The Dantzig selector: statistical estimation when p is much larger than n

Emmanuel Candes† and Terence Tao♯

† Applied and Computational Mathematics, Caltech, Pasadena, CA 91125

♯ Department of Mathematics, University of California, Los Angeles, CA 90095

May 2005

Abstract

In many important statistical applications, the number of variables or parameters p is much larger than the number of observations n. Suppose then that we have observations $y = Ax + z$, where $x \in \mathbf{R}^p$ is a parameter vector of interest, A is a data matrix with possibly far fewer rows than columns, $n \ll p$, and the z_i's are i.i.d. $N(0, \sigma^2)$. Is it possible to estimate x reliably based on the noisy data y?

To estimate x, we introduce a new estimator—we call the *Dantzig selector*—which is solution to the ℓ_1-regularization problem

$$\min_{\tilde{x} \in \mathbf{R}^p} \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad \|A^T r\|_{\ell_\infty} \le (1 + t^{-1}) \sqrt{2 \log p} \cdot \sigma,$$

where r is the residual vector $y - A\tilde{x}$ and t is a positive scalar. We show that if A obeys a uniform uncertainty principle (with unit-normed columns) and if the true parameter vector x is sufficiently sparse (which here roughly guarantees that the model is identifiable), then with very large probability

$$\|\hat{x} - x\|_{\ell_2}^2 \le C^2 \cdot 2 \log p \cdot \left(\sigma^2 + \sum_i \min(x_i^2, \sigma^2) \right).$$

Our results are nonasymptotic and we give values for the constant C. In short, our estimator achieves a loss within a logarithmic factor of the ideal mean squared error one would achieve with an *oracle* which would supply perfect information about which coordinates are nonzero, and which were above the noise level.

In multivariate regression and from a model selection viewpoint, our result says that it is possible nearly to select the best subset of variables, by solving a very simple convex program, which in fact can easily be recast as a convenient linear program (LP).

Keywords. Statistical linear model, model selection, ideal estimation, oracle inequalities, sparse solutions to underdetermined systems, ℓ_1-minimization, linear programming, restricted orthonormality, geometry in high dimensions, random matrices.

Acknowledgments. E. C. is partially supported in part by a National Science Foundation grant DMS 01-40698 (FRG) and by an Alfred P. Sloan Fellowship. T. T. is supported in part by a grant from the Packard Foundation. E. C. thanks Rob Nowak for sending him an early preprint, Hannes Helgason for bibliographical research on this project, Justin Romberg for his help with numerical simulations, and Anestis Antoniadis for comments on an early version of the manuscript.

1 Introduction

In many important statistical applications, the number of variables or parameters p is now much larger than the number of observations n. In radiology and biomedical imaging for instance, one is typically able to collect far fewer measurements about an image of interest than the unknown number of pixels. Examples in functional MRI and tomography all come to mind. High dimensional data frequently arise in genomics. Gene expression studies are a typical example: a relatively low number of observations (in the tens) is available, while the total number of gene assayed (and considered as possible regressors) is easily in the thousands. Other examples in statistical signal processing and nonparametric estimation include the recovery of a continuous-time curve or surface from a finite number of noisy samples. Estimation in this setting is generally acknowledged as an important challenge in contemporary statistics, see the recent conference held in Leiden, The Netherlands (September 2002) "On high-dimensional data $p \gg n$ in mathematical statistics and bio-medical applications." It is believed that progress may have the potential for impact across many areas of statistics [25].

In many research fields then, scientists work with data matrices with many variables p and comparably few observations n. This paper is about this important situation, and considers the problem of estimating a parameter $x \in \mathbf{R}^p$ from the linear model

$$y = Ax + z; \tag{1.1}$$

$y \in \mathbf{R}^n$ is a vector of observations, A is an $n \times p$ predictor matrix, and z a vector of stochastic measurement errors. Unless specified otherwise, we will assume that $z \sim N(0, \sigma^2 I_n)$ is a vector of independent normal random variables, although it is clear that our methods and results may be extended to other distributions. Throughout this paper, we will of course typically assume that p is much larger than n.

1.1 Uniform uncertainty principles and the noiseless case

At first, reliably estimating x from y may seem impossible. Even in the noiseless case, one may wonder how one could possibly do this as one would need to solve an underdetermined system of equations with fewer equations than unknowns. But suppose now that x is known to be structured in the sense that it is sparse or compressible. For example, suppose that x is S-sparse so that only S of its entries are nonzero. This premise radically changes the problem, making the search for solutions feasible. In fact, [10] showed that in the noiseless case, one could actually recover x *exactly* by solving the convex program[1] ($\|\tilde{x}\|_{\ell_1} := \sum_{i=1}^p |\tilde{x}_i|$)

$$(P_1) \qquad \min_{\tilde{x} \in \mathbf{R}^p} \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad A\tilde{x} = y, \tag{1.2}$$

provided that the matrix $A \in \mathbf{R}^{n \times p}$ obeys a *uniform uncertainty principle*. That is, ℓ_1-minimization finds without error both the location and amplitudes—which we emphasize are a-priori completely unknown—of the nonzero components of the vector $x \in \mathbf{R}^p$.

To understand this result, we introduce the notion of uniform uncertainty principle (UUP) proposed in [9] and refined in [10]. This principle will play an important role throughout although we emphasize that this paper *is not* about the exact recovery of noiseless data. The

[1] (P_1) can even be recast as a linear program [14].

UUP essentially states that the $n \times p$ measurement or design matrix A obeys a "restricted isometry hypothesis." Let A_T, $T \subset \{1, \ldots, p\}$ be the $n \times |T|$ submatrix obtained by extracting the columns of A corresponding to the indices in T; then [10] defines the S-restricted isometry constant δ_S of A which is the smallest quantity such that

$$(1 - \delta_S) \|c\|_{\ell_2}^2 \leq \|A_T c\|_{\ell_2}^2 \leq (1 + \delta_S) \|c\|_{\ell_2}^2 \tag{1.3}$$

for all subsets T with $|T| \leq S$ and coefficient sequences $(c_j)_{j \in T}$. This property essentially requires that every set of columns with cardinality less than S approximately behaves like an orthonormal system. It was shown (also in [10]) that if S obeys

$$\delta_S + \delta_{2S} + \delta_{3S} < 1, \tag{1.4}$$

then solving (P_1) recovers *any* sparse signal x with support size obeying $|T| \leq S$.

Actually, [10] derived a slightly stronger result. Introduce the S, S'*-restricted orthogonality constants* $\theta_{S,S'}$ for $S + S' \leq p$ to be the smallest quantity such that

$$|\langle A_T c, A_{T'} c' \rangle| \leq \theta_{S,S'} \cdot \|c\|_{\ell_2} \|c'\|_{\ell_2} \tag{1.5}$$

holds for all *disjoint* sets $T, T' \subseteq \{1, \ldots, p\}$ of cardinality $|T| \leq S$ and $|T'| \leq S'$. Small values of restricted orthogonality constants indicate that disjoint subsets of covariates span nearly orthogonal subspaces. Then [10] showed that the recovery is exact provided

$$\delta_S + \theta_{S,S} + \theta_{S,2S} < 1, \tag{1.6}$$

which is a little better since it is not hard to see that $\delta_{S+S'} - \delta_{S'} \leq \theta_{S,S'} \leq \delta_{S+S'}$ for $S' \geq S$ [10].

1.2 Uniform uncertainty principles and statistical estimation

Any real-world sensor or measurement device is subject to at least a small amount of noise. And now one asks whether it is possible to reliably estimate the parameter $x \in \mathbf{R}^p$ from the noisy data $y \in \mathbf{R}^n$ and the model (1.1). Frankly, this may seem like an impossible task. How can one hope to estimate x, when in addition to having too few observations, these are also contaminated with noise?

To estimate x with noisy data, we consider nevertheless solving the convex program

$$(DS) \qquad \min_{\tilde{x} \in \mathbf{R}^p} \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad \|A^T r\|_{\ell_\infty} := \sup_{1 \leq i \leq p} |(A^T r)_i| \leq \lambda_p \cdot \sigma \tag{1.7}$$

for some $\lambda_p > 0$, and where r is the vector of residuals

$$r = y - A\tilde{x}. \tag{1.8}$$

In other words, we seek an estimator $\hat{x}$ with minimum complexity (as measured by the ℓ_1-norm) among all objects that are consistent with the data. The constraint on the residual vector imposes that for each $i \in \{1, \ldots, p\}$, $|(A^T r)_i| \leq \lambda_p \cdot \sigma$, and guarantees that the residuals are within the noise level. As we shall see later, this proposal makes sense provided that the columns of A have the same Euclidean size and in this paper, we will always assume they are unit-normed; our results would equally apply to matrices with different column

sizes—one would only need to change the right hand-side to $|(A^T r)_i|$ less or equal to $\lambda_p \cdot \sigma$ times the Euclidean norm of the ith column of A, or to $|(A^T r)_i| \le \sqrt{1+\delta_1} \cdot \lambda_p \cdot \sigma$ since all the columns have norm less than $\sqrt{1+\delta_1}$.

Again, the program (DS) is convex, and can easily be recast as a linear program (LP)

$$\min \sum_i u_i \quad \text{subject to} \quad -u \le \tilde{x} \le u \ \text{ and } \ -\lambda_p \sigma \mathbf{1} \le A^T(y - A\tilde{x}) \le \lambda_p \sigma \mathbf{1} \tag{1.9}$$

where the optimization variables are $u, \tilde{x} \in R^p$, and $\mathbf{1}$ is a p-dimensional vector of ones. Hence, our estimation procedure is computationally tractable, see Section 4.4 for details. There is indeed a growing family of ever more efficient algorithms for solving such problems (even for problems with tens or even hundreds of thousands of observations) [7].

We call the estimator (1.7) the *Dantzig selector*; with this name, we intend to pay tribute to the father of linear programming who passed away while we were finalizing this manuscript, and to underscore that the convex program (DS) is effectively a variable selection technique.

The first result of this paper is that the Dantzig selector is surprisingly accurate.

Theorem 1.1 *Suppose $x \in \mathbf{R}^p$ is any S-sparse vector of parameters obeying $\delta_{2S}+\theta_{S,2S} < 1$. Choose $\lambda_p = \sqrt{2 \log p}$ in (1.7). Then with large probability, $\hat{x}$ obeys*

$$\|\hat{x} - x\|_{\ell_2}^2 \le C_1^2 \cdot (2 \log p) \cdot S \cdot \sigma^2, \tag{1.10}$$

with $C_1 = 4/(1-\delta_S-\theta_{S,2S})$. Hence, for small values of $\delta_S+\theta_{S,2S}$, $C_1 \approx 4$. For concreteness, if one chooses $\lambda_p := \sqrt{2(1+\beta)\log p}$ for each $\beta \ge 0$, the bound holds with probability exceeding $1 - (\sqrt{\pi \log p} \cdot p^{\beta})^{-1}$ with the proviso that λ_p^2 substitutes $2 \log p$ in (1.10).

We will discuss the condition $\delta_{2S} + \theta_{S,2S} < 1$ later but for the moment, observe that (1.10) describes a striking phenomenon: not only are we able to reliably estimate the vector of parameters from limited observations, but the mean squared error is simply proportional—up to a logarithmic factor—to the true number of unknowns times the noise level σ^2. What is of interest here is that one can achieve this feat by solving a simple linear program. Moreover, and ignoring the log-like factor, statistical common sense tells us that (1.10) is, in general, unimprovable.

To see why this is true, suppose one had available an oracle letting us know in advance, the location of the S nonzero entries of the parameter vector, i.e. $T_0 := \{i : x_i \neq 0\}$. That is, in the language of model selection, one would know the right model ahead of time. Then one could use this valuable information and construct an ideal estimator $x^\star$ by using the least-squares projection

$$x^\star_{T_0} = (A^T_{T_0} A_{T_0})^{-1} A^T_{T_0} y,$$

where $x^\star_{T_0}$ is the restriction of $x^\star$ to the set T_0, and set $x^\star$ to zero outside of T_0. (At times, we will abuse notations and also let x_I be the truncated vector equal to x_i for $i \in I$ and zero otherwise). Clearly

$$x^\star = x + (A^T_{T_0} A_{T_0})^{-1} A^T_{T_0} z,$$

and

$$\mathbf{E}\|x^\star - x\|_{\ell_2}^2 = \mathbf{E}\|(A^T_{T_0} A_{T_0})^{-1} A^T_{T_0} z\|_{\ell_2}^2 = \sigma^2 \mathrm{Tr}((A^T_{T_0} A_{T_0})^{-1}).$$

Now since all the eigenvalues of $A_{T_0}^T A_{T_0}$ belong to the interval $[1-\delta_S, 1+\delta_S]$, the ideal expected mean squared error would obey

$$\mathbf{E}\|x^\star - x\|_{\ell_2}^2 \geq \frac{1}{1+\delta_S} \cdot S \cdot \sigma^2.$$

Hence, Theorem 1.1 says that the minimum ℓ_1 estimator achieves a loss within a logarithmic factor of the ideal mean squared error; the logarithmic factor is the price we pay for adaptivity, i.e. for not knowing ahead of time where the nonzero parameter values actually are.

In short, the recovery procedure, although extremely nonlinear, is *stable in the presence of noise*. This is especially interesting because the matrix A in (1.1) is rectangular; it has many more columns than rows. As such, most of its singular values are zero. In solving (DS), we are essentially trying to invert the action of A on our hidden x in the presence of noise. The fact that this matrix inversion process keeps the perturbation from"blowing up"—even though it is severely ill-posed—is perhaps unexpected.

