Asymptotic analysis of high-dimensional LAD regression with Lasso

By: Xiaoli Gao and Jian Huang


© Academia Sinica, Institute of Statistical Science.

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.

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

Article:

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

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

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

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

where $\lambda_n$ is a penalty parameter. We are interested in the statistical properties of $\hat{\beta}_n$ in sparse, high-dimensional settings. We provide conditions under which $\hat{\beta}_n$ 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 \( \ell_1 \) 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 \( \ell_1 \) 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 \( \ell_1 \) 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 $\ell_2$ 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 $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})'$ 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 $\beta_0 = (\beta_{10}, \beta_{20})'$, where $\beta_{10} = (\beta_{0j}, j \in A_1)'$ consists of all important covariates, and $\beta_{20} = 0$ consists of all 0 elements. Let $x_j = (x_{1j}, \ldots, x_{nj})'$ and $x_i' = (x_{i1}, \ldots, x_{ip})'$, designed matrix $X_n = (x_1, \ldots, x_p)$. Covariates are assumed fixed. We make the following assumptions.

(A1) The $\varepsilon_i$’s are independent and identical distributed with median 0 and a continuous, positive density $f$ in a neighborhood of 0.

(A2) With $\Sigma^n = n^{-1}X'_nX_n$, there exists a positive definite matrix $\Sigma$ such that $\Sigma^n \to \Sigma$. If $\tau_{1n}$ and $\tau_{2n}$ are the minimum and maximum eigenvalues of $\Sigma^n$, 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 $\hat{\beta}_n$ be the LAD-Lasso estimator corresponding to a sequence $\lambda_n$ in (1.2). If (A1) and (A2) hold, then $\hat{\beta}_n$ has the following asymptotic properties.
consistency) If \( \lambda_n = o(n) \), then \( \hat{\beta}_n \to_p \beta_0 \).

- (Asymptotic distribution) If \( \lambda_n / \sqrt{n} \to \lambda_0 \geq 0 \), then \( \hat{\beta}_n \) has the limiting distribution
  \[ \sqrt{n}(\hat{\beta}_n - \beta_0) \to_d \arg\min(V(\mathbf{u})), \]
  where
  \[ V(\mathbf{u}) = -2\mathbf{u}'\mathbf{W} + f(0)\mathbf{u}'\Sigma\mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j \text{sign}(\beta_{0j})I(\beta_{0j} \neq 0) + |u_j|I(\beta_{0j} = 0)|. \]

Here \( \mathbf{W} \) is a random vector with a \( N(0, \Sigma/4) \) 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 \( \lambda_n \) 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

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

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

\[ \sum_{i=1}^{n} x_{ij}^2 = n \quad \text{for all} \quad j. \]
(B3) (Sparse model) There is a positive $M_1$ such that $|A_1| \leq M_1n^2/\lambda_n^2$.

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

\begin{align}
(a) & \quad 0 < c_{\min}(\min\{n, p_n\}) \leq c_{\max}(\min\{n, p_n\}) < c^* < \infty; \\
(b) & \quad c_{\min}(d_n) > c_s \text{ for } d_n \leq Mn^2/\lambda_n^2, \text{ where } M \text{ is a positive constant.}
\end{align}

(B2) assumes that the covariates are bounded. (B3) restricts the number of nonzero coefficients in the true model. Although $p_n \asymp n$, 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

$$
\|\hat{\beta}_n - \beta_0\|^2 = O \left( \lambda_n^4|A_1|n^{-2}c_s^{-2}f^{-1}(0) \right) + Op \left( d_n \log(2p_n)n^{-1}c_s^{-2}f^{-1}(0) \right), \tag{2.1}
$$

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

The theorem states that the LAD-Lasso estimator is consistent for properly selected values of $\lambda_n$, 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 $\lambda_n$: 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 $\lambda_n$ 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 $X_{1n} = (x_j, j \in A_1)$ and $s_{1n} = (s_j, j \in A_1)'$, where $s_j = \text{sign}(\beta_0j)$. Let $\Sigma_{11}^n = n^{-1}X_{1n}'X_{1n}$, the correlation matrix resulting from all the important covariates, and $\Sigma_{11}^n$ is invertible. Zhao and Yu (2006) introduced the following irrepresentable conditions.

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

$$
|n^{-1}x_j'X_{1n}(\Sigma_{11}^n)^{-1}s_{1n}| \leq 1 - \delta
$$

for $\forall j \notin A_1$, then the covariates satisfy the strong irrepresentable condition.
Definition 2. If \( |n^{-1}x_jX_1^\top(\Sigma_{11}^{-1}s_{1n})| < 1 \) for all \( j \notin A_1 \), the covariates satisfy the weak irreproducible 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 \( \lambda_n \) such that \( \lim_{n \to \infty} P(\tilde{\beta}_n(\lambda_n) = s\beta_0) = 1 \), where \( \tilde{\beta}_n(\lambda_n) = s\beta_0 \) means that \( \tilde{\beta}_n(\lambda_n) \) and \( \beta_0 \) have the same sign component-wisely.

