1 Introduction to compressed sensing

Mark A. Davenport, Marco F. Duarte, Yonina C. Eldar, and Gitta Kutyniok

Compressed sensing (CS) is an exciting, rapidly growing, field that has attracted considerable attention in signal processing, statistics, and computer science, as well as the broader scientific community. Since its initial development only a few years ago, thousands of papers have appeared in this area, and hundreds of conferences, workshops, and special sessions have been dedicated to this growing research field. In this chapter, we provide an up-to-date review of the basics of the theory underlying CS. This chapter should serve as a review to practitioners wanting to join this emerging field, and as a reference for researchers. We focus primarily on the theory and algorithms for sparse recovery in finite dimensions. In subsequent chapters of the book, we will see how the fundamentals presented in this chapter are expanded and extended in many exciting directions, including new models for describing structure in both analog and discrete-time signals, new sensing design techniques, more advanced recovery results and powerful new recovery algorithms, and emerging applications of the basic theory and its extensions.

1.1 Introduction

We are in the midst of a digital revolution that is driving the development and deployment of new kinds of sensing systems with ever-increasing fidelity and resolution. The theoretical foundation of this revolution is the pioneering work of Kotelnikov, Nyquist, Shannon, and Whittaker on sampling continuous-time bandlimited signals [162, 195, 209, 247]. Their results demonstrate that signals, images, videos, and other data can be exactly recovered from a set of uniformly spaced samples taken at the so-called *Nyquist rate* of twice the highest frequency present in the signal of interest. Capitalizing on this discovery, much of signal processing has moved from the analog to the digital domain and ridden the wave of Moore's law. Digitization has enabled the creation of sensing and processing systems that are more robust, flexible, cheaper and, consequently, more widely used than their analog counterparts.

As a result of this success, the amount of data generated by sensing systems has grown from a trickle to a torrent. Unfortunately, in many important and emerging applications,

Compressed Sensing: Theory and Applications, ed. Yonina C. Eldar and Gitta Kutyniok. Published by Cambridge University Press.

the resulting Nyquist rate is so high that we end up with far too many samples. Alternatively, it may simply be too costly, or even physically impossible, to build devices capable of acquiring samples at the necessary rate [146,241]. Thus, despite extraordinary advances in computational power, the acquisition and processing of signals in application areas such as imaging, video, medical imaging, remote surveillance, spectroscopy, and genomic data analysis continues to pose a tremendous challenge.

To address the logistical and computational challenges involved in dealing with such high-dimensional data, we often depend on compression, which aims at finding the most concise representation of a signal that is able to achieve a target level of acceptable distortion. One of the most popular techniques for signal compression is known as *transform coding*, and typically relies on finding a basis or frame that provides *sparse* or *compressible* representations for signals in a class of interest [31, 77, 106]. By a sparse representation, we mean that for a signal of length n, we can represent it with $k \ll n$ nonzero coefficients; by a compressible representation, we mean that the signal is well-approximated by a signal with only k nonzero coefficients. Both sparse and compressible signals can be represented with high fidelity by preserving only the values and locations of the largest coefficients of the signal. This process is called *sparse approximation*, and forms the foundation of transform coding schemes that exploit signal sparsity and compressibility, including the JPEG, JPEG2000, MPEG, and MP3 standards.

Leveraging the concept of transform coding, *compressed sensing* has emerged as a new framework for signal acquisition and sensor design that enables a potentially large reduction in the sampling and computation costs for sensing signals that have a sparse or compressible representation. While the Nyquist–Shannon sampling theorem states that a certain minimum number of samples is required in order to perfectly capture an arbitrary bandlimited signal, when the signal is sparse in a known basis we can vastly reduce the number of measurements that need to be stored. Consequently, when sensing sparse signals we might be able to do better than suggested by classical results. This is the fundamental idea behind CS: rather than first sampling at a high rate and then compressing the sampled data, we would like to find ways to *directly* sense the data in a compressed form – i.e., at a lower sampling rate. The field of CS grew out of the work of Candès, Romberg, and Tao and of Donoho, who showed that a finite-dimensional signal having a sparse or compressible representation can be recovered from a small set of linear, non-adaptive measurements [3, 33, 40–42, 44, 82]. The design of these measurement schemes and their extensions to practical data models and acquisition systems are central challenges in the field of CS.

