Near-Oracle Performance of Basis Pursuit under Random Noise

Zvika Ben-Haim, Yonina C. Eldar, and Michael Elad

Abstract

We consider the problem of estimating a deterministic sparse vector $\mathbf{x}_0$ from underdetermined noisy measurements, in which the noise is a Gaussian random vector. Two techniques which are commonly used in this setting are the Dantzig selector and basis pursuit denoising (BPDN). It has previously been shown that, with high probability, the Dantzig selector comes close to the performance of the oracle estimator which knows the locations of the nonzero components of $\mathbf{x}_0$, when the performance is measured by the ℓ_2 distance between $\mathbf{x}_0$ and its estimate. In this paper, we demonstrate that BPDN achieves analogous results, but that the constants involved in the BPDN analysis are significantly better.

Submitted to *IEEE Trans. Signal Processing*, March 2009.
Index terms: Sparse estimation, basis pursuit denoising, Dantzig selector.

I. Introduction

Estimation problems with sparsity constraints have attracted considerable attention in recent years because of their potential use in numerous signal processing applications, such as denoising, compression and sampling. In a typical setup, an unknown deterministic parameter $\mathbf{x}_0 \in \mathbb{R}^m$ is to be estimated from measurements $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$. The dictionary $\mathbf{A} \in \mathbb{R}^{n\times m}$ is deterministic and known, but typically consists of more columns than rows (i.e., $m > n$), so that without further assumptions, $\mathbf{x}_0$ is unidentifiable from $\mathbf{b}$. The impassé is resolved by assuming that the parameter vector is sparse, i.e., that most elements of $\mathbf{x}_0$ are zero. Under the assumption of sparsity, several estimation approaches can be used; these include basis pursuit denoising (BPDN) [1] (also known as the Lasso in the context of regression problems in machine learning) and the Dantzig selector [2]. Both techniques seek estimates having a low ℓ_1 norm; the difference between these approaches is in the method of ensuring consistency between the estimate and the measurements. A large body of theoretical and empirical evidence demonstrates the ability of BPDN and the Dantzig selector to approximate $\mathbf{x}_0$, as long as $\mathbf{x}_0$ is sufficiently sparse [2]–[4].

To completely specify the problem, we must define in some way the behavior of the measurement noise. This is usually done using one of two approaches. The first is to analyze performance assuming the noise is unknown but bounded. We refer to such a setting as adversarial, since in this case one must demonstrate successful operation even if $\mathbf{w}$ is chosen so as to maximally harm the estimator. An arguably more realistic scenario is to model $\mathbf{w}$ as a random vector, and seek performance guarantees which hold with high probability. This last option should be distinguished from Bayesian performance analysis, as practiced in [5], [6], where on top of the stochastic model for $\mathbf{w}$, a probabilistic model for $\mathbf{x}_0$ is also used. In this paper, we assume a deterministic parameter $\mathbf{x}_0$, which necessarily leads to weaker bounds.

As noted previously [2], [3], the assumption of random noise yields significant improvements in performance guarantees, compared with the adversarial setting. Indeed, Candès and Tao [2] have demonstrated that the Dantzig selector is close to optimal under the assumption of Gaussian random noise. More specifically, they showed that, with high probability, the ℓ_2 distance between $\mathbf{x}_0$ and the Dantzig estimate is within a constant times $\log m$ of the performance of an ideal "oracle" estimator, which knows the locations of the nonzero elements of $\mathbf{x}_0$. The $\log m$ factor is an unavoidable consequence of the fact that the nonzero locations in $\mathbf{x}_0$ are unknown [7]. Thus, the error of the Dantzig selector comes within a constant multiple of the minimum achievable estimation error.

Z. Ben-Haim and Y. C. Eldar are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: {zvikabh@tx, yonina@ee}.technion.ac.il). M. Elad is with the Department of Computer Science, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: elad@cs.technion.ac.il). Contact information for Z. Ben-Haim: phone +972-4-8294700, fax +972-4-8295757. This work was supported in part by the Israel Science Foundation under Grants 1081/07 and 599/08, and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract no. 216715).

Can such a performance guarantee be provided for the BPDN approach as well? This question was raised immediately after the publication of the Dantzig selector result [8], [9]. Recently, Bickel et al. [10] have demonstrated that, indeed, the performance of BPDN is bounded, with high probability, by $C \log m$ times the oracle performance, for a constant C. However, the constant involved in this analysis is considerably larger than that of the Dantzig selector.

In this paper, we provide a performance guarantee for BPDN which is better than that of the Dantzig selector, sometimes by more than an order of magnitude. Thus, in terms of guaranteed performance under random noise, BPDN is at least as successful as the Dantzig selector. Our analysis also provides a rule of thumb for choosing the regularization parameter of BPDN, and we show numerically that this rule is useful even beyond the range of settings supported by the theoretical analysis. Our results are formulated using only the mutual coherence of the dictionary, and thus do not require knowledge of any computationally intractable properties of the problem setting.

The rest of this paper is organized as follows. We begin in Section II with an overview of the problem setting and an analysis of several measures of the suitability of given dictionaries for sparse estimation. In Section III, we review some convex relaxation techniques for sparse estimation. In Section IV, we analyze the performance of BPDN under adversarial noise. This motivates the introduction of random noise, for which substantially better guarantees can be obtained in Section V. Finally, the validity of these results is examined in a practical estimation scenario in Section VI.

The following notation is used throughout the paper. Vectors and matrices are denoted, respectively, by boldface lowercase and boldface uppercase letters. The indices of the nonzero entries of a vector $\mathbf{x}$ are called its support set and denoted $\mathrm{supp}(\mathbf{x})$. Given an index set Λ and a matrix $\mathbf{A}$, the notation $\mathbf{A}_\Lambda$ refers to the submatrix formed from the columns of $\mathbf{A}$ indexed by Λ. The ℓ_p norm of a vector $\mathbf{x}$, for $1 \leq p \leq \infty$, is denoted $\|\mathbf{x}\|_p$, while $\|\mathbf{x}\|_0$ refers to the number of nonzero elements in $\mathbf{x}$.