Presumably, our result would be especially interesting if one could estimate of the order of n parameters with as few as n observations. That is, we would like the condition $\delta_{2S} + \theta_{S,2S} < 1$ to hold for very large values of S, e.g. as close as possible to n. Now, this paper is part of a larger body of work [8–10] which shows that for "generic" or random design matrices A, the condition holds for very significant values of S. Suppose for instance that A is a random matrix with i.i.d. Gaussian entries. Then with overwhelming probability, the condition holds for $S = O(n/\log(p/n))$. In other words, this setup only requires $O(\log(p/n))$ observations per nonzero parameter value; e.g. when n is a nonnegligible fraction of p, one only needs a handful of observations per nonzero coefficient. In practice, this number is quite small as fewer than 5 or 6 observations per unknown generally suffice (over a large range of the ratio p/n), see Section 4. Many design matrices have a similar behavior and Section 3 discusses a few of these.

As an aside, it is interesting to note that for S obeying the condition of the theorem, the reconstruction from noiseless data ($\sigma = 0$) is exact and that our condition is slightly better than (1.6).

1.3 Oracle inequalities

Theorem 1.1 is certainly remarkable but there are instances, however, in which it may still be a little naive. Suppose for example that x is very small so that x is well below the noise level, i.e. $|x_i| \ll \sigma$ for all i. Then with this information, we could set $\hat{x} = 0$, and the squared error loss would then simply be $\sum_{i=1}^p |x_i|^2$ which may potentially be much smaller than σ^2 times the number of nonzero coordinates of x. In some sense, this is a situation in which the squared bias is much smaller than the variance.

A more ambitious proposal might then ask for a near-optimal trade-off *coordinate by coordinate*. To explain this idea, suppose for simplicity that A is the identity matrix so that $y \sim N(x, \sigma^2 I_p)$. Suppose then that one had available an *oracle* letting us know ahead of time, which coordinates of x are significant, i.e. the set of indices for which $|x_i| > \sigma$. Then equipped with this oracle, we would set $x_i^* = y_i$ for each index in the significant set, and

$x_i^* = 0$ otherwise. The expected mean squared-error of this ideal estimator is then

$$\mathbf{E}\|x^* - x\|_{\ell_2}^2 = \sum_{i=1}^{p} \min(x_i^2, \sigma^2). \tag{1.11}$$

Here and below, we will refer to (1.11) as the ideal MSE. As is well-known, thresholding rules with threshold level at about $\sqrt{2\log p} \cdot \sigma$ achieve the ideal MSE to within a multiplicative factor proportional to $\log p$ [17, 18].

In the context of the linear model, we might think about the ideal estimation as follows: consider the least-squares estimator $\hat{x}_I = (A_I^T A_I)^{-1} A_I^T y$ as before and consider the ideal least-squares estimator $x^\star$ which minimizes the expected mean squared error

$$x^\star = \underset{I \subset \{1,\ldots,p\}}{\operatorname{argmin}} \ \mathbf{E}\|x - \hat{x}_I\|_{\ell_2}^2.$$

In other words, one would fit all least squares model and rely on an oracle to tell us which model to choose. This is *ideal* because we can of course not evaluate $\mathbf{E}\|x - \hat{x}_I\|_{\ell_2}^2$ since we do not know x (we are trying to estimate it after all). But we can view this as a benchmark and ask whether any real estimator would obey

$$\|\hat{x} - x\|_{\ell_2}^2 = O(\log p) \cdot \mathbf{E}\|x - x^\star\|_{\ell_2}^2. \tag{1.12}$$

with large probability.

In some sense, (1.11) is a proxy for the ideal risk $\mathbf{E}\|x - x^\star\|_{\ell_2}^2$. Indeed, let I be a fixed subset of indices and consider regressing y onto this subset (we again denote by x_I the restriction of x to the set I). The error of this estimator is given by

$$\|\hat{x}_I - x\|_{\ell_2}^2 = \|\hat{x}_I - x_I\|_{\ell_2}^2 + \|x_I - x\|_{\ell_2}^2.$$

The first term is equal to

$$\hat{x}_I - x_I = (A_I^T A_I)^{-1} A_I^T A x_{I^c} + (A_I^T A_I)^{-1} A_I^T z.$$

and its expected mean squared error is given by the formula

$$\mathbf{E}\|\hat{x}_I - x_I\|^2 = \|(A_I^T A_I)^{-1} A_I^T A x_{I^c}\|_{\ell_2}^2 + \sigma^2 \mathrm{Tr}((A_I^T A_I)^{-1}).$$

Thus this term obeys

$$\mathbf{E}\|\hat{x}_I - x_I\|^2 \geq \frac{1}{1 + \delta_{|I|}} \cdot |I| \cdot \sigma^2$$

for the same reasons as before. In short, for all sets I, $|I| \leq S$, with $\delta_S < 1$, say,

$$\mathbf{E}\|x - x_I\|^2 \geq \frac{1}{2} \cdot \left(\sum_{i \in I^c} x_i^2 + |I| \cdot \sigma^2 \right),$$

which gives that the ideal mean squared error is bounded below by

$$\mathbf{E}\|x - x^*\|_{\ell_2}^2 \geq \frac{1}{2} \cdot \min_I \left(\sum_{i \in I^c} x_i^2 + |I| \cdot \sigma^2 \right) = \frac{1}{2} \cdot \sum_i \min(x_i^2, \sigma^2)$$

In that sense, the ideal risk is lower bounded by the proxy (1.12). As we have seen, the proxy is meaningful since it has a natural interpretation in terms of the ideal squared bias and variance:

$$\sum_i \min(x_i^2, \sigma^2) = \min_{I \subset \{1,\dots,p\}} \|x - x_I\|_{\ell_2}^2 + |I| \cdot \sigma^2.$$

This raises a fundamental question: given data y and the linear model (1.1), not knowing anything about the significant coordinates of x, not being able to observe directly the parameter values, can we design an estimator which nearly achieves (1.11)? Our main result is that the Dantzig selector (1.7) does... just that.

Theorem 1.2 *Choose $t > 0$ and set $\lambda_p := (1+t^{-1})\sqrt{2\log p}$ in (1.7). Then if x is S-sparse with $\delta_{2S} + \theta_{S,2S} < 1 - t$, our estimator obeys*

$$\|\hat{x} - x\|_{\ell_2}^2 \le C_2^2 \cdot \lambda_p^2 \cdot \left(\sigma^2 + \sum_{i=1}^{p} \min(x_i^2, \sigma^2)\right), \tag{1.13}$$

with large probability (the probability is as before for $\lambda_p := (\sqrt{1+\beta} + t^{-1})\sqrt{2\log p}$). Here, C_2 may only depend on δ_{2S} and $\theta_{S,2S}$, see below.

We emphasize that (1.13) is nonasymptotic and our analysis actually yields explicit constants. For instance, we also prove that

$$C_2 = 2\frac{C_0}{1-\delta-\theta} + 2\frac{\theta(1+\delta)}{(1-\delta-\theta)^2} + \frac{1+\delta}{1-\delta-\theta}$$

and

$$C_0 := 2\sqrt{2}(1 + \frac{1-\delta^2}{1-\delta-\theta}) + (1+1/\sqrt{2})\frac{(1+\delta)^2}{1-\delta-\theta}, \tag{1.14}$$

where above and below, we put $\delta := \delta_{2S}$ and $\theta := \theta_{S,2S}$ for convenience. For δ and θ small, C_2 is close to

$$C_2 \approx 2(4\sqrt{2} + 1 + 1/\sqrt{2}) + 1 \le 16.$$

The condition imposing $\delta_{2S} + \theta_{S,2S} < 1$ (or less than $1-t$) has a rather natural interpretation in terms of model identifiability. Consider a rank deficient submatrix $A_{T_1 \cup T_2}$ with $2S$ columns (lowest eigenvalue is $0 = 1 - \delta_{2S}$), and with indices in T_1 and T_2, each of size S. Then there is a vector h obeying $Ah = 0$ and which can be decomposed as $h = x_1 - x_2$ where x_1 is supported on T_1 and likewise for x_2; that is,

$$Ax_1 = Ax_2.$$

In short, this says that the model is not identifiable since both x_1 and x_2 are S-sparse. In other words, we need $\delta_{2S} < 1$. The requirement $\delta_{2S} + \theta_{S,2S} < 1$ (or less than $1-t$) is only slightly stronger than the identifiability condition, roughly two times stronger. It puts a lower bound on the singular values of submatrices and in effect, prevents situations where multicollinearity between competitive subsets of predictors could occur.

1.4 Ideal model selection by linear programming

Our estimation procedure is of course an implicit method for choosing a desirable subset of predictors, based on the noisy data $y = Ax + z$, from among all subsets of variables. As the reader will see, there is nothing in our arguments that requires p to be larger than n and thus, the Dantzig selector can be construed as a very general variable selection strategy—hence the name.

There is of course a huge literature on model selection, and many procedures motivated by a wide array of criteria have been proposed over the years—among which [1, 6, 23, 26, 31]. By and large, the most commonly discussed approach—the "canonical selection procedure" according to [23]—is defined as

$$\underset{\tilde{x}\in\mathbf{R}^p}{\operatorname{argmin}} \|y - A\tilde{x}\|_{\ell_2}^2 + \Lambda \cdot \sigma^2 \cdot \|\tilde{x}\|_{\ell_0}, \qquad \|\tilde{x}\|_{\ell_0} := |\{i : \tilde{x}_i \neq 0\}|, \tag{1.15}$$

which best trades-off between the goodness of fit and the complexity of the model, the so-called bias and variance terms. Popular selection procedures such as AIC, C_p, BIC and RIC are all of this form with different values of the parameter Λ, see also [3–6] for related proposals. To cut a long story short, model selection is an important area in parts because of the thousands of people routinely fitting large linear models or designing statistical experiments. As such, it has and still receives a lot of attention, and progress in this field is likely to have a large impact. Now despite the size of the current literature, we believe there are two major problems in this field.

- First, finding the minimum of (1.15) is in general NP-hard [27]. To the best of our knowledge, solving this problem essentially require exhaustive searches over all subsets of columns of A, a procedure which clearly is combinatorial in nature and has exponential complexity since for p of size about n, there are about 2^p such subsets. (We are of course aware that in a few exceptional circumstances, e.g. when A is an orthonormal matrix, the solution is computationally feasible and given by thresholding rules [6, 18].)

  In other words, solving the model selection problem might be possible only when p ranges in the few dozens. This is especially problematic especially when one considers that we now live in a "data-driven" era marked by ever larger datasets.

- Second, only few quantitative results are available, essentially corresponding to restricted situations where either $A^T A$ is diagonal or close to being diagonal. Important and beautiful results here include [2, 3, 5, 6, 19, 23].

In recent years, researchers have of course developed alternatives to overcome these computational difficulties, and we would like to single out the popular lasso also known as Basis Pursuit [14, 33] which relaxes the counting norm $\|\tilde{x}\|_{\ell_0}$ into the convex ℓ_1-norm $\|\tilde{x}\|_{\ell_1}$. Not much is known about the performance of such strategies although they seem to work well in practice, e.g. see [30].

Against this background, our work clearly marks a significant departure from the current literature, both in terms of what it achieves and of its methods. Indeed, our paper introduces a method for selecting variables based on linear programming, and obtains decisive quantitative results in fairly general settings.

1.5 Extension to nearly sparse parameters

We considered thus far the estimation of sparse parameter vectors, i.e. with a number S of nonzero entries obeying $\delta_{2S} + \theta_{S,2S}$. We already explained that this condition is in some sense necessary as otherwise one might have an "aliasing" problem, a situation in which $Ax \approx Ax'$ although x and x' might be completely different. However, extensions of our results to nonsparse objects are possible provided that one imposes other types of constraints to remove the possibility of strong aliasing.

Many such constraints may exist and we consider one of them which imposes some decay condition on the entries of x. Rearrange the entries of x by decreasing order of magnitude $|x_{(1)}| \geq |x_{(2)}| \geq \ldots \geq |x_{(p)}|$ and suppose the kth largest entry obeys

$$|x_{(k)}| \leq R \cdot k^{-1/s}, \tag{1.16}$$

for some positive R and $s \leq 1$, say. Can we show that our estimator achieves an error close to the proxy (1.11)? The first observation is that to mimic this proxy, we need to be able to estimate reliably all the coordinates which are significantly above the noise level, i.e. roughly such that $|x_i| \geq \sigma$. Let $S = |\{i : |x_i| > \sigma\}|$. Then if $\delta_{2S} + \theta_{S,2S} < 1$, this might be possible but otherwise, we may simply not have enough observations to estimate that many coefficients. The second observation is that for $x \in \mathbf{R}^p$ obeying (1.16)

$$\sum_i \min(x_i^2, \sigma^2) = S \cdot \sigma^2 + \sum_{i \geq S+1} |x_{(i)}|^2 \leq C \cdot (S \cdot \sigma^2 + R^2 S^{-2r}) \tag{1.17}$$

with $r = 1/s - 1/2$. With this in mind, we have the following result.

Theorem 1.3 *Suppose $x \in \mathbf{R}^p$ obeys* (1.16) *and let S_* be fixed such that $\delta_{2S_*} + \theta_{S_*,2S_*} < 1$. Choose λ_p as in Theorem 1.1. Then $\hat{x}$ obeys*

$$\|\hat{x} - x\|_{\ell_2}^2 \leq \min_{1 \leq S \leq S^*} C_3 \cdot 2 \log p \cdot (S \cdot \sigma^2 + R^2 S^{-2r}) \tag{1.18}$$

with large probability.

Note that for each x obeying (1.16), $|\{i : |x_i| > \sigma\}| \leq (R/\sigma)^{1/s}$. Then if $S_* \geq (R/\sigma)^{1/s}$, it is not hard to see that (1.18) becomes

$$\|\hat{x} - x\|_{\ell_2}^2 \leq O(\log p) \cdot R^{\frac{2}{2r+1}} \cdot (\sigma^2)^{\frac{2r}{2r+1}} \tag{1.19}$$

which is the well-known minimax rate for classes of objects exhibiting the decay (1.16). Even though we have $n \ll p$, the Dantzig selector recovers the minimax rate that one would get if we were able to measure *all* the coordinates of x *directly* via $\tilde{y} \sim N(x, \sigma^2 I_p)$. In the case where $S_* \leq (R/\sigma)^{1/s}$, the method saturates because we do not have enough data to recover the minimax rate, and can only guarantee a squared loss of about $O(\log p)(R^2 S_*^{-2r} + S_* \cdot \sigma^2)$. Note, however, that the error is well-controlled.

