Journal of Multivariate Analysis 120 (2013) 135–151

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

The L_1 penalized LAD estimator for high dimensional linear regression

Lie Wang

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 15 May 2012
Available online 12 April 2013

AMS subject classifications:
62J05
62F99

Keywords:
High dimensional regression
LAD estimator
L_1 penalization
Variable selection

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 the 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.

1. Introduction

The 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 $Y = (y_1, y_2, \ldots, y_n)'$ is the n dimensional vector of outcomes, X is the $n \times p$ design matrix, and $z = (z_1, z_2, \ldots, z_n)'$ is the n dimensional vector of measurement errors (or noises). We assume $X = (X_1, X_2, \ldots, X_p)$ 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 $i = 1, 2, \ldots, p$. 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,

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

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.

The ordinary least squares 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. Among them, the lasso (L_1 penalized least squares) type methods have been studied in a number of papers; for example, [23,5,20]. The classical lasso estimator can be written as

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

E-mail addresses: liewang@math.mit.edu, liewang@gmail.com.

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.04.001

where λ is the penalty level (tuning parameter). In the setting of Gaussian noise and known variance, it is suggested in [5] that the penalty could be

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

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. $\|\hat{\beta}_{\text{lasso}} - \beta\|_2 \leq C(k\log(2p/\alpha)/n)^{1/2}$ for some constant $C > 0$ with probability at least $1-\alpha$.

The lasso method has nice properties, but it also relies 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, [4] proposed the square-root lasso method, where the knowledge of the distribution or variance is not required. Instead, some moment assumptions of the errors and the design matrix are needed. Other than the constrained optimization or penalized optimization methods, the stepwise algorithm are also studied; see for example [25,7]. 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 literature, 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 have a 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 are present. These methods have desired robust properties in linear regression models; see for example [2,16,21].

Recently, the penalized version of the LAD method was studied in several papers and the variable selection and estimation properties were discussed. In [14], 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 was discussed in, for example [24,18], where the number of variables p is assumed to be fixed. It is worth noting that in the proof of Lemma 1 of [24], the authors did not prove that the convergence in the last step is uniform, hence the proof is incomplete. In a recent paper [3], the quantile (and median) regression models were considered and the 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. More results about penalized median or quantile regression can be found in, for example [12,17]. In our paper, we consider the fixed design case, allow for new interesting general noise structure, and also the noiseless case. Besides, we will discuss the conditions on matrix X for the case of Gaussian random design in the Appendix.

In this paper, we present analysis for the L_1 PLAD method and we discuss the selection of the 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 is 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

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

It is important to notice that we do not have any assumptions on the 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 the ordinary LAD method package to solve the L_1 penalized LAD estimator. This is because we can consider the penalty terms as new observations, i.e. $Y_{n+i} = 0$ and $x_{n+i,j} = \lambda \times I(j = i)$ for $i, j = 1, 2, \ldots, p$. 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 the 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 that the L_1 penalized LAD method has very good numerical performance regardless of the noise distribution. Technical lemmas and 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 $i = 1, 2, \ldots, n$. 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

$$S = X'(\text{sign}(z_1), \text{sign}(z_2), \ldots, \text{sign}(z_n))',$$

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 $I = \text{sign}(z)$, then $I = (I_1, I_2, \ldots, I_n)'$ where $I_i = \text{sign}(z_i)$. 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, $S = X'I$, summarizes the estimation error in the setting of the 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 [5,3,4]. The intuition of this choice is that when the true coefficient β is a vector of 0, then the estimator should also be 0 with a given high probability. This is a general principle of choosing the penalty and can be applied to many other problems; see [4,3] for more discussions. 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. 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. Note that this penalty is provided and discussed in [3]. To approximate this value, 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 $j = 1, 2, \ldots, p$. This can be done by using Hoeffding's inequality, see for example [15], 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. Note that these choices are based on union bound and concentration inequalities. Thus when the sample size n is relatively small, these inequalities are not very tight. Hence in practice, these penalty levels tend to be relatively large and can cause additional bias to the estimator. From practical point of view, we suggest to use a smaller penalty level when the sample size is not large. See the numerical study section for more discussions. It is worth pointing out that if there exists an $i \in \{1, 2, \ldots, p\}$ 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 $\Phi^{-1}(1-\alpha/2p) \le (q-2)\sqrt{\log n}$. Then the choice of penalty $\lambda = c\sqrt{n}\Phi^{-1}(1-\frac{\alpha}{2p})$ satisfies

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

where ω_n goes to 0 as n goes to infinity.

