The L[subscript 1] penalized LAD estimator for high dimensional linear regression

Citation: Wang, Lie. "The L[subscript 1] Penalized LAD Estimator for High Dimensional Linear Regression." Journal of Multivariate Analysis 120 (September 2013): 135–151.

As Published: http://dx.doi.org/10.1016/j.jmva.2013.04.001

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/99451

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

L_1 penalized LAD estimator for high dimensional linear regression

Lie Wang*

Abstract

In this paper, the high-dimensional sparse linear regression model is considered, where the overall number of variables is larger than the number of observations. We investigate the L_1 penalized least absolute deviation method. Different from most of other methods, the L_1 penalized LAD method does not need any knowledge of standard deviation of the noises or any moment assumptions of the noises. Our analysis shows that the method achieves near oracle performance, i.e. with large probability, the L_2 norm of the estimation error is of order $O(\sqrt{k \log p/n})$. The result is true for a wide range of noise distributions, even for the Cauchy distribution. Numerical results are also presented.

Keywords: high dimensional regression, LAD estimator, L_1 penalization, variable selection.

*Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; e-mail: liewang@math.mit.edu Research supported by NSF Grant DMS-1005539.

1 Introduction

High dimensional linear regression model, where the number of observations is much less than the number of unknown coefficients, has attracted much recent interests in a number of fields such as applied math, electronic engineering, and statistics. In this paper, we consider the following classical high dimensional linear model:

$$Y = X\beta + z. \tag{1}$$

where $Y = (y_1, y_2, \cdots, y_n)'$ is the n dimensional vector of outcomes, X is the $n \times p$ design matrix, and $z = (z_1, z_2, \cdots, z_n)'$ is the n dimensional vector of measurement errors (or noises). We assume $X = (X_1, X_2, \cdots, X_p)$ where $X_i \in R^n$ denotes the ith regressor or variable. Throughout, we assume that each vector X_i is normalized such that $\|X_i\|_2^2 = n$ for $i = 1, 2, \cdots, p$. We will focus on the high dimensional case where $p \geq n$ and our goal is to reconstruct the unknown vector $\beta \in R^p$.

Since we are considering a high dimensional linear regression problem, a key assumption is the sparsity of the true coefficient β. Here we assume,

$$T = supp(\beta) \text{ has } k < n \text{ elements.}$$

The set T of nonzero coefficients or significant variables is unknown. In what follows, the true parameter value β and p and k are implicitly indexed by the sample size n, but we omit the index in our notation whenever this does not cause confusion.

Ordinary least square method is not consistent in the setting of $p > n$. In recent years, many new methods have been proposed to solve the high dimensional linear regression problem. Methods based on L_1 penalization or constrained L_1 minimization have been extensively studied. Dantzig selector was proposed in [10], which can be written as

$$\hat{\beta}_{DS} = \arg\min_{\gamma \in R^p} \|\gamma\|_1, \text{ subject to } \|X'(Y - X\gamma)\|_\infty \leq c\sigma\sqrt{2n \log p},$$

for some constant $c > 1$. It is clear that the Dantzig selector depends on the standard deviation of the noise and the Gaussian assumption. General constrained L_1 minimization methods for noiseless case and Gaussian noise were studied in [7]. More results about the constrained L_1 minimization can be found in for example [9], [12], [8] and the references therein.

Besides the constrained minimization methods, the lasso (L_1 penalized least square) type methods have been studied in a number of papers, for example, [22], [4], and [18]. The classical lasso estimator can be written as

$$\hat{\beta}_{lasso} = \arg\min_{\gamma} \frac{1}{2}\|Y - X\gamma\|_2^2 + \lambda\|\gamma\|_1,$$

where λ is the penalty level (tuning parameter). In the setting of Gaussian noise and known variance, it is suggested in [4] that the penalty could be

$$\lambda = 2c\sigma\sqrt{n\Phi^{-1}(1-\alpha/2p)},$$

where $c > 1$ is a constant and α is small chosen probability. By using this penalty value, it was shown that the lasso estimator can achieve near oracle performance, i.e. $\|\hat{\beta}_{lasso} - \beta\|_2 \leq C(k\log(2p/\alpha)/n)^{1/2}$ for some constant $C > 0$ with probability at least $1 - \alpha$.

The lasso method has nice properties, but it also replies heavily on the Gaussian assumption and a known variance. In practice, the Gaussian assumption may not hold and the estimation of the standard deviation σ is not a trivial problem. In a recent paper, [3] proposed the square-root lasso method, where the knowledge of the distribution or variance are not required. Instead, some moment assumptions of the errors and design matrix are needed. Other than the constrained optimization or penalized optimization methods, the stepwise algorithm are also studied, see for example [24] and [6]. It is worth noting that to properly apply the stepwise methods, we also need assumptions on the noise structure or standard deviation of the noises.

It is now seen that for most of the proposed methods, the noise structure plays an important role in the estimation of the unknown coefficients. In most of the existing literatures, either an assumption on the error distribution or a known variance is required. Unfortunately, in the high dimensional setup, these assumptions are not always true. Moreover, in cases where heavy-tailed errors or outliers are found in the response, the variance of the errors may be unbounded. Hence the above methods cannot be applied.

To deal with the cases where the error distribution is unknown or may has heavy tail. We propose the following L_1 penalized least absolute deviation (L_1 PLAD) estimator,

$$\hat{\beta} \in \arg\min\{\gamma : \|Y - X\gamma\|_1 + \lambda\|\gamma\|_1\}. \tag{2}$$

The least absolute deviation (LAD) type of methods are important when heavy-tailed errors present. These methods have desired robust properties in linear regression models, see for example [1], [15] and [19].

Recently, the penalized version of the LAD method was studied in several papers and the variable selection and estimation properties were discussed. In [13], the asymptotic properties of variable selection consistency were discussed under strong conditions such that the entries of the design matrix X are uniformly bounded. Also, how to find the tuning parameter that will generate the consistent estimator is still unclear. The estimation consistency of the penalized LAD estimator were discussed in for example [23] and [16], where the number of variables p is assumed to be fixed. It is worth noting that in the proof of lemma 1 of [23], the authors did not prove the convergence in the last step is uniform, hence the proof is incomplete. In a recent paper [2], the quantile regression model was considered and L_1 penalized method was proposed. Properties of the estimator were presented under restricted eigenvalue type conditions and smooth assumptions on the density function of the noise. It is worth pointing out that in our paper, we discuss both the noisy and noiseless cases and a more general noise structure is considered. Besides, we

will discuss the conditions on matrix X under the case of Gaussian random design in the Appendix.

Here in this paper, we present analysis for the L_1 PLAD method and we discuss the selection of penalty level, which does not depend on any unknown parameters or the noise distribution. Our analysis shows that the L_1 PLAD method has surprisingly good properties. The main contribution of the present paper has twofold. (1) We proposed a rule for setting the penalty level, it is simply

$$\lambda = c\sqrt{2A(\alpha)n\log p},$$

where $c > 1$ is a constant, α is a chosen small probability, and $A(\alpha)$ is a constant such that $2p^{-(A(\alpha)-1)} \leq \alpha$. In practice, we can simply choose $\lambda = \sqrt{2n\log p}$, see the numerical study section for more discussions. This choice of penalty is universal and we only assume that the noises have median 0 and $P(z_i = 0) = 0$ for all i. (2) We show that with high probability, the estimator has near oracle performance, i.e. with high probability

$$\|\hat{\beta} - \beta\|_2 = O(\sqrt{\frac{k\log p}{n}}).$$

It is important to notice that we do not have any assumptions on the distribution or moments of the noise, we only need a scale parameter to control the tail probability of the noise. Actually, even for Cauchy distributed noise, where the first order moment does not exist, our results still hold.

Importantly, the problem retains global convexity, making the method computationally efficient. Actually, we can use ordinary LAD method package to solve the L_1 penalized LAD estimator. This is because if we consider the penalty terms as new observations, i.e. $Y_{n+i} = 0$ and $x_{n+i,j} = \lambda \times I(j = i)$ for $i, j = 1, 2, \cdots, p$. Here $I(j = i)$ is the indicator function such that $I(j = i) = 1$ if $j = i$ and $I(j = i) = 0$ if not. Then our L_1 penalized estimator can be considered as an ordinary LAD estimator with p unknown coefficients and

$p + n$ observations. Hence it can be solved efficiently.

The rest of the paper is organized as follows. Section 2 discusses the choice of penalty level. In section 3, the main results about the estimation error and several critical lemmas are presented. We also briefly explain the main idea of the proofs. Section 4 presents the simulation study results, which shows the L_1 penalized LAD method has very good numerical performance regardless the noise distribution. Technical lemmas and the proofs of theorems are given in section 5. The Appendix presents the discussion of conditions on matrix X under Gaussian random design.

