linearly dependent columns of A .*

It is useful to reformulate this notion in terms of the null space of A , which we will throughout denote by $\mathcal{N}(A)$, and state its range. The proof is obvious.

Lemma 3.1 *Let A be an $m \times n$ matrix. Then*

$$\text{spark}(A) = \min\{k : \mathcal{N}(A) \cap \Sigma_k \neq \{0\}\}$$

and $\text{spark}(A) \in [2, m + 1]$.

This notion enables us to derive an equivalent condition on unique solvability of (P_0) . Since the proof is short, we state it for clarity purposes.

Theorem 3.1 ([19]) *Let A be an $m \times n$ matrix, and let $k \in \mathbb{N}$. Then the following conditions are equivalent.*

- (i) *If a solution x of (P_0) satisfies $\|x\|_0 \leq k$, then this is the unique solution.*
- (ii) *$k < \text{spark}(A)/2$.*

Proof. (i) \Rightarrow (ii). We argue by contradiction. If (ii) does not hold, by Lemma 3.1, there exists some $h \in \mathcal{N}(A)$, $h \neq 0$ such that $\|h\|_0 \leq 2k$. Thus, there exist x and \tilde{x} satisfying $h = x - \tilde{x}$ and $\|x\|_0, \|\tilde{x}\|_0 \leq k$, but $Ax = A\tilde{x}$, a contradiction to (i).

(ii) \Rightarrow (i). Let x and \tilde{x} satisfy $y = Ax = A\tilde{x}$ and $\|x\|_0, \|\tilde{x}\|_0 \leq k$. Thus $x - \tilde{x} \in \mathcal{N}(A)$ and $\|x - \tilde{x}\|_0 \leq 2k < \text{spark}(A)$. By Lemma 3.1, it follows that $x - \tilde{x} = 0$, which implies (i).

□

3.1.2 Uniqueness of (P_1)

Due to the underdeterminedness of A and hence the ill-posedness of the recovery problem, in the analysis of uniqueness of the minimization problem (P_1) , the null space of A also plays a particular role. The related so-called null space property, first introduced in [15], is defined as follows.

Definition 3.2 *Let A be an $m \times n$ matrix. Then A has the null space property (NSP) of order k , if, for all $h \in \mathcal{N}(A) \setminus \{0\}$ and for all index sets $|\Lambda| \leq k$,*

$$\|1_\Lambda h\|_1 < \frac{1}{2}\|h\|_1.$$

An equivalent condition for the existence of a unique sparse solution of (P_1) can now be stated in terms of the null space property. For the proof, we refer to [15].

Theorem 3.2 ([15]) *Let A be an $m \times n$ matrix, and let $k \in \mathbb{N}$. Then the following conditions are equivalent.*

- (i) *If a solution x of (P_1) satisfies $\|x\|_0 \leq k$, then this is the unique solution.*
- (ii) *A satisfies the null space property of order k .*

It should be emphasized that [15] studies the Compressed Sensing Problem in a much more general way by analyzing quite general encoding-decoding strategies.

3.2 Sufficient Conditions

The core of compressed sensing is to determine when ' $\ell_0 = \ell_1$ ', i.e., when the solutions of (P_0) and (P_1) coincide. The most well-known sufficient conditions for this to hold true are phrased in terms of mutual coherence and of the restricted isometry property.

3.2.1 Mutual Coherence

The mutual coherence of a matrix, initially introduced in [21], measures the smallest angle between each pair of its columns.

Definition 3.3 *Let $A = (a_i)_{i=1}^n$ be an $m \times n$ matrix. Then its mutual coherence $\mu(A)$ is defined as*

$$\mu(A) = \max_{i \neq j} \frac{|\langle a_i, a_j \rangle|}{\|a_i\|_2 \|a_j\|_2}.$$

The maximal mutual coherence of a matrix is certainly 1 in the case that two columns coincide. The lower bound presented in the next result, known also as the *Welch bound*, is more interesting. It can be shown that it is attained by so-called *Grassmannian frames* [56], see also Section 4.

Lemma 3.2 *Let A be an $m \times n$ matrix. Then we have*

$$\mu(A) \in \left[\sqrt{\frac{n-m}{m(n-1)}}, 1 \right].$$

Let us mention that different variants of mutual coherence exist, in particular, the *Babel function* [19], the *cumulative coherence function* [57], the *structured p -Babel function* [4], the *fusion coherence* [6], and *cluster coherence* [22]. The notion of cluster coherence will in fact be later discussed in Section 6 for a particular application.

Imposing a bound on the sparsity of the original vector by the mutual coherence of the sensing matrix, the following result can be shown; its proof can be found in [19].

