*Submitted to the Annals of Statistics*

# ADAPTIVE ROBUST VARIABLE SELECTION

By Jianqing Fan*, Yingying Fan† and Emre Barut

*Princeton University, University of Southern California and IBM T.J. Watson Research Center*

Heavy-tailed high-dimensional data are commonly encountered in various scientific fields and pose great challenges to modern statistical analysis. A natural procedure to address this problem is to use penalized quantile regression with weighted $L_1$-penalty, called weighted robust Lasso (WR-Lasso), in which weights are introduced to ameliorate the bias problem induced by the $L_1$-penalty. In the ultra-high dimensional setting, where the dimensionality can grow exponentially with the sample size, we investigate the model selection oracle property and establish the asymptotic normality of the WR-Lasso. We show that only mild conditions on the model error distribution are needed. Our theoretical results also reveal that adaptive choice of the weight vector is essential for the WR-Lasso to enjoy these nice asymptotic properties. To make the WR-Lasso practically feasible, we propose a two-step procedure, called adaptive robust Lasso (AR-Lasso), in which the weight vector in the second step is constructed based on the $L_1$-penalized quantile regression estimate from the first step. This two-step procedure is justified theoretically to possess the oracle property and the asymptotic normality. Numerical studies demonstrate the favorable finite-sample performance of the AR-Lasso.

**1. Introduction.** The advent of modern technology makes it easier to collect massive, large-scale data sets. A common feature of these data sets is that the number of covariates greatly exceeds the number of observations, a regime opposite to conventional statistical settings. For example, portfolio allocation with hundreds of stocks in finance involves a covariance matrix of about tens of thousands of parameters, but the sample sizes are often only in the order of hundreds (e.g., daily data over a year period (Fan et al., 2008)). Genome-wide association studies in biology involve hundreds of thousands of single-nucleotide polymorphisms (SNPs), but the available

*Jianqing Fan's research was supported by the National Institutes of Health grant R01-GM072611 and National Science Foundation grants DMS-1206464 and DMS-0704337.

†Yingying Fan's research was supported by NSF CAREER Award DMS-1150318 and Grant DMS-0906784, the 2010 Zumberge Individual Award from USC's James H. Zumberge Faculty Research and Innovation Fund, and USC Marshall summer research funding.

*AMS 2000 subject classifications:* Primary 62J07; secondary 62H12

*Keywords and phrases:* Adaptive weighted $L_1$, High dimensions, Oracle properties, Robust regularization

sample size is usually in hundreds too. Data-sets with large number of variables but relatively small sample size pose great unprecedented challenges, and opportunities, for statistical analysis.

Regularization methods have been widely used for high-dimensional variable selection (Bickel and Li, 2006; Bickel et al., 2009; Efron et al., 2007; Fan and Li, 2001; Lv and Fan, 2009; Tibshirani, 1996; Zhang, 2010; Zou, 2006). Yet, most existing methods such as penalized least-squares or penalized likelihood (Fan and Lv, 2011) are designed for light-tailed distributions. Zhao and Yu (2006) established the irrepresentability conditions for the model selection consistency of the Lasso estimator. Fan and Li (2001) studied the oracle properties of nonconcave penalized likelihood estimators for fixed dimensionality. Lv and Fan (2009) investigated the penalized least-squares estimator with folded-concave penalty functions in the ultra-high dimensional setting and established a nonasymptotic weak oracle property. Fan and Lv (2008) proposed and investigated the sure independence screening method in the setting of light-tailed distributions. The robustness of the aforementioned methods have not yet been thoroughly studied and well understood.

Robust regularization methods such as the least absolute deviation (LAD) regression and quantile regression have been used for variable selection in the case of fixed dimensionality. See, for example, Li and Zhu (2008); Wang, Li and Jiang (2007); Wu and Liu (2009); Zou and Yuan (2008). The penalized composite likelihood method was proposed in Bradic et al. (2011) for robust estimation in ultra-high dimensions with focus on the efficiency of the method. They still assumed sub-Gaussian tails. Belloni and Chernozhukov (2011) studied the $L_1$-penalized quantile regression in high-dimensional sparse models where the dimensionality could be larger than the sample size. We refer to their method as robust Lasso (R-Lasso). They showed that the R-Lasso estimate is consistent at the near-oracle rate, and gave conditions under which the selected model includes the true model, and derived bounds on the size of the selected model, uniformly in a compact set of quantile indices. Wang (2012) studied the $L_1$-penalized LAD regression and showed that the estimate achieves near oracle risk performance with a nearly universal penalty parameter and established also a sure screening property for such an estimator. van de Geer and Müller (2012) obtained bounds on the prediction error of a large class of $L_1$ penalized estimators, including quantile regression. Wang et al. (2012) considered the nonconvex penalized quantile regression in the ultra-high dimensional setting and showed that the oracle estimate belongs to the set of local minima of the nonconvex penalized quantile regression, under mild assumptions on the er-

ror distribution.

In this paper, we introduce the penalized quantile regression with the weighted $L_1$-penalty (WR-Lasso) for robust regularization, as in Bradic et al. (2011). The weights are introduced to reduce the bias problem induced by the $L_1$-penalty. The flexibility of the choice of the weights provides flexibility in shrinkage estimation of the regression coefficient. WR-Lasso shares a similar spirit to the folded-concave penalized quantile-regression (Wang et al., 2012; Zou and Li, 2008), but avoids the nonconvex optimization problem. We establish conditions on the error distribution in order for the WR-Lasso to successfully recover the true underlying sparse model with asymptotic probability one. It turns out that the required condition is much weaker than the sub-Gaussian assumption in Bradic et al. (2011). The only conditions we impose is that the density function of error has Lipschitz property in a neighborhood around 0. This includes a large class of heavy-tailed distributions such as the stable distributions, including the Cauchy distribution. It also covers the double exponential distribution whose density function is nondifferentiable at the origin.

Unfortunately, because of the penalized nature of the estimator, WR-Lasso estimate has a bias. In order to reduce the bias, the weights in WR-Lasso need to be chosen adaptively according to the magnitudes of the unknown true regression coefficients, which makes the bias reduction infeasible for practical applications.

To make the bias reduction feasible, we introduce the adaptive robust Lasso (AR-Lasso). The AR-Lasso first runs R-Lasso to obtain an initial estimate, and then computes the weight vector of the weighted $L_1$-penalty according to a decreasing function of the magnitude of the initial estimate. After that, AR-Lasso runs WR-Lasso with the computed weights. We formally establish the model selection oracle property of AR-Lasso in the context of Fan and Li (2001) with no assumptions made on the tail distribution of the model error. In particular, the asymptotic normality of the AR-Lasso is formally established.

This paper is organized as follows. First, we introduce our robust estimators in Section 2. Then, to demonstrate the advantages of our estimator, we show in Section 3 with a simple example that Lasso behaves sub-optimally when noise has heavy tails. In Section 4.1, we study the performance of the oracle-assisted regularization estimator. Then in Section 4.2, we show that when the weights are adaptively chosen, WR-Lasso has the model selection oracle property, and performs as well as the oracle-assisted regularization estimate. In Section 4.3, we prove the asymptotic normality of our proposed estimator. The feasible estimator, AR-Lasso, is investigated in Section 5.

Section 6 presents the results of the simulation studies. Finally, in Section 7 we present the proofs of the main theorems. Additional proofs, as well as the results of a genome-wide association study, are provided in the supplementary Appendix (Fan et al., 2013).

**2. Adaptive Robust Lasso.** Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y}$ is an $n$-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T = (\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_p)$ is an $n \times p$ fixed design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n$-dimensional error vector whose components are independently distributed and satisfy $P(\varepsilon_i \leq 0) = \tau$ for some known constant $\tau \in (0, 1)$. Under this model, $\mathbf{x}_i^T\boldsymbol{\beta}$ is the conditional $\tau$th-quantile of $y_i$ given $\mathbf{x}_i$. We impose no conditions on the heaviness of the tail probability or the homoscedasticity of $\varepsilon_i$. We consider a challenging setting in which $\log p = o(n^b)$ with some constant $b > 0$. To ensure the model identifiability and to enhance the model fitting accuracy and interpretability, the true regression coefficient vector $\boldsymbol{\beta}^*$ is commonly imposed to be sparse with only a small proportion of nonzeros (Fan and Li, 2001; Tibshirani, 1996). Denoting the number of nonzero elements of the true regression coefficients by $s_n$, we allow $s_n$ to slowly diverge with the sample size $n$ and assume that $s_n = o(n)$. To ease the presentation, we suppress the dependence of $s_n$ on $n$ whenever there is no confusion. Without loss of generality, we write $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \mathbf{0}^T)^T$, i.e. only the first $s$ entries are non-vanishing. The true model is denoted by

$$\mathcal{M}_* = \text{supp}(\boldsymbol{\beta}^*) = \{1, \cdots, s\},$$

and its complement, $\mathcal{M}_*^c = \{s+1, \cdots, p\}$, represents the set of noise variables.

We consider a fixed design matrix in this paper and denote by $\mathbf{S} = (\mathbf{S}_1, \cdots, \mathbf{S}_n)^T = (\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_s)$ the submatrix of $\mathbf{X}$ corresponding to the covariates whose coefficients are non-vanishing. These variables will be referred to as the signal covariates and the rest will be called noise covariates. The set of columns that correspond to the noise covariates is denoted by $\mathbf{Q} = (\mathbf{Q}_1, \cdots, \mathbf{Q}_n)^T = (\tilde{\mathbf{x}}_{s+1}, \cdots, \tilde{\mathbf{x}}_p)$. We standardize each column of $\mathbf{X}$ to have $L_2$-norm $\sqrt{n}$.

To recover the true model and estimate $\boldsymbol{\beta}^*$, we consider the following regularization problem

$$\min_{\boldsymbol{\beta} \in \mathbf{R}^p} \Big\{ \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + n\lambda_n \sum_{j=1}^p p_{\lambda_n}(|\beta_j|) \Big\}, \tag{2.2}$$

where $\rho_\tau(u) = u(\tau - 1\{u \le 0\})$ is the quantile loss function, and $p_{\lambda_n}(\cdot)$ is a nonnegative penalty function on $[0, \infty)$ with a regularization parameter $\lambda_n \ge 0$. The use of quantile loss function in (2.2) is to overcome the difficulty of heavy tails of the error distribution. Since $P(\varepsilon \le 0) = \tau$, (2.2) can be interpreted as the sparse estimation of the conditional $\tau$th quantile. Regarding the choice of $p_{\lambda_n}(\cdot)$, it was demonstrated in Lv and Fan (2009) and Fan and Lv (2011) that folded-concave penalties are more advantageous for variable selection in high dimensions than the convex ones such as the $L_1$-penalty. It is, however, computationally more challenging to minimize the objective function in (2.2) when $p_\lambda(\cdot)$ is folded-concave. Noting that with a good initial estimate $\widehat{\boldsymbol{\beta}}^{ini} = (\widehat{\beta}_1^{ini}, \cdots, \widehat{\beta}_p^{ini})^T$ of the true coefficient vector, we have

$$p_{\lambda_n}(|\beta_j|) \approx p_{\lambda_n}(|\widehat{\beta}_j^{ini}|) + p'_{\lambda_n}(|\widehat{\beta}_j^{ini}|)\Big(|\beta_j| - |\widehat{\beta}_j^{ini}|\Big).$$

Thus, instead of (2.2) we consider the following weighted $L_1$-regularized quantile regression

$$L_n(\boldsymbol{\beta}) = \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + n\lambda_n\|\mathbf{d} \circ \boldsymbol{\beta}\|_1, \tag{2.3}$$

where $\mathbf{d} = (d_1, \cdots, d_p)^T$ is the vector of non-negative weights, and $\circ$ is the Hadamard product, i.e., the componentwise product of two vectors. This motivates us to define the weighted robust Lasso (WR-Lasso) estimate as the global minimizer of the convex function $L_n(\boldsymbol{\beta})$ for a given non-stochastic weight vector:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L_n(\boldsymbol{\beta}). \tag{2.4}$$

The uniqueness of the global minimizer is easily guaranteed by adding a negligible $L_2$-regularization in implementation. In particular, when $d_j = 1$ for all $j$, the method will be referred to as robust Lasso (R-Lasso).