Definition 4. If \( \lim_{n \to \infty} P(\exists \lambda > 0, \tilde{\beta}_n(\lambda) = s\beta_0) = 1 \), the LAD-Lasso estimator is general sign consistent.

We note that variable selection consistency only requires all zeros in \( \beta_0 \) 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 irreproducible condition are satisfied. We have \( P(\tilde{\beta}_n(\lambda_n) = s\beta_0) = 1 \) 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 irreproducible 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 irreproducible 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 \( \Sigma^n \) 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 $0 \leq c_1 < 1/2$ such that $|A_1| = O(n^{c_1})$;
(b) there exist positive constants $M_0$ and $c_2 > c_1$ such that $n^{(1-c_2)/2} b_{n1} \geq M_0$.

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

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

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 $p_n = O\left(\exp(n^{c_3})\right)$ 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 $P(\hat{\beta}_n(\lambda_n) = \beta_0) \to 1$ 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 $\lambda_n$, we consider an augmented data set $\{(y_i^*, x_{i1}^*, \ldots, x_{ip}^*)\}$, $1 \leq i \leq n + p$. Here $\{(y_i^*, x_{i1}^*, \ldots, x_{ip}^*)\} = \{(y_i, x_{i1}, \ldots, x_{ip})\}$, $1 \leq i \leq n$ and $\{(y_i^*, x_{i1}^*, \ldots, x_{ip}^*)\} = \{(0, \lambda_n e_{i-n})\}$, $n + 1 \leq i \leq n + p$, and $e_i$ is a $p$-dimensional unit vector with $i$th element equal to 1. We can rewrite $L_n(\beta)$ as

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

Thus, we can compute the penalized estimator using any method for solving standard $\ell_1$ minimization problem. For instance, if $p$ is not very large, the QUANTREG package can be used to find $\hat{\beta}_n(\lambda_n)$. 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 $\ell_1$ regularization parameter. If we assume that model errors are double exponentially and independently distributed, then AIC and BIC scores can be calculated using $AIC = n \log(RSA/n) + df$, and $BIC = n \log(RSA/n) + df \log(n)/2$, where $RSA = \sum_{i=1}^{n} |y_i - \hat{\beta}_0 x_i|$, 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 

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

We choose the optimal $\lambda_n$ 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 $x_i$ and $x_j$ to be $0.5^{|i-j|}$.

**Example 1.** We generated $\varepsilon_i$’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 $\varepsilon_i$’s in Examples 1 and 2 have the same variance.
Example 3. The same as Example 1, except that $\varepsilon_i$ 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 $\ell_1$ 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. $\epsilon_i \sim N(0, 1^2)$.

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

$^1$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. $\epsilon_i \sim \text{dbexp}(0, 1/\sqrt{2})$.

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

$^1$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. \( \varepsilon \sim \text{Cauchy}(0, 1) \).

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

\(^1\)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 \( \hat{\beta}_n \) is divided into three steps.

**Step 1.** We approximate \( L_n(\beta) \) by

\[
M_n(\beta) \equiv f(0)(\beta - \beta_0)'X'X(\beta - \beta_0) - \eta'X(\beta - \beta_0) + \lambda_n \sum_{j=1}^{p_n} |\beta_j|,
\]

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

\[
R_{i,n}(\beta) \equiv |\varepsilon_i - (\beta - \beta_0)'x^i| - |\varepsilon_i| + \eta_i(\beta - \beta_0)'x^i,
\]
and \( \xi_n(\beta) \equiv R_{i,n}(\beta) - E_{\beta} R_{i,n}(\beta) \). Then

\[
|L_n(\beta) - \|y - X_n \beta_0\|_1 - M_n(\beta)| = o \left( \|X(\beta - \beta_0)\|_2^2 \right) + \left| \sum_{i=1}^{n} \xi_i(\beta) \right|. \tag{7.1}
\]

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

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

**Step 3.** We bound the \( \ell_2 \) distance between \( \hat{\beta}_n \) and \( \tilde{\beta}_n \).

**Lemma 1.** Concentration Theorem \cite{Bousquet2002}. Let \( Z_1, \ldots, Z_n \) be independent random variables with values in some space \( Z \). Let \( \mathcal{F} \) be a class of real-valued functions on \( Z \), satisfying, for some positive constants \( \kappa_n \) and \( \kappa_n \), \( \|v\|_{\infty} \leq \kappa, \forall v \in \mathcal{F} \), and \( n^{-1} \sum_{i=1}^{n} \text{Var}(v(Z_i)) \leq \kappa_n^2, \forall v \in \mathcal{F} \). If \( Z = \sup_{v \in \mathcal{F}} n^{-1} \sum_{i=1}^{n} v(Z_i) - E(v(Z_i)) \), then

\[
P \left( Z \geq E(Z) + z \sqrt{2(\kappa_n^2 + \kappa_n E(Z)) + \frac{z^2 \kappa_n}{3}} \right) \leq \exp\{-n z^2\}, \quad \text{for } z > 0.
\]