Theorem 3.3 ([30, 19]) *Let A be an $m \times n$ matrix, and let $x \in \mathbb{R}^n \setminus \{0\}$ be a solution of (P_0) satisfying*

$$\|x\|_0 < \frac{1}{2}(1 + \mu(A)^{-1}).$$

Then x is the unique solution of (P_0) and (P_1) .

3.2.2 Restricted Isometry Property

We next discuss the restricted isometry property, initially introduced in [11]. It measures the degree to which each subset of k column vectors of A is close to being an isometry. Notice that this notion automatically ensures stability, as will become evident in the next theorem.

Definition 3.4 *Let A be an $m \times n$ matrix. Then A has the Restricted Isometry Property (RIP) of order k , if there exists a $\delta_k \in (0, 1)$ such that*

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2 \quad \text{for all } x \in \Sigma_k.$$

Several derivations of this notion were introduced during the last years, of which examples are the *fusion RIP* [6] and the *D-RIP* [9].

Although also for mutual coherence, error estimates for recovery from noisy data are known, in the setting of the RIP those are very natural. In fact, the error can be phrased in terms of the best k -term approximation (cf. Definition 2.2) as follows.

Theorem 3.4 ([8, 15]) *Let A be an $m \times n$ matrix which satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$. Let $x \in \mathbb{R}^n$, and let \hat{x} be a solution of the associated (P_1) problem. Then*

$$\|x - \hat{x}\|_2 \leq C \cdot \left(\frac{\sigma_k(x)_1}{\sqrt{k}} \right).$$

3.3 Necessary Conditions

Meaningful necessary conditions are significantly harder to achieve. An interesting string of research was initiated by Donoho and Tanner with the two papers [25, 26]. The main idea is to derive equivalent conditions utilizing the theory of convex polytopes. For this, let C^n be defined by

$$C^n = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}.$$

A condition equivalent to ' $\ell_0 = \ell_1$ ' can then be formulated in terms of properties of a particular related polytope.

Theorem 3.5 ([25, 26]) *Retain the previous assumptions and notations, let A be an $m \times n$ matrix, and let the polytope P be defined by $P = AC^n \subseteq \mathbb{R}^m$. Then the following conditions are equivalent.*

- (i) *The number of k -faces of P equals the number of k -faces of C^n .*
- (ii) *$(P_0) = (P_1)$.*

The geometric intuition behind this result is the fact that the number of k -faces of P equals the number of indexing sets $\Lambda \subseteq \{1, \dots, n\}$ with $|\Lambda| = k$ such that vectors x satisfying $\text{supp } x = \Lambda$ can be recovered via (P_1) .

Extending these techniques, Donoho and Tanner were also able to provide highly accurate analytical descriptions of the occurring phase transition when considering the area of exact recovery dependent on the ratio of the number of equations to the number of unknowns n/m versus the ratio of the number of nonzeros to the number of equations k/n . The interested reader might want to consult [27] for further details.

4 Sensing Matrices

After presenting some general relations between the introduced notions of spark, NSP, mutual coherence, and RIP, we will discuss some explicit constructions for, in particular, mutual coherence and RIP.

4.1 Relations between Spark, NSP, Mutual Coherence, and RIP

Before discussing different approaches to construct a sensing matrix, we first present several known relations between the introduced notions spark, NSP, mutual coherence, and RIP. This allows to easily compute or at least estimate other measures, if a sensing matrix is designed for a particular one. For the proofs of the different statements, we refer to [32, Chapter 1].

Lemma 4.1 *Let A be an $m \times n$ matrix with normalized columns.*

- (i) *We have*

$$\text{spark}(A) \geq 1 + \frac{1}{\mu(A)}.$$

- (ii) *A satisfies the RIP of order k with $\delta = k\mu(A)$ for all $k < \mu(A)^{-1}$.*
- (iii) *Suppose A satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$. If*

$$\frac{\sqrt{2}\delta_{2k}}{1 - (1 + \sqrt{2})\delta_{2k}} < \sqrt{\frac{k}{n}},$$

then A satisfies the NSP of order $2k$.

4.2 Spark and Mutual Coherence

Let us now provide some exemplary classes of sensing matrices with advantageous spark and mutual coherence properties.

The first observation one can make (see also [15]) is that an $m \times n$ Vandermonde matrix A satisfies

$$\text{spark}(A) = m + 1.$$

One serious drawback though is the fact that these matrices become badly conditioned as $n \rightarrow \infty$.

Turning to the weaker notion of mutual coherence, of particular interest – compare Subsection 6.1 – are sensing matrices composed of two orthonormal bases or frames for \mathbb{R}^m . If the two orthonormal bases Φ_1 and Φ_2 , say, are chosen to be mutual unbiased such as the Fourier and the Dirac basis, then

$$\mu([\Phi_1|\Phi_2]) = \frac{1}{\sqrt{m}},$$

which is the optimal bound on mutual coherence for such types of $m \times 2m$ sensing matrix. Other constructions are known for $m \times m^2$ matrices A generated from the Alltop sequence [36] or by using Grassmannian frames [56], in which cases the optimal lower bound is attained:

$$\mu(A) = \frac{1}{\sqrt{m}}.$$

The number of measurements required for recovery of a k -sparse signal can then determined to be $m = O(k^2 \log n)$.