2 Choice of Penalty

In this section, we discuss the choice of the penalty level for the L_1 PLAD estimator. For any $\gamma \in R^p$, let $Q(\gamma) = \|Y - X\gamma\|_1$. Then the L_1 PLAD estimator can be written as

$$\hat{\beta} \in \arg\min\{\gamma : Q(\gamma) + \lambda\|\gamma\|_1\}.$$

An important quantity to determine the penalty level is the sub-differential of Q evaluated at the point of true coefficient β. Here we assume that the measurement errors z_i satisfy $P(z_i = 0) = 0$ and the median of z_i is 0 for $i = 1, 2, \cdots, n$. Assume that $z_i \neq 0$ for all i, then the sub-differential of $Q(\gamma) = \|Y - X\gamma\|_1$ at point $\gamma = \beta$ can be written as

$$S = X'(sign(z_1), sign(z_2), \cdots, sign(z_n))',$$

where $sign(x)$ denotes the sign of x, i.e. $sign(x) = 1$ if $x > 0$, $sign(x) = -1$ if $x < 0$, and $sign(0) = 0$. Let $I = sign(z)$, then $I = (I_1, I_2, \cdots, I_n)'$ where $I_i = sign(z_i)$. Since z_i's are independent and have median 0, we know that $P(I_i = 1) = P(I_i = -1) = 0.5$ and I_i are independent.

The sub-differential of $Q(\gamma)$ at the point of β, $S = X'I$, summaries the estimation error in the setting of linear regression model. We will choose a penalty λ that dominates the

estimation error with large probability. This principle of selecting the penalty λ is motivated by [4] and [3]. It is worth noting that this is a general principle of choosing the penalty and can be applied to many other problems. To be more specific, we will choose a penalty λ such that it is greater than the maximum absolute value of S with high probability, i.e. we need to find a penalty level λ such that

$$P(\lambda \geq c\|S\|_\infty) \geq 1 - \alpha, \tag{3}$$

for a given constant $c > 1$ and a given small probability α. Note that c is a theoretical constant and in practice we can simply take $c = 1.1$. Since the distribution of I is known, the distribution of $\|S\|_\infty$ is known for any given X and does not depend on any unknown parameters.

Now for any random variable W let $q_\alpha(W)$ denote the $1 - \alpha$ quantile of W. Then in theory, $q_\alpha(\|S\|_\infty)$ is known for any given X. Therefore if we choose $\lambda = cq_\alpha(\|S\|_\infty)$, inequality (3) is satisfied.

In practice, it might be hard to calculate the exact quantile $q_\alpha(\|S\|_\infty)$ for a given X. One possible way to calculate or approximate it is by simulation, but this will cause additional computation time. Here we propose the following choice of penalty.

$$\lambda = c\sqrt{2A(\alpha)n \log p}, \tag{4}$$

where $A(\alpha) > 0$ is a constant such that $2p^{-(A(\alpha)-1)} \leq \alpha$.

To show that the above choice of penalty satisfies (3), we need to bound the tail probability of $\sum_{i=1}^n X_{ij} I_i$ for $i = 1, 2, \cdots, p$. This can be done by using the Hoeffding's inequality, see for example [14], and union bounds. We have the following lemma.

Lemma 1 *The choice of penalty $\lambda = c\sqrt{2A(\alpha)n \log p}$ as in (4) satisfies*

$$P(\lambda \geq c\|S\|_\infty) \geq 1 - \alpha.$$

From the proof of the previous lemma, we can see that if we use the following special choice of λ,

$$\lambda = 2c\sqrt{n \log p}, \tag{5}$$

Then we have that

$$P(\lambda \geq c\|S\|_\infty) \geq 1 - \frac{2}{p}. \tag{6}$$

The above penalties are simple and have good theoretical properties. Moreover, they do not require any conditions on matrix X or value of p and n. But in practice, since the bounds here are not very tight, these penalty levels tend to be relatively large and can cause additional bias to the estimator. It is worth pointing out that if there exists an $i \in \{1, 2, \cdots, p\}$ such that $\|X_i\|_1 < \lambda$, then $\hat{\beta}_i$ must be 0. Otherwise we can replace $\hat{\beta}_i$ by 0, and the value of $Q(\hat{\beta}) + \lambda\|\hat{\beta}\|_1$ will reduce by at least $(\lambda - \|X_i\|_1)|\hat{\beta}_i|$. This means if the penalty level λ is too large, the L_1 PLAD method may kill some of the significant variables. To deal with this issue, we propose the following refined asymptotic choice of penalty level, provided some moment conditions on design matrix X.

Lemma 2 *Suppose*

$$B = \sup_n \sup_{1 \leq j \leq p} \frac{1}{n}\|X_j\|_q^q < \infty, \tag{7}$$

for some constant $q > 2$. Assume $\Phi^{-1}(1 - \alpha/2p) \leq (q-2)\sqrt{\log n}$. Then the choice of penalty $\lambda = c\sqrt{n}\Phi^{-1}(1 - \frac{\alpha}{2p})$ satisfies

$$P(\lambda \geq c\|S\|_\infty) \geq 1 - \alpha(1 + \omega_n),$$

where ω_n goes to 0 as n goes to infinity.

This choice of penalty replies on moment conditions of X and relative size of p and n, but it could be smaller than the previous ones and in practice it will cause less bias. We investigate the effect of different penalties in the numerical study section.

To simplify our arguments, in the following theoretical discussion we will use (5) as the default choice of penalty. It can be seen that the above choices of penalty levels do not depend on the distribution of measurement errors z_i or unknown coefficient β. As long as z_i's are independent random variables with median 0 and $P(z_i = 0) = 0$, the choices satisfy our requirement. This is a big advantage over the traditional lasso method, which significantly relies on the Gaussian assumption and the variance of the errors.

3 Properties of the Estimator

In this section, we present the properties of the L_1 PLAD estimator. We shall state the upper bound for estimation error $h = \hat{\beta} - \beta$ under L_2 norm $\|h\|_2$. We shall also present the variable selection properties for both noisy and noiseless cases. The choice of penalty is described in the previous section. Throughout the discussion in this section, we assume the penalty λ satisfies $\lambda \geq c\|S\|_\infty$ for some fixed constant $c > 1$. In what follows, for any set $E \subset \{1, 2, \cdots, p\}$ and vector $h \in R^p$, let $h_E = hI(E)$ denote the p dimensional vector such that we only keep the coordinates of h when their indexes are in E and replace others by 0.

3.1 Conditions on design matrix X

We will first introduce some conditions on design matrix X. Recall that we assume $\lambda \geq c\|S\|_\infty$, this implies the following event, namely $h = \hat{\beta} - \beta$ belongs to the restricted set $\Delta_{\bar{C}}$, where

$$\begin{aligned}\Delta_{\bar{C}} = \quad & \{\delta \in R^p : \|\delta_T\|_1 \geq \bar{C}\|\delta_{T^c}\|_1, \\ & \text{where } T \subset \{1, 2, \cdots, p\} \text{ and } T \text{ contains at most } k \text{ elements.}\},\end{aligned}$$

and $\bar{C} = (c-1)/(c+1)$. To show this important property of the L_1 PLAD estimator,

recall that $\hat{\beta}$ minimizes $\|X\gamma - Y\|_1 + \lambda\|\gamma\|_1$. Hence

$$\|Xh + z\|_1 + \lambda\|\hat{\beta}\|_1 \leq \|z\|_1 + \lambda\|\beta\|_1.$$

Let T denote the set of significant coefficients. Then

$$\|Xh + z\|_1 - \|z\|_1 \leq \lambda(\|h_T\|_1 - \|h_{T^c}\|_1). \tag{8}$$

Since the sub-differential of $Q(\gamma)$ at the point of β is $X'I$, where $I = sign(z)$.

$$\|Xh + z\|_1 - \|z\|_1 \geq (Xh)'I \geq h'X'I \geq -\|h\|_1\|X'I\|_\infty \geq -\frac{\lambda}{c}(\|h_T\|_1 - \|h_{T^c}\|_1).$$

So

$$\|h_T\|_1 \geq \bar{C}\|h_{T^c}\|_1, \tag{9}$$

where $\bar{C} = \frac{c-1}{c+1}$.

The fact that $h \in \Delta_{\bar{C}}$ is extremely important for our arguments. This fact is also important for the arguments of classical lasso method and the square-root lasso method, see for example, [4] and [3].

