symmetry

Article

Robust Variable Selection Based on Relaxed Lad Lasso

Hongyu Li [1], Xieting Xu [2], Yajun Lu [3,*], Xi Yu [4,*], Tong Zhao [2] and Rufei Zhang [5,6,7]

[1] The Graduate School, Woosuk University, Wanju-gun 55338, Korea
[2] College of Economics, Hebei GEO University, Shijiazhuang 050031, China
[3] School of Business Administration, Chongqing Technology and Business University, Chongqing 400067, China
[4] HBIS Supply Chain Management Co., Ltd., Shijiazhuang 050001, China
[5] Hebei Center for Ecological and Environmental Geology Research, Hebei GEO University, Shijiazhuang 050031, China
[6] Reaserch Center of Nutural Resources Assets, Hebei GEO University, Shijiazhuang 050031, China
[7] Hebei Province Mineral Resources Development and Management and the Transformation and Upgrading of Resources Industry Soft Science Resrarch Base, Shijiazhuang 050031, China
* Correspondence: luyajun@email.ctbu.edu.cn (Y.L.); yuxi@hggyl.wecom.work (X.Y.)

Abstract: Least absolute deviation is proposed as a robust estimator to solve the problem when the error has an asymmetric heavy-tailed distribution or outliers. In order to be insensitive to the above situation and select the truly important variables from a large number of predictors in the linear regression, this paper introduces a two-stage variable selection method named relaxed lad lasso, which enables the model to obtain robust sparse solutions in the presence of outliers or heavy-tailed errors by combining least absolute deviation with relaxed lasso. Compared with lasso, this method is not only immune to the rapid growth of noise variables but also maintains a better convergence rate, which is $O_p\left(n^{-1/2}\right)$. In addition, we prove that the relaxed lad lasso estimator has the property of consistency at large samples; that is, the model selects the number of important variables with a high probability of convergence to one. Through the simulation and empirical results, we further verify the outstanding performance of relaxed lad lasso in terms of prediction accuracy and the correct selection of informative variables under the heavy-tailed distribution.

Keywords: variable selection; relaxed lasso; least absolute deviation; consistency; heavy-tailed

Citation: Li, H.; Xu, X.; Lu, Y.; Yu, X.; Zhao, T.; Zhang, R. Robust Variable Selection Based on Relaxed Lad Lasso. *Symmetry* **2022**, *14*, 2161. https://doi.org/10.3390/sym14102161

Academic Editor: Luis Manuel Sánchez Ruiz

Received: 11 September 2022
Accepted: 11 October 2022
Published: 15 October 2022

1. Introduction

With the expansion of datasets, selecting factors that truly affect the response variable from enormous predictors has been a topic of interest for statisticians for many years. However, the response variable commonly contains heavy-tailed errors or outliers in practice. In such a situation, traditional variable selection techniques may fail to produce robust sparse solutions. In this paper, a new estimator for the heavy-tailed distribution data is suggested as a way to deal with this problem.

In the past two decades, Tibshirani [1] first combined ordinary least square (OLS) with L_1 penalty and proposed a new variable selection method named least absolute shrinkage and selection operator (lasso). Lasso is a convex regularization method by adding L_1 norm, which avoids the influence of the sign of OLS on the prediction results. The method can also perform simultaneously model selection and shrinkage estimation in high-dimensional data. However, lasso is sensitive in the case of heavy tails on the model distribution, which arises from the problem of heterogeneity due to the data coming from different sets [2]. So, any small changes in the data can cause the solution path of lasso to contain many irrelevant noise variables. The above instability can also occur when a single relevant covariate is randomly selected. It means that applying lasso to the same data may generate widely different results [3]. In addition, the convergence speed of lasso can be affected by the rapid growth of noise variables, and the convergence speed itself is slow. Relaxed lasso was proposed to overcome the influence of noise variables and perform variable selection at a faster and more stable speed. Meinshausen [4] defined the relaxed lasso estimator for $\lambda \in [0, \infty]$ and $\phi \in (0, 1]$ as

$$\hat{\beta}^{Rlasso} = \arg\min_{\beta} \left\| \boldsymbol{Y} - \sum_{j=1}^{p} X_j^T \{\beta_j \cdot \mathbf{1}_M\} \right\|_2^2 + \phi\lambda \sum_{j=1}^{p} |\beta_j|, \tag{1}$$

where $M \subseteq \{1, \ldots, p\}$, p is the number of variables with nonzero coefficients selected into the model, and $\mathbf{1}_M$ is an indicator function, that is $\mathbf{1}_M = \begin{cases} 0, & t \in M \\ 1, & t \notin M \end{cases}$, for all $t \in \{1, \ldots, p\}$. Hastie et al. [5] extended the work of Bertsimas et al. by comparing the lasso, forward stepwise and relaxed lasso methods with different signal-to-noise ratios (SNRs) scenarios. The results show that relaxed lasso has an overall outstanding performance at any SNR level. This superiority is reflected in the relaxation parameter ϕ. By appropriately modifying the parameter ϕ, relaxed lasso ensures that the resulting model is consistent with the true model, neither favoring excessive compression that would result in the exclusion of essential variables nor selecting redundant noise variables. This serves as the main reason why we add the relaxation parameter ϕ to the lad lasso. Compared to lasso, relaxed lasso greatly reduces the number of false positives while also achieving a trade-off between low computational complexity and fast convergence rates [6]. From the perspective of the closed-form solution, Mentch and Zhou [7] indicate that the relaxed lasso estimator can be expressed as a weighted average of lasso and least squares. When the weight of the lasso is increased, it provides a greater amount of regularization, hence reducing the degree of freedom of the variables in the final model to achieve sparse solutions. Bloise et al. [8] demonstrated that relaxed lasso has higher predictive power because it is able to avoid overfitting by tuning two separate parameters. He [9] concluded that relaxed lasso improves prediction accuracy since it avoids selecting unimportant variables and excessively removing informative variables. Extensive research has demonstrated that relaxed lasso has advantages in terms of variable selection, prediction accuracy, convergence speed, and computational complexity. However, relaxed lasso, like OLS cannot produce reliable solutions when the response variable contains heavy-tailed errors or outliers.

In order to solve the problem of poor fitting results of relaxed lasso to the heavy-tailed distribution or outliers, least absolute deviation (LAD) based on robust regression is introduced. It estimates coefficients by minimizing the sum of the absolute values of the prediction errors. The traditional squared loss in the objective function used by classic regularization methods is unsuitable for heavy-tailed distributions and outliers, but LAD performs admirably in these situations. Gao [10] showed that the LAD loss could provide a powerful alternative to the squared loss. In recent years, some researchers have combined robust regression with popular penalty regularization methods. The most typical method is the lad lasso of Wang et al. [11], which combines lad and adaptive lasso so that the model can perform robust variable selection. Then, the theoretical properties of lad lasso under large samples have been systematically studied by Gao and Huang [12] and Xu and Ying [13]. Arslan [14] proposed a weighted lad lasso to mitigate the effect of outliers on explanatory and response variables. In addition, lad lasso also has a wide range of practical applications. For example, Rahardiantoro and Kurnia [15] showed that lad lasso has a more minor standard error than lasso in the presence of outliers in high-dimensional data via simulation. Zhou and Liu [16] also applied the lad lasso to the double-truncated data and showed that it is more accurate to select the real model than the best subset selection procedure. Li and Wang [17] applied lad lasso to the change point problem in fields such as statistics and econometrics. Thanks to the superior performance of lad lasso, we consider proposing a new estimator that can not only perform variable selection but is also insensitive to the heavy-tailed distribution or outliers in the response variable.

In this article, we combine lad lasso and the relaxation parameter of relaxed lasso to propose relaxed lad lasso and study its asymptotic properties in the case of large samples.

