Asymptotic analysis of high-dimensional LAD regression with Lasso

By: Xiaoli Gao and Jian Huang

Gao, X.L. and Huang, J. (2010). Asymptotic analysis of high-dimensional LAD regression with Lasso. Statistica Sinica, 20, 1485-1506.

Abstract:

The Lasso is an attractive approach to variable selection in sparse, highdimensional regression models. Much work has been done to study the selection and estimation properties of the Lasso in the context of least squares regression. However, the least squares based method is sensitive to outliers. An alternative to the least squares method is the least absolute deviations (LAD) method which is robust to outliers in the responses. In this paper, we study the selection and estimation properties of the Lasso in LAD regression. We provide sufficient conditions under which the LAD-Lasso is estimation or selection consistent in sparse, high-dimensional settings. We use simulation studies to evaluate the performance of the LAD-Lasso, and compare the proposed method with the LS-Lasso in a range of generating models.

Keywords: Consistency | high-dimensional model | Lasso | robust regression | sparsity | variable selection

Article:

***Note: Full text of article below**

Statistica Sinica **20** (2010), 1485-1506

ASYMPTOTIC ANALYSIS OF HIGH-DIMENSIONAL LAD REGRESSION WITH LASSO

Xiaoli Gao and Jian Huang

Oakland University and University of Iowa

Abstract: The Lasso is an attractive approach to variable selection in sparse, high-dimensional regression models. Much work has been done to study the selection and estimation properties of the Lasso in the context of least squares regression. However, the least squares based method is sensitive to outliers. An alternative to the least squares method is the least absolute deviations (LAD) method which is robust to outliers in the responses. In this paper, we study the selection and estimation properties of the Lasso in LAD regression. We provide sufficient conditions under which the LAD-Lasso is estimation or selection consistent in sparse, high-dimensional settings. We use simulation studies to evaluate the performance of the LAD-Lasso, and compare the proposed method with the LS-Lasso in a range of generating models.

Key words and phrases: Consistency, high-dimensional model, Lasso, robust regression, sparsity, variable selection.

1. Introduction

Consider a linear regression model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where y_i is the response variable, x_{ij}'s are covariates or design variables, and ε_i is the error term. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. The LAD-Lasso estimator is the value $\widehat{\boldsymbol{\beta}}_n$ that minimizes the criterion

$$L_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} x_{ij}\beta_j| + \lambda_n \sum_{j=1}^{p} |\beta_j|, \tag{1.2}$$

where λ_n is a penalty parameter. We are interested in the statistical properties of $\widehat{\boldsymbol{\beta}}_n$ in sparse, high-dimensional settings. We provide conditions under which $\widehat{\boldsymbol{\beta}}_n$ is estimation consistent and/or variable-selection consistent.

High-dimensional data arise in many important applications. For example, in studies involving microarray gene expression data, the total number of covariates

p is much larger than sample size n, but the number of important covariates is typically smaller than n. Penalized methods have emerged as effective in analyzing such data. A popular approach is the Lasso (Tibshirani (1996)) that uses the ℓ_1 penalty (Chen and Donoho (1995)). This method was proposed in the context of least squares (LS) regression and likelihood estimation in generalized linear models, but it is conceptually straightforward to apply the ℓ_1 penalty to other models, such as LAD regression.

LAD regression is an interesting and robust alternative to the LS method, which is known to be sensitive to outliers. There is a large body of literature on the theoretical properties of and computational methods for the LAD estimators (see e.g., Bassett and Koenker (1978); Koenker and Bassett (1978); Pollard (1991); Portnoy and Koenker (1997)). These studies have focused on the "small p, large n" settings. Several studies have investigated the properties of M-estimators with a divergent number of covariates (Huber (1981) and Portnoy (1984, 1985)). In particular, Portnoy (1984, 1985) studied both the consistency and the asymptotic normality of a class of M-estimators under certain conditions on the growth rate of p as a function n. However, Portnoy did not consider penalized regression or selection of variables in sparse models.

Recently there has been much work on least squares (LS) regression with the Lasso. Many interesting results have been obtained regarding to its variable selection, estimation, and prediction properties in both "small p, large n" and "large p, small n" settings. Examples include Knight and Fu (2000); Greenshtein and Ritov (2004); Leng, Lin, and Wahba (2006); Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Meinshausen and Yu (2009); van de Geer (2008); and Zhang and Huang (2008), among others. In particular, van de Geer (2008) studied the Lasso in high-dimensional generalized linear models. She obtained results on the ℓ_1 error of the Lasso estimator by focusing on prediction error. Zhang and Huang (2008) introduced a sparse Riesz condition on the correlation of designed covariates. They showed that the LS-Lasso selects a model of the right order of dimensionality, controls the bias of the selected model at a level determined by the contributions of small regression coefficients and threshold bias, and selects all coefficients of greater order than the bias of the selected model. An important aspect of the results of Zhang and Huang (2008) is that the logarithm of the number of variables can be of the same order as the sample size under certain conditions. Zhao and Yu (2006) showed that the irrepresentable condition is sufficient and almost necessary for the LS-Lasso to possess the model selection consistency property.

The aforementioned work significantly advanced our understanding of the Lasso in high-dimensional settings. However, those results are obtained in the context of least squares (LS) regression and, in particular, make use of the nice

properties of a least squares loss function, e.g., the convenient geometry associated with the ℓ_2 norm and the characterization of the whole path of the LS-Lasso estimator developed in the least angle regression (LARS) algorithm. It is more difficult to deal with the LAD loss function since it is not differentiable at zero, and there is no simple geometry associated with it. Even in the standard fixed dimensional settings, rigorous analysis of the (non-penalized) LAD estimator is quite involved.

The remainder of this paper is organized as follows. In Section 2, we study the asymptotic properties of the LAD-Lasso estimator in high-dimensional settings. In Section 3, we present an almost sufficient and necessary condition under which the LAD-Lasso estimator is model selection consistent. In Section 4, we consider the computation and tuning parameter selection of the LAD-Lasso. Simulation studies are reported in Section 5. The discussion and proofs of main results are given in Section 6 and 7, respectively.

2. Asymptotic Properties of the LAD-Lasso

In this section, we study the estimation consistency of the LAD-Lasso estimator in both fixed p and large p cases.

Take $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})'$ as the true model in (1.1). Let $A_1 = \{j : \beta_{0j} \neq 0, 1 \leq j \leq p\}$. We can rearrange the covariates such that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}'_{\mathbf{10}}, \boldsymbol{\beta}'_{\mathbf{20}})'$, where $\boldsymbol{\beta}_{\mathbf{10}} = (\beta_{0j}, j \in A_1)'$ consists of all important covariates, and $\boldsymbol{\beta}_{\mathbf{20}} = \mathbf{0}$ consists of all 0 elements. Let $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})'$ and $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})'$, designed matrix $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Covariates are assumed fixed. We make the following assumptions.

(A1) The ε_i's are independent and identical distributed with median 0 and a continuous, positive density f in a neighborhood of 0.

(A2) With $\boldsymbol{\Sigma}^n \equiv n^{-1}\mathbf{X}'_n\mathbf{X}_n$, there exists a positive definite matrix $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma}^n \to \boldsymbol{\Sigma}$. If τ_{1n} and τ_{2n} are the minimum and maximum eigenvalues of $\boldsymbol{\Sigma}^n$, there exist constants $0 < \tau_1 < \tau_2 < \infty$ such that $\tau_1 \leq \tau_{1n} \leq \tau_{2n} \leq \tau_2$ for all n.

Condition (A1) is standard in the LAD regression literature, and covers very general error distributions without assuming the existence of moments. For instance, (A1) covers the t, double exponential and Cauchy distributions. (A2) is a common condition in linear regression that ensures identifiability of the regression parameters.

Theorem 1. *Let $\widehat{\boldsymbol{\beta}}_n$ be the LAD-Lasso estimator corresponding to a sequence λ_n in* (1.2). *If* (A1) *and* (A2) *hold, then $\widehat{\boldsymbol{\beta}}_n$ has the following asymptotic properties.*

- (*Consistency*) *If* $\lambda_n = o(n)$, *then* $\widehat{\boldsymbol{\beta}}_n \to_P \boldsymbol{\beta}_0$.
- (*Asymptotic distribution*) *If* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, *then* $\widehat{\boldsymbol{\beta}}_n$ *has the limiting distribution*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to_d \arg\min(V(\mathbf{u})),$$

where

$$V(\mathbf{u}) = -2\mathbf{u}'\mathbf{W} + f(0)\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u} + \lambda_0 \sum_{j=1}^{p} [u_j sign(\beta_{0j}) I(\beta_{0j} \neq 0) + |u_j| I(\beta_{0j} = 0)].$$

Here $\mathbf{W}$ *is a random vector with a* $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}/4)$ *distribution.*

Theorem 1 can be proved using an approach similar to that of Knight and Fu (2000), who studied the asymptotic properties of the LS-Lasso estimator for fixed p. We omit the proof.

Theorem 1 indicates that the right rate of growth for λ_n is $\sqrt{n}$. If $\lambda_n/\sqrt{n} \to 0$, then asymptotically, the LAD-Lasso behaves like the non-penalized LAD estimator: it is asymptotically normal with root-n rate of convergence, but does not perform variable selection. If $\lambda_n/\sqrt{n} \to \infty$, then the LAD-Lasso can be quite biased and does not have a proper asymptotic distribution at the root-n rate. When $\lambda_n/\sqrt{n} \to \lambda_0$ with $0 < \lambda_0 < \infty$, the asymptotic distribution of the LAD-Lasso exists but is in general not normal; it puts positive probability at zero when $\beta_{0j} = 0$, which reflects the fact that the LAD-Lasso estimates may take exact zero values and does variable selection.