**Lemma 2.** Symmetrization Theorem \cite{van der Vaart1996}. Let the \( Z_i \)'s be a sequence of independent random variables with values in space \( Z \). Let the \( \mu_i \)'s be a Rademacher sequence independent of the \( Z_i \)'s. Let \( \mathcal{F} \) be a class of real-valued functions on \( Z \). Then

\[
E \left( \sup_{v \in \mathcal{F}} \left| \sum_{i=1}^{n} v(Z_i) - E(v(Z_i)) \right| \right) \leq 2E \left( \sup_{v \in \mathcal{F}} \left| \sum_{i=1}^{n} \mu_i v(Z_i) \right| \right).
\]

**Lemma 3.** Contraction Theorem \cite{Ledoux1991}. Let \( z_1, \ldots, z_n \) be non-random elements in some space \( Z \). Let \( v_1 : R \rightarrow R \) be any Lipschitz function and the \( \mu_i \)'s be a Rademacher sequence. Then for any function \( f^* : Z \rightarrow R \),

\[
E \left( \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \mu_i [v_1(f(z_i)) - v_1(f^*(z_i))] \right| \right) \leq 2E \left( \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \mu_i [f(z_i) - f^*(z_i)] \right| \right).
\]

**Lemma 4.** \cite{van de Geer2008}. Let \( Z_1, \ldots, Z_n \) be independent random variables with values in \( Z \) and \( v_1, \ldots, v_m \) be real-valued functions on \( Z \), such that, for \( k = 1, \ldots, m \),

\[
E(v_k(Z_i)) = 0, \quad \forall i; \quad \|v_k\|_{\infty} \leq \kappa_n; \quad n^{-1} \sum_{i=1}^{n} E(v_k^2(Z_i)) \leq \kappa_n^2.
\]
Then
\[
E \left( \max_{1 \leq k \leq m} \left| n^{-1} \sum_{i=1}^{n} u_k(Z_i) \right| \right) \leq \sqrt{\frac{2n^2 \log(2m)}{n} + \sigma_n \log(2m)}.
\]

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 \( \hat{\beta}_n \) of (7.2) and LAD-Lasso estimator \( \hat{\beta}_n \) of (11) for the same \( \lambda_n \).

Thus, \( A_2 = A_2(\lambda_n) \equiv \{ j : \hat{\beta}_{nj} = \beta_{nj}(\lambda_n) \neq 0, 1 \leq j \leq p_n \} \) and \( A_3 = A_3(\lambda_n) \equiv \{ j : \hat{\beta}_{nj} = \beta_{nj}(\lambda_n) \neq 0, 1 \leq j \leq p_n \} \). If \( B = B(\lambda_n) \equiv A_1 \cup A_2 \cup A_3 = \{ j : \beta_{nj} \neq 0 \text{ or } \hat{\beta}_{nj} \neq 0, 1 \leq j \leq p_n \} \), then \( \beta_B = (\beta_j, j \in B) ', \hat{\beta}_B = (\hat{\beta}_{nj}, j \in B) ', \) and \( \beta_B = (\beta_{nj}, j \in B) ' \). Similarly \( x_B = (x_j, j \in B) ', \) \( X_B = (x_j, j \in B) \).

**Lemma 5.** Under conditions (B1)–(B4), with probability converging to 1, we have
\[
|A_2| \leq \left( \frac{c_{\text{max}}(|A_2|)}{4} \right) \left( \frac{n^2}{\lambda_n^2} \right) \leq \left( \frac{c_{\text{max}}(\min\{n, p_n\})}{4} \right) \left( \frac{n^2}{\lambda_n^2} \right),
\]
\[
|A_3| \leq c_{\text{max}}(|A_3|) \left( \frac{n^2}{\lambda_n^2} \right) \leq c_{\text{max}}(\min\{n, p_n\}) \left( \frac{n^2}{\lambda_n^2} \right).
\]

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

**Proof.** Let \( \sigma_n^2 = \text{Var}(z_i) = (4f(0))^{-1} \) 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 \( \beta_n \) to be a LAD-Lasso estimator is
\[
\begin{align*}
x_j \text{sign}(y - X_n \hat{\beta}_n) &= \lambda_n \text{sign}(\hat{\beta}_{nj}) \text{ if } \hat{\beta}_{nj} \neq 0, \\
x_j \text{sign}(y - X_n \hat{\beta}_n) &< \lambda_n \text{ if } \hat{\beta}_{nj} = 0.
\end{align*}
\]

Then we have
\[
|A_3| \lambda_n^2 = \sum_{j \in A_3} \left( x_j \text{sign}(y - X_n \hat{\beta}_n) \right)^2 
\leq n^2 c_{\text{max}}(|A_3|) \leq n^2 c_{\text{max}}(\min\{n, p_n\}).
\]

Thus we have the second inequality.

**Lemma 6.** Under conditions (B1)–(B4),
\[
\| \hat{\beta}_n - \beta_0 \|_2^2 \leq O(\lambda_n^2 |A_1| n^{-2} c^2 f^{-1}(0)) + O_P \left( d_{1n} \log p_n n^{-1} c^2 f^{-1}(0) \right),
\]
where \( d_{1n} = (M_1 + c^4/4)(n^2/\lambda_n^2) \).
\textbf{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 \cite{Meinshausen2009}.

