Estimating Unknown Sparsity in Compressed Sensing

Miles E. Lopes MLOPES@STAT.BERKELEY.EDU
UC Berkeley, Dept. Statistics, 367 Evans Hall, Berkeley, CA 94720-3860

Abstract

In the theory of compressed sensing (CS), the sparsity $\|x\|_0$ of the unknown signal $x \in \mathbb{R}^p$ is commonly assumed to be a known parameter. However, it is typically unknown in practice. Due to the fact that many aspects of CS depend on knowing $\|x\|_0$, it is important to estimate this parameter in a data-driven way. A second practical concern is that $\|x\|_0$ is a highly unstable function of x. In particular, for real signals with entries not exactly equal to 0, the value $\|x\|_0 = p$ is not a useful description of the effective number of coordinates. In this paper, we propose to estimate a stable measure of sparsity $s(x) := \|x\|_1^2/\|x\|_2^2$, which is a sharp lower bound on $\|x\|_0$. Our estimation procedure uses only a small number of linear measurements, does not rely on any sparsity assumptions, and requires very little computation. A confidence interval for $s(x)$ is provided, and its width is shown to have no dependence on the signal dimension p. Moreover, this result extends naturally to the matrix recovery setting, where a soft version of matrix rank can be estimated with analogous guarantees. Finally, we show that the use of randomized measurements is essential to estimating $s(x)$. This is accomplished by proving that the minimax risk for estimating $s(x)$ with deterministic measurements is large when $n \ll p$.

1. Introduction

The central problem of compressed sensing (CS) is to estimate an unknown signal $x \in \mathbb{R}^p$ from n linear measurements $y = (y_1, \ldots, y_n)$ given by

$$y = Ax + \epsilon, \tag{1}$$

where $A \in \mathbb{R}^{n \times p}$ is a user-specified measurement matrix, $\epsilon \in \mathbb{R}^n$ is a random noise vector, and n is much smaller than the signal dimension p. During the last several years, the theory of CS has drawn widespread attention to the fact that this seemingly ill-posed problem can be solved reliably when x is sparse — in the sense that the parameter $\|x\|_0 := \text{card}\{j : x_j \neq 0\}$ is much less than p. For instance, if n is approximately $\|x\|_0 \log(p/\|x\|_0)$, then accurate recovery can be achieved with high probability when A is drawn from a Gaussian ensemble (Donoho, 2006; Candès et al., 2006). Along these lines, the value of the parameter $\|x\|_0$ is commonly assumed to be known in the analysis of recovery algorithms — even though it is typically *unknown* in practice. Due to the fundamental role that sparsity plays in CS, this issue has been recognized as a significant gap between theory and practice by several authors (Ward, 2009; Eldar, 2009; Malioutov et al., 2008). Nevertheless, the literature has been relatively quiet about the problems of estimating this parameter and quantifying its uncertainty.

1.1. Motivations and the role of sparsity

At a conceptual level, the problem of estimating $\|x\|_0$ is quite different from the more well-studied problems of estimating the full signal x or its support set $S := \{j : x_j \neq 0\}$. The difference arises from sparsity assumptions. On one hand, a procedure for estimating $\|x\|_0$ should make very few assumptions about sparsity (if any). On the other hand, methods for estimating x or S often assume that a sparsity level is given, and then *impose* this value on the solution $\widehat{x}$ or $\widehat{S}$. Consequently, a simple plug-in estimate of $\|x\|_0$, such as $\|\widehat{x}\|_0$ or $\text{card}(\widehat{S})$, may fail when the sparsity assumptions underlying $\widehat{x}$ or $\widehat{S}$ are invalid.

To emphasize that there are many aspects of CS that depend on knowing $\|x\|_0$, we provide several examples below. Our main point here is that a method for estimating $\|x\|_0$ is valuable because it can help to address a broad range of issues.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

- **Modeling assumptions.** One of the core modeling assumptions invoked in applications of CS is that the signal of interest has a sparse representation. Likewise, the problem of checking whether or not this assumption is supported by data has been an active research topic, particularly in in areas of face recognition and image classification (Rigamonti et al., 2011; Shi et al., 2011). In this type of situation, an estimate $\widehat{\|x\|}_0$ that does not rely on any sparsity assumptions is a natural device for validating the use of sparse representations.

- **The number of measurements.** If the choice of n is too small compared to the "critical" number $n^*(x) := \|x\|_0 \log(p/\|x\|_0)$, then there are known information-theoretic barriers to the accurate reconstruction of x (Arias-Castro et al., 2011). At the same time, if n is chosen to be much larger than $n^*(x)$, then the measurement process is wasteful, as there are known algorithms that can reliably recover x with approximately $n^*(x)$ measurements (Davenport et al., 2011).

  To deal with the selection of n, a sparsity estimate $\widehat{\|x\|}_0$ may be used in two different ways, depending on whether measurements are collected sequentially, or in a single batch. In the sequential case, an estimate of $\|x\|_0$ can be computed from a set of "preliminary" measurements, and then the estimated value $\widehat{\|x\|}_0$ determines how many additional measurements should be collected to recover the full signal. Also, it is not always necessary to take additional measurements, since the preliminary set may be re-used to compute $\widehat{x}$ (as discussed in Section 5). Alternatively, if all of the measurements must be taken in one batch, the value $\widehat{\|x\|}_0$ can be used to certify whether or not enough measurements were actually taken.

- **The measurement matrix.** Two of the most well-known design characteristics of the matrix A are defined explicitly in terms of sparsity. These are the *restricted isometry property of order k* (RIP-k), and the *restricted null-space property of order k* (NSP-k), where k is a presumed upper bound on the sparsity level of the true signal. Since many recovery guarantees are closely tied to RIP-k and NSP-k, a growing body of work has been devoted to certifying whether or not a given matrix satisfies these properties (d'Aspremont & El Ghaoui, 2011; Juditsky & Nemirovski, 2011; Tang & Nehorai, 2011). When k is treated as given, this problem is already computationally difficult. Yet, when the sparsity of x is unknown, we must also remember that such a "certificate" is not meaningful unless we can check that k is consistent with the true signal.