II. Characterizing the Dictionary

Let $\mathbf{x}_0 \in \mathbb{R}^m$ be an unknown vector, and denote its support set by $\Lambda_0 = \mathrm{supp}(\mathbf{x}_0)$. Let $k = \|\mathbf{x}_0\|_0$ be the number of nonzero entries in $\mathbf{x}_0$. In our setting, it is typically assumed that k is much smaller than m, i.e., that most elements in $\mathbf{x}_0$ are zero. Suppose we obtain noisy measurements

$$\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a known overcomplete dictionary ($m > n$). We refer to the columns $\mathbf{a}_i$ of $\mathbf{A}$ as the *atoms* of the dictionary, and assume throughout our work that the atoms are normalized, $\|\mathbf{a}_i\|_2 = 1$. We will consider primarily the situation in which the noise $\mathbf{w}$ is random, though for comparison we will also examine the case of a bounded deterministic noise vector; a precise definition of $\mathbf{w}$ is deferred to subsequent sections.

For $\mathbf{x}_0$ to be identifiable, one must guarantee that different values of $\mathbf{x}_0$ produce significantly different values of $\mathbf{b}$. One way to ensure this is to examine all possible *subdictionaries*, or k-element sets of atoms, and verify that the subspaces spanned by these subdictionaries differ substantially from one another.

More specifically, several methods have been proposed to formalize the notion of the suitability of a dictionary for sparse estimation. These include the mutual coherence, the cumulative coherence [3], the exact recovery coefficient (ERC) [3], the spark [4], and the uniform uncertainty principle (UUP) [2], [11]. Except for the mutual coherence and cumulative coherence, none of these measures can be efficiently calculated, in general, for a given matrix $\mathbf{A}$. Since the values of the cumulative and mutual coherence are quite close, our focus in this paper will be on the mutual coherence $\mu = \mu(\mathbf{A})$, which is defined as

$$\mu \triangleq \max_{i \neq j} |\mathbf{a}_i^* \mathbf{a}_j| \,. \tag{2}$$

While the mutual coherence can be efficiently calculated directly from (2), it is not immediately clear in what way μ is related to the requirement that subdictionaries must span different subspaces. Indeed, μ ensures a lack of correlation between single atoms, while we require a distinction between k-element subdictionaries. To explore this relation, we will now define the UUP, which is more directly related to the subdictionaries of $\mathbf{A}$. We will then show that the mutual coherence can be used to bound the constants involved in the UUP, a fact which will also prove useful in our subsequent analysis. This strategy is inspired by earlier works, which have used the mutual

coherence to bound the ERC [3] and the spark [4]. Thus, the coherence can be viewed as a tractable proxy for more accurate measures of the quality of a dictionary, which cannot themselves be calculated efficiently.

By the UUP we refer to two properties describing "good" dictionaries, namely, the restricted isometry property (RIP) and the restricted orthogonality property (ROP), which we now define. A dictionary $\mathbf{A}$ is said to satisfy the RIP [11] of order k with parameter δ_k if, for every index set Λ of size k, we have

$$(1-\delta_k)\|\mathbf{y}\|_2^2 \leq \|\mathbf{A}_\Lambda \mathbf{y}\|_2^2 \leq (1+\delta_k)\|\mathbf{y}\|_2^2 \tag{3}$$

for all $\mathbf{y} \in \mathbb{R}^k$. Thus, when δ_k is small, the RIP ensures that any k-atom subdictionary is nearly orthogonal, which in turn implies that any two disjoint $k/2$-atom subdictionaries are well-separated.

Similarly, $\mathbf{A}$ is said to satisfy the ROP [2] of order (k_1, k_2) with parameter θ_{k_1,k_2} if, for every pair of disjoint index sets Λ_1 and Λ_2 having cardinalities k_1 and k_2, respectively, we have

$$\left|\mathbf{y}_1^* \mathbf{A}_{\Lambda_1}^* \mathbf{A}_{\Lambda_2} \mathbf{y}_2\right| \leq \theta_{k_1,k_2}\|\mathbf{y}_1\|_2\|\mathbf{y}_2\|_2 \tag{4}$$

for all $\mathbf{y}_1 \in \mathbb{R}^{k_1}$ and for all $\mathbf{y}_2 \in \mathbb{R}^{k_2}$. In words, the ROP requires any two disjoint subdictionaries containing k_1 and k_2 elements, respectively, to be nearly orthogonal to each other. These two properties are therefore closely related to the requirement that distinct subdictionaries of $\mathbf{A}$ behave dissimilarly.

In recent years, it has been demonstrated that various practical estimation techniques successfully approximate $\mathbf{x}_0$ from $\mathbf{b}$, if the constants δ_k and θ_{k_1,k_2} are sufficiently small [2], [11], [12]. This occurs, for example, when the entries in $\mathbf{A}$ are chosen randomly according to a Gaussian IID distribution, as well as in some specific deterministic dictionary constructions.

Unfortunately, in the standard estimation setting, one cannot design the system matrix $\mathbf{A}$ according to these specific rules. In general, if one is given a particular dictionary $\mathbf{A}$, then there is no known algorithm for efficiently determining its UUP constants. Indeed, the very nature of the UUP properties seems to require enumerating over an exponential number of subspaces in order to find the "worst" subdictionary. While the mutual coherence μ of (2) tends to be far less accurate in capturing the accuracy of a dictionary, it is still useful to be able to say something about the UUP constants based only on μ. Such a result is given in the following lemma.

Lemma 1: For any matrix $\mathbf{A}$, the RIP constant δ_k of (3) and the ROP constant θ_{k_1,k_2} of (4) satisfy the bounds

$$\delta_k \leq (k-1)\mu, \tag{5}$$

$$\theta_{k_1,k_2} \leq \mu\sqrt{k_1 k_2} \tag{6}$$

where μ is the mutual coherence (2).