Now we shall define some important quantities of design matrix X. Let λ_k^u be the smallest number such that for any k sparse vector $d \in R^p$,

$$\|Xd\|_2^2 \leq n\lambda_k^u\|d\|_2^2.$$

Here k sparse vector d means that the vector d has at most k nonzero coordinates, or $\|d\|_0 \leq k$. Similarly, let λ_k^l be the largest number such that for any k sparse vector $d \in R^p$,

$$\|Xd\|_2^2 \geq n\lambda_k^l\|d\|_2^2.$$

Let θ_{k_1,k_2} be the smallest number such that for any k_1 and k_2 sparse vector c_1 and c_2 with disjoint support,

$$|\langle Xc_1, Xc_2\rangle| \leq n\theta_{k_1,k_2}\|c_1\|_2\|c_2\|_2.$$

The definition of the above constants are essentially the Restricted Isometry Constants, see for example [11], but we use different notations for upper and lower bounds. We also need to define the following restricted eigenvalue of design matrix X. These definitions are based on the idea of [4]. Let

$$\kappa_k^l(\bar{C}) = \min_{h \in \Delta_{\bar{C}}} \frac{\|Xh\|_1}{n\|h_T\|_2}.$$

To show the properties of the L_1 penalized LAD estimator, we need $\kappa_k^l(\bar{C})$ to be bounded away from 0. To simplify the notations, when it is not causing any confusion, we will simply write $\kappa_k^l(\bar{C})$ as κ_k^l.

3.2 Important Lemmas

Before presenting the main theorem, we first state a few critical lemmas. From (8), we know that

$$\|Xh + z\|_1 - \|z\|_1 \leq \lambda\|h_T\|_1.$$

To bound the estimation error, we shall first investigate the random variable $\frac{1}{\sqrt{n}}(\|Xh + z\|_1 - \|z\|_1)$. For any vector $d \in R^p$, let

$$B(d) = \frac{1}{\sqrt{n}} \left| (\|Xd + z\|_1 - \|z\|_1) - E(\|Xd + z\|_1 - \|z\|_1) \right|.$$

We introduce the following important result.

Lemma 3 *Suppose z_i's are independent random variables. Assume $p > n$ and $p > 3\kappa_k^u$ then*

$$P\left(\sup_{\|d\|_0 = k, \|d\|_2 = 1} B(d) \geq (1 + 2C_1\sqrt{\lambda_k^u})\sqrt{2k \log p} \right) \leq 2p^{-4k(C_1^2 - 1)}, \tag{10}$$

where $C_1 > 1$ is a constant.

From the above lemma, we know that with probability at least $1 - 2p^{-4k(C_1^2-1)}$, for any k sparse vector $d \in R^p$,

$$\frac{1}{\sqrt{n}}(\|Xd + z\|_1 - \|z\|_1) \geq \frac{1}{\sqrt{n}}E(\|Xd + z\|_1 - \|z\|_1) - C\sqrt{2k \log p}\|h\|_2, \tag{11}$$

where $C = 1 + 2C_1\sqrt{\lambda_k^u}$. This lemma shows that with high probability, the value of the random variable $\frac{1}{\sqrt{n}}(\|Xd+z\|_1-\|z\|_1)$ is very close to its expectation. Since the expectation is fixed and much easier to analysis than the random variable itself, this lemma plays an important role in our proof of the main theorem.

Next, we will investigate the properties of $E(\|Xd + z\|_1 - \|z\|_1)$. We have the following lemmas.

Lemma 4 *For any continuous random variable z_i, we have that*

$$\frac{dE(|z_i + x| - |z_i|)}{dx} = 1 - 2P(z_i \leq -x).$$

Now we will introduce the scale assumptions on the measurement errors z_i. suppose there exists a constant $a > 0$ such that

$$\begin{aligned} P(z_i \geq x) &\leq \frac{1}{2 + ax} \text{ for all } x \geq 0 \\ P(z_i \leq x) &\leq \frac{1}{2 + a|x|} \text{ for all } x < 0. \end{aligned} \tag{12}$$

Here a served as a scale parameter of the distribution of z_i. This is a very weak condition and even Cauchy distribution satisfies it. Based on this assumption, we have that for any $c > 0$,

$$\begin{aligned} & E(|z_i + c| - |z_i|) = c - 2\int_0^c P(z_i < -x)dx \\ \geq \quad & c - 2\int_0^c \frac{1}{2 + ax}dx = c - \frac{2}{a}\log(1 + \frac{a}{2}c). \end{aligned}$$

Hence we have the following lemma.

Lemma 5 *Suppose random variable z satisfies condition (12), then*

$$E(|z_i + c| - |z_i|) \geq \frac{a}{16}|c|(|c| \wedge \frac{6}{a}). \tag{13}$$

Remark 1 *This is just a weak bound and can be improved easily. But for simplicity, we use this one in our discussion.*

3.3 Main Theorems

Now we shall propose our main result. Here we assume that the measurement errors z_i are independent and identically distributed random variables with median 0. We also assume that z_is satisfy condition (12). Moreover, we assume $\lambda_k^l > \theta_{k,k}(\frac{1}{C} + \frac{1}{4})$ and

$$\frac{3\sqrt{n}}{16}\kappa_k^l > \lambda\sqrt{k/n} + C_1\sqrt{2k\log p}(\frac{5}{4} + \frac{1}{\bar{C}}), \tag{14}$$

for some constant C_1 such that $C_1 > 1 + 2\sqrt{\lambda_k^u}$. We have the following theorem.

Theorem 1 *Under the previous assumptions, the L_1 penalized LAD estimator $\hat{\beta}$ satisfies with probability at least $1 - 2p^{-4k(C_2^2-1)+1}$*

$$\|\hat{\beta} - \beta\|_2 \leq \sqrt{\frac{2k\log p}{n}}\frac{16(c\sqrt{2} + 1.25C_1 + C_1/\bar{C})}{a(\lambda_k^l - \theta_{k,k}(\frac{1}{\bar{C}} + \frac{1}{4}))^2/\lambda_k^u}\sqrt{1 + \frac{1}{\bar{C}}}.$$

where $C_1 = 1 + 2C_2\sqrt{\lambda_k^u}$ and $C_2 > 1$ is a constant.

Remark 2 *From the proof of the theorem, we can see that the identically distributed assumption of the measurement errors is not essential. We just need that there exist a constant $a > 0$ such that for all i, $P(z_i \geq x) \leq \frac{1}{2+ax}$ for $x \geq 0$ and $P(z_i \leq x) \leq \frac{1}{2+a|x|}$ for $x < 0$. This is also verified in the section of simulation study.*

Remark 3 *Actually, $\theta_{k,k}$ can be bounded by λ_k^l and λ_k^u and the condition $\lambda_k^l > \theta_{k,k}(\frac{1}{\bar{C}} + \frac{1}{4})$ can be replaced by a number of similar RIP conditions, see for example [8]. We keep it here just to simplify the arguments.*

Remark 4 *Condition (14) implies that the columns of X cannot be too sparse. This is because if the columns of X are sparse then the L_1 norm of columns of X will be small, hence the value κ_k^l will be small.*

From the theorem we can easily see that asymptotically, with high probability,

$$\|\hat{\beta} - \beta\|_2 = O(\sqrt{\frac{2k \log p}{n}}). \tag{15}$$

This means that asymptotically, the L_1 PLAD estimator has near oracle performance and hence it matches the asymptotic performance of the lasso method with known variance.

A simple consequence of the main theorem is that the L_1 PLAD estimator will select most of the significant variables with high probability. We have the following theorem.

Theorem 2 *Suppose $\hat{T} = supp(\hat{\beta})$ be the estimated support of the coefficients. Then under the same conditions as in Theorem 1, with probability at least $1 - 2p^{-4k(C_2^2-1)+1}$,*

$$\left\{ i : |\beta_i| \geq \sqrt{\frac{2k \log p}{n}} \frac{16(c\sqrt{2} + 1.25C_1 + C_1/\bar{C})}{a(\lambda_k^l - \theta_{k,k}(\frac{1}{C} + \frac{1}{4}))^2/\lambda_k^u} \right\} \subset \hat{T}, \tag{16}$$

where $C_1 = 1 + 2C_2\sqrt{\lambda_k^u}$ and $C_2 > 1$ is a constant.

Remark 5 *This theorem shows that the L_1 PLAD method will select a model that contains all the variables with large coefficients. If in the main model, all the nonzero coefficients are large enough in terms of absolute value, then the L_1 PLAD method can select all of them into the model.*

A special but important case in high dimensional linear regression is the noiseless case. The next theorem shows that the L_1 PLAD estimator has nice variable selection property in the noiseless case.

Theorem 3 *Consider the noiseless case. Suppose we use a penalty level λ such that $\lambda < n\kappa_k^l(1)$, the L_1 penalized LAD estimator $\hat{\beta}$ satisfies $\hat{\beta} = \beta$.*

Remark 6 *Suppose $\kappa_k^l(1)$ are bounded away from 0 for all n and we use the penalty level $\lambda = 2\sqrt{n \log p}$. Then when $\sqrt{\log p} = o(n)$ and n large enough. The L_1 penalized LAD estimator $\hat{\beta}$ satisfies $\hat{\beta} = \beta$.*

Remark 7 *From the discussion in the Appendix we know that if we use the i.i.d. Gaussian random design and $\log p = o(n^{1/6})$, then for n large enough the L_1 PLAD estimator satisfies $\hat{\beta} = \beta$ with high probability.*

4 Numerical Study

In this section, we will show some numerical results. Throughout this section, we use $n = 200$, $p = 400$ and $k = 5$ and set $\beta = (3, 3, 3, 3, 3, 0, \cdots, 0)$. We will study both the estimation properties and variable selection properties of the L_1 PLAD estimator under various noise structures. In our simulation study, we generate the design matrix X by i.i.d. $N(0, 1)$ random variables and then normalize the columns.