- **Recovery algorithms.** When recovery algorithms are implemented, the sparsity level of x is often treated as a tuning parameter. For example, if k is a presumed bound on $\|x\|_0$, then the Orthogonal Matching Pursuit algorithm (OMP) is typically initialized to run for k iterations. A second example is the Lasso algorithm, which computes the solution $\widehat{x} \in \operatorname{argmin}\{\|y - Av\|_2^2 + \lambda\|v\|_1 : v \in \mathbb{R}^p\}$, for some choice of $\lambda \geq 0$. The sparsity of $\widehat{x}$ is determined by the size of λ, and in order to select the appropriate value, a family of solutions is examined over a range of λ values. In the case of either OMP or Lasso, a sparsity estimate $\widehat{\|x\|}_0$ would reduce computation by restricting the possible choices of λ or k, and it would also ensure that the chosen values conform to the true signal.

1.2. An alternative measure of sparsity

Despite the important theoretical role of the parameter $\|x\|_0$ in many aspects of CS, it has the practical drawback of being a highly unstable function of x. In particular, for real signals $x \in \mathbb{R}^p$ whose entries are not exactly equal to 0, the value $\|x\|_0 = p$ is not a useful description of the effective number of coordinates.

In order to estimate sparsity in a way that accounts for the instability of $\|x\|_0$, it is desirable to replace the ℓ_0 norm with a "soft" version. More precisely, we would like to identify a function of x that can be interpreted like $\|x\|_0$, but remains stable under small perturbations of x. A natural quantity that serves this purpose is the *numerical sparsity*

$$s(x) := \frac{\|x\|_1^2}{\|x\|_2^2}, \tag{2}$$

which always satisfies $1 \leq s(x) \leq p$ for any non-zero x. Although the ratio $\|x\|_1^2/\|x\|_2^2$ appears sporadically in different areas (Tang & Nehorai, 2011; Hurley & Rickard, 2009; Hoyer, 2004; Lopes et al., 2011), it does not seem to be well known as a sparsity measure in CS.

A key property of $s(x)$ is that it is a sharp lower bound on $\|x\|_0$ for all non-zero x,

$$s(x) \leq \|x\|_0, \tag{3}$$

which follows from applying the Cauchy-Schwarz inequality to the relation $\|x\|_1 = \langle x, \operatorname{sgn}(x)\rangle$. (Equality in (3) is attained iff the non-zero coordinates of x are

equal in magnitude.) We also note that this inequality is invariant to scaling of x, since $s(x)$ and $\|x\|_0$ are individually scale invariant. In the opposite direction, it is easy to see that the only continuous upper bound on $\|x\|_0$ is the trivial one: If a continuous function f satisfies $\|x\|_0 \le f(x) \le p$ for all x in some open subset of $\mathbb{R}^p$, then f must be identically equal to p. (There is a dense set of points where $\|x\|_0 = p$.) Therefore, we must be content with a continuous lower bound.

The fact that $s(x)$ is a sensible measure of sparsity for non-idealized signals is illustrated in Figure 1. In essence, if x has k large coordinates and $p-k$ small coordinates, then $s(x) \approx k$, whereas $\|x\|_0 = p$. In the left panel, the sorted coordinates of three different vectors in $\mathbb{R}^{100}$ are plotted. The value of $s(x)$ for each vector is marked with a triangle on the x-axis, which shows that $s(x)$ adapts well to the decay profile. This idea can be seen in a more geometric way in the middle and right panels, which plot the the sub-level sets $\mathcal{S}_c := \{x \in \mathbb{R}^p : s(x) \le c\}$ with $c \in [1, p]$. When $c \approx 1$, the vectors in $\mathcal{S}_c$ are closely aligned with the coordinate axes, and hence contain one effective coordinate. As $c \uparrow p$, the set $\mathcal{S}_c$ includes more dense vectors until $\mathcal{S}_p = \mathbb{R}^p$.

1.3. Related work.

Some of the challenges described in Section 1.1 can be approached with the general tools of cross-validation (CV) and empirical risk minimization (ERM). This approach has been used to select various parameters, such as the number of measurements n (Malioutov et al., 2008; Ward, 2009), the number of OMP iterations k (Ward, 2009), or the Lasso regularization parameter λ (Eldar, 2009). At a high level, these methods consider a collection of (say m) solutions $\widehat{x}^{(1)}, \dots, \widehat{x}^{(m)}$ obtained from different values $\theta_1, \dots, \theta_m$ of some tuning parameter of interest. For each solution, an empirical error estimate $\widehat{\mathrm{err}}(\widehat{x}^{(j)})$ is computed, and the value θ_{j^*} corresponding to the smallest $\widehat{\mathrm{err}}(\widehat{x}^{(j)})$ is chosen.

Although methods based on CV/ERM share common motivations with our work here, these methods differ from our approach in several ways. In particular, the problem of estimating a soft measure of sparsity, such as $s(x)$, has not been considered from that angle. Also, the cited methods do not give any theoretical guarantees to ensure that the estimated sparsity level is close to the true one. Note that even if an estimate $\widehat{x}$ has small error $\|\widehat{x} - x\|_2$, it is not necessary for $\|\widehat{x}\|_0$ to be close to $\|x\|_0$. This point is especially relevant when one is interested in identifying a set of important variables or interpreting features.

From a computational point view, the CV/ERM approaches can also be costly — since $\widehat{x}^{(j)}$ is typically computed from a separate optimization problem for for each choice of the tuning parameter. By contrast, our method for estimating $s(x)$ requires no optimization, and can be computed easily from just a small set of preliminary measurements.