Theorem 1 provides some insight into the behavior of the LAD-Lasso when p is fixed. However, it is not applicable when p diverges with n. Here the problem is more difficult. Indeed, when $p > n$, the model is in general not identifiable. To ensure identifiability of the model and consistency of the LAD-Lasso estimator, assumptions on the sparsity of the model and other regularity conditions are needed.

Let A be any subset of $\{1, \ldots, p\}$ and $\mathbf{X}_A = (\mathbf{x}_j, j \in A)$. We sometimes write $p_n = p$ to indicate that p can diverge with n. For any positive integer $m \leq p_n$, let

$$c_{\min}(m) = \min_{|A|=m} \min_{\|\boldsymbol{\nu}\|_2=1} \frac{1}{n} \boldsymbol{\nu}'\mathbf{X}_A'\mathbf{X}_A\boldsymbol{\nu} \text{ and } c_{\max}(m) = \max_{|A|=m} \max_{\|\boldsymbol{\nu}\|_2=1} \frac{1}{n} \boldsymbol{\nu}'\mathbf{X}_A'\mathbf{X}_A\boldsymbol{\nu}.$$

We make the following assumptions.

(B1) (Random errors) (A1) holds.

(B2) There is a positive constant b_0, such that $|x_{ij}| < b_0$ for all i and j, and $\sum_{i=1}^{n} x_{ij}^2 = n$ for all j.

(B3) (Sparse model) There is a positive M_1 such that $|A_1| \le M_1 n^2/\lambda_n^2$.

(B4) There exist constants $0 < c_* < c^* < \infty$ such that, for any sufficiently large n,

(a) $0 < c_{\min}(\min\{n, p_n\}) \le c_{\max}(\min\{n, p_n\}) < c^* < \infty$;

(b) $c_{\min}(d_n) > c_*$ for $d_n \le Mn^2/\lambda_n^2$, where M is a positive constant.

(B2) assumes that the covariates are bounded. (B3) restricts the number of nonzero coefficients in the true model. Although $p_n \gg n$, the number of true important covariates is bounded at some rate. (B4) assumes that the eigenvalues of the correlation matrix are bounded and that the eigenvalues of any submatrix of the correlation matrix with dimension $O(n^2/\lambda_n^2)$ are bounded away from zero.

Theorem 2. *If* $\lambda_n^4/n^3 = O(1)$ *and* (B1)−(B4) *hold, then*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 = O\left(\lambda_n^2 |A_1| n^{-2} c_*^{-2} f^{-1}(0)\right) + O_P\left(d_n \log(2p_n) n^{-1} c_*^{-2} f^{-1}(0)\right), \quad (2.1)$$

where $d_n = 2(M_1 + 5c^*/4)(n^2/\lambda_n^2)$, *and* A_1, M_1, c_*, *and* c^* *are defined in* (B3) *and* (B4).

The theorem states that the LAD-Lasso estimator is consistent for properly selected values of λ_n, even when p_n increases almost exponentially with n as long as $\log(p_n) = O(n^\alpha)$ for some $0 < \alpha < 1$. It also makes it clear that there is a trade-off between bias and variance in the selection of the tuning parameter λ_n: the first term on the right side of (2.1) is the bias of the LAD-Lasso estimator, and the second term is the variance; a smaller λ_n means smaller bias but larger variance. The proof of Theorem 2 is given in Section 7.

3. Model Selection Consistency of the LAD-Lasso

The estimation consistency result in Section 2 does not imply that the LAD-Lasso estimator is model selection consistent. In this section, we assume the irrepresentable condition for the LAD-Lasso estimator and generalize the model selection consistency of the LS-Lasso to the LAD-Lasso in both fixed p and $p \gg n$ cases. Let $\mathbf{X_1}_n = (\mathbf{x}_j,\ j \in A_1)$ and $\mathbf{s_1}_n = (s_j,\ \ j \in A_1)'$, where $s_j = \operatorname{sign}(\beta_{0j})$. Let $\boldsymbol{\Sigma}_{\mathbf{11}}^n = n^{-1}\mathbf{X}_{\mathbf{1}n}'\mathbf{X_1}_n$, the correlation matrix resulting from all the important covariates, and $\boldsymbol{\Sigma}_{\mathbf{11}}^n$ is invertible. Zhao and Yu (2006) introduced the following irrepresentable conditions.

Definition 1. If there exists a constant $0 < \delta < 1$, such that

$$|n^{-1}\mathbf{x}_j'\mathbf{X_1}_n(\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1}s_{\mathbf{1}n}| \le 1 - \delta$$

for $\forall\ j \notin A_1$, then the covariates satisfy the strong irrepresentable condition.

Definition 2. If $|n^{-1}\mathbf{x}_j'\mathbf{X}_{\mathbf{1}1}(\mathbf{\Sigma}_{\mathbf{11}}^n)^{-1}s_{\mathbf{1}n}| < 1 \quad \text{for } \forall\, j \notin A_1$, the covariates satisfy the weak irrepresentable condition.

Following Zhao and Yu (2006), we define the sign consistency of the LAD-Lasso estimator as follows.

Definition 3. A LAD-Lasso estimator is strongly sign consistent if there exists a sequence of λ_n such that $\lim_{n\to\infty}\mathbf{P}(\widehat{\boldsymbol{\beta}}_n(\lambda_n) =_s \boldsymbol{\beta}_0) = 1$, where $\widehat{\boldsymbol{\beta}}_n(\lambda_n) =_s \boldsymbol{\beta}_0$ means that $\widehat{\boldsymbol{\beta}}_n(\lambda_n)$ and $\boldsymbol{\beta}_0$ have the same sign component-wisely.

Definition 4. If $\lim_{n\to\infty}\mathbf{P}(\exists\ \lambda > 0,\ \widehat{\boldsymbol{\beta}}_n(\lambda) =_s \boldsymbol{\beta}_0) = 1$, the LAD-Lasso estimator is general sign consistent.

We note that variable selection consistency only requires all zeros in $\boldsymbol{\beta}_0$ to be matched, not the signs. So sign consistency defined here is stronger than the usual variable selection consistency. We evaluate the model selection properties of the LAD-Lasso by investigating sign consistency.

Theorem 3. *Let p be fixed. Suppose that* (A1), (A2), *and the strong irrepresentable condition are satisfied. We have $\mathbf{P}(\widehat{\boldsymbol{\beta}}_n(\lambda_n) =_s \boldsymbol{\beta}_0) = 1$ for $\lambda_n = O(n^{\pi_2})$, where $(1+\pi_1)/2 < \pi_2 < 1$ for some $0 < \pi_1 < 1$.*

For fixed p, Theorem 3 provides sufficient conditions under which the LAD-Lasso is strongly sign consistent. In particular, under those conditions, the LAD-Lasso can distinguish zero coefficients from nonzero coefficients with probability converging to one.

Theorem 4. *For fixed p, under* (A1) *and* (A2), *the LAD-Lasso cannot be general sign consistent if the weak irrepresentable condition fails.*

The proof of this result is similar to the proof in Zhao and Yu (2006). However, the non-differentiability of the absolute value function at zero requires careful attention, and we omit the proof. By Theorems 3 and 4, the irrepresentable condition is almost sufficient and necessary for sign consistency of the LAD-Lasso.

When $p \gg n$, the assumptions and regularity conditions in (A2) are inappropriate since $\mathbf{\Sigma}^n$ may not converge as n grows. In this case, some structural conditions on the model are required.

(C1) (A1) holds.

(C2) (B2) holds.

(C3) If $b_{n1} = \min_{j\in A_1}|\beta_{0j}|$, then

(a) there exists $0 \le c_1 < 1/2$ such that $|A_1| = O(n^{c_1})$;

(b) there exist positive constants M_0 and $c_2 > c_1$ such that $n^{(1-c_2)/2} b_{n1} \ge M_0$.

(C4) There exist constants c_* and c^* such that, for any $m_n = O(n^{c_1})$, we have

$$0 < c_* < c_{\min}(m_n) < c_{\max}(m_n) \le c_{\max}(n) < c^* < \infty.$$

In (C3), (a) assumes that the number of nonzero coefficients increases with n at a slower rate than root n; (b) assumes that the true nonzero coefficients cannot be too small. (C4) assumes that the correlation matrix satisfies the sparse Riesz condition on the rank of $O(|A_1|)$.

Theorem 5. *Suppose* (C1)−(C4) *and the strong irrepresentable condition are satisfied. The LAD-Lasso is strong sign consistent even if* $p_n = O(\exp\{n^{c_3}\})$ *for* $c_3 < \min\{c_2 - c_1, 1 - 2c_1, 1/2\}$. *In particular, if* $\lambda_n = O(n^{(1+c_4)/2})$ *with* $c_3 < c_4 < \min\{c_2 - c_1, 1 - 2c_1, 1/2\}$, *then* $\mathbf{P}(\widehat{\boldsymbol{\beta}}_n(\lambda_n) =_s \boldsymbol{\beta}_0) \to 1$ as $n \to \infty$.

This theorem investigates the model selection property of the LAD-Lasso in high-dimensional settings where p may grow with n at an almost exponential rate. In addition to the usual regularity conditions, the keys for the LAD-Lasso estimator to be selection consistent are the sparsity of the model, the sparse Riesz condition, and the strong irrepresentable condition. The sparsity condition restricts the growth rate of the number of important covariates in the true model; the sparse Riesz condition ensures identifiability of the nonzero coefficients.

4. Computation of the LAD-Lasso

In this section, we describe an approach to computing the LAD-Lasso estimator and discuss how to choose the tuning parameter.