While this idea has only recently gained significant attention in the signal processing community, there have been hints in this direction dating back as far as the eighteenth century. In 1795, Prony proposed an algorithm for the estimation of the parameters associated with a small number of complex exponentials sampled in the presence of noise [201]. The next theoretical leap came in the early 1900s, when Carathéodory showed that a positive linear combination of *any* k sinusoids is uniquely determined by its value at $t = 0$ and at *any* other $2k$ points in time [46,47]. This represents far fewer samples than the number of Nyquist-rate samples when k is small and the range of possible frequencies is large. In the

1990s, this work was generalized by George, Gorodnitsky, and Rao, who studied sparsity in biomagnetic imaging and other contexts [134–136, 202]. Simultaneously, Bresler, Feng, and Venkataramani proposed a sampling scheme for acquiring certain classes of signals consisting of k components with nonzero bandwidth (as opposed to pure sinusoids) under restrictions on the possible spectral supports, although exact recovery was not guaranteed in general [29, 117, 118, 237]. In the early 2000s Blu, Marziliano, and Vetterli developed sampling methods for certain classes of parametric signals that are governed by only k parameters, showing that these signals can be sampled and recovered from just $2k$ samples [239].

A related problem focuses on recovery of a signal from partial observation of its Fourier transform. Beurling proposed a method for extrapolating these observations to determine the entire Fourier transform [22]. One can show that if the signal consists of a finite number of impulses, then Beurling's approach will correctly recover the entire Fourier transform (of this non-bandlimited signal) from *any* sufficiently large piece of its Fourier transform. His approach – to find the signal with smallest ℓ_1 norm among all signals agreeing with the acquired Fourier measurements – bears a remarkable resemblance to some of the algorithms used in CS.

More recently, Candès, Romberg, Tao [33, 40–42, 44], and Donoho [82] showed that a signal having a sparse representation can be recovered *exactly* from a small set of linear, non-adaptive measurements. This result suggests that it may be possible to sense sparse signals by taking far fewer measurements, hence the name *compressed* sensing. Note, however, that CS differs from classical sampling in three important respects. First, sampling theory typically considers infinite-length, continuous-time signals. In contrast, CS is a mathematical theory focused on measuring finite-dimensional vectors in $\mathbb{R}^n$. Second, rather than sampling the signal at specific points in time, CS systems typically acquire measurements in the form of inner products between the signal and more general test functions. This is in fact in the spirit of modern sampling methods which similarly acquire signals by more general linear measurements [113, 230]. We will see throughout this book that *randomness* often plays a key role in the design of these test functions. Third, the two frameworks differ in the manner in which they deal with *signal recovery*, i.e., the problem of recovering the original signal from the compressive measurements. In the Nyquist–Shannon framework, signal recovery is achieved through sinc interpolation – a linear process that requires little computation and has a simple interpretation. In CS, however, signal recovery is typically achieved using highly nonlinear methods.[1] See Section 1.6 as well as the survey in [226] for an overview of these techniques.

Compressed sensing has already had a notable impact on several applications. One example is medical imaging [178–180, 227], where it has enabled speedups by a factor of seven in pediatric MRI while preserving diagnostic quality [236]. Moreover, the broad applicability of this framework has inspired research that extends

[1] It is also worth noting that it has recently been shown that nonlinear methods can be used in the context of traditional sampling as well, when the sampling mechanism is nonlinear [105].

the CS framework by proposing practical implementations for numerous applications, including sub-Nyquist sampling systems [125, 126, 186–188, 219, 224, 225, 228], compressive imaging architectures [99, 184, 205], and compressive sensor networks [7, 72, 141].

The aim of this book is to provide an up-to-date review of some of the important results in CS. Many of the results and ideas in the various chapters rely on the fundamental concepts of CS. Since the focus of the remaining chapters is on more recent advances, we concentrate here on many of the basic results in CS that will serve as background material to the rest of the book. Our goal in this chapter is to provide an overview of the field and highlight some of the key technical results, which are then more fully explored in subsequent chapters. We begin with a brief review of the relevant mathematical tools, and then survey many of the low-dimensional models commonly used in CS, with an emphasis on sparsity and the union of subspaces models. We next focus attention on the theory and algorithms for sparse recovery in finite dimensions. To facilitate our goal of providing both an elementary introduction as well as a comprehensive overview of many of the results in CS, we provide proofs of some of the more technical lemmas and theorems in the Appendix.