1.4. Our contributions.

The primary contribution of this paper is our treatment of unknown sparsity as a *parameter estimation problem*. Specifically, we identify a stable measure of sparsity that is relevant to CS, and propose an efficient estimator with provable guarantees. Secondly, we are not aware of any other papers that have demonstrated a distinction between random and deterministic mea-

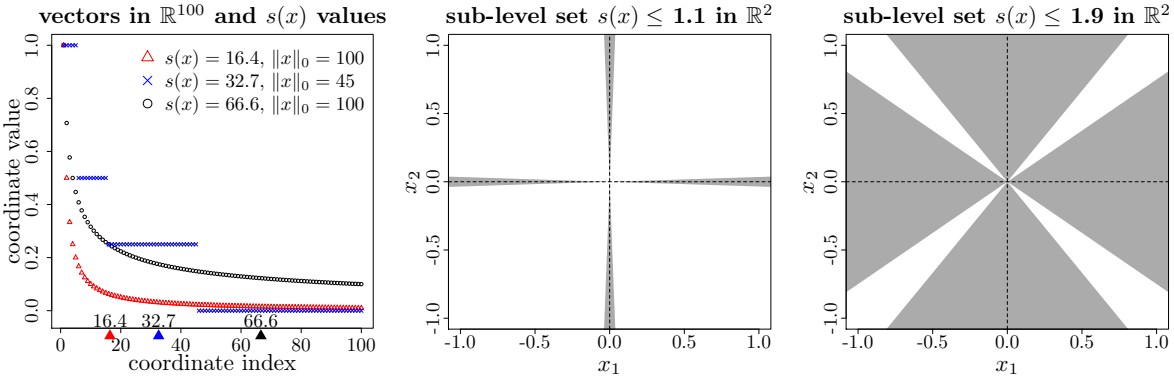

Figure 1. Characteristics of $s(x)$. Left panel: Three vectors (red, blue, black) in $\mathbb{R}^{100}$ have been plotted with their coordinates in order of decreasing size (maximum entry normalized to 1). Two of the vectors have power-law decay profiles, and one is a dyadic vector with exactly 45 positive coordinates (red: $x_i \propto i^{-1}$, blue: dyadic, black: $x_i \propto i^{-1/2}$). Color-coded triangles on the bottom axis indicate that the $s(x)$ value represents the "effective" number of coordinates.

surements with regard to unknown sparsity (as in Section 4).

The remainder of the paper is organized as follows. In Section 2, we show that a principled choice of n can be made if $s(x)$ is known. This is accomplished by formulating a recovery condition for the Basis Pursuit algorithm directly in terms of $s(x)$. Next, in Section 3, we propose an estimator $\widehat{s}(x)$, and derive a dimension-free confidence interval for $s(x)$. The procedure is also shown to extend to the problem of estimating a soft measure of rank for matrix-valued signals. In Section 4 we show that the use of randomized measurements is essential to estimating $s(x)$. Finally, we present simulations in Section 5 to validate the consequences of our theoretical results. Due to space constraints, we defer all of our proofs to the supplement.

Notation. We define $\|x\|_q^q := \sum_{j=1}^p |x_j|^q$ for any $q > 0$ and $x \in \mathbb{R}^p$, which only corresponds to a genuine norm for $q \geq 1$. For sequences of numbers a_n and b_n, we write $a_n \lesssim b_n$ or $a_n = \mathcal{O}(b_n)$ if there is an absolute constant $c > 0$ such that $a_n \leq c b_n$ for all large n. If $a_n/b_n \to 0$, we write $a_n = o(b_n)$. For a matrix M, we define the Frobenius norm $\|M\|_F = \sqrt{\sum_{i,j} M_{ij}^2}$, the matrix ℓ_1-norm $\|M\|_1 = \sum_{i,j} |M_{ij}|$. Finally, for two matrices A and B of the same size, we define the inner product $\langle A, B\rangle := \mathrm{tr}(A^\top B)$.

2. Recovery conditions in terms of $s(x)$

The purpose of this section is to present a simple proposition that links $s(x)$ with recovery conditions for the Basis Pursuit algorithm (BP). This is an important motivation for studying $s(x)$, since it implies that if $s(x)$ can be estimated well, then n can be chosen appropriately. In other words, we offer an *adaptive* choice of n.

In order to explain the connection between $s(x)$ and recovery, we first recall a standard result (Candès et al., 2006) that describes the ℓ_2 error rate of the BP algorithm. Informally, the result assumes that the noise is bounded as $\|\epsilon\|_2 \leq \epsilon_0$ for some constant $\epsilon_0 > 0$, the matrix $A \in \mathbb{R}^{n\times p}$ is drawn from a suitable ensemble, and n satisfies $n \gtrsim T\log(pe/T)$ for some $T \in \{1,\dots,p\}$ with $e = \exp(1)$. The conclusion is that with high probability, the solution $\widehat{x} \in \mathrm{argmin}\{\|v\|_1 : \|Av - y\|_2 \leq \epsilon_0, v \in \mathbb{R}^p\}$ satisfies

$$\|\widehat{x} - x\|_2 \leq c_1\,\epsilon_0 + c_2\,\frac{\|x - x_T\|_1}{\sqrt{T}}, \tag{4}$$

where $x_T \in \mathbb{R}^p$ is the best T-term approximation[1] to x, and $c_1, c_2 > 0$ are constants. This bound is a fundamental point of reference, since it matches the minimax optimal rate under certain conditions (Candès, 2006), and applies to all signals $x \in \mathbb{R}^p$ (rather than just k-sparse signals). Additional details may be found in (Cai et al., 2010) [Theorem 3.3], (Vershynin, 2010) [Theorem 5.65].

We now aim to answer the question, "If $s(x)$ were known, how large should n be in order for $\widehat{x}$ to be close to x?" Since the bound (4) assumes $n \gtrsim T\log(pe/T)$, our question amounts to choosing T. For this purpose, it is natural to consider the *relative* ℓ_2 error

$$\frac{\|\widehat{x}-x\|_2}{\|x\|_2} \leq c_1\,\frac{\epsilon_0}{\|x\|_2} + c_2\,\frac{1}{\sqrt{T}}\,\frac{\|x-x_T\|_1}{\|x\|_2}, \tag{5}$$

so that the *T-term approximation error* $\frac{1}{\sqrt{T}}\frac{\|x-x_T\|_1}{\|x\|_2}$ does not depend on the scale of x (i.e. invariant under $x \mapsto cx$ with $c \neq 0$).

Proposition 1 below shows how knowledge of $s(x)$ allows us to control the approximation error. Specifically, the result shows that the condition $T \gtrsim s(x)$ is necessary for the approximation error to be small, and the condition $T \gtrsim s(x)\log(p)$ is sufficient.