4.3 RIP

We begin by discussing some deterministic constructions of matrices satisfying the RIP. The first noteworthy construction was presented by DeVore and requires $m \gtrsim k^2$, see [17]. A very recent, highly sophisticated approach [5] by Bourgain et al. still requires $m \gtrsim k^{2-\alpha}$ with some small constant α . Hence up to now deterministic constructions require a large m , which is typically not feasible for applications.

The construction of random sensing matrices satisfying RIP is a possibility to circumvent this problem. Such constructions are closely linked to the famous Johnson-Lindenstrauss Lemma, which is extensively utilized in numerical linear algebra, machine learning, or other areas requiring dimension reduction.

Theorem 4.1 (Johnson-Lindenstrauss Lemma [38]) *Let $\varepsilon \in (0, 1)$, let $x_1, \dots, x_p \in \mathbb{R}^n$, and let $m = O(\varepsilon^{-2} \log p)$ be a positive integer. Then there exists a Lipschitz map $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that*

$$(1 - \varepsilon)\|x_i - x_j\|_2^2 \leq \|f(x_i) - f(x_j)\|_2^2 \leq (1 + \varepsilon)\|x_i - x_j\|_2^2 \quad \text{for all } i, j \in \{1, \dots, p\}.$$

The key requirement for a matrix to satisfy the Johnson-Lindenstrauss Lemma with high probability is the following concentration inequality for an arbitrarily fixed $x \in \mathbb{R}^n$:

$$\mathbb{P} \left((1 - \varepsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2 \right) \leq 1 - 2e^{-c_0\varepsilon^2 m}. \quad (1)$$

The relation of RIP to the Johnson-Lindenstrauss Lemma is established in the following result. We also mention that recently even a converse of the following theorem was proved in [40].

Theorem 4.2 ([3]) *Let $\delta \in (0, 1)$. If the probability distribution generating the $m \times n$ matrices A satisfies the concentration inequality (1) with $\varepsilon = \delta$, then there exist constants c_1, c_2 such that, with probability $\leq 1 - 2e^{-c_2\delta^2 m}$, A satisfies the RIP of order k with δ for all $k \leq c_1\delta^2 m / \log(n/k)$.*

This observation was then used in [3] to prove that Gaussian and Bernoulli random matrices satisfy the RIP of order k with δ provided that $m \gtrsim \delta^{-2}k \log(n/k)$. Up to a constant, lower bounds for Gelfand widths of ℓ_1 -balls [34] show that this dependence on k and n is indeed optimal.

5 Recovery Algorithms

In this section, we will provide a brief overview of the different types of algorithms typically used for sparse recovery. Convex optimization algorithms require very few measurements but are computationally more complex. On the other extreme are combinatorial algorithms, which are very fast – often sublinear – but require many measurements that are sometimes difficult to obtain. Greedy algorithms are in some sense a good compromise between those extremes concerning computational complexity and the required number of measurements.

5.1 Convex Optimization

In Subsection 1.3, we already stated the convex optimization problem

$$\min_x \|x\|_1 \text{ subject to } y = Ax$$

most commonly used. If the measurements are affected by noise, a conic constraint is required; i.e., the minimization problem needs to be changed to

$$\min_x \|x\|_1 \text{ subject to } \|Ax - y\|_2^2 \leq \varepsilon,$$

for a carefully chosen $\varepsilon > 0$. For a particular parameter $\lambda > 0$, this problem is equivalent to the unconstrained version given by

$$\min_x \frac{1}{2} \|Ax - y\|_2^2 + \lambda \|x\|_1.$$

Developed convex optimization algorithms specifically adapted to the compressed sensing setting include interior-point methods [11], projected gradient methods [33], and iterative thresholding [16]. The reader might also be interested to check the webpages www-stat.stanford.edu/~candes/l1magic and sparselab.stanford.edu for available code.

5.2 Greedy Algorithms

Greedy algorithms iteratively approximate the coefficients and the support of the original signal. They have the advantage of being very fast and easy to implement. Often the theoretical performance guarantees are very similar to, for instance, ℓ_1 minimization results.

The most well-known greedy approach is *Orthogonal Matching Pursuit*, which is described in Figure 2. OMP was introduced in [52] as an improved successor of *Matching Pursuit* [46].