1.2 Review of vector spaces

For much of its history, signal processing has focused on signals produced by physical systems. Many natural and man-made systems can be modeled as linear. Thus, it is natural to consider signal models that complement this kind of linear structure. This notion has been incorporated into modern signal processing by modeling signals as *vectors* living in an appropriate *vector space*. This captures the linear structure that we often desire, namely that if we add two signals together then we obtain a new, physically meaningful signal. Moreover, vector spaces allow us to apply intuitions and tools from geometry in $\mathbb{R}^3$, such as lengths, distances, and angles, to describe and compare signals of interest. This is useful even when our signals live in high-dimensional or infinite-dimensional spaces. This book assumes that the reader is relatively comfortable with vector spaces. We now provide a brief review of some of the key concepts in vector spaces that will be required in developing the CS theory.

1.2.1 Normed vector spaces

Throughout this book, we will treat signals as real-valued functions having domains that are either continuous or discrete, and either infinite or finite. These assumptions will be made clear as necessary in each chapter. We will typically be concerned with *normed vector spaces*, i.e., vector spaces endowed with a *norm*.

In the case of a discrete, finite domain, we can view our signals as vectors in an n-dimensional Euclidean space, denoted by $\mathbb{R}^n$. When dealing with vectors in $\mathbb{R}^n$, we will

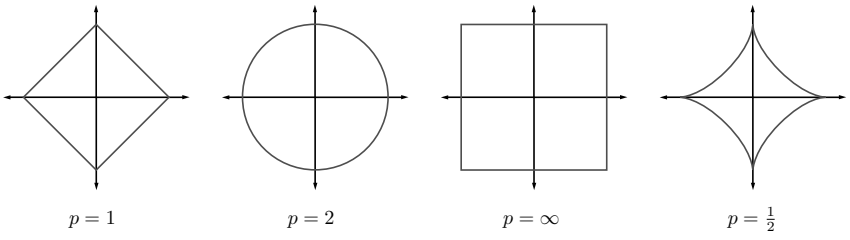


Figure 1.1 Unit spheres in $\mathbb{R}^2$ for the ℓ_p norms with $p = 1, 2, \infty$, and for the ℓ_p quasinorm with $p = \frac{1}{2}$.

make frequent use of the ℓ_p norms, which are defined for $p \in [1, \infty]$ as

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, & p \in [1, \infty); \\ \max\limits_{i=1,2,\ldots,n} |x_i|, & p = \infty. \end{cases} \tag{1.1}$$

In Euclidean space we can also consider the standard *inner product* in $\mathbb{R}^n$, which we denote

$$\langle x, z \rangle = z^T x = \sum_{i=1}^n x_i z_i.$$

This inner product leads to the ℓ_2 norm: $\|x\|_2 = \sqrt{\langle x, x \rangle}$.

In some contexts it is useful to extend the notion of ℓ_p norms to the case where $p < 1$. In this case, the "norm" defined in (1.1) fails to satisfy the triangle inequality, so it is actually a quasinorm. We will also make frequent use of the notation $\|x\|_0 := |\mathrm{supp}(x)|$, where $\mathrm{supp}(x) = \{i : x_i \neq 0\}$ denotes the support of x and $|\mathrm{supp}(x)|$ denotes the cardinality of $\mathrm{supp}(x)$. Note that $\|\cdot\|_0$ is not even a quasinorm, but one can easily show that

$$\lim_{p \to 0} \|x\|_p^p = |\mathrm{supp}(x)|,$$

justifying this choice of notation. The ℓ_p (quasi-)norms have notably different properties for different values of p. To illustrate this, in Figure 1.1 we show the unit sphere, i.e., $\{x : \|x\|_p = 1\}$, induced by each of these norms in $\mathbb{R}^2$.

We typically use norms as a measure of the strength of a signal, or the size of an error. For example, suppose we are given a signal $x \in \mathbb{R}^2$ and wish to approximate it using a point in a one-dimensional affine space A. If we measure the approximation error using an ℓ_p norm, then our task is to find the $\widehat{x} \in A$ that minimizes $\|x - \widehat{x}\|_p$. The choice of p will have a significant effect on the properties of the resulting approximation error. An example is illustrated in Figure 1.2. To compute the closest point in A to x using each ℓ_p norm, we can imagine growing an ℓ_p sphere centered on x until it intersects with A. This will be the point $\widehat{x} \in A$ that is closest to x in the corresponding ℓ_p norm. We observe that larger p tends to spread out the error more evenly among the two coefficients, while smaller p leads to an error that is more unevenly distributed and tends to be sparse. This intuition generalizes to higher dimensions, and plays an important role in the development of CS theory.