This choice of penalty relies 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 that the penalty λ satisfies $\lambda \ge c\|S\|_\infty$ for some fixed constant $c > 1$. In what follows, for any set $E \subset \{1, 2, \ldots, p\}$ 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 indices 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 $\lambda \ge c\|S\|_\infty$, this implies the following event, namely $h = \hat\beta - \beta$ belongs to the restricted set $\Delta_{\bar C}$, where

$$\Delta_{\bar C} = \{\delta \in R^p : \|\delta_T\|_1 \ge \bar C\|\delta_{T^c}\|_1, \text{ where } T \subset \{1, 2, \ldots, p\} \text{ and } T \text{ contains at most } k \text{ elements.}\},$$

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

$$\|Xh + z\|_1 + \lambda\|\hat\beta\|_1 \le \|z\|_1 + \lambda\|\beta\|_1.$$

Let T denote the set of significant coefficients. Then

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

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

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

So

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

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 the classical lasso method and the square-root lasso method; see for example, [5,4].

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 $d \in R^p$,

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

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

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

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

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

The definition of the above constants is essentially the Restricted Isometry Constants. The maximum value of $1-\lambda_k^l$ and $\lambda_k^u - 1$ is called k-restricted isometry property (RIP) and θ_{k_1,k_2} is called the k_1, k_2-restricted orthogonality constant; see for example [11,8] for more discussion about these constants. Here we use different notations for upper and lower bounds of the restricted isometry property. It follows from [10,13] or [1] that for i.i.d. Gaussian random design, i.e. $X_{ij} \sim N(0, 1)$, for any $0 < c < 1$, there exist constants $C_1, C_2 > 0$ such that when $k \le C_1\frac{n}{\log p}$,

$$P(\max\{1-\lambda_k^l, \lambda_k^u - 1\} \le c) \ge 1 - O(e^{-C_2 p}). \tag{10}$$

Therefore when $k \log p = o(n)$ and n large enough, with high probability, λ_k^l will be bounded away from zero by any given constant between 0 and 1, and λ_k^u will be bounded above by any given constant greater than 1. Moreover, from the proof of Lemma 12 in [19] we know that in the case of Gaussian random design, the normalizing constant in our setting will not affect the above results.

We also need to define the following restricted eigenvalue of design matrix X. These definitions are based on the idea of [5]. 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 or goes to 0 slow enough. To simplify the notations, when it is not causing any confusion, we will simply write $\kappa_k^l(\bar{C})$ as κ_k^l. Please see the Appendix for more discussion about κ_k^l in the random design case.

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 $p > 3\sqrt{k}$ then*

$$P\left(\sup_{\|d\|_0 = k, \|d\|_2 = 1} B(d) \geq \left(1 + 2C_2\sqrt{\lambda_k^u}\right) \sqrt{2k \log p} \right) \leq 2p^{-4k(C_2^2 - 1)}, \tag{11}$$

where $C_2 > 1$ is a constant.

From the above lemma, we know that with probability at least $1 - 2p^{-4k(C_2^2-1)}$, for any k sparse vector $d \in R^p$,

$$\frac{1}{\sqrt{n}}(\|Xd + z\|_1 - \|z\|_1) \geq \frac{1}{\sqrt{n}} E(\|Xd + z\|_1 - \|z\|_1) - C_1\sqrt{2k \log p}\|h\|_2, \tag{12}$$

where $C_1 = 1 + 2C_2\sqrt{\lambda_k^u}$. 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 analyze than the random variable itself, this lemma plays an important role in our proof of the main theorem. Also note that the above lemma does not require the random variables z_i to have mean zero or median zero, hence it can be applied to any independent random variables.

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

$$\begin{aligned} P(z_i \geq x) &\leq \frac{1}{2 + ax} \quad \text{for all } x \geq 0 \\ P(z_i \leq x) &\leq \frac{1}{2 + a|x|} \quad \text{for all } x < 0. \end{aligned} \tag{13}$$

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 $c > 0$,

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

Hence we have the following lemma.

Lemma 5. *Suppose that the random variable z satisfies condition (13), then*

$$E(|z_i + c| - |z_i|) \geq \frac{a}{16}|c|\left(|c| \wedge \frac{6}{a}\right). \tag{14}$$

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. In order to formulate our main result we will use the following condition

$$\frac{3\sqrt{n}}{16}\kappa_k^l > \lambda\sqrt{k/n} + C_1\sqrt{2k\log p}\left(\frac{5}{4} + \frac{1}{\bar{C}}\right), \tag{15}$$

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

Theorem 1. *Consider model (1), assume $z_1, z_2, \ldots, z_n$ are independent and identically distributed random variables satisfying (13). Suppose $\lambda_k^l > \theta_{k,k}(\frac{1}{\bar{C}} + \frac{1}{4})$ and (15) holds, then the L_1 penalized LAD estimator $\hat{\beta}$ satisfies with probability at least $1 - 2p^{-4k(C_2^2-1)+1}$*

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

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 restricted isometry property (RIP) conditions; see for example [9]. We keep it here just to simplify the arguments. Note that the values of λ_k^l, λ_k^u, $\theta_{k,k}$ and κ_k^l may depend on n, p and k. For more properties of λ_k^l, λ_k^u, $\theta_{k,k}$, please see [9,1,10] and the references therein. For the value of κ_k^l, please see the Appendix for more discussion.