Proposition 1. *Let $x \in \mathbb{R}^p \setminus \{0\}$, and $T \in \{1,\dots,p\}$. The following statements hold for any $c_0, \varepsilon > 0$.*

(i) If the T-term approximation error satisfies $\frac{1}{\sqrt{T}}\frac{\|x-x_T\|_1}{\|x\|_2} \leq \varepsilon$, then

$$T \geq \frac{1}{(1+\varepsilon)^2}\cdot s(x).$$

(ii) If $T \geq c_0\, s(x)\log(p)$, then the T-term approximation error satisfies

$$\frac{1}{\sqrt{T}}\frac{\|x-x_T\|_1}{\|x\|_2} \leq \frac{1}{\sqrt{c_0\log(p)}}\Big(1 - \frac{T}{p}\Big).$$

In particular, if $T \geq 2s(x)\log(p)$ with $p \geq 100$, then

$$\frac{1}{\sqrt{T}}\frac{\|x-x_T\|_1}{\|x\|_2} \leq \frac{1}{3}.$$

Remarks. A notable feature of these bounds is that they hold for all non-zero signals. In our simulations in Section 5, we show that choosing $n = 2\lceil \widehat{s}(x)\rceil \log(p/\lceil \widehat{s}(x)\rceil)$ based on an estimate $\widehat{s}(x)$ leads to accurate reconstruction across many sparsity levels.

3. Estimation results for $s(x)$

In this section, we give a simple procedure to estimate $s(x)$ for any $x \in \mathbb{R}^p \setminus \{0\}$. The procedure uses a small number of measurements, makes no sparsity assumptions, and requires very little computation. The measurements we prescribe may also be *re-used* to recover the full signal after $s(x)$ has been estimated.

[1] The vector $x_T \in \mathbb{R}^p$ is obtained by setting to 0 all coordinates of x except the T largest ones (in magnitude).

The results in this section are based on the measurement model (1), written in scalar notation as

$$y_i = \langle a_i, x\rangle + \epsilon_i, \quad i = 1, \dots, n. \tag{6}$$

We assume only that the noise variables ϵ_i are independent, and bounded by $|\epsilon_i| \le \sigma_0$, for some constant $\sigma_0 > 0$. No additional structure on the noise is needed.

3.1. Sketching with stable laws

Our estimation procedure derives from a technique known as *sketching* in the streaming computation literature (Indyk, 2006). Although this area deals with problems that have mathematical connections to CS, the use of sketching techniques in CS does not seem to be well known.

For any $q \in (0, 2]$, the sketching technique offers a way to estimate $\|x\|_q$ from a set of randomized linear measurements. In our approach, we estimate $s(x) = \|x\|_1^2/\|x\|_2^2$ by estimating $\|x\|_1$ and $\|x\|_2$ from separate sets of measurements. The core idea is to generate the measurement vectors $a_i \in \mathbb{R}^p$ using *stable laws* (Zolotarev, 1986).

Definition 1. A random variable V has a *symmetric stable distribution* if its characteristic function is of the form $\mathbb{E}[\exp(\sqrt{-1}tV)] = \exp(-|\gamma t|^q)$ for some $q \in (0, 2]$ and some $\gamma > 0$. We denote the distribution by $V \sim S_q(\gamma)$, and γ is referred to as the *scale* parameter.

The most well-known examples of symmetric stable laws are the cases of $q = 2, 1$, namely the Gaussian distribution $N(0, \gamma^2) = S_2(\gamma)$, and the Cauchy distribution $C(0, \gamma) = S_1(\gamma)$. To fix some notation, if a vector $a_1 = (a_{1,1}, \dots, a_{1,p}) \in \mathbb{R}^p$ has i.i.d. entries drawn from $S_q(\gamma)$, we write $a_1 \sim S_q(\gamma)^{\otimes p}$. The connection with ℓ_q norms hinges on the following property of stable distributions (Zolotarev, 1986).

Lemma 1. *Suppose $x \in \mathbb{R}^p$, and $a_1 \sim S_q(\gamma)^{\otimes p}$ with parameters $q \in (0, 2]$ and $\gamma > 0$. Then, the random variable $\langle x, a_1\rangle$ is distributed according to $S_q(\gamma\|x\|_q)$.*

Using this fact, if we generate a set of i.i.d. vectors $a_1, \dots, a_n$ from $S_q(\gamma)^{\otimes p}$ and let $y_i = \langle a_i, x\rangle$, then $y_1, \dots, y_n$ is an i.i.d. sample from $S_q(\gamma\|x\|_q)$. Hence, in the special case of noiseless linear measurements, the task of estimating $\|x\|_q$ is equivalent to a well-studied univariate problem: *estimating the scale parameter of a stable law from an i.i.d. sample.*

When the y_i are corrupted with noise, our analysis shows that standard estimators for scale parameters are only moderately affected. The impact of the noise can also be reduced via the choice of γ when generating $a_i \sim S_q(\gamma)^{\otimes p}$. The γ parameter controls the "energy level" of the measurement vectors a_i. (Note that in the Gaussian case, if $a_1 \sim S_2(\gamma)^{\otimes p}$, then $\mathbb{E}\|a_1\|_2^2 = \gamma^2 p$.) In our results, we leave γ as a free parameter to show how the effect of noise is reduced as γ is increased.

3.2. Estimation procedure for $s(x)$

Two sets of measurements are used to estimate $s(x)$, and we write the total number as $n = n_1 + n_2$. The first set is obtained by generating i.i.d. measurement vectors from a Cauchy distribution,

$$a_i \sim C(0, \gamma)^{\otimes p}, \quad i = 1, \dots, n_1. \tag{7}$$

The corresponding values y_i are then used to estimate $\|x\|_1$ via the statistic

$$\widehat{T}_1 := \tfrac{1}{\gamma}\text{median}(|y_1|, \dots, |y_{n_1}|), \tag{8}$$

which is a standard estimator of the scale parameter of the Cauchy distribution (Fama & Roll, 1971; Li et al., 2007). Next, a second set of i.i.d. measurement vectors are generated from a Gaussian distribution

$$a_i \sim N(0, \gamma^2)^{\otimes p}, \quad i = n_1 + 1, \dots, n_1 + n_2. \tag{9}$$

