Guaranteed Sparse Recovery under Linear Transformation

Ji Liu JI-LIU@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

Lei Yuan LEI.YUAN@ASU.EDU
Jieping Ye JIEPING.YE@ASU.EDU
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA

Abstract

We consider the following signal recovery problem: given a measurement matrix $\Phi \in \mathbb{R}^{n\times p}$ and a noisy observation vector $c \in \mathbb{R}^n$ constructed from $c = \Phi\theta^* + \epsilon$ where $\epsilon \in \mathbb{R}^n$ is the noise vector whose entries follow i.i.d. centered sub-Gaussian distribution, how to recover the signal θ^* if $D\theta^*$ is sparse under a linear transformation $D \in \mathbb{R}^{m\times p}$? One natural method using convex optimization is to solve the following problem:

$$\min_{\theta} \frac{1}{2}\|\Phi\theta - c\|^2 + \lambda\|D\theta\|_1.$$

This paper provides an upper bound of the estimate error and shows the consistency property of this method by assuming that the design matrix Φ is a Gaussian random matrix. Specifically, we show 1) in the noiseless case, if the condition number of D is bounded and the measurement number $n \geq \Omega(s\log(p))$ where s is the sparsity number, then the true solution can be recovered with high probability; and 2) in the noisy case, if the condition number of D is bounded and the measurement increases faster than $s\log(p)$, that is, $s\log(p) = o(n)$, the estimate error converges to zero with probability 1 when p and s go to infinity. Our results are consistent with those for the special case $D = \mathbf{I}_{p\times p}$ (equivalently LASSO) and improve the existing analysis. The condition number of D plays a critical role in our analysis. We consider the condition numbers in two cases including the fused LASSO and the random graph: the condition number in the fused LASSO case is bounded by a constant, while the condition number in the random graph case is bounded with high probability if $\frac{m}{p}$ (i.e., $\frac{\#\text{edge}}{\#\text{vertex}}$) is larger than a certain constant. Numerical simulations are consistent with our theoretical results.

Appearing in *Proceedings of the 30^{th} International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013.

1. Introduction

The sparse signal recovery problem has been well studied recently from the theory aspect to the application aspect in many areas including compressive sensing (Candès & Plan, 2009; Candès & Tao, 2007), statistics (Ravikumar et al., 2008; Bunea et al., 2007; Koltchinskii & Yuan, 2008; Lounici, 2008; Meinshausen et al., 2006), machine learning (Zhao & Yu, 2006; Zhang, 2009b; Wainwright, 2009; Liu et al., 2012), and signal processing (Romberg, 2008; Donoho et al., 2006; Zhang, 2009a). The key idea is to use the ℓ_1 norm to relax the ℓ_0 norm (the number of nonzero entries). This paper considers a specific type of sparse signal recovery problems, that is, the signal is assumed to be sparse under a linear transformation D. It includes the well-known fused LASSO (Tibshirani et al., 2005) as a special case. The theoretical property of such problem has not been well understood yet, although it has achieved success in many applications (Chan, 1998; Tibshirani et al., 2005; Candès et al., 2006; Sharpnack et al., 2012). Formally, we define the problem as follows: given a measurement matrix $\Phi \in \mathbb{R}^{n\times p}$ ($p \gg n$) and a noisy observation vector $c \in \mathbb{R}^n$ constructed from $c = \Phi\theta^* + \epsilon$ where $\epsilon \in \mathbb{R}^n$ is the noise vector whose entries follow i.i.d. centered sub-Gaussian distribution[1], how to recover the signal θ^* if $D\theta^*$ is sparse where $D \in \mathbb{R}^{m\times p}$ is a constant matrix dependent on the specific application[2]? A natural model for such type

[1]Note that this "identical distribution" assumption can be removed; see Zhang (2009a). For simplification of analysis, we enforce this condition throughout this paper.

[2]We study the most general case of D, and thus our analysis is applicable for both $m \geq p$ or $m \leq p$.

of sparsity recovery problems is:

$$\min_{\theta} : \frac{1}{2}\|\Phi\theta - c\|^2 + \lambda\|D\theta\|_0. \quad (1)$$

The least square term is from the sub-Gaussian noise assumption and the second term is due to the sparsity requirement. Since this combinatorial optimization problem is NP-hard, the conventional ℓ_1 relaxation technique can be applied to make it tractable, resulting in the following convex model:

$$\min_{\theta} : \frac{1}{2}\|\Phi\theta - c\|^2 + \lambda\|D\theta\|_1. \quad (2)$$

Such model includes many well-known sparse formulations as special cases:

- The fused LASSO (Tibshirani et al., 2005; Friedman et al., 2007) solves

$$\min_{\theta} : \frac{1}{2}\|\Phi\theta - c\|^2 + \lambda_1\|\theta\|_1 + \lambda_2\|Q\theta\|_1 \quad (3)$$

where the total variance matrix $Q \in \mathbb{R}^{(p-1)\times p}$ is defined as $Q = [\mathbf{I}_{(p-1)\times(p-1)}; \mathbf{0}_{p-1}] - [\mathbf{0}_{p-1}; \mathbf{I}_{(p-1)\times(p-1)}]$, that is,

$$Q = \begin{bmatrix} 1 & -1 & 0 & ... & 0 \\ 0 & 1 & -1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & 1 & -1 \end{bmatrix}.$$