For any given λ_n, we consider an augmented data set $\{(y_i^*, x_{i1}^*, \cdots x_{ip}^*)\}$, $1 \le i \le n + p$. Here $\{(y_i^*, x_{i1}^*, \cdots x_{ip}^*)\} = \{(y_i, x_{i1}, \cdots x_{ip})\}$, $1 \le i \le n$ and $\{(y_i^*, x_{i1}^*, \cdots x_{ip}^*)\} = \{(0, \lambda_n \mathbf{e}'_{i-n})\}$, $n + 1 \le i \le n + p$, and $\mathbf{e}_i$ is a p-dimensional unit vector with ith element equal to 1. We can rewrite $L_n(\boldsymbol{\beta})$ as

$$\sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} x_{ij}\beta_j| + \lambda_n \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n+p} \Big| y_i^* - \sum_{j=1}^{p} x_{ij}^* \beta_j \Big|.$$

Thus, we can compute the penalized estimator using any method for solving standard ℓ_1 minimization problem. For instance, if p is not very large, the QUANTREG package can be used to find $\widehat{\boldsymbol{\beta}}_n(\lambda_n)$. For large p, Wu and Lange (2008) proposed a very fast and efficient greedy descent algorithm.

The Akaike Information Criterion (AIC, Akaike (1973)) and the Bayesian Information Criterion (BIC, Schwarz (1978)) are two criteria to choose the prediction optimal ℓ_1 regularization parameter. If we assume that model errors are double exponentially and independently distributed, then AIC and BIC scores can be calculated using AIC $= n\log(\text{RSA}/n) + df$, and BIC $= n\log(\text{RSA}/n) + df\log(n)/2$, where RSA $= \sum_{i=1}^{n} |y_i - \widehat{\boldsymbol{\beta}}_n' x_i|$, and df is degrees of freedom of the model. Zou, Hastie, and Tibshirani (2007) proved that the number of nonzero coefficients in a LS-Lasso estimate is an unbiased estimate of degrees of freedom for the LS-Lasso. Similarly, we estimate degrees of freedom of the selected LAD-Lasso model by

$$\widehat{df} = \widehat{df}(\lambda_n) = \text{ the size of the } \{j : \ \widehat{\beta}_{nj}(\lambda_n) \neq 0, \ 1 \leq j \leq p\}.$$

We choose the optimal λ_n by minimizing either the AIC score or the BIC score. Our simulation studies suggest that the BIC works well when the objective of the analysis is to select important variables, even when p is larger than n. The AIC tends to choose more variables in the generating model in order to achieve a better prediction performance. In general, AIC-type criteria are better suited if the purpose of the analysis is to minimize the difference between the true distribution and the estimate from a candidate model, and the BIC-type criteria are appropriate if the purpose is to uncover the model structure, but none of these criteria can achieve both goals (Shao (1997) and Yang (2005)).

5. Simulation Studies

In this section, we use six simulated examples to evaluate the finite sample performance of the LAD-Lasso in high-dimensional settings. In each example, the data was simulated from a linear regression model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, \ldots, n.$$

In order to have a design matrix that satisfies the strong irrepresentable condition, we generated the covariates from the multivariate normal distribution and set the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ to be $0.5^{|i-j|}$.

Example 1. We generated ε_i's from $N(0, 1)$. The true regression coefficients were $\beta_{0j} = 2$ for $96 \leq j \leq 100$, and 0 otherwise. Thus the number of true nonzero coefficients was $q = 5$.

Example 2. The same as Example 1, except that $\varepsilon_i \sim \text{dbexp}(0, 1/\sqrt{2})$, a double exponential distribution with the location and scale parameters 0 and $1/\sqrt{2}$, respectively. Notice that the ε_i's in Examples 1 and 2 have the same variance.

Example 3. The same as Example 1, except that ε_i has a standard Cauchy distribution.

We first investigated the performance of the LAD-Lasso in the cases where $p < n$, $p = n$, and $p > n$ with $p = 200$ and $n = 500$, 200, and 100, respectively. Then we fixed the sample size n to be 200 and increased the number of coefficients from $p = 1,000$ to $2,000$ and then to $5,000$. In each case, we generated $1,000$ datasets. Out of $1,000$ iterations, we first computed the average number of total estimated nonzero coefficients (TN), correctly estimated nonzero number (CN), and incorrectly estimated nonzero number (IN). The ratio of correctly fitted models (CFR), the ratio of over-fitted models (OFR) and the ratio of under-fitted models (UFR) were also computed. Similar to Wang, Li and Jiang (2007), we evaluated the prediction accuracies of the LAD-Lasso using the mean and median of the *mean absolute prediction error* (MAPE). In each iteration, we generated 1000 testing data sets. The mean and median of MAPE were calculated from those testing datasets. In order to compare the efficiency of LAD-Lasso and the LS-Lasso, we applied the BIC to both methods, since AIC tends to obtain slightly lower MAPE by over-fitting the model.

The robustness property of the LAD-Lasso can be observed from the simulation results listed in Table 1−3. Table 1 shows that the LS-Lasso worked better for normal cases than it did for the double exponential cases. Both Table 1 and 2 show that the LAD-Lasso performed better than the LS-Lasso did. The advantage margins of the LAD-Lasso became more obvious when the model error was more heavy-tailed. Table 3 lists the results for the Cauchy random error, which does not have a finite moment. The LS-Lasso failed to detect the correct model in most of the replications. However, the LAD-Lasso still performed well in this case.

6. Concluding Remarks

We have investigated the theoretical properties of the LAD-Lasso for estimation and model selection in the cases when p is fixed, possibly larger than n. The computation of the LAD-Lasso can be carried out using existing programs because both the loss and penalty functions use the ℓ_1 norm. In the high-dimensional setting with $p > n$, in addition to the standard assumptions for LAD regression, conditions on model sparsity and the design matrix structure are needed for the estimation and selection consistency of the LAD-Lasso. We also assumed that the penalty parameter was fixed at a certain growth rate. An important question is whether results obtained in this paper still hold when the penalty parameter is selected using a data-driven criterion, such as the AIC or BIC. This is a challenging and unsolved problem, even for the LS-Lasso, in high-dimensional settings, and deserves further investigation.

Table 1. Results for Example 1. $\varepsilon_i \sim \mathrm{N}(0, 1^2)$.

Model	n	p	TN[1] (CN[2] , IN[3])	CFR[4] (OFR[5] , UFR[6])	Median (Mean)[7]
LAD-Lasso	200,	5,000	5.772 (5.000, 0.772)	48.7% (51.3%, 0%)	1.855 (1.877)
LS-Lasso			6.752 (5.000, 1.752)	30.0% (70.0%, 0%)	1.778 (1.779)
LAD-Lasso	200,	2,000	5.733 (5.000, 0.733)	48.4% (51.6%, 0%)	1.851 (1.848)
LS-Lasso			6.754 (5.000, 1.754)	31.7% (68.3%, 0%)	1.747 (1.749)
LAD-Lasso	200,	1,000	5.562 (5.000, 0.562)	61.4% (38.6%, 0%)	1.780 (1.788)
LS-Lasso			6.483 (5.000, 1.483)	33.1% (66.9%, 0%)	1.724 (1.721)
LAD-Lasso	500,	200	5.103 (5.000, 0.103)	91.1% (8.9%, 0%)	1.680 (1.681)
LS-Lasso			6.063 (5.000, 1.063)	59.1% (40.9%, 0%)	1.647 (1.648)
LAD-Lasso	200,	200	5.337 (5.000, 0.337)	73.2% (26.8%, 0%)	1.760 (1.768)
LS-Lasso			6.355 (5.000, 1.355)	35.2% (64.8%, 0%)	1.684 (1.691)
LAD-Lasso	100,	200	5.872 (5.000, 1.872)	44.1% (55.9%, 0%)	2.076 (2.097)
LS-Lasso			8.061 (5.000, 3.061)	14.2% (85.8%, 0%)	1.811 (1.817)

[1]The total estimated nonzero coefficients on average.

[2] The correctly estimated nonzero number on average.

[3] The incorrectly estimated nonzero number on average.

[4]The ratio of correctly fitted models.

[5] The ratio of over-fitted models.

[6]The ratio of under-fitted models.

[7]The median (mean) of the mean absolute prediction error.

Table 2. Results for Example 2. $\varepsilon_i \sim \mathrm{dbexp}(0, 1/\sqrt{2})$.

Model	n	p	TN[1] (CN[2] , IN[3])	CFR[4] (OFR[5] , UFR[6])	Median (Mean)[7]
LAD-Lasso	200,	5,000	5.822 (5.000, 0.822)	54.7% (45.3%, 0%)	1.653 (1.647)
LS-Lasso			6.882 (5.000, 1.882)	28.1% (71.9%, 0%)	1.662 (1.667)
LAD-Lasso	200,	2,000	5.718 (5.000, 0.718)	48.4% (51.6%, 0%)	1.612 (1.616)
LS-Lasso			6.641 (5.000, 1.641)	28.3% (71.3%, 0%)	1.627 (1.628)
LAD-Lasso	200,	1,000	5.678 (5.000, 0.678)	53.4% (46.6%, 0%)	1.548 (1.558)
LS-Lasso			6.513 (5.000, 1.513)	33.7% (66.3%, 0%)	1.589 (1.593)
LAD-Lasso	500,	200	5.156 (5.000, 0.156)	87.9% (12.1%, 0%)	1.464 (1.465)
LS-Lasso			6.333 (5.000, 1.333)	48.9% (51.1%, 0%)	1.489 (1.490)
LAD-Lasso	200,	200	5.157 (5.000, 0.157)	86.1% (13.9%, 0%)	1.563 (1.563)
LS-Lasso			6.090 (5.000, 1.090)	44.2% (55.8%, 0%)	1.563 (1.560)
LAD-Lasso	100,	200	5.695 (5.000, 0.695)	43.5% (56.5%, 0%)	1.883 (1.900)
LS-Lasso			8.024 (5.000, 3.024)	15.3% (84.7%, 0%)	1.739 (1.725)

[1]The total estimated nonzero coefficients on average.