The adaptive robust Lasso (AR-Lasso) refers specifically to the two-stage procedure in which the stochastic weights $\hat{d}_j = p'_{\lambda_n}(|\widehat{\beta}_j^{ini}|)$ for $j = 1, \cdots, p$ are used in the second step for WR-Lasso and are constructed using a concave penalty $p_{\lambda_n}(\cdot)$ and the initial estimates, $\widehat{\beta}_j^{ini}$, from the first step. In practice, we recommend using R-Lasso as the initial estimate and then using SCAD to compute the weights in AR-Lasso. The asymptotic result of this specific AR-Lasso is summarized in Corollary 1 in Section 5 for the ultra-high dimensional robust regression problem. This is a main contribution of the paper.

**3. Suboptimality of Lasso.** In this section, we use a specific example to illustrate that, in the case of heavy-tailed error distribution, Lasso fails at model selection unless the non-zero coefficients, $\beta_1^*, \cdots, \beta_s^*$, have a very large magnitude. We assume that the errors $\varepsilon_1, \cdots, \varepsilon_n$ have the identical symmetric stable distribution and the characteristic function of $\varepsilon_1$ is given by

$$E[\exp(iu\varepsilon_1)] = \exp(-|u|^\alpha),$$

where $\alpha \in (0, 2)$. By Nolan (2012), $E|\varepsilon_1|^p$ is finite for $0 < p < \alpha$, and $E|\varepsilon_1|^p = \infty$ for $p \geq \alpha$. Furthermore as $z \to \infty$,

$$P(|\varepsilon_1| \geq z) \sim c_\alpha z^{-\alpha},$$

where $c_\alpha = \sin(\frac{\pi\alpha}{2})\Gamma(\alpha)/\pi$ is a constant depending only on $\alpha$, and we use the notation $\sim$ to denote that two terms are equivalent up to some constant. Moreover, for any constant vector $\mathbf{a} = (a_1, \cdots, a_n)^T$, the linear combination $\mathbf{a}^T\boldsymbol{\varepsilon}$ has the following tail behavior

$$P(|\mathbf{a}^T\boldsymbol{\varepsilon}| > z) \sim \|\mathbf{a}\|_\alpha^\alpha c_\alpha z^{-\alpha}, \tag{3.1}$$

with $\|\cdot\|_\alpha$ denoting the $L_\alpha$-norm of a vector.

To demonstrate the suboptimality of Lasso, we consider a simple case in which the design matrix satisfies the conditions that $\mathbf{S}^T\mathbf{Q} = \mathbf{0}$, $\frac{1}{n}\mathbf{S}^T\mathbf{S} = \mathbf{I}_s$, the columns of $\mathbf{Q}$ satisfy $|\text{supp}(\tilde{\mathbf{x}}_j)| = m_n = O(n^{1/2})$ and $\text{supp}(\tilde{\mathbf{x}}_k) \cap \text{supp}(\tilde{\mathbf{x}}_j) = \varnothing$ for any $k \neq j$ and $k, j \in \{s+1, \cdots, p\}$. Here, $m_n$ is a positive integer measuring the sparsity level of the columns of $\mathbf{Q}$. We assume that there are only fixed number of true variables, i.e., $s$ is finite, and that $\max_{ij} |x_{ij}| = O(n^{1/4})$. Thus, it is easy to see that $p = O(n^{1/2})$. In addition, we assume further that all nonzero regression coefficients are the same and $\beta_1^* = \cdots = \beta_s^* = \beta_0 > 0$.

We first consider R-Lasso, which is the global minimizer of (2.4). We will later see in Theorem 2 that by choosing the tuning parameter

$$\lambda_n = O((\log n)^2\sqrt{(\log p)/n}),$$

R-Lasso can recover the true support $\mathcal{M}_* = \{1, \cdots, s\}$ with probability tending to 1. Moreover, the signs of the true regression coefficients can also be recovered with asymptotic probability one as long as the following condition on signal strength is satisfied

$$\lambda_n^{-1}\beta_0 \to \infty, \quad \text{i.e.} \quad (\log n)^{-2}\sqrt{n/(\log p)}\beta_0 \to \infty. \tag{3.2}$$

Now, consider Lasso, which minimizes

$$\tilde{L}_n(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + n\lambda_n\|\boldsymbol{\beta}\|_1. \tag{3.3}$$

We will see that for (3.3) to recover the true model and the correct signs of coefficients, we need a much stronger signal level than that is given in (3.2). By results in optimization theory, the Karush–Kuhn–Tucker (KKT) conditions guaranteeing the necessary and sufficient conditions for $\widetilde{\boldsymbol{\beta}}$ with $\mathcal{M} = \operatorname{supp}(\widetilde{\boldsymbol{\beta}})$ being a minimizer to (3.3) are

$$\begin{gathered}\widetilde{\boldsymbol{\beta}}_{\mathcal{M}} + n\lambda_n(\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}})^{-1}\operatorname{sgn}(\widetilde{\boldsymbol{\beta}}_{\mathcal{M}}) = (\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}})^{-1}\mathbf{X}_{\mathcal{M}}^T\mathbf{y},\\ \|\mathbf{X}_{\mathcal{M}^c}^T(\mathbf{y} - \mathbf{X}_{\mathcal{M}}\widetilde{\boldsymbol{\beta}}_{\mathcal{M}})\|_\infty \le n\lambda_n,\end{gathered}$$

where $\mathcal{M}^c$ is the complement of $\mathcal{M}$, $\boldsymbol{\beta}_{\mathcal{M}}$ is the subvector formed by entries of $\boldsymbol{\beta}$ with indices in $\mathcal{M}$, and $\mathbf{X}_{\mathcal{M}}$ and $\mathbf{X}_{\mathcal{M}^c}$ are the submatrices formed by columns of $\mathbf{X}$ with indices in $\mathcal{M}$ and $\mathcal{M}^c$, respectively. It is easy to see from the above two conditions that for Lasso to enjoy the sign consistency, $\operatorname{sgn}(\widetilde{\boldsymbol{\beta}}) = \operatorname{sgn}(\boldsymbol{\beta}^*)$ with asymptotic probability one, we must have these two conditions satisfied with $\mathcal{M} = \mathcal{M}^*$ with probability tending to 1. Since we have assumed that $\mathbf{Q}^T\mathbf{S} = \mathbf{0}$ and $n^{-1}\mathbf{S}^T\mathbf{S} = \mathbf{I}$, the above sufficient and necessary conditions can also be written as

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{M}^*} + \lambda_n\operatorname{sgn}(\widetilde{\boldsymbol{\beta}}_{\mathcal{M}^*}) = \boldsymbol{\beta}^*_{\mathcal{M}^*} + n^{-1}\mathbf{S}^T\boldsymbol{\varepsilon}, \tag{3.4}$$

$$\|\mathbf{Q}^T\boldsymbol{\varepsilon}\|_\infty \le n\lambda_n, \tag{3.5}$$

Conditions (3.4) and (3.5) are hard for Lasso to hold simultaneously. The following proposition summarizes the necessary condition, whose proof is given in the supplementary material (Fan et al., 2013).

PROPOSITION 1. *In the above model, with probability at least* $1 - e^{-\tilde{c}_0}$, *where* $\tilde{c}_0$ *is some positive constant, Lasso does not have sign consistency, unless the following signal condition is satisfied*

$$n^{\frac{3}{4}-\frac{1}{\alpha}}\beta_0 \to \infty. \tag{3.6}$$

Comparing this with (3.2), it is easy to see that even in this simple case, Lasso needs much stronger signal levels than R-Lasso in order to have a sign consistency in the presence of a heavy-tailed distribution.

**4. Model Selection Oracle Property.** In this section, we establish the model selection oracle property of WR-Lasso. The study enables us to see the bias due to penalization, and that an adaptive weighting scheme is needed in order to eliminate such a bias. We need the following condition on the distribution of noise.

CONDITION 1. *There exist universal constants $c_1 > 0$ and $c_2 > 0$ such that for any $u$ satisfying $|u| \le c_1$, $f_i(u)$'s are uniformly bounded away from 0 and $\infty$ and*

$$|F_i(u) - F_i(0) - uf_i(0)| \le c_2 u^2,$$

*where $f_i(u)$ and $F_i(u)$ are the density function and distribution function of the error $\varepsilon_i$, respectively.*

Condition 1 implies basically that each $f_i(u)$ is Lipschitz around the origin. Commonly used distributions such as the double-exponential distribution and stable distributions including the Cauchy distribution all satisfy this condition.

Denote by $\mathbf{H} = \mathrm{diag}\{f_1(0), \cdots, f_n(0)\}$. The next condition is on the submatrix of $\mathbf{X}$ that corresponds to signal covariates and the magnitude of the entries of $\mathbf{X}$.

CONDITION 2. *The eigenvalues of $\frac{1}{n}\mathbf{S}^T\mathbf{H}\mathbf{S}$ are bounded from below and above by some positive constants $c_0$ and $1/c_0$, respectively. Furthermore,*

$$\kappa_n \equiv \max_{ij} |x_{ij}| = o(\sqrt{n}s^{-1}).$$

Although Condition 2 is on the fixed design matrix, we note that the above condition on $\kappa_n$ is satisfied with asymptotic probability one when the design matrix is generated from some distributions. For instance, if the entries of $\mathbf{X}$ are independent copies from a sub-exponential distribution, the bound on $\kappa_n$ is satisfied with asymptotic probability one as long as $s = o\big(\sqrt{n}/(\log p)\big)$; if the components are generated from sub-Gaussian distribution, then the condition on $\kappa_n$ is satisfied with probability tending to one when $s = o\big(\sqrt{n/(\log p)}\big)$.

4.1. *Oracle Regularized Estimator.* To evaluate our newly proposed method, we first study how well one can do with the assistance of the oracle information on the locations of signal covariates. Then, we use this to establish the asymptotic property of our estimator without the oracle assistance. Denote by $\widehat{\boldsymbol{\beta}}^o = ((\widehat{\boldsymbol{\beta}}_1^o)^T, \mathbf{0}^T)^T$ the oracle regularized estimator (ORE) with $\widehat{\boldsymbol{\beta}}_1^o \in \mathbf{R}^s$ and $\mathbf{0}$ being the vector of all zeros, which minimizes $L_n(\boldsymbol{\beta})$ over the space

$\{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathbf{R}^p : \boldsymbol{\beta}_2 = \mathbf{0} \in \mathbf{R}^{p-s}\}$. The next theorem shows that ORE is consistent, and estimates the correct sign of the true coefficient vector with probability tending to one. We use $\mathbf{d}_0$ to denote the first $s$ elements of $\mathbf{d}$.

THEOREM 1. *Let $\gamma_n = C_1(\sqrt{s(\log n)/n} + \lambda_n \|\mathbf{d}_0\|_2)$ with $C_1 > 0$ a constant. If Conditions 1 and 2 hold and $\lambda_n \|\mathbf{d}_0\|_2 \sqrt{s}\kappa_n \to 0$, then there exists some constant $c > 0$ such that*

$$P\left(\|\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n\right) \ge 1 - n^{-cs}. \tag{4.1}$$

*If in addition $\gamma_n^{-1} \min_{1\le j\le s} |\beta_j^*| \to \infty$, then with probability at least $1 - n^{-cs}$,*

$$\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_1^o) = \operatorname{sgn}(\boldsymbol{\beta}_1^*),$$

*where the above equation should be understood componentwisely.*

As shown in Theorem 1, the consistency rate of $\widehat{\boldsymbol{\beta}}_1^o$ in terms of the vector $L_2$-norm is given by $\gamma_n$. The first component of $\gamma_n$, $C_1\sqrt{s(\log n)/n}$, is the oracle rate within a factor of $\log n$, and the second component $C_1\lambda_n\|\mathbf{d}_0\|_2$ reflects the bias due to penalization. If no prior information is available, one may choose equal weights $\mathbf{d}_0 = (1, 1, \cdots, 1)^T$, which corresponds to R-Lasso. Thus for R-Lasso, with probability at least $1 - n^{-cs}$, it holds that

$$\|\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*\|_2 \le C_1(\sqrt{s(\log n)/n} + \sqrt{s}\lambda_n). \tag{4.2}$$

4.2. *WR-Lasso.* In this section, we show that even without the oracle information, WR-Lasso enjoys the same asymptotic property as in Theorem 1 when the weight vector is appropriately chosen. Since the regularized estimator $\widehat{\boldsymbol{\beta}}$ in (2.4) depends on the full design matrix $\mathbf{X}$, we need to impose the following conditions on the design matrix to control the correlation of columns in $\mathbf{Q}$ and $\mathbf{S}$.