\textbf{Lemma 7.} Let

$$Z(\delta) \equiv \sup_{\beta_B \in S_0^\delta} n^{-1} \left| \sum_{i=1}^n R_{i,n}(\beta_B) - E_\varepsilon(R_{i,n}(\beta_B)) \right| \quad \text{for } \forall \delta > 0,$$

where $S_0^\delta \equiv \{ \beta_B \in R^{|B|}, \|\beta_B - \beta_{B_0}\|_2 \leq \delta \}$. Then under (B1) and (B2),

$$E_\varepsilon Z(\delta) \leq 8\delta a_n,$$

where $a_n = \sqrt{|B|} \left( \sqrt{2n^{-1}\log(2|B|)} + 8\delta b_0 n^{-1} \log(2|B|) \right)$, and

$$P\left( Z(\delta) \geq 8\delta a_n + z \sqrt{2(\kappa_n^2 + 8\delta a_n \iota_n) + \frac{z^2\iota_n}{3}} \right) \leq \exp\{-nz^2\}, \quad \forall \ z > 0,$$

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

\textbf{Proof.} Let $f_{\beta_B}(x^i_B) = (\beta_B - \beta_{B_0})^T x^i_B$, $i = 1, \ldots, n$, and $v_i(t) = \varepsilon_i t = \int_0^t [I(0 < \varepsilon_i \leq s)] ds$, $i = 1, \ldots, n$. Then we have $f_{\beta_{B_0}}(x^i_B) = 0$, $|v_i(t) - v_i(\tilde{t})| \leq |t - \tilde{t}|$, $\forall \ t, \tilde{t} \in R$, and

$$R_{i,n}(\beta_B) = 2 \int_0^\infty (\beta_B - \beta_{B_0})^T x^i_B [I(0 < \varepsilon_i \leq s)] ds = 2v_i (f_{\beta_B}(x^i_B)).$$

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

$$E_\varepsilon Z(\delta) = E_\varepsilon \sup_{\beta_B \in S_0^\delta} n^{-1} \left| \sum_{i=1}^n R_{i,n}(\beta_B) - E_\varepsilon(R_{i,n}(\beta_B)) \right|$$

$$\leq 4E_\varepsilon \sup_{\beta_B \in S_0^\delta} n^{-1} \left| \sum_{i=1}^n \mu_i \left[ v_i(f_{\beta_B}(x^i_B)) - v_i (f_{\beta_{B_0}}(x^i_B)) \right] \right|$$

$$\leq 8E_\varepsilon \sup_{\beta_B \in S_0^\delta} n^{-1} \left| \sum_{i=1}^n \mu_i \left[ f_{\beta_B}(x^i_B) - f_{\beta_{B_0}}(x^i_B) \right] \right|, \quad (7.5)$$

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

$$n^{-1} \left| \sum_{i=1}^n \mu_i \left[ f_{\beta_B}(x^i_B) - f_{\beta_{B_0}}(x^i_B) \right] \right|$$

$$\leq \max_{j \in B} n^{-1} \left| \sum_{i=1}^n \mu_i x_{ij} \sqrt{|B|}/\|\beta_B - \beta_{B_0}\|_2, \quad (7.6)$$
and, with probability 1,
\[
\max_j |x_j^* \mu| < b_0 < \infty; \quad n^{-1} \sum_{i=1}^n E(x_{ij} \mu_i)^2 = n^{-1} \sum_{i=1}^n x_{ij}^2 = 1.
\]

Then from Lemma 4 we obtain
\[
E_x \left( \max_{j \in B} \left| n^{-1} \sum_{i=1}^n \mu_i x_{ij} \right| \right) \leq \sqrt{2n^{-1} \log(2|B|)} + b_0 n^{-1} \log(2|B|). \tag{7.7}
\]

Combining (7.5) and (7.7),
\[
E_x (\mathbf{Z}(\delta)) \leq 8\sqrt{|B|} \delta \left( \sqrt{2n^{-1} \log(2|B|)} + b_0 n^{-1} \log(2|B|) \right). \tag{7.8}
\]

Thus (7.5) holds. On space $S_0^4$, we have
\[
\max_{1 \leq i \leq n} |R_{i,n}(\beta_B) - E_x(R_{i,n}(\beta_B))| \leq 4b_0 \sqrt{|B|} \delta,
\]
\[
n^{-1} \sum_{i=1}^n \text{Var}(R_{i,n}(\beta_B) - E_x[R_{i,n}(\beta_B)]) \leq 4n^{-1} \sum_{i=1}^n ((\beta_B - \beta_{B0}) x_i^*)^2 \leq 4c^* \delta^2.
\]

Let $\kappa_n^2 = 4c^* \delta^2$ and $\tau_n = 4b_0 \sqrt{|B|}$ in Lemma 1. Then for $\forall z > 0$,
\[
P \left( \mathbf{Z}(\delta) \geq E\mathbf{Z}(\delta) + z \sqrt{2(\kappa_n^2 + n \tau_n E\mathbf{Z}(\delta))} + \frac{z^2 \tau_n}{3} \right) \leq \exp \{-nz^2 \}. \tag{7.9}
\]