The proof of Lemma 1 can be found in Appendix I. We will apply this lemma in Section V, when examining the performance of the Dantzig selector. The lemma can also be used in conjunction with other results that rely on the RIP and ROP.

In the next section, we highlight some of the methods used for approximating $\mathbf{x}_0$ from $\mathbf{b}$. The ability of these algorithms to accurately estimate $\mathbf{x}_0$ turns out to be related to the merit of the dictionary, as quantified above. We will be interested, in particular, in performance guarantees which are a function of the mutual coherence; such results will be presented in Sections IV and V.

III. Estimation Techniques

We now briefly review several approaches for estimating $\mathbf{x}_0$ from noisy measurements $\mathbf{b}$ given by (1). Our main focus in this paper is the ℓ_1-penalty version of BPDN, which estimates $\mathbf{x}_0$ as a solution $\hat{\mathbf{x}}_{\mathrm{BP}}$ to the quadratic program

$$\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 + \gamma\|\mathbf{x}\|_1 \tag{7}$$

for some regularization parameter γ. We refer to the optimization problem (7) as BPDN; this is not to be confused with the related ℓ_1-error version of BPDN, given by

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 \leq \delta \tag{8}$$

for a given value of δ. Since (7) is the Lagrangian of (8), the two problems are related in that any solution of the ℓ_1-penalty problem for a particular γ corresponds to a solution of the ℓ_1-error version with an appropriate choice of δ. Nevertheless, strictly speaking, (7) and (8) are distinct.

A more recently proposed estimator is the Dantzig selector [2], defined as a solution $\hat{\mathbf{x}}_{\mathrm{DS}}$ to the optimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{A}^*(\mathbf{b}-\mathbf{A}\mathbf{x})\|_\infty \leq \tau \tag{9}$$

where τ is again a user-selected parameter. The Dantzig selector, like BPDN, is a convex relaxation method, but rather than penalizing the ℓ_2 norm of the residual $\mathbf{b}-\mathbf{A}\mathbf{x}$, the Dantzig selector ensures that the residual is weakly correlated with all dictionary atoms.

Finally, we also mention the so-called oracle estimator, which is based both on $\mathbf{b}$ and on the support set Λ_0 of $\mathbf{x}_0$; the support set is assumed to have been provided by an "oracle". The oracle estimator finds the least squares solution among vectors $\mathbf{x}$ whose support coincides with Λ_0. The resulting estimator is given by

$$\hat{\mathbf{x}}_{\mathrm{or}} = \begin{cases} \mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{b} & \text{on the support set } \Lambda_0, \\ \mathbf{0} & \text{elsewhere.} \end{cases} \tag{10}$$

Since Λ_0 is unknown, the oracle estimator cannot be applied in practice. However, it is often used as a gold standard against which the performance of practical estimators can be compared.

IV. Performance under Adversarial Noise

Our problem is to estimate $\mathbf{x}_0$ from noisy measurements $\mathbf{b}$ given by (1). In this section, we discuss the case in which the noise $\mathbf{w}$ is an unknown deterministic vector which satisfies $\|\mathbf{w}\|_2 \leq \varepsilon$. In Section V, we will compare this setting with the results which can be obtained when $\mathbf{w}$ is random.

Typical "stability" results under adversarial noise guarantee that if the mutual coherence μ of $\mathbf{A}$ is sufficiently small, and if $\mathbf{x}_0$ is sufficiently sparse, then the distance between $\hat{\mathbf{x}}_{\mathrm{BP}}$ and $\mathbf{x}_0$ is on the order of the noise magnitude. Consider, for example, the following theorem, which is based on the work of Tropp [3, §IV-C].[1]

Theorem 1 (Tropp): Let $\mathbf{x}_0$ be an unknown deterministic vector with known sparsity $\|\mathbf{x}_0\|_0 = k$, and let $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\|\mathbf{w}\|_2 \leq \varepsilon$. Suppose the mutual coherence μ of the dictionary $\mathbf{A}$ satisfies $k < 1/(3\mu)$. Let $\hat{\mathbf{x}}_{\mathrm{BP}}$ denote a solution of BPDN (7) with regularization parameter $\gamma = 2\varepsilon$. Then, $\hat{\mathbf{x}}_{\mathrm{BP}}$ is unique, the support of $\hat{\mathbf{x}}_{\mathrm{BP}}$ is a subset of the support of $\mathbf{x}_0$, and

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_\infty < \left(3 + \sqrt{\tfrac{3}{2}}\right)\varepsilon \approx 4.22\varepsilon. \tag{11}$$

Results similar to Theorem 1 have also been obtained [4], [11, Th. 1.1], [12, Th. 1.2] for the related ℓ_1-error estimation approach (8). In all cases, unless one considers matrices having a particular structure, the theorems hold under the assumption that $\|\mathbf{x}_0\|_0 = O(1/\mu)$, and the stability result guarantees that the distance between $\hat{\mathbf{x}}_{\mathrm{BP}}$ and $\mathbf{x}_0$ is on the order of the noise power ε. To the best of our knowledge, no stability results are available for the Dantzig selector under this setting.

The adversarial noise results are somewhat disappointing, because one would expect the knowledge that $\mathbf{x}_0$ is sparse to assist in denoising; yet Theorem 1 guarantees only that the distance between each coordinate of $\hat{\mathbf{x}}_{\mathrm{BP}}$ and the corresponding coordinate of $\mathbf{x}_0$ is less than about four times the maximum noise level. However, the fact that no denoising has occurred is a consequence of the problem setting itself, rather than a limitation of the algorithms proposed above. In the adversarial case, even the oracle estimator (10) can only guarantee an estimation error on the order of ε. This is because $\mathbf{w}$ can be chosen so that $\mathbf{w} \in \mathrm{span}(\mathbf{A}_{\Lambda_0})$, in which case projection onto $\mathrm{span}(\mathbf{A}_{\Lambda_0})$, as performed by the oracle estimator, does not remove any portion of the noise.