CONDITION 3. *With $\gamma_n$ defined in Theorem 1, it holds that*

$$\left\|\frac{1}{n}\mathbf{Q}^T\mathbf{H}\mathbf{S}\right\|_{2,\infty} < \frac{\lambda_n}{2\|\mathbf{d}_1^{-1}\|_\infty \gamma_n},$$

*where $\|\mathbf{A}\|_{2,\infty} = \sup_{\mathbf{x}\neq 0} \|\mathbf{A}\mathbf{x}\|_\infty / \|\mathbf{x}\|_2$ for a matrix $\mathbf{A}$ and vector $\mathbf{x}$, and $\mathbf{d}_1^{-1} = (d_{s+1}^{-1}, \cdots, d_p^{-1})^T$. Furthermore, $\log(p) = o(n^b)$ for some constant $b \in (0, 1)$.*

To understand the implications of Condition 3, we consider the case of $f_1(0) = \cdots = f_n(0) \equiv f(0)$. In the special case of $\mathbf{Q}^T\mathbf{S} = \mathbf{0}$, Condition 3 is satisfied automatically. In the case of equal correlation, that is, $n^{-1}\mathbf{X}^T\mathbf{X}$ having off-diagonal elements all equal to $\rho$, the above Condition 3 reduces to

$$|\rho| < \frac{\lambda_n}{4f(0)\|\mathbf{d}_1^{-1}\|_\infty \sqrt{s}\gamma_n}.$$

This puts an upper bound on the correlation coefficient $\rho$ for such a dense matrix.

It is well known that for Gaussian errors, the optimal choice of regularization parameter $\lambda_n$ has the order $\sqrt{(\log p)/n}$ (Bickel et al., 2009). The distribution of the model noise with heavy tails demands a larger choice of $\lambda_n$ to filter the noise for R-lasso. When $\lambda_n \geq \sqrt{(\log n)/n}$, $\gamma_n$ given in (4.2) is in the order of $C_1\lambda_n\sqrt{s}$. In this case, Condition 3 reduces to

$$\|n^{-1}\mathbf{Q}^T\mathbf{H}\mathbf{S}\|_{2,\infty} < O(s^{-1/2}). \tag{4.3}$$

For WR-Lasso, if the weights are chosen such that $\|\mathbf{d}_0\|_2 = O(\sqrt{s(\log n)/n}/\lambda_n)$ and $\|\mathbf{d}_1\|_\infty = O(1)$, then $\gamma_n$ is in the order of $C_1\sqrt{s(\log n)/n}$, and correspondingly, Condition 3 becomes

$$\|n^{-1}\mathbf{Q}^T\mathbf{H}\mathbf{S}\|_{2,\infty} < O\big(\lambda_n\sqrt{n/(s(\log n))}\big).$$

This is a more relaxed condition than (4.3), since with heavy-tailed errors, the optimal $\lambda_n$ should be larger than $\sqrt{(\log p)/n}$. In other words, WR-Lasso not only reduces the bias of the estimate, but also allows for stronger correlations among the signal and noise covariates. However, the above choice of weights depends on unknown locations of signals. A data-driven choice will be given in Section 5, in which the resulting AR-Lasso estimator will be studied.

The following theorem shows the model selection oracle property of the WR-Lasso estimator.

THEOREM 2. *Suppose Conditions 1 – 3 hold. In addition, assume that* $\min_{j\geq s+1} d_j > c_3$ *with some constant* $c_3 > 0$,

$$\gamma_n s^{3/2}\kappa_n^2(\log_2 n)^2 = o(n\lambda_n^2), \quad \lambda_n\|\mathbf{d}_0\|_2\kappa_n \max\{\sqrt{s}, \|\mathbf{d}_0\|_2\} \to 0, \tag{4.4}$$

*and* $\lambda_n > 2\sqrt{(1+c)(\log p)/n}$, *where* $\kappa_n$ *is defined in Condition 2,* $\gamma_n$ *is defined in Theorem 1, and* $c$ *is some positive constant. Then, with probability at least* $1 - O(n^{-cs})$, *there exists a global minimizer* $\widehat{\boldsymbol{\beta}} = ((\widehat{\boldsymbol{\beta}}_1^o)^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ *of* $L_n(\boldsymbol{\beta})$ *which satisfies*

*1)* $\widehat{\boldsymbol{\beta}}_2 = 0$*;*
*2)* $\|\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n$*.*

Theorem 2 shows that the WR-Lasso estimator enjoys the same property as ORE with probability tending to one. However, we impose non-adaptive assumptions on the weight vector $\mathbf{d} = (\mathbf{d}_0^T, \mathbf{d}_1^T)^T$. For noise covariates, we assume $\min_{j>s} d_j > c_3$, which implies that each coordinate needs to be penalized. For the signal covariates, we impose (4.4), which requires $\|\mathbf{d}_0\|_2$ to be small.

When studying the nonconvex penalized quantile regression, Wang et al. (2012) assumed that $\kappa_n$ is bounded and the density functions of $\varepsilon_i$'s are uniformly bounded away from 0 and $\infty$ in a small neighborhood of 0. Their assumption on the error distribution is weaker than our Condition 1. We remark that the difference is because we have weaker conditions on $\kappa_n$ and the penalty function (See Condition 2 and (4.4)). In fact, our Condition 1 can be weakened to the same condition as that in Wang et al. (2012) at the cost of imposing stronger assumptions on $\kappa_n$ and the weight vector $\mathbf{d}$.

Belloni and Chernozhukov (2011) and Wang (2012) imposed the restricted eigenvalue assumption of the design matrix and studied the $L_1$-penalized quantile regression and LAD regression, respectively. We impose different conditions on the design matrix and allow flexible shrinkage by choosing $\mathbf{d}$. In addition, our Theorem 2 provides a stronger result than consistency; we establish model selection oracle property of the estimator.

4.3. *Asymptotic Normality.* We now present the asymptotic normality of our estimator. Define $\mathbf{V}_n = (\mathbf{S}^T\mathbf{H}\mathbf{S})^{-1/2}$ and $\mathbf{Z}_n = (\mathbf{Z}_{n1}, \cdots, \mathbf{Z}_{nn})^T = \mathbf{S}\mathbf{V}_n$ with $\mathbf{Z}_{nj} \in \mathbf{R}^s$ for $j = 1, \cdots, n$.

THEOREM 3. *Assume the conditions of Theorem 2 hold, the first and second order derivatives* $f_i'(u)$ *and* $f_i''(u)$ *are uniformly bounded in a small neighborhood of 0 for all* $i = 1, \cdots, n$*, and that* $\|\mathbf{d}_0\|_2 = O(\sqrt{s/n}\ /\lambda_n)$*,* $\max_i \|\mathbf{H}^{1/2}\mathbf{Z}_{ni}\|_2 = o(s^{-7/2}(\log s)^{-1})$*, and* $\sqrt{n/s}\min_{1\le j\le s}|\beta_j^*| \to \infty$*. Then, with probability tending to 1 there exists a global minimizer* $\widehat{\boldsymbol{\beta}} = ((\widehat{\boldsymbol{\beta}}_1^o)^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ *of* $L_n(\boldsymbol{\beta})$ *such that* $\widehat{\boldsymbol{\beta}}_2 = 0$*. Moreover,*

$$\mathbf{c}^T(\mathbf{Z}_n^T\mathbf{Z}_n)^{-1/2}\mathbf{V}_n^{-1}\big[(\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*) + \frac{n\lambda_n}{2}\mathbf{V}_n^2\tilde{\mathbf{d}}_0\big] \xrightarrow{\mathscr{D}} N\Big(0, \tau(1-\tau)\Big),$$

*where* $\mathbf{c}$ *is an arbitrary* $s$*-dimensional vector satisfying* $\mathbf{c}^T\mathbf{c} = 1$*, and* $\tilde{\mathbf{d}}_0$ *is an* $s$*-dimensional vector with the* $j$*th element* $d_j\mathrm{sgn}(\beta_j^*)$*.*

The proof of Theorem 3 is an extension of the proof on the asymptotic normality theorem for the LAD estimator in Pollard (1990), in which the theorem is proved for fixed dimensionality. The idea is to approximate $L_n(\boldsymbol{\beta}_1, \mathbf{0})$ in (2.4) by a sequence of quadratic functions, whose minimizers converge to normal distribution. Since $L_n(\boldsymbol{\beta}_1, \mathbf{0})$ and the quadratic approximation are close, their minimizers are also close, which results in the asymptotic normality in Theorem 3.

Theorem 3 assumes that $\max_i \|\mathbf{H}^{1/2}\mathbf{Z}_{ni}\|_2 = o(s^{-7/2}(\log s)^{-1})$. Since by definition $\sum_{i=1}^n \|\mathbf{H}^{1/2}\mathbf{Z}_{ni}\|_2^2 = s$, it is seen that the condition implies $s = o(n^{1/8})$. This assumption is made to guarantee that the quadratic approximation is close enough to $L_n(\boldsymbol{\beta}_1, \mathbf{0})$. When $s$ is finite, the condition becomes $\max_i \|\mathbf{Z}_{ni}\|_2 = o(1)$, as in Pollard (1990). Another important assumption is $\lambda_n\sqrt{n}\|\mathbf{d}_0\|_2 = O(\sqrt{s})$, which is imposed to make sure that the bias $2^{-1}n\lambda_n\mathbf{c}^T\mathbf{V}_n\tilde{\mathbf{d}}_0$ caused by the penalty term does not diverge. For instance, using R-Lasso will create a non-diminishing bias and thus cannot be guaranteed to have asymptotic normality.

Note that we do not assume a parametric form of the error distribution. Thus, our oracle estimator is in fact a semiparametric estimator with the error density as the nuisance parameter. Although Theorem 3 is proved for fixed design matrix, similar results hold if we assume that $(\mathbf{x}_i, y_i)$ are i.i.d. observations and that conditions of Theorem 3 on design matrix hold with asymptotic probability one. In this case, if additionally $\varepsilon_i$ is independent of $\mathbf{x}_i$ with density function $f_\varepsilon(\varepsilon)$, $s$ is fixed and $E[\mathbf{S}_1\mathbf{S}_1^T]$ is positive definite, then using similar arguments it can be proved that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*)$ is asymptotically normal with covariance matrix $\tau(1-\tau)(f_\varepsilon^2(0)E[\mathbf{S}_1\mathbf{S}_1^T])^{-1}$, which is the semiparametric efficiency bound derived in Newey and Powell (1990) under some additional assumptions of the distributions of $\mathbf{x}_i$ and $\varepsilon_i$. See their paper for more details on these assumptions. Hence, our oracle estimator is semiparametric efficient.

**5. Properties of the Adaptive Robust Lasso.** In previous sections, we have seen that the choice of the weight vector $\mathbf{d}$ plays a pivotal role for the WR-Lasso estimate to enjoy the model selection oracle property and asymptotic normality. In fact, conditions in Theorem 2 require that $\min_{j\geq s+1} d_j > c_3$ and that $\|\mathbf{d}_0\|_2$ does not diverge too fast. Theorem 3 imposes an even more stringent condition, $\|\mathbf{d}_0\|_2 = O(\sqrt{s/n}\ /\lambda_n)$, on the weight vector $\mathbf{d}_0$. For R-Lasso, $\|\mathbf{d}_0\|_2 = \sqrt{s}$ and these conditions become very restrictive. For example, the condition in Theorem 3 becomes $\lambda_n = O(n^{-1/2})$, which is too low for a thresholding level even for Gaussian errors. Hence, an adaptive choice of weights is needed to ensure that those

conditions are satisfied. To this end, we propose a two-step procedure.