In this case, the associated y_i values are used to compute an estimate of $\|x\|_2^2$ given by

$$\widehat{T}_2^2 := \tfrac{1}{\gamma^2 n_2}(y_{n_1+1}^2 + \cdots + y_{n_1+n_2}^2), \tag{10}$$

which is a natural estimator of the variance of a Gaussian distribution. Combining these two statistics, our estimate of $s(x) = \|x\|_1^2/\|x\|_2^2$ is defined as

$$\widehat{s}(x) := \widehat{T}_1^2/\widehat{T}_2^2. \tag{11}$$

3.3. Confidence interval.

The following theorem describes the relative error $\left|\frac{\widehat{s}(x)}{s(x)} - 1\right|$ via an asymptotic confidence interval for $s(x)$. Our result is stated in terms of the noise-to-signal ratio

$$\rho := \tfrac{\sigma_0}{\gamma\|x\|_2},$$

and the standard Gaussian quantile $z_{1-\alpha}$, which satisfies $\Phi(z_{1-\alpha}) = 1 - \alpha$ for any coverage level $\alpha \in (0, 1)$. In this notation, the following parameters govern the width of the confidence interval,

$$\eta_n(\alpha, \rho) := \tfrac{z_{1-\alpha}}{\sqrt{n}} + \rho \quad \text{and} \quad \delta_n(\alpha, \rho) := \tfrac{\pi z_{1-\alpha}}{\sqrt{2n}} + \rho,$$

and we write these simply as δ_n and η_n. As is standard in high-dimensional statistics, we allow all of the model parameters p, x, σ_0 and γ to vary implicitly as functions of (n_1, n_2), and let $(n_1, n_2, p) \to \infty$. For simplicity, we choose to take measurement sets of equal sizes,

$n_1 = n_2 = n/2$, and we place a mild constraint on ρ, namely $\eta_n(\alpha, \rho) < 1$. (Note that standard algorithms such as Basis Pursuit are not expected to perform well unless $\rho \ll 1$, as is clear from the bound (5).) Lastly, we make no restriction on the growth of p/n, which makes $\widehat{s}(x)$ well-suited to high-dimensional problems.

Theorem 1. *Let $\alpha \in (0, 1/2)$ and $x \in \mathbb{R}^p \setminus \{0\}$. Assume that $\widehat{s}(x)$ is constructed as above, and that the model (6) holds. Suppose also that $n_1 = n_2 = n/2$ and $\eta_n(\alpha, \rho) < 1$ for all n. Then as $(n, p) \to \infty$, we have*

$$\mathbb{P}\Big(\sqrt{\tfrac{\widehat{s}(x)}{s(x)}} \in \big[\tfrac{1-\delta_n}{1+\eta_n}, \tfrac{1+\delta_n}{1-\eta_n}\big]\Big) \geq (1-2\alpha)^2 + o(1). \quad (12)$$

Remarks. The most important feature of this result is that the width of the confidence interval does *not* depend on the dimension or sparsity of the unknown signal. Concretely, this means that the number of measurements needed to estimate $s(x)$ to a fixed precision is only $\mathcal{O}(1)$ with respect to the size of the problem. Our simulations in Section 5 also show that the relative error of $\widehat{s}(x)$ does not depend on dimension or sparsity of x. Lastly, when δ_n and η_n are small, we note that the relative error $|\widehat{s}(x)/s(x) - 1|$ is at most of order $(n^{-1/2} + \rho)$ with high probability, which follows from the simple Taylor expansion $\frac{(1+\varepsilon)^2}{(1-\varepsilon)^2} = 1 + 4\varepsilon + o(\varepsilon)$.

3.4. Estimating rank and sparsity of matrices

The framework of CS naturally extends to the problem of recovering an unknown matrix $X \in \mathbb{R}^{p_1 \times p_2}$ on the basis of the measurement model

$$y = \mathcal{A}(X) + \epsilon, \quad (13)$$

where $y \in \mathbb{R}^n$, $\mathcal{A}$ is a user-specified linear operator from $\mathbb{R}^{p_1 \times p_2}$ to $\mathbb{R}^n$, and $\epsilon \in \mathbb{R}^n$ is a vector of noise variables. In recent years, many researchers have explored the recovery of X when it is assumed to have sparse or low rank structure. We refer to the papers (Candès & Plan, 2011; Chandrasekaran et al., 2010) for descriptions of numerous applications. In analogy with the previous section, the parameters $\mathrm{rank}(X)$ or $\|X\|_0$ play important theoretical roles, but are very sensitive to perturbations of X. Likewise, it is of basic interest to estimate stable measures of rank and sparsity for matrices. Since the sparsity analogue $s(X) := \|X\|_1^2/\|X\|_F^2$ can be estimated as a straightforward extension of Section 3.2, we restrict our attention to the more distinct problem of rank estimation.

3.4.1. The rank of semidefinite matrices

In the context of recovering a low-rank positive semidefinite matrix $X \in \mathbb{S}_+^{p \times p} \setminus \{0\}$, the quantity $\mathrm{rank}(X)$ plays the role of $\|x\|_0$ in the recovery of a sparse vector. If we let $\lambda(X) \in \mathbb{R}_+^p$ denote the vector of ordered eigenvalues of X, the connection can be made explicit by writing $\mathrm{rank}(X) = \|\lambda(X)\|_0$. As in Section 3.2, our approach is to consider a robust alternative to the rank. Motivated by the quantity $s(x) = \|x\|_1^2/\|x\|_2^2$ in the vector case, we now consider

$$r(X) := \tfrac{\|\lambda(X)\|_1^2}{\|\lambda(X)\|_2^2} = \tfrac{\mathrm{tr}(X)^2}{\|X\|_F^2}$$

as our measure of the effective rank for non-zero X, which always satisfies $1 \leq r(X) \leq p$. The quantity $r(X)$ has appeared elsewhere as a measure of rank (Lopes et al., 2011; Tang & Nehorai, 2010), but is less well known than other rank relaxations, such as the *numerical rank* $\|X\|_F^2/\|X\|_{\mathrm{op}}^2$ (Rudelson & Vershynin, 2007). The relationship between $r(X)$ and $\mathrm{rank}(X)$ is completely analogous to $s(x)$ and $\|x\|_0$. Namely, we have a sharp, scale-invariant inequality

$$r(X) \leq \mathrm{rank}(X).$$

The quantity $r(X)$ is more stable than $\mathrm{rank}(X)$ in the sense that if X has k large eigenvalues, and $p-k$ small eigenvalues, then $r(X) \approx k$, whereas $\mathrm{rank}(X) = p$.