We first investigate the effect of different choices of penalty levels. Then we compare the L_1 PLAD method and the lasso method in the Gaussian noise case. We also study the numerical properties of L_1 PLAD estimator under different noise structures, including the heteroscedastic cases. We use the quantreg package and lars package in R to run the simulation.

4.1 Effect of Penalty levels

Section 2 discusses the choice of penalty levels. It is known that our desired choice is $cq_\alpha(\|S\|_\infty)$. But since this value is hard to calculate, we propose several upper bounds and asymptotic choices. Now we will investigate the effect of different choices of penalty levels on the L_1 PLAD estimator. To be specific, we consider the following four penalties, $\lambda_1 = \sqrt{1.5n \log p}$, $\lambda_2 = \sqrt{2n \log p}$, $\lambda_3 = \sqrt{3n \log p}$, and $\lambda_4 = \sqrt{4n \log p}$. Note that they are

Table 1: The average of estimation error $\|\hat{\beta} - \beta\|_2^2$ over 200 simulations under different penalty levels and error distributions. Numbers in the parentheses are the medians of the estimation errors of post L_1 PLAD method, i.e. results of ordinary LAD estimators on the selected subset.

	λ_1	λ_2	λ_3	λ_4
$N(0,1)$ noise	0.658 (0.356)	1.054 (0.239)	3.189 (0.095)	23.730 (4.586)
$t(2)$ noise	1.263 (0.552)	2.351 (0.299)	10.121 (0.081)	33.018 (18.771)
Cauchy noise	2.176 (0.861)	4.736 (0.334)	21.417 (0.103)	39.351 (26.241)

all fixed choices and do not depend on any assumptions or parameters. For noises, we use (a) $N(0,1)$ noise, (b) $t(2)$ noise, and (c) Cauchy noise. For each setting, we run the simulation 200 times and the average L_2 norm square of the estimation errors are summarized in the following table.

From table 1 we can see that λ_4 is too large in our setup and it kills most of the variables. (It is worth noting that if we increase the sample size to for example $n = 400$ and $p = 800$, λ_4 becomes a reasonable choice.) Moreover, larger λ cause more bias to the estimator. In practice, an ordinary least square method or least absolute deviation method could be applied to the selected variables to correct the bias (post L_1 PLAD method). We summarized the median of the ordinary LAD estimators on the selected subset in the above table. It can be seen that among the four penalty levels, λ_1 has the best results in terms of the estimation error $\|\hat{\beta} - \beta\|_2^2$, and λ_3 has the best results in terms of post L_1 PLAD estimation error. The post L_1 PLAD results are very good for all three noise distributions even though the $t(2)$ distribution does not have bounded variance and Cauchy distribution does not have bounded expectation.

4.2 Gaussian Noise

Now consider the Gaussian noise case, i.e. z_i are independent and identically normal random variables. The standard deviation σ of z_i is varied between 0 and 1. Here we also include the noiseless case (where the traditional lasso cannot select the model correctly) and the Cauchy distribution case (to compare). We will use penalty level $\lambda = \sqrt{2n \log p}$ and run 200 times for each value of σ. For each simulation, we use both the L_1 PLAD method and the classical lasso method. For the lasso method, we consider to ways to select the penalty. One is to use $\sigma \times \lambda$ as the penalty, where we assume the standard deviation is known. The other one is by cross validation. In the noiseless case, we use $0.01 \times \lambda$ or the cross validation to select the penalty levels for the lasso method. For the Cauchy distribution case, only the cross validation is considered. Here we summaries the average estimation error and the variable selection results of both methods for different distributions.

In table 2, the average type I error means the average number of significant variables that are unselected over 200 runs. The average type II error means the average number of insignificant variables that are selected over 200 runs. The results show that in terms of estimation, the classical lasso method with known standard deviation does better than L_1 PLAD method, except the noiseless case. This is partly because that lasso knows the standard deviation and L_1 PLAD does not. Also, the penalty level for L_1 PLAD method has stronger shrinkage effect and hence cause more bias. The lasso with cross validation did a fine job in terms of estimation in the Gaussian noise case, but it performs poorly in the noiseless case and Cauchy distribution case.

In term of variable selection, the L_1 PLAD method does better than classical lasso method. The two methods both select all the significant variables in all the 200 simulations for the Gaussian noise cases. The L_1 PLAD method has smaller average type II errors which means the lasso method tends to select more incorrect variables than the L_1 PLAD method.

On the other hand, the lasso with cross validation selects a large amount of variables into the model, its average type II errors is huge. It is worth noting that L_1 PLAD method does a perfect job in noiseless case, it selects the perfect model in every run. While the lasso method never have a correct variable selection result.

4.3 Heavy tail and Heteroscedastic Noise

In the proof of Theorem 1 and all the discussions, the identically distribution assumption is not essential for our arguments. Now we will study the performance of the L_1 PLAD estimator when the noises z_i are just independent and not identically distributed. We will consider three cases: (a) Half of the z_i are $N(0,1)$ random variables and half of them are $N(0,4)$ random variables. (b) Half of the z_i are $t(2)$ random variables and half of them are

Table 2: The average of estimation error $\|\hat{\beta}-\beta\|_2^2$ over 200 replications and the variable selection results for lasso and L_1 penalized LAD method.

Distribution	$N(0,0)$	$N(0,0.25^2)$	$N(0,0.5^2)$	$N(0,1)$	Cauchy
L_1 PLAD: Average of $\|\hat{\beta}-\beta\|_2$	0	0.065	0.269	1.057	4.731
L_1 PLAD: Average type I error	0	0	0	0	0.002
L_1 PLAD: Average type II error	0	0.185	0.150	0.120	0.161
Lasso: Average of $\|\hat{\beta}-\beta\|_2$	11.419	0.062	0.106	0.344	NA
Lasso: Average type I error	0	0	0	0	NA
Lasso: Average type II error	24.125	0.825	0.875	0.710	NA
CV Lasso: Average of $\|\hat{\beta}-\beta\|_2$	2.788	0.122	0.241	0.455	6.862
CV Lasso: Average type I error	0.05	0	0	0	2.18
CV Lasso: Average type II error	31.775	65.960	61.875	53.035	20.8

$t(2)$ random variables multiple by 2. (c) One third of the z_i are $N(0,1)$ random variables; one third of them are $t(2)$ random variables; the rest of them follow exponential distribution with parameter 1 (relocated such that the median is 0). We use penalty $\lambda = \sqrt{2n \log p}$ for all cases. It is worth noting that in all the cases, traditional lasso method and the constrained minimization methods cannot be properly applied since the variances of the noises are unbounded.

Table 3 summaries the average estimation errors and variable selection properties of the L_1 PLAD method over 200 runs. We also summarize the estimation errors of the post L_1 PLAD method in the parentheses. It can be seen that the L_1 PLAD method has very nice estimation and variable selection properties for all cases. Compare the variable selection results here with the Gaussian noise case in table 2, we can see that although we have many different noise structures, the L_1 PLAD method can always select a good model. Its variable selection results here are comparable to the Gaussian noise case.

5 Proofs

We will first show some technical lemmas and then prove the main results.

Table 3: The average of estimation error $\|\hat{\beta} - \beta\|_2$ over 200 replications and the variable selection results for the L_1 PLAD method. Numbers in the parentheses are the medians of the estimation errors of post L_1 PLAD method.

	Case (a)	Case (b)	Case (c)
Average of $\|\hat{\beta} - \beta\|_2^2$	1.701 (0.234)	3.545 (0.228)	1.808 (0.210)
Average type I error	0	0	0.004
Average type II error	0.141	0.152	0.161

5.1 Technical Lemmas

We first state the Slastnikov-Rubin-Sethuraman Moderate Deviation Theorem. Let $X_{ni}, i = 1, \ldots, k_n; n \geq 1$ be a double sequence of row-wise independent random variables with $E(X_{ni}) = 0$, $E(X_{ni}^2) < \infty$, $i = 1, \ldots, k_n$; $n \geq 1$, and $B_n^2 = \sum_{i=1}^{k_n} E(X_{ni}^2) \to \infty$ as $n \to \infty$. Let $F_n(x) = P\left(\sum_{i=1}^{k_n} X_{ni} < xB_n\right)$. We have

Lemma 6 *(Slastnikov, Theorem 1.1) If for sufficiently large n and some positive constant c,*

$$\sum_{i=1}^{k_n} E(|X_{ni}|^{2+c^2})\rho(|X_{ni}|)\log^{-(1+c^2)/2}(3+|X_{ni}|) \leq g(B_n)B_n^2,$$

where $\rho(t)$ is slowly varying function monotonically growing to infinity and $g(t) = o(\rho(t))$ as $t \to \infty$, then

$$1 - F_n(x) \sim 1 - \Phi(x), F_n(-x) \sim \Phi(-x), \quad n \to \infty,$$

uniformly in the region $0 \leq x \leq c\sqrt{\log B_n^2}$.