In the first step, we use R-Lasso, which gives the estimate $\widehat{\boldsymbol{\beta}}^{ini}$. As has been shown in Belloni and Chernozhukov (2011) and Wang (2012), R-Lasso is consistent at a near-oracle rate $\sqrt{s(\log p)/n}$ and selects the true model $\mathcal{M}^*$ as a submodel (in other words, R-Lasso has the sure screening property using the terminology of Fan and Lv (2008)) with asymptotic probability one, namely,

$$\operatorname{supp}\big(\widehat{\boldsymbol{\beta}}^{ini}\big) \supseteq \operatorname{supp}(\boldsymbol{\beta}^*) \text{ and } \|\widehat{\boldsymbol{\beta}}_1^{ini} - \boldsymbol{\beta}_1^*\|_2 = O(\sqrt{s(\log p)/n}).$$

We remark that our Theorem 2 also ensures the consistency of R-Lasso. Compared to Belloni and Chernozhukov (2011), Theorem 2 presents stronger results but also needs more restrictive conditions for R-Lasso. As will be shown in latter theorems, only the consistency of R-Lasso is needed in the study of AR-Lasso, so we quote the results and conditions on R-Lasso in Belloni and Chernozhukov (2011) with the mind of imposing weaker conditions.

In the second step, we set $\widehat{\mathbf{d}} = (\hat{d}_1, \cdots, \hat{d}_p)^T$ with $\hat{d}_j = p'_{\lambda_n}(|\widehat{\beta}_j^{ini}|)$ where $p_{\lambda_n}(|\cdot|)$ is a folded concave penalty function, and then solve the regularization problem (2.4) with a newly computed weight vector. Thus, vector $\widehat{\mathbf{d}}_0$ is expected to be close to the vector $(p'_{\lambda_n}(|\beta_1^*|), \cdots, p'_{\lambda_n}(|\beta_s^*|))^T$ under $L_2$-norm. If a folded concave penalty such as SCAD is used, then $p'_{\lambda_n}(|\beta_j^*|)$ will be close, or even equal, to zero for $1 \le j \le s$ and thus the magnitude of $\|\widehat{\mathbf{d}}_0\|_2$ is negligible.

Now, we formally establish the asymptotic properties of AR-Lasso. We first present a more general result and then highlight our recommended procedure, which uses R-Lasso as the initial estimate and then uses SCAD to compute the stochastic weights, in Corollary 1. Denote by $\mathbf{d}^* = (d_1^*, \cdots, d_p^*)$ with $d_j^* = p'_{\lambda_n}(|\beta_j^*|)$. Using the weight vector $\widehat{\mathbf{d}}$, AR-Lasso minimizes the following objective function

$$\widehat{L}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + n\lambda_n\|\widehat{\mathbf{d}} \circ \boldsymbol{\beta}\|_1. \tag{5.1}$$

We also need the following conditions to show the model selection oracle property of the two-step procedure.

Condition 4. *With asymptotic probability one, the initial estimate satisfies* $\|\widehat{\boldsymbol{\beta}}^{ini} - \boldsymbol{\beta}^*\|_2 \le C_2\sqrt{s(\log p)/n}$ *with some constant* $C_2 > 0$.

As discussed above, if R-Lasso is used to obtain the initial estimate, it satisfies the above condition. Our second condition is on the penalty function.

CONDITION 5. *$p'_{\lambda_n}(t)$ is non-increasing in $t \in (0, \infty)$ and is Lipschitz with constant $c_5$, that is,*

$$|p'_{\lambda_n}(|\beta_1|) - p'_{\lambda_n}(|\beta_2|)| \le c_5|\beta_1 - \beta_2|,$$

*for any $\beta_1, \beta_2 \in \mathbf{R}$. Moreover, $p'_{\lambda_n}(C_2\sqrt{s(\log p)/n}) > \frac{1}{2}p'_{\lambda_n}(0+)$ for large enough $n$, where $C_2$ is defined in Condition 4.*

For the SCAD (Fan and Li, 2001) penalty, $p'_{\lambda_n}(\beta)$ is given by

$$p'_{\lambda_n}(\beta) = 1\{\beta \le \lambda_n\} + \frac{(a\lambda_n - \beta)_+}{(a-1)\lambda_n} 1\{\beta > \lambda_n\}, \tag{5.2}$$

for a given constant $a > 2$, and it can be easily verified that Condition 5 holds if $\lambda_n > 2(a+1)^{-1}C_2\sqrt{s(\log p)/n}$.

THEOREM 4. *Assume conditions of Theorem 2 hold with $\mathbf{d} = \mathbf{d}^*$ and $\gamma_n = a_n$, where*

$$a_n = C_3\Big(\sqrt{s(\log n)/n} + \lambda_n\big(\|\mathbf{d}_0^*\|_2 + C_2c_5\sqrt{s(\log p)/n}\big)\Big),$$

*with some constant $C_3 > 0$, and $\lambda_n s\kappa_n\sqrt{(\log p)/n} \to 0$. Then, under Conditions 4 and 5, with probability tending to one, there exists a global minimizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ of (5.1) such that $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \le a_n$.*

The results in Theorem 4 are analogous to those in Theorem 2. The extra term $\lambda_n\sqrt{s(\log p)/n}$ in the convergence rate $a_n$, compared to the convergence rate $\gamma_n$ in Theorem 2, is caused by the bias of the initial estimate $\widehat{\boldsymbol{\beta}}^{ini}$. Since the regularization parameter $\lambda_n$ goes to zero, the bias of AR-Lasso is much smaller than that of the initial estimator $\widehat{\boldsymbol{\beta}}^{ini}$. Moreover, the AR-Lasso $\widehat{\boldsymbol{\beta}}$ possesses the model selection oracle property.

Now we present the asymptotic normality of the AR-Lasso estimate.

CONDITION 6. *The smallest signal satisfies $\min_{1\le j\le s}|\beta_j^*| > 2C_2\sqrt{(s\log p)/n}$. Moreover, it holds that $p''_{\lambda_n}(|\beta|) = o(s^{-1}\lambda_n^{-1}(n\log p)^{-1/2})$ for any $|\beta| > 2^{-1}\min_{1\le j\le s}|\beta_j^*|$.*

The above condition on the penalty function is satisfied when the SCAD penalty is used and $\min_{1\le j\le s}|\beta_j^*| \ge 2a\lambda_n$ where $a$ is the parameter in the SCAD penalty (5.2).

THEOREM 5. *Assume conditions of Theorem 3 hold with* $\mathbf{d} = \mathbf{d}^*$ *and* $\gamma_n = a_n$*, where* $a_n$ *is defined in Theorem 4. Then, under Conditions 4 – 6, with asymptotic probability one, there exists a global minimizer* $\widehat{\boldsymbol{\beta}}$ *of (5.1) having the same asymptotic properties as those in Theorem 3.*

With the SCAD penalty, conditions in Theorems 4 and 5 can be simplified and AR-Lasso still enjoys the same asymptotic properties, as presented in the following corollary.

COROLLARY 1. *Assume* $\lambda_n = O(\sqrt{s(\log p)(\log\log n)/n})$, $\log p = o(\sqrt{n})$, $\min_{1\le j\le s}|\beta_j^*| \ge 2a\lambda_n$ *with* $a$ *the parameter in the SCAD penalty, and* $\kappa_n = o(n^{1/4}s^{-1/2}(\log n)^{-3/2}(\log p)^{1/2})$*. Further assume that* $\|n^{-1}\mathbf{Q}^T\mathbf{H}\mathbf{S}\|_{2,\infty} < C_4\sqrt{(\log p)(\log\log n)/\log n}$ *with* $C_4$ *some positive constant. Then, under Conditions 1 and 2, with asymptotic probability one, there exists a global minimizer* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ *of* $\widehat{L}_n(\boldsymbol{\beta})$ *such that*

$$\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \le O(\sqrt{s(\log n)/n}), \quad \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_1) = \operatorname{sgn}(\boldsymbol{\beta}_1^*), \quad \textit{and } \widehat{\boldsymbol{\beta}}_2 = \mathbf{0}.$$

*If in addition,* $\max_i \|\mathbf{H}^{1/2}\mathbf{Z}_{ni}\|_2 = o(s^{-7/2}(\log s)^{-1})$*, then we also have*

$$\mathbf{c}^T(\mathbf{Z}_n^T\mathbf{Z}_n)^{-1/2}\mathbf{V}_n^{-1}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) \xrightarrow{\mathscr{D}} N\left(0, \tau(1-\tau)\right),$$

*where* $\mathbf{c}$ *is an arbitrary* $s$*-dimensional vector satisfying* $\mathbf{c}^T\mathbf{c} = 1$.

Corollary 1 provides sufficient conditions for ensuring the variable selection sign consistency of AR-Lasso. These conditions require that R-Lasso in the initial step has the sure screening property. We remark that in implementation, AR-Lasso is able to select the variables missed by R-Lasso, as demonstrated in our numerical studies in the next section. The theoretical comparison of the variable selection results of R-Lasso and AR-Lasso would be an interesting topic for future study. One set of $(p, n, s, \kappa_n)$ satisfying conditions in Corollary 1 is $\log p = O(n^{b_1})$, $s = o(n^{(1-b_1)/2})$ and $\kappa_n = o(n^{b_1/4}(\log n)^{-3/2})$ with $b_1 \in (0, 1/2)$ some constant. Corollary 1 gives one specific choice of $\lambda_n$, not necessarily the smallest $\lambda_n$, which makes our procedure work. In fact, the condition on $\lambda_n$ can be weakened to $\lambda_n > 2(a+1)^{-1}\|\widehat{\boldsymbol{\beta}}_1^{ini} - \boldsymbol{\beta}_1\|_\infty$. Currently, we use the $L_2$-norm $\|\widehat{\boldsymbol{\beta}}_1^{ini} - \boldsymbol{\beta}_1\|_2$ to bound this $L_\infty$-norm, which is too crude. If one can establish $\|\widehat{\boldsymbol{\beta}}_1^{ini} - \boldsymbol{\beta}_1\|_\infty = O_p(\sqrt{n^{-1}\log p})$ for an initial

estimator $\widehat{\boldsymbol{\beta}}_1^{ini}$, then the choice of $\lambda_n$ can be as small as $O(\sqrt{n^{-1}\log p})$, the same order as that used in Wang (2012). On the other hand, since we are using AR-Lasso, the choice of $\lambda_n$ is not as sensitive as R-Lasso.

**6. Numerical Studies.** In this section we evaluate the finite sample property of our proposed estimator with synthetic data. Please see the supplementary material (Fan et al., 2013) for a real life data set analysis, where we provide results of an eQTL study on the *CHRNA6* gene.

To assess the performance of the proposed estimator and compare it with other methods, we simulated data from the high-dimensional linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_0 + \varepsilon_i, \quad \mathbf{x} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_{\boldsymbol{x}}\right),$$

where the data had $n = 100$ observations and the number of parameters was chosen as $p = 400$. We fixed the true regression coefficient vector as

$$\boldsymbol{\beta}_0 = \{2, 0, 1.5, 0, .80, 0, 0, 1, 0, 1.75, 0, 0, .75, 0, 0, 0.3, 0, \ldots, 0\}.$$

For the distribution of the noise, $\varepsilon$, we considered six symmetric distributions: Normal with variance 2 ($\mathcal{N}(0, 2)$), a scale mixture of Normals for which $\sigma_i^2 = 1$ with probability 0.9 and $\sigma_i^2 = 25$ otherwise ($MN_1$), a different scale mixture model where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and $\sigma_i \sim$ Unif$(1, 5)$ ($MN_2$), Laplace, Student's $t$ with degrees of freedom 4 with doubled variance ($\sqrt{2} \times t_4$) and Cauchy. We take $\tau = 0.5$, corresponding to $L_1$-regression, throughout the simulation. Correlation of the covariates, $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ were either chosen to be identity (i.e. $\boldsymbol{\Sigma}_{\boldsymbol{x}} = I_p$) or they were generated from an AR(1) model with correlation 0.5, that is $\boldsymbol{\Sigma}_{\boldsymbol{x}(i,j)} = 0.5^{|i-j|}$.