Our procedure for estimating $r(X)$ is based on estimating $\mathrm{tr}(X)$ and $\|X\|_F^2$ from separate sets of measurements. The semidefinite condition is exploited through the basic relation $\langle I_{p\times p}, X\rangle = \mathrm{tr}(X) = \|\lambda(X)\|_1$. To estimate $\mathrm{tr}(X)$, we use n_1 linear measurements of the form

$$y_i = \langle \gamma I_{p\times p}, X\rangle + \epsilon_i, \quad i = 1, \ldots, n_1 \quad (14)$$

and compute the estimator $\breve{T}_1 := \frac{1}{\gamma}\frac{1}{n_1}\sum_{i=1}^{n_1} y_i$, where $\gamma > 0$ is again the measurement energy parameter. Next, to estimate $\|X\|_F^2$, we note that if $Z \in \mathbb{R}^{p\times p}$ has i.i.d. $N(0,1)$ entries, then $\mathbb{E}\langle X, Z\rangle^2 = \|X\|_F^2$. Hence, if we collect n_2 additional measurements of the form

$$y_i = \langle \gamma Z_i, X\rangle + \epsilon_i, \quad i = n_1 + 1, \ldots, n_1 + n_2, \quad (15)$$

where the $Z_i \in \mathbb{R}^{p\times p}$ are independent random matrices with i.i.d. $N(0,1)$ entries, then a suitable estimator of $\|X\|_F^2$ is $\breve{T}_2^2 := \frac{1}{\gamma^2 n_2}\sum_{i=n_1+1}^{n_1+n_2} y_i^2$. Combining these statistics, we propose

$$\widehat{r}(X) := \breve{T}_1^2/\breve{T}_2^2$$

as our estimate of $r(X)$. In principle, this procedure can be refined by using the measurements (14) to estimate the noise distribution, but we omit these details. Also, we retain the assumptions of the previous section, and assume only that the ϵ_i are independent and bounded by $|\epsilon_i| \leq \sigma_0$. The next theorem shows that the estimator $\widehat{r}(X)$ mirrors $\widehat{s}(X)$ as in Theorem 1, but with ρ being replaced by $\varrho := \sigma_0/(\gamma\|X\|_F)$, and with η_n being replaced by $\zeta_n = \zeta_n(\varrho, \alpha) := z_{1-\alpha}/\sqrt{n} + \varrho$.

Theorem 2. *Let $\alpha \in (0, 1/2)$ and $X \in \mathbb{S}_+^{p\times p} \setminus \{0\}$. Assume that $\widehat{r}(X)$ is constructed as above, and that the model (13) holds. Suppose also that $n_1 = n_2 = n/2$ and $\zeta_n(\alpha, \rho) < 1$ for all n. Then as $(n, p) \to \infty$, we have*

$$\mathbb{P}\Big(\sqrt{\tfrac{\widehat{r}(X)}{r(X)}} \in \big[\tfrac{1-\varrho}{1+\zeta_n}, \tfrac{1+\varrho}{1-\zeta_n}\big]\Big) \geq 1 - 2\alpha + o(1). \quad (16)$$

Remarks. In parallel with Theorem 1, this confidence interval has the valuable property that its width does not depend on the rank or dimension of X, but merely on the noise-to-signal ratio $\varrho = \sigma_0/(\gamma\|X\|_F)$. The relative error $|\widehat{r}(X)/r(X) - 1|$ is at most of order $(n^{-1/2} + \varrho)$ with high probability when ζ_n is small.

4. Deterministic measurement matrices

The problem of constructing deterministic matrices A with good recovery properties (e.g. RIP-k or NSP-k) has been a longstanding topic within CS. Since our procedure in Section 3.2 selects A at random, it is natural to ask if randomization is essential to the estimation of unknown sparsity. In this section, we show that estimating $s(x)$ with a deterministic matrix A leads to results that are inherently different from our randomized procedure.

At an informal level, the difference between random and deterministic matrices makes sense if we think of the estimation problem as a game between nature and a statistician. Namely, the statistician first chooses a matrix $A \in \mathbb{R}^{n\times p}$ and an estimation rule $\delta : \mathbb{R}^n \to \mathbb{R}$. (The function δ takes $y \in \mathbb{R}^n$ as input and returns an estimate of $s(x)$.) In turn, nature chooses a signal $x \in \mathbb{R}^p \setminus \{0\}$, with the goal of maximizing the statistician's error. When the statistician chooses A deterministically, nature has the freedom to adversarially select an x that is ill-suited to the fixed matrix A. By contrast, if the statistician draws A at random, then nature does not know what value A will take, and therefore has less knowledge to choose a "bad" signal.

In the case of noiseless *random* measurements, Theorem 1 implies that our particular estimation rule $\widehat{s}(x)$ can achieve a relative error of order $|\widehat{s}(x)/s(x) - 1| = \mathcal{O}(n^{-1/2})$ with high probability for any non-zero x. (cf. Remarks for Theorem 1.) Our aim is now to show that for noiseless *deterministic* measurements, *all* estimation rules δ have a worst-case relative error $|\delta(Ax)/s(x) - 1|$ that is much larger than than $n^{-1/2}$. In other words, there is always a choice of x that can defeat a deterministic procedure, whereas $\widehat{s}(x)$ is likely to succeed under any choice of x.

In stating the following result, we note that it involves no randomness whatsoever — since we assume that the observed measurements $y = Ax$ are noiseless and obtained from a deterministic matrix A.

Theorem 3. *The minimax relative error for estimating $s(x)$ from noiseless deterministic measurements $y = Ax$ satisfies*

$$\inf_{A\in\mathbb{R}^{n\times p}} \inf_{\delta:\mathbb{R}^n\to\mathbb{R}} \sup_{x\in\mathbb{R}^p\setminus\{0\}} \Big|\tfrac{\delta(Ax)}{s(x)} - 1\Big| \geq \tfrac{1-(n+1)/p}{2(1+2\sqrt{2\log(2p)})^2}.$$

Remarks. Under the typical high-dimensional scenario where there is some $\kappa \in (0, \infty)$ for which $p/n \to \kappa$ as $(n, p) \to \infty$, we have the lower bound $|\frac{\delta(Ax)}{s(x)} - 1| \gtrsim \frac{1}{\log(n)}$, which is much larger than $n^{-1/2}$.