In conclusion, results in this adversarial context must take into account values of $\mathbf{w}$ which are chosen so as to cause maximal damage to the estimation algorithm. In many practical situations, such a scenario is overly pessimistic. Thus, it is interesting to ask what guarantees can be made about the performance of BPDN under the assumption of random (and thus non-adversarial) noise. This scenario is considered in the next section.

[1] Tropp considers only the case in which the entries of $\mathbf{x}_0$ belong to the set $\{0, \pm 1\}$. However, since the analysis performed in [3, §IV-C] can readily be applied to the general setting considered here, we omit the proof of Theorem 1.

V. Performance under Random Noise

Assume now that we obtain measurements $\mathbf{b}$ as in (1), and that $\mathbf{w}$ is a Gaussian random vector with mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}$. Let us begin by examining the oracle estimator $\hat{\mathbf{x}}_{\rm or}$ of (10). It is readily seen [2] that the MSE of $\hat{\mathbf{x}}_{\rm or}$ is $E\{\|\hat{\mathbf{x}}_{\rm or}-\mathbf{x}_0\|_2^2\} = \sigma^2\,\mathrm{Tr}((\mathbf{A}_{\Lambda_0}^*\mathbf{A}_{\Lambda_0})^{-1})$. It follows from the Gershgorin disc theorem [13] that all eigenvalues of $\mathbf{A}_{\Lambda_0}^*\mathbf{A}_{\Lambda_0}$ are between $1-(k-1)\mu$ and $1+(k+1)\mu$. Therefore, for reasonable sparsity levels, $\mathrm{Tr}((\mathbf{A}_{\Lambda_0}^*\mathbf{A}_{\Lambda_0})^{-1})$ is not much larger than k; for example, if we assume, as in Theorem 1, that $k<1/3\mu$, then $\mathrm{Tr}((\mathbf{A}_{\Lambda_0}^*\mathbf{A}_{\Lambda_0})^{-1}) < \frac{3}{2}k$, and thus the oracle's MSE in this case is no larger than $\frac{3}{2}k\sigma^2$. Considering that the mean power of $\mathbf{w}$ is $n\sigma^2$, it is evident that the oracle estimator has substantially reduced the noise level. One would like to know whether comparable performance gains can be achieved by practical methods, which do not have access to the oracle.

This question was first addressed in the context of the Dantzig selector (9). The result, due to Candès and Tao [2], is derived using the UUP constants (3)–(4). For a given matrix $\mathbf{A}$, however, it is usually intractable to determine the UUP constants. Instead, we obtain the following result by applying Lemma 1 to [2, Th. 1.1].

Theorem 2 (Candès and Tao): Let $\mathbf{x}_0$ be an unknown deterministic vector such that $\|\mathbf{x}_0\|_0 = k$, and let $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is a random noise vector. Assume that

$$k < 1 + \frac{1}{(1+\sqrt{2})\mu} \tag{12}$$

and consider the Dantzig selector (9) with parameter

$$\tau = \sigma\sqrt{2(1+\alpha)\log m} \tag{13}$$

for some constant $\alpha > 0$. Then, with probability exceeding

$$1 - \frac{1}{m^\alpha\sqrt{\pi\log m}}, \tag{14}$$

the Dantzig selector $\hat{\mathbf{x}}_{\rm DS}$ satisfies

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm DS}\|_2^2 \le 2c_1^2(1+\alpha)\cdot\log m\cdot k\sigma^2 \tag{15}$$

where

$$c_1 = \frac{4}{1-\big((1+\sqrt{2})k-1\big)\,\mu}. \tag{16}$$

This theorem is significant because it demonstrates that, while $\hat{\mathbf{x}}_{\rm DS}$ does not quite reach the performance of the oracle estimator, it does come within a constant factor multiplied by $\log m$, with high probability. Note that the $\log m$ factor is an unavoidable result of the fact that the locations of the nonzero elements in $\mathbf{x}_0$ are unknown (see [7, §7.4] and the references therein).

It is clearly of interest to determine whether results similar to Theorem 2 can be obtained for BPDN [8], [9]. Bickel et al. [10] have recently answered this question affirmatively; they showed that, with high probability, BPDN comes within a factor of $C\log m$ of the oracle performance, for a constant C. In fact, their analysis is quite versatile, and simultaneously provides a result for both the Dantzig selector and BPDN. However, the constant C obtained in their BPDN guarantee is always larger than 128, which is considerably weaker than that of Theorem 2.

In the following, we obtain an improved performance guarantee for BPDN. In particular, we demonstrate that, for an appropriate choice of the regularization parameter γ, the squared error of the BPDN estimate is bounded, with high probability, by a small constant times $k\sigma^2\log(m-k)$, and that this constant is lower than that of Theorem 2. We begin by stating the following somewhat more general result, whose proof is found in Appendix II.

Theorem 3: Let $\mathbf{x}_0$ be an unknown deterministic vector with known sparsity $\|\mathbf{x}_0\|_0 = k$, and let $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is a random noise vector. Suppose that[2]

$$k < \frac{1}{3\mu}. \tag{17}$$

[2]As in [3], analogous findings can also be obtained under the weaker requirement $k < 1/(2\mu)$, but the resulting expressions are somewhat more involved.

Then, with probability exceeding

$$\left(1-(m-k)\exp\left(-\frac{\gamma^2}{8\sigma^2}\right)\right)\left(1-e^{-k/7}\right), \tag{18}$$

the solution $\hat{\mathbf{x}}_{\mathrm{BP}}$ of BPDN (7) is unique and satisfies

$$\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq \left(\sigma\sqrt{3}+\tfrac{3}{2}\gamma\right)^2 k. \tag{19}$$

To compare the results for BPDN and the Dantzig selector, we now derive from Theorem 3 a result which holds with a probability on the order of (14). Observe that in order for (18) to be a high probability, we require $\exp(-\gamma^2/(8\sigma^2))$ to be substantially smaller than $1/(m-k)$. This requirement can be used to select a value for the regularization parameter γ. In particular, one requires γ to be at least on the order of $\sqrt{8\sigma^2\log(m-k)}$. However, γ should not be much larger than this value, as this will increase the error bound (19). We propose to use the value

$$\gamma=\sqrt{8\sigma^2(1+\alpha)\log(m-k)} \tag{20}$$

for some fairly small $\alpha>0$. Substituting this value of γ into Theorem 3 yields the following result.