Corollary 1 *(Slastnikov, Rubin-Sethuraman) If $q > c^2 + 2$ and*

$$\sum_{i=1}^{k_n} E[|X_{ni}|^q] \leq KB_n^2,$$

then there is a sequence $\gamma_n \to 1$, such that

$$\left|\frac{1 - F_n(x) + F_n(-x)}{2(1 - \Phi(x))} - 1\right| \leq \gamma_n - 1 \to 0, \quad n \to \infty,$$

uniformly in the region $0 \leq x \leq c\sqrt{\log B_n^2}$.

Remark. Rubin-Sethuraman derived the corollary for $x = t\sqrt{\log B_n^2}$ for fixed t. Slastnikov's result adds uniformity and relaxes the moment assumption. We refer to [21] for proofs.

Next, we will state a couple of simple yet useful results. Suppose $U > 0$ is a fixed constant. For any $x = (x_1, x_2, \cdots, x_n) \in R^n$, let

$$G(x) = \sum_{i=1}^{n} |x_i|(|x_i| \wedge U),$$

where $a \wedge b$ denotes the minimum of a and b. Then we have the following results.

Lemma 7 *For any $x = (x_1, x_2, \cdots, x_n) \in R^n$, we have that*

$$G(x) \geq \begin{cases} \frac{U\|x\|_1}{2} & \text{if} \quad \|x\|_1 \geq nU/2 \\ \|x\|_2^2 & \text{if} \quad \|x\|_1 < nU/2. \end{cases}$$

Proof. Let $y = x/U$, then it is easy to see that

$$\frac{G(x)}{U^2} = \sum_{i=1}^{n} |y_i|(|y_i| \wedge 1).$$

We first consider the case where $\|y\|_1 \geq n/2$. Now suppose $|y_i| < 1$ for $i = 1, 2, \cdots, k$ (note that k might be 0 or n), and $|y_i| > 1$ for $i > k$. Then

$$\frac{G(x)}{U^2} = \|y\|_1 + \sum_{i=1}^{k} y_i^2 - \sum_{i=1}^{k} |y_i| \geq \|y\|_1 - \frac{k}{4} \geq \frac{\|y\|_1}{2}.$$

Now let us consider the case where $\|y\|_1 < n/2$. Suppose there exists an i such that $|y_i| > 1$, then there must be a j such that $|y_j| < 1/2$. If we replace y_i and y_j by $y_i' = |y_i| - \epsilon \geq 1$ and $y_j' = |y_j| + \epsilon \leq 1/2$ for some $\epsilon > 0$, the value of $G(x)/U^2$ decreases. This means that if $G(x)/U^2$ is minimized, all the y_i must satisfy that $|y_i| \leq 1$. In this case,

$$G(x)/U^2 = \|y\|_2^2.$$

Putting the above inequalities together, the lemma is proved. ■

The following lemma is from [8].

Lemma 8 *For any $x \in R^n$,*

$$\|x\|_2 - \frac{\|x\|_1}{\sqrt{n}} \leq \frac{\sqrt{n}}{4}\big(\max_{1 \leq i \leq n} |x_i| - \min_{1 \leq i \leq n} |x_i|\big).$$

Remark 8 *A interesting consequence of the above lemma is: for any* $x \in R^n$,

$$\|x\|_2 \leq \frac{\|x\|_1}{\sqrt{n}} + \frac{\sqrt{n}\|x\|_\infty}{4}$$

5.2 Proof of Lemma 1

In this section, we will prove lemma 1 by union bound and Hoeffding's inequality. Firstly, by the union bound, it can be seen that

$$P(c\sqrt{2A(\alpha)n\log p} \leq c\|S\|_\infty) \leq \sum_{i=1}^{p} P(\sqrt{2A(\alpha)n\log p} \leq |X_i'I|).$$

For each i, by Hoeffiding inequality,

$$P(\sqrt{2A(\alpha)n\log p} \leq |X_i'I|) \leq 2\exp\{-\frac{4A(\alpha)n\log p}{4\|X_i\|_2^2}\} = 2p^{-A(\alpha)},$$

since $\|X_i\|_2^2 = n$ for all i. Therefore,

$$P(c\sqrt{2A(\alpha)n\log p} \leq c\|S\|_\infty) \leq p2p^{-A(\alpha)} \leq \alpha.$$

Hence the lemma is proved.

5.3 Proof of Lemma 2

By the union bound, it can be seen that

$$P(c\sqrt{n}\Phi^{-1}(1-\alpha/(2p)) \leq c\|S\|_\infty) \leq \sum_{i=1}^{p} P(\sqrt{n}\Phi^{-1}(1-\alpha/(2p)) \leq |X_i'I|).$$

For each i, from Corollary 1,

$$\begin{aligned} & P(\sqrt{n}\Phi^{-1}(1-\alpha/(2p)) \leq |X_i'I|) \\ \leq \quad & 2(1-\Phi(\Phi^{-1}(1-\alpha/(2p))))(1+\omega_n) = \alpha/p(1+\omega_n), \end{aligned}$$

where ω_n goes to 0 as n goes to infinity, provided that $\Phi^{-1}(1-\alpha/2p) \leq (q-2)\sqrt{\log n}$. Hence

$$P(c\sqrt{n}\Phi^{-1}(1-\alpha/(2p)) \leq c\|S\|_\infty) \leq \alpha(1+\omega_n).$$

5.4 Proof of Lemma 5

It is easy to see that when $c \geq \frac{6}{a}$,

$$c - \frac{2}{a}\log(1+\frac{a}{2}c) \geq c - \frac{2}{a}\frac{ac}{4} = \frac{c}{2},$$

and when $c \leq \frac{6}{a}$,

$$c - \frac{2}{a}\log(1+\frac{a}{2}c) \geq c - \frac{2}{a}(\frac{ac}{2} - \frac{1}{8}(\frac{ac}{2})^2) = \frac{ac^2}{16}.$$

Similarly, we can show that for any real number c, when $|c| \geq \frac{6}{a}$,

$$E(|z_i + c| - |z_i|) \geq \frac{|c|}{2},$$

and when $|c| \leq \frac{6}{a}$,

$$E(|z_i + c| - |z_i|) \geq \frac{ac^2}{16}.$$

Putting the above inequalities together, the lemma is proved.

5.5 Proof of Lemma 3

First, it can be seen that for any $1 \leq i \leq n$, $||(Xd)_i - z_i| - |z_i|| \leq |(Xd)_i|$. So $|(Xd)_i - z_i| - |z_i|$ is a bounded random variable for any fixed d. Hence for any fixed k sparse signal $d \in R^p$, by Hoeffding's inequality, we have

$$P\left(B(d) \geq t\right) \leq 2\exp\{-\frac{t^2 n}{2\|Xd\|_2^2}\},$$

for all $t > 0$. From the definition of λ_k^u, we know that

$$P\left(B(d) \geq t\right) \leq 2\exp\{-\frac{t^2}{2\lambda_k^u \|d\|_2^2}\}.$$

In the above inequality, let $t = C\sqrt{2k\log p}\|d\|_2$, we have

$$P\left(B(d) \geq C\sqrt{2k\log p}\|d\|_2\right) \leq 2p^{-kC^2/\lambda_k^u}, \tag{17}$$

for all $C > 0$. Next we will find an upper bound for $\sup_{d\in R^P, \|d\|_0=k, \|d\|_2=1} |B(d)|$. We shall use the ϵ-Net and covering number argument. Consider the ϵ-Net of the set $\{d \in R^P, \|d\|_0 = k, \|d\|_2 = 1\}$. From the standard results of covering number, see for example [5], we know that the covering number of $\{d \in R^k, \|d\|_2 = 1\}$ by ϵ balls (i.e. $\{y \in R^k : \|y - x\|_2 \leq \epsilon\}$) is at most $(3/\epsilon)^k$ for $\epsilon < 1$. So the covering number of $\{d \in R^P, \|d\|_0 = k, \|d\|_2 = 1\}$ by ϵ balls is at most $(3p/\epsilon)^k$ for $\epsilon < 1$. Suppose N is such a ϵ-Net of $\{d \in R^P, \|d\|_0 = k, \|d\|_2 = 1\}$. By union bound,

$$P(\sup_{d\in N} |B(d)| \geq C\sqrt{2k\log p}) \leq 2(3/\epsilon)^k p^k p^{-kC^2/\lambda_k^u},$$

for all $C > 0$. Moreover, it can be seen that,

$$\sup_{d_1,d_2\in R^p, \|d_1-d_2\|_0\leq k, \|d_1-d_2\|_2\leq\epsilon} |B(d_1) - B(d_2)| \leq \frac{2}{\sqrt{n}}\|X(d_1 - d_2)\|_1 \leq 2\sqrt{n}\kappa_k^u\epsilon.$$