1.6 Variations

When A is an orthogonal matrix, the Dantzig selector $\hat{x}$ is then the ℓ_1-minimizer subject to the constraint $\|A^T y - \hat{x}\|_{\ell_\infty} \leq \lambda_p \cdot \sigma$. This implies that $\hat{x}$ is simply the soft-thresholded version of $A^T y$ at level $\lambda_p \cdot \sigma$, thus

$$\hat{x}_i = \max(|(A^T y)_i| - \lambda_p \cdot \sigma, 0) \mathrm{sgn}((A^T y)_i).$$

In other words, $A^T y$ is shifted toward the origin. In Section 4, we will see that for arbitrary A's, the method continues to exhibit a soft-thresholding type of behavior and as a result, may slightly underestimate the true value of the nonzero parameters.

There are several simple methods which can correct for correct for this bias, and increase performance in practical settings. We consider one of these based on a two-stage procedure.

1. Estimate $I = \{i : x_i \neq 0\}$ with $\hat{I} = \{i : \hat{x}_i \neq 0\}$ with x as in (1.7) (or more generally with $\hat{I} = \{i : |\hat{x}_i| > \alpha \cdot \sigma\}$ for some $\alpha \geq 0$).
2. Construct the estimator
$$\hat{x}_{\hat{I}} = (A_{\hat{I}}^T A_{\hat{I}})^{-1} A_{\hat{I}}^T y, \tag{1.20}$$
and set the other coordinates to zero.

Hence, we rely on the Dantzig selector to estimate the model I, and constructs a new estimator by regressing the data y onto the model $\hat{I}$. We will refer to this variation as the Gauss-Dantzig selector. As we will see in Section 4, this recenters the estimate and generally yields higher statistical accuracy. We anticipate that all our theorems hold with some such variations.

1.7 Organization of the paper

The paper is organized as follows. We prove our main results, namely, Theorems 1.1 and 1.2 in Section 2. We discuss the implications of this work for experimental design in Section 3. and for model selection in Section 5.1. Section 4 introduces numerical experiments showing that our approach is effective in practical applications. Finally, Section 5.2 closes the paper with a short summary of our findings and of their consequences for model selection, and with a discussion of other related works.

2 Proof of Theorems

We introduce some notations that we will use throughout this section. We let $v_1, \ldots, v_p \in \mathbf{R}^n$ be the p columns of A (the exploratory variables) so that $Ax = x_1 v_1 + \ldots + x_p v_p$ and $(A^T y)_j = \langle y, v_j \rangle$, $1 \leq j \leq p$. We recall that the columns of A are normalized to have unit-norm, i.e. $\|v_j\|_{\ell_2} = 1$.

Note that it is sufficient to prove our theorems with $\sigma = 1$ as the general case would follow from a simple rescaling argument. Therefore, we assume $\sigma = 1$ from now on. Now a key observation is that with large probability, $z \sim N(0, I_n)$ obeys the *orthogonality condition*

$$|\langle z, v_j \rangle| \leq \lambda_p \text{ for all } 1 \leq j \leq p, \tag{2.1}$$

for $\lambda_p = \sqrt{2 \log p}$. This is standard and simply follows from the fact that for each j, $X_j := \langle z, v_j \rangle \sim N(0, 1)$. We will see that if (2.1) holds, then (1.10) holds. Note that for each $u > 0$, $\mathbf{P}(\sup_j |X_j| > u) \leq p \cdot \phi(u)/u$ where $\phi(u) := (2\pi)^{-1/2} e^{-u^2/2}$, and our quantitative probabilistic statement just follows from this bound. Better bounds are possible but we will not pursue these refinements here. As remarked earlier, if the columns were not unit normed, one would obtain the same conclusion with $\lambda_p = \sqrt{1 + \delta_1} \cdot \sqrt{2 \log p}$ since $\|v_j\|_{\ell_2} \leq \sqrt{1 + \delta_1}$.

2.1 High-dimensional geometry

It is probably best to start by introducing intuitive arguments underlying Theorems 1.1 and 1.2. These ideas are very geometrical and we hope they will convey the gist of the proof.

Consider Theorem 1.1 first, and suppose that $y = Ax + z$, where z obeys the orthogonality condition (2.1) for some λ_p. Let $\hat{x}$ be the minimizer of (1.7). Clearly, the true vector of parameters x is feasible and hence

$$\|\hat{x}\|_{\ell_1} \le \|x\|_{\ell_1}.$$

Decompose $\hat{x}$ as $\hat{x} = x + h$ and let T_0 be the support of x, $T_0 = \{i : x_i \neq 0\}$. Then h obeys two geometric constraints:

1. First, as essentially observed in [20],

$$\|x\|_{\ell_1} - \|h_{T_0}\|_{\ell_1} + \|h_{T_0^c}\|_{\ell_1} \le \|x + h\|_{\ell_1} \le \|x\|_{\ell_1},$$

   where $h_{T_0}(t) = h(t)$ for $t \in T_0$ and zero elsewhere (similarly for $h_{T_0^c}$). Hence, h obeys the cone constraint

$$\|h_{T_0^c}\|_{\ell_1} \le \|h_{T_0}\|_{\ell_1}. \tag{2.2}$$

2. Second, since

$$\langle z - r, v_j \rangle = \langle A\hat{x} - Ax, v_j \rangle = \langle Ah, v_j \rangle,$$

   it follows from the triangle inequality that

$$\|A^T Ah\|_{\ell_\infty} \le 2\lambda_p. \tag{2.3}$$

We will see that these two geometrical constraints imply that h is small in the ℓ_2-norm. In other words, we will show that

$$\sup_{h \in \mathbf{R}^p} \|h\|_{\ell_2}^2 \quad \text{subject to } \|h_{T_0^c}\|_{\ell_1} \le \|h_{T_0}\|_{\ell_1} \text{ and } \|A^T Ah\|_{\ell_\infty} \le 2\lambda_p \tag{2.4}$$

obeys the desired bound, i.e. $O(\lambda_p^2 \cdot |T_0|)$. This is illustrated in Figure 1 (a). Hence, our statistical question is deeply connected with the geometry of high-dimensional Banach spaces, and of high-dimensional spaces in general.

To think about the general case, consider the set of indices $T_0 := \{i : |x_i| > \sigma\}$ and let x_{T_0} be the vector equal to x on T_0 and zero outside, $x = x_{T_0} + x_{T_0^c}$. Suppose now that x_{T_0} were feasible, then we would have $\|\hat{x}\|_{\ell_1} \le \|x_{T_0}\|_{\ell_1}$; writing $x = x_{T_0} + h$, the same analysis as that of Theorem 1.1—and outlined above—would give

$$\|\hat{x} - x_{T_0}\|_{\ell_2}^2 = O(\log p) \cdot |T_0| \cdot \sigma^2.$$

From $\|\hat{x} - x\|_{\ell_2}^2 \le 2\|\hat{x} - x_{T_0}\|_{\ell_2}^2 + 2\|x - x_{T_0}\|_{\ell_2}^2$, one would get

$$\|\hat{x} - x_{T_0}\|_{\ell_2}^2 = O(\log p) \cdot |T_0| \cdot \sigma^2 + 2 \sum_{i:|x_i|<\sigma} x_i^2$$

which is the content of (1.13). Unfortunately, while x_{T_0} may be feasible for "most" S-sparse vectors x, it is not for some, and the argument is considerably more involved.

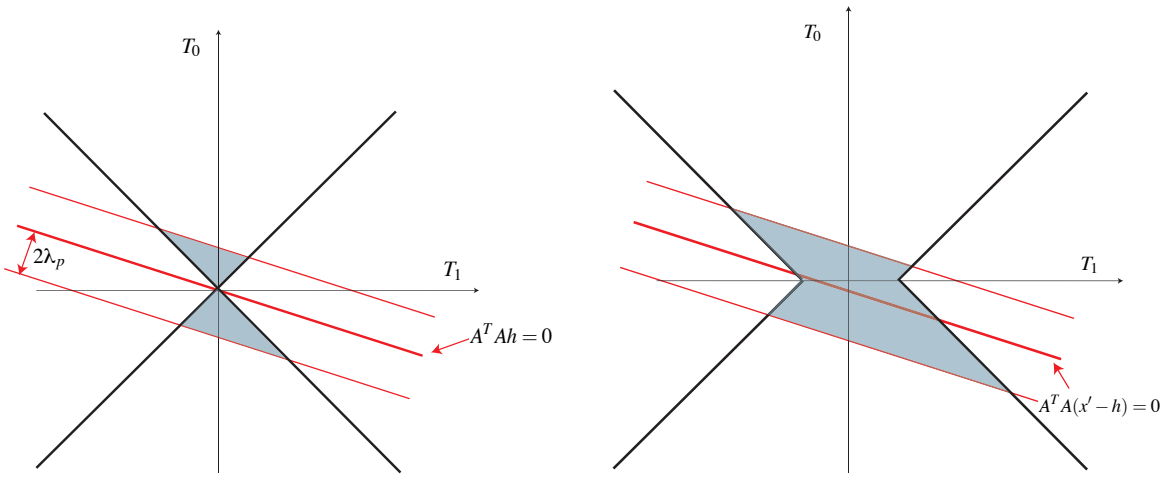


Figure 1: *This figure represents the geometry of the constraints. On the left, the shaded area represents the set of h obeying both (2.2) (hourglass region) and (2.3) (slab region). The right figure represents the situation in the more general case.*

2.2 Proof of Theorem 1.1

Lemma 2.1 *Suppose T_0 is a set of cardinality S with $\delta + \theta < 1$. For a vector $h \in \mathbf{R}^p$, we let T_1 be the S largest positions of h outside of T_0. Put $T_{01} = T_0 \cup T_1$. Then*

$$\|h\|_{\ell_2(T_{01})} \le \frac{1}{1-\delta}\|A^T_{T_{01}} Ah\|_{\ell_2} + \frac{\theta}{(1-\delta)S^{1/2}}\|h\|_{\ell_1(T_0^c)}$$

and

$$\|h\|^2_{\ell_2} \le \|h\|^2_{\ell_2(T_{01})} + S^{-1}\|h\|^2_{\ell_1(T_0^c)}.$$

Proof Consider the restricted transformation $A_{T_{01}} : \mathbf{R}^{T_{01}} \to \mathbf{R}^n$, $A_{T_{01}} c := \sum_{j \in T_{01}} c_j v_j$. Let $V \subset \mathbf{R}^n$ be the span of $\{v_j : j \in T_{01}\}$. Then V is of course the range of $A_{T_{01}}$ and also the orthogonal complement of the kernel of $A^T_{T_{01}}$ which says that $\mathbf{R}^n$ is the orthogonal sum $V \oplus V^\perp$. Because $\delta < 1$, we know that the operator $A_{T_{01}}$ is a bijection from $\mathbf{R}^{T_{01}}$ to V, with singular values between $\sqrt{1-\delta}$ and $\sqrt{1+\delta}$. As a consequence, for any $c \in \ell_2(T_{01})$, we have

$$\sqrt{1-\delta}\,\|c\|_{\ell_2} \le \|A_{T_{01}} c\|_{\ell_2} \le \sqrt{1+\delta}\,\|c\|_{\ell_2}.$$

Moreover, letting P_V denote the orthogonal projection onto V, we have for each $w \in \mathbf{R}^n$, $A^T_{T_{01}} w = A^T_{T_{01}} P_V w$ and it follows that

$$\sqrt{1-\delta}\,\|P_V w\|_{\ell_2} \le \|A^T_{T_{01}} w\|_{\ell_2} \le \sqrt{1+\delta}\,\|P_V w\|_{\ell_2}. \tag{2.5}$$

We apply this to $w := Ah$ and conclude in particular that

$$\|P_V Ah\|_{\ell_2} \le (1-\delta)^{-1/2}\|A^T_{T_{01}} Ah\|_{\ell_2}. \tag{2.6}$$

The next step is to derive a lower bound on $P_V Ah$. To do this, we begin by dividing T_0^c into subsets of size S and enumerate T_0^c as $n_1, n_2, \ldots, n_{m-|T_0|}$ in decreasing order of magnitude

of $h_{T_0^c}$. Set $T_j = \{n_\ell, (j-1)S+1 \le \ell \le jS\}$. That is, T_1 is as before and contains the indices of the S largest coefficients of $h_{T_0^c}$, T_2 contains the indices of the next S largest coefficients, and so on.

