Scaling law for recovering the sparsest element in a subspace

Citation: Demanet, L., and P. Hand. "Scaling Law for Recovering the Sparsest Element in a Subspace." Information and Inference 3, 4 (July 2014): 295–309 © 2014 The Authors

As Published: http://dx.doi.org/10.1093/IMAIAI/IAU007

Publisher: Oxford University Press (OUP)

Persistent URL: http://hdl.handle.net/1721.1/115483

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Scaling Law for Recovering the Sparsest Element in a Subspace

Laurent Demanet and Paul Hand

Massachusetts Institute of Technology, Department of Mathematics,
77 Massachusetts Avenue, Cambridge, MA 02139

October 2013, Revised May 2014

Abstract

We address the problem of recovering a sparse n-vector within a given subspace. This problem is a subtask of some approaches to dictionary learning and sparse principal component analysis. Hence, if we can prove scaling laws for recovery of sparse vectors, it will be easier to derive and prove recovery results in these applications. In this paper, we present a scaling law for recovering the sparse vector from a subspace that is spanned by the sparse vector and k random vectors. We prove that the sparse vector will be the output to one of n linear programs with high probability if its support size s satisfies $s \lesssim n/\sqrt{k \log n}$. The scaling law still holds when the desired vector is approximately sparse. To get a single estimate for the sparse vector from the n linear programs, we must select which output is the sparsest. This selection process can be based on any proxy for sparsity, and the specific proxy has the potential to improve or worsen the scaling law. If sparsity is interpreted in an ℓ_1/ℓ_∞ sense, then the scaling law can not be better than $s \lesssim n/\sqrt{k}$. Computer simulations show that selecting the sparsest output in the ℓ_1/ℓ_2 or thresholded-ℓ_0 senses can lead to a larger parameter range for successful recovery than that given by the ℓ_1/ℓ_∞ sense. sparsity, linear programming, signal recovery, sparse principal component analysis, dictionary learning.

1 Introduction

We consider the task of finding the sparsest nonzero element in a subspace. That is, given a basis for the subspace $W \subset \mathbb{R}^n$, we seek the minimizer of the problem

$$\min \|z\|_0 \text{ s.t. } z \in W, z \neq 0, \tag{1}$$

where $\|z\|_0$ is the number of nonzeros in the coefficients of z. Because problem (1) is NP-hard over an arbitrary subspace [**?**], we study a convex relaxation over particular subspace models for which we can prove recovery results.

This problem is a special case of the more general task of finding many independent sparse vectors in a subspace [**?**]. This general task has many applications, of which some will require a full basis of sparse vectors and others will require only a few sparse vectors.

1.1 Applications

Finding sparse vectors in a subspace is an important subtask in several sparse structure recovery problems. Two examples are dictionary learning and sparse principal component analysis (PCA).

- Dictionary Learning with a square, sparsely used dictionary [**?**]. Consider an $m \times n$ matrix Y, where $n > m$ and each column of Y can be written as a sparse linear combination of m unknown and independent dictionary elements. That is to say, $Y = AX$, where A is invertible, the columns of A are the dictionary elements, and the rows of X are sparse. The goal is to find A and X from only the knowledge of Y. In order to find this decomposition, we note that the row span of Y is the same as that of X. Hence, a process that finds a sparse basis of a subspace could be used to find X from the row space of Y. The dictionary A can then be directly computed. Such a process based on linear programming was used in [**?**] to study this dictionary learning problem.

- Sparse PCA in the infinite data limit. Consider the following noisy data:

$$x_i \sim N(0, \beta P_W + I), \text{ for } i = 1...m, \tag{2}$$

  where β is a signal-to-noise ratio and P_W is the projection matrix onto a subspace W. The task of PCA is to identify the directions along which there is the most variability in the data. For (2), these directions are the subspace W, which we interpret as the signal. In applications involving gene expression data, these directions may consist of the combinations of genes that distinguish different types of cancer cells [**?**], for example. In some applications, it is desirable that the discovered signal involve few variables [**?**]. Hence, sparse signal vectors are of particular interest. For example, this sparsity allows easier interpretation of the relevant genes for the cancer discrimination task. It is thus useful to find the sparse signal directions within W based on the measurements x_i, which is a sparse PCA problem [**?**, **?**, **?**]. This problem is most interesting for a small number of samples, but it is nontrivial even in the infinite data limit. In this limit, P_W is measured exactly, and sparse PCA reduces to finding the sparsest element in the subspace W.

To prove rigorous recovery results for problems like these, it is important to have scaling laws for when the sparsest element subtask succeeds.

1.2 Recovery by n linear programs

The difficulty of finding sparse elements in a subspace stems from the nonconvexity of both the ℓ_0 objective and the nonzero constraint in (1). As is standard in sparse recovery, we consider the ℓ_1 relaxation of the ℓ_0 sparsity objective. The nonzero constraint must now be replaced with some proper normalization in order to prevent outcomes arbitrarily close to zero.

A natural normalization is to search for an element with unit Euclidean length. Unfortunately, the resulting problem is nonconvex, and methods for solving it directly get stuck in local minima [**?**]. This approach may be convexified; for more details, see Section 1.7 of the present paper.

A convex normalization is to search for an element where a particular coefficient is set to unity. Such an approach should work best if we could normalize the largest coefficient to be 1. Unfortunately, we do not know which component corresponds to the largest coefficient of the sparsest element. Hence, we attempt to recover the sparse element by separately normalizing each of the n components. That is, we attempt to find the sparsest element in W by solving the collection of n linear programs

$$\min \|z\|_1 \text{ such that } z \in W, z(i) = 1, \tag{3}$$

where $1 \leq i \leq n$. This is the approach introduced by Spielman et al. in [**?**]. In order to get a single estimate from the optimizers of these n linear programs, we need a selector that returns the output that is the 'sparsest.' Here, the sparsity of a vector z may be interpreted in one many precise senses, such as the strict-ℓ_0 sense, $\|z\|_0$; a thresholded-ℓ_0 sense, $\#\{i \mid |z(i)| \geq \epsilon\}$; the ℓ_1/ℓ_∞ sense, $s = \|z\|_1/\|z\|_\infty$; or the ℓ_1/ℓ_2 sense, $\|z\|_1/\|z\|_2$. In principle, this interpretation of sparsity may alter the scaling law of the overall recovery process.