Therefore

$$\sup_{d\in R^P, \|d\|_0=k, \|d\|_2=1} |B(d)| \leq \sup_{d\in N} |B(d)| + 2\sqrt{n}\kappa_k^u\epsilon.$$

Let $\epsilon = \sqrt{\frac{2k\log p}{n}}\frac{1}{2\kappa_k^u}$, we know that

$$\begin{aligned} & P\left(\sup_{d\in R^P, \|d\|_0=k, \|d\|_2=1} |B(d)| \geq C\sqrt{2k\log p}\right) \\ \leq \quad & P\left(\sup_{d\in N} |B(d)| \geq (C-1)\sqrt{2k\log p}\right) \leq 2(\frac{3p\sqrt{n}\kappa_k^u}{p^{(C-1)^2/\lambda_k^u}})^k. \end{aligned}$$

Under the assumption that $p > n$ and $p > 3\kappa_k^u$, let $C = 1 + 2C_1\sqrt{\lambda_k^u}$ for some $C_1 > 1$, we know that

$$P\left(\sup_{d\in R^P, \|d\|_0=k, \|d\|_2=1} |B(d)| \geq (1 + 2C_1\sqrt{\lambda_k^u})\sqrt{2k\log p}\right) \leq 2p^{-4k(C_1^2-1)}. \tag{18}$$

Hence the lemma is proved.

5.6 Proof of Lemma 4

Since $||z_i + x| - |z_i|| \leq |x|$ is bounded, the expectation always exists. Suppose the density function of z_i is $f(z)$ and $x > 0$. It is easy to see that

$$\begin{aligned} E(|z_i + x| - |z_i|) &= \int_0^\infty f(t)xdt + \int_{-x}^0 f(t)(2t + x)dt - \int_{-\infty}^{-x} f(t)xdt \\ &= x(\int_{-x}^\infty f(t)dt - \int_{-\infty}^{-x} f(t)dt) + 2\int_{-x}^0 tf(t)dt \\ &= x(1 - 2P(z_i \leq -x)) + 2\int_{-x}^0 tf(t)dt. \end{aligned}$$

Hence it is easy to see that

$$\frac{dE(|z_i + x| - |z_i|)}{dx} = 1 - 2P(z_i \leq -x).$$

5.7 Proof of Theorem 1 and 3

Now we will bound the estimation error of the L_1 penalized LAD estimator. Recall that $h = \beta - \hat{\beta}$ and $h \in \Delta_{\bar{C}} = \{\delta \in R^p : \|\delta_T\|_1 \geq \bar{C}\|\delta_{T^c}\|_1\}$. Without loss of generality, assume $|h_1| \geq |h_2| \geq \cdots, \geq |h_p|$. Let $S_0 = \{1, 2, \cdots, k\}$, we have $h_{S_0} \geq \bar{C}h_{S_0^c}$. Partition $\{1, 2, \cdots, p\}$ into the following sets:

$$S_0 = \{1, 2, \cdots, k\}, S_1 = \{k + 1, \cdots, 2k\}, S_2 = \{2k + 1, \cdots, 3k\}, \cdots.$$

Then it follows from lemma 8 that

$$\begin{aligned} \sum_{i\geq 1} \|h_{S_i}\|_2 &\leq \sum_{i\geq 1} \frac{\|h_{S_i}\|_1}{\sqrt{k}} + \frac{\sqrt{k}}{4}|h_{k+1}| \leq \frac{1}{\sqrt{k}}\|h_{S_0^c}\|_1 + \frac{1}{4\sqrt{k}}\|h_{S_0}\|_1 \\ &\leq (\frac{1}{\sqrt{k}\bar{C}} + \frac{1}{4\sqrt{k}})\|h_{S_0}\|_1 \leq (\frac{1}{4} + \frac{1}{\bar{C}})\|h_{S_0}\|_2. \end{aligned} \tag{19}$$

It is easy to see that

$$\begin{aligned} &\frac{1}{\sqrt{n}}(\|Xh + z\|_1 - \|z\|_1) \geq \frac{1}{\sqrt{n}}(\|Xh_{S_0} + z\|_1 - \|z\|_1) \\ &+ \sum_{i\geq 1} \frac{1}{\sqrt{n}}(\|X(\sum_{j=0}^{i} h_{S_j}) + z\|_1 - \|X(\sum_{j=0}^{i-1} h_{S_j}) + z\|_1) \end{aligned} \tag{20}$$

Now for any fixed vector d, let

$$M(d) = \frac{1}{\sqrt{n}} E(\|Xd + z\|_1 - \|z\|_1).$$

By lemma 3, we know that with probability at least $1 - 2p^{-4k(C_2^2-1)}$,

$$\frac{1}{\sqrt{n}}(\|Xh_{S_0} + z\|_1 - \|z\|_1) \geq M(h_{S_0}) - C_1\sqrt{2k\log p}\|h_{S_0}\|_2,$$

and for $i \geq 1$ with probability at least $1 - 2p^{-4k(C_2^2-1)}$,

$$\frac{1}{\sqrt{n}}(\|X(\sum_{j=0}^{i} h_{S_j}) + z\|_1 - \|X(\sum_{j=0}^{i-1} h_{S_j}) + z\|_1) \geq M(h_{S_i}) - C_1\sqrt{2k\log p}\|h_{S_i}\|_2,$$

where $C_1 = 1 + 2C_2\sqrt{\lambda_k^u}$ and $C_2 > 1$ is a constant. Put the above inequalities together, we know that with probability at least $1 - 2p^{-4k(C_2^2-1)+1}$,

$$\frac{1}{\sqrt{n}}(\|Xh + z\|_1 - \|z\|_1) \geq M(h) - C_1\sqrt{2k\log p}\sum_{i\geq 0}\|h_{S_i}\|_2. \tag{21}$$

By this and inequality (8) and (19), we have that with probability at least $1-2p^{-4k(C_2^2-1)+1}$,

$$M(h) \leq \frac{\lambda\sqrt{k}}{\sqrt{n}}\|h_{S_0}\|_2 + C_1\sqrt{2k\log p}(1.25 + \frac{1}{\bar{C}})\|h_{S_0}\|_2. \tag{22}$$

Next, we consider two cases. First, if $\|Xh\|_1 \geq 3n/a$, then from lemma 7 and inequality (13),

$$\frac{1}{\sqrt{n}}E(\|Xh + z\|_1 - \|z\|_1) \geq \frac{3}{16\sqrt{n}}\|Xh\|_1 \geq \frac{3\sqrt{n}}{16}\kappa_k^l\|h_{S_0}\|_2. \tag{23}$$

From assumption (14), we must have $\|h_{S_0}\|_2 = 0$ and hence $\hat{\beta} = \beta$.

On the other hand, if $\|Xh\|_1 < 3n/a$, from lemma 7 and inequality (13),

$$\frac{1}{\sqrt{n}}E(\|Xh + z\|_1 - \|z\|_1) \geq \frac{a}{16\sqrt{n}}\|Xh\|_2^2. \tag{24}$$

By the same argument as in the proofs of Theorem 3.1 and Theorem 3.2 in [8], we know that

$$|\langle Xh_{S_0}, Xh\rangle| \geq n\lambda_k^l\|h_{S_0}\|_2^2 - n\theta_{k,k}\|h_{S_0}\|_2\sum_{i\geq 1}\|h_{S_i}\|_2 \geq n(\lambda_k^l - \theta_{k,k}(\frac{1}{\bar{C}} + \frac{1}{4}))\|h_{S_0}\|_2^2.$$

And

$$|\langle Xh_{S_0}, Xh\rangle| \le \|Xh_{S_0}\|_2\|Xh\|_2 \le \|Xh\|_2\sqrt{n\lambda_k^u}\|h_{S_0}\|_2.$$

Therefore

$$\|Xh\|_2^2 \ge n\frac{(\lambda_k^l - \theta_{k,k}(\frac{1}{\bar{C}} + \frac{1}{4}))^2}{\lambda_k^u}\|h_{S_0}\|_2^2.$$

Hence by (22) and (24), we know that with probability at least $1 - 2p^{-4k(C_2^2-1)+1}$,

$$\|h_{S_0}\|_2 \le \frac{16\lambda\sqrt{k}}{na\eta_k^l} + \sqrt{\frac{2k\log p}{n}}\frac{16C_1(1.25 + 1/\bar{C})}{a\eta_k^l}, \tag{25}$$

where $\eta_k^l = (\lambda_k^l - \theta_{k,k}(\frac{1}{\bar{C}} + \frac{1}{4}))^2/\lambda_k^u$. In particular, when $\lambda = 2c\sqrt{n\log p}$. Putting the above discussion together, we have

$$\|h_{S_0}\|_2 \le \sqrt{\frac{2k\log p}{n}}\frac{16(c\sqrt{2} + 1.25C_1 + C_1/\bar{C})}{a\eta_k^l}. \tag{26}$$

Since

$$\sum_{i\ge 1}\|h_{S_i}\|_2^2 \le |h_{k+1}|\sum_{i\ge 1}\|h_{S_i}\|_1 \le \frac{1}{\bar{C}}\|h_{S_0}\|_2^2,$$

We know that with probability at least $1 - 2p^{-4k(C_2^2-1)+1}$,

$$\|\hat{\beta} - \beta\|_2 \le \sqrt{\frac{2k\log p}{n}}\frac{16(c\sqrt{2} + 1.25C_1 + C_1/\bar{C})}{a\eta_k^l}\sqrt{1 + \frac{1}{\bar{C}}}.$$

where $\eta_k^l = (\lambda_k^l - \theta_{k,k}(\frac{1}{\bar{C}} + \frac{1}{4}))^2/\lambda_k^u$, $C_1 = 1 + 2C_2\sqrt{\lambda_k^u}$ and $C_2 > 1$ is a constant.