One can write Eq. (3) in the form of Eq. (2) by letting $\lambda = 1$ and D be the conjunction of the identity matrix and the total variance matrix, that is,

$$D = \begin{bmatrix} \lambda_1\mathbf{I}_{p\times p} \\ \lambda_2 Q \end{bmatrix}.$$

- The general K dimensional changing point detection problem (Candès et al., 2006) can be expressed by

$$\min_{\theta} : \frac{1}{2}\sum_{(i_1,i_2,\cdots,i_K)\in S}(\theta_{i_1,i_2,\cdots,i_K} - c_{i_1,i_2,\cdots,i_K})^2 + \lambda\sum_{i_1=1}^{I_1-1}\sum_{i_2=1}^{I_2-1}\cdots\sum_{i_K=1}^{I_K-1}(|\theta_{i_1,i_2,\cdots,i_K} - \theta_{i_1+1,i_2,\cdots,i_K}| + \cdots + |\theta_{i_1,i_2,\cdots,i_K} - \theta_{i_1,i_2,\cdots,i_K+1}|) \quad (4)$$

where $\theta \in \mathbb{R}^{I_1\times I_2\times\cdots I_K}$ is a K dimensional tensor with a stepwise structure and S is the set of indices. The second term is used to measure the total variance. The changing point is defined as the point where the signal changes. One can properly define D to rewrite Eq. (4) in the form of Eq. (2). In addition, if the structure of the signal is piecewise constant, then one can replace the second term by

$$\lambda\sum_{i_1=2}^{I_1-1}\sum_{i_2=2}^{I_2-1}\cdots\sum_{i_K=2}^{I_K-1}(|2\theta_{i_1,i_2,\cdots,i_K} - \theta_{i_1+1,i_2,\cdots,i_K} - \theta_{i_1-1,i_2,\cdots,i_K}| + \cdots + |2\theta_{i_1,i_2,\cdots,i_K} - \theta_{i_1,i_2,\cdots,i_K+1} - \theta_{i_1,i_2,\cdots,i_K-1}|).$$

It can be written in the form of Eq. (2) as well.

- The second term of (4), that is, the total variance, is defined as the sum of differences between two neighboring entries (or nodes). A graph can generalize this definition by using edges to define neighboring entries rather than entry indexes. Let $G(V, E)$ be a graph. One has

$$\min_{\theta\in\mathbb{R}^{|V|}} : \frac{1}{2}\|\Phi\theta - c\|^2 + \lambda\sum_{(i,j)\in E}|\theta_i - \theta_j|, \quad (5)$$

where $\sum_{(i,j)\in E}|\theta_i - \theta_j|$ defines the total variance over the graph G. The k^{th} edge between nodes i and j corresponds to the k^{th} row of the matrix $D \in \mathbb{R}^{|E|\times|V|}$ with zero at all entries except $D_{ki} = 1$ and $D_{kj} = -1$. Taking $\Phi = \mathbf{I}_{p\times p}$, one obtains the edge LASSO (Sharpnack et al., 2012).

This paper studies the theoretical properties of problem (2) by providing an upper bound of the estimate error, that is, $\|\hat{\theta} - \theta^*\|$ where $\hat{\theta}$ denotes the estimation. The consistency property of this model is shown by assuming that the design matrix Φ is a Gaussian random matrix. Specifically, we show 1) in the noiseless case, if the condition number of D is bounded and the measurement number $n \geq \Omega(s\log(p))$ where s is the sparsity number, then the true solution can be recovered under some mild conditions with high probability; and 2) in the noisy case, if the condition number of D is bounded and the measurement number increases faster than $s\log(p)$, that is, $n = O(s\log(p))$, then the estimate error converges to zero with probability 1 under some mild conditions when p goes to infinity. Our results are consistent with those for the special case $D = \mathbf{I}_{p\times p}$ (equivalently LASSO) and improve the existing analysis in (Candès et al., 2011; Vaiter et al., 2013). To the best of our knowledge, this is the first work that establishes the consistency properties for the general problem (2). The condition number of D plays a critical role in our analysis. We consider the condition numbers in two cases including the fused LASSO and the random graph: the condition number in the fused LASSO case is bounded by a constant, while the condition number in the random graph case is bounded with high probability if $\frac{m}{p}$ (that is, $\frac{\#\text{edge}}{\#\text{vertex}}$) is larger than a certain constant. Numerical simulations are consistent with our theoretical results.

1.1. Notations and Assumptions

Define

$$\rho^{+}_{\Psi,Y}(l_1,l_2) = \max_{h\in\mathbb{R}^{l_1}\times\mathcal{H}(Y,l_2)} \frac{\|\Psi h\|^2}{\|h\|^2},$$
$$\rho^{-}_{\Psi,Y}(l_1,l_2) = \min_{h\in\mathbb{R}^{l_1}\times\mathcal{H}(Y,l_2)} \frac{\|\Psi h\|^2}{\|h\|^2},$$

where l_1 and l_2 are nonnegative integers, Y is the dictionary matrix, and $\mathcal{H}(Y,l_2)$ is the union of all subspaces spanned by l_2 columns of Y:

$$\mathcal{H}(Y,l_2) = \{Yv \mid \|v\|_0 \le l_2\}.$$

Note that the length of h is the sum of l_1 and the dimension of the subspace $\mathcal{H}$ (which is in general not equal to l_2). The definition of $\rho^{+}_{\Psi,Y}(l_1,l_2)$ and $\rho^{-}_{\Psi,Y}(l_1,l_2)$ is inspired by the D-RIP constant (Candès et al., 2011). Recall that the D-RIP constant δ_d is defined by the smallest quantity such that

$$(1-\delta_d)\|h\|^2 \le \|\Psi h\|^2 \le (1+\delta_d)\|h\|^2 \quad \forall\, h\in\mathcal{H}(Y,l_2).$$