1.3 Subspace Models

Because finding the sparsest element in an arbitrary space is NP-hard, we restrict our attention to certain subspace models. Two such models are the Bernoulli-Gaussian model and a planted-random model.

- Bernoulli-Gaussian model [**?**]. Consider a collection of vectors where each entry is nonzero with a fixed probability. Suppose the values of these nonzero coefficients are given by independent normal random variables. The Bernoulli-Gaussian subspace is given by the span of these vectors. If the subspace dimension and the expected sparsity of each vector is small enough, then the sparsest basis provides exactly these vectors, up to scaling. Spielman et al. [**?**] provide a scaling law for successful recovery with the linear programs (3) under this subspace model.

- Planted-random model. Consider a sparse vector and a collection of independent Gaussian random vectors. The planted-random subspace is given by the span of all these vectors. If the sparse vector is sparse enough, then it is the unique sparsest element with probability 1, up to scaling. To the best of our knowledge, this model has not been studied in the literature.

In the present paper we will provide scaling laws for successful recovery under the planted-random subspace model. Concisely, the problem is as follows:

$$\begin{aligned} \text{Let} &: v \in \mathbb{R}^n, \tilde{v}_j \sim N(0, I_n) \text{ for } 1 \leq j \leq n \\ \text{Given} &: \{w_1, \ldots, w_{k+1}\} \text{ a basis for } W = \text{span}\{v, \tilde{v}_1, \ldots, \tilde{v}_k\} \\ \text{Find} &: v \end{aligned} \tag{4}$$

The vector v is the unique sparsest element in the space with probability 1, up to scaling, provided that $\|v\|_0 < n - k$.

1.4 Exact Recovery by a Single Program

Solving (4) by the n linear programs (3) requires that at least one of those programs recovers v. Roughly speaking, the program of form (3) that is most likely to succeed corresponds to an $i^* \in \text{argmax}_i |v(i)|$. This is because $v/v(i)$ is feasible for (3) and has the smallest ℓ_1 norm when $i = i^*$. The following theorem provides a scaling law under which v is exactly recovered by this instance of (3).

Theorem 1. *Let $k \leq n/32$. There exists a universal constant c such that for sufficiently large n,*

$$\|v\|_0 \leq c\frac{n/\sqrt{\log n}}{\sqrt{k}} \Rightarrow \frac{v}{v(i^*)} \text{ is the unique solution to (3) for } i = i^*, \tag{5}$$

with probability at least $1 - \tilde{\gamma}_1 e^{-\tilde{\gamma}_2 n/2} - ke^{-\lfloor c\sqrt{n/\log n}\rfloor} - \frac{k}{n^2}$. Here, $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are universal constants.

From the scaling law, we observe the following scaling limits on the permissible sparsity in terms of the dimensionality of the search space:

$$k \text{ on the order of } 1 \Longrightarrow \|v\|_0 \lesssim n/\sqrt{\log n} \tag{6}$$

$$k \text{ on the order of } n \Longrightarrow \|v\|_0 \lesssim \sqrt{n}/\sqrt{\log n} \tag{7}$$

That is, a search space of constant size permits the discovery of a vector whose support size is almost a constant fraction of n. Similarly, a search space of fixed and sufficiently small fraction of the ambient dimension allows recovery of a vector whose support size is almost on the the order of the square root of that dimension.

Roughly speaking, sparse recovery succeeds because the planted vector is smaller in ℓ_1 than any item in the random part of the subspace. The full minimization problem then reduces to computing the minimal value of (3) for a random subspace with no planted vector. In more precise terms, recall that $|v(i^*)| = \|v\|_\infty$ and $v/v(i^*)$ is feasible for (3) with $i = i^*$. A necessary condition for successful recovery is that the sparse vector is smaller in ℓ_1 than the minimum value attainable by the span of the random vectors:

$$\frac{\|v\|_1}{\|v\|_\infty} \leq \min \|z\|_1 \text{ such that } z \in \text{span}\{\tilde{v}_1, \ldots, \tilde{v}_k\}, z(i^*) = 1, \tag{8}$$

We will show in Section 2.1 that the right hand side of (8) scales like $n/\sqrt{k}$ when k is at most some constant fraction of n. As $\|v\|_1/\|v\|_\infty \leq \|v\|_0$ for all v, and the equality is attained for some v, we conclude that high probability recovery of arbitrary v is possible only if $\|v\|_0 \lesssim n/\sqrt{k}$.

Because of this necessary condition (8), the scaling law between n, k, and $\|v\|_0$ in Theorem 1 can not be improved, except for the logarithmic factor. The scaling could conceivably be improved in the related context where we only seek some i for which $v/v(i)$ is the unique solution to (3).

1.5 Stable Recovery by a Single Program

The linear programs (3) can also recover an approximately sparse v. That is, if v is sufficiently close to a vector that is sparse enough in an ℓ_0 sense, we expect to recover something close to v by solving (3) with $i = i^* \in \text{argmax}_i |v(i)|$. Let v_s be the best s-sparse approximation of v. The following theorem provides a scaling law under which v is approximately the output to the $i = i^*$ instance of the program (3).

Theorem 2. *Let $k \leq n/32$. There exists universal constants c, C such that for sufficiently large n, for $s = \lfloor c\frac{n/\sqrt{\log n}}{\sqrt{k}} \rfloor$, and for $i = i^*$, any minimizer $z^\#$ of (3) satisfies*

$$\left\| z^\# - \frac{v}{v(i^*)} \right\|_2 \leq C \frac{\sqrt{k \log n}}{\sqrt{n}} \frac{\|v - v_s\|_1}{\|v\|_\infty} \tag{9}$$

with probability at least $1 - \tilde{\gamma}_1 e^{-\tilde{\gamma}_2 n/2} - k e^{-\lfloor c\sqrt{n/\log n} \rfloor} - k/n^2$.

The dependence on k and n in (9) is favorable provided that $k \lesssim n/\log n$. In the case that $k \sim n$, the error bound has a mildly unfavorable constant, growing like $\sqrt{\log n}$. The $\sqrt{k/n}$ behavior of the error constant plays the roll of the $1/\sqrt{s}$ term that arises in the noisy compressed sensing problem [**?**]. The estimate (9) is slightly worse, as $\sqrt{k/n} \sim k^{1/4}/\sqrt{s}$, ignoring logarithmic factors. We believe that the k and n dependence of the error bound could be improved.