Decompose now $P_V Ah$ as

$$P_V Ah = P_V Ah_{T_{01}} + \sum_{j\ge 2} P_V Ah_{T_j}. \tag{2.7}$$

By definition, $Ah_{T_{01}} \in V$ and $P_V Ah_{T_{01}} = Ah_{T_{01}}$. Further, since P_V is an orthogonal projection onto the span of the v_j's for $j \in T_{01}$, $P_V Ah_{T_j} = \sum_{j\in T_{01}} c_j v_j$ for some coefficients c_j, and the following identity holds

$$\|P_V Ah_{T_j}\|_{\ell_2}^2 = \langle P_V Ah_{T_j}, Ah_{T_j}\rangle. \tag{2.8}$$

By restricted orthogonality followed by restricted isometry, this gives

$$\begin{aligned}\langle P_V Ah_{T_j}, Ah_{T_j}\rangle &\le \theta\,(\sum_{j\in T_{01}} |c_j|^2)^{1/2}\,\|h_{T_j}\|_{\ell_2} \\ &\le \frac{\theta}{\sqrt{1-\delta}}\,\|P_V Ah_{T_j}\|_{\ell_2}\,\|h_{T_j}\|_{\ell_2}\end{aligned}$$

which upon combining with (2.8) gives

$$\|P_V Ah_{T_j}\|_{\ell_2} \le \frac{\theta}{\sqrt{1-\delta}}\,\|h_{T_j}\|_{\ell_2}. \tag{2.9}$$

We then develop an upper bound on $\sum_{j\ge 2}\|h_{T_j}\|_{\ell_2}$ as in [13]. By construction, the magnitude of each coefficient in T_{j+1} is less than the average of the magnitudes in T_j:

$$|h_{T_{j+1}}(t)| \;\le \|h_{T_j}\|_{\ell_1}/S.$$

Then $\|h_{T_{j+1}}\|_{\ell_2}^2 \;\le\; \|h_{T_j}\|_{\ell_1}^2/S$ and, therefore,

$$\sum_{j\ge 2}\|h_{T_j}\|_{\ell_2} \;\le\; S^{-1/2}\sum_{j\ge 1}\|h_{T_j}\|_{\ell_1} = S^{-1/2}\|h\|_{\ell_1(T_0^c)}. \tag{2.10}$$

To summarize, $Ah_{T_{01}}$ obeys $\|Ah_{T_{01}}\|_{\ell_2} \ge \sqrt{1-\delta}\,\|h_{T_{01}}\|_{\ell_2}$ by restricted isometry, and since $\sum_{j\ge 2}\|P_V Ah_{T_j}\|_{\ell_2} \le \theta(1-\delta)^{-1/2}S^{-1/2}\|h\|_{\ell_1(T_0^c)}$,

$$\|P_V Ah\|_{\ell_2} \ge \sqrt{1-\delta}\|h\|_{\ell_2(T_{01})} - \frac{\theta}{\sqrt{1-\delta}}\,S^{-1/2}\,\|h\|_{\ell_1(T_0^c)}.$$

Combining this with (2.6) proves the first part of the lemma.

It remains to argue about the second part. Observe that the kth largest value of $h_{T_0^c}$ obeys

$$|h_{T_0^c}|_{(k)} \;\le\; \|h_{T_0^c}\|_{\ell_1}/k$$

and, therefore,

$$\|h_{T_{01}^c}\|_{\ell_2}^2 \;\le\; \|h_{T_0^c}\|_{\ell_1}^2 \sum_{k\ge S+1} 1/k^2 \;\le\; \|h_{T_0^c}\|_{\ell_1}^2/S,$$

which is what we needed to establish. The lemma is proven. ∎

Theorem 1.1 is now an easy consequence of this lemma. Observe that on the one hand, (2.2) gives

$$\|h_{T_0^c}\|_{\ell_1} \le \|h_{T_0}\|_{\ell_1} \le S^{1/2} \|h_{T_0}\|_{\ell_2}$$

while on the other hand, (2.3) gives

$$\|A^T_{T_{01}} Ah\|_{\ell_2} \le (2S)^{1/2} \cdot 2\lambda_p$$

since each of the $2S$ coefficients of $A^*_{T_{01}} Ah$ is at most $2\lambda_p$ (2.3). In conclusion, we apply Lemma 2.1 and obtain

$$\|h\|_{\ell_2(T_{01})} \le \frac{1}{1-\delta-\theta} \cdot \sqrt{2S} \cdot 2\lambda_p.$$

The theorem follows since $\|h\|^2_{\ell_2} \le 2\|h\|^2_{\ell_2(T_{01})}$.

2.3 Proof of Theorem 1.3

The argument underlying Theorem 1.3 is almost the same as that of Theorem 1.1. We let T_0 be the set of the S largest entries of x, and write $\hat{x} = x + h$ as before. If z obeys the orthogonality condition (1.5), x is feasible and

$$\|x_{T_0}\|_{\ell_1} - \|h_{T_0}\|_{\ell_1} + \|h_{T_0^c}\|_{\ell_1} - \|x_{T_0^c}\|_{\ell_1} \le \|x+h\|_{\ell_1} \le \|x\|_{\ell_1},$$

which gives

$$\|h_{T_0^c}\|_{\ell_1} \le \|h_{T_0}\|_{\ell_1} + 2\|x_{T_0^c}\|_{\ell_1}.$$

The presence of the extra-term is the only difference in the argument. We then conclude with Lemma 2.1 and (2.3) that

$$\|h_{T_0^c}\|_{\ell_2} \le \frac{C}{1-\delta-\theta} \cdot (\lambda_p \cdot S^{1/2} + \|x_{T_0^c}\|_{\ell_1} \cdot S^{-1/2}).$$

The second part of Lemma 2.1 gives $\|h\|_{\ell_2} \le 2\|h\|_{\ell_2(T_{01})} + \|x_{T_0^c}\|_{\ell_1} \cdot S^{-1/2}$. Since for x obeying the decay condition (1.16), $\|x_{T_0^c}\|_{\ell_1} \cdot S^{-1/2} \le C \cdot R \cdot S^{-r}$, with $r = 1/s - 1/2$, we established that for all $S \le S_*$.

$$\|h_{T_0^c}\|_{\ell_2} \le \frac{C}{1-\delta_{S_*}-\theta_{\theta_*,2S_*}} \cdot (\lambda_p \cdot S^{1/2} + R \cdot S^{-r}).$$

The theorem follows.

2.4 Proof of Theorem 1.2

We begin with an auxiliary lemma.

Lemma 2.2 *For any vector x, we have*

$$\|Ax\|_{\ell_2} \le \sqrt{1+\delta}(\|x\|_{\ell_2} + (2S)^{-1/2}\|x\|_{\ell_1}).$$

Proof Let T_1 be the $2S$ largest positions of x, then T_2 be the next largest, and so forth. Then

$$\|Ax\|_{\ell_2} \leq \|Ax_{T_1}\|_{\ell_2} + \sum_{j \geq 2} \|Ax_{T_j}\|_{\ell_2}.$$

From restricted isometry we have

$$\|Ax_{T_1}\|_{\ell_2} \leq (1+\delta)^{1/2} \|x_{T_1}\|_{\ell_2} \leq (1+\delta)^{1/2} \|x\|_{\ell_2}.$$

and

$$\|Ax_{T_j}\|_{\ell_2} \leq (1+\delta)^{1/2} \|x_{T_j}\|_{\ell_2} \leq (1+\delta)^{1/2} (2S)^{-1/2} \|x_{T_{j-1}}\|_{\ell_1}.$$

The claim follows. ∎

We now turn to the proof of Theorem 1.2. As usual we let $\hat{x}$ be the ℓ_1 minimizer subject to the constraints

$$\|A^T(A\hat{x} - y)\|_{\ell_\infty} = \sup_{1 \leq j \leq p} |\langle A\hat{x} - y, v_j \rangle| \leq (1 + t^{-1})\lambda, \tag{2.11}$$

where $\lambda := \sqrt{2 \log p}$ for short.

Without loss of generality we may order the x_j's in decreasing order of magnitude

$$|x_1| \geq |x_2| \geq \ldots \geq |x_p|. \tag{2.12}$$

In particular, by the sparsity assumption on x we know that

$$x_j = 0 \text{ for all } j > S. \tag{2.13}$$

In particular we see that

$$\sum_j \min(x_j^2, \lambda^2) \leq S \cdot \lambda^2.$$

Let S_0 be the smallest integer such that

$$\sum_j \min(x_j^2, \lambda^2) \leq S_0 \cdot \lambda^2, \tag{2.14}$$

thus $0 \leq S_0 \leq S$ and

$$S_0 \cdot \lambda^2 \leq \lambda^2 + \sum_j \min(x_j^2, \lambda^2). \tag{2.15}$$

Also, observe from (2.12) that

$$S_0 \cdot \lambda^2 \geq \sum_{j=1}^{S_0+1} \min(x_j^2, \lambda^2) \geq (S_0 + 1) \min(x_{S_0+1}^2, \lambda^2)$$

and hence $\min(x_{S_0+1}^2, \lambda^2)$ is strictly less than λ^2. By (2.12) we conclude that

$$x_j < \lambda \text{ for all } j > S_0. \tag{2.16}$$

Write $x = x^{(1)} + x^{(2)}$ where

$$x_j^{(1)} = x_j \cdot 1_{1 \leq j \leq S_0}$$
$$x_j^{(2)} = x_j \cdot 1_{j > S_0}.$$

Thus $x^{(1)}$ is a hard-thresholded version of x, localized to the set

$$T_0 := \{1, \ldots, S_0\}.$$

By (2.16), $x^{(2)}$ is S-sparse with

$$\|x^{(2)}\|_{\ell_2}^2 = \sum_{j>S_0} \min(x_j^2, \lambda^2) \leq S_0 \cdot \lambda^2.$$

As we shall see in the next section, Corollary 2.5 allows the decomposition $x^{(2)} = x' + x''$ where

$$\|x'\|_{\ell_2} \leq \frac{1+\delta}{1-\delta-\theta}\,\lambda \cdot S_0^{1/2} \tag{2.17}$$

and

$$\|x'\|_{\ell_1} \leq \frac{1+\delta}{1-\delta-\theta}\,\lambda \cdot S_0 \tag{2.18}$$

and

$$\|A^T A x''\|_{\ell_\infty} < \frac{1-\delta^2}{1-\delta-\theta}\,\lambda. \tag{2.19}$$

We use this decomposition and observe that

$$A^T(A(x^{(1)} + x') - y) = -A^T A x'' - A^T z$$

and hence by (2.1), (2.19)

$$\|A^T(A(x^{(1)} + x') - y)\|_{\ell_\infty} \leq (1 + \frac{1-\delta^2}{1-\delta-\theta})\lambda. \tag{2.20}$$

By assumption $(1-\delta-\theta)^{-1} \leq t^{-1}$ and, therefore, $x^{(1)} + x'$ is feasible which in turn implies

$$\|\hat{x}\|_{\ell_1} \leq \|x^{(1)} + x'\|_{\ell_1} \leq \|x^{(1)}\|_{\ell_1} + \frac{(1+\delta)}{1-\delta-\theta} S_0 \cdot \lambda.$$

Put $\hat{x} = x^{(1)} + h$. Then $\|\hat{x}\|_{\ell_1} \geq \|x^{(1)}\|_{\ell_1} - \|h\|_{\ell_1(T_0)} + \|h\|_{\ell_1(T_0^c)}$ so that

$$\|h\|_{\ell_1(T_0^c)} \leq \|h\|_{\ell_1(T_0)} + \frac{1+\delta}{1-\delta-\theta} S_0 \cdot \lambda, \tag{2.21}$$

and from (2.11) and (2.20), we conclude that

$$\|A^T A(x' - h)\|_{\ell_\infty} \leq 2(1 + \frac{1-\delta^2}{1-\delta-\theta})\lambda. \tag{2.22}$$

Figure 1 (b) schematically illustrates both these constraints.

The rest of the proof is essentially as that of Theorem 1.1. By Lemma 2.1 we have

$$\|h_{01}\|_{\ell_2} \leq \frac{1}{1-\delta}\|A_{T_{01}}^T A h\|_{\ell_2} + \frac{\theta}{(1-\delta)S_0^{1/2}}\|h\|_{\ell_1(T_0^c)}$$

On the other hand, from (2.22) we have

$$\|A_{T_{01}}^T A(x' - h)\|_{\ell_2} \leq 2\sqrt{2}(1 + \frac{1-\delta^2}{1-\delta-\theta})S_0^{1/2} \cdot \lambda$$

while from Lemma 2.2 and (2.18), (2.17) we have

$$\|Ax'\|_{\ell_2} \leq (1+1/\sqrt{2})\frac{(1+\delta)^{3/2}}{1-\delta-\theta} S_0^{1/2} \cdot \lambda$$

and hence by restricted isometry

$$\|A_{T_{01}}^T Ax'\|_{\ell_2} \leq (1+1/\sqrt{2})\frac{(1+\delta)^2}{1-\delta-\theta}\, S_0^{1/2} \cdot \lambda.$$

In short,

$$\|A_{T_{01}}^T Ah\|_{\ell_2} \leq C_0 \cdot S_0^{1/2} \cdot \lambda$$

where C_0 was defined in (1.14). We conclude that

$$\|h_{01}\|_{\ell_2} \leq \frac{C_0}{1-\delta}\, S_0^{1/2} \cdot \lambda + \frac{\theta}{(1-\delta)S_0^{1/2}}\, \|h\|_{\ell_1(T_0^c)}.$$

Finally, the bound (2.21) gives

$$\|h\|_{\ell_1(T_0^c)} \leq S_0^{1/2}\, \|h_{01}\|_{\ell_2} + \frac{1+\delta}{1-\delta-\theta}\, S_0 \cdot \lambda$$

and hence

$$\|h_{01}\|_{\ell_2} \leq C_0' \cdot S_0^{1/2} \cdot \lambda$$

where

$$C_0' := \frac{C_0}{1-\delta-\theta} + \frac{\theta(1+\delta)}{(1-\delta-\theta)^2}.$$

Applying the second part of Lemma 2.1 and (2.21), we conclude

$$\|h\|_{\ell_2} \leq 2\|h_{01}\|_{\ell_2} + \frac{1+\delta}{1-\delta-\theta} S_0^{1/2} \cdot \lambda \leq C_2 \cdot S_0^{1/2} \cdot \lambda.$$

and the claim follows from (2.15).