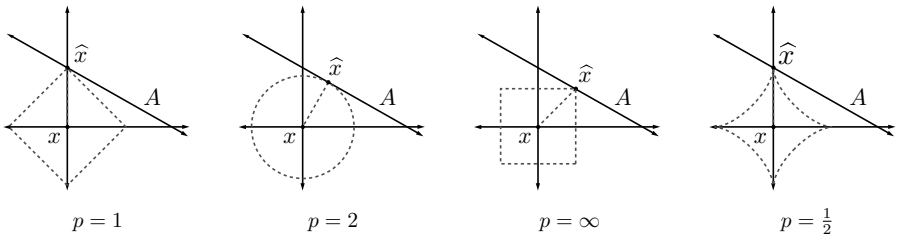


Figure 1.2 Best approximation of a point in $\mathbb{R}^2$ by a one-dimensional subspace using the ℓ_p norms for $p = 1, 2, \infty$, and the ℓ_p quasinorm with $p = \frac{1}{2}$.

1.2.2 Bases and frames

A set $\{\phi_i\}_{i=1}^n$ is called a basis for $\mathbb{R}^n$ if the vectors in the set span $\mathbb{R}^n$ and are linearly independent.[2] This implies that each vector in the space has a unique representation as a linear combination of these basis vectors. Specifically, for any $x \in \mathbb{R}^n$, there exist (unique) coefficients $\{c_i\}_{i=1}^n$ such that

$$x = \sum_{i=1}^{n} c_i \phi_i.$$

Note that if we let Φ denote the $n \times n$ matrix with columns given by ϕ_i and let c denote the length-n vector with entries c_i, then we can represent this relation more compactly as

$$x = \Phi c.$$

An important special case of a basis is an orthonormal basis, defined as a set of vectors $\{\phi_i\}_{i=1}^n$ satisfying

$$\langle \phi_i, \phi_j \rangle = \begin{cases} 1, & i = j; \\ 0, & i \neq j. \end{cases}$$

An orthonormal basis has the advantage that the coefficients c can be easily calculated as

$$c_i = \langle x, \phi_i \rangle,$$

or

$$c = \Phi^T x$$

in matrix notation. This can easily be verified since the orthonormality of the columns of Φ means that $\Phi^T \Phi = I$, where I denotes the $n \times n$ identity matrix.

It is often useful to generalize the concept of a basis to allow for sets of possibly linearly dependent vectors, resulting in what is known as a *frame* [48,55,65,163,164,182]. More

[2] In any n-dimensional vector space, a basis will always consist of exactly n vectors. Fewer vectors are not sufficient to span the space, while additional vectors are guaranteed to be linearly dependent.

formally, a frame is a set of vectors $\{\phi_i\}_{i=1}^n$ in $\mathbb{R}^d$, $d < n$ corresponding to a matrix $\Phi \in \mathbb{R}^{d\times n}$, such that for all vectors $x \in \mathbb{R}^d$,

$$A\|x\|_2^2 \le \left\|\Phi^T x\right\|_2^2 \le B\|x\|_2^2$$

with $0 < A \le B < \infty$. Note that the condition $A > 0$ implies that the rows of Φ must be linearly independent. When A is chosen as the largest possible value and B as the smallest for these inequalities to hold, then we call them the *(optimal) frame bounds*. If A and B can be chosen as $A = B$, then the frame is called *A-tight*, and if $A = B = 1$, then Φ is a *Parseval frame*. A frame is called *equal-norm*, if there exists some $\lambda > 0$ such that $\|\phi_i\|_2 = \lambda$ for all $i = 1, \dots, n$, and it is *unit-norm* if $\lambda = 1$. Note also that while the concept of a frame is very general and can be defined in infinite-dimensional spaces, in the case where Φ is a $d \times n$ matrix A and B simply correspond to the smallest and largest eigenvalues of $\Phi\Phi^T$, respectively.