1.6 Selecting the Sparsest Output

Successful recovery of v by solving the n linear programs (3) requires both that v is the output to (3) for some i, and that v is selected as the 'sparsest' output among all n linear programs. We now comment on effects of selecting the sparsest output under several different interpretations of sparsity: ℓ_1/ℓ_∞, ℓ_1/ℓ_2, and thresholded-ℓ_0. Computer simulations in Section 3 show that the precise sense in which sparsity is interpreted can substantially affect recovery performance, especially when the planted-random subspace has small dimension. The worst empirical performance is exhibited by ℓ_1/ℓ_∞, and the best is by thresholded-ℓ_0, though this method has the drawback of requiring a threshold parameter.

1.7 Discussion

We now compare several approaches for solving the sparsest element problem (4). The main difference among these approaches is the way they relax sparsity into something tractable for optimization problems. We then compare the sparsest element problem to compressed sensing.

As the ℓ_1/ℓ_∞ ratio is a proxy for sparsity, it is natural to try to find the sparsest nonzero element in the subspace W by solving

$$\min \frac{\|z\|_1}{\|z\|_\infty} \text{ such that } z \in W, z \neq 0. \tag{10}$$

While this problem is not convex, it is the same as solving the n programs (3) and selecting the smallest output in an ℓ_1 or an ℓ_1/ℓ_∞ sense. Geometrically, this approach corresponds to replacing an ℓ_∞ constraint ball by separate hyperplanes along each of its faces.

We note that the scaling for successful recovery with (10) can be no better than than $s \lesssim n/\sqrt{k}$. To see this, observe that the right hand side of (8) provides an upper bound on the minimal value of $\|z\|_1/\|z\|_\infty$ for $z \in \text{span}\{\tilde{v}_1, \ldots, \tilde{v}_k\}$. Hence, outside the scaling $s \lesssim n/\sqrt{k}$, there would be random vectors that are sparser than v in the ℓ_1/ℓ_∞ sense. We do not recommend simply solving (10) as an approach for finding the sparsest vector in a subspace because computer simulations reveal that recovery can be improved merely by changing the sparsity selector to ℓ_1/ℓ_2 or thresholded-ℓ_0.

As the ℓ_1/ℓ_2 norm ratio is also a proxy for sparsity, it is also natural to try to solve the sparsest element problem by optimizing

$$\min \frac{\|z\|_1}{\|z\|_2} \text{ such that } z \in W, z \neq 0. \tag{11}$$

This problem is also nonconvex, though it can be convexified by a lifting procedure similar to [**?**]. After such a procedure, it becomes an $n \times n$ semidefinite program. While the procedure squares the dimensionality, it may give rise to provable recovery guarantees under a scaling law. We are unaware of any such results in the literature. In a sense, the present paper can be viewed as a simplification of this $n \times n$ matrix recovery problem into n linear programs on vectors, provided that we are willing to change the precise proxy for sparsity that we are optimizing.

A more elementary approach to solving the sparsest element problem (4) is to simply choose the largest diagonal elements of the projector matrix onto W. If the sparse element v is sparse enough, then this process will recover its support with high probability. While this method may have a similar scaling as that in Theorem 1, we anticipate that it is less robust because it capitalizes

directly on specific properties of the distribution of planted-random subspaces. As an analogy, consider the problem of identifying a planted clique in a random graph. A thresholding based algorithm [**?**] performs at the best known tractable scaling, up to logarithmic factors; however, its performance is much worse with semirandom models [**?**], where an adversary is able to modify the planted random graph within some constraints. For the sparsest element problem, we leave a detailed study of the thresholding approach to future research.

We now compare the sparsest element problem to compressed sensing. The linear programs (3) could be viewed as separate compressed sensing problems. From this perspective, the sensing matrix of each would have many rows that are orthogonal to the planted sparse vector, hence complicating an analysis based on restricted isometries. Because of this difficulty, we do not use a compressed sensing perspective for the proofs of the recovery theorems.

The sparsest element problem results in different qualitative scalings than those of compressed sensing. For example, if sparsity is minimized in the ℓ_1/ℓ_∞ sense, the scaling law can not be better than $s \lesssim n/\sqrt{k}$. This scaling depends strongly on k. From a compressed sensing perspective, one might guess that the recoverable sparsity is on the order of the number of measurements. Because the problems (3) can be written with $n-k+1$ equality constraints, this guess would provide the scaling $s \lesssim n$, which is impossible when minimizing sparsity in an ℓ_1/ℓ_∞ sense.

1.8 Notation

Let W be the planted-random subspace spanned by v and $\tilde{v}_1, \ldots, \tilde{v}_k$, as given in (4). Let the matrix $V = [v, \tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k]$ have these vectors as columns. Let $\tilde{V} = [\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k]$, which gives $V = [v, \tilde{V}]$. Let $v(i)$ be the ith component of the vector v. Let $V_{i^*,:}$ be the i^*-th row of V. Write $V_{i^*,:} = [1, \tilde{a}^t]$, where $\tilde{a} \in \mathbb{R}^k$. For a set S, write $\tilde{V}_S$ and $\tilde{V}_{S^c}$ as the restrictions of $\tilde{V}$ to the rows given by S and S^c, respectively. Write $x \asymp y$ when there exists positive c and C, which are independent of n and k, such that $cy \leq x \leq Cy$.

2 Proofs

To prove the theorems, we note that (3), (5), (9), and the value of i^* are all invariant to any rescaling of v. Without loss of generality, it suffices to take $\|v\|_\infty = 1$ and $v(i^*) = 1$. Our aim is to prove that v is or is near the solution to (3) when $i = i^*$. We begin by noting that $W = \text{range}(V)$. Hence, changing variables by $z = Vx$, (3) is equivalent to

$$\min \|Vx\|_1 \text{ such that } Vx(i^*) = 1 \tag{12}$$

when $i = i^*$. Note that $Vx(i^*)$ refers to the i^* component of Vx. We will show that $x = e_1$ is the solution to (12) in the exact case and is near the solution in the noisy case. Write $x = [x(1), \tilde{x}]$ in order to separately study the behavior of x on and away from the first coefficient. Our overall proof approach is to show that if n is larger than the given scaling, a nonzero $\tilde{x}$ gives rise to a large contribution to the ℓ_1 norm of Vx from coefficients off the support of v.