One can verify that $\delta_d = \max\{\rho^{+}_{\Psi,Y}(0,l_2)-1, 1-\rho^{-}_{\Psi,Y}(0,l_2)\}$ if Ψ satisfies the D-RIP condition in terms of the sparsity l_2 and the dictionary Y. Denote $\rho^{+}_{\Psi,Y}(0,l_2)$ and $\rho^{-}_{\Psi,Y}(0,l_2)$ as $\rho^{+}_{\Psi,Y}(l_2)$ and $\rho^{-}_{\Psi,Y}(l_2)$ respectively for short.

Denote the compact singular value decomposition (SVD) of D as $D = U\Sigma V_\beta^T$. Let $Z = U\Sigma$ and its pseudo-inverse be $Z^+ = \Sigma^{-1}U^T$. One can verify that $Z^+Z = I$. $\sigma_{\min}(D)$ denotes the minimal nonzero singular value of D and $\sigma_{\max}(D)$ denotes the maximal one, that is, the spectral norm $\|D\|$. One has $\sigma_{\min}(D) = \sigma_{\min}(Z) = \sigma^{-1}_{\max}(Z^+)$ and $\sigma_{\max}(D) = \sigma_{\max}(Z) = \sigma^{-1}_{\min}(Z^+)$. Define

$$\kappa := \frac{\sigma_{\max}(D)}{\sigma_{\min}(D)} = \frac{\sigma_{\max}(Z)}{\sigma_{\min}(Z)}.$$

Let T_0 be the support set of $D\theta^*$, that is, a subset of $\{1,2,\cdots,m\}$, with $s := |T_0|$. Denote T_0^c as its complementary index set with respect to $\{1,2,\cdots,m\}$. Without loss of generality, we assume that D does not contain zero rows. Assume that $c = \Phi\theta + \epsilon$ where $\epsilon\in\mathbb{R}^n$ and all entries ϵ_i's are i.i.d. centered sub-Gaussian random variables with sub-Gaussian norm Δ (Readers who are not familiar with the sub-Gaussian norm can treat Δ as the standard derivation in Gaussian random variable). In discussing the dimensions of the problem and how they are related to each other in the limit (as n and p both approach ∞), we make use of order notation. If α and β are both positive quantities that depend on the dimensions, we write $\alpha = O(\beta)$ if α can be bounded by a fixed multiple of β for all sufficiently large dimensions. We write $\alpha = o(\beta)$ if for *any* positive constant $\phi > 0$, we have $\alpha \le \phi\beta$ for all sufficiently large dimensions. We write $\alpha = \Omega(\beta)$ if both $\alpha = O(\beta)$ and $\beta = O(\alpha)$. Throughout this paper, a Gaussian random matrix means that all entries follow i.i.d. standard Gaussian distribution $\mathcal{N}(0,1)$.

1.2. Related Work

Candès et al. (2011) proposed the following formulation to solve the problem in this paper:

$$\min_\theta : \|D\theta\|_1 \quad s.t.: \ \|\Phi\theta - c\| \le \varepsilon, \tag{6}$$

where $D \in \mathbb{R}^{m\times p}$ is assumed to have orthogonal columns and ε is taken as the upper bound of $\|\epsilon\|$. They showed that the estimate error is bounded by $C_0\varepsilon + C_1\|(D\theta^*)_{T^c}\|_1/\sqrt{|T|}$ with high probability if $\sqrt{n}\Phi \in \mathbb{R}^{n\times p}$ is a Gaussian random matrix[3] with $n \ge \Omega(s\log m)$, where C_0 and C_1 are two constants. Letting $T = T_0$ and $\varepsilon = \|\epsilon\|$, the error bound turns out to be $C_0\|\epsilon\|$. This result shows that in the noiseless case, with high probability, the true signal can be exactly recovered. In the noisy case, assume that ϵ_i's $(i = 1,\cdots,n)$ are i.i.d centered sub-Gaussian random variables, which implies that $\|\epsilon\|^2$ is bounded by $\Omega(n)$ with high probability. Note that since the measurement matrix Φ is scaled by $1/\sqrt{n}$ from the definition of "Gaussian random matrix" in (Candès et al., 2011), the noise vector should be corrected similarly. In other words, $\|\epsilon\|^2$ should be bounded by $\Omega(1)$ rather than $\Omega(n)$, which implies that the estimate error in (Candès et al., 2011) converges to a constant asymptotically.

Nama et al. (2012) considered the noiseless case and analyzed the formulation

$$\min_\theta : \|D\theta\|_1 \quad s.t.: \ \Phi\theta = c, \tag{7}$$

assuming all rows of D to be in the general position, that is, any p rows of D are linearly independent, which is violated by the fused LASSO. An sufficient condition was proposed to recover the true signal θ^* using the cosparse analysis.

Vaiter et al. (2013) also considered the formulation in Eq. (2) but mainly gave robustness analysis for this model using the cosparse technique. A sufficient condition [different from Nama et al. (2012)] to exactly recover the true signal was given in the noiseless case. In the noisy case, they took λ to be a value proportional to $\|\epsilon\|$ and proved that the estimate error is bounded by $\Omega(\|\epsilon\|)$ under certain conditions. However, they did not consider the Gaussian ensembles for Φ; see (Vaiter et al., 2013, Section 3.B).