The proof of Theorem 3 is simple. In the noiseless case, we know that

$$\|Xh\|_1 \le \lambda(\|h_T\|_1 - \|h_{T^c}\|_1).$$

This means $\|h_T\|_1 \ge \|h_{T^c}\|_1$ and hence $h \in \Delta_1$. So

$$\|Xh\|_1 \ge n\kappa_k^l(1)\|h_T\|_1.$$

Since we assume that $n\kappa_k^l(1) > \lambda$, we must have $\|h\|_1 = 0$. Therefore $\hat{\beta} = \beta$.

Appendix

In the appendix, we consider the value of $\kappa_k^l(\bar{C})$ under the Gaussian random design case.

Sufficient Conditions

We will first give a set of sufficient conditions under which the value $\kappa_k^l(\bar{C})$ can be bounded below. Let $\Phi = (X_T'X_T)^{-1}X_T'$ and

$$M = \|(X_T'X_T)^{-1}X_T'X_{T^c}\|_1.$$

Here for any $n \times m$ matrix A, the matrix norm $\|A\|_1 = \sup_{h \in R^m} \frac{\|Ah\|_1}{\|h\|_1}$ denotes the maximum absolute column sum of A. Suppose $h \in \Delta_c$, i.e. $\|h_T\|_1 \geq \frac{1}{\bar{C}}\|h_{T^c}\|_1$. We have

$$\|\Phi Xh\|_1 \geq \|\Phi X_T h_T\|_1 - \|\Phi X_{T^c} h_{T^c}\|_1 \geq \|h_T\|_1 - M\|h_{T^c}\|_1 \geq (1 - M\bar{C})\|h_T\|_1.$$

Note that $M < 1$ is called exact recovery condition (ERC), see for example [6]. It is also the irrepresentable condition discussed in for example [18]. On the other hand

$$\|\Phi Xh\|_1 \leq \|\Phi\|_1 \|Xh\|_1.$$

So,

$$\|Xh\|_1 \geq \frac{1 - M\bar{C}}{\|\Phi\|_1}\|h_T\|_1 \geq \frac{1 - M\bar{C}}{\|\Phi\|_1}\|h_T\|_2.$$

We have the following lemma.

Lemma 9 *Suppose $M\bar{C} < 1$, then $\kappa_k^l(\bar{C}) \geq \frac{(1-M\bar{C})}{n\|\Phi\|_1}$.*

The values of M can be bounded by the mutual incoherence constant μ defined as $\mu = \max_{i,j} |\frac{\langle X_i, X_j \rangle}{n}|$. The following lemma can be found in [6].

Lemma 10 *Assume $\mu < \frac{1}{k-1}$, then $M < \frac{k\mu}{1-(k-1)\mu}$.*

Gaussian Random Design

Now suppose X is generated by Gaussian random design, i.e. we generate the entries of X by i.i.d. $N(0,1)$ random variables and then normalize it. We will show some asymptotical bounds on the value of $\kappa_k^l(\bar{C})$.

Lemma 11 *Under the previous assumptions, when* $\log p = o(n^{1/6})$*, we know that for any constant* $c > 0$*.*

$$P(\mu < \sqrt{\frac{2c\log p}{n}}) \geq 1 - 4p^{-(c-2)}(1 + o(1)).$$

Proof. By lemma 14, we know that for any i, j and $b = o(n^{1/6})$,

$$P(|\frac{\langle X_i, X_j\rangle}{n}| > \frac{b}{\sqrt{n}}) \leq 4e^{-b^2/2}(1 + o(1)).$$

This means

$$P(\mu > \frac{b}{\sqrt{n}}) \leq 4p^2 e^{-b^2/2}(1 + o(1)).$$

Hence when $\log p = o(n^{1/6})$, we know that

$$P(\mu < \sqrt{\frac{2c\log p}{n}}) \geq 1 - 4p^{-(c-2)}(1 + o(1)).$$

■

A simple consequence of the previous lemma is

$$P(M \leq \frac{k\sqrt{2c\log p/n}}{1 - (k-1)\sqrt{2c\log p/n}}) \geq 1 - 4p^{-(c-2)}(1 + o(1)). \tag{27}$$

Asymptotically, when $\log p = o(n^{1/n})$, we know that $M \to 0$ as n goes to infinity.

Lemma 12 *Under the previous assumptions, we have that for any* $c > 0$*,*

$$P(\|\Phi\|_1 \leq \frac{k(1 + \sqrt{\frac{(1+c)\log n}{k}})(1 + \sqrt{\frac{(1+c)\log n}{n}})}{(\sqrt{n} - \sqrt{k} - \sqrt{2c\log n})^2}) \geq 1 - \frac{3}{n^c}.$$

Proof. Since X is generate randomly, we assume N is a $n \times k$ matrix such that N_{ij} are i.i.d. $N(0,1)$ random variables for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, k$. The we generate X_T by $X_T = N \times Diag(1/l_1, 1/l_2, \cdots, 1/l_k)$ where $l_j = \sqrt{\frac{N_{1j}^2 + N_{2j}^2 + \cdots + N_{nj}^2}{n}}$. Let $L = Diag(1/l_1, 1/l_2, \cdots, 1/l_k)$, we have $\Phi = L^{-1}(N'N)^{-1}N'$. Let $N' = (N_1, N_2, \cdots, N_n)$ where N_i' is the ith row of N. Then

$$\|\Phi\|_1 = \max_i\{\|L^{-1}(N'N)^{-1}N_i\|_1\} \leq \max_i\{\sqrt{k}\|L^{-1}(N'N)^{-1}N_i\|_2\}.$$

By χ^2 tail bound, see for example [17], we know that for any $r \geq 0$,

$$P(\|N_i\|_2^2 \leq k(1 + r/2)^2) \geq 1 - e^{-kr^2/4} \text{ and } P(l_i \leq (1 + r/2)) \geq 1 - e^{-nr^2/4}.$$

By union bounds,

$$P(\sup_i \|N_i\|_2 \leq k(1 + r/2)^2) \geq 1 - ne^{-kr^2/4},$$

$$P(\|L^{-1}\|_2 \leq (1 + r/2)) \geq 1 - ke^{-nr^2/4}.$$

Also, by standard result of Gaussian random matrix (see for example [20]),

$$P(\|(N'N)^{-1}\|_2 \leq \frac{1}{(\sqrt{n} - \sqrt{k} - t)^2}) \geq 1 - e^{-t^2/2}.$$

Therefore

$$P(\|\Phi\|_1 \leq \frac{k(1 + r_1/2)(1 + r_2/2)}{(\sqrt{n} - \sqrt{k} - t)^2}) \geq 1 - ne^{-kr_1^2/4} - ke^{-nr_2^2/4} - e^{-t^2/2},$$

given that $r_1 \geq 0, r_2 \geq 0$. Specially, let $r_1 = 2\sqrt{\frac{(1+c)\log n}{k}}$, $r_2 = 2\sqrt{\frac{(1+c)\log n}{n}}$, and $t = \sqrt{2c \log n}$ for some $c > 0$,

$$P(\|\Phi\|_1 \leq \frac{k(1 + \sqrt{\frac{(1+c)\log n}{k}})(1 + \sqrt{\frac{(1+c)\log n}{n}})}{(\sqrt{n} - \sqrt{k} - \sqrt{2c \log n})^2}) \geq 1 - \frac{3}{n^c}.$$

■

Hence asymptotically, when $k = o(n)$, $\|\Phi\|_1 = O_p(\frac{k\sqrt{\log n}}{n})$. Combine the about results together, we have that for the Gaussian random design case, when $\log p = o(n^{1/6})$ and $k = o(n)$, asymptotically,

$$\kappa_k^l(\bar{C}) = O(\frac{1}{k\sqrt{\log n}}). \tag{28}$$

The following lemma shows the tail bound for sample covariance.