Remark 4. Condition (15) 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,

$$\|\hat{\beta} - \beta\|_2 = O\left(\sqrt{\frac{2k\log p}{n}}\right). \tag{16}$$

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

$$\left\{i : |\beta_i| \geq \sqrt{\frac{2k\log p}{n}}\frac{16(c\sqrt{2} + 1.25C_1 + C_1/\bar{C})}{a\left(\lambda_k^l - \theta_{k,k}\left(\frac{1}{\bar{C}} + \frac{1}{4}\right)\right)^2/\lambda_k^u}\right\} \subset \hat{T}, \tag{17}$$

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.

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

		$N(0, 1)$	$t(2)$	Cauchy
λ_1	Estimation	0.310 (0.236)	0.424 (0.330)	0.668 (0.431)
	Prediction	0.174 (0.072)	0.238 (0.101)	0.394 (0.136)
λ_2	Estimation	0.342 (0.215)	0.500 (0.279)	0.822 (0.364)
	Prediction	0.225 (0.051)	0.323 (0.067)	0.566 (0.090)
λ_3	Estimation	0.545 (0.194)	0.904 (0.248)	1.790 (0.341)
	Prediction	0.492 (0.041)	0.814 (0.049)	1.665 (0.074)
λ_4	Estimation	3.667 (0.385)	9.801 (3.580)	19.003 (22.727)
	Prediction	4.401 (0.073)	12.403 (0.583)	28.988 (3.835)

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 a 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 $k \log p = o(n)$, 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 $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)$. 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, each row of the design matrix X is generated by $N(0, \Sigma)$ distribution with Toeplitz correlation matrix $\Sigma_{ij} = (1/2)^{|j-k|}$. We then normalize the columns of X such that each column has L_2 norm $\sqrt{n}$.

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, $\lambda_1 = \sqrt{1.5n \log p}$, $\lambda_2 = \sqrt{2n \log p}$, $\lambda_3 = \sqrt{4n \log p}$, and $\lambda_4 = \sqrt{10n \log p}$. Note that they are 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 and prediction errors are summarized in Table 1.

From Table 1 we can see that λ_4 is too large in our setup and it kills a lot of significant variables. (It is worth noting that if we increase the sample size to for example $n = 500$ and $p = 1000$, λ_4 becomes a reasonable choice.) Moreover, larger λ cause more bias to the estimator. In practice, an ordinary least squares 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 average 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 prediction error $\|X\hat{\beta} - X\beta\|_2^2/n$. But λ_3 has the best results in terms of the post L_1 PLAD estimation error, which indicates that λ_3 has the best variable selection properties. We can see that the performances of λ_1, λ_2 and λ_3 are all reasonably good, which shows the L_1 PLAD method can work for a wide range of penalty levels. The post L_1 PLAD results are 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

Table 2
The average of estimation error $\|\hat{\beta} - \beta\|_2^2$ and prediction error $\|X(\hat{\beta} - \beta)\|_2^2/n$ over 200 replications and the variable selection results for lasso and the 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 $\|\hat{\beta} - \beta\|_2^2$	0	0.021	0.085	0.357	0.791
L_1 PLAD: average of $\|X(\hat{\beta} - \beta)\|_2^2/n$	0	0.014	0.060	0.223	0.544
L_1 PLAD: average type I error	0	0	0	0	0
L_1 PLAD: average type II error	0	0.170	0.265	0.250	0.23
Lasso: average of $\|\hat{\beta} - \beta\|_2$	11.419	0.012	0.049	0.209	NA
Lasso: average of $\|X(\hat{\beta} - \beta)\|_2^2/n$	0.0002	0.008	0.032	0.125	NA
Lasso: average type I error	0	0	0	0	NA
Lasso: average type II error	12.800	0.345	0.385	0.350	NA
CV lasso: average of $\|\hat{\beta} - \beta\|_2$	1.410	0.178	0.313	0.569	11.862
CV lasso: average of $\|X(\hat{\beta} - \beta)\|_2^2/n$	0.018	0.010	0.018	0.033	0.854
CV lasso: average type I error	0.035	0	0	0	1.62
CV lasso: average type II error	31.930	87.230	76.575	63.695	22.480

Table 3
The average of estimation error $\|\hat{\beta} - \beta\|_2^2$ and prediction error $\|X(\hat{\beta} - \beta)\|_2^2/n$ over 200 replications and the variable selection results for the L_1 PLAD method. Numbers in the parentheses are the averages of the estimation errors of the post L_1 PLAD method.

	Case (a)	Case (b)	Case (c)
Average of $\|\hat{\beta} - \beta\|_2^2$	0.652 (0.369)	1.091 (0.554)	0.321 (0.171)
Average of $\|X(\hat{\beta} - \beta)\|_2^2/n$	0.407 (0.085)	0.727 (0.132)	0.267 (0.041)
Average type I error	0	0	0
Average type II error	0.205	0.225	0.210

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 two ways to select the penalty. One is to use $\sigma \times \lambda$ as the penalty, where we assume that 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 summarizes the average estimation error, prediction 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 the 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 the 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 the Cauchy distribution case.