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

**Proof of Theorem 2.** For any $\delta > 0$, let $S_0^4 = \{\beta_B \in \mathbb{R}^{|B|}, \|\beta_B - \bar{\beta}_B\|_2 = \delta \}$ and $S_{\delta} = \{\beta_B \in \mathbb{R}^{|B|}, \|\beta_B - \beta_B\|_2 \leq \delta \}$. Let
\[
h_n(\delta) = \inf_{\beta_B \in S_0^4} n^{-1} \sum_{i=1}^n \xi_i \beta_B - M_n(\beta_B),
\]
\[
\Delta_n(\delta) = \sup_{\beta_B \in S_{\delta}} n^{-1} \sum_{i=1}^n \left| L_n(\beta_B) - \|y - X_B \beta_B\|_1 - M_n(\beta_B) \right|.
\]

From (A1), (B4), and $\|\beta_B - \beta_{B0}\|_2^2 \leq 2\|\beta_B - \bar{\beta}_B\|_2^2 + 2\|\bar{\beta}_B - \beta_{B0}\|_2^2$, we have
\[
\Delta_n(\delta) \leq o_P(\delta^2 c^*) + o_P \left( \sup_{\beta_B \in S_{\delta}} \|\bar{\beta}_B - \beta_{B0}\|_2^2 \right) + \sup_{\beta_B \in S_{\delta}} \left| \sum_{i=1}^n \xi_i(\beta_B) \right|,
\]
\[
h_n(\delta) \geq \inf_{\beta_B \in S_0^4} (f(0)/n)(\beta_B - \bar{\beta}_B)'X_B(\beta_B - \bar{\beta}_B) \geq \delta^2 f(0)c_n > 0.
\]
Let \( \delta_1 = \delta/2 > 0 \). From the convex minimization theorem in [Hjort and Pollard 1993], we get

\[
\begin{align*}
P \left( \| \tilde{\beta}_B - \beta_B \|_2 > \delta_1 \right) \\
\leq P \left( \Delta_n(\delta_1) > \frac{h_n(\delta_1)}{2} \right) \\
\leq P \left( \sup_{\beta_B \in S_0^d} n^{-1} \sum_{i=1}^n \xi_i(\beta_B) > \frac{\delta^2 f(0) c_\epsilon}{4} \right) + P \left( \| \tilde{\beta}_B - \beta_B \|_2 > \delta_1 \right) \\
= P \left( Z(\delta) > \frac{\delta^2 f(0) c_\epsilon}{16} \right) + P \left( \| \tilde{\beta}_B - \beta_B \|_2 > \delta_1 \right).
\end{align*}
\]

From Lemma 7,

\[
P \left( Z(\delta) \geq 8 \delta a_n + z \sqrt{2(\kappa^2_n + \iota_n 8 \delta a_n)} + \frac{z^2 \delta^2 c_\epsilon}{3} \right) \leq e^{-nz^2}, \quad \forall z > 0,
\]

where \( \kappa^2_n = 4c^* \delta^2 \) and \( \iota_n = 4 b_0 \sqrt{\| \delta \|} \). As defined, \( d_n = 2(M_1 + 5c^*/4)(n^2/\lambda^2_n) \). Let \( \delta = c_1 d_n \log(2p_n) n^{-1} c_\epsilon^{-2} f^{-1}(0) \) 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

\[
\lim_{n \to \infty} \delta a_n + z(c^* \delta^2 + 8 \sqrt{B} b_0 a_n \delta^2)^{-1/2} + \frac{z^2 \sqrt{B} b_0}{6} \leq \frac{f(0) \delta^2 c_\epsilon}{128},
\]

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

\[
\lim_{n \to \infty} P \left( \| \tilde{\beta}_B - \beta_B \|_2 > \delta_1 \right) \leq \lim_{n \to \infty} P \left( \| \tilde{\beta}_B - \beta_B \|_2 > \delta_1 \right).
\]

From Lemma 6 and \( d_{1n} < d_n \), we know that

\[
\| \tilde{\beta}_B - \beta_B \|_2^2 \leq O \left( \lambda^2_n A_1 n^{-2} c_\epsilon^{-2} f^{-1}(0) \right) + O_P \left( d_n \log(2p_n) n^{-1} c_\epsilon^{-2} f^{-1}(0) \right).
\]

Thus we have

\[
\| \tilde{\beta}_{B_2} - \beta_{B_0} \|_2^2 \leq O \left( \lambda^2_n A_1 n^{-2} c_\epsilon^{-2} f^{-1}(0) \right) + O_P \left( d_n \log(2p_n) n^{-1} c_\epsilon^{-2} f^{-1}(0) \right).
\]

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

Let \( X_{2n} = (x_j, j \not\in A_1) \) and \( \Sigma_{2n} \equiv X_{2n}^\prime X_{2n}/n \). The main results in Section 3 are established with the following lemmas.