2.1 Scaling Without a Sparse Vector

In this section, we derive the scaling law for the minimal ℓ_1 norm attainable in a random subspace when one of the components is set to unity. This law is the key part of the justification that the

scaling in the theorems can not be improved except for the logarithmic factor. As per Section 1.7, it also provides the proof of the best possible scaling when minimizing sparsity in an ℓ_1/ℓ_∞ sense. We also present the derivation for pedagogical purposes, as it contains the key ideas and probabilistic tools we will use when proving the theorems. The rest of this section will prove the following lemma.

Lemma 3. *Let $\tilde{V}$ be an $n \times k$ matrix with i.i.d. $N(0,1)$ entries, where $k \leq n/16$. With high probability,*

$$\frac{n}{\sqrt{k}} \asymp \min \|\tilde{V}\tilde{x}\|_1 \text{ such that } \tilde{V}\tilde{x}(i^*) = 1. \tag{13}$$

The failure probability is exponentially small in n and k.

Proof. Because range$(\tilde{V})$ is a k-dimensional random subspace, we can appeal to the uniform equivalence of the ℓ_1 and ℓ_2 norms, as given by the following lemma.

Lemma 4. *Fix $\eta < 1$. For every y in a randomly chosen (with respect to the natural Grassmannian measure) ηn-dimensional subspace of $\mathbb{R}^n$,*

$$c_\eta \sqrt{n}\|y\|_2 \leq \|y\|_1 \leq \sqrt{n}\|y\|_2$$

with probability $1 - \gamma_1 e^{-\gamma_2 n}$ for universal constants γ_1, γ_2.

This result is well known [**?**, **?**]. Related results with different types of random subspaces can be found at [**?**, **?**, **?**, **?**]. Thus, with high probability,

$$\|\tilde{V}\tilde{x}\|_1 \asymp \sqrt{n}\|\tilde{V}\tilde{x}\|_2 \text{ for all } \tilde{x}. \tag{14}$$

We now appeal to nonasymptotic estimates of the singular values of $\tilde{V}$. Corollary 5.35 in [**?**] gives that for a matrix $A \in \mathbb{R}^{n\times k}$ with $k \leq n/16$ and i.i.d. $N(0,1)$ entries,

$$\mathbb{P}\Big(\frac{\sqrt{n}}{2} \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \frac{3\sqrt{n}}{2}\Big) \geq 1 - 2e^{-n/32}. \tag{15}$$

Thus, with high probability,

$$\|\tilde{V}\tilde{x}\|_2 \asymp \sqrt{n}\|\tilde{x}\|_2. \tag{16}$$

Combining (14) and (16), we get $\|\tilde{V}\tilde{x}\|_1 \asymp n\|\tilde{x}\|_2$ with high probability. Hence, the minimum values of the following two programs are within fixed constant multiples of each other:

$$\min \|\tilde{V}\tilde{x}\|_1 \text{ such that } \tilde{V}_{i^*,:}\tilde{x} = 1 \quad \asymp \quad \min n\|\tilde{x}\|_2 \text{ such that } \tilde{V}_{i^*,:}\tilde{x} = 1. \tag{17}$$

By the Cauchy-Schwarz inequality and concentration estimates of the length of a Gaussian vector, any feasible point in the programs (17) satisfies

$$\|\tilde{x}\|_2 \geq \frac{1}{\|\tilde{V}_{i^*,:}\|_2} \asymp \frac{1}{\sqrt{k}}, \tag{18}$$

with failure probability that decays exponentially in k. Considering the $\tilde{x}$ for which the inequality in (18) is achieved, we get that the minimal value to the right hand program in (17) scales like $n/\sqrt{k}$, proving the lemma. □

2.2 Proof of Theorem 1

The proof of Theorem 1 hinges on the following lemma. Let S be a superset of the support of v. Relative to the candidate $x = e_1$, any nonzero $\tilde{x}$ gives components on S^c that can only increase $\|Vx\|_1$. Nonzero $\tilde{x}$ can give components on S that decrease $\|Vx\|_1$. If the ℓ_1 norm of $\tilde{V}\tilde{x}$ on S^c is large enough and the ℓ_1 norm of $\tilde{V}\tilde{x}$ on S is small enough, then the minimizer to (12) must be e_1.

Lemma 5. *Let $V = [v, \tilde{V}]$ with $\|v\|_\infty = 1$, $V_{i^*,:} = [1, \tilde{a}^t]$, supp$(v) \subseteq S$, and $|S| = s$. Suppose that $\|\tilde{V}_S\tilde{x}\|_1 \leq 2s\|\tilde{x}\|_1$ and $\|\tilde{V}_{S^c}\tilde{x}\|_1 \geq (2\|\tilde{a}\|_\infty + 2)s\|\tilde{x}\|_1$ for all $\tilde{x}$. Then, e_1 is the unique solution to (12).*

Proof. For any x, observe that

$$\|Vx\|_1 = \|v\ x(1) + \tilde{V}_S\tilde{x}\|_1 + \|\tilde{V}_{S^c}\tilde{x}\|_1 \tag{19}$$

$$\geq \|v\|_1|x(1)| - 2s\|\tilde{x}\|_1 + \|\tilde{V}_{S^c}\tilde{x}\|_1 \tag{20}$$

$$\geq \|v\|_1|x(1)| + 2\|\tilde{a}\|_\infty s\|\tilde{x}\|_1 \tag{21}$$

where the first inequality is from the upper bound on $\|\tilde{V}_S\tilde{x}\|_1$ and the second inequality is from the lower bound on $\|\tilde{V}_{S^c}\tilde{x}\|_1$. Note that $x = e_1$ is feasible and has value $\|Ve_1\|_1 = \|v\|_1$. Hence, at a minimizer $\tilde{x}^\#$,

$$\|v\|_1|x^\#(1)| + 2\|\tilde{a}\|_\infty s\|\tilde{x}^\#\|_1 \leq \|v\|_1. \tag{22}$$