Corollary 1: Under the conditions of Theorem 3, let $\hat{\mathbf{x}}_{\mathrm{BP}}$ be a solution of BPDN (7) with γ given by (20). Then, with probability exceeding

$$\left(1-\frac{1}{(m-k)^{\alpha}}\right)\left(1-e^{-k/7}\right) \tag{21}$$

we have

$$\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq \left(\sqrt{3}+3\sqrt{2(1+\alpha)\log(m-k)}\right)^2 k\sigma^2. \tag{22}$$

Let us examine the probability (21) with which Corollary 1 holds, to verify that it is indeed roughly equal to (14). The expression (21) consists of a product of two terms, both of which converge to 1 as the problem dimensions increase. The right-hand term may seem odd because it appears to favor non-sparse signals; however, this is an artifact of the method of proof, which requires a sufficient number of nonzero coefficients for large number approximations to hold. This right-hand term converges to 1 exponentially and therefore typically has a negligible effect on the overall probability of success; for example, for $k\geq 100$ this term is larger than $1-10^{-6}$.

The left-hand term in (21) tends to 1 polynomially as $m-k$ increases. This is a slightly lower rate than the probability (14) with which the Dantzig selector bound holds; however, this difference is compensated for by a correspondingly lower multiplicative factor of $\log(m-k)$ in the BPDN error bound (22), as opposed to the $\log m$ factor in the Dantzig selector. In any case, for both theorems to hold, m must increase much more quickly than k, so that these differences are negligible.

For large k and $m-k$, Corollary 1 ensures that, with high probability, $\|\hat{\mathbf{x}}_{\mathrm{BP}}-\mathbf{x}_0\|_2^2$ is no larger than a constant multiplied by $k\sigma^2\log(m-k)$. Up to a multiplicative constant, this error bound is essentially identical to the result (15) for the Dantzig selector. As we have seen, the probabilities with which these bounds hold are likewise almost identical. However, the constants involved in the BPDN, as demonstrated by Corollary 1, are substantially lower than those previously known for the Dantzig selector. To see this, consider a situation in which $k=1/(4\mu)$. In this case, for large k, the bound (15) on the Dantzig selector rapidly converges to

$$\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 \leq 203.6(1+\alpha)\cdot\log m\cdot k\sigma^2. \tag{23}$$

By comparison, the performance of BPDN in the same setting, as bounded by Corollary 1, is

$$\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq 18(1+\alpha)\cdot\log(m-k)\cdot k\sigma^2 \tag{24}$$

which is over 10 times lower. This improvement is not merely a result of the particular choice of k or μ. Indeed, the multiplicative factor of 18 which appeared in the BPDN bound (24) holds for large k with any value of μ, as long as $k<1/(3\mu)$; whereas it can be seen from (15)–(16) that the multiplicative factor of the Dantzig selector is always larger than 32.

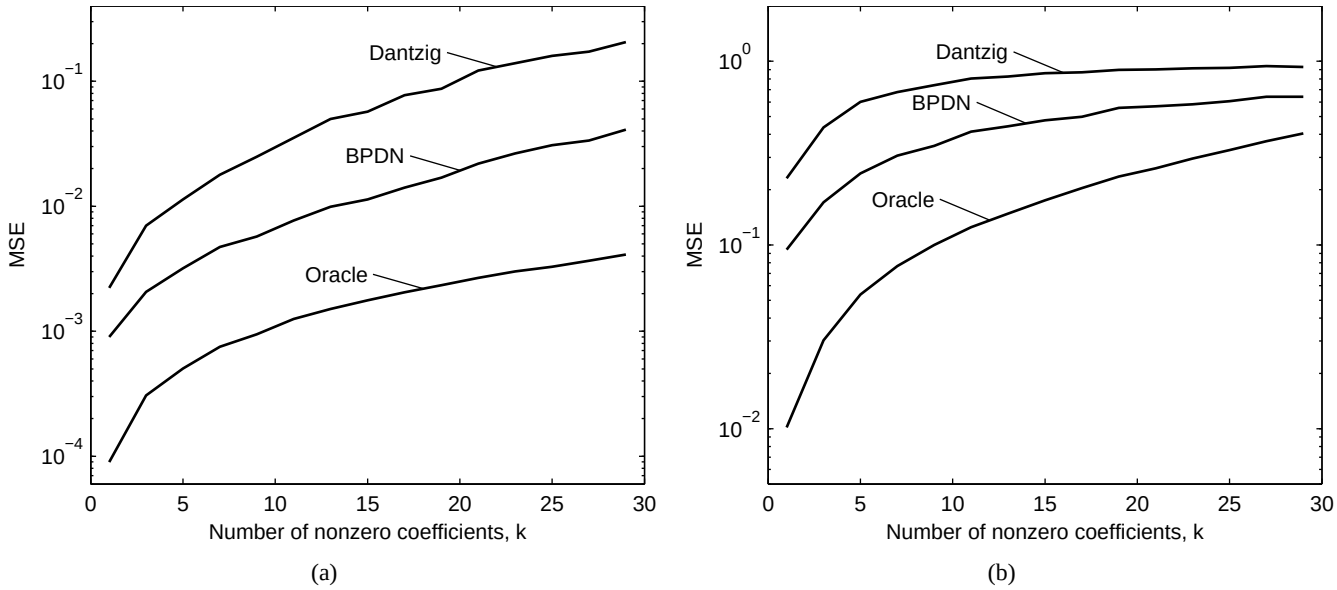

Fig. 1. MSE of the Dantzig selector (9), BPDN (7), and the oracle estimator (10), for various sparsity levels. The standard deviation of the noise is $\sigma = 0.01$ in (a) and $\sigma = 0.1$ in (b).