We implemented five methods for each setting:

1. $L_2$-*Oracle*, which is the least squares estimator based on the signal covariates.
2. *Lasso*, the penalized least-squares estimator with $L_1$-penalty as in Tibshirani (1996).
3. *SCAD*, the penalized least-squares estimator with SCAD penalty as in Fan and Li (2001).
4. *R-Lasso*, the robust Lasso defined as the minimizer of (2.4) with $\mathbf{d} = \mathbf{1}$.
5. *AR-Lasso*, which is the adaptive robust lasso whose adaptive weights on the penalty function were computed based on the SCAD penalty using the R-Lasso estimate as an initial value.

The tuning parameter, $\lambda_n$, was chosen optimally based on 100 validation data-sets. For each of these data-sets, we ran a grid search to find the best $\lambda_n$ (with the lowest $L_2$ error for $\boldsymbol{\beta}$) for the particular setting. This optimal

$\lambda_n$ was recorded for each of the 100 validation data-sets. The median of these 100 optimal $\lambda_n$ were used in the simulation studies. We preferred this procedure over cross-validation because of the instability of the $L_2$ loss under heavy tails.

The following four performance measures were calculated:

1. $L_2$ loss, which is defined as $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2$.
2. $L_1$ loss, which is defined as $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1$.
3. Number of noise covariates that are included in the model, that is the number of false positives (FP).
4. Number of signal covariates that are not included, i.e. the number of false negatives (FN).

For each setting, we present the average of the performance measure based on 100 simulations. The results are depicted in Tables 1 and 2. A boxplot of the $L_2$ losses under different noise settings is also given in Figure 1 (the $L_2$ loss boxplot for the independent covariate setting is similar and omitted). For the results in Tables 1 and 2, one should compare the performance between Lasso and R-Lasso and that between SCAD and AR-Lasso. This comparison reflects the effectiveness of $L_1$-regression in dealing with heavy-tail distributions. Furthermore, comparing Lasso with SCAD, and R-Lasso with AR-Lasso, shows the effectiveness of using adaptive weights in the penalty function.

Our simulation results reveal the following facts. The quantile based estimators were more robust in dealing with the outliers. For example, for the first mixture model ($MN_1$) and Cauchy, R-Lasso outperformed Lasso, and AR-Lasso outperformed SCAD in all of the four metrics, and significantly so when the error distribution is the Cauchy distribution. On the other hand, for the light-tail distributions such as the normal distribution, the efficiency loss was limited. When the tails get heavier, for instance for the Laplace distribution, quantile based methods started to outperform the least-squares based approaches, more so when the tails got heavier.

The effectiveness of weights in AR-Lasso is self-evident. SCAD outperformed Lasso and AR-Lasso outperformed R-Lasso in almost all of the settings. Furthermore, for all of the error settings AR-Lasso had significantly lower $L_2$ and $L_1$ loss as well as a smaller model size compared to other estimators.

It is seen that when the noise does not have heavy tails, that is for the normal and the Laplace distribution, all the estimators are comparable in terms of $L_1$ loss. As expected, estimators that minimize squared loss worked better than R-Lasso and AR-Lasso estimators under Gaussian noise, but

their performances deteriorated as the tails got heavier. In addition, in the two heteroscedastic settings, AR-Lasso had the best performance among others.

For Cauchy noise, least squares methods could only recover 1 or 2 of the true variables on average. On the other hand, $L_1$-estimators (R-Lasso and AR-Lasso) had very few false negatives, and as evident from $L_2$ loss values, these estimators only missed variables with smaller magnitudes.

In addition, AR-Lasso consistently selected a smaller set of variables than R-Lasso. For instance, for the setting with independent covariates, under the Laplace distribution, R-Lasso and AR-Lasso had on average 34.76 and 18.81 false positives, respectively. Also note that AR-Lasso consistently outperformed R-Lasso: It estimated $\boldsymbol{\beta}^*$ (lower $L_1$ and $L_2$ losses), and the support of $\boldsymbol{\beta}^*$ (lower averages for the number of false positives) more efficiently.

| | | $L_2$ Oracle | Lasso | SCAD | R-Lasso | AR-Lasso |
|---|---|---|---|---|---|---|
| $\mathcal{N}(0,2)$ | $L_2$ loss | 0.833 | 4.114 | 3.412 | 5.342 | 2.662 |
| | $L_1$ loss | 0.380 | 1.047 | 0.819 | 1.169 | 0.785 |
| | FP, FN | - | 27.00, 0.49 | 29.60, 0.51 | 36.81, 0.62 | 17.27, 0.70 |
| $MN_1$ | $L_2$ loss | 0.977 | 5.232 | 4.736 | 4.525 | 2.039 |
| | $L_1$ loss | 0.446 | 1.304 | 1.113 | 1.028 | 0.598 |
| | FP, FN | - | 26.80, 0.73 | 29.29, 0.68 | 34.26, 0.51 | 16.76, 0.51 |
| $MN_2$ | $L_2$ loss | 1.886 | 7.563 | 7.583 | 8.121 | 5.647 |
| | $L_1$ loss | 0.861 | 2.085 | 2.007 | 2.083 | 1.845 |
| | FP, FN | - | 20.39, 2.28 | 23.25, 2.19 | 24.64, 2.29 | 11.97, 2.57 |
| Laplace | $L_2$ loss | 0.795 | 4.056 | 3.395 | 4.610 | 2.025 |
| | $L_1$ loss | 0.366 | 1.016 | 0.799 | 1.039 | 0.573 |
| | FP, FN | - | 26.87, 0.62 | 29.98, 0.49 | 34.76, 0.48 | 18.81, 0.40 |
| $\sqrt{2} \times t_4$ | $L_2$ loss | 1.087 | 5.303 | 5.859 | 6.185 | 3.266 |
| | $L_1$ loss | 0.502 | 1.378 | 1.256 | 1.403 | 0.951 |
| | FP, FN | - | 24.61, 0.85 | 36.95, 0.76 | 33.84, 0.84 | 18.53, 0.82 |
| Cauchy | $L_2$ loss | 37.451 | 211.699 | 266.088 | 6.647 | 3.587 |
| | $L_1$ loss | 17.136 | 30.052 | 40.041 | 1.646 | 1.081 |
| | FP, FN | - | 27.39, 5.78 | 34.32, 5.94 | 27.33, 1.41 | 17.28, 1.10 |

TABLE 1

*Simulation Results with Independent Covariates*

**7. Proofs.** In this section, we prove Theorems 1, 2 and 4 and provide the lemmas used in these proofs. The proofs of Theorems 3 and 5 and Proposition 1 are given in the supplementary Appendix (Fan et al., 2013).

We use techniques from empirical process theory to prove the theoretical results. Let $v_n(\boldsymbol{\beta}) = \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})$. Then $L_n(\boldsymbol{\beta}) = v_n(\boldsymbol{\beta}) + n\lambda_n \sum_{j=1}^p d_j|\beta_j|$. For a given deterministic $M > 0$, define the set

$$\mathcal{B}_0(M) = \{\boldsymbol{\beta} \in \mathbf{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le M, \operatorname{supp}(\boldsymbol{\beta}) \subseteq \operatorname{supp}(\boldsymbol{\beta}^*)\}.$$

| | | $L_2$ Oracle | Lasso | SCAD | R-Lasso | AR-Lasso |
|---|---|---|---|---|---|---|
| $\mathcal{N}(0,2)$ | $L_2$ loss | 0.836 | 3.440 | 3.003 | 4.185 | 2.580 |
| | $L_1$ loss | 0.375 | 0.943 | 0.803 | 1.079 | 0.806 |
| | FP, FN | - | 20.62, 0.59 | 23.13, 0.56 | 22.72, 0.77 | 14.49, 0.74 |
| $MN_1$ | $L_2$ loss | 1.081 | 4.415 | 3.589 | 3.652 | 1.829 |
| | $L_1$ loss | 0.495 | 1.211 | 1.055 | 0.901 | 0.593 |
| | FP, FN | - | 18.66, 0.77 | 15.71, 0.75 | 26.65, 0.60 | 13.29, 0.51 |
| $MN_2$ | $L_2$ loss | 1.858 | 6.427 | 6.249 | 6.882 | 4.890 |
| | $L_1$ loss | 0.844 | 1.899 | 1.876 | 1.916 | 1.785 |
| | FP, FN | - | 15.16, 2.08 | 14.77, 1.96 | 18.22, 1.91 | 7.86, 2.71 |
| Laplace | $L_2$ loss | 0.803 | 3.341 | 2.909 | 3.606 | 1.785 |
| | $L_1$ loss | 0.371 | 0.931 | 0.781 | 0.927 | 0.573 |
| | FP, FN | - | 19.32, 0.62 | 21.60, 0.38 | 24.44, 0.46 | 12.90, 0.55 |
| $\sqrt{2} \times t_4$ | $L_2$ loss | 1.122 | 4.474 | 4.259 | 4.980 | 2.855 |
| | $L_1$ loss | 0.518 | 1.222 | 1.201 | 1.299 | 0.946 |
| | FP, FN | - | 20.00, 0.76 | 18.49, 0.91 | 23.56, 0.79 | 13.40, 1.05 |
| Cauchy | $L_2$ loss | 31.095 | 217.395 | 243.141 | 5.388 | 3.286 |
| | $L_1$ loss | 13.978 | 31.361 | 36.624 | 1.461 | 1.074 |
| | FP, FN | - | 25.59, 5.48 | 32.01, 5.43 | 20.80, 1.16 | 12.45, 1.17 |

TABLE 2

*Simulation Results with Correlated Covariates*

Then, define the function

$$Z_n(M) = \sup_{\boldsymbol{\beta} \in \mathcal{B}_0(M)} \frac{1}{n} \big| (v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}^*)) - E\big(v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}^*)\big) \big|. \tag{7.1}$$

Lemma 1 in Section 7.4 gives the rate of convergence for $Z_n(M)$.

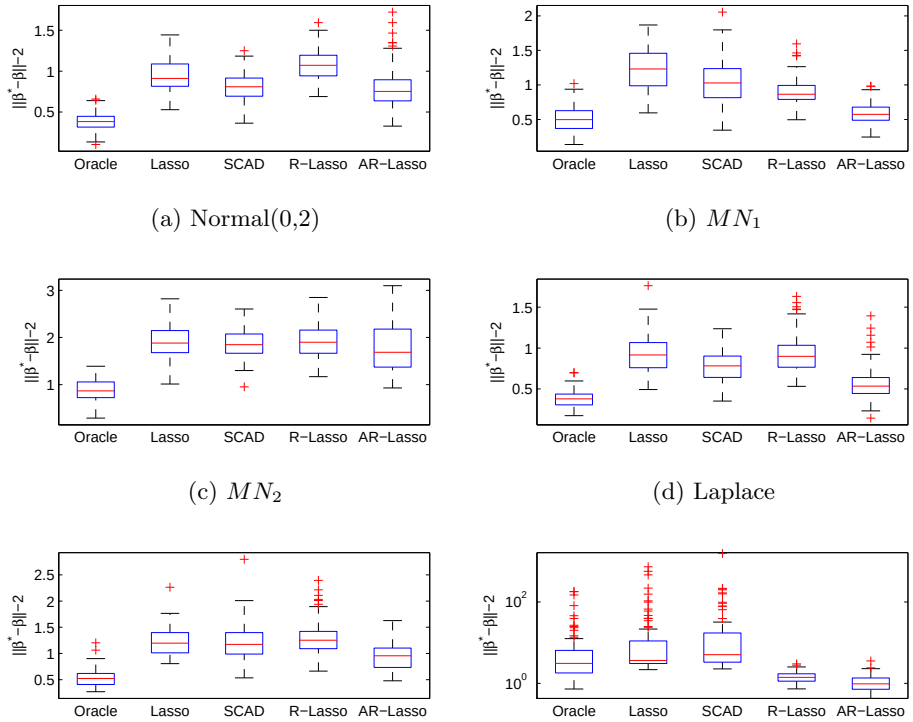
7.1. *Proof of Theorem 1.* We first show that for any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T \in \mathcal{B}_0(M)$ with $M = o(\kappa_n^{-1} s^{-1/2})$,

$$E[v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}^*)] \geq \frac{1}{2} c_0 c n \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2, \tag{7.2}$$

for sufficiently large $n$, where $c$ is the lower bound for $f_i(\cdot)$ in the neighborhood of 0. The intuition follows from the fact that $\boldsymbol{\beta}^*$ is the minimizer of the function $Ev_n(\boldsymbol{\beta})$ and hence in Taylor's expansion of $E[v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}^*)]$ around $\boldsymbol{\beta}^*$, the first order derivative is zero at the point $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. The left-hand side of (7.2) will be controlled by $Z_n(M)$. This yields the $L_2$-rate of convergence in Theorem 1.