Using the constraint $x^\#(1) + \tilde{a}^t\tilde{x}^\# = 1$, a minimizer must satisfy

$$\|v\|_1(1 - \|\tilde{a}\|_\infty\|\tilde{x}^\#\|_1) + 2\|\tilde{a}\|_\infty s\|\tilde{x}^\#\|_1 \leq \|v\|_1. \tag{23}$$

Noting that $\|v\|_1 \leq s$, a minimizer must satisfy

$$2\|\tilde{a}\|_\infty s\|\tilde{x}^\#\|_1 \leq \|\tilde{a}\|_\infty s\|\tilde{x}^\#\|_1. \tag{24}$$

Hence, $\tilde{x}^\# = 0$. The constraint provides $x^\#(1) = 1$, which proves that e_1 is the unique solution to (12). □

To prove Theorem 1 by applying Lemma 5, we need to study the minimum value of $\|\tilde{V}_{S^c}\tilde{x}\|_1/\|\tilde{x}\|_1$ for matrices $\tilde{V}_{S^c}$ with i.i.d. $N(0,1)$ entries. Precisely, we will show the following lemma.

Lemma 6. *Let A be a $n \times k$ matrix with i.i.d. $N(0,1)$ entries, with $k \leq n/16$. There is a universal constant $\tilde{c}$, such that with high probability, $\|Ax\|_1/\|x\|_1 \geq \tilde{c}n/\sqrt{k}$ for all $x \neq 0$. This probability is at least $1 - 2e^{-n/32} - \gamma_1 e^{-\gamma_2 n}$.*

Proof of Lemma 6. We are to study the problem

$$\min \|Ax\|_1 \text{ such that } \|x\|_1 = 1, \tag{25}$$

which is equivalent to

$$\min \|Ax\|_1 \text{ such that } \|x\|_1 \geq 1. \tag{26}$$

The minimum value of (26) can be bounded from below by that of

$$\min \|Ax\|_1 \text{ such that } \|x\|_2 \geq 1/\sqrt{k}, \tag{27}$$

because the feasible set of (26) is included in the feasible set of (27). We now write both the objective and constraint in terms of Ax. To that end, we apply the lower bound in (15) to get

$$\mathbb{P}(\|x\|_2 \leq 2\frac{\|Ax\|_2}{\sqrt{n}} \text{ for all } x) \geq 1 - 2e^{-n/32}. \tag{28}$$

The feasible set of (27) is contained by the set $\{x \mid \|Ax\|_2 \geq \frac{1}{2}\sqrt{\frac{n}{k}}\}$ with high probability. Hence, a lower bound to (27) is with high probability given by

$$\min \|Ax\|_1 \text{ such that } \|Ax\|_2 \geq \frac{1}{2}\sqrt{\frac{n}{k}}. \tag{29}$$

In order to find a lower bound on (29), we apply Lemma 4 to the range of A, which is a k-dimensional random subspace of $\mathbb{R}^n$ with $k \leq n/16$. Taking $\eta = 1/16$, we see that with high probability, the minimal value of (29) is bounded from below by $\frac{c_\eta}{2}\frac{n}{\sqrt{k}}$. The minimal value of (29), and hence of (25), is bounded from below by $\tilde{c}n/\sqrt{k}$ for some universal constant $\tilde{c}$ with probability at least $1 - 2e^{-n/32} - \gamma_1 e^{-\gamma_2 n}$. □

To prove the theorem by applying Lemma 5, we also need to study the maximum value of $\|\tilde{V}_S\tilde{x}\|_1/\|\tilde{x}\|_1$ for matrices $\tilde{V}_S$ with i.i.d. $N(0,1)$ entries.

Lemma 7. *Let A be a $s \times k$ matrix with i.i.d. $N(0,1)$ entries. Then $\sup_{x\neq 0} \|Ax\|_1/\|x\|_1 \leq 2s$ with probability at least $1 - ke^{-s}$.*

Proof. Note that elementary matrix theory gives that the $\ell_1 \to \ell_1$ operator norm of A is

$$\max_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{1\leq i\leq k} \|Ae_i\|_1. \tag{30}$$

As Ae_i is an $s \times 1$ vector of i.i.d. standard normals, we have

$$\mathbb{P}(\|Ae_i\|_1 > t) \leq 2^s e^{-t^2/2s}. \tag{31}$$

Hence,

$$\mathbb{P}(\max_i \|Ae_i\|_1 > t) \leq k2^s e^{-t^2/2s}. \tag{32}$$

Taking $t = 2s$, we conclude

$$\mathbb{P}(\max_i \|Ae_i\|_1 > 2s) \leq k2^s e^{-2s} \leq ke^{-s}. \tag{33}$$

□

We can now combine Lemmas 5, 6, and 7 to prove Theorem 1.

Proof of Theorem 1. Let $\tilde{c}$ be the universal constant given by Lemma 6 and let $c = \tilde{c}/5$. We will show that for $\|v\|_0 \leq c\frac{n/\sqrt{\log n}}{\sqrt{k}}$, the minimizer to (3) is v with at least the stated probability.

Let S be any superset of $\mathrm{supp}(v)$ with cardinality $s = \lfloor c\frac{n/\sqrt{\log n}}{\sqrt{k}} \rfloor$. As per Lemma 5, e_1 is the solution to (12), and hence v is the unique solution to (3), if the following events occur simultaneously:

$$\|\tilde{V}_S \tilde{x}\|_1 \leq 2s\|\tilde{x}\|_1 \text{ for all } \tilde{x} \tag{34}$$

$$\|\tilde{a}\|_\infty \leq 2\sqrt{\log n} \tag{35}$$

$$\|\tilde{V}_{S^c} \tilde{x}\|_1 \geq 5\sqrt{\log n} s\|\tilde{x}\|_1 \text{ for all } \tilde{x} \tag{36}$$