2.5 Refinements

The constant C_2 obtained by this argument is not best possible; it is possible to lower it further but at the cost of making the arguments slightly more complicated. For instance, in Lemma 2.2 one can exploit the approximate orthogonality between Ax_{T_1} and the Ax_{T_j}'s to improve over the triangle inequality. Also, instead of defining $T_1, T_2, \ldots$ to have cardinality S, one can instead choose these sets to have cardinality ρS for some parameter ρ to optimize in later. We will not pursue these refinements here. However, we observe that in the limiting case $\delta = \theta = 0$, then A is an orthogonal matrix, and as we have seen earlier $\hat{x}_j = \max(|(A^T y)_j| - \lambda, 0)\mathrm{sgn}((A^T y)_j)$. In this case one easily verifies that

$$\|x - \hat{x}\|_{\ell^2}^2 \leq \sum_{i=1}^p \min(x_i^2, 4\lambda^2).$$

This would correspond, roughly speaking, to a value of $C_2 = 2$ in Theorem 1.2, and therefore shows that there is room to improve C_2 by a factor of roughly eight.

2.6 Dual sparse reconstructions

This section justifies the construction of a pseudo-hard thresholded vector which obeys the constraints, see (2.17), (2.18), and (2.19) in the proof of Theorem 1.2.

Lemma 2.3 (Dual sparse reconstruction, ℓ_2 version) *Let S s.t. $\delta+\theta<1$, and let c_T be supported on T for some $|T| \leq 2S$, then there exists x supported on T, and an exceptional set E disjoint from T with*

$$|E| \leq S, \tag{2.23}$$

such that

$$\langle Ax, v_j \rangle = c_j \text{ for all } j \in T \tag{2.24}$$

and

$$|\langle Ax, v_j \rangle| \leq \frac{\theta}{(1-\delta)\sqrt{S}} \|c_T\|_{\ell_2} \text{ for all } j \notin (T \cup E) \tag{2.25}$$

and

$$(\sum_{j \in E} |\langle Ax, v_j \rangle|^2)^{1/2} \leq \frac{\theta}{1-\delta} \|c_T\|_{\ell_2}. \tag{2.26}$$

Also we have

$$\|x\|_{\ell_2} \leq \frac{1}{1-\delta} \|c_T\|_{\ell_2} \tag{2.27}$$

and

$$\|x\|_{\ell_1} \leq \frac{S^{1/2}}{1-\delta} \|c_T\|_{\ell_2} \tag{2.28}$$

Proof We let

$$x := (A_T^T A_T)^{-1} c$$

which gives (2.24), (2.27), (2.28) by Cauchy-Schwarz. We then set

$$E := \{j \notin T : |\langle Ax, v_j \rangle| > \frac{\theta}{(1-\delta)\sqrt{S}} \|c_T\|_{\ell_2}\},$$

so (2.25) holds.

Now if T' is disjoint from T with $|T'| \leq S$ and $d_{T'}$ is supported on T', then

$$|\langle A_{T'}^T A_T x, d_{T'} \rangle| \leq \theta \|x\|_{\ell_2} \|d_{T'}\|_{\ell_2}$$

and hence by duality

$$\|A^T A_T x\|_{\ell_2(T')} \leq \theta \|x\|_{\ell_2} \leq \frac{\theta}{1-\delta} \|c_T\|_{\ell_2}. \tag{2.29}$$

If $|E| \geq S$, then we can find $T' \subset E$ with $|T'| = S$. But then we have

$$\|A^T A_T x\|_{\ell_2(T')} > \frac{\theta}{(1-\delta)S^{1/2}} \|c_T\|_{\ell_2} |T'|^{1/2}$$

which contradicts (2.29). Thus we have (2.23). Now we can apply (2.29) with $T' := E$ to obtain (2.26). ∎

Corollary 2.4 (Dual sparse reconstruction, ℓ_∞ version) *Let c_T be supported on T for some $|T| \leq S$, then there exists x obeying (2.24) such that*

$$|\langle Ax, v_j \rangle| \leq \frac{\theta}{(1-\delta-\theta)\sqrt{S}} \|c_T\|_{\ell_2} \text{ for all } j \notin T. \tag{2.30}$$

Furthermore we have

$$\|x\|_{\ell_2} \leq \frac{1}{1-\delta-\theta} \|c_T\|_{\ell_2} \tag{2.31}$$

and

$$\|x\|_{\ell_1} \leq \frac{S^{1/2}}{1-\delta-\theta} \|c_T\|_{\ell_2} \tag{2.32}$$

Proof The proof of this lemma operates by iterating the preceding lemma as in Lemma 2.2 of [10]. We simply rehearse the main ingredients and refer the reader to [10] for details.

We may normalize $\sum_{j \in T} |c_j|^2 = 1$. Write $T_0 := T$. Using Lemma 2.3, we can find a vector x_1 and a set $T_1 \subseteq \{1, \ldots, p\}$ such that

$$\begin{aligned}
T_0 \cap T_1 &= \emptyset \\
|T_1| &\leq S \\
\langle Ax_1, v_j \rangle &= c_j \text{ for all } j \in T_0 \\
|\langle Ax_1, v_j \rangle| &\leq \frac{\theta}{(1-\delta)S^{1/2}} \text{ for all } j \notin T_0 \cup T_1 \\
(\sum_{j \in T_1} |\langle Ax_1, v_j \rangle|^2)^{1/2} &\leq \frac{\theta}{1-\delta} \\
\|x_1\|_{\ell_2} &\leq \frac{1}{1-\delta} \\
\|x_1\|_{\ell_1} &\leq \frac{S^{1/2}}{1-\delta}
\end{aligned}$$

Applying Lemma 2.3 iteratively gives a sequence of vectors $x_{n+1} \in \mathbf{R}^p$ and sets $T_{n+1} \subseteq \{1, \ldots, p\}$ for all $n \geq 1$ with the properties

$$\begin{aligned}
T_n \cap (T_0 \cup T_{n+1}) &= \emptyset \\
|T_{n+1}| &\leq S \\
\langle Ax_{n+1}, v_j \rangle &= \langle Ax_n, v_j \rangle \text{ for all } j \in T_n \\
\langle Ax_{n+1}, v_j \rangle &= 0 \text{ for all } j \in T_0 \\
|\langle Ax_{n+1}, v_j \rangle| &\leq \frac{\theta}{(1-\delta)S^{1/2}} \left(\frac{\theta}{1-\delta} \right)^n \forall j \notin T_0 \cup T_n \cup T_{n+1} \\
(\sum_{j \in T_{n+1}} |\langle Ax_{n+1}, v_j \rangle|^2)^{1/2} &\leq \frac{\theta}{1-\delta} \left(\frac{\theta}{1-\delta} \right)^n \\
\|x_{n+1}\|_{\ell_2} &\leq \frac{1}{1-\delta} \left(\frac{\theta}{1-\delta} \right)^{n-1} \\
\|x_{n+1}\|_{\ell_1} &\leq \frac{S^{1/2}}{1-\delta} \left(\frac{\theta}{1-\delta} \right)^{n-1}.
\end{aligned}$$

By hypothesis, we have $\frac{\theta}{1-\delta} \leq 1$. Thus if we set

$$x := \sum_{n=1}^{\infty} (-1)^{n-1} x_n$$

then the series is absolutely convergent and, therefore, x is a well-defined vector. And it turns out that x obeys the desired properties, see Lemma 2.2 in [10]. ∎

Corollary 2.5 (Constrained thresholding) *Let x be S-sparse such that*

$$\|x\|_{\ell_2} < \lambda \cdot S^{1/2}$$

for some $\lambda > 0$. Then there exists a decomposition $x = x' + x''$ such that

$$\|x'\|_{\ell_2} \leq \frac{1+\delta}{1-\delta-\theta} \|x\|_{\ell_2}$$

and

$$\|x'\|_{\ell_1} \leq \frac{1+\delta}{1-\delta-\theta} \frac{\|x\|_{\ell_2}^2}{\lambda}$$

and

$$\|A^T A x''\|_{\ell_\infty} < \frac{1-\delta^2}{1-\delta-\theta} \lambda.$$

Proof Let

$$T := \{j : |\langle Ax, v_j \rangle| \geq (1+\delta)\lambda\}.$$

Suppose that $|T| \geq S$, then we can find a subset T' of T with $|T'| = S$. Then by restricted isometry we have

$$(1+\delta)^2 \lambda^2 S \leq \sum_{j \in T'} |\langle Ax, v_j \rangle|^2 \leq (1+\delta) \|Ax\|_{\ell_2}^2 \leq (1+\delta)^2 \|x\|_{\ell_2}^2,$$

contradicting the hypothesis. Thus $|T| < S$. Applying restricted isometry again, we conclude

$$(1+\delta)^2 \lambda^2 |T| \leq \sum_{j \in T} |\langle Ax, v_j \rangle|^2 \leq (1+\delta)^2 \|x\|_{\ell_2}^2$$

and hence

$$|T| \leq S := \frac{\|x\|_{\ell_2}^2}{\lambda^2}.$$

Applying Corollary 2.4 with $c_j := \langle Ax, v_j \rangle$, we can find an x' such that

$$\langle Ax', v_j \rangle = \langle Ax, v_j \rangle \text{ for all } j \in T$$

and

$$\|x'\|_{\ell_2} \leq \frac{1+\delta}{1-\delta-\theta} \|x\|_{\ell_2}$$

and

$$\|x'\|_{\ell_1} \leq \frac{(1+\delta)\sqrt{S}}{1-\delta-\theta} \|x\|_{\ell_2} = \frac{(1+\delta)}{1-\delta-\theta} \frac{\|x\|_{\ell_2}^2}{\lambda}$$

and

$$|\langle Ax', v_j\rangle| \le \frac{\theta(1+\delta)}{(1-\delta-\theta)\sqrt{S}}\|x\|_{\ell_2} \text{ for all } j \notin T.$$

By definition of S we thus have

$$|\langle Ax', v_j\rangle| \le \frac{\theta}{1-\delta-\theta}(1+\delta)\lambda \text{ for all } j \notin T.$$

Meanwhile, by definition of T we have

$$|\langle Ax, v_j\rangle| < (1+\delta)\lambda \text{ for all } j \notin T.$$

Setting $x'' := x - x'$, the claims follow. ∎

3 Significance for Experimental Design

Our method might be of interest to anyone seeking to measure or sense a sparse high-dimensional vector using as few measurements as possible. In the noiseless case, our earlier results showed that if x is S-sparse, then it can be reconstructed *exactly* from n measurements $y = Ax$ provided that $\delta + \theta < 1$ [8, 10]. These were later extended to include wider classes of objects, i.e. the so called *compressible* objects. Now, any real-world sensor or measuring device is subject to at least a small amount of noise. Our results indicate that the Dantzig selector is robust, making it ideal for practical applications.

3.1 Random matrices and designs

An interesting aspect of this theory is that random matrices A are in some sense ideal for recovering an object from a few projections. For example, if A is a properly normalized Gaussian matrix with i.i.d. entries, then the conditions of our theorems hold with

$$S \asymp n/\log(p/n) \tag{3.1}$$

with overwhelming probability [9, 10, 21, 32]. The same relation is also conjectured to be true for other types of random matrices such as normalized binary arrays with i.i.d. entries taking values ± 1 with probability $1/2$. Other interesting strategies for recovering a sparse signal in the time domain might be to sense a comparably small number of its Fourier coefficients. In fact, [9] shows that in this case, our main condition holds with

$$S \asymp n/\log^6 p,$$

for nearly all subsets of observed coefficients of size n. Vershynin has informed us that $S \asymp n/\log^4 p$ also holds, and we believe that $S \asymp n/\log p$ is also true. More generally, suppose that A is obtained by randomly sampling n rows of a p by p orthonormal matrix (and renormalizing the columns so that they are unit-normed). Then we can take

$$S \asymp n/[\mu^2 \log^4 p]$$

with μ the coherence $\mu := \sup_{ij} \sqrt{n}|A_{ij}|$ [9].

Of course, all these calculations have implications for random designs. For example, suppose that in an *idealized application*, one could—in a first experiment—observe x directly and measure $y^{(1)} \sim N(x, \sigma^2 I_p)$. Consider a second experiment where one measures instead $y \sim N(Ax, \sigma^2 I_n)$ where A is a renormalized random design matrix with i.i.d. entries taking values ± 1 with probability 1/2. Suppose that the signal is S-sparse (note that $\|Ax\|_{\ell_2} \asymp \|x\|_{\ell_2}$). Then reversing (3.1) we see that with about

$$n \asymp S \cdot \log(p/S)$$

observations, one would get just about the same mean squared error than that one would achieve by measuring *all* the coordinates of x directly (and apply thresholding).

Such procedures are not foreign to statisticians. Combining parameters by random design or otherwise goes back a long way, see for example the long history of blood pooling strategies. The theoretical analysis needs of course to be validated with numerical simulations, which may give further insights about the practical behavior of our methods. Section 4 presents a first series of experiments to complement our study.

3.2 Applications

The ability to recover a sparse or nearly sparse parameter vector from a few observations raises tantalizing opportunities and we mention just a few to give concrete ideas.