[2] The correctly estimated nonzero number on average.

[3] The incorrectly estimated nonzero number on average.

[4]The ratio of correctly fitted models.

[5] The ratio of over-fitted models.

[6]The ratio of under-fitted models.

[7]The median (mean) of the mean absolute prediction error.

Table 3. Results for Example 3. $\varepsilon_i \sim \text{Cauchy}(0,1)$.

Model	n	p	TN[1]	(CN[2],	IN[3])	CFR[4]	(OFR[5],	UFR[6])	Median	(Mean)[7]
LAD-Lasso	200,	5,000	5.272	(4.911,	0.361)	67.1%	(24.3%,	4.0%)	2.997	(43.216)
LS-Lasso			172.262	(1.442,	170.820)	0 %	(1.0%,	2.0%)	10.232	(65.668)
LAD-Lasso	200,	2,000	5.108	(4.827,	0.281)	61.4%	(23.6%,	13.0%)	2.972	(24.137)
LS-Lasso			166.787	(1.464,	165.323)	0 %	(3.1%,	2.2%)	11.187	(42.856)
LAD-Lasso	200,	1,000	5.000	(4.906,	0.094)	82.4%	(9.6%,	8.0%)	2.691	(13.735)
LS-Lasso			151.333	(1.821,	149.512)	0 %	(2.0%,	3.0%)	6.894	(22.994)
LAD-Lasso	500,	200	5.015	(5.000,	0.015)	98.9%	(1.1%,	0 %)	2.202	(44.305)
LS-Lasso			55.176	(2.866,	52.310)	0 %	(9.2%,	4.5%)	54.094	(104.723)
LAD-Lasso	200,	200	5.071	(5.000,	0.071)	93.1%	(6.9%,	0 %)	2.526	(22.311)
LS-Lasso			62.866	(2.822,	60.044)	2.0%	(16.6%,	7.1%)	151.826	(233.993)
LAD-Lasso	100,	200	4.728	(4.424,	0.304)	39.1%	(3.0%,	33.9%)	3.098	(21.692)
LS-Lasso			60.882	(2.042,	58.840)	0.9%	(10.5%,	3.1%)	9.346	(33.770)

[1]The total estimated nonzero coefficients on average.

[2] The correctly estimated nonzero number on average.

[3] The incorrectly estimated nonzero number on average.

[4]The ratio of correctly fitted models.

[5] The ratio of over-fitted models.

[6]The ratio of under-fitted models.

[7]The median (mean) of the mean absolute prediction error.

Much work on penalized LS regression has been done under the assumption of Gaussian errors. The study of the LAD-Lasso provides us a robust sparse solution by relaxing the sub-Gaussian assumption. We hope that this paper help facilitate future studies of other penalized LAD methods.

7. Proofs

In this section, we provide the proofs of results in Section 2 and 3. Throughout the proofs, more properties of the LAD-Lasso are investigated. For instance, given the sparse Riesz condition, the number of nonzero elements in a LAD-Lasso estimator is bounded at a certain rate, even though p grows with n very quickly.

The proof of the estimation consistency of $\widehat{\boldsymbol{\beta}}_n$ is divided into three steps.

Step 1. We approximate $L_n(\boldsymbol{\beta})$ by

$$M_n(\boldsymbol{\beta}) \equiv f(0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \boldsymbol{\eta}'\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \lambda_n \sum_{j=1}^{p_n} |\beta_j|,$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ and $\eta_i = \text{sign}(\varepsilon_i)$. Let

$$R_{i,n}(\boldsymbol{\beta}) \equiv \left|\varepsilon_i - (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{x}^i\right| - |\varepsilon_i| + \eta_i(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{x}^i,$$

and $\xi_{in}(\boldsymbol{\beta}) \equiv R_{i,n}(\boldsymbol{\beta}) - \mathbf{E}_{\boldsymbol{\varepsilon}} R_{i,n}(\boldsymbol{\beta})$. Then

$$|L_n(\boldsymbol{\beta}) - \|\mathbf{y} - \mathbf{X}_n\boldsymbol{\beta}_0\|_1 - M_n(\boldsymbol{\beta})| = o\left(\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2\right) + \Big|\sum_{i=1}^{n} \xi_{in}(\boldsymbol{\beta})\Big|. \quad (7.1)$$

Step 2. We study the rate consistency of $\widetilde{\boldsymbol{\beta}}_n \equiv \arg\min\{M_n(\boldsymbol{\beta})\}$. In fact, $\widetilde{\boldsymbol{\beta}}_n$ is the LS-Lasso estimator of the new regression model

$$z_i = \sqrt{f(0)} \sum_{j=1}^{p_n} x_{ij}\beta_j + (2\sqrt{f(0)})^{-1}\eta_i, \quad 1 \le i \le n. \quad (7.2)$$

Step 3. We bound the ℓ_2 distance between $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n$.

Lemma 1.Concentration Theorem (Bousquet (2002)). *Let $Z_1, \ldots, Z_n$ be independent random variables with values in some space $\mathcal{Z}$. Let Υ be a class of real-valued functions on $\mathcal{Z}$, satisfying, for some positive constants ι_n and κ_n, $\|\upsilon\|_\infty \le \iota_n$, $\forall\ \upsilon \in \Upsilon$, and $n^{-1}\sum_{i=1}^n Var(\upsilon(Z_i)) \le \kappa_n^2$, $\forall\ \upsilon \in \Upsilon$. If $\mathbf{Z} = \sup_{\upsilon\in\Upsilon} \left|n^{-1}\sum_{i=1}^n \upsilon(Z_i) - \mathbf{E}(\upsilon(Z_i))\right|$, then*

$$\mathbf{P}\left(\mathbf{Z} \ge \mathbf{E}(\mathbf{Z}) + z\sqrt{2(\kappa_n^2 + \iota_n\mathbf{E}(\mathbf{Z}))} + \frac{z^2\iota_n}{3}\right) \le \exp\{-nz^2\}, \text{ for } z > 0.$$

Lemma 2. Symmetrization Theorem (van der Vaart and Wellner (1996)). *Let the Z_i's be a sequence of independent random variables with values in space $\mathcal{Z}$. Let the μ_i's be a Rademacher sequence independent of the Z_i's. Let Υ be a class of real-valued functions on $\mathcal{Z}$. Then*

$$\mathbf{E}\Big(\sup_{\upsilon\in\Upsilon}\Big|\sum_{i=1}^n \upsilon(Z_i) - \mathbf{E}(\upsilon(Z_i))\Big|\Big) \le 2\mathbf{E}\Big(\sup_{\upsilon\in\Upsilon}\Big|\sum_{i=1}^n \mu_i\upsilon(Z_i)\Big|\Big).$$

Lemma 3. Contraction Theorem (Ledoux and Talagrand (1991)). *Let $z_1, \ldots, z_n$ be non-random elements in some space $\mathcal{Z}$. Let $\upsilon_i : \mathbf{R} \to \mathbf{R}$ be any Lipschitz function and the μ_i's be a Rademacher sequence. Then for any function $f^* : \mathcal{Z} \to \mathbf{R}$,*

$$\mathbf{E}\Big(\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \mu_i[\upsilon_i(f(z_i)) - \upsilon_i(f^*(z_i))]\Big|\Big) \le 2\mathbf{E}\Big(\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \mu_i[f(z_i) - f^*(z_i)]\Big|\Big).$$

Lemma 4.(van de Geer (2008)). *Let $Z_1, \ldots, Z_n$ be independent random variables with values in $\mathcal{Z}$ and $\upsilon_1, \ldots, \upsilon_m$ be real-valued functions on $\mathcal{Z}$, such that, for $k = 1, \ldots, m$,*

$$\mathbf{E}(\upsilon_k(Z_i)) = 0,\ \forall\ i; \quad \|\upsilon_k\|_\infty \le \iota_n; \quad n^{-1}\sum_{i=1}^n \mathbf{E}(\upsilon_k^2(Z_i)) \le \kappa_n^2.$$

Then

$$\mathbf{E}\Big(\max_{1\le k\le m}\Big|n^{-1}\sum_{i=1}^{n}\upsilon_k(Z_i)\Big|\Big)\le\sqrt{\frac{2\kappa_n^2\log(2m)}{n}}+\frac{\iota_n\log(2m)}{n}.$$

These results are our main tools. We first bound the number of nonzero elements of a LAD-Lasso estimate at some rate, and this is used to calculate the rate of convergence. Let A_2 and A_3 consist of all nonzero elements in the LS-Lasso estimate $\widetilde{\boldsymbol{\beta}}_n$ of (7.2) and LAD-Lasso estimator $\widehat{\boldsymbol{\beta}}_n$ of (1.1) for the same λ_n. Thus, $A_2=A_2(\lambda_n)\equiv\{j:\ \widetilde{\beta}_{nj}=\widetilde{\beta}_{nj}(\lambda_n)\neq 0,\ 1\le j\le p_n\}$ and $A_3=A_3(\lambda_n)\equiv\{j:\ \widehat{\beta}_{nj}=\widehat{\beta}_{nj}(\lambda_n)\neq 0,\ 1\le j\le p_n\}$. If $B=B(\lambda_n)\equiv A_1\cup A_2\cup A_3=\{j:\ \beta_{0j}\neq 0 \text{ or } \widetilde{\beta}_{nj}\neq 0 \text{ or } \widehat{\beta}_{nj}\neq 0,\ 1\le j\le p_n\}$, then $\boldsymbol{\beta}_B\equiv(\beta_j,j\in B)'$, $\boldsymbol{\beta}_{B0}\equiv(\beta_{0j},j\in B)'$, $\widehat{\boldsymbol{\beta}}_B\equiv(\widehat{\beta}_{nj},j\in B)'$, and $\widetilde{\boldsymbol{\beta}}_B\equiv(\widetilde{\beta}_{nj},j\in B)'$. Similarly $\mathbf{x}_B^i\equiv(x_{ij},j\in B)'$, $\mathbf{X}_B\equiv(\mathbf{x}_j,j\in B)$.