**Lemma 8.** Let \( H_n = (h_1, \ldots, h_n) = (h_{ik})_{n \times n} \) be symmetric and idempotent. Let \( \gamma = (\gamma_1, \ldots, \gamma_n)' \), where \( \gamma_i \)'s are independent and identical distributed, assumed centered and bounded in probability. Then

\[
E \max_{j \in A} |w^j H_n \gamma| \leq O \left( \sqrt{n \log(2|A|)} \right) + O \left( \log(2|A|) \right),
\]
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 \( v_j(\gamma_i) = \mathbf{w}_j' \mathbf{h}_i \gamma_i = \sum_{k=1}^n w_{kj} h_{ki} \gamma_i \). Without loss of generality, we take \( P(\gamma \leq 1) = 1 \). Then

\[
|v_j(\gamma_i)| \leq |\mathbf{w}_j' \mathbf{h}_i| = \sqrt{\mathbf{w}_j' \mathbf{h}_i \mathbf{w}_j} \leq b \text{ with probability 1,}
\]

\[
n^{-1} \sum_{i=1}^n v_j^2(\gamma_i) \leq n^{-1} \sum_{i=1}^n \mathbf{w}_j' \mathbf{h}_i \mathbf{w}_j = n^{-1} \mathbf{w}_j' \mathbf{H} \mathbf{w}_j \leq b^2.
\]

From Lemma 4, \( n^{-1} \mathbb{E} \max_{j \in A} |\mathbf{w}_j' \mathbf{H}_n \gamma| \leq \sqrt{2b^2 n^{-1} \log(2|A|)} + bn^{-1} \log(2|A|) \).

**Lemma 9.** Let \( r_{ij} \equiv x_{ij} \left[ \text{sign}(\epsilon_i - (\hat{\beta}_n - \beta_0)' \mathbf{x}_i) - \text{sign}(\epsilon_i) \right] \), \( \xi_{ij} \equiv r_{ij} - \mathbb{E}_x[r_{ij}] \), \( h_{nj}^{(1)} = \sum_{i=1}^n \xi_{ij} \), and \( h_{nj}^{(2)} = \sum_{i=1}^n \mathbb{E}_x[r_{ij}] + 2f(0) \sum_{i=1}^n x_{ij}(\hat{\beta}_n - \beta_0)' \mathbf{x}_i \), \( \mathbf{h}_1^{(1)} = (h_{nj}^{(1)}, j \in A_1)' \), \( \mathbf{h}_1^{(2)} = (h_{nj}^{(2)}, j \in A_1)' \).

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

\[
\mathbb{E}_x \left( \max_{j \in A_1} |h_{nj}^{(1)}| \right) \leq O \left( \sqrt{n \log(2|A_1|)} + O(\log(2|A_1|)) \right),
\]

\[
\mathbb{E}_x \left( \max_{j \in A_1} n^{-1} |\mathbf{x}_j' \mathbf{X}_1 \mathbf{H}_n^{-1} \mathbf{h}_1^{(1)}| \right) \leq O \left( \sqrt{n \log(2|A_1|)} + O(\log(2|A_1|)) \right).
\]

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

\[
|h_{nj}^{(2)}| \leq O \left( \frac{\lambda_n^2 |A_1|}{n} \right) + \mathbb{O}_P \left( n^{c_1} \log p_n \right),
\]

\[
|c_j' (\mathbf{X}_{11}^n)^{-1} \mathbf{h}_1^{(2)}| \leq O \left( \frac{\lambda_n^2 |A_1|^{3/2}}{n} \right) + \mathbb{O}_P \left( n^{c_1} |A_1|^{1/2} \log p_n \right),
\]

\[
|\mathbf{x}_j' \mathbf{X}_1 (\mathbf{X}_{11}^n)^{-1} \mathbf{h}_1^{(2)}| = O \left( \frac{\lambda_n^2 |A_1|^{2}}{n^{1/2}} \right) + \mathbb{O}_P \left( n^{c_1+1/2} |A_1| \log p_n \right).
\]

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

\[
\mathbb{E}_x \max_{j \in A_1} |c_j' (\mathbf{X}_{11}^n)^{-1} \mathbf{X}_1^* \gamma| \leq O(\sqrt{n \log(2|A_1|))}.
\]
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

\[
h_{nj}^{(2)} = O_P \left( b_0(\hat{\beta}_{1n} - \beta_{10})'X_1'X_1(\hat{\beta}_{1n} - \beta_{10}) \right).
\]

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

\[
\|\hat{\beta}_n - \beta_0\|^2_2 = O \left( \frac{\lambda_n^2|A_1|}{n^2} \right) + O_P \left( \frac{n^{c_1}(\log p_n)}{n} \right).
\]

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

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

Proof of Theorem 5. By the Karush-Kuhn-Tucker condition, \( \hat{\beta}_n \) is a LAD-Lasso estimate if and only if

\[
\left\{ \begin{array}{l}
\sum_{i=1}^{n} x_{ij}\text{sign}(y_i - \hat{\beta}_n'x^i) = \lambda_n\text{sign}(\hat{\beta}_{nj}) \quad \text{for } \hat{\beta}_{nj} \neq 0, \\
\sum_{i=1}^{n} x_{ij}\text{sign}(y_i - \hat{\beta}'_n x^i) < \lambda_n \quad \text{for } \hat{\beta}_{nj} = 0.
\end{array} \right. \tag{7.10}
\]