In terms of variable selection, the L_1 PLAD method does better than the classical lasso method. Both the two methods 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 in the model, its average type II errors is huge. It is worth noting that the L_1 PLAD method does a perfect job in the 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 distributed 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 $t(2)$ random variables multiplied 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 follows 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, the traditional lasso method and the constrained minimization methods cannot be properly applied since the variances of the noises are unbounded.

Table 3 summarizes the average estimation errors, prediction errors and variable selection properties of the L_1 PLAD method over 200 runs. We also summarize the estimation errors and prediction 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.

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 the following.

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

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

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 the reader to [22] for proofs.

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

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

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 $i = 1, 2, \ldots, k$ (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 $|y_i| \le 1$. In this case,

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

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

The following lemma is from [9].

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

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

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

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

5.2. Proof of Lemma 1

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

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

For each i, by Hoeffding's inequality,

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

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

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

Hence the lemma is proved.

5.3. Proof of Lemma 2

By the union bound, it can be seen that

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

For each i, from Corollary 1,

$$P(\sqrt{n}\Phi^{-1}(1-\alpha/(2p)) \le |X_i'I|) \le 2(1-\Phi(\Phi^{-1}(1-\alpha/(2p))))(1+\omega_n) = \alpha/p(1+\omega_n),$$

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

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

5.4. Proof of Lemma 5

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

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

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

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

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

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

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

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

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

$$P\left(B(d) \geq C\sqrt{2k\log p}\|d\|_2\right) \leq 2p^{-kC^2/\lambda_k^u}, \tag{18}$$

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 [6], 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,

$$P\left(\sup_{d\in N} |B(d)| \geq C\sqrt{2k\log p}\right) \leq 2(3/\epsilon)^k p^k p^{-kC^2/\lambda_k^u}, \tag{19}$$

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

$$\begin{aligned}\sup_{d_1, d_2 \in R^p, \|d_1 - d_2\|_0 \leq k, \|d_1 - d_2\|_2 \leq \epsilon} |B(d_1) - B(d_2)| &\leq \frac{2}{\sqrt{n}}\|X(d_1 - d_2)\|_1 \\ &\leq \frac{2}{\sqrt{n}} \max_{i=1,2,\ldots,p}\{\|X_i\|_1\}\|d_1 - d_2\|_1 \leq 2\sqrt{n}\sqrt{k}\epsilon.\end{aligned}$$

Therefore

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

Let $\epsilon = \sqrt{\frac{2k\log p}{n}}\frac{1}{2\sqrt{k}}$, by (19) we know that

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

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

$$P\left(\sup_{d\in R^p, \|d\|_0 = k, \|d\|_2 = 1} |B(d)| \geq (2C_1\sqrt{\lambda_k^u})\sqrt{2k\log p}\right) \leq 2p^{-4k(C_1^2 - 1)}. \tag{20}$$

Hence the lemma is proved.

5.6. Proof of Lemma 4

Since $\|z_i + x| - |z_i\| \le |x|$ 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

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

Hence it is easy to see that

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

5.7. Proof of Theorems 1 and 3

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

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

Then it follows from Lemma 8 that

$$\begin{aligned} \sum_{i\ge1} \|h_{S_i}\|_2 &\le \sum_{i\ge1} \frac{\|h_{S_i}\|_1}{\sqrt{k}} + \frac{\sqrt{k}}{4}|h_{k+1}| \le \frac{1}{\sqrt{k}}\|h_{S_0^c}\|_1 + \frac{1}{4\sqrt{k}}\|h_{S_0}\|_1 \\ &\le \left(\frac{1}{\sqrt{k}\bar{C}} + \frac{1}{4\sqrt{k}}\right)\|h_{S_0}\|_1 \le \left(\frac{1}{4} + \frac{1}{\bar{C}}\right)\|h_{S_0}\|_2. \end{aligned} \tag{21}$$

It is easy to see that

$$\begin{aligned} \frac{1}{\sqrt{n}}(\|Xh + z\|_1 - \|z\|_1) \ge{}& \frac{1}{\sqrt{n}}(\|Xh_{S_0} + z\|_1 - \|z\|_1) \\ &+ \sum_{i\ge1} \frac{1}{\sqrt{n}}\left(\left\|X\left(\sum_{j=0}^{i} h_{S_j}\right) + z\right\|_1 - \left\|X\left(\sum_{j=0}^{i-1} h_{S_j}\right) + z\right\|_1\right). \end{aligned} \tag{22}$$

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 $1 - 2p^{-4k(C_2^2-1)}$,

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

Again by Lemma 3, for any $i \ge 1$ with probability at least $1 - 2p^{-4k(C_2^2-1)}$,

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

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 $1 - 2p^{-4k(C_2^2-1)+1}$,

$$\frac{1}{\sqrt{n}}(\|Xh + z\|_1 - \|z\|_1) \ge M(h) - C_1\sqrt{2k\log p}\sum_{i\ge0}\|h_{S_i}\|_2. \tag{23}$$