It integrates the advantages of relaxed lasso and lad lasso methods into the following three points. Firstly, the relaxed lad lasso estimator has the same consistency property as the lad lasso, i.e., the method selects important variables with a high probability of convergence to one. Secondly, since relaxed lasso has a closed-form solution, solving relaxed lad lasso is eventually equivalent to solving the LAD program, so we can employ a simple and efficient algorithm. Thirdly, relaxed lad lasso possesses the robustness of lad lasso to heavy-tailed errors or outliers in the response variable. In theory, we prove the $\sqrt{n}$-consistency of relaxed lad lasso under some mild assumptions and illustrate its advantages in convergence speed. Although the convergence speed of the relaxed lad lasso $O_p\left(n^{-1/2}\right)$ is slower than that of relaxed lasso $O_p\left(n^{-1}\right)$, our method handles outliers and heavy-tailed errors well because it is not affected by the rapid growth of noise variables. The simulation shows that relaxed lad lasso has the highest prediction accuracy and probability of the correct selection of important variables under heavy-tailed distributions compared to other methods. We also apply relaxed lad lasso to financial data and obtain the same results as the simulation regarding prediction accuracy.

However, our method has room for improvement, as LAD cannot handle the presence of outliers in the explanatory variables and is sensitive to leverage points [18]. Hence, our method suffers from the same problem. Under the framework of LAD regression, researchers have proposed many new methods to improve robustness by reducing the weight of leverage points. Giloni et al. [19] proposed a weighted least absolute deviation process (WLAD) to overcome the shortcomings of the LAD method. However, as the proportion of outliers increases, the robustness of the WLAD estimator significantly decreases [20]. To obtain a high robustness estimator and abnormal information of observations, Gao and Feng [21] proposed a penalized weighted least absolute deviation (PWLAD) regression method. Jiang et al. [22] combined the PWLAD estimator and the lasso method to detect outliers and select variables robustly. However, it is worth noting that these methods mainly address the robustness problem when there are leverage points or outliers in the explanatory variables. Still, our method is suitable in situations with heavy-tailed errors or outliers in the response variable. In the simulation, we assume that the model error follows a heavy-tailed distribution such as the t-distribution. Therefore, we do not compare relaxed lad lasso with the above methods due to the different application scenarios. More specific details can be found in Section 4.

The remainder of the paper is organized as follows: Section 2 defines the estimator of relaxed lad lasso and interprets the parameters in the model. In addition, we give the detailed procedure of the algorithm. Section 3 describes the asymptotic properties of the loss function and provides the theorems' assumptions. Section 4 compares the performance of relaxed lad lasso with conventional lasso methods (such as classical lasso, adaptive lasso, and relaxed lasso) through simulations under different heavy-tailed distribution scenarios. Section 5 analyzes empirical data to confirm the robustness of the proposed method to heavy-tailed distributions. Section 6 summarizes the advantages of the new method as well as suggestions for further research. The proofs of the theorems are given in Appendixes A–E.

2. Relaxed Lad Lasso

2.1. Definition

This article considers the linear model

$$\boldsymbol{Y} = \boldsymbol{X}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2}$$

The random error term $\boldsymbol{\varepsilon}$ does not require it to obey a certain normal distribution like the traditional regression model. In this model, the condition of random error on the distribution is relaxed, and only the median is 0. $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)$ is an $n \times p$ dimensional matrix from a normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}$, where X_i is the predictor matrix of the ith variable and $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables. $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_j\right)^T$

is the regression coefficient of the model. In addition, the regression coefficient is nonzero when $j \le p$.

Next, we define relaxed lad lasso, which combines the L_1 penalty term with the relaxation parameter ϕ, so that the new model can still maintain excellent convergence speed and variable selection ability when there are outliers in the heavy-tailed distribution and the response variable.

Definition 1. *The solution to relaxed lad lasso is*

$$\hat{\beta}^{Reladlasso} = \arg\min_{\beta} \left| Y - \sum_{j=1}^{p} X_j^T \{\beta_j \cdot \mathbf{1}_{S^\lambda}\} \right| + n\lambda\phi \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

where $\mathbf{1}_{S^\lambda}$ is an indicator function, $S^\lambda = \left\{1 \le t \le p \mid \hat{\beta}_t^\lambda \ne 0\right\}$ is the set of nonzero coefficients; the penalty parameter $\lambda \in [0, \infty)$ and $\phi \in [0, 1]$.

When a regression coefficient belongs to the set S^λ, it is selected into the true model. If the parameters within the range of the S^λ set take different values, the model will have different functions. The value of the penalty parameter λ indicates the degree of compression applied to the coefficients so that it controls the number of predictors entering the model. When either λ or ϕ takes 0, minimizing the objective function for relaxed lad lasso is equivalent to solving the LAD method, and the original intention of variable selection is lost, so the parameters always start from a value far from 0. In this paper, the optimal parameters λ, ϕ are chosen through cross-validation, and the relaxed lad lasso estimator will be consistent if the important variables are correctly selected.

We define the loss function of relaxed lad lasso as

$$L(\lambda, \phi) = E\left| \boldsymbol{Y} - \boldsymbol{X}^T \hat{\boldsymbol{\beta}} \right| - \sigma^2. \tag{4}$$

2.2. Algorithm

In the following, we provide a detailed algorithm for solving relaxed lad lasso as defined in (3). It is well known that the closed-form solution of relaxed lasso is a linear combination of the lasso and least squares estimator. The same form of the solution can be extended to relaxed lad lasso. It turns out that the relaxed lad lasso estimator $\hat{\beta}^{Reladlasso}$ is the combination of the lad lasso estimator $\hat{\beta}^{Ladlasso}$ and the LAD estimator $\hat{\beta}^{Lad}$, so we can solve them separately.

Computationally, the relaxed lad lasso estimator can be written as

$$\hat{\beta}^{Reladlasso} = \phi\hat{\beta}^{Ladlasso} + (1-\phi)\hat{\beta}^{Lad}, \tag{5}$$

where the parameter $\phi \in [0, 1]$. Firstly, we are interested in estimating $\hat{\beta}^{Ladlasso}$ by minimizing the convex problem

$$\hat{\beta}^{Ladlasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - X_i\beta| + n\lambda \sum_{j=1}^{p} |\beta_j|. \tag{6}$$

A new dataset $\left\{\left(Y_i', X_i'\right)\right\}$ with $i = 1, \ldots, n+p$ can be considered to transform the lad lasso solution in (6) to the conventional LAD citerion. We set $\left(Y_i', X_i'\right) = (Y, X_i)$ for $1 < i < n$ and $\left(Y_{n+k}', X_{n+k}'\right)$=$(0, n\lambda d_k)$ for $1 \le k \le p$, where $d_k = (0, \ldots, 0, 1_{k\text{th}}, 0, \ldots, 0)$ such that the kth component is equal to 1 and the remaining components are equal to 0. It should be noted that the lad lasso estimator can be expressed as follows:

$$\hat{\beta}^{Ladlasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n+p} \left| Y_i' - X_i'\beta \right|. \tag{7}$$

Therefore, the computational effort of solving all lad lasso solutions in (7) is identical to that of computing any unpenalized LAD program. Then, we consider the following lad lasso solution

$$\hat{\beta}^{Lad} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - X_i\beta|. \tag{8}$$

If $\hat{\beta}_j^{lad} \neq 0, 1 \leq j \leq p$, then the subgradient of (8) is given by

$$\frac{d\|\boldsymbol{Y} - \boldsymbol{X\beta}\|_1}{d\beta_j} = -X^T \operatorname{sgn}\left(Y - X\hat{\beta}_j^{lad}\right). \tag{9}$$

So, the solution of the LAD is given by iterating

$$\hat{\beta}_{k+1}^{Lad} = \hat{\beta}_k^{Lad} - \alpha\left[-X^T \operatorname{sgn}\left(Y - X\hat{\beta}_k^{Lad}\right)\right], \tag{10}$$

where k is the number of iterations and $\alpha > 0$ is a suitable step size.