[3] Note that the "Gaussian random matrix" defined in (Candès et al., 2011) is slightly different from ours. In (Candès et al., 2011), $\Phi\in\mathbb{R}^{n\times p}$ is a Gaussian random matrix if each entry of Φ is generated from $\mathcal{N}(0,1/n)$. Please refer to Section 1.5 in (Candès et al., 2011). Here we only restate the result in (Candès et al., 2011) by using our definition for Gaussian random matrices.

The fused LASSO, a special case of Eq. (2), was also studied recently. The sufficient condition of detecting jumping points is given by Kolar et al. (2009). A special fused LASSO formulation was considered by Rinaldo (2009) in which Φ was set to be the identity matrix and D to be the combination of the identity matrix and the total variance matrix. Sharpnack et al. (2012) proposed and studied the edge LASSO by letting Φ be the identity matrix and D be the matrix corresponding to the edges of a graph.

1.3. Organization

The remaining of this paper is organized as follows. To build up a uniform analysis framework, we simplify the formulation (2) in Section 2. The main result is presented in Section 3. Section 4 analyzes the value of an important parameter in our main results in two cases: the fused LASSO and the random graph. The numerical simulation is presented to verify the relationship between the estimate error and the condition number in Section 5. We conclude this paper in Section 6. All proofs are provided in the long version of this paper (Liu et al., 2013).

2. Simplification

As highlighted by Vaiter et al. (2013), the analysis for a wide $D \in \mathbb{R}^{m\times p}$ (that is, $p > m$) significantly differs from a tall D (that is, $p < m$). To build up a uniform analysis framework, we use the singular value decomposition (SVD) of D to simplify Eq. (2), which leads to an equivalent formulation.

Consider the compact SVD of D: $D = U\Sigma V_\beta^T$ where $U \in \mathbb{R}^{n\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$($r$ is the rank of D), and $V_\beta \in \mathbb{R}^{p\times r}$. We then construct $V_\alpha \in \mathbb{R}^{p\times(p-r)}$ such that

$$V := \left[\begin{array}{cc} V_\alpha & V_\beta \end{array}\right] \in \mathbb{R}^{p\times p}$$

is a unitary matrix. Let $\beta = V_\beta^T\theta$ and $\alpha = V_\alpha^T\theta$. These two linear transformations split the original signal into two parts as follows:

$$\min_{\alpha,\beta}: \ \frac{1}{2}\left\|\Phi\left[V_\alpha\ V_\beta\right]\left[\begin{array}{c} V_\alpha^T \\ V_\beta^T \end{array}\right]\theta - c\right\|^2 + \lambda\|U\Sigma V_\beta^T\theta\|_1 \tag{8}$$

$$\equiv \frac{1}{2}\left\|\left[\Phi V_\alpha\ \Phi V_\beta\right]\left[\begin{array}{c} \alpha \\ \beta \end{array}\right] - c\right\|^2 + \lambda\|U\Sigma\beta\|_1 \tag{9}$$

$$\equiv \frac{1}{2}\left\|A\alpha + B\beta - c\right\|^2 + \lambda\|Z\beta\|_1 \tag{10}$$

where $A = \Phi V_\alpha \in \mathbb{R}^{n\times(p-r)}$, $B = \Phi V_\beta \in \mathbb{R}^{n\times r}$, and $Z = U\Sigma \in \mathbb{R}^{m\times r}$. Let $\hat\alpha, \hat\beta$ be the solution of Eq. (10). One can see the relationship between $\hat\alpha$ and $\hat\beta$: $\hat\alpha = -(A^TA)^{-1}A^T(B\hat\beta - c)$,[4] which can be used to further simplify Eq. (10):

$$\min_\beta: \ f(\beta) := \frac{1}{2}\|(I - A(A^TA)^{-1}A^T)(B\beta - c)\|^2 + \lambda\|Z\beta\|_1.$$

Let

$$X = (I - A(A^TA)^{-1}A^T)B$$

and

$$y = (I - A(A^TA)^{-1}A^T)c.$$

We obtain the following simplified formulation:

$$\min_\beta: \ f(\beta) = \frac{1}{2}\|X\beta - y\|^2 + \lambda\|Z\beta\|_1, \tag{11}$$

where $X \in \mathbb{R}^{n\times r}$ and $Z \in \mathbb{R}^{m\times r}$.

Denote the solution of Eq. (2) as $\hat\theta$ and the ground truth as θ^*. One can verify $\hat\theta = V[\hat\alpha^T\ \hat\beta^T]^T$. Define $\alpha^* := V_\alpha^T\theta^*$ and $\beta^* := V_\beta^T\theta^*$. Note that unlike $\hat\alpha$ and $\hat\beta$ the following usually does not hold: $\alpha^* = -(A^TA)^{-1}A^T(B\beta^* - c)$. Let $h = \hat\beta - \beta^*$ and $d = \hat\alpha - \alpha^*$. We will study the upper bound of $\|\hat\theta - \theta^*\|$ in terms of $\|h\|$ and $\|d\|$ based on the relationship $\|\hat\theta - \theta^*\| \le \|h\| + \|d\|$.