By this and inequality (8) and (21), we have that with probability at least $1 - 2p^{-4k(C_2^2-1)+1}$,

$$M(h) \le \frac{\lambda\sqrt{k}}{\sqrt{n}}\|h_{S_0}\|_2 + C_1\sqrt{2k\log p}\left(1.25 + \frac{1}{\bar{C}}\right)\|h_{S_0}\|_2. \tag{24}$$

Next, we consider two cases. First, if $\|Xh\|_1 \ge 3n/a$, then from Lemma 7 and inequality (14),

$$\frac{1}{\sqrt{n}}E(\|Xh+z\|_1 - \|z\|_1) \ge \frac{3}{16\sqrt{n}}\|Xh\|_1 \ge \frac{3\sqrt{n}}{16}\kappa_k^l\|h_{S_0}\|_2. \tag{25}$$

From assumption (15), 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 (14),

$$\frac{1}{\sqrt{n}}E(\|Xh+z\|_1 - \|z\|_1) \ge \frac{a}{16\sqrt{n}}\|Xh\|_2^2. \tag{26}$$

By the same argument as in the proofs of Theorems 3.1 and 3.2 in [9], we know that

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

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

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

Hence by (24) and (26), we know that with probability at least $1 - 2p^{-4k(C_2^2-1)+1}$,

$$\|h_{S_0}\|_2 \le \frac{16\lambda\sqrt{k}}{na\eta_k^l} + \sqrt{\frac{2k\log p}{n}}\frac{16C_1(1.25 + 1/\bar{C})}{a\eta_k^l}, \tag{27}$$

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

$$\|h_{S_0}\|_2 \le \sqrt{\frac{2k\log p}{n}}\frac{16(c\sqrt{2} + 1.25C_1 + C_1/\bar{C})}{a\eta_k^l}. \tag{28}$$

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 $\hat{\beta} = \beta$.

Acknowledgment

This research was supported by NSF Grant DMS-1005539.

Appendix

In the appendix, we consider the value of $\kappa_k^l(\bar{C})$ under the Gaussian random design case. We will show that if $k\log p = o(n)$ and $p > 3$ and n large enough, $\kappa_k^l(\bar{C})$ will be bounded away from 0 by a constant with high probability, provided that $0 < \frac{1}{\bar{C}} < \sqrt{2}$. First, we define the following k sparse subset of p dimensional vectors.

$$S_k = \{h \in R^p : \|h\|_0 \le k\}, \tag{29}$$

where $\|h\|_0$ denotes the number of nonzero coordinates of vector h. Suppose X is an $n \times p$ matrix, define the following quantities.

$$\rho_k^u = \sup_{h\in S_k} \frac{\|Xh\|_1}{n\|h\|_2}, \tag{30}$$

$$\rho_k^l = \inf_{h\in S_k} \frac{\|Xh\|_1}{n\|h\|_2}. \tag{31}$$

Suppose now the entries of X are generated by independent and identically distributed $N(0, 1)$ random variables and normalized such that the L_2 norm of each column is $\sqrt{n}$. We will first bound the value of $\kappa_k^l(\bar{C})$ in terms of ρ_k^u and ρ_l^u.

Lemma 9. *For any positive integers a, b such that $a \ge k + \frac{b}{4}$. We have*

$$\kappa_k^l(\bar{C}) \ge \rho_a^l - \frac{1}{\bar{C}}\sqrt{\frac{a}{b}}\rho_b^u. \tag{32}$$

Proof. For any $h \in \Delta_{\bar{C}}$, without loss of generality, assume $|h_1| \ge |h_2| \ge \cdots \ge |h_p|$. Then we know that

$$\sum_{i=1}^{k} |h_i| \ge \bar{C} \sum_{i>k} |h_i|. \tag{33}$$

Now let $T_0 = \{1, 2, \ldots, a\}$ and $T_i = \{a + (i-1)b + 1, a + (i-1)b + 2, \ldots, a + ib\}$ for $i \ge 1$. Let $h(i) = hI_{T_i}$ for $i = 0, 1, \ldots$. It can be seen that

$$\|Xh\|_1 \ge \|Xh(0)\|_1 - \sum_{i\ge 1} \|Xh(i)\|_1 \ge n\left(\rho_a^l\|h(0)\|_2 - \rho_b^u \sum_{i\ge 1} \|h(i)\|_2\right). \tag{34}$$

Now by the Shifting inequality in [8], we know that

$$\sum_{i\ge 1} \|h(i)\|_2 \le \frac{1}{\sqrt{b}}\|h_{T^c}\|_1 \le \frac{1}{\sqrt{b}}\frac{1}{\bar{C}}\|h_T\|_1 \le \frac{1}{\bar{C}}\sqrt{\frac{a}{b}}\|h(0)\|_2. \tag{35}$$

Therefore,

$$\|Xh\|_1 \ge n\|h(0)\|_2\left(\rho_a^l - \frac{1}{\bar{C}}\sqrt{\frac{a}{b}}\rho_b^u\right). \tag{36}$$

This means $\kappa_k^l(r) \ge \rho_a^l - \frac{1}{\bar{C}}\sqrt{\frac{a}{b}}\rho_b^u$. □

The above lemma means that if there exist a and b such that $a \ge k + \frac{b}{4}$ and $\rho_a^l - \frac{1}{\bar{C}}\sqrt{\frac{a}{b}}\rho_b^u > 0$, then $\kappa_k^l(r)$ is bounded away from 0. Next, we will bound ρ_k^u and ρ_k^l. First of all, for any fixed $h \in S_k$, we have the following lemma.