To prove (7.2), we set $a_i = |\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)|$. Then, for $\boldsymbol{\beta} \in \mathcal{B}_0(M)$,

$$|a_i| \leq \|\mathbf{S}_i\|_2 \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq \sqrt{s} \kappa_n M \to 0.$$




(a) Normal(0,2) (b) $MN_1$

(c) $MN_2$ (d) Laplace

(e) $\sqrt{2} \times t_4$ (f) Cauchy

Fig 1: Boxplots for $L_2$ Loss with Correlated Covariates

Thus if $\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) > 0$, by $E1\{\varepsilon_i \le 0\} = \tau$, Fubini's theorem, mean value theorem, and Condition 1 it is easy to derive that

(7.3)

$$E\big[\rho_\tau(\varepsilon_i - a_i) - \rho_\tau(\varepsilon_i)\big] = E[a_i(1\{\varepsilon_i \le a_i\} - \tau) - \varepsilon_i 1\{0 \le \varepsilon_i \le a_i\}]$$
$$= E[\int_0^{a_i} 1\{0 \le \varepsilon_i \le s\}ds] = \int_0^{a_i}(F_i(s) - F_i(0))ds = \frac{1}{2}f_i(0)a_i^2 + o(1)a_i^2,$$

where the $o(1)$ is uniformly over all $i = 1, \cdots, n$. When $\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) < 0$, the same result can be obtained. Furthermore, by Condition 2,

$$\sum_{i=1}^n f_i(0)a_i^2 = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)^T\mathbf{S}^T\mathbf{H}\mathbf{S}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \ge c_0 n\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2.$$

This together with (7.3) and the definition of $v_n(\boldsymbol{\beta})$ proves (7.2).

The inequality (7.2) holds for any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T \in \mathcal{B}_0(M)$, yet $\widehat{\boldsymbol{\beta}}^o = ((\widehat{\boldsymbol{\beta}}_1^o)^T, \mathbf{0}^T)^T$ may not be in the set. Thus, we let $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$, where

$$\widetilde{\boldsymbol{\beta}}_1 = u\widehat{\boldsymbol{\beta}}_1^o + (1-u)\boldsymbol{\beta}_1^*, \qquad \text{with} \quad u = M/(M + \|\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*\|_2),$$

which falls in the set $\mathcal{B}_0(M)$. Then, by the convexity and the definition of $\widehat{\boldsymbol{\beta}}_1^o$,

$$L_n(\widetilde{\boldsymbol{\beta}}) \le uL_n(\widehat{\boldsymbol{\beta}}_1^o, \mathbf{0}) + (1-u)L_n(\boldsymbol{\beta}_1^*, \mathbf{0}) \le L_n(\boldsymbol{\beta}_1^*, \mathbf{0}) = L_n(\boldsymbol{\beta}^*).$$

Using this and the triangle inequality we have

$$\begin{aligned} E[v_n(\widetilde{\boldsymbol{\beta}}) - v_n(\boldsymbol{\beta}^*)] &= \{v_n(\boldsymbol{\beta}^*) - Ev_n(\boldsymbol{\beta}^*)\} - \{v_n(\widetilde{\boldsymbol{\beta}}) - Ev_n(\widetilde{\boldsymbol{\beta}})\} \\ &\quad + L_n(\widetilde{\boldsymbol{\beta}}) - L_n(\boldsymbol{\beta}^*) + n\lambda_n\|\mathbf{d}_0 \circ \boldsymbol{\beta}_1^*\|_1 - n\lambda_n\|\mathbf{d}_0 \circ \widetilde{\boldsymbol{\beta}}_1\|_1 \\ &\le nZ_n(M) + n\lambda_n\|\mathbf{d}_0 \circ (\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1)\|_1. \end{aligned} \tag{7.4}$$

By the Cauchy-Schwarz inequality, the very last term is bounded by $n\lambda_n\|\mathbf{d}_0\|_2\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \le n\lambda_n\|\mathbf{d}_0\|_2 M$.

Define the event $\mathcal{E}_n = \{Z_n(M) \le 2Mn^{-1/2}\sqrt{s\log n}\}$. Then by Lemma 1,

$$P(\mathcal{E}_n) \ge 1 - \exp(-c_0 s(\log n)/8). \tag{7.5}$$

On the event $\mathcal{E}_n$, by (7.4), we have

$$E[v_n(\widetilde{\boldsymbol{\beta}}) - v_n(\boldsymbol{\beta}^*)] \le 2M\sqrt{sn(\log n)} + n\lambda_n\|\mathbf{d}_0\|_2 M.$$

Taking $M = 2\sqrt{s/n} + \lambda_n\|\mathbf{d}_0\|_2$. By Condition 2 and the assumption $\lambda_n\|\mathbf{d}_0\|_2\sqrt{s}\kappa_n \to 0$, it is easy to check that $M = o(\kappa_n^{-1}s^{-1/2})$. Combining these two results with (7.2), we obtain that on the event $\mathcal{E}_n$,

$$\frac{1}{2}c_0 n\|\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2^2 \le (2\sqrt{sn(\log n)} + n\lambda_n\|\mathbf{d}_0\|_2)(2\sqrt{s/n} + \lambda_n\|\mathbf{d}_0\|_2),$$

which entails that

$$\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}_1\|_2 \le O(\lambda_n\|\mathbf{d}_0\|_2 + \sqrt{s(\log n)/n}).$$

Note that $\|\boldsymbol{\beta}_1^* - \widetilde{\boldsymbol{\beta}}\|_2 \le M$ implies $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \le 2M$. Thus, on the event $\mathcal{E}_n$,

$$\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \le O(\lambda_n\|\mathbf{d}_0\|_2 + \sqrt{s(\log n)/n}).$$

The second result follows trivially.

7.2. *Proof of Theorem 2.* Since $\widehat{\boldsymbol{\beta}}_1^o$ defined in Theorem 1 is a minimizer of $L_n(\boldsymbol{\beta}_1, \mathbf{0})$, it satisfies the KKT conditions. To prove that $\widehat{\boldsymbol{\beta}} = ((\widehat{\boldsymbol{\beta}}_1^o)^T, \mathbf{0}^T)^T \in \mathbf{R}^p$ is a global minimizer of $L_n(\boldsymbol{\beta})$ in the original $\mathbf{R}^p$ space, we only need to check the following condition

$$\|\mathbf{d}_1^{-1} \circ \mathbf{Q}^T \rho_\tau'(\mathbf{y} - \mathbf{S}\widehat{\boldsymbol{\beta}}_1^o)\|_\infty < n\lambda_n, \tag{7.6}$$

where $\rho_\tau'(\mathbf{u}) = (\rho_\tau'(u_i), \cdots, \rho_\tau'(u_n))^T$ for any $n$-vector $\mathbf{u} = (u_1, \cdots, u_n)^T$ with $\rho_\tau'(u_i) = \tau - 1\{u_i \le 0\}$. Here, $\mathbf{d}_1^{-1}$ denotes the vector $(d_{s+1}^{-1}, \cdots, d_p^{-1})^T$. Then the KKT conditions and the convexity of $L_n(\boldsymbol{\beta})$ together ensure that $\widehat{\boldsymbol{\beta}}$ is a global minimizer of $L(\boldsymbol{\beta})$.

Define events

$$A_1 = \{\|\widehat{\boldsymbol{\beta}}_1^o - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n\} \qquad A_2 = \{\sup_{\boldsymbol{\beta} \in \mathcal{N}} \|\mathbf{d}_1^{-1} \circ \mathbf{Q}^T \rho_\tau'(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}_1)\|_\infty < n\lambda_n\},$$

where $\gamma_n$ is defined in Theorem 1 and

$$\mathcal{N} = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in R^p : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \le \gamma_n,\ \boldsymbol{\beta}_2 = \mathbf{0} \in \mathbf{R}^{p-s}\}.$$

Then by Theorem 1 and Lemma 2 in Section 7.4, $P(A_1 \cap A_2) \ge 1 - o(n^{-cs})$. Since $\widehat{\boldsymbol{\beta}} \in \mathcal{N}$ on the event $A_1$, the inequality (7.6) holds on the event $A_1 \cap A_2$. This completes the proof of Theorem 2.

7.3. *Proof of Theorem 4.* The idea of the proof follows those used in the proof of Theorems 1 and 2. We first consider the minimizer of $\widehat{L}_n(\boldsymbol{\beta})$ in the subspace $\{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathbf{R}^p : \boldsymbol{\beta}_2 = \mathbf{0}\}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0})^T$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^* + \tilde{a}_n \mathbf{v}_1 \in \mathbf{R}^s$ with $\tilde{a}_n = \sqrt{s(\log n)/n} + \lambda_n\big(\|\mathbf{d}_0^*\|_2 + C_2 c_5 \sqrt{s(\log p)/n}\big)$, $\|\mathbf{v}_1\|_2 = C$, and $C > 0$ is some large enough constant. By the assumptions in the theorem we have $\tilde{a}_n = o(\kappa_n^{-1} s^{-1/2})$. Note that

$$\widehat{L}_n(\boldsymbol{\beta}_1^* + \tilde{a}_n \mathbf{v}_1, \mathbf{0}) - \widehat{L}_n(\boldsymbol{\beta}_1^*, \mathbf{0}) = I_1(\mathbf{v}_1) + I_2(\mathbf{v}_1), \tag{7.7}$$

where $I_1(\mathbf{v}_1) = \|\rho_\tau(\mathbf{y} - \mathbf{S}(\boldsymbol{\beta}_1^* + \tilde{a}_n \mathbf{v}_1))\|_1 - \|\rho_\tau(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}_1^*)\|_1$ and $I_2(\mathbf{v}_1) = n\lambda_n(\|\widehat{\mathbf{d}}_0 \circ (\boldsymbol{\beta}_1^* + \tilde{a}_n \mathbf{v}_1)\|_1 - \|\widehat{\mathbf{d}}_0 \circ \boldsymbol{\beta}_1^*\|_1)$ with $\|\rho_\tau(\mathbf{u})\|_1 = \sum_{i=1}^n \rho_\tau(u_i)$ for any vector $\mathbf{u} = (u_1, \cdots, u_n)^T$. By the results in the proof of Theorem 1, $E[I_1(\mathbf{v}_1)] \ge 2^{-1} c_0 n \|\tilde{a}_n \mathbf{v}_1\|_2^2$, and moreover, with probability at least $1 - n^{-cs}$,

$$|I_1(\mathbf{v}_1) - E[I_1(\mathbf{v}_1)]| \le n Z_n(C a_n) \le 2\tilde{a}_n \sqrt{s(\log n)n} \|\mathbf{v}_1\|_2.$$

Thus, by the triangle inequality,

$$I_1(\mathbf{v}_1) \ge 2^{-1} c_0 \tilde{a}_n^2 n \|\mathbf{v}_1\|_2^2 - 2\tilde{a}_n \sqrt{s(\log n)n} \|\mathbf{v}_1\|_2. \tag{7.8}$$