Lemma 13 *Suppose (X_i, Y_i) for $i = 1, 2, \cdots, n$ are i.i.d. bivariate normal random vectors such that $E(X_i) = E(Y_i) = 0$, $Var(X_i) = Var(Y_i) = 1$ and $Cov(X_i, Y_i) = \rho$. Let $\hat{\rho} = \sum_{i=1}^n X_i Y_i$ be the sample covariance, then we have*

$$P(|\hat{\rho} - \rho| > \frac{b}{\sqrt{n}}) \leq 2\exp\{-\frac{b^2}{2(1+3\rho^2)}\}(1 + o(1)), \tag{29}$$

provided that $b = o(n^{1/6})$.

Proof. Suppose $Z_i = X_i Y_i - \rho$ and let $Z = \frac{1}{n}\sum_{i=1}^n Z_i$. Then the moment generating function of Z_i is

$$\begin{aligned} E(e^{t(X_iY_i-\rho)}) &= \int\int \frac{e^{-t\rho}}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{x^2+y^2-2(\rho+t(1-\rho^2))xy}{2(1-\rho^2)}\}dxdy \\ &= \frac{e^{-t\rho}}{\sqrt{1-2\rho t-(1-\rho^2)t^2}}. \end{aligned} \tag{30}$$

Hence the moment generating function of Z is

$$E(e^{tZ}) = e^{-\rho t}(1 - 2\rho t/n - (1-\rho^2)t^2/n^2)^{-n/2}.$$

Let $\psi(t) = \log E(e^{tZ}) = -\frac{n}{2}\log(1 - 2\rho t/n - (1-\rho^2)t^2/n^2) - \rho t$. We know that for any $a > 0$ and $t > 0$,

$$P(Z > a) = P(e^{tz} > e^{ta}) \leq E(e^{tZ})e^{-ta} = e^{\psi(t)-at}.$$

Now due to the fact that $-\log(1-x) \leq x/(1-x)$, we know that

$$\begin{aligned} \psi(t) - at &\leq \frac{n}{2}\frac{2\rho t/n + (1-\rho^2)t^2/n^2}{1 - 2\rho t/n - (1-\rho^2)t^2/n^2} - \rho t - at \\ &= \frac{\frac{1}{2n}(1+3\rho^2)t^2 + (1-\rho^2)t^3/n^2}{1 - 2\rho t/n - (1-\rho^2)t^2/n^2} - at. \end{aligned}$$

Now let $t = na/(1+3\rho^2)$ and let $a = b/\sqrt{n}$, we have

$$P(Z > \frac{b}{\sqrt{n}}) \leq \exp\left\{\left(-\frac{b^2}{2(1+3\rho^2)} + \frac{1-\rho^2+2\rho(1+3\rho^2)}{(1+3\rho^2)^3}\frac{b^3}{\sqrt{n}} + \frac{1-\rho^2}{(1+3\rho^2)^3}\frac{b^4}{n}\right) \times C\right\},$$

where

$$C = \left(1 - \frac{2\rho}{1+3\rho^2}\frac{b^2}{n} - \frac{1-\rho^2}{(1+3\rho^2)^2}\frac{b^3}{n^3}\right).$$

Therefore it is easy to see that if $b = o(n^{1/6})$, then

$$P(Z > \frac{b}{\sqrt{n}}) \leq \exp\{-\frac{b^2}{2(1+3\rho^2)}\}(1+o(1)).$$

By similar argument, we can show that

$$P(|Z| > \frac{b}{\sqrt{n}}) \leq 2\exp\{-\frac{b^2}{2(1+3\rho^2)}\}(1+o(1)).$$

■

Next, we have the following tails bounds for sample correlation

Lemma 14 *Suppose (X_i, Y_i) for $i = 1, 2, \cdots, n$ are i.i.d. bivariate normal random vectors such that $E(X_i) = E(Y_i) = 0$, $Var(X_i) = \sigma_x^2$, $Var(Y_i) = \sigma_y^2$ and $Cov(X_i, Y_i) = 0$. Let*

$$\hat{\rho} = \frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i^2}}.$$

Then we have

$$P(|\hat{\rho}| \geq \frac{b}{\sqrt{n}}) \leq 4\exp\{-\frac{b^2}{2}\}(1+o(1)), \tag{31}$$

provided that $b = o(n^{1/6})$.

Proof. Suppose $r_n = b/(b+\sqrt{n/2})$, then we know that $b^2 r_n = o(1)$ and $\frac{n}{4}r_n^2 = \frac{b^2(1-r_n)^2}{2}$. We have

$$\begin{aligned} P(|\hat{\rho}| \geq \frac{b}{\sqrt{n}}) &\leq P(|\frac{\sum_{i=1}^n X_i Y_i}{n(1-r_n)\sigma_x\sigma_y}| \geq \frac{b}{\sqrt{n}}) \\ + \quad P(\frac{\sum_{i=1}^n X_i^2}{n} &< (1-r_n)\sigma_x^2) + P(\frac{\sum_{i=1}^n Y_i^2}{n} < (1-r_n)\sigma_y^2). \end{aligned}$$

Since $b = o(n^{1/6})$, we know that

$$P(|\frac{\sum_{i=1}^n X_i Y_i}{n(1-r_n)\sigma_x\sigma_y}| \geq \frac{b}{\sqrt{n}}) \leq 2\exp\{-\frac{b^2(1-r_n)^2}{2}\}(1+o(1)),$$

and the Chi-square tail bound (see for example lemma 1 in [17])

$$P(\sum_{i=1}^n X_i^2/n < (1-r_n)\sigma_x^2) \leq \exp\{-\frac{n}{4}r_n^2\},$$

$$P(\sum_{i=1}^n Y_i^2/n < (1-r_n)\sigma_y^2) \leq \exp\{-\frac{n}{4}r_n^2\}.$$

Putting the above terms together, we have

$$P(|\hat{\rho}| \geq \frac{b}{\sqrt{n}}) \leq 4\exp\{-\frac{b^2(1-r_n)^2}{2}\}(1+o(1)) = 4\exp\{-\frac{b^2}{2}\}(1+o(1)),$$

where the last equality is due to the fact that $b^2 r_n = o(1)$. Therefore the lemma is proved.

■

References

[1] Bassett, G., and Koenker, R. (1978), Asymptotic Theory of Least Absolute Error Regression. *j. amer. statist. assoc.* **73**, 618C621.

[2] Belloni, A., and Chernozhukov, V. (2011). L1-Penalized Quantile Regression in High-Dimensional Sparse Models. *Ann. Statist.* **39**, 82-130.

[3] Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, to appear.

[4] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

[5] Bourgain, J., and Milman, V. D. (1987). New volume ratio properties for convex symmetric bodies in r^n. *Invent. Math.* **88**, 319-340.

[6] Cai, T., and Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery. *IEEE Trans. Inf. Theory.* **57**, 4680-4688.

[7] Cai, T., Wang, L., and Xu, G. (2010a). Shifting Inequality and Recovery of Sparse Signals. *IEEE Trans. Signal Process.* **58**, 1300-1308.

[8] Cai, T., Wang, L., and Xu, G. (2010b). New Bounds for Restricted Isometry Constants. *IEEE Trans. Inf. Theory.* **56**, 4388-4394.

[9] Candès, E. J., Romberg, J., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59**, 1207-1223.

[10] Candès, E. J., and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313-2351.

[11] Candès, E. J., and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory.* **51**, 4203-4215.

[12] Donoho, D. (2006). Compressed sensing. *IEEE Trans. Inf. Theory.* **52**, 1289-1306.

[13] Gao, X., and Huang, J. (2010). Asymptotic Analysis of high-dimensional LAD regression with Lasso. *Statistica Sinica.* **20**, 1485-1506.

[14] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *j. amer. statist. assoc.* **58**, 13-30.

[15] Huber, P. (1981). *Robust Statistics.* Wiley, New York.

[16] Lambert-Lacroix, S., and Zwald, L. (2011). Robust regression through the Hubers criterion and adaptive lasso penalty. *Electronic Journal of Statistics.* **5**, 1015-1053.

[17] Laurent, B., and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, 1302-1338.

[18] Meinshausen, N., and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37(1)**, 2246-2270.

[19] Portnoy, S., and Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science.* **12**, 279-300.

[20] Rudelson, M., and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. *Proceedings of the International Congress of Mathematicians.* **3**, 1576-1602.

[21] Slastnikov, A. D. (1979). Limit theorems for moderate deviation probabilities. *Theo. Prob. Appl.* **23**, 322-340.

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

[23] Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection via the LAD-Lasso. *J. Business and Economic Statistics.* **25**, 347-355.

[24] Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *J. Machine Learning Res.* **10**, 555-568.