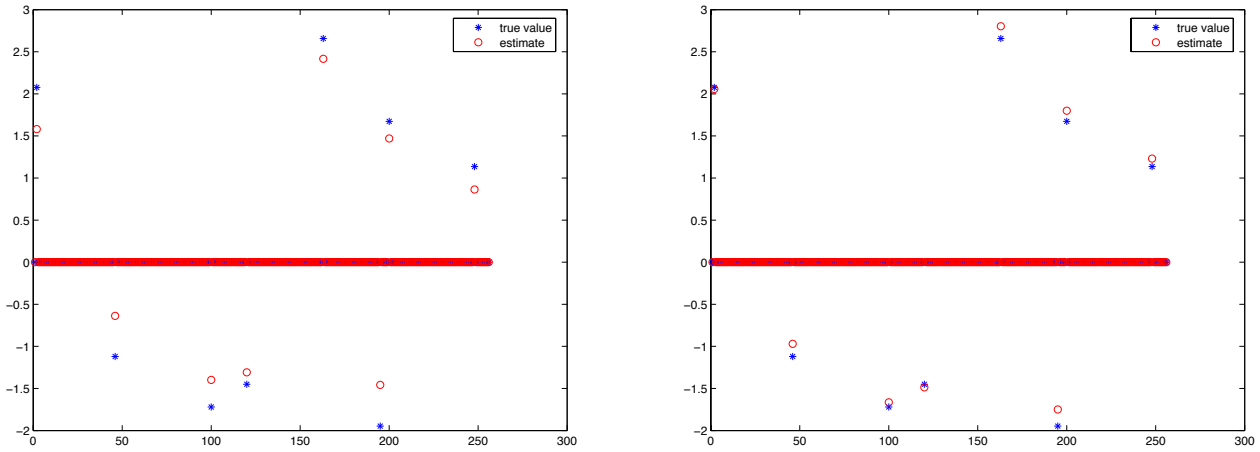
1. *Biomedical imaging.* In the field of bio-medical imaging, one is often only able to collect far fewer measurements than the number of pixels. In Magnetic Resonance Imaging (MRI) for instance, one would like to reconstruct high-resolution images from heavily undersampled frequency data as this would allow image acquisition speeds far beyond those offered by current technologies [15, 28]. If the image is sparse as is the case in Magnetic Resonance Angiography (or if its gradient is sparse or more generally, if the image is sparse in a fixed basis [11]), then ℓ_1-minimization may have a chance to be very effective in such challenging settings.

2. *Analog to digital.* By making a number n of general linear measurements rather than measuring the usual pixels, one could in principle reconstruct a compressible or sparse image with essentially the same resolution as that one would obtain by measuring all the pixels. Now suppose one could design analog sensors able to make measurements by correlating the signal we wish to acquire against incoherent waveforms as discussed in the previous sections. Then one would effectively be able to make up a digital image with far fewer sensors than what is usually considered necessary [9, 22].

3. *Sensor networks.* There are promising applications in sensor networks where taking random projections may yield the same distortion (the same quality of reconstruction) but using much less power than what is traditionally required [24].

4 Numerical Experiments

This section presents numerical experiments to illustrate the Dantzig selector and gives some insights about the numerical method for solving (1.9).



(a) (b)

Figure 2: *Estimation from $y = Ax + z$ with A a 72 by 256 matrix with independent Gaussian entries. A blue star indicates the true value of the parameter and a red circle the estimate. In this example, $\sigma = .11$ and $\lambda = 3.5$ so that the threshold is at $\delta = \lambda \cdot \sigma = .39$. (a) Dantzig selector* (1.7)*. Note that our procedure correctly identifies all the nonzero components of x, and correctly sets to zero all the others. Observe the soft-thresholding-like behavior. (b) Estimation based on the two-stage strategy* (1.20)*. The signal and estimator are very sparse, which is why there is a solid red line at zero.*

4.1 An illustrative example

In this first example, the design matrix A has $n = 72$ rows and $p = 256$ columns, with independent Gaussian entries (and then normalized so that the columns have unit-norm). We then select x with $S := |\{i : x_i \neq 0\}| = 8$, and form $y = Ax + z$, where the z_i's are i.i.d. $N(0, \sigma^2)$. The noise level is adjusted so that

$$\sigma = \frac{1}{3}\sqrt{\frac{S}{n}}.$$

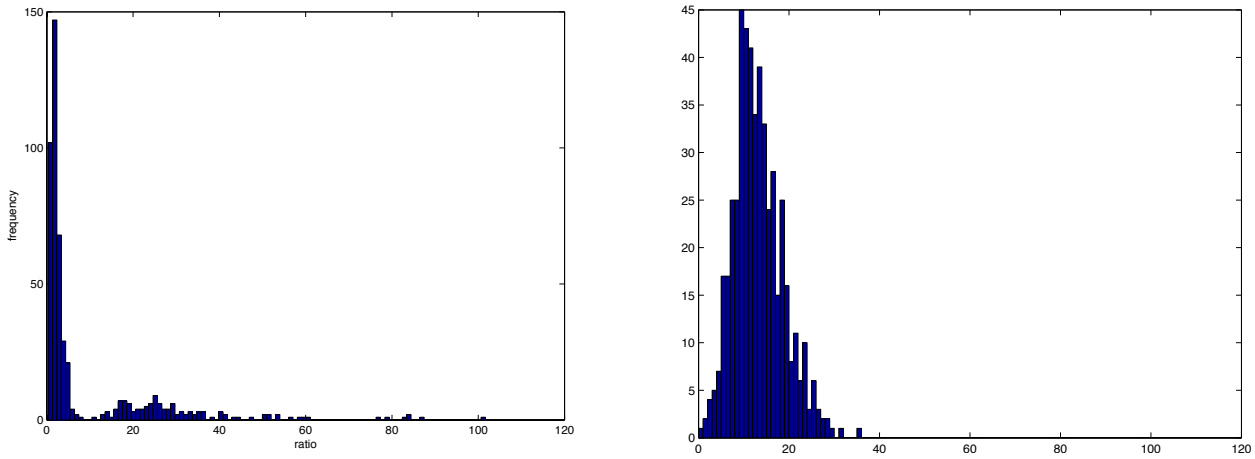
Here and below, the regularizing parameter λ_p in (DS) is chosen via Monte Carlo simulations, i.e. as the empirical maximum of $|A^T z|_i$ over several realizations of $z \sim N(0, I_n)$. The results are presented in Figure 2.

First, we note that in this example, our procedure correctly identifies all the nonzero components of x, and correctly sets to zero all the others. Quantitatively speaking, the ratio ρ^2 between the squared error loss and the ideal squared error (1.11) is equal to

$$\rho^2 := \frac{\sum_i (\hat{x}_i - x_i)^2}{\sum_i \min(x_i^2, \sigma^2)} = 10.28. \tag{4.1}$$

(Note that here $2 \log p = 11.01$.) Second, and as essentially observed earlier, the method clearly exhibits a soft-thresholding type of behavior and as a result, tends to underestimate the true value of the nonzero parameters. However, the two-stage Dantzig selector (1.20) introduced in Section 1.6 corrects for this bias. When applied to the same dataset, it recenters the estimator, and yields an improved squared error since now $\rho^2 = 1.14$, compare with the results in Figure 2 (b).

In our practical experience, the two-stage or Gauss-Dantzig selector procedure tends to



(a) (b)

Figure 3: *Statistics of the ratio between the squared error* $\sum_i (\hat{x}_i - x_i)^2$ *and the ideal mean-squared error* $\sum_i \min(x_i^2, \sigma^2)$. *(a)* $S = 8$, $\sigma = 1/3\sqrt{S/n} = .11$, *and the threshold is here* $\lambda \cdot \sigma = .5814$. *(b)* $S = 8$, $\sigma = \sqrt{S/n} = .33$, *and the threshold is now* $\lambda \cdot \sigma = 1.73$.

outperform our original proposal, and to study its typical quantitative performance, we performed a series of experiments designed as follows:

1. A is a 72 by 256 matrix, sampled as before (A is fixed throughout);
2. select a support set T of size $|T| = S$ uniformly at random, and sample a vector x on T with independent and identically distributed entries according to the model

$$x_i = \varepsilon_i(1 + |a_i|)$$

   where the sign $\varepsilon_i = \pm 1$ with probability 1/2, and $a_i \sim N(0, 1)$ (the moduli and the signs of the amplitudes are independent);
3. make $\tilde{y} = Ax + z$, with $z \sim N(0, \sigma^2 I_n)$ and compute $\hat{x}$ by means of the two-stage procedure (1.20);
4. repeat 500 times for each S, and for different noise levels σ.

The results are presented in Figure 3, and show that our approach works well. With the squared ratio ρ^2 as in (4.1), the median and the mean of ρ^2 are 2.35 and 9.42 respectively for a noise level σ set at $1/3\sqrt{S/n}$. In addition, 75% of the time, ρ^2 is less than 10. With $\sigma = \sqrt{S/n}$, the mean and median are 12.38 and 13.78 respectively, with a bell-shaped distribution.

The reader may wonder why the two histograms in Figure 3 look somewhat different. The answer is simply due to the fact that in Figure 3(a), $\sigma = .11$ and the threshold is about $\lambda \cdot \sigma = .5814$ which means that the nonzero x_i's are above the noise level. In Figure 3(b), however, $\sigma = .33$, and the threshold is $\lambda \cdot \sigma = 1.73$. This means that a fraction of the nonzero components of x are within the noise level and will be set to zero. This explains the observed difference in the quantitative behavior.

4.2 Binary design matrices

We now consider the case where the design matrix has i.i.d entries taking on values ± 1, each with probability 1/2 (the entries are then divided by $\sqrt{n}$ so that the columns have unit-norm). This simply amounts to measuring differences among randomly selected subsets of coordinates. Note that if A had 0/1 entries, one would measure the aggregate effect of randomly selected subsets of coordinates, much like in pooling design schemes. The number of predictors is set to $p = 5,000$ and the number of observations to $n = 1,000$. Of interest here is the estimation accuracy as the number S of significant parameters increases. The results are presented in Table 1. In all these experiments, the nonzero coordinates of the parameter vector are sampled as in Section 4.1 and the noise level is adjusted to $\sigma = 1/3\sqrt{S/n})$ so that with $y = Ax + z$, the variance $\mathbf{E}\|z\|^2 = n\sigma^2$ is proportional to $\mathbf{E}\|x\|^2$ with the same constant of proportionality (fixed signal to noise ratio).

S	5	10	20	50	100	150	200
Dantzig selector	22.81	17.30	28.85	18.49	25.71	49.73	74.93
Gauss-Dantzig selector	.36	.65	1.04	1.09	1.53	13.71	48.74

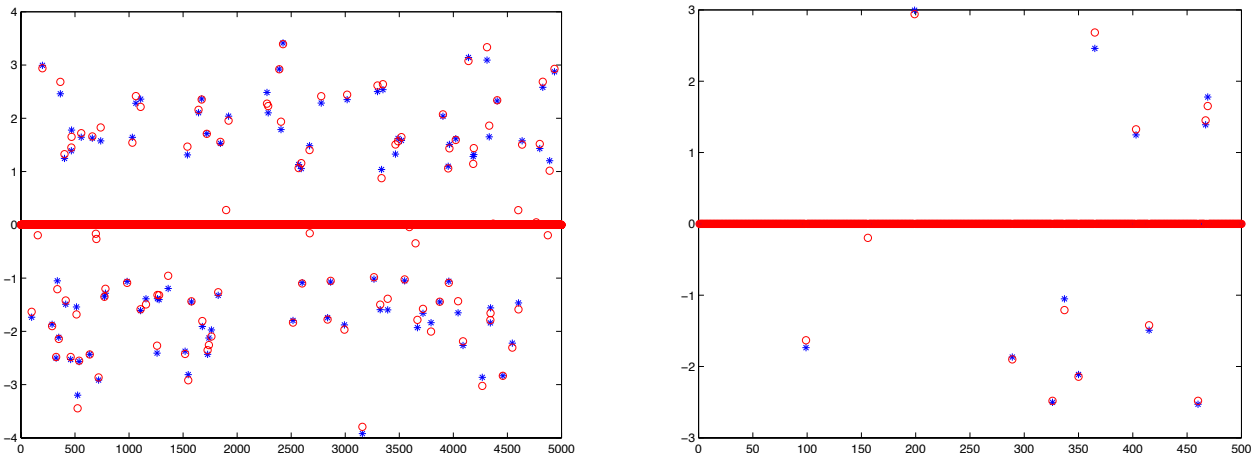
Table 1: *Ratio between the squared error* $\sum_i (\hat{x}_i - x_i)^2$ *and the ideal mean-squared error* $\sum_i \min(x_i^2, \sigma^2)$. *The binary matrix A is the same in all these experiments, and the noise level* σ *is adjusted so that the signal to noise ratio is nearly constant. Both estimators and especially the Dantzig selector exhibit a remarkable performance until a breakdown point around* $S = 200$

Our estimation procedures and most notably the Gauss-Dantzig selector are remarkably accurate as long as the number of significant parameters is not too large, here about $S = 200$. For example, for $S = 100$, the ratio between the Gauss-Dantzig selector's squared error loss and the ideal mean-squared error is only 1.53. Figure 4 illustrates the estimation precision in this case.

To confirm these findings, we now sample the amplitudes of the parameter vector x according to a Cauchy distribution in order to have a wide range of component values x_i, some of which are within the noise level while others are way above; A is fixed and we now vary the number S of nonzero components of x as before, while $\sigma = .5$ is now held constant. The results are presented in Table 2. Again, the Gauss-Dantzig selector performs well.