Lemma 10. *Suppose the entries of X are i.i.d. $N(0, 1)$ random variables and $k\log p = o(n)$. For any fixed $h \in S_k$ and a constant $c > 0$, we have that there exists a constant $C > 0$ such that,*

$$P\left(\frac{\|Xh\|_1}{n\|h\|_2} \ge \sqrt{\frac{2}{\pi}} - c\sqrt{\frac{k\log p}{n}}\right) \ge 1 - C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2}-\frac{1}{\pi}\right)k}, \tag{37}$$

$$P\left(\frac{\|Xh\|_1}{n\|h\|_2} \le \sqrt{\frac{2}{\pi}} + c\sqrt{\frac{k\log p}{n}}\right) \ge 1 - C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2}-\frac{1}{\pi}\right)k}. \tag{38}$$

Proof. Let $Xh = z = (z_1, z_2, \ldots, z_n)^T$, then z_i are i.i.d. $N(0, \|h\|_2^2)$ random variables. Let $w_i = |z_i|/\|h\|_2 - \sqrt{2/\pi}$, then we know that $E(w_i) = 0$ and $E(w_i^2) = 1 - \frac{2}{\pi}$. Let $M(s)$ denote the moment generating function of w_i, then

$$M(s) = E(e^{s(|z_i|/\|h\|_2 - \sqrt{2/\pi})}) = e^{-s\sqrt{2/\pi}} E(e^{s|z_i|/\|h\|_2}) = e^{\frac{s^2}{2} - \sqrt{2/\pi}s} 2\Phi(s). \tag{39}$$

Therefore for any $s \geq 0$ and $t \geq 0$,

$$P\left(\sum_{i=1}^n w_i \geq t\right) = P\left(e^{s\sum_{i=1}^n w_i} \geq e^{st}\right) \leq e^{-st} E\left(e^{s\sum_{i=1}^n w_i}\right) \tag{40}$$

$$= e^{-st} e^{\frac{ns^2}{2} - n\sqrt{2/\pi}s} 2^n \Phi^n(s) = \exp\left\{\frac{ns^2}{2} - \left(t + \sqrt{\frac{2}{\pi}}\right)s + n\log(2\Phi(s))\right\}. \tag{41}$$

In the above inequality, let $s = t/n$, we have

$$P\left(\sum_{i=1}^n w_i \geq t\right) \leq \exp\left\{-\frac{t^2}{2n} - \sqrt{\frac{2}{\pi}}t + n\log\left(2\Phi\left(\frac{t}{n}\right)\right)\right\}. \tag{42}$$

By Taylor series, it can be seen that when $t = o(n)$,

$$\log\left(2\Phi\left(\frac{t}{n}\right)\right) - \sqrt{\frac{2}{\pi}}\frac{t}{n} = \frac{1}{\pi}\frac{t^2}{n^2} + O\left(\left(\frac{t}{n}\right)^3\right). \tag{43}$$

So when $t = o(n)$

$$P\left(\sum_{i=1}^n w_i \geq t\right) \leq \exp\left\{-\left(\frac{1}{2} - \frac{1}{\pi}\right)\frac{t^2}{n} + n \times O\left(\left(\frac{t}{n}\right)^3\right)\right\}. \tag{44}$$

Let $t = c\sqrt{nk\log p}$ for some constant $c > 0$. Since $k\log p = o(n)$, we know that $t = o(n)$ and $\frac{t^3}{n^2} = o(\frac{t^2}{n})$. So we know that there exists a constant $C > 0$ such that when n is large enough,

$$P\left(\sum_{i=1}^n w_i \geq c\sqrt{nk\log p}\right) \leq C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2} - \frac{1}{\pi}\right)k}. \tag{45}$$

By a similar argument, we can prove that when $k\log p = o(n)$, for any $c > 0$ there exists a $C > 0$ such that

$$P\left(\sum_{i=1}^n w_i \leq -c\sqrt{nk\log p}\right) \leq C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2} - \frac{1}{\pi}\right)k}. \tag{46}$$