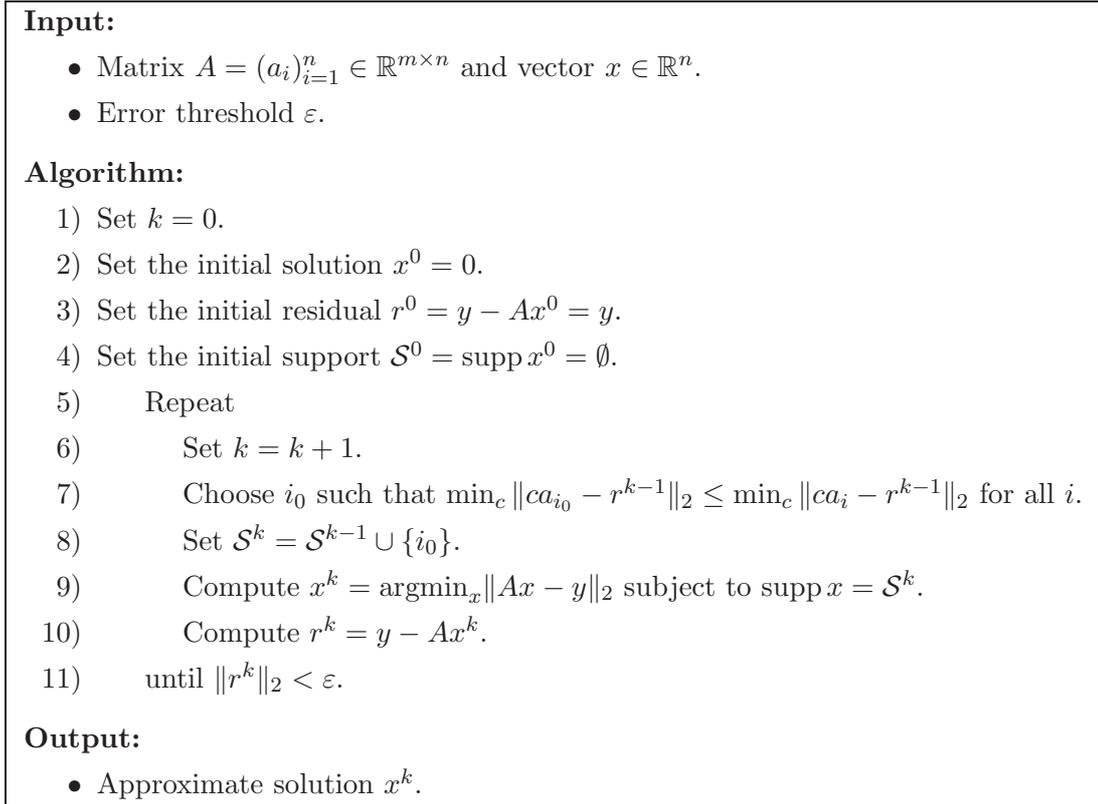


Figure 2: Orthogonal Matching Pursuit (OMP): Approximation of the solution of (P_0)

Interestingly, a theorem similar to Theorem 3.3 can be proven for OMP.

Theorem 5.1 ([57, 20]) *Let A be an $m \times n$ matrix, and let $x \in \mathbb{R}^n \setminus \{0\}$ be a solution of (P_0) satisfying*

$$\|x\|_0 < \frac{1}{2}(1 + \mu(A)^{-1}).$$

Then OMP with error threshold $\varepsilon = 0$ recovers x .

Other prominent examples of greedy algorithms are stagewise OMP (StOMP) [28], regularized OMP (ROMP) [51], and compressive sampling MP (CoSamp) [50]. For a survey of these methods, we wish to refer to [32, Chapter 8]. A very recent development are message passing algorithms for compressed sensing pioneered in [23]; a survey can be found in [32, Chapter 9].

5.3 Combinatorial Algorithms

These methods apply group testing to highly structured samples of the original signal. From the various types of algorithms, we mention the HHS pursuit [35] and a sub-linear Fourier transform [37].

6 Applications

We now turn to some applications of compressed sensing. Two of those we will discuss in more detail, namely data separation and recovery of missing data.

6.1 Data Separation

The data separation problem can be stated in the following way. Let $x = x_1 + x_2 \in \mathbb{R}^n$. Assuming we are just given x , how can we extract x_1 and x_2 from it? At first glance, this seems not possible, since there are two unknowns for one datum.