Lemma 5. *Under conditions* (B1)−(B4)*, with probability converging to* 1*, we have*

$$|A_2|\le\Big(\frac{c_{\max}(|A_2|)}{4}\Big)\Big(\frac{n^2}{\lambda_n^2}\Big)\le(\frac{c_{\max}(\min\{n,p_n\})}{4})\Big(\frac{n^2}{\lambda_n^2}\Big),$$
$$|A_3|\le c_{\max}(|A_3|)\Big(\frac{n^2}{\lambda_n^2}\Big)\le c_{\max}(\min\{n,p_n\})\Big(\frac{n^2}{\lambda_n^2}\Big).$$

Thus $|B|\le(M_1+5c_{\max}(\min\{n,p_n\})/4)\,(n^2/\lambda_n^2)$*, where* M_1 *is the constant in* (B3) *and* $|\cdot|$ *is the cardinal value function.*

Proof. Let $\sigma_z^2\equiv\mathrm{Var}(z_i)=(4f(0))^{-1}$ at (7.2). The first inequality follows from the study of the LS-Lasso in Meinshausen and Yu (2009). By the Karush-Kuhn-Tucker condition, a necessary and sufficient condition for $\widehat{\boldsymbol{\beta}}_n$ to be a LAD-Lasso estimator is

$$\begin{cases}\mathbf{x}_j'\mathrm{sign}(\mathbf{y}-\mathbf{X}_n\widehat{\boldsymbol{\beta}}_n)=\lambda_n\mathrm{sign}(\widehat{\beta}_{nj}) & \text{if } \widehat{\beta}_{nj}\neq 0,\\ \Big|\mathbf{x}_j'\mathrm{sign}(\mathbf{y}-\mathbf{X}_n\widehat{\boldsymbol{\beta}}_n)\Big|<\lambda_n & \text{if } \widehat{\beta}_{nj}=0.\end{cases}$$

Then we have

$$\begin{aligned}|A_3|\lambda_n^2&=\sum_{j\in A_3}\Big(\mathbf{x}_j'\mathrm{sign}(\mathbf{y}-\mathbf{X}_n\widehat{\boldsymbol{\beta}}_n)\Big)^2\\&\le n^2c_{\max}(|A_3|)\le n^2c_{\max}(\min\{n,p_n\}).\end{aligned}$$

Thus we have the second inequality.

Lemma 6. *Under conditions* (B1)−(B4)*,*

$$\|\widetilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2^2\le O\left(\lambda_n^2|A_1|n^{-2}c_*^{-2}f^{-1}(0)\right)+O_P\left(d_{1n}\log p_nn^{-1}c_*^{-2}f^{-1}(0)\right),$$

where $d_{1n}=(M_1+c^*/4)(n^2/\lambda_n^2)$.

Proof. Notice that $(2\sqrt{f(0)})^{-1}\eta_i$ has a sub-Gaussian distribution with constant $(\sqrt{2f(0)})^{-1}$. Thus Lemma 6 is the direct result of the rate consistency of the LS-Lasso in Meinshausen and Yu (2009).

Lemma 7. *Let*

$$\mathbf{Z}(\delta) \equiv \sup_{\boldsymbol{\beta}_B \in \mathbf{S}_\delta^0} n^{-1} \Big| \sum_{i=1}^n R_{i,n}(\boldsymbol{\beta}_B) - \mathbf{E}_{\boldsymbol{\varepsilon}}(R_{i,n}(\boldsymbol{\beta}_B)) \Big| \text{ for } \forall \delta > 0,$$

where $\mathbf{S}_\delta^0 \equiv \{\boldsymbol{\beta}_B \in \mathbf{R}^{|B|},\ \|\boldsymbol{\beta}_B - \boldsymbol{\beta}_{B0}\|_2 \le \delta\}$. *Then under* (B1) *and* (B2),

$$\mathbf{E}_{\boldsymbol{\varepsilon}} \mathbf{Z}(\delta) \le 8\delta a_n, \tag{7.3}$$

where $a_n = \sqrt{|B|}\left(\sqrt{2n^{-1}\log(2|B|)} + b_0 n^{-1}\log(2|B|)\right)$, *and*

$$\mathbf{P}\Big(\mathbf{Z}(\delta) \ge 8\delta a_n + z\sqrt{2(\kappa_n^2 + 8\delta a_n \iota_n)} + \frac{z^2 \iota_n}{3}\Big) \le \exp\{-nz^2\}, \quad \forall\ z > 0, \tag{7.4}$$

where $\kappa_n^2 = 4c^*\delta^2$ *and* $\iota_n = 4\delta b_0\sqrt{|B|}$.

Proof. Let $f_{\boldsymbol{\beta}_B}(\mathbf{x}_B^i) = (\boldsymbol{\beta}_B - \boldsymbol{\beta}_{B0})'\mathbf{x}_B^i$, $\quad i = 1, \ldots, n$, and $\upsilon_i(t) = \upsilon(\varepsilon_i, t) = \int_0^t [I(0 < \varepsilon_i \le s)]ds$, $i = 1, \ldots, n$. Then we have $f_{\boldsymbol{\beta}_{B0}}(x_B^i) = 0$, $|\upsilon_i(t) - \upsilon_i(\tilde{t})| \le |t - \tilde{t}|$, $\forall\ t, \tilde{t} \in \mathbf{R}$, and

$$R_{i,n}(\boldsymbol{\beta}_B) = 2\int_0^{(\boldsymbol{\beta}_B - \boldsymbol{\beta}_{B0})'\mathbf{x}_B^i} [I(0 < \varepsilon_i \le s)]ds = 2\upsilon_i\left(f_{\boldsymbol{\beta}_B}(\mathbf{x}_B^i)\right).$$

If $\mu_1, \ldots, \mu_n$ be a Rademacher sequence independent of $y_1, \ldots, y_n$, then we have

$$\begin{aligned}
\mathbf{E}_{\boldsymbol{\varepsilon}} \mathbf{Z}(\delta) &= \mathbf{E}_\epsilon \sup_{\boldsymbol{\beta}_B \in \mathbf{S}_\delta^0} n^{-1} \Big| \sum_{i=1}^n [R_{i,n}(\boldsymbol{\beta}_B) - \mathbf{E}_{\boldsymbol{\varepsilon}}(R_{i,n}(\boldsymbol{\beta}_B))\ \Big| \\
&\le 4\mathbf{E}_{\boldsymbol{\varepsilon}} \sup_{\boldsymbol{\beta}_B \in \mathbf{S}_\delta^0} n^{-1} \Big| \sum_{i=1}^n \mu_i \left[\upsilon_i(f_{\boldsymbol{\beta}_B}(\mathbf{x}_B^i)) - \upsilon_i\left(f_{\boldsymbol{\beta}_{B0}}(\mathbf{x}_B^i)\right)\right]\ \Big| \\
&\le 8\mathbf{E}_{\boldsymbol{\varepsilon}} \sup_{\boldsymbol{\beta}_B \in \mathbf{S}_\delta^0} n^{-1} \Big| \sum_{i=1}^n \mu_i \left[f_{\boldsymbol{\beta}_B}(\mathbf{x}_B^i) - f_{\boldsymbol{\beta}_{B0}}(\mathbf{x}_B^i)\right]\ \Big|,
\end{aligned} \tag{7.5}$$

where the first and second inequalities are obtained from Lemmas 2 and 3. Furthermore,

$$\begin{aligned}
& n^{-1} \Big| \sum_{i=1}^n \mu_i \left[f_{\boldsymbol{\beta}_B}(\mathbf{x}_B^i) - f_{\boldsymbol{\beta}_{B0}}(\mathbf{x}_B^i)\right]\ \Big| \\
& \quad \le \max_{j \in B} n^{-1} \Big| \sum_{i=1}^n \mu_i x_{ij} \Big| \sqrt{|B|}\|\boldsymbol{\beta}_B - \boldsymbol{\beta}_{B0}\|_2,
\end{aligned} \tag{7.6}$$

and, with probability 1,

$$\max_j |\mathbf{x}_j'\mu| < b_0 < \infty; \quad n^{-1}\sum_{i=1}^n E(x_{ij}\mu_i)^2 = n^{-1}\sum_{i=1}^n x_{ij}^2 = 1.$$

Then from Lemma 4 we obtain

$$\mathbf{E}_{\boldsymbol{\varepsilon}}\Big(\max_{j\in B}\Big|n^{-1}\sum_{i=1}^n \mu_i x_{ij}\Big|\Big) \le \sqrt{2n^{-1}\log(2|B|)} + b_0 n^{-1}\log(2|B|). \tag{7.7}$$

Combining (7.5)–(7.7),

$$\mathbf{E}_{\boldsymbol{\varepsilon}}(\mathbf{Z}(\delta)) \le 8\sqrt{|B|}\delta\Big(\sqrt{2n^{-1}\log(2|B|)} + b_0 n^{-1}\log(2|B|)\Big). \tag{7.8}$$

Thus (7.3) holds. On space $\mathbf{S}_\delta^0$, we have

$$\max_{1\le i\le n} |R_{i,n}(\boldsymbol{\beta}_B) - \mathbf{E}_{\boldsymbol{\varepsilon}}(R_{i,n}(\boldsymbol{\beta}_B))| \le 4b_0\sqrt{|B|}\delta,$$

$$n^{-1}\sum_{i=1}^n \mathrm{Var}(R_{i,n}(\boldsymbol{\beta}_B) - \mathbf{E}_{\boldsymbol{\varepsilon}}[R_{i,n}(\boldsymbol{\beta}_B)] \le 4n^{-1}\sum_{i=1}^n((\boldsymbol{\beta}_B - \boldsymbol{\beta}_{B0})'x_B^i)^2 \le 4c^*\delta^2.$$

Let $\kappa_n^2 = 4c^*\delta^2$ and $\iota_n = 4\delta b_0\sqrt{|B|}$ in Lemma 1. Then for $\forall\ z > 0$,

$$\mathbf{P}\left(\mathbf{Z}(\delta) \ge \mathbf{E}\mathbf{Z}(\delta) + z\sqrt{2(\kappa_n^2 + \iota_n\mathbf{E}\mathbf{Z}(\delta))} + \frac{z^2\iota_n}{3}\right) \le \exp\{-nz^2\}. \tag{7.9}$$

Combining (7.8) and (7.9), we obtain (7.4).