We can rewrite (7.10) as,

\[
\left\{ \begin{array}{l}
\sum_{i=1}^{n} x_{ij}\text{sign}(\varepsilon_i) + \sum_{i=1}^{n} \xi_{ij} + \sum_{i=1}^{n} E_{\varepsilon_i}[r_{ij}] = \lambda_n\text{sign}(\hat{\beta}_{nj}) \quad \text{for } \hat{\beta}_{nj} \neq 0, \\
\sum_{i=1}^{n} x_{ij}\text{sign}(\varepsilon_i) + \sum_{i=1}^{n} \xi_{ij} + \sum_{i=1}^{n} E_{\varepsilon_i}[r_{ij}] < \lambda_n \quad \text{for } \hat{\beta}_{nj} = 0.
\end{array} \right.
\]

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

\[
\left\{ \begin{array}{l}
\sum_{i=1}^{n} x_{ij}\text{sign}(\varepsilon_i) - 2f(0) \sum_{i=1}^{n} x_{ij}(\hat{\beta}_n - \beta_0)'x^i + h_{nj} = \lambda_n\text{sign}(\hat{\beta}_{nj}), \ \hat{\beta}_{nj} \neq 0, \\
\sum_{i=1}^{n} x_{ij}\text{sign}(\varepsilon_i) - 2f(0) \sum_{i=1}^{n} x_{ij}(\hat{\beta}_n - \beta_0)'x^i + h_{nj} < \lambda_n, \ \hat{\beta}_{nj} = 0.
\end{array} \right.
\]

Let \( \hat{\beta}_{1n} \) and \( \beta_{10} \) satisfy

\[
\hat{\beta}_{1n} = \beta_{10} + \frac{\left( \Sigma_{11}^n \right)^{-1}(X_{1n}'\text{sign}(\varepsilon) - \lambda_n s_{1n} + h_{1n})}{2nf(0)}, \tag{7.11}
\]
and \( \hat{\beta}_n = (\hat{\beta}_{1n}, 0')' \). Then \( X_n \hat{\beta}_n = X_{1n} \hat{\beta}_{1n} \) and \( x_j, \ j \in A_1 \) are linearly independent. From (7.11), \( \hat{\beta}_n = \beta_0 \) if \( \hat{\beta}_{1n} = \beta_{10} \), and

\[
\left| \sum_{i=1}^{n} x_{ij} \text{sign}(\varepsilon_i) - 2f(0) \sum_{i=1}^{n} x_{ij} (\hat{\beta}_n - \beta_0)' x_i + h_{nj} \right| < \lambda_n, \quad \text{for } \hat{\beta}_{nj} = 0.
\]

Let \( G_n = I_n - X_{1n} (\Sigma_{11}^n)^{-1} X_{1n}'/n \) such that

\[
x_j' \text{sign}(\varepsilon) - 2f(0)x_j X_{1n}(\hat{\beta}_{1n} - \beta_{10})
= x_j' G_n \text{sign}(\varepsilon) + \frac{x_j' X_{1n}(\Sigma_{11}^n)^{-1} \lambda_n s_{1n}}{n} - \frac{x_j' X_{1n}(\Sigma_{11}^n)^{-1} h_{1n}}{n}.
\]

(7.12)

If \( |\text{sign}(\beta_{0j})(\beta_{0j} - \hat{\beta}_{nj})| < |\beta_{0j}| \) for \( j \in A_1 \), then \( \hat{\beta}_{1n} = \beta_{10} \). Thus from (7.12) we have \( \hat{\beta}_n = s \beta_0 \) if

\[
\left\{ \begin{array}{l}
|e_j'(\Sigma_{11}^n)^{-1} X_{1n}' \text{sign}(\varepsilon) + \lambda_n e_j'(\Sigma_{11}^n)^{-1} s_{1n} + e_j'(\Sigma_{11}^n)^{-1} h_{1n}|, \\
< 2nf(0)|\beta_{0j}|, \quad j \in A_1, \\
x_j' G_n \text{sign}(\varepsilon) + \frac{x_j' X_{1n}(\Sigma_{11}^n)^{-1} \lambda_n s_{1n}}{n} - \frac{x_j' X_{1n}(\Sigma_{11}^n)^{-1} h_{1n}}{n} + h_{nj} < \lambda_n, \quad j \notin A_1.
\end{array} \right.
\]