6.1.1 An Orthonormal Basis Approach

The first approach to apply compressed sensing techniques consists in choosing appropriate orthonormal bases Φ_1 and Φ_2 for \mathbb{R}^n such that the coefficient vectors $\Phi_i^T x_i$ ($i = 1, 2$) are sparse. This leads to the following underdetermined linear system of equations:

$$x = [\Phi_1 \mid \Phi_2] \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

Compressed sensing now suggests to solve

$$\min_{c_1, c_2} \left\| \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \right\|_1 \quad \text{subject to } x = [\Phi_1 \mid \Phi_2] \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \quad (2)$$

If the sparse vector $[\Phi_1^T x_1, \Phi_2^T x_2]^T$ can be recovered, the data separation problem can be solved by computing

$$x_1 = \Phi_1(\Phi_1^T x_1) \quad \text{and} \quad x_2 = \Phi_2(\Phi_2^T x_2).$$

Obviously, separation can only be achieved provided that the components x_1 and x_2 are in some sense morphologically distinct. Notice that this property is indeed encoded in the problem through the required incoherence of the matrix $[\Phi_1 \mid \Phi_2]$.

In fact, this type of problem can be regarded as the birth of compressed sensing, since the fundamental paper [21] by Donoho and Huo analyzed a particular data separation problem, namely the separation of sinusoids and spikes. In this setting, x_1 consists of n samples of a continuum domain signal which is a superposition of sinusoids:

$$x_1 = \left(\frac{1}{\sqrt{n}} \sum_{\omega=0}^{n-1} c_{1,\omega} e^{2\pi i \omega t / n} \right)_{0 \leq t \leq n-1}$$

Letting Φ_1 be the Fourier basis, the coefficient vector

$$\Phi_1^T x_1 = c_1, \quad \text{where } \Phi_1 = [\varphi_{1,0} \mid \dots \mid \varphi_{1,n-1}] \quad \text{with } \varphi_{1,\omega} = \left(\frac{1}{\sqrt{n}} e^{2\pi i \omega t / n} \right)_{0 \leq t \leq n-1},$$

is sparse. The vector x_2 consists of n samples of a continuum domain signal which is a superposition of spikes, i.e., has few non-zero coefficients. Thus, letting Φ_2 denote the Dirac basis, the coefficient vector

$$\Phi_2^T x_2 = x_2 = c_2$$

is also sparse. Since the mutual coherence of the matrix $[\Phi_1 \mid \Phi_2]$ can be computed to be $\frac{1}{\sqrt{n}}$, Theorem 3.3 implies the following result.

Theorem 6.1 ([21, 30]) *Retain the previous assumptions and notations, and assume that $\|\Phi_1^T x_1\|_0 + \|\Phi_2^T x_2\|_0 < \frac{1}{2}(1 + \sqrt{n})$. Then*

$$\begin{bmatrix} \Phi_1^T x_1 \\ \Phi_2^T x_2 \end{bmatrix} = \operatorname{argmin}_{c_1, c_2} \left\| \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \right\|_1 \quad \text{subject to } x = [\Phi_1 \mid \Phi_2] \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

6.1.2 A Frame Approach

Now assume that we cannot find sparsifying orthonormal bases but Parseval frames Φ_1 and Φ_2 – notice that this situation is much more likely due to the advantageous redundancy of a frame. In this case, the minimization problem we stated in (2) faces the following problem: We are merely interested in the separation $x = x_1 + x_2$. However, for each such separation, due to the redundancy of the frames the minimization problem searches through infinitely many coefficients $[c_1, c_2]^T$ satisfying $x_i = \Phi_i c_i$, $i = 1, 2$. Thus it computes not only much more than necessary – in fact, it even computes the sparsest coefficient sequence of x with respect to the dictionary $[\Phi_1 \mid \Phi_2]$ – but this also causes numerical instabilities if the redundancy of the frames is too high.

To avoid this problem, we place the ℓ_1 norm on the *analysis*, rather than the *synthesis* side as already announced before in Subsection 1.5. Utilizing the fact that Φ_1 and Φ_2 are Parseval frames, i.e., that $\Phi_i \Phi_i^T = I$ ($i = 1, 2$), we can write

$$x = x_1 + x_2 = \Phi_1(\Phi_1^T x_1) + \Phi_2(\Phi_2^T x_2).$$

This particular choice of coefficients – which are in frame theory language termed *analysis coefficients* – leads to the minimization problem

$$\min_{\tilde{x}_1, \tilde{x}_2} \|\Phi_1^T \tilde{x}_1\|_1 + \|\Phi_2^T \tilde{x}_2\|_1 \quad \text{subject to } x = \tilde{x}_1 + \tilde{x}_2. \quad (3)$$

Interestingly, the associated recovery results employ structured sparsity, wherefore we will also briefly present those. First, the notion of relative sparsity (cf. Definition 2.3) is adapted to this situation.