The unpunished LAD program in (8) can be solved using the rq function in the quantreg package of R. The overview of the algorithm is described in Algorithm 1.

Algorithm 1 The algorithm for relaxed lad lasso

```
Input: Design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, response vector $\boldsymbol{Y} \in \mathbb{R}^n$, parameter $\phi \in [0,1]$, iteration
    number $k$, stepsize $\alpha$
Output: The relaxed lad lasso estimator $\hat{\beta}^{Reladlasso}$
Initialization: Define $\hat{\beta}^{Reladlasso} = \phi\hat{\beta}^{Ladlasso} + (1-\phi)\hat{\beta}^{Lad}$
Compute
    Set $\left\{\left(Y_i^{'}, X_i^{'}\right)\right\}$ with $i = 1, \ldots, n+p$ to be the new dataset of lad lasso
    Set $\left(Y_i^{'}, X_i^{'}\right) = (Y_i, X_i)$ for $1 \leq i \leq n$ and $\left(Y_{n+k}^{'}, X_{n+k}^{'}\right)$=$(0, n\lambda d_k)$ for $1 \leq k \leq p$, where
        $d_k = (0, \ldots, 0, 1_{\text{kth}}, 0, \ldots, 0)$
    The objective functions of Ladlasso and LAD are $Q^{'}(\beta) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n+p} \left|Y_i^{'} - X_i^{'}\beta\right|$ and
        $Q(\beta) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - X_i\beta|$
    Set $k = 0$
Repeat
    Update $\hat{\beta}_{k+1}^{Lad} \leftarrow \hat{\beta}_k^{Lad} - \alpha\left[-X^T \operatorname{sgn}\left(Y - X\hat{\beta}_k^{Lad}\right)\right]$
    Update $\hat{\beta}_{k+1}^{Ladlasso} \leftarrow \hat{\beta}_k^{Ladlasso} - \alpha\left[-X^T \operatorname{sgn}\left(Y - X\hat{\beta}_k^{Ladlasso}\right)\right]$
    Update $k = k + 1$
Until convergence
```

3. The Asymptotic Properties of Relaxed Lad Lasso

Before obtaining the asymptotic properties, we must first set certain conditions. Regarding the covariance matrix $\boldsymbol{\Sigma}$, we consider the settings in Fu and Knight [23] and Meinshausen [4] and put forward the first hypothesis:

Assumption 1. *For all $n \in \mathbb{N}$, the covariance matrix $cov(\boldsymbol{X}) = \boldsymbol{\Sigma}$ is diagonally dominant. According to the setting of Fu and Knight [23]:*

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \to \Sigma, \text{ as } n \to \infty, \tag{11}$$

and then, it can be deduced to obtain

$$\frac{1}{n} \underset{1 \le i \le n}{max} X_i^T X_i \to 0, \text{ as } n \to \infty. \tag{12}$$

Obviously, the default precondition for diagonal dominance of the covariance matrix is that the covariance matrix exists. When the strong condition of diagonal dominance is satisfied, the covariance matrix is positive definite, and the hidden condition is that its inverse matrix still exists.

Assumption 2. *There exist constants $c > 0$ and $s \in (0, 1)$ such that the number of predictors p grows exponentially with the number of observed variables n. It can be written as*

$$p_n \sim se^{cn}. \tag{13}$$

Assumption 2 sets the growth mode of p to satisfy the requirement that relaxed lad lasso still retains a better convergence speed in variable selection.

Assumption 3. *Define the range $\mathcal{L}$ of the penalty parameter λ. For a constant $c > 0$, we have*

$$\mathcal{L} = \{\lambda \ge 0 : ce^{p_n} \le n\}. \tag{14}$$

Assumption 3 sets the range of penalty parameters necessary to prove consistency.

Assumption 4. *The random error term ϵ_i does not follow any distribution and has a median of* 0.

In other variable selection models, such as lasso and adaptive lasso, the random error term ϵ_i usually obeys the normal distribution. However, for the study of relaxed lad lasso in this paper, the distribution conditions for the random error term are relaxed, and only the median is imposed. All of the above assumptions are necessary for proving the consistency of relaxed lad lasso.

Lemma 1. *Let* $\lim \inf\limits_{n \to \infty} \frac{n^*}{n} \to \frac{1}{R}$ *with* $R \ge 2$. $L_{n^*}(\lambda, \phi)$ *be an empirical loss function of* $L(\lambda, \phi)$, *where n^* is its sample size. Then, under Assumptions 1–4,*

$$\sup_{\lambda \in \mathcal{L}, \phi > 0} |L(\lambda, \phi) - L_{n^*}(\lambda, \phi)| = O_p\left(n^{-1/2} \log n\right), n \to \infty. \tag{15}$$

Lemma 1 will be used to prove the key conclusion in Theorem 4.

According to Lemma 1 of Wang et al. [11], lad lasso's oracle property is dependent on the $\sqrt{n}$-consistency, that is, $\sqrt{n}a_n \to 0$. Therefore, a_n is in a sequence with $o(n^{-1/2})$ as $n \to \infty$. The lad lasso model in this article uses a fixed λ because a_n is the largest λ in the nonzero parameters; then, you can obtain $\lambda = o(n^{-1/2})$.

Theorem 1. *In order to describe the loss under the lad lasso estimator when $n \to \infty$, according to Assumptions 1–4, we have:*

$$\inf_{\lambda} L(\lambda) = O_p\left(n^{-1/2}\right). \tag{16}$$

Theorem 1 first proves the convergence rate of lad lasso. Lad lasso uses the L_1 loss function. According to Pesme and Flammarion [24], it is shown that the L_1 loss function is non-strongly convex. Since the loss function does not have non-derivable points, we can still think that the algorithm is convex. The L_1 loss's non-strong convexity can guarantee $O\left(n^{-1/2}\right)$ convergence speed before and after iteration, and smoothness has no effect on the above conclusion, which indirectly proves that our conclusion is correct.

Theorem 2. *In order to describe the loss under the relaxed lad lasso estimator when $n \to \infty$, according to Assumptions 1–4, we have:*

$$\inf_{\lambda,\phi} L(\lambda,\phi) = O_p\left(n^{-1/2}\right). \tag{17}$$

One of the main contributions of our paper is to prove that the convergence speed of relaxed lad lasso is equivalent to that of lad lasso, even although adding the relaxation parameter ϕ does not improve the convergence speed of relaxed lad lasso. However, when the number of variables p grows exponentially with the sample n, the number of potential noise variables likewise increases significantly, but this will not slow down the convergence speed of relaxed lad lasso. Although the convergence speed of relaxed lad lasso is not ideally as fast as relaxed lasso, it still outperforms lasso due to the existence of L_1 loss and relaxation parameter ϕ, which offers good stability.

Theorem 3. *Under the condition that the design matrix is positive definite and the prediction error ε_i is continuous and has a positive density at the origin, when $\sqrt{n}a_n \to 0$, the estimator of relaxed lad lasso is $\sqrt{n}$-consistency for $\epsilon > 0$, which is*

$$\lim_{n\to\infty} P\left(\left|Q(\hat{\beta}) - Q(\beta)\right| > \epsilon\right) = 0. \tag{18}$$

Another major contribution of this paper is to prove that the the relaxed lad lasso estimator is consistent, where the conclusion of Lemma 1 of Wang et al. [11] in lad lasso is an essential precondition that the important variable's penalty parameter can converge to 0 faster than $n^{-1/2}$. It guarantees the consistency of lad lasso, and our proof is also based on this conclusion.