Proof of Theorem 2. For any $\delta > 0$, let $\mathbf{S}_\delta^d = \{\boldsymbol{\beta}_B \in \mathbf{R}^{|B|},\ \|\boldsymbol{\beta}_B - \widetilde{\boldsymbol{\beta}}_B\|_2 = \delta\}$ and $\mathbf{S}_\delta = \{\boldsymbol{\beta}_B \in \mathbf{R}^{|B|},\ \|\boldsymbol{\beta}_B - \widetilde{\boldsymbol{\beta}}_B\|_2 \le \delta\}$. Let

$$h_n(\delta) \equiv \inf_{\boldsymbol{\beta}_B\in\mathbf{S}_\delta^d} n^{-1}M_n(\boldsymbol{\beta}_B) - n^{-1}M_n(\widetilde{\boldsymbol{\beta}}_B),$$
$$\Delta_n(\delta) \equiv \sup_{\boldsymbol{\beta}_B\in\mathbf{S}_\delta} n^{-1}\left|L_n(\boldsymbol{\beta}_B) - \|\mathbf{y} - \mathbf{X}_B\boldsymbol{\beta}_{B0}\|_1 - M_n(\boldsymbol{\beta}_B)\right|.$$

From (7.1), (B4), and $\|\boldsymbol{\beta}_B - \boldsymbol{\beta}_{B0}\|_2^2 \le 2\|\boldsymbol{\beta}_B - \widetilde{\boldsymbol{\beta}}_B\|_2^2 + 2\|\widetilde{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{B0}\|_2^2$, we have

$$\Delta_n(\delta) \le o_P(\delta^2 c^*) + o_P\Big(\sup_{\boldsymbol{\beta}_B\in\mathbf{S}_\delta}\|\widetilde{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{B0}\|_2^2\Big) + \sup_{\boldsymbol{\beta}_B\in\mathbf{S}_\delta}\Big|\sum_{i=1}^n \xi_{in}(\boldsymbol{\beta}_B)\Big|,$$

$$h_n(\delta) \ge \inf_{\boldsymbol{\beta}_B\in\mathbf{S}_\delta^d}(f(0)/n)(\boldsymbol{\beta}_B - \widetilde{\boldsymbol{\beta}}_B)'\mathbf{X}_B'\mathbf{X}_B(\boldsymbol{\beta}_B - \widetilde{\boldsymbol{\beta}}_B) \ge \delta^2 f(0)c_* > 0.$$

Let $\delta_1 = \delta/2 > 0$. From the convex minimization theorem in Hjort and Pollard (1993) we get

$$\begin{aligned}
&\mathbf{P}\left(\|\widehat{\boldsymbol{\beta}}_B - \widetilde{\boldsymbol{\beta}}_B\|_2 > \delta_1\right) \\
&\quad \le \mathbf{P}\left(\Delta_n(\delta_1) > \frac{h_n(\delta_1)}{2}\right) \\
&\quad \le \mathbf{P}\Big(\sup_{\boldsymbol{\beta}_B \in \mathbf{S}_\delta^0} \Big| n^{-1}\sum_{i=1}^n \xi_{in}(\boldsymbol{\beta}_B)\Big| > \frac{\delta_1^2 f(0) c_*}{4}\Big) + \mathbf{P}\left(\|\widetilde{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{B0}\|_2 > \delta_1\right) \\
&\quad = \mathbf{P}\Big(\mathbf{Z}(\delta) > \frac{\delta^2 f(0) c_*}{16}\Big) + \mathbf{P}\left(\|\widetilde{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{B0}\|_2 > \delta_1\right).
\end{aligned}$$

From Lemma 7,

$$\mathbf{P}\left(\mathbf{Z}(\delta) \ge 8\delta a_n + z\sqrt{2(\kappa_n^2 + \iota_n 8\delta a_n)} + \frac{z^2\iota_{2n}}{3}\right) \le e^{-nz^2},\ \forall\ z > 0,$$

where $\kappa_n^2 = 4c^*\delta^2$ and $\iota_n = 4b_0\sqrt{|B|}\delta$. As defined, $d_n = 2\left(M_1 + 5c^*/4\right)(n^2/\lambda_n^2)$. Let $\delta = c_1(d_n \log(2p_n) n^{-1} c_*^{-2} f^{-1}(0))^{-1/2}$ for some constant $c_1 > 0$ and $z = c_2 n^{-1/4}(\log(2p_n))^{1/4}$ for some constant $c_2 > 0$. Then

$$\overline{\lim_{n\to\infty}}\ \delta a_n + z(c^*\delta^2 + 8\sqrt{|B|}b_0 a_n \delta^2)^{-1/2} + \frac{z^2\sqrt{|B|}\delta b_0}{6} \le \frac{f(0)\delta^2 c_*}{128},$$

for $a_n = \sqrt{|B|}\left(\sqrt{2n^{-1}\log(2|B|)} + b_0 n^{-1}\log(2|B|)\right)$. Thus

$$\lim_{n\to\infty} \mathbf{P}\left(\|\widehat{\boldsymbol{\beta}}_B - \widetilde{\boldsymbol{\beta}}_B\|_2 > \delta_1\right) \le \lim_{n\to\infty} \mathbf{P}\left(\|\widetilde{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{B0}\|_2 > \delta_1\right).$$

From Lemma 6 and $d_{1n} < d_n$, we know that

$$\|\widetilde{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{B0}\|_2^2 \le O\left(\lambda_n^2 |A_1| n^{-2} c_*^{-2} f^{-1}(0)\right) + O_P\left(d_n \log(2p_n) n^{-1} c_*^{-2} f^{-1}(0)\right).$$

Thus we have

$$\|\widehat{\boldsymbol{\beta}}_{B_2} - \widetilde{\boldsymbol{\beta}}_{B_20}\|_2^2 \le O\left(\lambda_n^2 |A_1| n^{-2} c_*^{-2} f^{-1}(0)\right) + O_P\left(d_n \log(2p_n) n^{-1} c_*^{-2} f^{-1}(0)\right).$$

Using the triangle inequality, the consistency of the LAD-Lasso can be obtained by combining the above two inequalities with Lemma 6.

Let $\mathbf{X}_{\mathbf{2}n} \equiv (\mathbf{x}_j,\ j \notin A_1)$ and $\boldsymbol{\Sigma}_{\mathbf{21}}^n \equiv \mathbf{X}_{\mathbf{2}n}'\mathbf{X}_{\mathbf{1}n}/n$. The main results in Section 3 are established with the following lemmas.

Lemma 8. *Let $\mathbf{H}_n (= (\mathbf{h}_1, \ldots, \mathbf{h}_n) = (h_{ik})_{n\times n})$ be symmetric and idempotent. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$, where γ_i's are independent and identical distributed, assumed centered and bounded in probability. Then*

$$\mathrm{E}\max_{j\in A}\left|\mathbf{w}_j'\mathbf{H}_n\boldsymbol{\gamma}\right| \le O\left(\sqrt{n\log(2|A|)}\right) + O\left(\log(2|A|)\right),$$

where $\mathbf{w}_j = (w_{1j}, \ldots, w_{nj})'$, $j \in A$, *and* $\max_{i,j} |w_{ij}| < b$ *for some positive constant* b.

Proof. Let $\upsilon_j(\gamma_i) = \mathbf{w}_j' \mathbf{h}_i \gamma_i = \sum_{k=1}^n w_{kj} h_{ki} \gamma_i$. Without loss of generality, we take $\mathrm{P}(\gamma \le 1) = 1$. Then

$$|\upsilon_j(\gamma_i)| \le |\mathbf{w}_j' \mathbf{h}_i| = \sqrt{|\mathbf{w}_j' \mathbf{h}_i \mathbf{h}_i' \mathbf{w}_j|} \le b \text{ with probability 1,}$$

$$n^{-1} \sum_{i=1}^n \upsilon_j^2(\gamma_i) \le n^{-1} \sum_{i=1}^n \mathbf{w}_j' \mathbf{h}_i \mathbf{h}_i' \mathbf{w}_j = n^{-1} \mathbf{w}_j' \mathbf{H} \mathbf{w}_j \le b^2.$$

From Lemma 4, $n^{-1} \mathrm{E} \max_{j \in A} \left| \mathbf{w}_j' \mathbf{H}_n \boldsymbol{\gamma} \right| \le \sqrt{2b^2 n^{-1} \log(2|A|)} + b n^{-1} \log(2|A|)$.