VI. Numerical Results

As we have seen, the performance guarantees of BPDN are better than the corresponding results for the Dantzig selector. However, since both bounds are far from tight, it is relevant to ask whether this gain also appears in the actual MSE achieved by the two estimators. We now address this question by simulation.

Several researchers have compared the performance of BPDN and the Dantzig selector numerically [8], [9], [14]. While the results of these analyses were mixed, they can be summarized by saying that the two approaches often yield similar results, perhaps with an occasional advantage for BPDN. These comparisons typically employ some data-driven method for selecting the regularization parameters γ of (7) and τ of (9). By contrast, our aim is to ascertain the applicability of the performance guarantees of the Dantzig selector and BPDN. Consequently, we will analyze the performance of the two approaches when the regularization is chosen as recommended by Theorem 2 and Corollary 1.

To perform the comparison, 100 random dictionaries $\mathbf{A}$ and 100 random parameter vectors $\mathbf{x}_0$ were constructed. Each dictionary consisted of $m = 200$ atoms of length $n = 100$; the dictionary entries were chosen from a zero-mean Gaussian IID distribution, and were then normalized to ensure that the ℓ_2 norm of each atom is 1. The support of $\mathbf{x}_0$ was selected uniformly at random, and the nonzero elements of the parameter were Gaussian IID variables with mean 0 and standard deviation 1.

For each dictionary and parameter vector, 10 noisy measurements $\mathbf{b}$ of the form (1) were generated by adding Gaussian random noise with mean 0 and standard deviation σ. The parameter $\mathbf{x}_0$ was then estimated from $\mathbf{b}$ using BPDN (7), the Dantzig selector (9), and the oracle estimator (10). The optimization problems (7) and (9) were solved in Matlab using the CVX package [15]. The regularization parameter used for BPDN was (20), as recommended by Corollary 1, and the threshold used for the Dantzig selector was (13), as recommended by Theorem 2 and [2]. In both cases, a value of $\alpha = 1$ was chosen; this ensures that the probability of success of BPDN and the Dantzig selector, as given by Theorems 2 and 3, has the same order of magnitude for large problems. Finally, the MSE of each estimate was calculated and averaged over all realizations of the noise, parameters and dictionaries. The experiment was repeated for 15 support sizes k in the range $1 \leq k \leq 30$ and for two different noise standard deviations, $\sigma = 0.01$ and $\sigma = 0.1$. The results are plotted in Fig. 1.

Note that for the given problem dimensions, the mutual coherence is at best 0.07, so that Theorems 2 and 3 apply only when $\mathbf{x}_0$ is quite sparse. Nevertheless, for most values of $\mathbf{x}_0$, both algorithms continue to perform well under these settings [6]. This is illustrated in Fig. 1, in which the average MSE of each estimator is plotted. It is also evident that, in the scenario under consideration, the MSE of BPDN is considerably lower than that of the Dantzig selector, even when the results of Section V no longer apply. Thus, the choice of the regularization parameter (20)

for BPDN appears to be a useful rule of thumb, even beyond the region of applicability of Corollary 1.

In a separate set of simulations, we compared the performance of BPDN and the Dantzig estimator for several deterministic dictionaries, including an overcomplete DCT [16] and a two-ortho dictionary $[\mathbf{I}, \mathbf{H}]$ in which $\mathbf{I}$ is the identity matrix and $\mathbf{H}$ is the Hadamard matrix. The results obtained in these cases were similar in essence to those of Fig. 1, so we refrain from describing these experiments in detail.

VII. Conclusion

In this paper, we analyzed the performance of BPDN for estimating a deterministic sparse parameter vector $\mathbf{x}_0$ under random noise. Specifically, we examined the distance between the BPDN estimate $\hat{\mathbf{x}}_{\mathrm{BP}}$ and $\mathbf{x}_0$. This distance was shown to be bounded, with high probability, by the error of the oracle estimator multiplied by $C \log m$, where C is a small constant, on the order of 18. This analysis improves a previous bound [10] in which it was shown that $C \geq 128$. Our performance guarantee is better than that of the Dantzig selector, for which $C \geq 32$. The result also provides a rule of thumb for selecting the BPDN regularization parameter, which appears to function well even when the analytic guarantees no longer hold.

Appendix I
Proof of Lemma 1

By Gershgorin's disc theorem [13], all eigenvalues of $\mathbf{A}_\Lambda^* \mathbf{A}_\Lambda$ are between $1 - (k-1)\mu$ and $1 + (k-1)\mu$. Combining this with the fact that, for all $\mathbf{y}$,

$$\lambda_{\min}(\mathbf{A}_\Lambda^* \mathbf{A}_\Lambda)\|\mathbf{y}\|_2^2 \leq \|\mathbf{A}_\Lambda \mathbf{y}\|_2^2 \leq \lambda_{\max}(\mathbf{A}_\Lambda^* \mathbf{A}_\Lambda)\|\mathbf{y}\|_2^2, \tag{25}$$

we obtain (5). Next, to demonstrate (6), observe that

$$\left|\mathbf{y}_1^* \mathbf{A}_{\Lambda_1}^* \mathbf{A}_{\Lambda_2} \mathbf{y}_2\right| \leq |\mathbf{y}_1^*| \cdot \left|\mathbf{A}_{\Lambda_1}^* \mathbf{A}_{\Lambda_2}\right| \cdot |\mathbf{y}_2| \tag{26}$$

where the absolute value of a matrix or vector is taken elementwise. Since $\mathbf{A}_{\Lambda_1}^* \mathbf{A}_{\Lambda_2}$ is a submatrix of $\mathbf{A}^*\mathbf{A}$ which does not contain any of the diagonal elements of $\mathbf{A}^*\mathbf{A}$, it follows that each element in $\mathbf{A}_{\Lambda_1}^* \mathbf{A}_{\Lambda_2}$ is smaller in absolute value than μ. Thus