Theorem 4. *Let $L(\hat{\lambda},\hat{\phi})$ be the loss of the relaxed lad lasso estimate and $(\hat{\lambda},\hat{\phi})$ chosen by K-fold cross-validation with $2 \le K \le \infty$. Under assumptions of relaxed lasso, it holds that*

$$L(\hat{\lambda},\hat{\phi}) = O_p\left(n^{-1/2}\log n\right). \tag{19}$$

We still use K-fold cross-validation when choosing the penalty parameters; that is, we select the optimal penalty parameters $\hat{\lambda}$ and $\hat{\phi}$ by minimizing the empirical loss function of cross-validation. First, define the empirical loss function on a different observation set from $R = 1,\ldots,K$ as

$$L_{cv}(\lambda,\phi) = K^{-1}\sum_{R=1}^{K} L_{R,\widetilde{n}}(\lambda,\phi), \tag{20}$$

where each partition of R consists of $\widetilde{n}$ observations and $L_{R,\widetilde{n}}(\lambda,\phi)$ is the empirical loss of the response variable.

4. Simulation

4.1. Setup

In this section, we present the results of extensive simulations that are conducted to assess how relaxed lad lasso performs in the presence of heavy-tailed errors. For comparison, the performances of the proposed relaxed lad lasso, lasso, lad lasso, and relaxed lasso are evaluated along with some metrics through the results of the mean and median of the mean absolute prediction error (MAPE), the average number of nonzero estimated coefficients (number of nonzeros), and the average number of correctly (mistakenly) estimated zeros. The above metrics are selected the same as Wang et al. [11] and Hastie et al. [5]. The procedure for setting parameters can be summarized as follows:

i. We consider the following regression model $\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$ in this simulation. The predictor matrix $\boldsymbol{X}$ is generated from a p-dimension multivariate normal distribution $N(\boldsymbol{0},\boldsymbol{\Sigma})$ where the covariance matrix $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$. $\boldsymbol{\varepsilon}$ is derived from several heavy-tailed distributions. The density function of the t-distribution shows a heavy tail compared to the standard normal distribution. Therefore, we set $\boldsymbol{\varepsilon}$ to follow t-distribution with 5 degrees of freedom df(t_5) and the standard t-distribution with 3 degrees of freedom df(t_3).

ii. We set the fixed dimension $p = 8$ and vary $n = 50, 100, 200$ to compare the performances of four methods under different sample sizes.

iii. The true regression coefficient $\boldsymbol{\beta} = (0.5, 1, 1.5, 2, 0, 0, 0, 0)$ is an eight-dimension vector where its first four elements are important variables taking nonzero coefficients and otherwise set as 0.

iv. The value of σ is adjusted to achieve different theoretical SNR values. We discuss $\sigma = 0.5$ and $\sigma = 1$ in order to test the effects of strong and weak SNR values on the results.

v. The parameter λ is selected by the five-fold cross-validation, and the mean absolute error is applied for the loss of cross-validation. In addition, 100 simulation iterations are completed for each situation to test the performance of relaxed lad lasso.

4.2. Evaluation Metrics

In order to select the optimal parameters by cross-validation, we divide the data into a training set and a test set. Assume that x_{test} represents a row from the test set's predictor matrix X and that $\hat{\beta}$ and $\hat{y}_{test}$ are the estimated coefficients and the fitting results of x_{test} on the test set. The evaluation metrics we use are shown as follows.

The mean absolute prediction error (MAPE):

$$\text{MAPE} = \text{E}(|Y_{test} - \hat{y}_{test}|) = \text{E}\left(\left|Y_{test} - x_{test}^T\hat{\beta}\right|\right). \tag{21}$$

The number of nonzero estimated coefficients (number of nonzeros):

$$\text{Number of nonzeros} = \|\hat{\beta}\|_0 = \sum_{i=1}^{p} 1\{\hat{\beta}_i \neq 0\}. \tag{22}$$

The number of correctly (mistakenly) estimated zeros (number of zeros):

$$\text{Correct} = \sum_{i=1}^{p} 1\{\hat{\beta}_i = 0,\beta_i = 0\}. \tag{23}$$

$$\text{Incorrect} = \sum_{i=1}^{p} 1\{\hat{\beta}_i = 0,\beta_i \neq 0\}. \tag{24}$$

4.3. Summary of Results

As can be seen, the results of the simulation are summarized in Tables 1 and 2. From the view of the number of selected nonzero and zero variables, all of the methods are shown to be comparable; however, the mean and median results of prediction accuracy are rather different. When the SNR is low (i.e., $\sigma = 0.5$), relaxed lad lasso outperforms lasso, lad lasso, and relaxed lasso with the lowest mean and median of MAPE. Additionally, relaxed lad lasso almost correctly identifies the number of noise variables in the sense that the variable selection results come close to the number of zero coefficients in the true model. In particular, since the real regression model has four variables that are nonzero, relaxed lad lasso correctly selects the number of important variables that is closest to the actual nonzero variables. When the SNR is high (i.e., $\sigma = 1.0$), all methods perform slightly worse than the low SNR situation; nevertheless, relaxed lad lasso remains a competitive method and consistently performs well on these evaluation metrics.

It is worth noting that as the number of observations n increases, the difference between the results of relaxed lad lasso and the worst methods with the same SNR value becomes smaller. For $\sigma = 0.5$ with t_5 error, the numerical difference between the relaxed lad lasso and the lad lasso's mean MAPE is 0.043 in a small number of observations, i.e., $n = 50$. When n increases to 200, the difference between them drops to 0.023. Therefore, relaxed lad lasso stands out when n is small, but as n increases, the advantage of relaxed lad lasso starts to decrease because, in that case, the data asymptotically follow a normal distribution that breaks the condition of heavy-tailed distributions we required. Therefore, we can conclude that in a normal distribution, the performance of relaxed lad lasso is comparable to the traditional robust regression and ordinary lasso method. However, when data have a heavy-tail distribution, relaxed lad lasso has an overall superior performance in terms of prediction accuracy and correct selection of the number of important information variables.

Table 1. Simulation results for t_5 error.

σ	n	Method	Mean MAPE	Median MAPE	Number of Nonzeros	Number of Zeros	
						Incorrect	Corerect
0.5	50	Lasso	0.150	0.148	4.3	0.02	3.67
		Ladlasso	0.176	0.163	2.8	1.23	4.00
		Rlasso	0.142	0.139	3.7	0.38	3.95
		Rladlasso	0.133	0.131	4.1	0.12	3.77
	100	Lasso	0.145	0.141	4.2	0.00	3.76
		Ladlasso	0.158	0.152	3.0	1.01	4.00
		Rlasso	0.137	0.131	3.8	0.19	3.98
		Rladlasso	0.133	0.127	4.2	0.00	3.78
	200	Lasso	0.138	0.130	4.1	0.00	3.88
		Ladlasso	0.151	0.124	3.1	0.88	4.00
		Rlasso	0.129	0.125	4.0	0.03	4.00
		Rladlasso	0.128	0.125	4.2	0.00	3.79
1	50	Lasso	0.307	0.306	4.1	0.25	3.65
		Ladlasso	0.314	0.313	2.3	1.74	4.00
		Rlasso	0.298	0.292	3.1	1.00	3.93
		Rladlasso	0.279	0.280	3.9	0.43	3.69
	100	Lasso	0.277	0.272	4.1	0.10	3.80
		Ladlasso	0.269	0.265	2.8	1.21	4.00
		Rlasso	0.267	0.262	3.3	0.76	3.96
		Rladlasso	0.258	0.255	4.0	0.20	3.79
	200	Lasso	0.251	0.249	4.1	0.02	3.86
		Ladlasso	0.248	0.247	3.0	1.01	4.00
		Rlasso	0.248	0.243	3.3	0.70	4.00
		Rladlasso	0.239	0.235	4.1	0.06	3.83

Table 2. Simulation results for t_3 error.