3. Main Results

This section presents the main results in this paper. The estimate error by Eq. (2), or equivalently Eq. (11), is given in Theorem 1:

Theorem 1. *Define*

$$\begin{aligned}
W_{Xh,1} :=& \rho^-_{X,Z^+}(s+l),\\
W_{Xh,2} :=& 6\sigma^{-1}_{\min}(Z)\rho^+_{X,Z^+}(s+l)\sqrt{s/l},\\
W_{d,1} :=& \frac{1}{2}\sigma^{-1}_{\min}(A^TA)(\bar\rho^+(s+l+p-r) -\\
& \bar\rho^-(s+l+p-r)),\\
W_{d,2} :=& \frac{3}{2}\sigma^{-1}_{\min}(A^TA)\sigma^{-1}_{\min}(Z)\sqrt{s/l}(\bar\rho^+(l+p-r) -\\
& \bar\rho^-(l+p-r)),\\
W_\sigma :=& \frac{\sigma_{\max}(Z)\sigma_{\min}(Z)}{\sigma_{\min}(Z) - 3\sqrt{s/l}\sigma_{\max}(Z)},\\
W_h :=& 3\sqrt{s/l}\sigma^{-1}_{\min}(Z),
\end{aligned}$$

where $\bar\rho^+(p-r,.)$ and $\bar\rho^-(p-r,.)$ denote $\rho^+_{[A,B],Z^+}(p-r,.)$ and $\rho^-_{[A,B],Z^+}(p-r,.)$ respectively for short. Taking $\lambda > 2\|(Z^+)^TX^T\epsilon\|_\infty$ in Eq. (2), we have if A^TA is invertible (apparently, $n \ge p - r$ is required) and there exists an integer $l > 9\kappa^2 s$ such that $W_{Xh,1} - W_{Xh,2}W_\sigma > 0$, then

$$\|\hat\theta - \theta^*\| \le W_\theta\sqrt{s}\lambda + \|(A^TA)^{-1}A^T\epsilon\| \tag{12}$$

where

$$W_\theta = 6\frac{(1 + W_{d,1})W_\sigma + (W_h + W_{d,2})W_\sigma^2}{W_{Xh,1} - W_{Xh,2}W_\sigma}.$$

[4] Here we assume that A^TA is invertible.

One can see from the proof that the first term of (12) is mainly due to the estimate error of the sparse part β and the second term is due to the estimate error of the free part α.

The upper bound in Eq. (12) strongly depends on parameters about X and Z^+ such as $\rho^+_{X,Z^+}(\cdot)$, $\rho^-_{X,Z^+}(\cdot)$, $\bar{\rho}^+(\cdot,\cdot)$, and $\bar{\rho}^-(\cdot,\cdot)$. Although for a given Φ and D, X and Z^+ are fixed, it is still challenging to evaluate these parameters. Similar to existing literature like (Candès & Tao, 2005), we assume Φ to be a Gaussian random matrix and estimate the values of these parameters in Theorem 2.

Theorem 2. *Assume that Φ is a Gaussian random matrix. The following holds with probability at least $1-2\exp\{-\Omega(k\log(em/k))\}$:*

$$\sqrt{\rho^+_{X,Z^+}(k)} \le \sqrt{n+r-p} + \Omega\left(\sqrt{k\log(em/k)}\right) \quad (13)$$

$$\sqrt{\rho^-_{X,Z^+}(k)} \ge \sqrt{n+r-p} - \Omega\left(\sqrt{k\log(em/k)}\right). \quad (14)$$

The following holds with probability at least $1-2\exp\{-\Omega\left((k+p-r)\log(ep/(k+p-r))\right)\}$:

$$\sqrt{\rho^+_{[A,B],Z^+}(p-r,k)} \le \sqrt{n} + \Omega(\sqrt{(k+p-r)\log(ep/(k+p-r))}) \quad (15)$$

$$\sqrt{\rho^-_{[A,B],Z^+}(p-r,k)} \ge \sqrt{n} - \Omega(\sqrt{(k+p-r)\log(ep/(k+p-r))}). \quad (16)$$

Now we are ready to analyze the estimate error given in Eq. (12). Two cases are considered in the following: the noiseless case $\epsilon = 0$ and the noisy case $\epsilon \neq 0$.

3.1. Noiseless Case $\epsilon = 0$

First let us consider the noiseless case. Since $\epsilon = 0$, the second term in Eq. (12) vanishes. We can choose a value of λ to make the first term in Eq. (12) arbitrarily small. Hence the true signal θ^* can be recovered with an arbitrary precision as long as $W_\theta > 0$, which is equivalent to requiring $W_{Xh,1} - W_{Xh,2}W_\sigma > 0$. Actually, when λ is extremely small, Eq. (2) approximately solves the problem in Eq. (7) with $\varepsilon = 0$.

Intuitively, the larger the measurement number n is, the easier the true signal θ^* can be recovered, since more measurements give a feasible subspace with a lower dimension. In order to estimate how many measurements are required, we consider the measurement matrix Φ to be a Gaussian random matrix (This is also a standard setup in compressive sensing.). Since this paper mainly focuses on the large scale case, one can treat the value of l as a number proportional to $\kappa^2 s$.

Using Eq. (13) and Eq. (14), we can estimate the lower bound of $W_{Xh,1} - W_{Xh,2}W_\sigma$ in Lemma 1.