Applying Lemma 7 to the $s \times k$ matrix $\tilde{V}_S$, we get that (34) holds with probability at least $1 - ke^{-s} \geq 1 - ke^{-\lfloor c\sqrt{n/\log n}\rfloor}$. Classical results on the maximum of a gaussian vector establishes that (35) holds with probability at least $1 - k/n^2$. Because $s \leq n/2$ and $k \leq n/32$, we have that $\tilde{V}_{S^c}$ has height at least $n/2$ and width at most $n/32$. Hence, Lemma 6 gives that $\|\tilde{V}_{S^c}\tilde{x}\|_1/\|\tilde{x}\|_1 \geq \tilde{c}n/\sqrt{k}$ for all $\tilde{x} \neq 0$ with probability at least $1 - 2e^{-n/64} - \gamma_1 e^{-\gamma_2 n/2}$. Because $s \leq \frac{\tilde{c}}{5}\frac{n/\sqrt{\log n}}{\sqrt{k}}$, we conclude (36), allowing us to apply Lemma 5. Hence, successful recovery occurs with probability at least $1 - 2e^{-n/64} - \gamma_1 e^{-\gamma_2 n/2} - ke^{-\lfloor c\sqrt{n/\log n}\rfloor} - k/n^2 \geq 1 - \tilde{\gamma}_1 e^{-\tilde{\gamma}_2 n} - ke^{-\lfloor c\sqrt{n/\log n}\rfloor} - k/n^2$, for some $\tilde{\gamma}_1, \tilde{\gamma}_2$.

□

2.3 Proof of Theorem 2

We will prove the following lemma, of which Theorem 2 is a special case.

Lemma 8. *Let $k \leq n/32$. There exists universal constants c, C such that for sufficiently large n, for all $s \leq c\frac{n/\sqrt{\log n}}{\sqrt{k}}$, and for $i = i^*$, any minimizer $z^\#$ of (3) satisfies*

$$\left\| z^\# - \frac{v}{v(i^*)} \right\|_2 \leq C\frac{\sqrt{n}}{s}\frac{\|v - v_s\|_1}{\|v\|_\infty} \tag{37}$$

with probability at least $1 - \tilde{\gamma}_1 e^{-\tilde{\gamma}_2 n} - ke^{-s} - k/n^2$.

At first glance, this lemma appears to have poor error bounds for large n and poor probabilistic guarantees for small s. On further inspection, the bounds can be improved by simply considering a larger s, possibly even larger than the size of the support of v. Larger values of s simultaneously increase the denominator and decrease the s-term approximation error in the numerator of (37). Taking the largest permissible value $s = \lfloor c\frac{n/\sqrt{\log n}}{\sqrt{k}} \rfloor$, we arrive at Theorem 2.

Lemma 8 hinges on the following analog of Lemma 5.

Lemma 9. *Fix $1 \leq s < n$ and $\alpha > 0$. Let $V = [v, \tilde{V}]$ with $\|v\|_\infty = 1$, $V_{i^*,:} = [1, \tilde{a}^t]$, $\delta = \|v - v_s\|_1$, $\mathrm{supp}(v) \subseteq S$, and $|S| = s$. If $\|\tilde{V}_S\tilde{x}\|_1 \leq 2s\|\tilde{x}\|_1$ and $\|\tilde{V}_{S^c}\tilde{x}\|_1 \geq (2\|\tilde{a}\|_\infty + 2 + \alpha)s\|\tilde{x}\|_1$ for all $\tilde{x} \in \mathbb{R}^k$, then any $x^\#$ minimizing (12) satisfies*

$$|x_1^\# - 1| \leq \frac{2\delta}{s}, \qquad \text{and} \qquad \|\tilde{x}^\#\|_1 \leq \frac{2\delta}{s(\|\tilde{a}\|_\infty + \alpha)}. \tag{38}$$

Proof. For any x, observe that

$$\|Vx\|_1 = \|v \cdot x(1) + \tilde{V}_S \tilde{x}\|_1 + \|\tilde{V}_{S^c}\tilde{x}\|_1 \tag{39}$$

$$\geq \|v\|_1 |x(1)| - 2s\|\tilde{x}\|_1 + \|\tilde{V}_{S^c}\tilde{x}\|_1 \tag{40}$$

$$\geq \|v\|_1 |x(1)| + (2\|\tilde{a}\|_\infty + \alpha)s\|\tilde{x}\|_1 \tag{41}$$

$$\geq (\|v_s\|_1 - \delta)|x(1)| + (2\|\tilde{a}\|_\infty + \alpha)s\|\tilde{x}\|_1 \tag{42}$$

where the first inequality is from the upper bound on $\|\tilde{V}_S\tilde{x}\|_1$ and the second inequality is from the lower bound on $\|\tilde{V}_{S^c}\tilde{x}\|_1$. Note that $x = e_1$ is feasible and has value $\|Ve_1\|_1 = \|v\|_1 \leq \|v_s\|_1 + \delta$. Hence, at a minimizer $\tilde{x}^\#$,

$$(\|v_s\|_1 - \delta)|x^\#(1)| + (2\|\tilde{a}\|_\infty + \alpha)s\|\tilde{x}^\#\|_1 \leq \|v_s\|_1 + \delta. \tag{43}$$

Using the constraint $x^\#(1) + \tilde{a}\tilde{x}^\# = 1$, a minimizer must satisfy

$$(\|v_s\|_1 - \delta)(1 - \|\tilde{a}\|_\infty \|\tilde{x}^\#\|_1) + (2\|\tilde{a}\|_\infty + \alpha)s\|\tilde{x}^\#\|_1 \leq \|v_s\|_1 + \delta. \tag{44}$$

Noting that $\|v_s\|_1 \leq s$, a minimizer must satisfy

$$\|\tilde{x}^\#\|_1 \leq \frac{2\delta}{(\|\tilde{a}\|_\infty + \alpha)s}. \tag{45}$$

Applying the constraint again, we get

$$|x^\#(1) - 1| \leq \frac{2\delta}{s}. \tag{46}$$

□

We now complete the proof of Theorem 2 by proving Lemma 8.

Proof of Lemma 8. Let $\tilde{c}$ be the universal constant given by Lemma 6 and let $c = \tilde{c}/6$. We will show that for any $s \leq c\frac{n/\sqrt{\log n}}{\sqrt{k}}$, the minimizer to (3) is near v with at least the stated probability.