5. Simulations

Relative error of $\widehat{s}(x)$. To validate the consequences of Theorem 1, we study how the relative error $|\widehat{s}(x)/s(x) - 1|$ depends on the parameters p, ρ, and $s(x)$. We generated measurements $y = Ax + \epsilon$ under a broad range of parameter settings, with $x \in \mathbb{R}^{10^4}$ in most cases. Note that although $p = 10^4$ is a very large dimension, it is not at all extreme from the viewpoint of applications (e.g. a megapixel image with $p = 10^6$). Details regarding parameter settings are given below. As anticipated by Theorem 1, the left and right panels in Figure 2 show that the relative error has no noticeable dependence on p or $s(x)$. The middle panel shows that for fixed $n_1 + n_2$, the relative error grows moderately with $\rho = \frac{\sigma_0}{\gamma\|x\|_2}$. Lastly, our theoretical bounds on $|\widehat{s}(x)/s(x) - 1|$ conform to the empirical curves in the case of low noise ($\rho = 10^{-2}$).

Reconstruction of x based on $\widehat{s}(x)$. For the problem of choosing n, we considered the choice of $\widehat{n} := 2\lceil\widehat{s}(x)\rceil \log(p/\lceil\widehat{s}(x)\rceil)$. The simulations show that $\widehat{n}$ adapts to the structure of the true signal, and is also sufficiently large for accurate reconstruction. First, to compute $\widehat{s}(x)$, we followed Section 3.2, and drew initial measurement sets of Cauchy and Gaussian vectors with $n_1 = n_2 = 500$ and $\gamma = 1$. If it happened to be the case that $500 \geq \widehat{n}$, then reconstruction was performed using only the initial 500 Gaussian measurements. Alternatively, if $\widehat{n} > 500$, then $(\widehat{n} - 500)$ additional measurements vectors a_i were drawn from $N(0, \widehat{n}^{-1/2} I_{p\times p})$ for reconstruction. Further details are given below. Figure 3 illustrates the results for three power-law signals in $\mathbb{R}^{10^4}$ with $x_{[i]} \propto i^{-\nu}$, $\nu = 0.7, 1.0, 1.3$ and $\|x\|_1 = 1$ (corresponding to $s(x) = 823, 58, 11$). In each panel, the coordinates of x are plotted in black, and those of $\widehat{x}$ are plotted in red. Clearly, there is good qualitative agreement in all cases. From left to right, the value of $\widehat{n} = 2\lceil\widehat{s}(x)\rceil \log(p/\lceil\widehat{s}(x)\rceil)$ was 4108, 590, and 150.

Settings for relative error of $\widehat{s}(x)$ (Figure 2). For each parameter setting labeled in the figures, we let $n_1 = n_2$ and averaged $|\widehat{s}(x)/s(x) - 1|$ over 200 problem instances of $y = Ax + \epsilon$. In all cases, the matrix A was chosen according to (7) and (9) with $\gamma = 1$ and $\epsilon_i \sim \text{Uniform}[-\sigma_0, \sigma_0]$. We always chose the normalization $\|x\|_2 = 1$, and $\gamma = 1$ so that $\rho = \sigma_0$. (In the left and right panels, $\rho = 10^{-2}$.) For the left and middle panels, all signals have the decay profile $x_{[i]} \propto i^{-1}$. For the right panel, the values $s(x) = 2, 58, 4028$, and 9878 were obtained using decay profiles $x_i \propto i^{-\nu}$ with $\nu = 2, 1, 1/2, 1/10$. In the left and right panels, we chose $p = 10^4$ for all curves. A theoretical bound on $|\widehat{s}(x)/s(x) - 1|$ in black was computed from Theorem 1 with $\alpha = \frac{1}{2} - \frac{1}{2\sqrt{2}}$. (This bound holds with probability at least $1/2$, and hence may be reasonably plotted against the average of $|\widehat{s}(x)/s(x) - 1|$.)

Settings for reconstruction (Figure 3). We computed reconstructions using the SPGL1 solver (van den Berg & Friedlander, 2007; 2008) for the BP problem $\widehat{x} \in \text{argmin}\{\|v\|_1 : \|Av - y\|_2 \le \epsilon_0, v \in \mathbb{R}^p\}$, with the choice $\epsilon_0 = \sigma_0\sqrt{\widehat{n}}$ being based on i.i.d. noises $\epsilon_i \sim \text{Uniform}[-\sigma_0, \sigma_0]$ and $\sigma_0 = 0.001$. When re-using the first 500 Gaussian measurements, we re-scaled the vectors a_i and the values y_i by $1/\sqrt{\widehat{n}}$ so that the matrix $A \in \mathbb{R}^{\widehat{n} \times p}$ would be expected to satisfy the RIP property. For each choice of $\nu = 0.7, 1.0, 1.3$, we generated 25 problem instances and plotted the vector $\widehat{x}$ corresponding to the median of $\|\widehat{x} - x\|_2$ over the 25 runs (so that the plots reflect typical performance).

Acknowledgements

MEL thanks the reviewers for constructive comments, as well as valuable references that substantially improved the paper. Peter Bickel is thanked for helpful discussions, and the DOE CSGF fellowship is gratefully acknowledged for support under grant DE-FG02-97ER25308.