$$\left|\mathbf{y}_1^* \mathbf{A}_{\Lambda_1}^* \mathbf{A}_{\Lambda_2} \mathbf{y}_2\right| \leq \mu \, |\mathbf{y}_1^*| \, \mathbb{1}\mathbb{1}^* \, |\mathbf{y}_2| = \mu \|\mathbf{y}_1\|_1 \|\mathbf{y}_2\|_1 \tag{27}$$

where $\mathbb{1}$ indicates a vector of ones. Using the fact that $\|\mathbf{y}\|_1 \leq \sqrt{k}\|\mathbf{y}\|_2$ for any k-vector $\mathbf{y}$, we obtain

$$\left|\mathbf{y}_1^* \mathbf{A}_{\Lambda_1}^* \mathbf{A}_{\Lambda_2} \mathbf{y}_2\right| \leq \mu \sqrt{k_1 k_2} \|\mathbf{y}_1\|_2 \|\mathbf{y}_2\|_2, \tag{28}$$

which implies that θ_{k_1,k_2} satisfies (6).

Appendix II
Proof of Theorem 3

The proof is based closely on the work of Tropp [3]. From the triangle inequality,

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2 \leq \|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{or}}\|_2 + \|\hat{\mathbf{x}}_{\mathrm{or}} - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2 \tag{29}$$

where $\hat{\mathbf{x}}_{\mathrm{or}}$ is the oracle estimator defined in (10). Our goal is to separately bound the two terms on the right-hand side of (29). Indeed, as we will see, the two constants $\sigma\sqrt{3}$ and $\frac{3}{2}\gamma$ in (19) arise, respectively, from the two terms in (29).

Beginning with the term $\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{or}}\|_2$, let $\mathbf{x}_{0,\Lambda}$ denote the k-vector containing the elements of $\mathbf{x}_0$ indexed by Λ_0, and similarly, let $\hat{\mathbf{x}}_{\mathrm{or},\Lambda}$ denote the corresponding subvector of $\hat{\mathbf{x}}_{\mathrm{or}}$. We then have

$$\begin{aligned} \mathbf{x}_{0,\Lambda} - \hat{\mathbf{x}}_{\mathrm{or},\Lambda} &= \mathbf{x}_{0,\Lambda} - \mathbf{A}_{\Lambda_0}^\dagger (\mathbf{A}\mathbf{x}_0 + \mathbf{w}) \\ &= \mathbf{x}_{0,\Lambda} - \mathbf{A}_{\Lambda_0}^\dagger (\mathbf{A}_{\Lambda_0}\mathbf{x}_{0,\Lambda} + \mathbf{w}) \\ &= \mathbf{A}_{\Lambda_0}^\dagger \mathbf{w}, \end{aligned} \tag{30}$$

where we have used the fact that $\mathbf{A}_{\Lambda_0}$ has full column rank, which is a consequence [17] of the condition (17). Thus, $\mathbf{x}_{0,\Lambda} - \hat{\mathbf{x}}_{\mathrm{or},\Lambda}$ is a Gaussian random vector with mean $\mathbf{0}$ and covariance $\sigma^2 \mathbf{A}_{\Lambda_0}^\dagger \mathbf{A}_{\Lambda_0}^{\dagger *} = \sigma^2 (\mathbf{A}_{\Lambda_0}^* \mathbf{A}_{\Lambda_0})^{-1}$.

For future use, we note that the cross-correlation between $\mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{w}$ and $(\mathbf{I}-\mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^{\dagger})\mathbf{w}$ is

$$\begin{aligned} E\Big\{\mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{w}\mathbf{w}^*(\mathbf{I}-\mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^{\dagger})^*\Big\} &= \sigma^2\mathbf{A}_{\Lambda_0}^{\dagger}(\mathbf{I}-\mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^{\dagger})^* \\ &= \mathbf{0}, \end{aligned} \tag{31}$$

where we have used the fact [18, Th. 1.2.1] that for any matrix $\mathbf{M}$

$$\mathbf{M}^{\dagger}\mathbf{M}^{\dagger *}\mathbf{M}^* = (\mathbf{M}^*\mathbf{M})^{\dagger}\mathbf{M}^* = \mathbf{M}^{\dagger}. \tag{32}$$

Since $\mathbf{w}$ is Gaussian, it follows that $\mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{w}$ and $(\mathbf{I}-\mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^{\dagger})\mathbf{w}$ are statistically independent. Furthermore, because $\mathbf{x}_{0,\Lambda}-\hat{\mathbf{x}}_{\mathrm{or},\Lambda}$ depends on $\mathbf{w}$ only through $\mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{w}$, we conclude that

$$\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{or}} \text{ is statistically independent of } (\mathbf{I}-\mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^{\dagger})\mathbf{w}. \tag{33}$$

We now wish to bound the probability that $\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 > 3k\sigma^2$. Let $\mathbf{z}$ be a normalized Gaussian random variable, $\mathbf{z}\sim N(\mathbf{0},\mathbf{I}_k)$. Then

$$\begin{aligned} &\Pr\left\{\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 > 3k\sigma^2\right\} \\ &= \Pr\left\{\left\|\sigma(\mathbf{A}_{\Lambda_0}^*\mathbf{A}_{\Lambda_0})^{-1/2}\mathbf{z}\right\|_2^2 \geq 3k\sigma^2\right\} \\ &\leq \Pr\left\{\left\|(\mathbf{A}_{\Lambda_0}^*\mathbf{A}_{\Lambda_0})^{-1/2}\right\|^2\|\mathbf{z}\|_2^2 \geq 3k\right\} \end{aligned} \tag{34}$$

where $\|\mathbf{M}\|$ denotes the maximum singular value of the matrix $\mathbf{M}$. Thus, $\|(\mathbf{A}_{\Lambda_0}^*\mathbf{A}_{\Lambda_0})^{-1/2}\| = 1/s_{\min}$, where $s_{\min}$ is the minimum singular value of $\mathbf{A}_{\Lambda_0}$. From the Gershgorin disc theorem [13, p. 320], it follows that $s_{\min}\geq\sqrt{1-(k-1)\mu}$. Using (17), this can be simplified to $s_{\min}\geq\sqrt{2/3}$, and therefore

$$\left\|(\mathbf{A}_{\Lambda_0}^*\mathbf{A}_{\Lambda_0})^{-1/2}\right\| \leq \sqrt{\frac{3}{2}}. \tag{35}$$