Let S be any superset of $\mathrm{supp}(v_s)$ with cardinality s. Applying Lemma 9 with $\alpha = \sqrt{\log n}$, we observe that a minimizer $x^\#$ to (12) satisfies $|x^\#(1) - 1| \leq 2\delta/s$ and $\|\tilde{x}^\#\|_1 \leq 2\delta/(s\sqrt{\log n})$ if the following events occur simultaneously:

$$\|\tilde{V}_S\tilde{x}\|_1 \leq 2s\|\tilde{x}\|_1 \text{ for all } \tilde{x} \tag{47}$$

$$\|\tilde{a}\|_\infty \leq 2\sqrt{\log n} \tag{48}$$

$$\|\tilde{V}_{S^c}\tilde{x}\|_1 \geq 6\sqrt{\log n}s\|\tilde{x}\|_1 \text{ for all } \tilde{x} \tag{49}$$

Applying Lemma 7 to the $s \times k$ matrix $\tilde{V}_S$, we get that (47) holds with probability at least $1 - ke^{-s}$. Classical results on the maximum of a gaussian vector establishes that (48) holds with probability at least $1 - k/n^2$. Because $s \leq n/2$ and $k \leq n/32$, we have that $\tilde{V}_{S^c}$ has height at least $n/2$ and width at most $n/32$. Hence, Lemma 6 gives that $\|\tilde{V}_{S^c}\tilde{x}\|_1/\|\tilde{x}\|_1 \geq \tilde{c}n/\sqrt{k}$ for all $\tilde{x} \neq 0$ with probability at least $1 - 2e^{-n/64} - \gamma_1 e^{-\gamma_2 n/2}$. If $s \leq \frac{\tilde{c}}{6}\frac{n/\sqrt{\log n}}{\sqrt{k}}$, we conclude (49), allowing us to apply Lemma 5.

It remains to show that $Vx^\#$ is near v. Observe that

$$\begin{aligned}
\|Vx^\# - v\|_2 &= \|Vx^\# - Ve_1\|_2 && (50)\\
&\le \|v\|_2|x^\#(1) - 1| + \|\tilde{V}\tilde{x}^\#\|_2 && (51)\\
&\le \|v\|_2|x^\#(1) - 1| + \sigma_{\max}(\tilde{V})\|\tilde{x}^\#\|_2 && (52)\\
&\le \sqrt{n}|x^\#(1) - 1| + \frac{3}{2}\sqrt{n}\|\tilde{x}^\#\|_1 && (53)\\
&\le \sqrt{n}\frac{2\delta}{s} + \frac{3}{2}\sqrt{n}\frac{2\delta}{s\sqrt{\log n}} && (54)\\
&\le C\frac{\sqrt{n}}{s}\delta && (55)
\end{aligned}$$

The the third inequality uses the fact that $\|v\|_\infty = 1$ and $\sigma_{\max}(\tilde{V}) \le \frac{3}{2}\sqrt{n}$, which occurs with probability at least $1-2e^{-n/32}$ due to the upper bound in (15). Hence, approximate recovery occurs with probability at least $1-2e^{-n/64}-2e^{-n/32}-\gamma_1 e^{-\gamma_2 n/2}-ke^{-s}-k/n^2 \ge 1-\tilde{\gamma}_1 e^{-\tilde{\gamma}_2 n} - ke^{-s} - k/n^2$, for some $\tilde{\gamma_1}, \tilde{\gamma_2}$.

□

3 Simulations

We now present computer simulations that demonstrate when solving the n linear programs (3) can recover the approximately sparse vector $v \in \mathbb{R}^n$ from a planted random subspace (4). In order to obtain a single output from the n programs, we select the output that is the 'sparsest' in one of several possible senses. In this section, we study the effect of these different senses. We also study the best possible recovery performance of these senses by simulating the behavior of purely random subspaces with no planted sparse vector.

The parameters for the linear programs (3) are as follows. Let $n = 100$, $1 \le s \le n$, and $S = \{1, \ldots, s\}$. Let 1_S be the vector that is 1 on S and 0 on S^c. We attempt to recover the approximately s-sparse vector $v = 1_S + \delta u$, where $\delta = 0.01$ and u has i.i.d. Gaussian entries and is normalized such that $\|u\|_1 = 1$. Let $i^* = \text{argmax}_i\, |v(i)|$. We solve (3) for $1 \le i \le n$ using YALMIP [**?**] with the SDPT3 solver [**?**, **?**]. Among these n outputs, we let $z^\#$ be the one that (a) corresponds to $i = i^*$; (b) is smallest in the sense of ℓ_1/ℓ_∞; (c) is smallest in the sense of ℓ_1/ℓ_2; or (d) is sparsest in the sense of thresholded-ℓ_0 at the level $\epsilon = 0.01$. We will refer to (a) as the oracle selector because it corresponds to an oracle that tells us the index of the largest component of v. We call a recovery successful if $\left\|z^\# - v/v(i^*)\right\|_2 \le 0.01$. Figure 1 shows the probability of successful recovery, as computed over 50 independent trials, for many values of k and the approximate sparsity s. Near and below the phase transitions, simulations were performed for all even values of k and s. In the large regions to the top-right of the phase transitions, simulations were performed only for values of k and s that are multiples of 5. In this region, the probably of recovery was always zero.

We also ran computer simulations in the noiseless case $\delta = 0$. The outputs of these simulations are not shown because they look the same as Figure 1 for the parameters as above.

To get an indication for when (3) can successfully recover an approximately sparse vector from a planted random subspace, we also study the behavior of the corresponding problems without a planted vector. Specifically, we compute

$$\min \|z\|_1 \text{ such that } z \in \text{span}\{\tilde{v}_1, \ldots, \tilde{v}_k\}, z(i) = 1, \quad (56)$$

for i.i.d. $\tilde{v}_j \sim N(0, I_n)$ and for all $1 \leq i \leq n$. The curve in Figure 1a shows the dependence on k of the minimal value of (56) for a single fixed i, as found by the median over 50 independent trials. The curve in Figure 1b shows the solution to (56) that is smallest in the ℓ_1/ℓ_∞ sense, also found by the median of 50 trials. We note that the least ℓ_1/ℓ_∞ solution to (56) is also a solution of the problem

$$\min \frac{\|z\|_1}{\|z\|_\infty} \text{ such that } z \in \text{span}\{\tilde{v}_1, \ldots, \tilde{v}_k\}, z \neq 0. \tag{57}$$

These curves represent upper bounds for the sparsity of a recoverable signal: for sparsities of v above this curve, there will be linear combinations of random vectors that are considered sparser than v in the provided sense. Hence, recovering v would be impossible. These curves are also upper bounds in the noiseless case of $\delta = 0$. Figures 1c and 1d do not have corresponding lines because we do not have a tractable method for directly optimizing sparsity in the ℓ_1/ℓ_2 or thresholded-ℓ_0 senses.