The second term on the right side of (7.7) can be bounded as

$$|I_2(\mathbf{v}_1)| \le n\lambda_n \|\widehat{\mathbf{d}} \circ (\tilde{a}_n \mathbf{v}_1)\|_1 \le n a_n \lambda_n \|\widehat{\mathbf{d}}_0\|_2 \|\mathbf{v}_1\|_2. \tag{7.9}$$

By triangle inequality and Conditions 4 and 5, it holds that

$$\|\widehat{\mathbf{d}}_0\|_2 \le \|\widehat{\mathbf{d}}_0 - \mathbf{d}_0^*\|_2 + \|\mathbf{d}_0^*\|_2 \le c_5 \|\widehat{\boldsymbol{\beta}}_1^{ini} - \boldsymbol{\beta}_1^*\|_2 + \|\mathbf{d}_0^*\|_2 \le C_2 c_5 \sqrt{s(\log p)/n} + \|\mathbf{d}_0^*\|_2. \tag{7.10}$$

Thus, combining (7.7)–(7.10) yields

$$\begin{aligned}\widehat{L}_n(\boldsymbol{\beta}^* + \tilde{a}_n \mathbf{v}_1) - \widehat{L}_n(\boldsymbol{\beta}^*) \ge & 2^{-1} c_0 n a_n^2 \|\mathbf{v}_1\|_2^2 - 2\tilde{a}_n \sqrt{s(\log n)n} \|\mathbf{v}_1\|_2 \\ & - n a_n \lambda_n \big(\|\mathbf{d}_0^*\|_2 + C_2 c_5 \sqrt{s(\log p)/n}\big) \|\mathbf{v}_1\|_2.\end{aligned}$$

Making $\|\mathbf{v}_1\|_2 = C$ large enough, we obtain that with probability tending to one, $\widehat{L}_n(\boldsymbol{\beta}^* + \tilde{a}_n \mathbf{v}) - \widehat{L}_n(\boldsymbol{\beta}^*) > 0$. Then, it follows immediately that with asymptotic probability one, there exists a minimizer $\widehat{\boldsymbol{\beta}}_1$ of $\widehat{L}_n(\boldsymbol{\beta}_1, \mathbf{0})$ such that $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 \le C_3 \tilde{a}_n \equiv a_n$ with some constant $C_3 > 0$.

It remains to prove that with asymptotic probability one,

$$\|\widehat{\mathbf{d}}_1^{-1} \circ \mathbf{Q}^T \rho_\tau'(\mathbf{y} - \mathbf{S}\widehat{\boldsymbol{\beta}}_1)\|_\infty < n\lambda_n. \tag{7.11}$$

Then by KKT conditions, $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$ is a global minimizer of $\widehat{L}_n(\boldsymbol{\beta})$.

Now we proceed to prove (7.11). Since $\beta_j^* = 0$ for all $j = s+1, \cdots, p$, we have that $d_j^* = p_{\lambda_n}'(0+)$. Furthermore, by Condition 4, it holds that $|\widehat{\beta}_j^{ini}| \le C_2 \sqrt{s(\log p)/n}$ with asymptotic probability one. Then, it follows that

$$\min_{j>s} p_{\lambda_n}'(|\widehat{\beta}_j^{ini}|) \ge p_{\lambda_n}'(C_2 \sqrt{s(\log p)/n}).$$

Therefore, by Condition 5 we conclude that

$$\|(\widehat{\mathbf{d}}_1)^{-1}\|_\infty = (\min_{j>s} p_{\lambda_n}'(|\widehat{\beta}_j^{ini}|))^{-1} < 2/p_{\lambda_n}'(0+) = 2\|(\mathbf{d}_1^*)^{-1}\|_\infty. \tag{7.12}$$

From the conditions of Theorem 2 with $\gamma_n = a_n$, it follows from Lemma 2 (inequality (7.20)) that, with probability at least $1 - o(p^{-c})$,

$$\sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \le C_3 a_n} \|\mathbf{Q}^T \rho_\tau'(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}_1)\|_\infty < \frac{n\lambda_n}{2\|(\mathbf{d}_1^*)^{-1}\|_\infty}(1 + o(1)). \tag{7.13}$$

Combining (7.12)–(7.13) and by the triangle inequality, it holds that with asymptotic probability one,

$$\sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \le C_3 a_n} \|(\widehat{\mathbf{d}}_1)^{-1} \circ \mathbf{Q}^T \rho_\tau'(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}_1)\|_\infty < n\lambda_n.$$

Since the minimizer $\widehat{\boldsymbol{\beta}}_1$ satisfies $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 < C_3 a_n$ with asymptotic probability one, the above inequality ensures that (7.11) holds with probability tending to one. This completes the proof.

7.4. *Lemmas.* This subsection contains Lemmas used in proofs of Theorems 1,2 and 4.

LEMMA 1. *Under Condition 2, for any $t > 0$, we have*

$$P(Z_n(M) \geq 4M\sqrt{s/n} + t) \leq \exp\big(-nc_0t^2/(8M^2)\big). \tag{7.14}$$

PROOF. Define $\rho(s, y) = (y-s)(\tau - 1\{y-s \leq 0\})$. Then, $v_n(\boldsymbol{\beta})$ in (7.1) can be rewritten as $v_n(\boldsymbol{\beta}) = \sum_{i=1}^n \rho(\mathbf{x}_i^T\boldsymbol{\beta}, y_i)$. Note that the following Lipschitz condition holds for $\rho(\cdot, y_i)$

$$\big|\rho(s_1, y_i) - \rho(s_2, y_i)\big| \leq \max\{\tau, 1-\tau\}|s_1 - s_2| \leq |s_1 - s_2|. \tag{7.15}$$

Let $W_1, \cdots, W_n$ be a Rademacher sequence, independent of model errors $\varepsilon_1, \cdots, \varepsilon_n$. The Lipschitz inequality (7.15) combined with the symmetrization theorem and Concentration inequality (see, for example, Theorems 14.3 and 14.4 in Bühlmann and van de Geer (2011)) yields that

$$\begin{aligned} E[Z_n(M)] &\leq 2E \sup_{\boldsymbol{\beta}\in\mathcal{B}_0(M)} \Big|\frac{1}{n}\sum_{i=1}^n W_i\big(\rho(\mathbf{x}_i^T\boldsymbol{\beta}, y_i) - \rho(\mathbf{x}_i^T\boldsymbol{\beta}^*, y_i)\big)\Big| \\ &\leq 4E \sup_{\boldsymbol{\beta}\in\mathcal{B}_0(M)} \Big|\frac{1}{n}\sum_{i=1}^n W_i\big(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}_i^T\boldsymbol{\beta}^*\big)\Big|. \end{aligned} \tag{7.16}$$

On the other hand, by the Cauchy-Schwarz inequality

$$\begin{aligned} \Big|\sum_{i=1}^n W_i\big(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}_i^T\boldsymbol{\beta}^*\big)\Big| &= \Big|\sum_{j=1}^s \big(\sum_{i=1}^n W_i x_{ij}\big)(\beta_j - \beta_j^*)\Big| \\ &\leq \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \big\{\sum_{j=1}^s \big|\sum_{i=1}^n W_i x_{ij}\big|^2\big\}^{1/2}. \end{aligned}$$

By Jensen's inequality and concavity of the square root function, $E\left(X^{1/2}\right) \leq (EX)^{1/2}$ for any non-negative random variable $X$. Thus, these two inequalities ensure that the very right hand side of (7.16) can be further bounded by

$$\begin{aligned} &\sup_{\boldsymbol{\beta}\in\mathcal{B}_0(M)} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 E\Big\{\sum\nolimits_{j=1}^s \big|\frac{1}{n}\sum_{i=1}^n W_i x_{ij}\big|^2\Big\}^{1/2} \\ \leq & M\Big\{\sum\nolimits_{j=1}^s E\big|\frac{1}{n}\sum_{i=1}^n W_i x_{ij}\big|^2\Big\}^{1/2} = M\sqrt{s/n}. \end{aligned} \tag{7.17}$$

Therefore, it follows from (7.16) and (7.17) that

$$E[Z_n(M)] \leq 4M\sqrt{s/n}. \tag{7.18}$$

Next since $n^{-1}\mathbf{S}^T\mathbf{S}$ has bounded eigenvalues, for any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \mathbf{0}^T) \in \mathcal{B}_0(M)$,

$$\frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*))^2 = \frac{1}{n}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)^T\mathbf{S}^T\mathbf{S}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*) \leq c_0^{-1}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2^2 \leq c_0^{-1}M^2.$$

Combining this with the Lipschitz inequality (7.15), (7.18), and applying Massart's concentration theorem (see Theorem 14.2 in Bühlmann and van de Geer (2011)) yields that for any $t > 0$,

$$P(Z_n(M) \geq 4M\sqrt{s/n} + t) \leq \exp\big(-nc_0t^2/(8M^2)\big).$$

This proves the Lemma. □

LEMMA 2. *Consider a ball in $R^s$ around $\boldsymbol{\beta}^*$: $\mathcal{N} = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in R^p : \boldsymbol{\beta}_2 = 0,\ \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq \gamma_n\}$ with some sequence $\gamma_n \to 0$. Assume that $\min_{j>s} d_j > c_3$, $\sqrt{1+\gamma_n s^{3/2}\kappa_n^2}\log_2 n = o(\sqrt{n}\ \lambda_n)$, $n^{1/2}\lambda_n(\log p)^{-1/2} \to \infty$, and $\kappa_n\gamma_n^2 = o(\lambda_n)$. Then under Conditions 1-3, there exists some constant $c > 0$ such that*

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{N}} \|\mathbf{d}_1^{-1} \circ \mathbf{Q}^T\rho_\tau'(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}_1)\|_\infty \geq n\lambda_n\Big) \leq o(p^{-c}),$$

*where $\rho_\tau'(u) = \tau - 1\{u \leq 0\}$.*

PROOF. For a fixed $j \in \{s+1, \cdots, p\}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathcal{N}$, define

$$\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i) = x_{ij}\big[\rho_\tau'(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \rho_\tau'(\varepsilon_i) - E[\rho_\tau'(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \rho_\tau'(\varepsilon_i)]\big],$$

where $\mathbf{x}_i^T = (x_{i1}, \cdots, x_{ip})$ is the $i$-th row of the design matrix. The key for the proof is to use the following decomposition

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathcal{N}} \Big\|\frac{1}{n}\mathbf{Q}^T\rho_\tau'(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}_1)\Big\|_\infty &\leq \sup_{\boldsymbol{\beta}\in\mathcal{N}} \Big\|\frac{1}{n}\mathbf{Q}^T E[\rho_\tau'(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}_1) - \rho_\tau'(\boldsymbol{\varepsilon})]\Big\|_\infty \\ &\quad + \Big\|\frac{1}{n}\mathbf{Q}^T\rho_\tau'(\boldsymbol{\varepsilon})\Big\|_\infty + \max_{j>s}\sup_{\boldsymbol{\beta}\in\mathcal{N}} \frac{1}{n}\sum_{i=1}^n \big|\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i)\big|.\end{aligned} \tag{7.19}$$

We will prove that with probability at least $1 - o(p^{-c})$,

$$I_1 \equiv \sup_{\boldsymbol{\beta}\in\mathcal{N}} \Big\| \frac{1}{n}\mathbf{Q}^T E[\rho'_\tau(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}_1) - \rho'_\tau(\boldsymbol{\varepsilon})] \Big\|_\infty < \frac{\lambda_n}{2\|\mathbf{d}_1^{-1}\|_\infty} + o(\lambda_n), \tag{7.20}$$

$$I_2 \equiv n^{-1}\|\mathbf{Q}^T \rho'_\tau(\boldsymbol{\varepsilon})\|_\infty = o(\lambda_n), \tag{7.21}$$

$$I_3 \equiv \max_{j>s} \sup_{\boldsymbol{\beta}\in\mathcal{N}} \Big| \frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i) \Big| = o_p(\lambda_n). \tag{7.22}$$

Combining (7.19)–(7.22) with the assumption $\min_{j>s} d_j > c_3$ completes the proof of the Lemma.