Lemma 1. *Assume Φ to be a Gaussian random matrix. Let $l = (10\kappa)^2 s$. With probability at least $1-2\exp\{-\Omega((s+l)\log(em/(s+l)))\}$, we have*

$$W_{Xh,1} - W_{Xh,2}W_\sigma \ge \frac{1}{7}(n+r-p) - \Omega\left(\sqrt{(n+r-p)(s+l)\log\left(\frac{em}{s+l}\right)}\right). \quad (17)$$

From Lemma 1, to recover the true signal, we only need

$$(n+r-p) > \Omega((s+l)\log(em/(s+l))).$$

To simplify the discussion, we propose several minor conditions first in Assumption 1.

Assumption 1. *Assume that*

- *$p-r \le \phi n$ ($\phi < 1$) in the noiseless case and $p-r \le \Omega(s)$ in the noisy case [5];*
- *the condition number $\kappa = \frac{\sigma_{\max}(D)}{\sigma_{\min}(D)} = \frac{\sigma_{\max}(Z)}{\sigma_{\min}(Z)}$ is bounded;*
- *$m = \Omega(p^i)$ where $i > 0$, that is, m can be a polynomial function in terms of p.*

One can verify that under Assumption 1, taking $l = (10\kappa)^2 s = \Omega(s)$, the right hand side of (17) is greater than

$$\Omega(n) - \Omega(\sqrt{ns\log(em/s)}) = \Omega(n) - \Omega(\sqrt{ns\log(ep/s)}).$$

Letting $n \ge \Omega(s\log(ep/s))$ [or $\Omega(s\log(em/s))$ if without assuming $m = \Omega(p^i)$], one can have that

$$W_{Xh,1} - W_{Xh,2}W_\sigma \ge \Omega(n) - \Omega(\sqrt{ns\log(ep/s)}) > 0$$

holds with high probability (since the probability in Lemma 1 converges to 1 while p goes to infinity). In other words, in the noiseless case the true signal can be recovered at an arbitrary precision with high probability.

To compare with existing results, we consider two special cases: $D = \mathbf{I}_{p\times p}$ (Candès & Tao, 2005) and D has orthogonal columns (Candès et al., 2011), that is, $D^T D = I$.

[5] This assumption indicates that the free dimension of the true signal θ^* (or the dimension of the free part $\alpha \in \mathbb{R}^{p-r}$) should not be too large. Intuitively, one needs more measurements to recover the free part because it has no sparse constraint and much fewer measurements to recover the sparse part. Thus, if only limited measurements are available, we have to restrict the dimension of the free part.

When $D = \mathbf{I}_{p\times p}$ and Φ is a Gaussian random matrix, the required measurements in (Candès & Tao, 2005) are $\Omega(s\log(ep/s))$, which is the same as ours. Also note that if $D = \mathbf{I}_{p\times p}$, Assumption 1 is satisfied automatically. Thus our result does not enforce any additional condition and is consistent with existing analysis for the special case $D = \mathbf{I}_{p\times p}$. Next we consider the case when D has orthogonal columns as in (Candès et al., 2011). In this situation, all conditions except $m = \Omega(p^i)$ in Assumption 1 are satisfied. One can easily verify that the required measurements to recover the true signal are $\Omega(s\log(em/s))$ without assuming $m = \Omega(p^i)$ from our analysis above, which is consistent with the result in (Candès et al., 2011).

3.2. Noisy Case $\epsilon \neq 0$

Next we consider the noisy case, that is, study the upper bound in (12) while $\epsilon \neq 0$. Similarly, we mainly focus on the large scale case and assume Gaussian ensembles for the measurement matrix Φ. Theorem 3 provides the upper bound of the estimate error under the conditions in Assumption 1.

Theorem 3. *Assume that the measurement matrix Φ is a Gaussian random matrix, the measurement satisfies $n = O(s\log p)$, and Assumption 1 holds. Taking $\lambda = C\|(Z^+)^T X^T \epsilon\|_\infty$ with $C > 2$ in Eq. (2), we have*

$$\|\hat{\theta} - \theta^*\| \leq \Omega\left(\sqrt{\frac{s\log p}{n}}\right), \tag{18}$$

with the probability at least $1 - \Omega(p^{-1}) - \Omega(m^{-1}) - \Omega(-s\log(ep/s))$.

One can verify that when p goes to infinity, the upper bound in (18) converges to 0 from $n = O(s\log p)$ and the probability converges to 1 due to $m = \Omega(p^i)$. It means that the estimate error converges to 0 asymptotically given the measures $n = O(s\log p)$.

This result shows the consistency property, that is, if the measurement number n grows faster than $s\log(p)$, the estimate error will vanish. This consistency property is consistent with the special case LASSO by taking $D = \mathbf{I}_{p\times p}$ (Zhang, 2009a). Candès et al. (2011) considered Eq. (6) and obtained an upper bound for the estimate error $\Omega(\varepsilon/\sqrt{n})$ which does not guarantee the consistency property like ours since $\varepsilon = \Omega(\|\epsilon\|) = \Omega(\sqrt{n})$. Their result only guarantees that the estimation error bound converges to a constant given $n = O(s\log p)$.

In addition, from the derivation of Eq. (18), one can simply verify that the boundedness requirement for κ can actually be removed, if we allow more observations, for example, $n = O(\kappa^4 s\log p)$. Here we enforce the boundedness condition just for simplification of analysis and a convenient comparison to the standard LASSO (it needs $n = O(s\log p)$ measurements).

4. The Condition Number of D

Since κ is a key factor from the derivation of Eq. (18), we consider the fused LASSO and the random graphs and estimate the values of κ in these two cases.