Our primary observation is that the selection process for the sparsest output among the n programs (3) can greatly effect recovery performance, especially for planted-random subspaces with low dimensionality. Figures 1a and 1b show that v can be recovered by the $i = i^*$ program yet discarded as not sparse enough in the ℓ_1/ℓ_∞ sense. In the parameter regime we simulated, the ℓ_1/ℓ_∞ selector exhibits the worst overall recovery performance. The thresholded-ℓ_0 selector gives the best performance, though it has the drawback of having a threshold parameter.

Our second observation is that recovery can succeed even when the $i = i^*$ instance of (3) fails. That is, solving all n programs of form (3) can outperform the result of a single program, even when an oracle tells us the index of the largest coefficient. For example, compare the phase transitions for small k of Figure 1a to Figures 1c and 1d. This effect is more likely when k is small and v has many large components. To see why, note that successful recovery is expected when $\|\tilde{V}_{i,:}\|_\infty$ is small. If k is small, $\|\tilde{V}_{i,:}\|_\infty$ is more likely to be small. If there are many indices i where $v(i)$ is large, it is likely that $\|\tilde{V}_{i,:}\|_\infty$ will be small enough for successful recovery with some other value of i.

Our third observation is that purely random subspaces provide good estimates for the best recoverable sparsity with a planted random subspace, under the oracle and ℓ_1/ℓ_∞ selectors. This agreement is revealed by the similarity of the phase transition and dashed lines in Figures 1a and 1b. In the plots, the region of successful recovery on average can not exceed the dashed lines because recovery requires that the planted sparse vector be smaller in the appropriate sense than any vector in the purely random part of the subspace. Because the dashed lines are also upper bounds in the noiseless case of $\delta = 0$, the performance of the recovery process can not be significantly better than that shown in Figures 1a and 1b.

Potentially, the recovery region could be improved significantly beyond that indicated by Figure 1. For example, one could consider more than n linear programs, each with a normalization against a different random direction in $\mathbb{R}^n$. Such an approach immediately gives rise to a natural tradeoff: recovery performance may be improved at the expense of more linear programs that need to be solved. We leave this relationship for future study.

Funding

This work was supported by the National Science Foundation to P.H. and L.D.; the Alfred P. Sloan Foundation to L.D.; TOTAL S.A. to L.D.; the Air Force Office of Scientific Research to L.D.; and

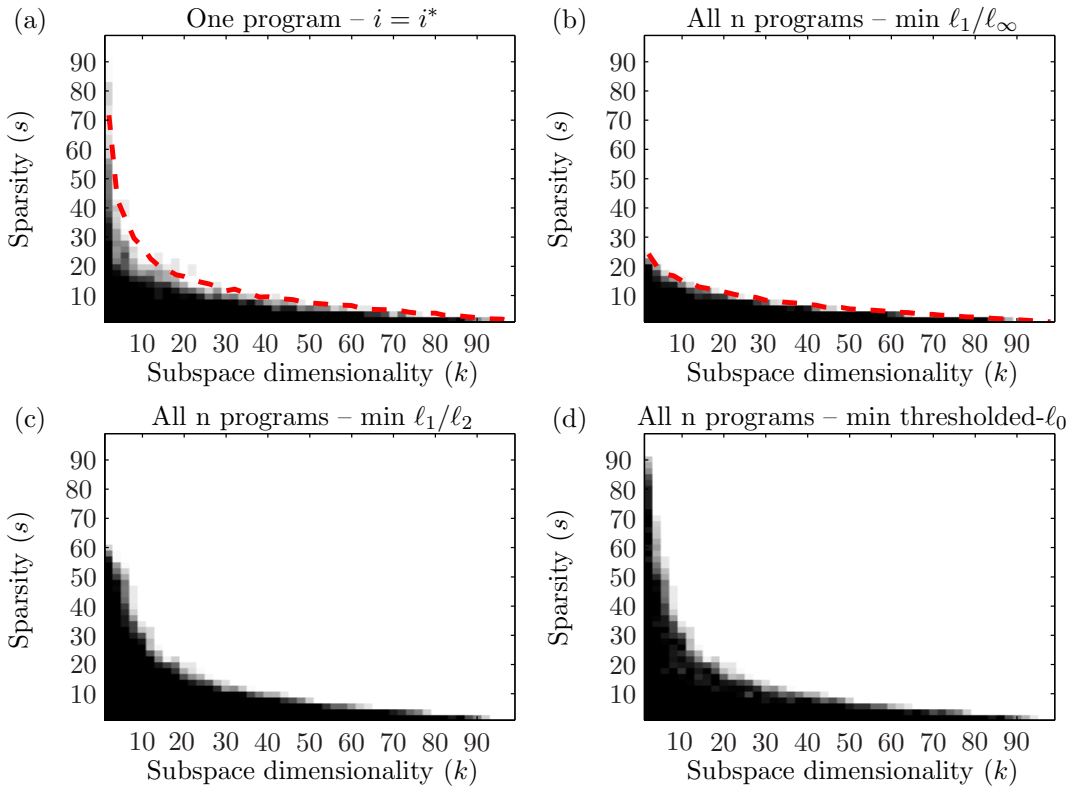

Figure 1: Empirical probability of recovery versus approximate sparsity and subspace dimensionality k. For a given s, the vector to be recovered is a noisy version of 1_S, where $|S| = s$. For all values of k and s, we solve n programs of form (3), one for each $1 \leq i \leq n$. Panel (a) shows the output corresponding to $i = i^*$. Panels (b), (c), and (d) show the output of the n programs that was smallest in the ℓ_1/ℓ_∞, ℓ_1/ℓ_2, and thresholded-ℓ_0 senses, respectively. The diagram shows the probability of recovery, as measured by 50 independent trials. White represents recovery with probability zero. Black represents recovery with probability 1. The dashed lines in panels (a) and (b) represent upper bounds for the maximal recoverable sparsity based on the behavior of random subspaces with no planted vector.

the Office of Naval Research to L.D.

Acknowledgements

The authors would like to thank Jonathan Kelner and Vladislav Voroninski for helpful discussions.