Lemma 9. *Let* $r_{ij} \equiv x_{ij} \left[sign(\varepsilon_i - (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)' \mathbf{x}^i) - sign(\varepsilon_i) \right]$, $\xi_{ij} \equiv r_{ij} - \mathrm{E}_{\varepsilon_i}[r_{ij}]$, $h_{nj}^{(1)} \equiv \sum_{i=1}^n \xi_{ij}$, *and* $h_{nj}^{(2)} \equiv \sum_{i=1}^n \mathrm{E}_{\varepsilon_i}[r_{ij}] + 2f(0) \sum_{i=1}^n x_{ij} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)' \mathbf{x}^i$, $\mathbf{h}_{\mathbf{1}n}^{(1)} = \left(h_{nj}^{(1)},\ j \in A_1 \right)'$, $\mathbf{h}_{\mathbf{1}n}^{(2)} = \left(h_{nj}^{(2)},\ j \in A_1 \right)'$.

(i) *Under* (C1) *and* (C2),

$$\mathrm{E}_{\boldsymbol{\varepsilon}} \left(\max_{j \in A_1} |h_{nj}^{(1)}| \right) \le O\left(\sqrt{n \log(2|A_1|)} \right) + O\left(\log(2|A_1|) \right),$$

$$\mathrm{E}_{\boldsymbol{\varepsilon}} \left(\max_{j \in A_1} n^{-1} \left| \mathbf{x}_j' \mathbf{X}_{\mathbf{1}n} (\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1} \mathbf{h}_{\mathbf{1}n}^{(1)} \right| \right) \le O\left(\sqrt{n \log(2|A_1|)} \right) + O\left(\log(2|A_1|) \right).$$

(ii) *If* $\lambda_n^4 / n^3 \le O(1)$ *then under* (C1), (C2), *and* (C4),

$$|h_{nj}^{(2)}| \le O\left(\frac{\lambda_n^2 |A_1|}{n} \right) + O_P\left(n^{c_1} \log p_n \right),$$

$$|\mathbf{e}_j' (\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1} \mathbf{h}_{\mathbf{1}n}^{(2)}| \le O\left(\frac{\lambda_n^2 |A_1|^{3/2}}{n} \right) + O_P\left(n^{c_1} |A_1|^{1/2} \log p_n \right),$$

$$|\mathbf{x}_j' \mathbf{X}_{\mathbf{1}} (\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1} \mathbf{h}_{\mathbf{1}n}^{(2)}| = O\left(\frac{\lambda_n^2 |A_1|^2}{n^{1/2}} \right) + O_P\left(n^{c_1 + 1/2} |A_1| \log p_n \right).$$

(iii) *If the* γ_i*'s are independent with zero mean and finite variance, then under* (C2) *and* (C4),

$$\mathrm{E}_{\boldsymbol{\varepsilon}} \max_{j \in A_1} \left| \mathbf{e}_j' (\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1} \mathbf{X}_{\mathbf{1}n}' \boldsymbol{\gamma} \right| \le O(\sqrt{n} \log(2|A_1|)).$$

Proof. Parts (i) and (ii) can be proved using Lemma 4. We prove (ii) below. Following Phillips (1991), we define a (generalized) delta function $\delta(x) = 2d(\text{sign}(x))/dx$. Thus we have

$$h_{nj}^{(2)} = O_P\left(b_0(\widehat{\boldsymbol{\beta}}_{\mathbf{1}n} - \boldsymbol{\beta}_{\mathbf{10}})'\mathbf{X}_{\mathbf{1}}'\mathbf{X}_{\mathbf{1}}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}n} - \boldsymbol{\beta}_{\mathbf{10}})\right).$$

By replacing $d_n = O(n^{c_1})$ in the proof of Theorem 2,

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 = O\left(\frac{\lambda_n^2|A_1|}{n^2}\right) + O_P\left(\frac{n^{c_1}(\log p_n)}{n}\right).$$

Thus for $1 \le j \le p_n$, $|h_{nj}^{(2)}| \le O\left(\lambda_n^2|A_1|/n\right) + O_P\left(n^{c_1}\log p_n\right)$. The two other inequalities in (ii) can be obtained similarly.

The proof of Theorem 3is a simplified version of that of Theorem 5. Thus we only prove Theorem 5.

Proof of Theorem 5. By the Karush-Kunh-Tucker condition, $\widehat{\boldsymbol{\beta}}_n$ is a LAD-Lasso estimate if and only if

$$\begin{cases} \sum_{i=1}^n x_{ij}\text{sign}(y_i - \widehat{\boldsymbol{\beta}}_n'\mathbf{x}^i) = \lambda_n\text{sign}(\widehat{\beta}_{nj}) & \text{for } \widehat{\beta}_{nj} \neq 0, \\ \left|\sum_{i=1}^n x_{ij}\text{sign}(y_i - \widehat{\boldsymbol{\beta}}_n'\mathbf{x}^i)\right| < \lambda_n & \text{for } \widehat{\beta}_{nj} = 0. \end{cases} \tag{7.10}$$

We can rewrite (7.10) as,

$$\begin{cases} \sum_{i=1}^n x_{ij}\text{sign}(\varepsilon_i) + \sum_{i=1}^n \xi_{ij} + \sum_{i=1}^n \text{E}_{\varepsilon_i}[r_{ij}] = \lambda_n\text{sign}(\widehat{\beta}_{nj}) & \text{for } \widehat{\beta}_{nj} \neq 0, \\ \left|\sum_{i=1}^n x_{ij}\text{sign}(\varepsilon_i) + \sum_{i=1}^n \xi_{ij} + \sum_{i=1}^n \text{E}_{\varepsilon_i}[r_{ij}]\right| < \lambda_n & \text{for } \widehat{\beta}_{nj} = 0. \end{cases}$$

Let $h_{nj} \equiv h_{nj}^{(1)} + h_{nj}^{(2)}$, $\mathbf{h_{1}}_n = (h_{nj},\ j \in A_1)'$, and $\mathbf{h_{2}}_n = (h_{nj},\ j \notin A_1)'$. Then (7.10) is also equivalent to

$$\begin{cases} \sum_{i=1}^n x_{ij}\text{sign}(\varepsilon_i) - 2f(0)\sum_{i=1}^n x_{ij}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)'\mathbf{x}^i + h_{nj} = \lambda_n\text{sign}(\widehat{\boldsymbol{\beta}}_{nj}),\ \widehat{\beta}_{nj} \neq 0, \\ \left|\sum_{i=1}^n x_{ij}\text{sign}(\varepsilon_i) - 2f(0)\sum_{i=1}^n x_{ij}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)'\mathbf{x}^i + h_{nj}\right| < \lambda_n, \qquad \widehat{\beta}_{nj} = 0. \end{cases}$$

Let $\widehat{\boldsymbol{\beta}}_{\mathbf{1}n}$ and $\boldsymbol{\beta}_{\mathbf{10}}$ satisfy

$$\widehat{\boldsymbol{\beta}}_{\mathbf{1}n} = \boldsymbol{\beta}_{\mathbf{10}} + \frac{(\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1}(\mathbf{X}_{\mathbf{1}n}'\text{sign}(\boldsymbol{\varepsilon}) - \lambda_n s_{\mathbf{1}n} + \mathbf{h}_{\mathbf{1}n})}{2nf(0)}, \tag{7.11}$$

and $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}'_{\mathbf{1}n}, \mathbf{0}')'$. Then $\mathbf{X}_n\widehat{\boldsymbol{\beta}}_n = \mathbf{X}_{\mathbf{1}n}\widehat{\boldsymbol{\beta}}_{\mathbf{1}n}$ and $\mathbf{x}_j, \quad j \in A_1$ are linearly independent. From (7.11), $\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0$ if $\widehat{\boldsymbol{\beta}}_{\mathbf{1}n} = s\ \boldsymbol{\beta}_{\mathbf{1}0}$, and

$$\left|\sum_{i=1}^{n} x_{ij}\mathrm{sign}(\varepsilon_i) - 2f(0)\sum_{i=1}^{n} x_{ij}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)'\mathbf{x}^i + h_{nj}\right| < \lambda_n, \quad \text{for } \widehat{\beta}_{nj} = 0.$$

Let $\mathbf{G}_n = \mathrm{I}_n - \mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{X}'_{\mathbf{1}n}/n$ such that

$$\begin{aligned}&\mathbf{x}'_j\mathrm{sign}(\varepsilon) - 2f(0)\mathbf{x}'_j\mathbf{X}_{\mathbf{1}n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}n} - \boldsymbol{\beta}_{\mathbf{1}0})\\ &\quad = \mathbf{x}'_j\mathbf{G}_n\mathrm{sign}(\boldsymbol{\varepsilon}) + \frac{\mathbf{x}'_j\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\lambda_n\mathbf{s}_{\mathbf{1}n}}{n} - \frac{\mathbf{x}'_j\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{h}_{\mathbf{1}n}}{n}. \qquad (7.12)\end{aligned}$$