Definition 6.1 *Let Φ_1 and Φ_2 be Parseval frames for \mathbb{R}^n with indexing sets $\{1, \dots, N_1\}$ and $\{1, \dots, N_2\}$, respectively, let $\Lambda_i \subset \{1, \dots, N_i\}$, $i = 1, 2$, and let $\delta > 0$. Then the vectors x_1 and x_2 are called δ -relatively sparse in Φ_1 and Φ_2 with respect to Λ_1 and Λ_2 , if*

$$\|1_{\Lambda_1^c} \Phi_1^T x_1\|_1 + \|1_{\Lambda_2^c} \Phi_2^T x_2\|_1 \leq \delta.$$

Second, the notion of mutual coherence is adapted to structured sparsity as already discussed in Subsection 3.2.1. This leads to the following definition of cluster coherence.

Definition 6.2 *Let $\Phi_1 = (\varphi_{1i})_{i=1}^{N_1}$ and $\Phi_2 = (\varphi_{2j})_{j=1}^{N_2}$ be Parseval frames for \mathbb{R}^n , respectively, and let $\Lambda_1 \subset \{1, \dots, N_1\}$. Then the cluster coherence $\mu_c(\Lambda_1, \Phi_1; \Phi_2)$ of Φ_1 and Φ_2 with respect to Λ_1 is defined by*

$$\mu_c(\Lambda_1, \Phi_1; \Phi_2) = \max_{j=1, \dots, N_2} \sum_{i \in \Lambda_1} |\langle \varphi_{1i}, \varphi_{2j} \rangle|.$$

The performance of the minimization problem (3) can then be analyzed as follows. It should be emphasized that the clusters of significant coefficients Λ_1 and Λ_2 are a mere analysis tool; the algorithm does not take those into account. Further, notice that the choice of those sets is highly delicate in its impact on the separation estimate. For the proof of the result, we refer to [22].

Theorem 6.2 ([22]) *Let $x = x_1 + x_2 \in \mathbb{R}^n$, let Φ_1 and Φ_2 be Parseval frames for \mathbb{R}^n with indexing sets $\{1, \dots, N_1\}$ and $\{1, \dots, N_2\}$, respectively, and let $\Lambda_i \subset \{1, \dots, N_i\}$, $i = 1, 2$. Further, suppose that x_1 and x_2 are δ -relatively sparse in Φ_1 and Φ_2 with respect to Λ_1 and Λ_2 , and let $[x_1^*, x_2^*]^T$ be a solution of the minimization problem (3). Then*

$$\|x_1^* - x_1\|_2 + \|x_2^* - x_2\|_2 \leq \frac{2\delta}{1 - 2\mu_c},$$

where $\mu = \max\{\mu_c(\Lambda_1, \Phi_1; \Phi_2), \mu_c(\Lambda_2, \Phi_2; \Phi_1)\}$.

Let us finally mention that data separation via compressed sensing has been applied, for instance, in imaging sciences for the separation of point- and curvelike objects, a problem appearing in several areas such as in astronomical imaging when separating stars from filaments and in neurobiological imaging when separating spines from dendrites. Figure 3 illustrates a numerical result from [43] using wavelets (see [45]) and shearlets (see [42, 41]) as sparsifying frames. A theoretical foundation is developed in [22].

For more details on data separation using compressed sensing techniques, we refer to [32, Chapter 11].

6.2 Recovery of Missing Data

The problem of recovery of missing data can be formulated as follows. Let $x = x_K + x_M \in \mathcal{W} \oplus \mathcal{W}^\perp$, where \mathcal{W} is a subspace of \mathbb{R}^n . We assume only x_K is known to us, and we aim to recover x . Again, this seems unfeasible unless we have additional information.

6.2.1 An Orthonormal Basis Approach

We now assume that – although x is not known to us – we at least know that it is sparsified by an orthonormal basis Φ , say. Letting $P_{\mathcal{W}}$ and $P_{\mathcal{W}^\perp}$ denote the orthogonal projections onto \mathcal{W} and \mathcal{W}^\perp , respectively, we are led to solve the underdetermined problem

$$P_{\mathcal{W}}\Phi c = P_{\mathcal{W}}x$$

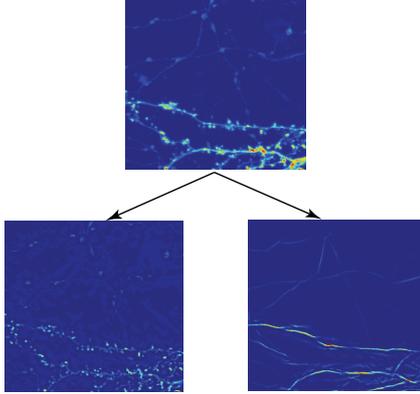


Figure 3: Separation of a neurobiological image using wavelets and shearlets

for the sparse solution c . As in the case of data separation, from a compressed sensing viewpoint it is suggestive to solve

$$\min_c \|c\|_1 \text{ subject to } P_{\mathcal{W}}\Phi c = P_{\mathcal{W}}x. \quad (4)$$

The original vector x can then be recovered via $x = \Phi c$. This minimization problem was first considered in [31] for inpainting – a terminology used for recovery of missing data in imaging science.