σ	n	Method	Mean MAPE	Median MAPE	Number of Nonzeros	Number of Zeros	
						Incorrect	Corerect
0.5	50	Lasso	0.184	0.178	4.2	0.10	3.68
		Ladlasso	0.205	0.198	2.7	1.33	4.00
		Rlasso	0.172	0.170	3.4	0.62	3.96
		Rladlasso	0.157	0.154	4.1	0.17	3.75
	100	Lasso	0.172	0.166	4.3	0.02	3.65
		Ladlasso	0.175	0.171	3.0	0.99	4.00
		Rlasso	0.164	0.159	3.6	0.36	4.00
		Rladlasso	0.154	0.150	4.2	0.00	3.79
	200	Lasso	0.162	0.160	4.1	0.00	3.88
		Ladlasso	0.170	0.171	3.1	0.89	4.00
		Rlasso	0.153	0.153	3.9	0.15	4.00
		Rladlasso	0.147	0.146	4.2	0.00	3.85
1	50	Lasso	0.339	0.324	3.8	0.59	3.58
		Ladlasso	0.340	0.320	2.1	1.89	4.00
		Rlasso	0.325	0.314	3.0	1.12	3.92
		Rladlasso	0.299	0.290	3.8	0.54	3.70
	100	Lasso	0.317	0.311	3.9	0.29	3.82
		Ladlasso	0.292	0.288	2.8	1.23	4.00
		Rlasso	0.303	0.286	3.0	1.02	4.00
		Rladlasso	0.280	0.274	4.0	0.22	3.80
	200	Lasso	0.308	0.308	4.0	0.10	3.90
		Ladlasso	0.286	0.283	3.0	1.00	4.00
		Rlasso	0.295	0.293	3.0	0.96	4.00
		Rladlasso	0.279	0.276	4.1	0.06	3.89

5. Application to Real Data

5.1. Dataset

The Research and Development (R&D) investment is critical for a company's operations in the current competitive environment, regardless of industry. The problem of identifying the primary factors affecting the R&D investment has been extensively researched to maintain competitiveness and improve innovation. The real data for this study came from the CSMAR database, which is considered one of the most professional and extensively used research databases available. The data have 2137 records, each of which corresponds to the financial data of a single publicly traded firm in 2021. We split the data into a training set and a test set with with a ratio of 7:3 so that the training set can be used to fit the model and the MAPE is measured on the test set. The R&D investment of a corporation is the response variable, and there are 86 predictor variables, such as management costs, operating costs, net profit, and other financial indicators that may impact a company's R&D expenditure. Table 3 provides a full overview of these factors. Due to the large variance in the R&D investment between industries, the response variables may have heavy-tailed errors or outliers. Therefore, we check to find that the residuals differentiated by an OLS fit have a kurtosis of 144.38, which is significantly greater than the normal distribution's value. Furthermore, we show the box plot of R&D investment and the QQ-plot of the OLS fit in Figures 1 and 2. The block dots in Figure 1 and the blue dots outside the 95% confidence interval in Figure 2 indicate that the response variable contains a large number of outliers. Note that we take the logarithm of the response value. Consequently, the dependability of conventional OLS-based estimators and model selection methods (e.g., lasso, relaxed lasso) is substantially compromised. To confirm the previously stated conclusion in Section 5, we continue to calculate the MAPE to compare the performances of the four methods that appear in the simulation.

Table 3. Description of variables in R&D investment data.

Variables	Description	Symbols
R&D investment	Research and Development Costs	Y
Profitability	Finance Costs (x_1), Payback On Assets (x_2), Operating Costs (x_3), …	$x_1, \ldots, x_{15}$
Business Capability	Net Accounts Receivable (x_{16}), Business Cycle (x_{17}), Current Assets (x_{18}), …	$x_{16}, \ldots, x_{30}$
Assets and Liabilities	Total Current Liabilities (x_{31}), Taxes Payable (x_{32}), Accounts Payable (x_{33}), …	$x_{31}, \ldots, x_{50}$
Profits	Operating Profit (x_{51}), Total Comprehensive Income (x_{52}), …	$x_{51}, \ldots, x_{70}$
Cash Flow	Cash Paid For Goods (x_{71}), Net Cash Flows From Investing Activities (x_{72}), …	$x_{71}, \ldots, x_{86}$

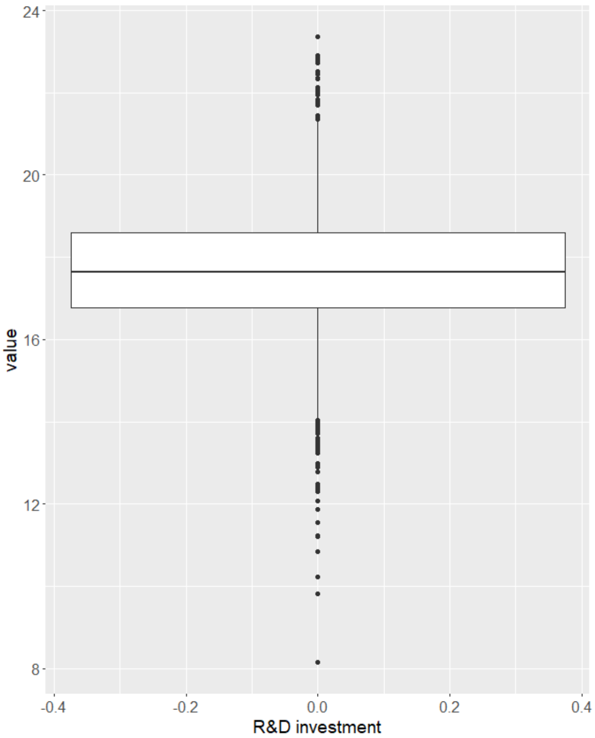

Figure 1. The box plot of the company's R&D investment. The box plot indicates outliers with black dots above the upper quartile plus 1.5 times the quartile difference or below the lower quartile minus 1.5 times the quartile difference.

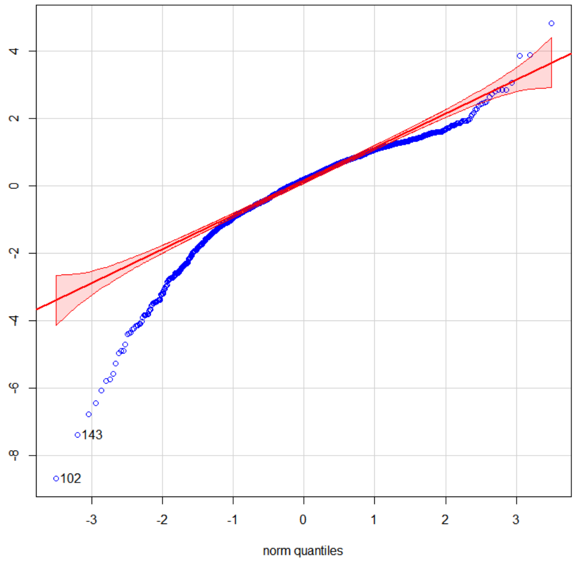

Figure 2. The QQ plot of the OLS fit. The red shaded area is the 95% confidence interval for the standard straight line $y = x$.

5.2. Analysis Results

In Table 4, relaxed lad lasso outperforms all competitors in terms of prediction accuracy, with the smallest MAPE of 0.184, as has been demonstrated in the simulation. Lad lasso and relaxed lasso have MAPEs of 0.201 and 0.191, respectively. Lasso is the worst method. Specifically, the MAPE of lasso is 0.203, which is slightly larger than that of lad lasso. The relevant variables selected by the best resulting model are listed in Table 5. We find that Net Accounts Receivable, Funds Paid to and for Staff, Other Income, Gains and Losses from Asset Disposition, Interest Income, and Basic Earnings Per Share are the most important factors influencing the R&D spending. Therefore, we can conclude that relaxed lad lasso obtains the sparse model with the highest prediction accuracy for data with heavy-tailed errors.

Table 4. Prediction accuracy for R&D investment study.