Now we proceed to prove (7.20). Note that $I_1$ can be rewritten as

$$I_1 = \max_{j>s} \sup_{\boldsymbol{\beta}\in\mathcal{N}} \Big| \frac{1}{n}\sum_{i=1}^n x_{ij} E[\rho'_\tau(\varepsilon_i) - \rho'_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})] \Big|. \tag{7.23}$$

By Condition 1,

$$E[\rho'_\tau(\varepsilon_i) - \rho'_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})] = F_i(\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)) - F_i(0) = f_i(0)\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) + \tilde{I}_i,$$

where $F(t)$ is the cumulative distribution function of $\varepsilon_i$, and $\tilde{I}_i = F_i(\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)) - F_i(0) - f_i(0)\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)$. Thus, for any $j > s$,

$$\sum_{i=1}^n x_{ij} E[\rho'_\tau(\varepsilon_i) - \rho'_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})] = \sum_{i=1}^n (f_i(0) x_{ij}\mathbf{S}_i^T)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) + \sum_{i=1}^n x_{ij}\tilde{I}_i.$$

This together with (7.23) and Cauchy-Schwartz inequality entails that

$$I_1 \le \|\frac{1}{n}\mathbf{Q}^T\mathbf{H}\mathbf{S}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\|_\infty + \max_{j>s} |\frac{1}{n}\sum_{i=1}^n x_{ij}\tilde{I}_i|, \tag{7.24}$$

where $\mathbf{H} = \mathrm{diag}\{f_1(0, \cdots, f_n(0))\}$. We consider the two terms on the right hand side of (7.24) one by one. By Condition 3, the first term can be bounded as

$$\|\frac{1}{n}\mathbf{Q}^T\mathbf{H}\mathbf{S}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\|_\infty \le \|\frac{1}{n}\mathbf{Q}^T\mathbf{H}\mathbf{S}\|_{2,\infty}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 < \frac{\lambda_n}{2\|\mathbf{d}_1^{-1}\|_\infty}. \tag{7.25}$$

By Condition 1, $|\tilde{I}_i| \le c(\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*))^2$. This together with Condition 2 ensures that the second term of (7.24) can be bounded as

$$\max_{j>s} \Big| \frac{1}{n}\sum_{i=1}^n x_{ij}\tilde{I}_i \Big| \le \frac{\kappa_n}{n}\sum_{i=1}^n \big|\tilde{I}_1\big| \le C\frac{\kappa_n}{n}\sum_{i=1}^n (\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*))^2 \le C\kappa_n\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2.$$

Since $\boldsymbol{\beta} \in \mathcal{N}$, it follows from the assumption $\lambda_n^{-1}\kappa_n\gamma_n^2 = o(1)$ that

$$\max_{j>s}\Big|\frac{1}{n}\sum_{i=1}^n x_{ij}\tilde{I}_i\Big| \le C\kappa_n\gamma_n^2 = o(\lambda_n).$$

Plugging the above inequality and (7.25) into (7.24) completes the proof of (7.20).

Next we prove (7.21). By Hoeffding's inequality, if $\lambda_n > 2\sqrt{(1+c)(\log p)/n}$ with $c$ is some positive constant, then

$$\begin{aligned} P(\|\mathbf{Q}^T\rho_\tau'(\boldsymbol{\varepsilon})\|_\infty \ge n\lambda_n) &\le \sum_{j=s+1}^p 2\exp\Big(-\frac{n^2\lambda_n^2}{4\sum_{i=1}^n x_{ij}^2}\Big) \\ &= 2\exp\Big(\log(p-s) - n\lambda_n^2/4\Big) \le O(p^{-c}). \end{aligned}$$

Thus, with probability at least $1 - O(p^{-c})$, (7.21) holds.

We now apply Corollary 14.4 in Bühlmann and van de Geer (2011) to prove (7.22). To this end, we need to check conditions of the Corollary. For each fixed $j$, define the functional space $\Gamma_j = \{\gamma_{\boldsymbol{\beta},j} : \boldsymbol{\beta} \in \mathcal{N}\}$. First note that $E[\gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i)] = 0$ for any $\gamma_{\boldsymbol{\beta},j} \in \Gamma_j$. Second, since the $\rho_\tau'$ function is bounded, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\beta},j}^2(\mathbf{x}_i, y_i) = \frac{1}{n}\sum_{i=1}^n x_{ij}^2\big(&\rho_\tau'(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \rho_\tau'(\varepsilon_i) \\ &- E\big(\rho_\tau'(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \rho_\tau(\varepsilon_i)\big)\big)^2 \le 4. \end{aligned}$$

Thus, $\|\gamma_{\boldsymbol{\beta},j}\|_n \equiv \Big(n^{-1}\sum_{i=1}^n \gamma_{\boldsymbol{\beta},j}^2(\mathbf{x}_i, y_i)^2\Big)^{1/2} \le 2$.

Third, we will calculate the covering number of the functional space $\Gamma_j$, $N(\cdot, \Gamma_j, \|\cdot\|_2)$. For any $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathcal{N}$ and $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1^T, \widetilde{\boldsymbol{\beta}}_2^T)^T \in \mathcal{N}$, by Condition 1 and the mean value theorem,

$$\begin{aligned} &E[\rho_\tau'(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \rho_\tau'(\varepsilon_i)] - E[\rho_\tau'(y_i - \mathbf{x}_i^T\widetilde{\boldsymbol{\beta}}) - \rho_\tau'(\varepsilon_i)] \\ &\quad = F_i(\mathbf{S}_i^T(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)) - F_i(\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)) = f_i(a_{1i})\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1), \end{aligned} \tag{7.26}$$

where $F(t)$ is the cumulative distribution function of $\varepsilon_i$, and $a_{1i}$ lies on the segment connecting $\mathbf{S}_i^T(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)$ and $\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)$. Let $\kappa_n = \max_{ij}|x_{ij}|$. Since $f_i(u)$'s are uniformly bounded, by (7.26),

$$\begin{aligned} &\big|x_{ij}E[\rho_\tau'(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \rho_\tau'(\varepsilon_i)] - x_{ij}E[\rho_\tau'(y_i - \mathbf{x}_i^T\widetilde{\boldsymbol{\beta}}) - \rho_\tau'(\varepsilon_i)]\big| \\ &\le C|x_{ij}\mathbf{S}_i^T(\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1)| \le C\|x_{ij}\mathbf{S}_i\|_2\|\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1\|_2 \le C\sqrt{s}\kappa_n^2\|\boldsymbol{\beta}_1 - \widetilde{\boldsymbol{\beta}}_1\|_2, \end{aligned} \tag{7.27}$$

where $C > 0$ is some generic constant. It is known (see, for example, Lemma 14.27 in Bühlmann and van de Geer (2011)) that the ball $\mathcal{N}$ in $\mathbf{R}^s$ can be covered by $(1 + 4\gamma_n/\delta)^s$ balls with radius $\delta$. Since $\rho'_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \rho'_\tau(\varepsilon_i)$ can only take 3 different values $\{-1, 0, 1\}$, it follows from (7.27) that the covering number of $\Gamma_j$ is $N(2^{2-k}, \Gamma_j, \|\cdot\|_2) = 3(1 + C^{-1}2^k\gamma_n s^{1/2}\kappa_n^2)^s$. Thus, by calculus, for any $0 \le k \le (\log_2 n)/2$,

$$\begin{aligned}&\log\big(1 + N(2^{2-k}, \Gamma, \|\cdot\|_2)\big) \le \log(6) + s\log\big(1 + C^{-1}2^k\gamma_n s^{1/2}\kappa_n^2\big)\\ &\le \log(6) + C^{-1}2^k\gamma_n s^{3/2}\kappa_n^2 \le 4\big(1 + C^{-1}\gamma_n s^{3/2}\kappa_n^2\big)2^{2k}.\end{aligned}$$

Hence, conditions of Corollary 14.4 in Bühlmann and van de Geer (2011) are checked and we obtain that for any $t > 0$,

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathcal{N}}\Big|\frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i)\Big| \ge \frac{8}{\sqrt{n}}\big(3\sqrt{1 + C^{-1}\gamma_n s^{3/2}\kappa_n^2}\log_2 n + 4 + 4t\big)\Big) \le 4\exp\Big(-\frac{nt^2}{8}\Big).$$

Taking $t = \sqrt{C(\log p)/n}$ with $C > 0$ large enough constant we obtain that

$$\begin{aligned}P\Big(\max_{j>s}\sup_{\boldsymbol{\beta}\in\mathcal{N}}\big|\frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\beta},j}(\mathbf{x}_i, y_i)\big| \ge \frac{24}{\sqrt{n}}\sqrt{1 + C^{-1}\gamma_n s^{3/2}\kappa_n^2}\log_2 n\Big)\\ \le 4(p - s)\exp\Big(-\frac{C\log p}{8}\Big) \to 0.\end{aligned}$$

Thus if $\sqrt{1 + \gamma_n s^{3/2}\kappa_n^2}\log_2 n = o(\sqrt{n}\ \lambda_n)$, then with probability at least $1 - o(p^{-c})$, (7.22) holds. This completes the proof of the Lemma.

□

**Acknowledgments.** The authors sincerely thank the Editor, Associate Editor, and three referees for their constructive comments that led to substantial improvement of the paper.

## SUPPLEMENTARY MATERIAL

**Supplementary Material for: Adaptive Robust Variable Selection**
(doi: COMPLETED BY THE TYPESETTER; .pdf). Due to space constraints, the proofs of Theorems 3 and 5 and the results of the real life data set study are relegated to the supplement (Fan et al., 2013).

## References.

Belloni, A. and Chernozhukov, V. (2011). $\ell_1$-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130.

Bickel, P. J. and Li, B. (2006). Regularization in statistics (with discussion). *Test* **15** 271–344.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* **37** 1705–1732.

Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. *J. Roy. Statist. Soc. Ser. B* **73** 325–349.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications* New York: Springer.

Candés, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when $p$ is much larger than $n$ (with discussion). *Ann. Statist.* **35** 2313–2351.

Efron, B., Hastie, T. and Tibshirani, R. (2007). Discussion of the "dantzig selector". *Ann. Statist.* **35**(6) 2358–2364.

Fan, J., Fan, Y. and Barut, E. (2013). Supplement to "Adaptive Robust Variable Selection."

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70** 849–911.

Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484.

Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32** 928–961.

Li, Y. and Zhu, J. (2008). L1-norm quantile regression. *J. Comput. Graph. Statist.* **17** 163–185.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528.

Meinshausen, N. and Bühlmann, B. (2010) Stability selection. *J. Roy. Statist. Soc. Ser. B* **72** 417–473.

Newey, W.K. and Powell, J.L. (1990) Efficient Estimation of Linear and Type I Censored Regression Models under Conditional Quantile Restrictions. *Econometric Theory* **3** 295–317.

Nolan, J. P. (2012). Stable Distributions - Models for Heavy Tailed Data. Birkhauser. (In progress, Chapter 1 online at academic2.american.edu/∼jpnolan).

Pollard, D. (1990). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199.

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

van de Geer, S. and Müller, P. (2012). Quasi-likelihood and/or robust estimation in high dimensions. *Stat. Sci.* **27** 469–480.

Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econom. Statist.* **25** 347–355.

Wang, L. (2012). L1 penalized LAD estimator for high dimensional linear regression. *Tentatively accepted by Journal of Multivariate Analysis.*

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultrahigh dimension. *Journal of American Statistical Association.* **107**, 214 – 222.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sin.* **37** 801–817.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty.

*Ann. Statist.* **38** 894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7** 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**(4) 1509–1533.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** 1108–1126.

J. Fan
Department of Operations Research and Fin Eng
Princeton University
Princeton, New Jersey 08544
USA
E-mail: jqfan@princeton.edu

Y. Fan
Data Sciences and Operations Department
Marshall School of Business
University of Southern California
Los Angeles, California 90089
USA
E-mail: fanyingy@marshall.usc.edu

E. Barut
IBM T.J. Watson Research Center
Yorktown Heights, New York 10598
USA
E-mail: abarut@us.ibm.com