Let us consider the fused LASSO first. The transformation matrix D is

$$\begin{bmatrix} \begin{bmatrix} \mathbf{I}_{(p-1)\times(p-1)} & \mathbf{0}_{p-1} \end{bmatrix} - \begin{bmatrix} \mathbf{0}_{p-1} & \mathbf{I}_{(p-1)\times(p-1)} \end{bmatrix} \\ \mathbf{I}_{p\times p} \end{bmatrix}.$$

One can verify that

$$\sigma_{\min}(D) = \min_{\|v\|=1} \|Dv\| \geq \min_{\|v\|=1} \|v\| = 1$$

and

$$\begin{aligned} &\sigma_{\max}(D) = \max_{\|v\|=1} \|Dv\| \\ \leq &\max_{\|v\|=1} \| \begin{bmatrix} \mathbf{I}_{(p-1)\times(p-1)} & \mathbf{0}_{p-1} \end{bmatrix} v - \\ &\begin{bmatrix} \mathbf{0}_{p-1} & \mathbf{I}_{(p-1)\times(p-1)} \end{bmatrix} v\| + \|v\| \\ \leq &\max_{\|v\|=1} \| \begin{bmatrix} \mathbf{I}_{(p-1)\times(p-1)} & \mathbf{0}_{p-1} \end{bmatrix} \|\|v\| + \\ &\| \begin{bmatrix} \mathbf{0}_{p-1} & \mathbf{I}_{(p-1)\times(p-1)} \end{bmatrix} \|\|v\| + \|v\| \\ \leq &3 \end{aligned}$$

which implies that $\sigma_{\min}(D) \geq 1$ and $\sigma_{\max}(D) \leq 3$. Hence we have $\kappa \leq 3$ in the fused LASSO case.

Next we consider the the random graph. The transformation matrix D corresponding to a random graph is generated in the following way: (1) each row is independent of the others; (2) two entries of each row are uniformly selected and are set to 1 and -1 respectively; (3) the remaining entries are set to 0. The following result shows that the condition number of D is bounded with high probability.

Theorem 4. *For any m and p satisfying that $m \geq cp$ where c is large enough, the following holds:*

$$\frac{\sigma_{\max}(D)}{\sigma_{\min}(D)} \leq \frac{\sqrt{m} + \Omega(\sqrt{p})}{\sqrt{m} - \Omega(\sqrt{p})},$$

with probability at least $1 - 2\exp\{-\Omega(p)\}$.

From this theorem, one can see that

- If $m = cp$ where c is large enough, then
$$\kappa = \frac{\sigma_{\max}(Z)}{\sigma_{\min}(Z)} = \frac{\sigma_{\max}(D)}{\sigma_{\min}(D)}$$
is bounded with high probability;
- If $m = p(p-1)/2$ which is the maximal possible m, then $\kappa \to 1$.

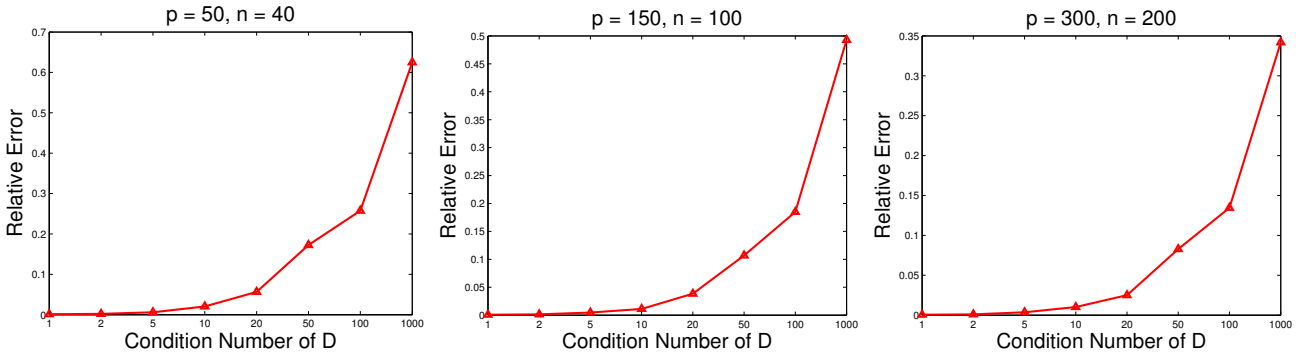

Figure 1. Illustration of the relationship between condition number and performance in terms of relative error. Three problem sizes are used as examples.

5. Numerical Simulations

In this section, we use numerical simulations to verify some of our theoretical results. Given a problem size n and p and condition number κ, we randomly generate D as follows. We first construct a $p \times p$ diagonal matrix D_0 such that

$$\text{Diag}(D_0) > 0 \text{ and } \frac{\max(\text{Diag}(D_0))}{\min(\text{Diag}(D_0))} = \kappa.$$

We then construct a random basis matrix $V \in \mathbb{R}^{p\times p}$, and let $D = D_0 V$. Clearly, D has independent columns and the condition number equals to κ. Next, a vector $x \in \mathbb{R}^p$ is generated such that $x_i \sim \mathcal{N}(0,1)$, $i = 1, \ldots, \frac{p}{10}$ and $x_j = 0$, $j = \frac{p}{10} + 1, \ldots, p$. θ^* is then obtained as $\theta^* = D^{-1}x$. Finally, we generate a matrix $\Phi \in \mathbb{R}^{n\times p}$ with $\Phi_{ij} \sim \mathcal{N}(0,1)$, noise $\epsilon \in \mathbb{R}^n$ with $\epsilon_i \sim \mathcal{N}(0, 0.001)$ and $y = \Phi\theta^* + \epsilon$.