Method	Lasso	Ladlasso	Rlasso	Rladlasso
MAPE	0.203	0.201	0.191	0.184

Table 5. Variables selected by relaxed lad lasso.

Order Number	Explanatory Variable	Coefficient
x_{16}	Net Accounts Receivable	0.297
x_{61}	Basic Earnings Per Share	0.023
x_{67}	Interest Income	0.115
x_{68}	Other Income	0.197
x_{70}	Gains and Losses from Asset Disposition	0.154
x_{74}	Funds Paid to and for Staff	0.251

Among the most important variables selected by relaxed lad lasso, Net Accounts Receivable indicates the volume of products sold by a business that have not been paid for; Gains and Losses from Asset Disposition, Interest Income, and Other Income measure the incomes of the company's operations; Basic Earnings Per Share reflects the profitability of the enterprise over a certain period; Funds Paid to and for Staff measures the company's benefits and rewards provided to its staff. The estimated coefficients of Net Accounts

Receivable and Funds Paid to and for Staff are 0.297 and 0.251, which both have relatively large positive effects on R&D investment. Then, the coefficients of Other Income, Profit and Loss from Asset Disposal, and Interest Income are 0.197, 0.154, and 0.115 as the decline of influence to the response variable. It is not surprising that the sales volume of products and profits influence the company's decision to promote innovation and improve technological development. To a certain extent, with a significant volume of sales and a consistent and large cash flow, the accumulated capital can be used to invest in the company's R&D investment. What is more, welfare-oriented businesses with attractive compensation will help executives act in the company's long-term interest to maximize shareholders' interests so that they will pay more attention to the innovation of their companies. In general, raising a company's R&D investment is heavily driven by a few critical factors, which can be summarized as sales volume, profitability, and staff welfare.

6. Conclusions

In this paper, we develop the relaxed lad lasso method for both variable selection and shrinkage estimation that is resistant to heavy-tailed errors or outliers in the response. As a combination of the ideas of relaxed lasso and lad lasso, the new estimator inherits good properties of lad lasso and can be solved using the same efficient algorithm for solving the LAD program. Theoretically, we have proven that relaxed lad lasso has the same convergence rate as lad lasso with $O_p\left(n^{-1/2}\right)$, and it is $\sqrt{n}$-consistent under mild conditions on predictors and the growth mode of the variable dimension. Additionally, we have shown that the rate to choose parameters $(\hat{\lambda}, \hat{\phi})$ by K-fold cross-validation is $O_p\left(n^{-1/2}\log n\right)$ fast. In the simulation, the proposed method produces more correct variable selection results and lower prediction errors than lasso, relaxed lasso, and lad lasso. It also performs well in the application of the company's R&D investment data. For further research, it is suggested that the comparable idea can be extended to Huber's M-estimation for a faster convergence rate. From the perspective of regression models, Contreras-Reyes et al. [25] uses a log-skew-t non-linear regression to analyze the Von Bertalanffy growth models (VBGMs). Motivated by this, the non-linear regression can also be improved by the proposed method under the heavy-tailed distribution.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L. and H.L.; software, H.L. and X.X.; validation, H.L., X.X., Y.L., X.Y., T.Z. and R.Z; writing—original draft preparation, X.X. and T.Z.; writing—review and editing, X.X., X.Y. and R.Z.; project administration, R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the S&T Program of Hebei (22557688D), the Research Project on the development of Military and Civil integration in Hebei Province (HB22JMRH025), and the Hebei Province social science development research project (20220202086).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Appendix Proof of Lemma 1

Proof. In Section 2, we have already known that the solutions of relaxed lad lasso can be viewed as a combination of lad lasso and the LAD estimator. Thus, we define the set of relaxed lad lasso solutions $H_1, \dots H_m$ as

$$H_t = \left\{\hat{\beta}^* = \phi\hat{\beta}^{Ladlasso} + (1-\phi)\hat{\beta}^{Lad}\right\}. \tag{A1}$$

where $0 < \phi < 1$. We arrange the λ sequences in descending order, that is $\lambda_m > \ldots > \lambda_1$. Let $\lambda_t, t = 1, \ldots, m$ be the largest penalty parameter selected such that $S_t = S_\lambda$, where S_t is the set of models estimated by lad lasso.
The loss function of relaxed lad lasso is as follows:

$$L(\lambda, \phi) = E\left|Y - \sum_{t \in \{1,\ldots,p\}} \hat{\beta}_t^* X^t\right|. \tag{A2}$$

Simplify Formula (A2) by using (A1) to obtain

$$L(\lambda, \phi) = E\left|Y - \sum_{t \in \{1,\ldots,p\}} \hat{\beta}^{Lad} X^t - \phi\left(\hat{\beta}^{Ladlasso} - \hat{\beta}^{Lad}\right) X^t\right|. \tag{A3}$$

To simplify the representation, for any λ, set

$$W_\lambda = Y - \sum_{t \in \{1,\ldots,p\}} \hat{\beta}^{Lad} X^t, \tag{A4}$$

and

$$Z_\lambda = \phi\left(\sum_{t \in \{1,\ldots,p\}} \hat{\beta}^{Ladlasso} - \sum_{t \in \{1,\ldots,p\}} \hat{\beta}^{Lad}\right) X^t. \tag{A5}$$

Then, we have

$$L(\lambda, \phi) \geq E|W_\lambda| - E|Z_\lambda|. \tag{A6}$$

We set $|W_\lambda| = c$. Bernstein's inequality indicates that there exists a small constant g such that

$$P\left(\frac{1}{n}\sum c_i - Ec < \frac{g}{n}\log(1-\theta) + \sqrt{\frac{-2\mathrm{var}(c)\log(\theta)}{n}}\right) \geq 1 - \theta. \tag{A7}$$

Let $\theta = \frac{1}{n}$; then, we have

$$P\left(E_{n^*}|W_\lambda| - E|W_\lambda| < kn^{*-1/2}\log n\right) \geq 1 - \frac{1}{n}, \tag{A8}$$

where $k > 0$, so for any $\varepsilon > 0$, taking the limit gives

$$\limsup_{n \to \infty} P\left(|E_{n^*}|W_\lambda| - E|W_\lambda|| > kn^{*-1/2}\log n\right) < \varepsilon. \tag{A9}$$

For Z_λ, we use the same steps to obtain

$$\limsup_{n \to \infty} P\left(|E_{n^*}|Z_\lambda| - E|Z_\lambda|| > kn^{*-1/2}\log n\right) < \varepsilon. \tag{A10}$$

From (A9), (A10) and simple algebraic operations, we obtain

$$\limsup_{n \to \infty} P\left(\left|\sup_{\lambda \in \mathcal{L}, \phi > 0} |L(\lambda, \phi) - L_{n^*}(\lambda, \phi)|\right| > kn^{*-1/2}\log n\right) < \varepsilon, \tag{A11}$$

which completes the proof. □

Appendix B. Appendix Proof of Theorem 1

Proof. We have defined the loss function $L(\lambda, \phi)$ for relaxed lad lasso. Similarly, the loss for the lad lasso estimator with the selected parameter λ can be written as

$$L(\lambda) = \sum_{t \in \{1,\cdots,p\}} \left|\hat{\beta}_t^\lambda - \beta_t\right|. \tag{A12}$$

Let λ_* denote the smallest penalty parameter so that unimportant variables can no longer enter the active set. The definition is as follows:

$$\lambda_* = \min_{\lambda \geq 0} \left\{ \lambda | \hat{\beta}_t^\lambda = 0, \forall t > q \right\}. \tag{A13}$$

Only nonzero coefficients, or components of $t \leq q$ in (A12), are included in our summation. When $\lambda \geq \lambda_*$, the lower bound of the loss $L(\lambda)$ satisfies

$$\inf_{\lambda \geq \lambda_*} L(\lambda) \geq q(1-\varepsilon)\lambda_*. \tag{A14}$$