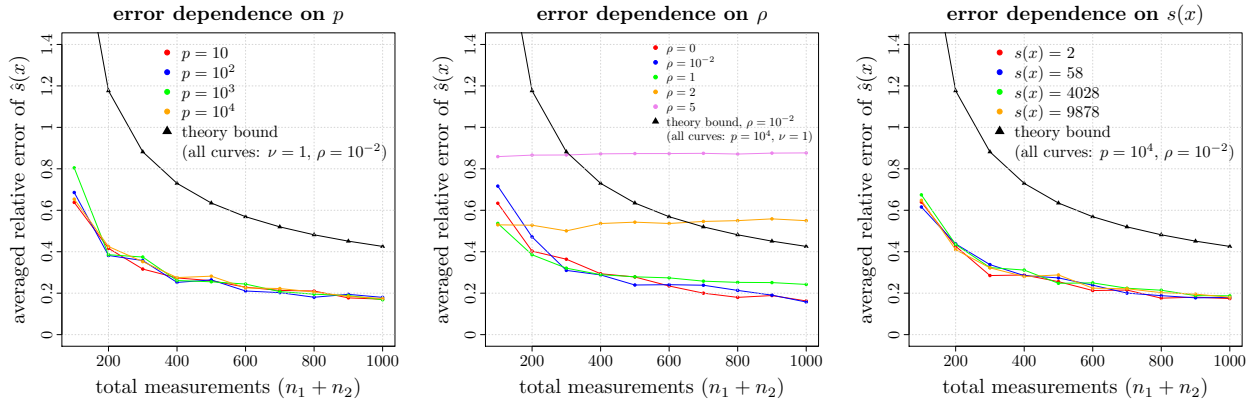

Figure 2. Performance of $\widehat{s}(x)$ as a function of p, ρ, $s(x)$, and number of measurements.

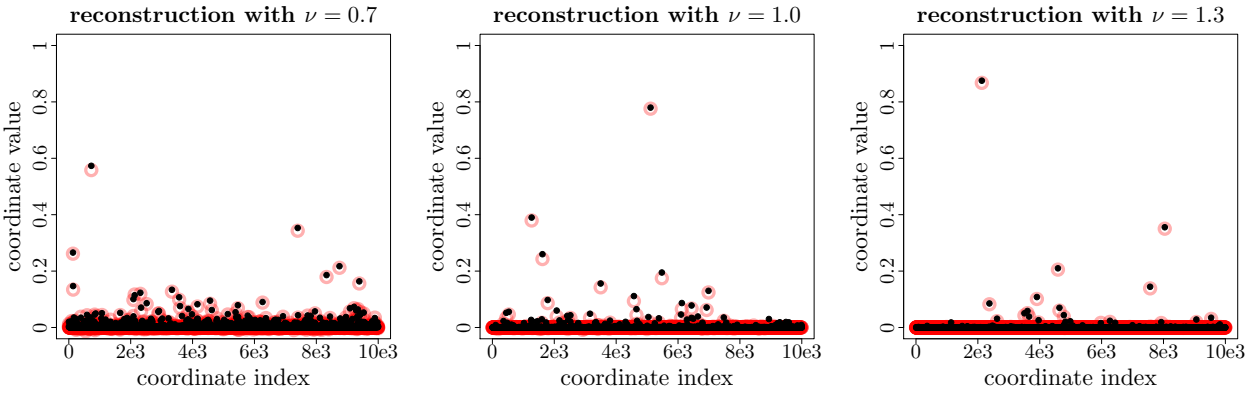

Figure 3. Signal recovery after choosing n based on $\widehat{s}(x)$. True signal x in black, and $\widehat{x}$ in red.

References

Arias-Castro, E., Candès, E. J., and Davenport, M. On the fundamental limits of adaptive sensing. *Arxiv preprint arXiv:1111.4646*, 2011.

Cai, T. T., Wang, L., and Xu, G. New bounds for restricted isometry constants. *Information Theory, IEEE Transactions on*, 56(9), 2010.

Candès, E. J. Compressive sampling. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pp. 1433–1452, 2006.

Candès, E. J. and Plan, Y. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Candès, E. J., Romberg, J.K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59 (8):1207–1223, 2006.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. The convex algebraic geometry of linear inverse problems. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pp. 699–703. IEEE, 2010.

d'Aspremont, A. and El Ghaoui, L. Testing the nullspace property using semidefinite programming. *Mathematical programming*, 127(1):123–144, 2011.

Davenport, M. A., Duarte, M. F., Eldar, Y. C., and Kutyniok, G. Introduction to compressed sensing. *Preprint*, 93, 2011.

Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Eldar, Y. C. Generalized sure for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009.

Fama, E. F. and Roll, Richard. Parameter estimates for symmetric stable distributions. *Journal of the American Statistical Association*, 66(334):331–338, 1971.

Hoyer, P.O. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

Hurley, N. and Rickard, S. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10): 4723–4741, 2009.

Indyk, P. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM (JACM)*, 53(3):307–323, 2006.

Juditsky, A. and Nemirovski, A. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Mathematical programming*, 127(1):57–88, 2011.

Li, P., Hastie, T., and Church, K. Nonlinear estimators and tail bounds for dimension reduction in l 1 using cauchy random projections. *Journal of Machine Learning Research*, pp. 2497–2532, 2007.

Lopes, M. E., Jacob, L., and Wainwright, M.J. A more powerful two-sample test in high dimensions using random projection. In *NIPS 24*, pp. 1206–1214. 2011.

Malioutov, D. M., Sanghavi, S., and Willsky, A. S. Compressed sensing with sequential observations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008.*, pp. 3357–3360. IEEE, 2008.

Rigamonti, R., Brown, M. A., and Lepetit, V. Are sparse representations really relevant for image classification? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1545–1552. IEEE, 2011.

Rudelson, M. and Vershynin, R. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21, 2007.

Shi, Q., Eriksson, A., van den Hengel, A., and Shen, C. Is face recognition really a compressive sensing problem? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 553–560. IEEE, 2011.

Tang, G. and Nehorai, A. The stability of low-rank matrix reconstruction: a constrained singular value view. *arXiv:1006.4088, submitted to IEEE Transactions on Information Theory*, 2010.

Tang, G. and Nehorai, A. Performance analysis of sparse recovery based on constrained minimal singular values. *IEEE Transactions on Signal Processing*, 59(12):5734–5745, 2011.

van den Berg, E. and Friedlander, M. P. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

van den Berg, E. and Friedlander, M. P. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008. doi: 10.1137/080714488. URL http://link.aip.org/link/?SCE/31/890.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *Arxiv preprint arxiv:1011.3027*, 2010.

Ward, R. Compressed sensing with cross validation. *IEEE Transactions on Information Theory*, 55(12):5773–5782, 2009.

Zolotarev, V. M. *One-dimensional stable distributions*, volume 65. Amer Mathematical Society, 1986.