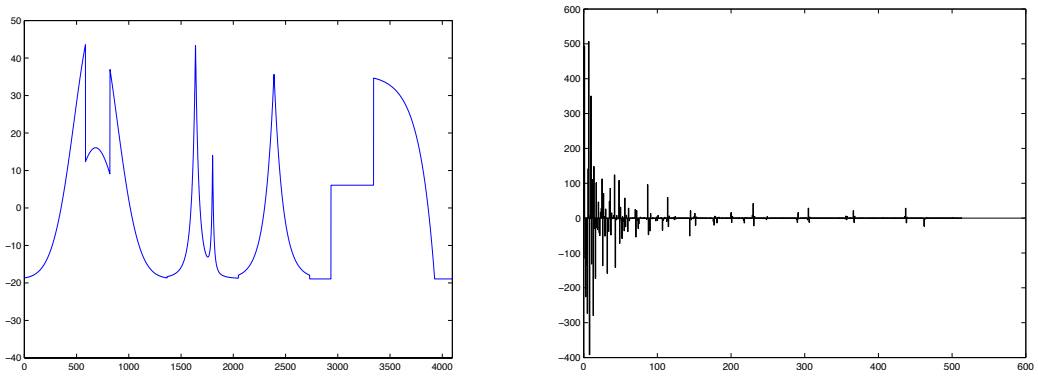
S	5	10	20	50	100	150	200
Gauss-Dantzig selector	3.70	4.52	2.78	3.52	4.09	6.56	5.11

Table 2: *Ratio between the squared error* $\sum_i (\hat{x}_i - x_i)^2$ *and the ideal mean-squared error* $\sum_i \min(x_i^2, \sigma^2)$. *The binary matrix A,* $\sigma = .5$ *and* $\lambda \cdot \sigma = 2.09$ *are the same in all these experiments.*



(a) (b)

Figure 4: *Estimation from $y = Ax + z$ with A a $1,000$ by $5,000$ matrix with independent binary entries. A blue star indicates the true value of the parameter and a red circle the estimate. (a) True parameter values and estimates obtained via the Dantzig selector* (1.20)*. There are $S = 100$ significant parameters. (b) Same plot but showing the first 500 coordinates for higher visibility.*



(a) (b)

Figure 5: *(a) One-dimensional signal f we wish to reconstruct. (b) First 512 wavelet coefficients of f.*

4.3 Examples in signal processing

We are interested in recovering a one-dimensional signal $f \in \mathbf{R}^p$ from noisy and undersampled Fourier coefficients of the form

$$y_j = \langle f, \phi_j \rangle + z_j, \quad 1 \le j \le n,$$

where $\phi_j(t)$, $t = 0, \ldots, p-1$ is a sinusoidal waveform $\phi_j(t) = \sqrt{2/n}\cos(\pi(k_j+1/2)(t+1/2))$, $k_j \in \{0, 1, \ldots, p-1\}$. Consider the signal f in Figure 5; f is not sparse but its wavelet coefficients sequence x is. Consequently, we may just as well estimate its coefficients in a nice wavelet basis. Letting Φ be the matrix with the ϕ_k's as rows, and W be the orthogonal wavelet matrix with wavelets as columns, we have $y = Ax + z$ where $A = \Phi W$, and our estimation procedure applies as is.

The test signal is of size $p = 4,096$ (Figure 5), and we sample a set of frequencies of size $n = 512$ by extracting the lowest 128 frequencies and randomly selecting the others. With

this set of observations, the goal is to study the quantitative behavior of the Gauss-Dantzig selector procedure for various noise levels. (Owing to the factor $\sqrt{2/n}$ in the definition of $\phi_j(t)$, the columns of A have size about one and for each column, individual thresholds λ_i—$|(A^T r)_i| \le \lambda_i \cdot \sigma$—are determined by looking at the empirical distribution of $|(A^T z)_i|$.) We adjust σ so that $\mathbf{E}\|z\|_{\ell_2}^2 = n\sigma^2 = \alpha^2\|Ax\|_{\ell_2}^2$ for various levels of the signal to noise ratio α. We use Daubechies' wavelets with 4 vanishing moments for the reconstruction. The result are presented in Table 3. As one can see, high statistical accuracy holds over a wide range of noise levels. Interestingly, the estimator is less accurate when the noise level is very small ($\alpha = 0.01$) which is not surprising since in this case, there are 178 wavelet coefficients exceeding σ in absolute value.

noise level (α)	.01	.05	.1	.5	1	5
$\sum_i(\hat{x}_i - x_i)^2 / \sum_i \min(x_i^2, \sigma^2)$	15.51	2.08	1.40	1.47	.91	1.00

Table 3: *Performance of the Gauss-Dantzig procedure in estimating a signal from undersampled and noisy Fourier coefficients. The subset of variables is here estimated by $|\hat{x}_i| > \sigma/4$ with $\hat{x}$ as in* (1.7)*. The noise level in the top row is the signal to noise ratio: $\alpha^2 = \|Ax\|_{\ell_2}^2/\mathbf{E}\|z\|_{\ell_2}^2$.*

In our last example, we consider the problem of reconstructing an image from undersampled Fourier coefficients. Here $x(t_1, t_2)$, $0 \le t_1, t_2 < N$, is an unknown N by N image so that p is the number of unknown pixels, $p = N^2$. As usual, the data is given by $y = Ax + z$, where

$$(Ax)_k = \sum_{t_1,t_2} x(t_1, t_2) \cos(2\pi(k_1 t_1 + k_2 t_2)/N), \quad k = (k_1, k_2), \tag{4.2}$$

or $(Ax)_k = \sum_{t_1,t_2} x(t_1, t_2) \sin(2\pi(k_1 t_1 + k_2 t_2)/N)$. In our example (see Figure 6(b)), the image x is not sparse but the gradient is. Therefore, to reconstruct the image, we apply our estimation strategy and minimize the ℓ_1-norm of the gradient size, also known as the total-variation of x

$$\min \|\tilde{x}\|_{TV} \quad \text{subject to} \quad |(A^T r)_i| \le \lambda_i \cdot \sigma \tag{4.3}$$

(the individual thresholds again depend on the column sizes as before); formally, the total-variation norm is of the form

$$\|\tilde{x}\|_{TV} = \sum_{t_1,t_2} \sqrt{|D_1\tilde{x}(t_1, t_2)|^2 + |D_2\tilde{x}(t_1, t_2)|^2},$$

where D_1 is the finite difference $D_1\tilde{x} = x(t_1, t_2) - x(t_1 - 1, t_2)$ and $D_2\tilde{x} = \tilde{x}(t_1, t_2) - \tilde{x}(t_1, t_2 - 1)$; in short, $\|\tilde{x}\|_{BV}$ is the ℓ_1-norm of the size of the gradient $D\tilde{x} = (D_1\tilde{x}, D_2\tilde{x})$, see also [29].

Our example follows the data acquisition patterns of many real imaging devices which can collect high-resolution samples along radial lines at relatively few angles. Figure 6(a) illustrates a typical case where one gathers $N = 256$ samples along each of 22 radial lines. In a first experiment then, we observe 22×256 noisy real-valued Fourier coefficients and use (4.3) for the recovery problem illustrated in Figure 6. The number of observations is then $n = 5,632$ whereas there are $p = 65,536$ observations. In other words, about 91.5% of the 2D Fourier coefficients of x are missing. The signal to noise ratio in this experiment is equal to $\|Ax\|_{\ell_2}/\|z\|_{\ell_2} = 5.85$. Figure 6(c) shows the reconstruction obtained by setting

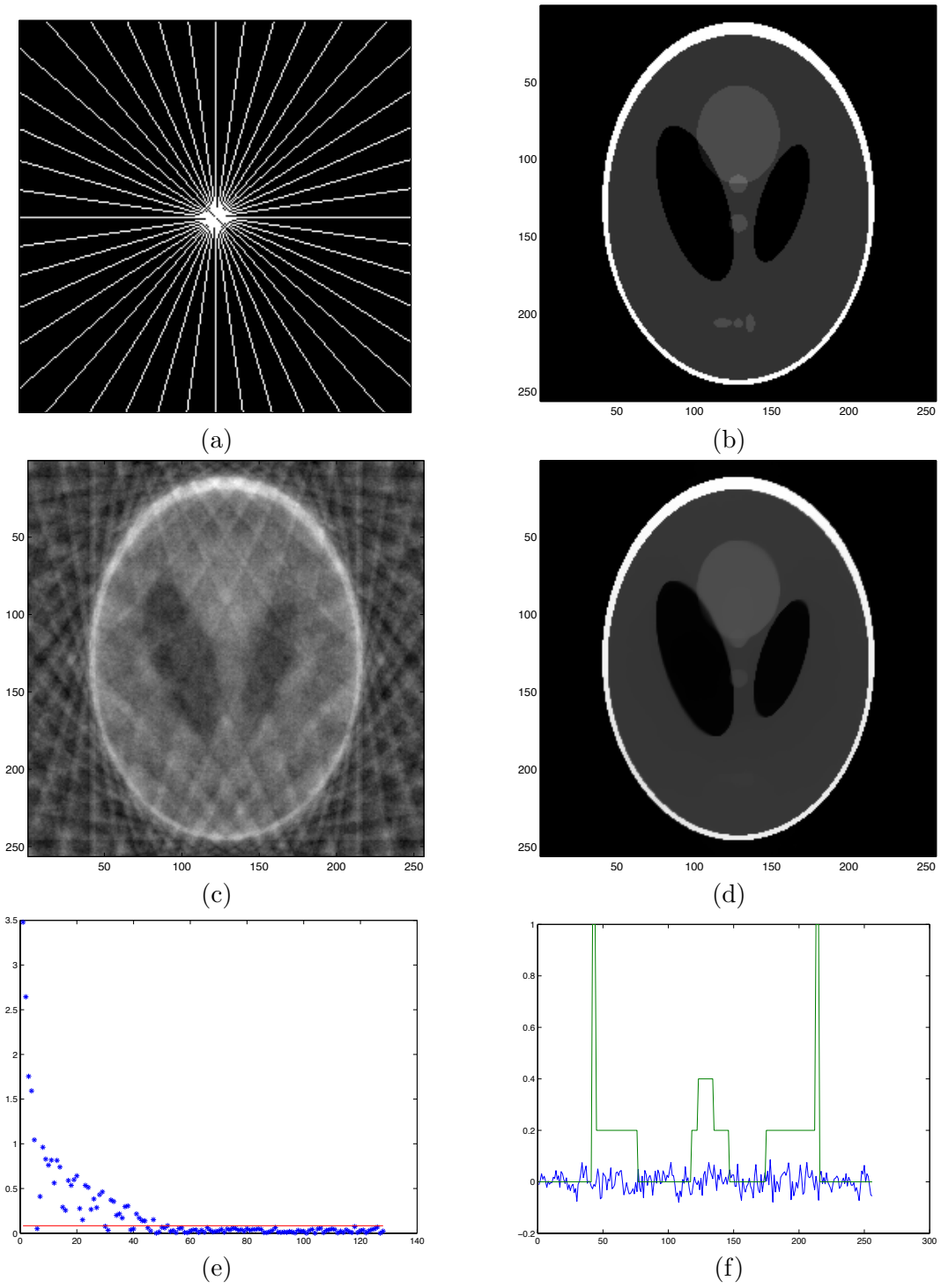


Figure 6: *(a) Sampling 'domain' in the frequency plane; Fourier coefficients are sampled along 22 approximately radial lines; here, $n \approx .086p$. (b) The Logan-Shepp phantom test image. (c) Minimum energy reconstruction obtained by setting unobserved Fourier coefficients to zero. (d) Reconstruction obtained by minimizing the total-variation, as in* (4.3)*. (e) Magnitude of the true Fourier coefficients along a radial line (frequency increases from left to right) on a logarithmic scale. Blue stars indicate values of* $\log(1 + |(Ax)_k|)$ *while the solid red line indicates the noise level* $\log(1 + \sigma)$*. Less than a third of the frequency samples exceed the noise level. (f)* $A^T z$ *and* x *are plotted along a scanline to convey a sense of the noise level.*

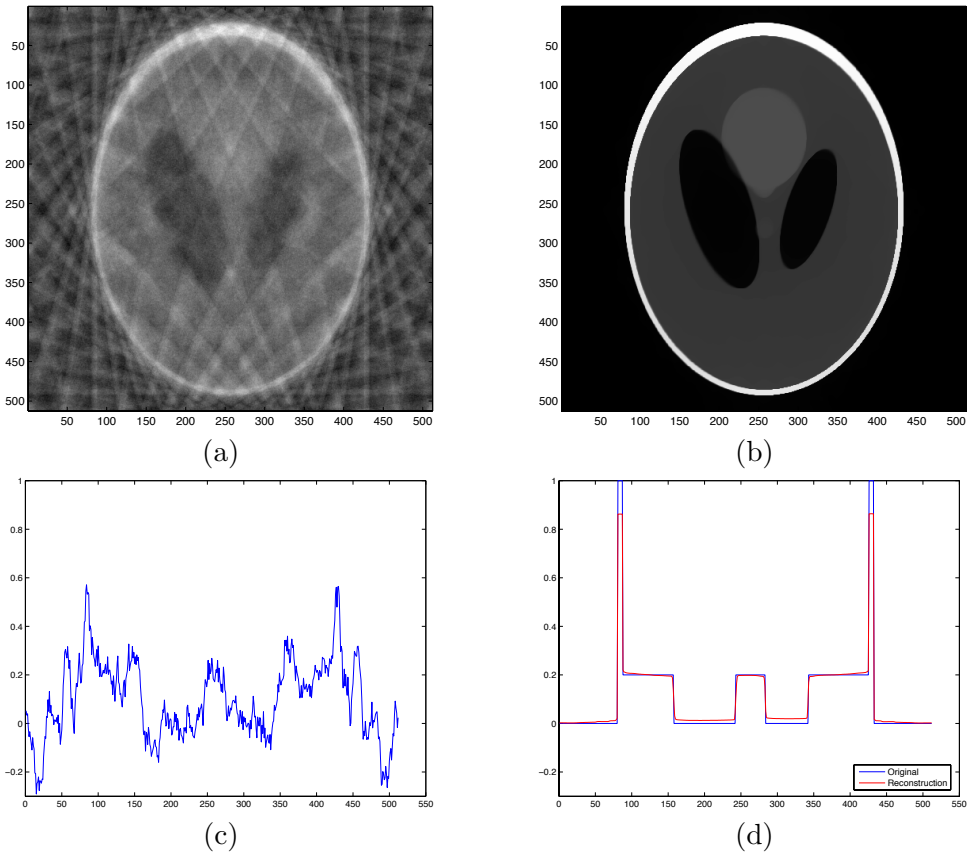


Figure 7: *(a) & (b). Similar experience than in Figure 6 but at a higher resolution ($p = 512^2$) so that now $n \approx .043p$. (c) & (d). Scanlines of both reconstructions.*

the unobserved Fourier coefficients to zero, while (d) shows the reconstruction (4.3). We follow up with a second experiment where the unknown image is now 512 by 512 so that $p = 262,144$ and $n = 22 \times 512 = 11,264$. The fraction of missing Fourier coefficients is now approaching 96%. The signal to noise ratio is about the same, $\|Ax\|_{\ell_2}/\|z\|_{\ell_2} = 5.77$. The reconstructions are of very good quality, especially when compared to the naive reconstruction which minimizes the energy of the reconstruction subject to matching the observed data. Figure 7 also shows the middle horizontal scanline of the phantom. As expected, we note a slight loss of contrast due to the nature of the estimator which here operates by "soft-thresholding" the gradient. There are, of course, ways of correcting for this bias but such issues are beyond the scope of this paper.