Let $M = \beta - \hat{\beta}^{\lambda_*}$, $N^\lambda = \hat{\beta}^\lambda - \hat{\beta}^{\lambda_*}$; then, we can write in another way, that is

$$\left| \hat{\beta}_t^\lambda - \beta_t \right| = \sqrt{M_t^2 - 2M_t N_t^\lambda + \left(N_t^\lambda\right)^2}. \tag{A15}$$

For $n \to \infty$, any $\delta > 0$, we have $P(|M_t| > (1-\delta)\lambda_*) = 1$. Then, $|M_t| < (1+\delta)\lambda_*$ is always established. Hence, for all $t \leq q$, there is

$$\left| \hat{\beta}_t^\lambda - \beta_t \right| \geq \sqrt{(1-\delta)^2 \lambda_*^2 - 2(1-\delta^2)\lambda_*(\lambda_* - \lambda) + (1-\delta)^2(\lambda_* - \lambda)^2}. \tag{A16}$$

Therefore, taking the lower bound on the right-hand side of the inequality yields

$$\inf_{\lambda \geq \lambda_*} L(\lambda) \geq \left[(1-\delta)^2 + 2\sqrt{q}\left(1-\delta^2\right) + q(1-\delta)^2 \right] \lambda_*. \tag{A17}$$

According to lad lasso's $\sqrt{n}$-consistency: $\lambda \sim n^{-\frac{1}{2}}$,

$$\inf_{\lambda \geq \lambda_*} L(\lambda) \sim O_p\left(n^{-\frac{1}{2}}\right), \tag{A18}$$

which completes the proof. □

Appendix C. Appendix Proof of Theorem 2

Proof. Let $S_* = \{1, \ldots, q\}$ represent the active set, i.e., the set of variables whose coefficients are nonzero. Define event A as

$$\exists \lambda : S_\lambda = S_*. \tag{A19}$$

Let constant $c > 0$. Using the conditional probability inequality, there is

$$P\left(\inf_{\lambda,\phi} L(\lambda, \phi) > cn^{-1/2} \right) \leq P\left(\inf_{\lambda,\phi} L(\lambda, \phi) > cn^{-1/2} | A \right) P(A) + P(A^c). \tag{A20}$$

We define the lad estimator's loss function as L_*; then, we have

$$P\left(\inf_{\lambda,\phi} L(\lambda, \phi) > cn^{-1/2} \right) \leq P\left(L_* > cn^{-1/2} \right) + P(A^c). \tag{A21}$$

The second term on the right-hand side of the above inequality is 0 because for $n \to \infty$, we have $P(A^c) \to 0$. From the property of the lad estimator which has been shown in Theorem 1, the first item on the right-hand side in (A21) satisfies

$$\limsup_{n \to \infty} P\left(L_* > cn^{-1/2} \right) < \varepsilon, \tag{A22}$$

which completes the proof. □

Appendix D. Appendix Proof of Theorem 3

Proof. To prove the consistency, we need to prove the following formula

$$P\left\{\inf_{\|\mathbf{v}\|=C} Q(\hat{\beta}) > Q(\beta)\right\} \geq 1-\epsilon, \tag{A23}$$

where $\mathbf{v} = \sqrt{n}(\hat{\beta}-\beta)$ is a vector with p dimensions such that $\|\mathbf{v}\| = C$, C is a large constant. $Q(\beta)$ is the relaxed lad lasso criterion. Define $D_n(\mathbf{v}) \equiv Q\left(\beta+\frac{\mathbf{v}}{\sqrt{n}}\right) - Q(\beta)$, then

$$\begin{aligned} D_n(\mathbf{v}) &= \sum_{i=1}^{n}\left\{\left|Y_i - X_i'\left(\beta+\frac{\mathbf{v}}{\sqrt{n}}\right)\right| - \left|Y_i - X_i'\beta\right|\right\} + n\lambda\phi\sum_{j=1}^{p}\left\{\left|\beta_j+\frac{\mathbf{v}}{\sqrt{n}}\right| - |\beta_j|\right\} \\ &\geq \sum_{i=1}^{n}\left\{\left|Y_i - X_i'\left(\beta+\frac{\mathbf{v}}{\sqrt{n}}\right)\right| - \left|Y_i - X_i'\beta\right|\right\} - \sqrt{n}a_n\phi\sum_{j=1}^{p}|\beta_j|. \end{aligned} \tag{A24}$$

According to Fu and Knight [23], for $a \neq 0$, it is true that

$$|a-b| - |a| = -b[I(a>0) - I(a<0)] + 2\int_0^b [I(a \leq s) - I(a \geq s)]ds. \tag{A25}$$

Applying the foregoing equation,

$$\sum_{i=1}^{n}\left\{\left|Y_i - X_i'\left(\beta+\frac{\mathbf{v}}{\sqrt{n}}\right)\right| - \left|Y_i - X_i'\beta\right|\right\} \tag{A26}$$

can be expressed as

$$-\frac{\mathbf{v}'}{\sqrt{n}}\sum_{i=1}^{n}[I(\varepsilon_i>0) - I(\varepsilon_i<0)] + 2\sum_{i=1}^{n}\int_0^{\frac{\mathbf{v}'X_i}{\sqrt{n}}}[I(\varepsilon_i>0) - I(\varepsilon_i<0)]ds. \tag{A27}$$

According to the central limit theorem, the distribution of the first item converges to $\mathbf{v}'W$, where W is a matrix with a mean of $\mathbf{0}$ and a variance of $\boldsymbol{\Sigma} = \mathrm{cov}(X_1)$. Denote the item $\int_0^{\frac{\mathbf{v}'X_i}{\sqrt{n}}}[I(\varepsilon_i>0) - I(\varepsilon_i<0)]ds$ by $F_{ni}(\mathbf{v})$. It is difficult to directly find what value $\sum_{i=1}^{n} F_{ni}(\mathbf{v})$ converges to according to the probability. We hope to use $\sum_{i=1}^{n}[F_{ni}(\mathbf{v}) - E(F_{ni}(\mathbf{v}))] = o_p(1)$ to transform the desired problem and then prove the "bridge". Hence,

$$\begin{aligned} nE\left[F_{ni}^2(\mathbf{v})I\left(\frac{|\mathbf{v}'X_i|}{\sqrt{n}} \geq c\right)\right] \leq & nE\left\{\left(\int_0^{\frac{\mathbf{v}'X_i}{\sqrt{n}}} 2ds\right)^2 I\left(\frac{|\mathbf{v}'X_i|}{\sqrt{n}} \geq c\right)\right\} \\ =& 4E\left[\left|\mathbf{v}'X_i\right|^2 I\left(\left|\mathbf{v}'X_i\right| \geq \sqrt{n}c\right)\right] \\ =& o(1). \end{aligned} \tag{A28}$$

Alternatively, owing to the continuity of $g(x)$, there exist $c > 0$ and $0 < d < \infty$ such that $\sup_{|x|<c} g(x) < g(0) + d$. Then, $nE\left[F_{ni}^2(\mathbf{v}) I\left(\frac{|\mathbf{v}'X_i|}{\sqrt{n}} < c\right)\right]$ is dominated by

$$\begin{aligned} nE\left[F_{ni}^2(\mathbf{v}) I\left(\frac{|\mathbf{v}'X_i|}{\sqrt{n}} < c\right)\right] &\leq 2ncE\left[F_{ni}^2(\mathbf{v}) I\left(\frac{|\mathbf{v}'X_i|}{\sqrt{n}} < c\right)\right] \\ &\leq 2ncE\left\{ \int_0^{\frac{\mathbf{v}'X_i}{\sqrt{n}}} [G(s) - G(0)]ds \cdot I\left(\left|\mathbf{v}'X_i\right| < \sqrt{n}c\right) \right\} \\ &\leq 2nc\{g(0) + d\}E\left\{ \int_0^{\frac{\mathbf{v}'X_i}{\sqrt{n}}} sds \cdot I\left(\left|\mathbf{v}'X_i\right| < \sqrt{n}c\right) \right\} \\ &\leq c\{g(0) + d\}E\left|\mathbf{v}'X_i\right|^2, \end{aligned} \tag{A29}$$

which converges to 0 as $c \to 0$. Therefore, as $n \to \infty$, $nE\left[F_{ni}^2(\mathbf{v})\right] \to 0$, we have

$$\operatorname{var}\left(\sum_{i=1}^n F_{ni}\right) = \sum_{i=1}^n \operatorname{var}(F_{ni}) \leq nE\left[F_{ni}^2(\mathbf{v})\right] \to 0. \tag{A30}$$