Frames can provide richer representations of data due to their redundancy [26]: for a given signal x, there exist infinitely many coefficient vectors c such that $x = \Phi c$. In order to obtain a set of feasible coefficients we exploit the *dual frame* $\widetilde{\Phi}$. Specifically, any frame satisfying

$$\Phi\widetilde{\Phi}^T = \widetilde{\Phi}\Phi^T = I$$

is called an (alternate) dual frame. The particular choice $\tilde{\Phi} = (\Phi\Phi^T)^{-1}\Phi$ is referred to as the *canonical dual frame*. It is also known as the Moore–Penrose pseudoinverse. Note that since $A > 0$ requires Φ to have linearly independent rows, this also ensures that $\Phi\Phi^T$ is invertible, so that $\tilde{\Phi}$ is well-defined. Thus, one way to obtain a set of feasible coefficients is via

$$c_d = \widetilde{\Phi}^T x = \Phi^T(\Phi\Phi^T)^{-1}x.$$

One can show that this sequence is the smallest coefficient sequence in ℓ_2 norm, i.e., $\|c_d\|_2 \le \|c\|_2$ for all c such that $x = \Phi c$.

Finally, note that in the sparse approximation literature, it is also common for a basis or frame to be referred to as a *dictionary* or *overcomplete dictionary* respectively, with the dictionary elements being called *atoms*.

1.3 Low-dimensional signal models

At its core, signal processing is concerned with efficient algorithms for acquiring, processing, and extracting information from different types of signals or data. In order to design such algorithms for a particular problem, we must have accurate models for the signals of interest. These can take the form of generative models, deterministic classes, or probabilistic Bayesian models. In general, models are useful for incorporating *a priori* knowledge to help distinguish classes of interesting or probable signals from uninteresting or improbable signals. This can help in efficiently and accurately acquiring, processing, compressing, and communicating data and information.

As noted in the introduction, much of classical signal processing is based on the notion that signals can be modeled as vectors living in an appropriate vector space (or

subspace). To a large extent, the notion that any possible vector is a valid signal has driven the explosion in the dimensionality of the data we must sample and process. However, such simple linear models often fail to capture much of the structure present in many common classes of signals – while it may be reasonable to model signals as vectors, in many cases not all possible vectors in the space represent valid signals. In response to these challenges, there has been a surge of interest in recent years, across many fields, in a variety of *low-dimensional signal models* that quantify the notion that the number of degrees of freedom in high-dimensional signals is often quite small compared to their ambient dimensionality.

In this section we provide a brief overview of the most common low-dimensional structures encountered in the field of CS. We will begin by considering the traditional sparse models for finite-dimensional signals, and then discuss methods for generalizing these classes to infinite-dimensional (continuous-time) signals. We will also briefly discuss low-rank matrix and manifold models and describe some interesting connections between CS and some other emerging problem areas.

1.3.1 Sparse models

Signals can often be well-approximated as a linear combination of just a few elements from a known basis or dictionary. When this representation is exact we say that the signal is *sparse*. Sparse signal models provide a mathematical framework for capturing the fact that in many cases these high-dimensional signals contain relatively little information compared to their ambient dimension. Sparsity can be thought of as one incarnation of *Occam's razor* — when faced with many possible ways to represent a signal, the simplest choice is the best one.

Sparsity and nonlinear approximation

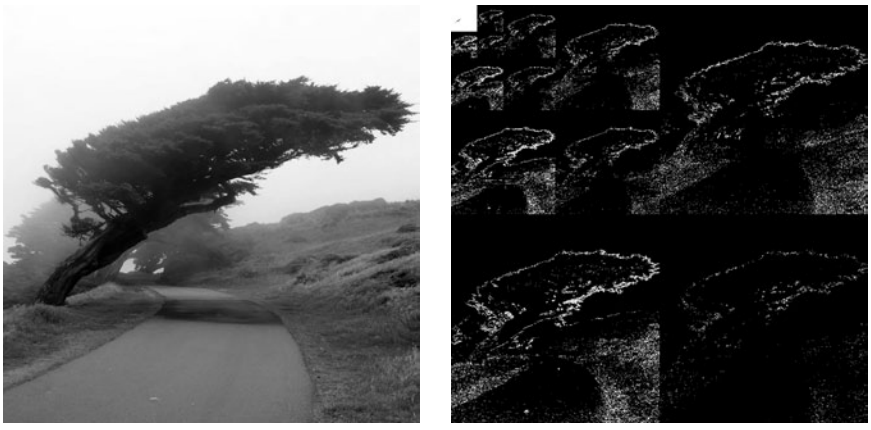
Mathematically, we say that a signal x is k-*sparse* when it has at most k nonzeros, i.e., $\|x\|_0 \leq k$. We let

$$\Sigma_k = \{x : \|x\|_0 \leq k\}$$

denote the set of all k-sparse signals. Typically, we will be dealing with signals that are not themselves sparse, but which admit a sparse representation in some basis Φ. In this case we will still refer to x as being k-sparse, with the understanding that we can express x as $x = \Phi c$ where $\|c\|_0 \leq k$.