The lemma is proved. □

Lemma 11. *Suppose X is an $n \times p$ matrix and the entries of X are i.i.d. $N(0, 1)$ random variables. Let Z be another $n \times p$ matrix such that $Z_{ij} = \frac{\sqrt{n}X_{ij}}{\sqrt{\sum_{k=1}^n X_{kj}^2}}$, i.e. we normalize the columns of X such that the L_2 norm of each column is $\sqrt{n}$. Suppose $k\log p = o(n)$, then for any constant $c > 0$, we have that there exists a constant $C > 0$ such that for any $0 \leq t_1 < 1/2$,*

$$P\left(\frac{\|Zh\|_1}{n\|h\|_2} \geq \frac{1}{1 - t_1}\left(\sqrt{\frac{2}{\pi}} - c\sqrt{\frac{k\log p}{n}}\right)\right) \leq C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2} - \frac{1}{\pi}\right)k} + pe^{-3nt_1^4/16}, \tag{47}$$

$$P\left(\frac{\|Zh\|_1}{n\|h\|_2} \leq \frac{1}{1 + t_1}\left(\sqrt{\frac{2}{\pi}} - c\sqrt{\frac{k\log p}{n}}\right)\right) \leq C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2} - \frac{1}{\pi}\right)k} + pe^{-3nt_1^4/16}. \tag{48}$$

Proof. Let D be a $p \times p$ diagonal matrix such that $D_{jj} = \frac{\sqrt{n}}{\sqrt{\sum_{k=1}^n X_{kj}^2}}$ for $j = 1, 2, \ldots, p$. Then we have $Z = XD$. This means

$$\frac{\|Zh\|_1}{n\|h\|_2} = \frac{\|XDh\|_1}{n\|h\|_2} = \frac{\|XDh\|_1}{n\|Dh\|_2}\frac{\|Dh\|_2}{\|h\|_2}. \tag{49}$$

By Lemma 8 from [19], we know that for any $0 \le t < 1/2$,

$$P\left(\left|\frac{\|h\|_2}{\|Dh\|_2} - 1\right| \ge t\right) \le 1 - pe^{-3nt^4/16}. \tag{50}$$

Therefore, by Lemma 10, the results are proved. □

Lemma 12. *Suppose the $n \times p$ matrix Z is generated in the same way as in the previous lemma. Suppose $k \log p = o(n)$, then for any constant $c > 0$, there exists a constant $C > 0$ such that for any $0 \le t_1 < 1/2$ and $0 < \epsilon < 1$,*

$$P\left(\rho_k^u \ge \frac{\sqrt{\frac{2}{\pi}} + c\sqrt{\frac{k\log p}{n}}}{(1-\epsilon)(1-t_1)}\right) \le (3p/\epsilon)^k \left(C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2}-\frac{1}{\pi}\right)k} + pe^{-3nt_1^4/16}\right), \tag{51}$$

$$P\left(\rho_k^l \le \frac{\sqrt{\frac{2}{\pi}} - c\sqrt{\frac{k\log p}{n}}}{1+t_1} - \epsilon\rho_k^u\right) \le (3p/\epsilon)^k \left(C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2}-\frac{1}{\pi}\right)k} + pe^{-3nt^4/16}\right). \tag{52}$$

Proof. We shall use the ϵ-Net and the covering number argument as in the proof of Lemma 3. The covering number of $\{h \in R^p, \|h\|_0 = k, \|h\|_2 = 1\}$ by ϵ balls is at most $(3p/\epsilon)^k$ for $\epsilon < 1$. Suppose N is such an ϵ-Net. By union bound,

$$P\left(\sup_{h\in N}\frac{\|Zh\|_1}{n\|h\|_2} \ge \frac{\sqrt{\frac{2}{\pi}} + c\sqrt{\frac{k\log p}{n}}}{1-t_1}\right) \le (3p/\epsilon)^k \left(C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2}-\frac{1}{\pi}\right)k} + pe^{-3nt_1^4/16}\right), \tag{53}$$

$$P\left(\inf_{h\in N}\frac{\|Zh\|_1}{n\|h\|_2} \le \frac{\sqrt{\frac{2}{\pi}} - c\sqrt{\frac{k\log p}{n}}}{1+t_1}\right) \le (3p/\epsilon)^k \left(C\left(\frac{1}{p}\right)^{c^2\left(\frac{1}{2}-\frac{1}{\pi}\right)k} + pe^{-3nt_1^4/16}\right). \tag{54}$$

Now for any h such that $\|h\|_0 = k$ and $\|h\|_2 = 1$, there exists a $h_c \in N$ such that

$$\frac{\|Zh_c\|_1}{n} - \frac{\|Z(h-h_c)\|_1}{n} \le \frac{\|Zh\|_1}{n} \le \frac{\|Zh_c\|_1}{n} + \frac{\|Z(h-h_c)\|_1}{n}. \tag{55}$$

Therefore, we know that

$$\rho_k^u \le \sup_{h\in N}\frac{\|Zh\|_1}{n\|h\|_2} + \epsilon\rho_k^u, \tag{56}$$

$$\rho_k^l \ge \inf_{h\in N}\frac{\|Zh\|_1}{n\|h\|_2} - \epsilon\rho_k^u. \tag{57}$$