Combining with (34) yields

$$\Pr\left\{\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 > 3k\sigma^2\right\} \leq \Pr\left\{\|\mathbf{z}\|_2^2 \geq 2k\right\}. \tag{36}$$

Observe that $\|\mathbf{z}\|_2^2$ is the sum of k independent normalized Gaussian random variables. The right-hand side of (36) is therefore $1-F_{\chi_k^2}(2k)$, where $F_{\chi_k^2}(\cdot)$ is the cumulative distribution function of the χ^2 distribution with k degrees of freedom. Using the formula [19, §16.3] for $F_{\chi_k^2}(\cdot)$, we have

$$\Pr\left\{\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 > 3k\sigma^2\right\} \leq Q\left(\tfrac{1}{2}k,k\right) \tag{37}$$

where $Q(a,z)$ is the regularized Gamma function

$$Q(a,z) = \frac{\int_z^{\infty}t^{a-1}e^{-t}dt}{\int_0^{\infty}t^{a-1}e^{-t}dt}. \tag{38}$$

$Q\left(\frac{1}{2}k,k\right)$ decays exponentially as $k\to\infty$, and it can be seen that

$$Q\left(\tfrac{1}{2}k,k\right) < e^{-k/7} \quad \text{for all } k. \tag{39}$$

We thus conclude that the event

$$\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 \leq 3k\sigma^2 \tag{40}$$

occurs with probability no smaller than $1-e^{-k/7}$. Note that the same technique can be applied to obtain bounds on the probability that $\|\mathbf{x}_0-\hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 > \alpha k\sigma^2$, for any $\alpha>\frac{2}{3}$. The only difference will be the rate of exponential decay in (39). However, the distance between $\mathbf{x}_0$ and $\hat{\mathbf{x}}_{\mathrm{or}}$ is usually small compared with the distance between $\hat{\mathbf{x}}_{\mathrm{or}}$ and $\hat{\mathbf{x}}_{\mathrm{BP}}$, so that such an approach does not significantly affect the overall result.

The above calculations provided a bound on the first term in (29). To address the second term $\|\hat{\mathbf{x}}_{\mathrm{or}}-\hat{\mathbf{x}}_{\mathrm{BP}}\|_2$, define the random event

$$G:\ \max_i\left|\mathbf{a}_i^*(\mathbf{I}-\mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^{\dagger})\mathbf{b}\right| \leq \tfrac{1}{2}\gamma \tag{41}$$

where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$. It is shown in [3, App. IV-A] that

$$\Pr\{G\} \geq 1 - (m-k)\exp\left(-\frac{\gamma^2}{8\sigma^2}\right). \tag{42}$$

If G indeed occurs, then the portion of the measurements $\mathbf{b}$ which do not belong to the range space of $\mathbf{A}_{\Lambda_0}$ are small, and consequently it has been shown [3, Cor. 9] that, in this case, the solution $\hat{\mathbf{x}}_{\mathrm{BP}}$ to (7) is unique, the support of $\hat{\mathbf{x}}_{\mathrm{BP}}$ is a subset of Λ_0, and

$$\|\hat{\mathbf{x}}_{\mathrm{BP}} - \hat{\mathbf{x}}_{\mathrm{or}}\|_\infty \leq \tfrac{3}{2}\gamma. \tag{43}$$

Since both $\hat{\mathbf{x}}_{\mathrm{BP}}$ and $\hat{\mathbf{x}}_{\mathrm{or}}$ are nonzero only in Λ_0, this implies that

$$\|\hat{\mathbf{x}}_{\mathrm{BP}} - \hat{\mathbf{x}}_{\mathrm{or}}\|_2 \leq \tfrac{3}{2}\gamma\sqrt{k}. \tag{44}$$

The event G depends on the random variable $\mathbf{w}$ only through $(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger)\mathbf{w}$. Thus, it follows from (33) that G is statistically independent of the event (34). The probability that both events occur simultaneously is therefore given by the product of their respective probabilities. In other words, with probability exceeding (18), both (44) and (40) hold. Using (29) completes the proof of the theorem.

REFERENCES

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[2] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[3] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[4] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[5] J. A. Tropp, "Norms of random submatrices and sparse approximation," *C. R. Acad. Sci. Paris, Ser. I*, vol. 346, no. 23-24, pp. 1271–1274, 2008.

[6] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Ann. Statist.*, 2009, to appear.

[7] E. J. Candès, "Modern statistical estimation via oracle inequalities," *Acta Numerica*, pp. 1–69, 2006.

[8] B. Efron, T. Hastie, and R. Tibshirani, "Discussion: The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2358–2364, 2007.

[9] E. Candes and T. Tao, "Rejoinder: The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2392–2404, 2007.

[10] P. J. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, 2008, to appear. [Online]. Available: http://arxiv.org/abs/0801.1095

[11] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. LIX, pp. 1207–1223, 2006.

[12] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, vol. 346, pp. 589–592, 2008. [Online]. Available: http://www.acm.caltech.edu/~emmanuel/papers/RIP.pdf

[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996.

[14] N. Meinshausen, G. Rocha, and B. Yu, "Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig," *Ann. Statist.*, vol. 35, no. 6, pp. 2373–2384, 2007.

[15] M. Grant and S. Boyd. (2009, Feb.) CVX: Matlab software for disciplined convex programming. [Online]. Available: http://stanford.edu/~boyd/cvx

[16] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[17] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, March 4, 2003.

[18] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*. London, UK: Pitman, 1979.

[19] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 6th ed. London: Edward Arnold, 1994, vol. 1.