If $|\mathrm{sign}(\beta_{0j})(\beta_{0j} - \widehat{\beta}_{nj})| < |\beta_{0j}|$ for $j \in A_1$, then $\widehat{\boldsymbol{\beta}}_{\mathbf{1}n} = s\ \boldsymbol{\beta}_{\mathbf{1}0}$. Thus from (7.12) we have $\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0$ if

$$\begin{cases} \left|\mathbf{e}'_j(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{X}'_{\mathbf{1}n}\mathrm{sign}(\boldsymbol{\varepsilon}) + \lambda_n\mathbf{e}'_j(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{s}_{\mathbf{1}n} + \mathbf{e}'_j(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{h}_{\mathbf{1}n}\right|, \\ \quad < |2nf(0)\beta_{0j}|, \quad j \in A_1, \\ \left|\mathbf{x}'_j\mathbf{G}_n\mathrm{sign}(\boldsymbol{\varepsilon}) + \frac{\mathbf{x}'_j\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\lambda_n\mathbf{s}_{\mathbf{1}n}}{n} - \frac{\mathbf{x}'_j\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}h_{\mathbf{1}n}}{n} + h_{nj}\right| < \lambda_n, \quad j \notin A_1. \end{cases}$$

Then for any $0 < k_1 < k_1 + k_2 < k_1 + k_2 + k_3 < 1$,

$$\begin{aligned}\mathbf{P}\{\widehat{\boldsymbol{\beta}}_n \neq_s \boldsymbol{\beta}_0\} \le\ & \mathbf{P}\Big\{|\mathbf{e}'_j(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{X}'_{\mathbf{1}n}\mathrm{sign}(\boldsymbol{\varepsilon})| \ge \frac{2nf(0)|\beta_{0j}|}{3} \text{ for some } j \in A_1\Big\}\\ &+\mathbf{P}\Big\{\lambda_n|\mathbf{e}'_j(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{s}_{\mathbf{1}n}| \ge \frac{2nf(0)|\beta_{0j}|}{3} \text{ for some } j \in A_1\Big\}\\ &+\mathbf{P}\Big\{|\mathbf{e}'_j(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{h}_{\mathbf{1}n}| \ge \frac{2nf(0)|\beta_{0j}|}{3} \text{ for some } j \in A_1\Big\}\\ &+\mathbf{P}\{|\mathbf{x}'_j\mathbf{G}_n\mathrm{sign}(\boldsymbol{\varepsilon})| \ge (1 - k_1 - k_2 - k_3)\lambda_n \text{ for some } j \notin A_1\}\\ &+\mathbf{P}\{(\frac{\lambda_n}{n})|\mathbf{x}'_j\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{s}_{\mathbf{1}n}| \ge k_1\lambda_n \text{ for some } j \notin A_1\}\\ &+\mathbf{P}\Big\{\frac{|\mathbf{x}'_j\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{h}_{\mathbf{1}n}|}{n} \ge k_2\lambda_n \text{ for some } j \notin A_1\Big\}\\ &+\mathbf{P}\{|h_{nj}| \ge k_3\lambda_n \text{ for some } j \notin A_1\}\\ =\ & \mathbf{P}\{I_1\} + \mathbf{P}\{I_2\} + \mathbf{P}\{I_3\} + \mathbf{P}\{I_4\} + \mathbf{P}\{I_5\} + \mathbf{P}\{I_6\} + \mathbf{P}\{I_7\}.\end{aligned}$$

In fact,

$$\begin{aligned}\mathbf{P}\{I_1\} &\le \mathbf{P}\{\max_{j\in A_1}\left|\mathbf{e}'_j(\boldsymbol{\Sigma}^n_{\mathbf{11}})^{-1}\mathbf{X}'_{\mathbf{1}n}\mathrm{sign}(\boldsymbol{\varepsilon})\right| \ge \frac{2nf(0)b_{n1}}{3}\}\\ &\le \left(\frac{2nf(0)b_{n1}}{3}\right)^{-1} O(\sqrt{n}\log(2|A_1|)) = o(1),\end{aligned}$$

where the last inequality is obtained from (iii) in Lemma 9. $\mathbf{P}\{I_2\} = o(1)$ since

$$\lambda_n(2nb_{n1}f(0))^{-1}\max_{j\in A_1}\left|\mathbf{e}_j'(\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1}\mathbf{s}_{\mathbf{1}n}\right| \le \lambda_n(2nb_{n1}f(0))^{-1}c_*^{-1}|A_1|^{1/2} = o(1).$$

$\mathbf{P}\{I_3\} = o(1)$ since

$$\begin{aligned}\mathbf{P}\{I_3\} &\le \mathbf{P}\Big\{\max_{j\in A_1}\Big|\mathbf{e}_j'(\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1}\mathbf{h}_{\mathbf{1}n}^{(1)}\Big| \ge \frac{f(0)nb_{n1}}{3}\Big\}\\ &\quad+\mathbf{P}\Big\{\max_{j\in A_1}\Big|\mathbf{e}_j'(\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1}\mathbf{h}_{\mathbf{1}n}^{(2)}\Big| \ge \frac{f(0)nb_{n1}}{3}\Big\}\\ &= \mathbf{P}\{I_{31}\} + \mathbf{P}\{I_{32}\},\end{aligned}$$

$$\mathbf{P}\{I_{31}\} \le O\left(\frac{\log(|A_1|)}{n^{1/2}b_{n1}}\right) = o(1),$$

$$\mathbf{P}\{I_{32}\} \le O\left(\frac{\lambda_n^2|A_1|^{3/2}}{n^2b_{n1}}\right) + O\left(\frac{n^{c_1}|A_1|^{1/2}(\log p_n)}{nb_{n1}}\right) = o(1).$$

Since $\mathbf{G}_n$ is an idempotent matrix and $\operatorname{sign}(\varepsilon_i) = 0$, Lemma 8 gives

$$\begin{aligned}\mathbf{P}\{I_4\} &\le (1-\textstyle\sum_{i=1}^3 k_i)^{-1}\lambda_n^{-1}\mathrm{E}_{\boldsymbol{\varepsilon}}\max_{j\notin A_1}\left|\mathbf{x}_j'\mathbf{G}_n\operatorname{sign}(\boldsymbol{\varepsilon})\right|\\ &\le O\left(\frac{\sqrt{n\log(2p_n)}}{\lambda_n}\right) + O\left(\frac{\log(2p_n)}{\lambda_n}\right) = o(1).\end{aligned}$$

Suppose the strong irrepresentable condition holds for $0 < \delta_0 < 1$. We can always choose $1-\delta_0 < k_1 < 1$. Thus we have

$$\mathbf{P}\{I_5\} \le \mathbf{P}\{n^{-1}\max_{j\notin A_1}\left|\mathbf{x}_j'\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1}\mathbf{s}_{\mathbf{1}n}\right| \ge k_1\} = o(1).$$

Furthermore,

$$\begin{aligned}\mathbf{P}\{I_6\} &= \mathbf{P}\Big\{n^{-1}\max_{j\notin A_1}|\mathbf{x}_j'\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1}\mathbf{h}_{\mathbf{1}n}^{(1)}| \ge \tfrac{\lambda_n k_2}{2}\Big\}\\ &\quad+\mathbf{P}\Big\{n^{-1}\max_{j\notin A_1}|\mathbf{x}_j'\mathbf{X}_{\mathbf{1}n}(\boldsymbol{\Sigma}_{\mathbf{11}}^n)^{-1}\mathbf{h}_{\mathbf{1}n}^{(2)}| \ge \tfrac{\lambda_n k_2}{2}\Big\}\\ &= \mathbf{P}\{I_{61}\} + \mathbf{P}\{I_{62}\}.\end{aligned}$$

The following inequalities hold from (i) and (ii) in Lemma 8.

$$\mathbf{P}\{I_{61}\} \le O\left(\lambda_n^{-1}\sqrt{n\log(2p_n)}\right) + O\left(\lambda_n^{-1}\log(2p_n)\right) = o(1),$$

$$\mathbf{P}\{I_{62}\} \le O\left(\frac{\lambda_n|A_1|^2}{n^{3/2}}\right) + O\left(\frac{n^{-1}n^{c_1+1/2}|A_1|(\log p_n)}{\lambda_n}\right),$$

$$\mathbf{P}\{I_7\} \le \mathbf{P}\{\max_{j\notin A_1}|h_{nj}^{(1)}| \ge \frac{\lambda_n k_3}{2}\} + \mathbf{P}\Big\{\max_{j\notin A_1}|h_{nj}^{(2)}| \ge \frac{\lambda_n k_3}{2}\Big\} = o(1).$$

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Second International Symposium on Information Theory*, 267-281.

Bassett, G. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *J. Amer. Statist. Assoc.* **73**, 618-622.

Bousquet, O. (2002). A Bennet concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathématique* **334**, 495-550.

Chen, S. and Donoho, D. (1995). Basis pursuit. Technical report, Statistics Department, Stanford University.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of over parametrization. *Bernoulli* **10**, 971-988.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. Statistical Research Report, University of Oslo.

Huber, P. J. (1981). *Robust Statist.*. Wiley, New York.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356-1378.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33-50.

Ledoux, M. and Talagrand, M. (1991). *Probability in Branch Spaces: isoperimetry and processes.* Springer Verlag, New York.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statist. Sinica* **16**, 1273-1284.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 246-270.

Phillips, P. C. B. (1991). A shortcut to LAD estimator asymptotics. *Econometric Theory* **7**, 450-463.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7**, 186-199.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12**, 1298-1309.

Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Ann. Statist.* **13**, 1403-1417.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221-264.

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

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614-645.

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Wiley, New York.

Wang, H. S., Li, G. D. and Jiang, G. H. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Business & Economic Statistics* **25**, 347-355.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Statist.* **2**, 224-244.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared?-A conflict between model identification and regression estimation. *Biometrika* **92**, 937-950.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Research* **7**, 2541-2563.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35**, 2173-2192.

Department of Mathematics and Statistics, Oakland University. 2200 N. Squirrel Road, Rochester, Michigan 48309-4401, U.S.A.

E-mail: gao2@oakland.edu

Department of Biostatistics and Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, U.S.A.

E-mail: jian-huang@uiowa.edu

(Received September 2008; accepted May 2009)