We solve Eq. (2) using the standard optimization package CVX[6] and λ is set as $\lambda = 2\|(Z^+)^T X^T \epsilon\|_\infty$ as suggested by Theorem 1. We use three different sizes of problems, with $n \in \{40, 100, 200\}$, $p \in \{50, 150, 300\}$ and κ ranging from 1 to 1000. For each problem setting, 100 random instances are generated and the average performance is reported. We use the relative error $\frac{\|\hat{\theta}-\theta^*\|}{\|\theta^*\|}$ for evaluation, and present the performance with respect to different condition numbers in Figure 1. We can observe from Figure 1 that in all three cases the relative error increases when the condition number increases. If we fix the condition number, by comparing the three curves, we can see that the relative error decreases when the problem size increases. These are consistent with our theoretical results in Section 3 [see Eq. (18)].

6. Conclusion and Future Work

This paper considers the problem of estimating a specific type of signals which is sparse under a given linear transformation D. A conventional convex relaxation technique is used to convert this NP-hard combinatorial optimization into a tractable problem. We develop a unified framework to analyze the convex formulation with a generic D and provide the estimate error bound. Our main results establish that 1) in the noiseless case, if the condition number of D is bounded and the measurement number $n \geq \Omega(s\log(p))$ where s is the sparsity number, then the true solution can be recovered with high probability; and 2) in the noisy case, if the condition number of D is bounded and the measurement number grows faster than $s\log(p)$ [that is, $s\log(p) = o(n)$], then the estimate error converges to zero when p and s go to infinity with probability 1. Our results are consistent with existing literature for the special case $D = \mathbf{I}_{p\times p}$ (equivalently LASSO) and improve the existing analysis for the same formulation. The condition number of D plays a critical role in our theoretical analysis. We consider the condition numbers in two cases including the fused LASSO and the random graph. The condition number in the fused LASSO case is bounded by a constant, while the condition number in the random graph case is bounded with high probability if $\frac{m}{p}$ (that is, $\frac{\#\text{edge}}{\#\text{vertex}}$) is larger than a certain constant. Numerical simulations are consistent with our theoretical results.

In future work, we plan to study a more general formulation of Eq. (2):

$$\min_\theta : \ f(\theta) + \lambda\|D\theta\|_1,$$

where D is an arbitrary matrix and $f(\theta)$ is a convex and smooth function satisfying the restricted strong convexity property. We expect to obtain similar consistency properties for this general formulation.

Acknowledgments

This work was supported in part by NSF grants IIS-0953662 and MCB-1026710. We would like to sincerely

[6] cvxr.com/cvx/

thank Professor Shijian Wang and Professor Eric Bach of the University of Wisconsin-Madison for useful discussion and helpful advice.

References

Bunea, F., Tsybakov, A., and Wegkamp, M. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

Candès, E. J. and Plan, Y. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5A):2145–2177, 2009.

Candès, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51 (12):4203–4215, 2005.

Candès, E. J. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

Candès, E. J., Romberg, J., and Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

Candès, E. J., Eldar, Y. C., Needell, D., and Randall, P. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31:59–73, 2011.

Chan, T. F. Total variation blind deconvolution. *IEEE Transactions on Image Processing*, 7(3):370–375, 1998.

Donoho, D. L., Elad, M., and Temlyakov, V. N. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

Kolar, M., Song, L., and Xing, E.P. Sparsistent learning of varying-coefficient models with structural changes. *NIPS*, pp. 1006–1014, 2009.

Koltchinskii, V. and Yuan, M. Sparse recovery in large ensembles of kernel machines on-line learning and bandits. *COLT*, pp. 229–238, 2008.

Liu, J., Wonka, P., and Ye, J. A multi-stage framework for Dantzig selector and Lasso. *Journal of Machine Learning Research*, 13:1189–1219, 2012.

Liu, J., Yuan, L., and Ye, J. Guaranteed sparse recovery under linear transformation. *arxiv.org/abs/1305.0047*, 2013.

Lounici, K. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

Meinshausen, N., Bhlmann, P., and Zrich, E. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

Nama, S., Daviesb, M. E., Eladc, M., and Gribonvala, R. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30–56, 2012.

Ravikumar, P., Raskutti, G., Wainwright, M. J., and Yu, B. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. *NIPS*, 2008.

Rinaldo, A. Properties and refinements of the fusedLasso. *The Annals of Statistics*, 37(5B):2922–2952, 2009.

Romberg, J. The Dantzig selector and generalized thresholding. *CISS*, pp. 22–25, 2008.

Sharpnack, J., Rinaldo, A., and Singh, A. Sparsistency of the edgeLasso over graphs. *AISTAT*, 2012.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fusedLasso. *Journal of the Royal Statistical Society Series B*, pp. 91–108, 2005.

Vaiter, S., Peyre, G., Dossal, C., and Fadili, J. Robust sparse analysis regularization. *IEEE Transaction on Information Theory*, 59(4):2001–2016, 2013.

Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Zhang, T. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *Annals of Statistics*, 37(5A):2109–2114, 2009a.

Zhang, T. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009b.

Zhao, P. and Yu, B. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.