This completes the proof of $\sum\limits_{i=1}^n [F_{ni}(\mathbf{v}) - E(F_{ni}(\mathbf{v}))] = o_p(1)$. Furthermore, we turn the problem to what value $E\left(\sum\limits_{i=1}^n F_{ni}\right)$ will converge to probabilistically, and $\sum\limits_{i=1}^n F_{ni}$ will also converge to this value.

$$\begin{aligned} E\left(\sum_{i=1}^n F_{ni}\right) &= nE[F_{ni}(\mathbf{v})] \\ &= nE\left\{ \int_0^{\frac{\mathbf{v}'X_i}{\sqrt{n}}} [G(s) - G(0)]ds \right\} \\ &= nE\left\{ \int_0^{\frac{\mathbf{v}'X_i}{\sqrt{n}}} sg(0)ds \right\} + o(1) \\ &= \frac{1}{2} g(0)\mathbf{v}' \frac{\left(X_i X_i'\right)}{n} \mathbf{v} \end{aligned} \tag{A31}$$

due to $P\left\{n^{-1/2}\max(|\mathbf{v}'X_1|, \ldots, |\mathbf{v}'X_n|) > c^*\right\} \to 0$. According to the law of large numbers,

$$\sum_{i=1}^n F_{ni} \to_p \frac{1}{2} g(0)\mathbf{v}'\Sigma\mathbf{v}. \tag{A32}$$

Therefore, the second item on the right side of (A27) converges to $g(0)\mathbf{v}'\Sigma\mathbf{v}$ according to probability. The proof is completed by choosing C large enough so that the second term of (A27) uniformly dominates the first term with $\|\mathbf{v}\| = C$. □

Appendix E. Appendix Proof of Theorem 4

Proof. Firstly, from Lemma 1, we can obtain the following inequality for $\hat{\lambda}$, $\hat{\varphi}$ and $c > 0$,

$$P\left(L(\hat{\lambda}, \hat{\varphi}) > cn^{-1/2} \log n\right) \leq 2\varepsilon. \tag{A33}$$

Then, we have

$$\begin{aligned} &P\Big(L(\hat{\lambda},\hat{\phi}) > cn^{-1/2}\log n\Big) \le P\Big(L_{cv}(\hat{\lambda},\hat{\phi}) > cn^{-1/2}\log n\Big) \\ &\le 2P\bigg(\sup\big|L(\hat{\lambda},\hat{\phi}) - L_{cv}(\hat{\lambda},\hat{\phi})\big| > \frac{1}{2}cn^{-1/2}\log n\bigg) + P\bigg(\inf L(\hat{\lambda},\hat{\phi}) > \frac{1}{2}cn^{-1/2}\log n\bigg). \end{aligned} \tag{A34}$$

The last term of this equation is given by the Bonferroni's inequality. Hence, for each $\varepsilon > 0$, there exists $c > 0$ such that

$$\lim\sup_{n\to\infty} P\Big(L(\hat{\lambda},\hat{\phi}) > cn^{-1/2}\log n\Big) < \varepsilon, \tag{A35}$$

which completes the proof. □

References

1. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288.
2. Wu, C.; Ma, S. A selective review of robust variable selection with applications in bioinformatics. *Briefings Bioinform.* **2015**, *16*, 873–883.
3. Uraibi, H.S. Weighted Lasso Subsampling for HighDimensional Regression. *Electron. J. Appl. Stat. Anal.* **2019**, *12*, 69–84.
4. Meinshausen, N. Relaxed lasso. *Comput. Stat. Data Anal.* **2007**, *52*, 374–393.
5. Hastie, T.; Tibshirani, R.; Tibshirani, R.J. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv* **2017**, arXiv:1707.08692.
6. Hastie, T.; Tibshirani, R.; Tibshirani, R.J. Rejoinder: Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Stat. Sci.* **2020**, *35*, 625–626.
7. Mentch, L.; Zhou, S. Randomization as regularization: A degrees of freedom explanation for random forest success. *J. Mach. Learn. Res.* **2020**, *21*, 1–36.
8. Bloise, F.; Brunori, P.; Piraino, P. Estimating intergenerational income mobility on sub-optimal data: A machine learning approach. *J. Econ. Inequal.* **2021**, *19*, 643–665.
9. He, Y. The Analysis of Impact Factors of Foreign Investment Based on Relaxed Lasso. *J. Appl. Math. Phys.* **2017**, *5*, 693–699.
10. Gao, X. Estimation and Selection Properties of the LAD Fused Lasso Signal Approximator. *arXiv* **2021**, arXiv:2105.00045.
11. Wang, H.; Li, G.; Jiang, G. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econ. Stat.* **2007**, *25*, 347–355.
12. Gao, X.; Huang, J. Asymptotic analysis of high-dimensional LAD regression with LASSO. *Stat. Sin.* **2010**, *20*, 1485–1506.
13. Xu, J.; Ying, Z. Simultaneous estimation and variable selection in median regression using Lasso-type penalty. *Ann. Inst. Stat. Math.* **2010**, *62*, 487–514.
14. Arslan, O. Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Comput. Stat. Data Anal.* **2012**, *56*, 1952–1965.
15. Rahardiantoro, S.; Kurnia, A. Lad-lasso: Simulation study of robust regression in high dimensional data. *Forum Statistika dan Komputasi.* **2020**, *20*.
16. Zhou, X.; Liu, G. LAD-lasso variable selection for doubly censored median regression models. *Commun. Stat. Theory Methods* **2016**, *45*, 3658–3667.
17. Li, Q.; Wang, L. Robust change point detection method via adaptive LAD-LASSO. *Stat. Pap.* **2020**, *61*, 109–121.
18. Croux, C.; Filzmoser, P.; Pison, G.; Rousseeuw, P.J. Fitting multiplicative models by robust alternating regressions. *Stat. Comput.* **2003**, *13*, 23–36.
19. Giloni, A.; Simonoff, J.S.; Sengupta, B. Robust weighted LAD regression. *Comput. Stat. Data Anal.* **2006**, *50*, 3124–3140.
20. Xue, F.; Qu, A. Variable selection for highly correlated predictors. *arXiv* **2017**, arXiv:1709.04840.
21. Gao, X.; Feng, Y. Penalized weighted least absolute deviation regression. *Stat. Interface* **2018**, *11*, 79–89.
22. Jiang, Y.; Wang, Y.; Zhang, J.; Xie, B.; Liao, J.; Liao, W. Outlier detection and robust variable selection via the penalized weighted LAD-LASSO method. *J. Appl. Stat.* **2021**, *48*, 234–246.
23. Fu, W.; Knight, K. Asymptotics for lasso-type estimators. *Ann. Stat.* **2000**, *28*, 1356–1378.
24. Pesme, S.; Flammarion, N. Online robust regression via sgd on the l1 loss. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2540–2552.
25. Contreras-Reyes, J.E.; Arellano-Valle, R.B.; Canales, T.M. Comparing growth curves with asymmetric heavy-tailed errors: Application to the southern blue whiting (*Micromesistius australis*). *Fish. Res.* **2014**, *159*, 88–94.