Sparsity has long been exploited in signal processing and approximation theory for tasks such as compression [77,199,215] and denoising [80], and in statistics and learning theory as a method for avoiding overfitting [234]. Sparsity also figures prominently in the theory of statistical estimation and model selection [139, 218], in the study of the human visual system [196], and has been exploited heavily in image processing tasks, since the multiscale wavelet transform [182] provides nearly sparse representations for natural images. An example is shown in Figure 1.3.

As a traditional application of sparse models, we consider the problems of image compression and image denoising. Most natural images are characterized by large smooth or



(a) (b)

Figure 1.3 Sparse representation of an image via a multiscale wavelet transform. (a) Original image. (b) Wavelet representation. Large coefficients are represented by light pixels, while small coefficients are represented by dark pixels. Observe that most of the wavelet coefficients are close to zero.

textured regions and relatively few sharp edges. Signals with this structure are known to be very nearly sparse when represented using a multiscale wavelet transform [182]. The wavelet transform consists of recursively dividing the image into its low- and high-frequency components. The lowest frequency components provide a coarse scale approximation of the image, while the higher frequency components fill in the detail and resolve edges. What we see when we compute a wavelet transform of a typical natural image, as shown in Figure 1.3, is that most coefficients are very small. Hence, we can obtain a good approximation of the signal by setting the small coefficients to zero, or *thresholding* the coefficients, to obtain a k-sparse representation. When measuring the approximation error using an ℓ_p norm, this procedure yields the *best k-term approximation* of the original signal, i.e., the best approximation of the signal using only k basis elements.[3]

Figure 1.4 shows an example of such an image and its best k-term approximation. This is the heart of nonlinear approximation [77] – nonlinear because the choice of which coefficients to keep in the approximation depends on the signal itself. Similarly, given the knowledge that natural images have approximately sparse wavelet transforms, this same thresholding operation serves as an effective method for rejecting certain common types of noise, which typically do *not* have sparse wavelet transforms [80].

[3] Thresholding yields the best k-term approximation of a signal with respect to an orthonormal basis. When redundant frames are used, we must rely on sparse approximation algorithms like those described in Section 1.6 [106, 182].

(a) (b)

Figure 1.4 Sparse approximation of a natural image. (a) Original image. (b) Approximation of image obtained by keeping only the largest 10% of the wavelet coefficients.

Geometry of sparse signals

Sparsity is a highly nonlinear model, since the choice of which dictionary elements are used can change from signal to signal [77]. This can be seen by observing that given a pair of k-sparse signals, a linear combination of the two signals will in general no longer be k-sparse, since their supports may not coincide. That is, for any $x, z \in \Sigma_k$, we do not necessarily have that $x + z \in \Sigma_k$ (although we do have that $x + z \in \Sigma_{2k}$). This is illustrated in Figure 1.5, which shows Σ_2 embedded in $\mathbb{R}^3$, i.e., the set of all 2-sparse signals in $\mathbb{R}^3$.

The set of sparse signals Σ_k does not form a linear space. Instead it consists of the union of all possible $\binom{n}{k}$ canonical subspaces. In Figure 1.5 we have only $\binom{3}{2} = 3$ possible subspaces, but for larger values of n and k we must consider a potentially huge number of subspaces. This will have significant algorithmic consequences in the development of the algorithms for sparse approximation and sparse recovery described in Sections 1.5 and 1.6.

Compressible signals

An important point in practice is that few real-world signals are *truly* sparse; rather they are compressible, meaning that they can be well-approximated by sparse signals. Such signals have been termed compressible, approximately sparse, or relatively sparse in various contexts. Compressible signals are well approximated by sparse signals in the same way that signals living close to a subspace are well approximated by the first few principal components [139]. In fact, we can quantify the compressibility by calculating the error incurred by approximating a signal x by some $\widehat{x} \in \Sigma_k$:

$$\sigma_k(x)_p = \min_{\widehat{x} \in \Sigma_k} \|x - \widehat{x}\|_p. \tag{1.2}$$