4.4 Implementation

In all the experiments above, we used a primal-dual interior point algorithm for solving the linear program (1.9). We used a specific implementation which we outline as this gives some insight about the computational workload of our method. For a general linear program with inequality constraints

$$\min c^T x \qquad \text{subject to} \qquad Fx \le b,$$

define

- $f(x) = Ax - b$,
- $r_{\text{dual}} = c + A^T\lambda$,
- $r_{\text{cent}} = -\text{diag}(\lambda)f(x) - \mathbf{1}/t$,

where $\lambda \in \mathbf{R}^m$ are the so-called dual variables, and t is a parameter whose value typically increases geometrically at each iteration; there are as many dual variables as inequality constraints. In a standard primal-dual method (with logarithmic barrier function) [7], one updates the current primal-dual pair (x, λ) by means of a Newton step, and solve

$$\begin{pmatrix} 0 & F^T \\ -\text{diag}(\lambda)F & -\text{diag}(f(x)) \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta\lambda \end{pmatrix} = - \begin{pmatrix} r_{\text{dual}} \\ r_{\text{cent}} \end{pmatrix};$$

that is,

$$\begin{aligned} \Delta\lambda &= -\text{diag}(1/f(x))(\text{diag}(\lambda)F\Delta_x - r_{\text{cent}}) \\ -[F^T\text{diag}(\lambda/f(x))F]\Delta x &= -(r_{\text{dual}} + F^T\text{diag}(1/f(x))r_{\text{cent}}). \end{aligned}$$

The current guess is then updated via $(x^+, \lambda^+) = (x, \lambda) + s(\Delta x, \Delta s)$ where the stepsize s is determined by line search or otherwise. Typically, the sequence of iterations stops once the primal-dual gap and the size of the residual vector fall below a specified tolerance level [7].

Letting $U = A^TA$ and $\tilde{y} = A^Ty$ in (1.9), our problem parameters have the following block structure

$$F = \begin{pmatrix} I & -I \\ -I & -I \\ U & 0 \\ -U & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \\ \delta + \tilde{y} \\ \delta - \tilde{y} \end{pmatrix}, \quad c = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

which gives

$$F^T\text{diag}(\lambda/f)F = \begin{pmatrix} D_1 + D_2 + U^T(D_3 + D_4)U & D_2 - D_1 \\ D_2 - D_1 & D_1 + D_2 \end{pmatrix},$$

where $D_i = \text{diag}(\lambda_i/f_i)$, $1 \le i \le 4$, and

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{pmatrix} \tilde{x} - u \\ -\tilde{x} - u \\ U\tilde{x} - \delta - \tilde{y} \\ -U\tilde{x} - \delta + \tilde{y} \end{pmatrix}.$$

Put $\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = r_{\text{dual}} + F^T\text{diag}(1/f(x))r_{\text{cent}}$. It is convenient to solve the system

$$F^T\text{diag}(\lambda/f)F \begin{pmatrix} \Delta\tilde{x} \\ \Delta u \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$$

by elimination and obtain

$$\begin{aligned} (4(D_1 + D_2)^{-1}D_1D_2 + U^T(D_3 + D_4)U)\Delta\tilde{x} &= r_1 - (D_1 + D_2)^{-1}(D_1 - D_2)r_2 \\ (D_1 + D_2)\Delta u &= r_2 - (D_2 - D_1)\Delta\tilde{x}. \end{aligned}$$

In other words, each step may involve solving a p by p system of linear equations. In fact, when n is less than p, it is possible to solve a smaller system thanks to the Sherman-Woodbury-Morrison formula. Indeed, write $U^T(D_3 + D_4)U = A^T B$ where $B = A(D_3 + D_4)A^T A$ and put $D_{12} = 4(D_1 + D_2)^{-1} D_1 D_2$. Then

$$(D_{12} + A^T B)^{-1} = D_{12}^{-1} - D_{12}^{-1} A^T (I + B D_{12}^{-1} A^T)^{-1} B D_{12}^{-1}.$$

The advantage is that one needs to solve the smaller n by n system $(I + B D_{12}^{-1} A^T)x' = b'$. Hence the cost of each Newton iteration is essentially that of solving an n by n system of linear questions plus that of forming the matrix $(I + B D_{12}^{-1} A^T)$. As far as the number of Newton iterations is concerned, we ran thousands of experiments and have never needed more than 45 Newton iterations for convergence.

Note that in some important applications, we may have fast algorithms for applying A and A^T to an arbitrary vector, as in the situation where A is a partial Fourier matrix since one can make use of FFT's. In such settings, one never forms $A^T A$ and uses iterative algorithms such as Conjugate Gradients for solving such linear systems; of course, this speeds up the computations.

Finally, a collection of MATLAB routines solving (1.9) for reproducing some of these experiments and testing these ideas in other setups is available at the following address: `http://www.acm.caltech.edu/~emmanuel/l1.shtml`.

5 Discussion

5.1 Significance for Model Selection

We would like to briefly discuss the implications of this work for model selection. Given a data matrix A with (unit-normed column) and observations $y = Ax + z$, our procedure will estimate x by that vector with minimum ℓ_1-norm among all objects $\tilde{x}$ obeying

$$|A^T r|_i \le (1 + t^{-1})\sqrt{2 \log p} \cdot \sigma, \text{ where } r = y - A\tilde{x}. \tag{5.1}$$

As the theory and the numerical experiments suggest, many of the coordinates $\hat{x}_i$ will typically be zero (at least under the assumption that the true unknown vector x is sparse) and, therefore, our estimation procedure effectively returns a candidate model $\hat{I} := \{i : \hat{x}_i \neq 0\}$.

As we have seen, the Dantzig selector is guaranteed to produce optimal results if

$$\delta(A)_{2S} + \theta(A)_{S,2S} < 1 \tag{5.2}$$

(note that since $\theta(A)_{S,2S} \le \delta_{3S}$, it would be sufficient to have $\delta(A)_{2S} + \delta(A)_{3S} < 1$). We have commented on the interpretation of this condition already. In a typical model selection problem, A is given; it is then natural to ask if this particular A obeys (5.2) for the assumed level S of sparsity. Unfortunately, obtaining an answer to this question might be computationally prohibitive as it may require checking the extremal singular values of exponentially many submatrices.

While this may represent a limitation, two observations are in order. First, there is empirical evidence and theoretical analysis suggesting approximate answers for certain types of

random matrices, and there is nowadays a significant amount of activity developing tools to address these questions [16]. Second, the failure of (5.2) to hold, is in general indicative of a structural difficulty of the problem, so that any procedure is likely to be unsuccessful for sparsity levels in the range of the critical one. To illustrate this point, let us consider the following example. Suppose that $\delta_{2S} + \theta_{S,2S} > 1$. Then this says that δ_{2S} is large and since δ_S is increasing in S, it may very well be that for S' in the range of S, e.g. $S' = 3S$, there might be submatrices (in fact possibly many submatrices) with S' columns which are either rank-deficient or which have very small singular values. In other words, the significant entries of the parameter vector might be arranged in such a way so that even if one knew their location, one would not be able to estimate their values because of rank-deficiency. This informally suggests that the Dantzig selector breaks down near the point where any estimation procedure, no matter how intractable, would fail.

Finally, it is worth emphasizing the connections between RIC [23] and (1.7). Both methods suggest a penalization which is proportional to the logarithm of the number of explanatory variables—a penalty that is well justified theoretically [6]. For example, in the very special case where A is orthonormal, $p = n$, RIC applies a hard-thresholding rule to the vector $A^T y$ at about $O(\sqrt{2 \log p})$, while our procedure translates in a soft-thresholding at about the same level; in our convex formulation (1.7), this threshold level is required to make sure that the true vector is feasible. In addition, the ideas developed in this paper have a broad applicability, and it is likely that they might be deployed and give similar bounds for RIC variable selection. Despite such possible similarities, our method differs substantially from RIC in terms of computational effort since (1.7) is tractable while RIC is not. In fact, we are not aware of any work in the model selection literature which is close in intent and in the results.

5.2 Connections with other works

In [12], the authors study a related problem where the goal is to recover a vector x from incomplete and contaminated measurements $y = Ax + e$, where the error vector (stochastic or deterministic) obeys $\|e\|_{\ell_2}^2 \le D$. There, the proposed reconstruction searches, among all objects consistent with the data y, for that with minimum ℓ_1-norm

$$(P_2) \qquad \min \ \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad \|y - A\tilde{x}\|_{\ell_2}^2 \le D. \tag{5.3}$$

Under essentially the same conditions as those of Theorems 1.1 and 1.2, [12] showed that the reconstruction error is bounded by

$$\|x^\sharp - x\|_{\ell_2}^2 \ \le \ C_3^2 \cdot D. \tag{5.4}$$

In addition, it is also argued that for arbitrary errors, the bound (5.4) is sharp and that in general, one cannot hope for a better accuracy.

If one assumes that the error is stochastic as in this paper, however, a mean squared error of about D might be far from optimal. Indeed, with the linear model (1.1), $D \sim \sigma^2 \chi_n$ and therefore D has size about $n\sigma^2$. But suppose now x is sparse and has only three nonzero coordinates, say, all exceeding the noise level. Then whereas Theorem 1.1 gives a loss of about $3\sigma^2$ (up to a log factor), (5.4) only guarantees an error of size about $n\sigma^2$. What is missing in [12] and is achieved here is the adaptivity to the unknown level of sparsity of the object we try to recover. Note that we do not claim that the program

(P_2) is ill-suited for adaptivity. It is possible that refined arguments would yield estimators based on quadratically constrained ℓ_1-minimization (variations of (5.3)) obeying the special adaptivity properties discussed in this paper.

Last but not least and while working on this manuscript, we became aware of related work by Haupt and Nowak [24]. Motivated by recent results such as [8, 9, 21], they studied the problem of reconstructing a signal from noisy random projections and obtained powerful quantitative estimates which resemble ours. Their setup is different though since they exclusively work with random design matrices, in fact random Rademacher projections, and do not study the case of fixed A's. In contrast, our model for A is deterministic and does not involve any kind of randomization, although our results can of course be specialized to random matrices. But more importantly, and perhaps this is the main difference, their estimation procedure requires solving a combinatorial problem much like (1.15) whereas we use linear programming.

References

[1] Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19** (1974), 716–723.

[2] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, **117** (2000), 467–493.

[3] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113** (1999), 301–413.

[4] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, **37** (1991), 1034–1054.

[5] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[6] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, **3** (2001), 203–268.

[7] S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[8] E. J. Candès, J. Romberg, and T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. Submitted to *IEEE Transactions on Information Theory*, June 2004. Available on the ArXiV preprint server: `math.GM/0409186`.

[9] E. J. Candès, and T. Tao, Near optimal signal recovery from random projections: universal encoding strategies? Submitted to *IEEE Transactions on Information Theory*, October 2004. Available on the ArXiV preprint server: `math.CA/0410542`.

[10] E. J. Candès and T. Tao. Decoding by linear programming. Submitted, December 2004. Available on the ArXiV preprint server: `math.MG/0502327`.

[11] E. Candès and J. Romberg. Practical signal recovery from random projections. *submitted to IEEE Trans. Signal Proc.*, January 2005.

[12] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. Submitted, February 2005. Available on the ArXiV preprint server: `math.NA/0503066`.

[13] E. J. Candès and T. Tao. Error Correction via linear programming. Submitted to FOCS 2005, April 2005.

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Scientific Computing* **20** (1999), 33–61.

[15] B. L. Daniel, Y. F. Yen, and G. H. Glover et al. Breast disease: dynamic spiral MR imaging. *Radiology* **209** (1998), 499–509.

[16] I. Daubechies. *Personal communication.* May 2005.

[17] D. L. Donoho and I. M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sci. Paris Sér. I Math.*, **319** (1994), 1317–1322.

[18] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81** (1994), 425–455.

[19] D. L. Donoho and I. M. Johnstone. Empirical atomic decomposition. Unpublished manuscript (1995).

[20] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, **47** (2001), 2845–2862.

[21] D. L. Donoho. For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ_1-norm Solution is also the Sparsest Solution. Manuscript, September 2004.

[22] D. L. Donoho. Compressed sensing. Manuscript, September 2004.

[23] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, **22** (1994), 1947–1975.

[24] J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. Submitted to *IEEE Trans. Info. Theory.* April 2005.

[25] Statistics: Challenges and Opportunities for the Twenty-First Century. NSF report. Edited by: Jon Kettenring, Bruce Lindsay, and David Siegmund, April 2003.

[26] C. L. Mallows. Some comments on C_p. *Technometrics* **15** (1973), 661–676.

[27] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.* **24** (1995), 227-234.

[28] D. C. Peters, T. M. Grist, F. R. Korosec, J. E. Holden, W. F. Block, K. L. Wedding, T. J. Carroll, and C. A. Mistretta. Undersampled projection reconstruction applied to MR angiography. *Magn. Reson. Med.* **43** (2000), 91–101.

[29] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation noise removal algorithm. *Physica D* **60** (1992), 259–68.

[30] S. Sardy, A. G. Bruce and P. Tseng. Block coordinate relaxation methods for nonparametric wavelet denoising. *J. Comput. Graph. Statist.* bf 9 (2000), 361–379.

[31] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

[32] S. J. Szarek. Condition numbers of random matrices. *J. Complexity* **7** (1991), 131–149.

[33] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** (1996), 267–288.