Then for any \( 0 < k_1 < k_1 + k_2 < k_1 + k_2 + k_3 < 1 \),

\[
P(\hat{\beta}_n \neq s \beta_0) \leq P\left\{ |e_j'(\Sigma_{11}^n)^{-1} X_{1n}' \text{sign}(\varepsilon)| \geq \frac{2nf(0)|\beta_{0j}|}{3} \right\}
+ P\left\{ |\lambda_n e_j'(\Sigma_{11}^n)^{-1} s_{1n}| \geq \frac{2nf(0)|\beta_{0j}|}{3} \right\}
+ P\left\{ |e_j'(\Sigma_{11}^n)^{-1} h_{1n}| \geq \frac{2nf(0)|\beta_{0j}|}{3} \right\}
+ P\left\{ |x_j' G_n \text{sign}(\varepsilon)| \geq (1 - k_1 - k_2 - k_3)\lambda_n \text{ for some } j \notin A_1 \right\}
+ P\left\{ |\lambda_n x_j' X_{1n}(\Sigma_{11}^n)^{-1} s_{1n}| \geq k_1 \lambda_n \text{ for some } j \notin A_1 \right\}
+ P\left\{ |\lambda_n x_j' X_{1n}(\Sigma_{11}^n)^{-1} h_{1n}| \geq k_2 \lambda_n \text{ for some } j \notin A_1 \right\}
+ P\left\{ |h_{nj}| \geq k_3 \lambda_n \text{ for some } j \notin A_1 \right\}
= P\{I_1\} + P\{I_2\} + P\{I_3\} + P\{I_4\} + P\{I_5\} + P\{I_6\} + P\{I_7\}.
\]

In fact,

\[
P\{I_1\} \leq P\left\{ \max_{j \in A_1} |e_j'(\Sigma_{11}^n)^{-1} X_{1n}' \text{sign}(\varepsilon)| \geq \frac{2nf(0)b_{n1}}{3} \right\}
\leq \left( \frac{2nf(0)b_{n1}}{3} \right)^{-1} O(\sqrt{n} \log(2|A_1|)) = o(1),
\]
where the last inequality is obtained from (iii) in Lemma 9. \( P\{I_2\} = o(1) \) since
\[
\lambda_n (2nb_1 f(0))^{-1} \max_{j \in A_1} |e_j'(\Sigma_{11}^n)^{-1}s_{1n}| \leq \lambda_n (2nb_1 f(0))^{-1}c^{-1} |A_1|^{1/2} = o(1).
\]

\( P\{I_3\} = o(1) \) since
\[
P\{I_3\} \leq P\left\{ \max_{j \in A_1} |e_j'(\Sigma_{11}^n)^{-1}h_{1n}^{(1)}| \geq \frac{f(0)nb_1}{3} \right\} + P\left\{ \max_{j \in A_1} |e_j'(\Sigma_{11}^n)^{-1}h_{1n}^{(2)}| \geq \frac{f(0)nb_1}{3} \right\} = P\{I_{31}\} + P\{I_{32}\},
\]
\[
P\{I_{31}\} \leq O\left( \frac{\log(|A_1|)}{n^{1/2}b_1} \right) = o(1),
\]
\[
P\{I_{32}\} \leq O\left( \frac{\lambda_n^2 |A_1|^{3/2}}{n^2 b_1} \right) + O\left( \frac{n^c |A_1|^{1/2} (\log p_n)}{nb_1} \right) = o(1).
\]
Since \( G_n \) is an idempotent matrix and \( \text{sign}(\varepsilon_i) = 0 \), Lemma 8 gives
\[
P\{I_4\} \leq (1 - \sum_{i=1}^{3} k_i)^{-1} \lambda_n^{-1} E_{\varepsilon} \max_{j \notin A_1} |x_j G_n \text{sign}(\varepsilon) | \leq O\left( \frac{n \log(2p_n)}{\lambda_n} \right) + O\left( \frac{\log(2p_n)}{\lambda_n} \right) = o(1).
\]

Suppose the strong irrepresentable condition holds for \( 0 < \delta_0 < 1 \). We can always choose \( 1 - \delta_0 < k_1 < 1 \). Thus we have
\[
P\{I_3\} \leq P\left\{ n^{-1} \max_{j \notin A_1} |x_j' X_{1n}(\Sigma_{11}^n)^{-1}s_{1n}| \geq k_1 \right\} = o(1).
\]
Furthermore,
\[
P\{I_6\} = P\left\{ n^{-1} \max_{j \notin A_1} |x_j' X_{1n}(\Sigma_{11}^n)^{-1}h_{1n}^{(1)}| \geq \frac{\lambda_n k_3}{2} \right\} + P\left\{ n^{-1} \max_{j \notin A_1} |x_j' X_{1n}(\Sigma_{11}^n)^{-1}h_{1n}^{(2)}| \geq \frac{\lambda_n k_3}{2} \right\} = P\{I_{61}\} + P\{I_{62}\}.
\]
The following inequalities hold from (i) and (ii) in Lemma 8.
\[
P\{I_{61}\} \leq O\left( \frac{\lambda_n^{-1} \sqrt{n \log(2p_n)}}{\lambda_n} \right) + O\left( \frac{\lambda_n^{-1} \log(2p_n)}{\lambda_n} \right) = o(1),
\]
\[
P\{I_{62}\} \leq O\left( \frac{\lambda_n |A_1|^2}{n^{3/2}} \right) + O\left( \frac{n^{-1}n^{c_1+1/2} |A_1| (\log p_n)}{\lambda_n} \right),
\]
\[
P\{I_7\} \leq P\left\{ \max_{j \notin A_1} |h_{nj}^{(1)}| \geq \frac{\lambda_n k_3}{2} \right\} + P\left\{ \max_{j \notin A_1} |h_{nj}^{(2)}| \geq \frac{\lambda_n k_3}{2} \right\} = o(1).
\]
References