Then by (53) and (54), the lemma is proved. □

Now assume that $k \log p = o(n)$ and $p > 3$. In the previous lemma, let $c = \frac{6\pi}{\pi-2}$, $t_1 = 2.5(\frac{k\log p}{n})^{1/4}$, and $\epsilon = 3/p$, we have that when n large enough,

$$P\left(\rho_k^u \le \frac{p}{(p-3)\left(1-2.5\left(\frac{k\log p}{n}\right)^{1/4}\right)}\left(\sqrt{\frac{2}{\pi}} + \frac{6\pi}{\pi-2}\sqrt{\frac{k\log p}{n}}\right)\right) \ge 1 - O(p^{-k}), \tag{58}$$

$$P\left(\rho_k^l \ge \frac{1}{1-2.5\left(\frac{k\log p}{n}\right)^{1/4}}\left(\sqrt{\frac{2}{\pi}} + \frac{6\pi}{\pi-2}\sqrt{\frac{k\log p}{n}}\right) - \frac{3\rho_k^u}{p}\right) \ge 1 - O(p^{-k}). \tag{59}$$

Therefore when n goes to infinity, both ρ_k^u and ρ_k^l converge to $\sqrt{\frac{2}{\pi}}$ in probability. Now for any $0 \le \frac{1}{C} < \sqrt{2}$, suppose we choose $a = \frac{k}{1-r^2/2}$ and $b = 2r^2a$ in Lemma 9, then we have that when n large enough, $\kappa_k^l(r)$ is bounded away from zero by a constant.

References

[1] R. Baraniuk, M.A. Davenport, M.F. Duarte, C. Hegde, An Introduction to Compressive Sensing, CONNEXIONS, Rice University, Houston, Texas, 2010.
[2] G. Bassett, R. Koenker, Asymptotic theory of least absolute error regression, J. Amer. Statist. Assoc. 73 (1978) 618–621.
[3] A. Belloni, V. Chernozhukov, L_1-penalized quantile regression in high-dimensional sparse models, Ann. Statist. 39 (2011) 82–130.
[4] A. Belloni, V. Chernozhukov, L. Wang, Square-root lasso: pivotal recovery of sparse signals via conic programming, Biometrika 98 (2011) 791–806.
[5] P.J. Bickel, Y. Ritov, A.B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. 37 (2009) 1705–1732.
[6] J. Bourgain, V.D. Milman, New volume ratio properties for convex symmetric bodies in r^n, Invent. Math. 88 (1987) 319–340.
[7] T. Cai, L. Wang, Orthogonal matching pursuit for sparse signal recovery, IEEE Trans. Inform. Theory 57 (2011) 4680–4688.
[8] T. Cai, L. Wang, G. Xu, Shifting inequality and recovery of sparse signals, IEEE Trans. Signal Process. 58 (2010) 1300–1308.
[9] T. Cai, L. Wang, G. Xu, New bounds for restricted isometry constants, IEEE Trans. Inform. Theory 56 (2010) 4388–4394.
[10] E.J. Candès, Compressive sampling, in: Proc. Int. Congr. Math., Vol. 3, 2006, pp. 1433–1452.
[11] E.J. Candès, T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory 51 (2005) 4203–4215.
[12] X. Chen, Z. Wang, M. McKeown, Asymptotic analysis of robust LASSOs in the presence of noise with large variance, IEEE Trans. Inform. Theory 56 (2010) 5131–5149.
[13] D. Donoho, Compressed sensing, IEEE Trans. Inform. Theory 52 (2006) 1289–1306.
[14] X. Gao, J. Huang, Asymptotic analysis of high-dimensional LAD regression with Lasso, Statist. Sinica 20 (2010) 1485–1506.
[15] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.
[16] P. Huber, Robust Statistics, Wiley, New York, 1981.
[17] R. Koenker, Quantile Regression, in: Econometric Society Monographs, Cambridge University Press, 2005.
[18] S. Lambert-Lacroix, L. Zwald, Robust regression through the Huber's criterion and adaptive lasso penalty, Electron. J. Stat. 5 (2011) 1015–1053.
[19] H. Liu, L. Wang, TIGER: a tuning-insensitive approach for optimally estimating large undirected graphs, Technical Report, 2012.
[20] N. Meinshausen, B. Yu, Lasso-type recovery of sparse representations for high-dimensional data, Ann. Statist. 37 (1) (2009) 2246–2270.
[21] S. Portnoy, R. Koenker, The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators, Statist. Sci. 12 (1997) 279–300.
[22] A.D. Slastnikov, Limit theorems for moderate deviation probabilities, Theory Probab. Appl. 23 (1979) 322–340.
[23] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[24] H. Wang, G. Li, G. Jiang, Robust regression shrinkage and consistent variable selection via the LAD-Lasso, J. Bus. Econom. Statist. 25 (2007) 347–355.
[25] T. Zhang, On the consistency of feature selection using greedy least squares regression, J. Mach. Learn. Res. 10 (2009) 555–568.