Application of Theorem 3.3 provides a sufficient condition for missing data recovery to succeed.

Theorem 6.3 ([19]) *Retain the previous assumptions and notations, and assume that $\|\Phi^T x\|_0 < \frac{1}{2}(1 + \mu(P_{\mathcal{W}}\Phi)^{-1})$. Then*

$$\Phi^T x = \operatorname{argmin}_c \|c\|_1 \text{ subject to } P_{\mathcal{W}}\Phi c = P_{\mathcal{W}}x.$$

6.2.2 A Frame Approach

As before, we now assume that the sparsifying system Φ is a redundant Parseval frame. The adapted version to (4), which places the ℓ_1 norm on the analysis side, reads

$$\min_{\tilde{x}} \|\Phi^T \tilde{x}\|_1 \text{ subject to } P_{\mathcal{W}}\tilde{x} = P_{\mathcal{W}}x. \quad (5)$$

Employing relative sparsity and cluster coherence, an error analysis can be derived in a similar way as before. For the proof, the reader might want to consult [39].

Theorem 6.4 ([39]) *Let $x \in \mathbb{R}^n$, let Φ be a Parseval frame for \mathbb{R}^n with indexing set $\{1, \dots, N\}$, and let $\Lambda \subset \{1, \dots, N\}$. Further, suppose that x is δ -relatively sparse in Φ with respect to Λ , and let x^* be a solution of the minimization problem (5). Then*

$$\|x^* - x\|_2 \leq \frac{2\delta}{1 - 2\mu_c},$$

where $\mu = \mu_c(\Lambda, P_{\mathcal{W}^\perp}\Phi; \Phi)$.

6.3 Further Applications

Other applications of compressed sensing include coding and information theory, machine learning, hyperspectral imaging, geophysical data analysis, computational biology, remote sensing, radar analysis, robotics and control, A/D conversion, and many more. Since an elaborate discussion of all those topics would go beyond the scope of this survey paper, we refer the interested reader to `dsp.rice.edu/cs`.

References

- [1] B. Adcock and A.C. Hansen. Generalized sampling and infinite dimensional compressed sensing. Preprint, 2012.
- [2] M. Aharon, M. Elad, and A.M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Proc.*, 54:4311–4322, 2006.
- [3] R.G. Baraniuk, M. Davenport, R.A. DeVore, and M. Wakin. A simple proof of the Restricted Isometry Property for random matrices. *Constr. Approx.*, 28:253–263, 2008.
- [4] L. Borup, R. Gribonval, and M. Nielsen. Beyond coherence: Recovering structured time-frequency representations. *Appl. Comput. Harmon. Anal.*, 14:120–128, 2008.
- [5] J. Bourgain, S. Dilworth, K. Ford, S. Konyagin, and D. Kutzarova. Explicit constructions of rip matrices and related problems. *Duke Math. J.*, to appear.
- [6] B. Boufounos, G. Kutyniok, and H. Rauhut. Sparse recovery from combined fusion frame measurements. *IEEE Trans. Inform. Theory*, 57:3864–387, 2011.
- [7] A.M. Bruckstein, D.L. Donoho, and A. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.*, 51:34–81, 2009.
- [8] E.J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci. I*, 346:589–592, 2008.
- [9] E.J. Candès, Y.C. Eldar, D. Needell, and P. Randall. Compressed Sensing with Coherent and Redundant Dictionaries. *Appl. Comput. Harmon. Anal.*, 31:59–73, 2011.
- [10] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9:717–772, 2008.
- [11] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information. *IEEE Trans. Inform. Theory*, 52:489–509, 2006.
- [12] P. G. Casazza and G. Kutyniok. *Finite Frames: Theory and Applications*, Birkhäuser, Boston, to appear.
- [13] P. G. Casazza, G. Kutyniok, and S. Li. Fusion Frames and Distributed Processing. *Appl. Comput. Harmon. Anal.* 25:114–132, 2008.

- [14] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, 1998.
- [15] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Am. Math. Soc.*, 22:211–231, 2009.
- [16] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57:1413-1457, 2004.
- [17] R. DeVore. Deterministic constructions of compressed sensing matrices. *J. Complex.*, 23:918–925, 2007.
- [18] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.
- [19] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization, *Proc. Natl. Acad. Sci. USA*, 100:2197–2202, 2003.
- [20] D. L. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52:6–18, 2006.
- [21] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, 47:2845–2862, 2001.
- [22] D. L. Donoho and G. Kutyniok. Microlocal analysis of the geometric separation problem. *Comm. Pure Appl. Math.*, to appear.
- [23] D. L. Donoho, A. Maleki, and A. Montanari. Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA*, 106:18914–18919, 2009.
- [24] D. L. Donoho and P. B. Starck. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 49:906–931, 1989.
- [25] D. L. Donoho and J. Tanner. Neighborliness of Randomly-Projected Simplices in High Dimensions. *Proc. Natl. Acad. Sci. USA*, 102:9452–9457, 2005.
- [26] D. L. Donoho and J. Tanner. Sparse Nonnegative Solutions of Underdetermined Linear Equations by Linear Programming. *Proc. Natl. Acad. Sci. USA*, 102:9446–9451, 2005.
- [27] D. L. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. Roy. Soc. S.-A*, 367:4273–4293, 2009.
- [28] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck. Sparse Solution of Underdetermined Linear Equations by Stagewise Orthogonal Matching Pursuit. Preprint, 2007.
- [29] M. Elad. *Sparse and Redundant Representations*. Springer, New York, 2010.
- [30] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Theory*, 48:2558–2567, 2002.

- [31] M. Elad, J.-L. Starck, P. Querre, and D.L. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl. Comput. Harmon. Anal.*, 19:340–358, 2005.
- [32] Y.C. Eldar and G. Kutyniok. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.
- [33] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signa.*, 1:586–597, 2007.
- [34] S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich. The Gelfand widths of ℓ_p -balls for $0 < p \leq 1$. *J. Complexity*, 26:629–640, 2010.
- [35] A.C. Gilbert, M.J. Strauss, and R. Vershynin. One sketch for all: Fast algorithms for Compressed Sensing. In *Proc. 39th ACM Symp. Theory of Computing (STOC)*, San Diego, CA, June 2007.
- [36] M. Herman and T. Strohmer. High Resolution Radar via Compressed Sensing. *IEEE Trans. Signal Proc.*, 57:2275–2284, 2009.
- [37] M.A. Iwen. Combinatorial Sublinear-Time Fourier Algorithms. *Found. of Comput. Math.*, 10:303–338, 2010.
- [38] W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math*, 26:189-206, 1984.
- [39] E. King, G. Kutyniok, and X. Zhuang. Clustered sparsity and microlocal analysis for data recovery. Preprint, 2012.
- [40] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM J. Math. Anal.*, 43:1269–1281, 2011.
- [41] G. Kutyniok and D. Labate. *Shearlets: Multiscale Analysis for Multivariate Data*. Birkhäuser, Boston, 2012.
- [42] G. Kutyniok and W.-Q Lim. Compactly supported shearlets are optimally sparse. *J. Approx. Theory*, 163:1564–1589, 2011.
- [43] G. Kutyniok and W.-Q Lim. Image separation using shearlets. In *Curves and Surfaces (Avignon, France, 2010), Lecture Notes in Computer Science 6920*, Springer, 2012.
- [44] Y. Lu and M. Do. Sampling signals from a union of subspaces. *IEEE Signal Proc. Mag.*, 25:41–47, 2008.
- [45] S.G. Mallat. *A wavelet tour of signal processing*. Academic Press, Inc., San Diego, CA, 1998.
- [46] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Proc.*, 41:3397–3415, 1993.

- [47] M. Mishali, Y. C. Eldar, and A. Elron. Xampling: Signal Acquisition and Processing in Union of Subspaces. *IEEE Trans. Signal Proc.*, 59:4719-4734, 2011.
- [48] S. Muthukrishnan. *Data Streams: Algorithms and Applications*. Now Publishers, Boston, MA, 2005.
- [49] S. Nam, M. E. Davies, M. Elad, and R. Gribonval. The Cospase Analysis Model and Algorithms. Preprint, 2012.
- [50] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmonic Anal.*, 26:301-321, 2008.
- [51] D. Needell and R. Vershynin. Uniform Uncertainty Principle and signal recovery via Regularized Orthogonal Matching Pursuit. *Found. of Comput. Math.*, 9:317-334, 2009.
- [52] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proc. of the 27th Asilomar Conference on Signals, Systems and Computers*, 1:40-44, 1993.
- [53] G. Pfander, H. Rauhut, and J. Tropp. The restricted isometry property for time-frequency structured random matrices. Preprint, 2011.
- [54] H. Rauhut, J. Romberg, and J. Tropp. Restricted isometries for partial random circulant matrices. *Appl. Comput. Harmonic Anal.*, 32:242-254, 2012.
- [55] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Trans. Inform. Theory*, 54:2210-2219, 2008.
- [56] T. Strohmer and R. W. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmonic Anal.*, 14:257-275, 2004.
- [57] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50:2231-2242, 2004.
- [58] X. Weiyu and B. Hassibi. Compressive Sensing over the Grassmann Manifold: a Unified Geometric Framework. *IEEE Trans. Inform. Theory*, to appear.