



A robust and efficient variable selection method for linear regression

Zhuoran Yang, Liya Fu, You-Gan Wang, Zhixiong Dong & Yunlu Jiang

To cite this article: Zhuoran Yang, Liya Fu, You-Gan Wang, Zhixiong Dong & Yunlu Jiang (2021): A robust and efficient variable selection method for linear regression, Journal of Applied Statistics, DOI: [10.1080/02664763.2021.1962259](https://doi.org/10.1080/02664763.2021.1962259)

To link to this article: <https://doi.org/10.1080/02664763.2021.1962259>



View supplementary material [↗](#)



Published online: 06 Aug 2021.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



A robust and efficient variable selection method for linear regression

Zhuoran Yang^a, Liya Fu^a, You-Gan Wang^b, Zhixiong Dong^a and Yunlu Jiang^c

^aSchool of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, People's Republic of China; ^bSchool of Mathematical Science, Queensland University of Technology, Brisbane, Australia; ^cDepartment of Statistics, College of Economics, Jinan University, Guangzhou, People's Republic of China

ABSTRACT

Variable selection is fundamental to high dimensional statistical modeling, and many approaches have been proposed. However, existing variable selection methods do not perform well in presence of outliers in response variable or/and covariates. In order to ensure a high probability of correct selection and efficient parameter estimation, we investigate a robust variable selection method based on a modified Huber's function with an exponential squared loss tail. We also prove that the proposed method has oracle properties. Furthermore, we carry out simulation studies to evaluate the performance of the proposed method for both $p < n$ and $p > n$. Our simulation results indicate that the proposed method is efficient and robust against outliers and heavy-tailed distributions. Finally, a real dataset from an air pollution mortality study is used to illustrate the proposed method.

ARTICLE HISTORY

Received 24 August 2020
Accepted 26 July 2021

KEYWORDS

Oracle properties; penalty function; robustness; variable selection

2010 MATHEMATICS SUBJECT CLASSIFICATIONS


62J05; 62J07

1. Introduction

In recent years, variable selection has been an important research area in statistical learning, and many approaches have been proposed for selecting variables [5,11,21]. The most popular method is regularization, which can simultaneously select variables and estimate regression parameters by minimizing a penalized loss function. The regularization method has high flexibility and thus plays an important role in the selection of variables.

Tibshirani [11] proposed penalizing the least squares by a least absolute shrinkage and selection operator (lasso). The lasso method can shrink some regression coefficients to exactly zero, but it results in biased parameter estimators. Fan and Li [5] constructed a variable selection method via nonconcave penalized likelihood by a smoothly clipped absolute deviation (SCAD), which has oracle properties. Zou [21] modified the lasso penalty and developed an adaptive lasso shrinkage with oracle properties. Zhang [20] proposed a minimax concave penalty. The methods mentioned above are mainly based on the least squares method. Therefore, these methods are sensitive to outliers and have low efficiency for heavy-tailed distributions. In real data sets, outliers commonly appear in the response

CONTACT Liya Fu  fuliya@mail.xjtu.edu.cn

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2021.1962259>

and/or covariates. Thus, robust variable selection methods are particularly important for data analysis.

To reduce the impact of underlying outliers in the data, a variety of penalized robust methods have been proposed. For example, Wang and Li [14] provided a weighted Wilcoxon-type smoothly clipped absolute deviation method. Wang *et al.* [13] considered penalized least absolute deviation (LAD), which is resistant to heavy-tailed errors or outliers in response. Johnson and Peng [8] and Leng [9] studied variable selection via the rank-based method. Wang *et al.* [16] proposed a class of penalized robust regression estimators based on an exponential squared loss, whose influence function is bounded with respect to the outliers in either the response or the covariate domain. Chang *et al.* [3] used Tukey's biweight criterion and proposed a Tukey-lasso method for variable selection, which is robust against outliers in both the response variable and covariates.

To achieve robustness and efficiency, we propose a penalized regression estimator based on the modified Huber's function with an exponential squared loss tail proposed by Jiang *et al.* [7]. The proposed method has the benefits of the squared loss and the exponential squared loss. Moreover, the proposed method is robust against outliers in response or heavy-tailed errors and also controls the influence of the leverage points in covariates. Furthermore, the proposed method has oracle properties.

The rest of the paper is organized as follows. The new method is presented in Section 2. The details of the choice of tuning parameters and the algorithm are introduced in Section 3. In Section 4, simulation studies are conducted to evaluate the performance of the proposed method. In Section 5, a real dataset is used to illustrate the proposed method. Finally, some conclusions are summarized in Section 6. The proof of the oracle properties is given in the Appendix.

2. Penalized robust regression estimator

Let Y_i be the response variable and \mathbf{X}_i be the corresponding p -dimensional covariates. We consider a linear regression model for (Y_i, \mathbf{X}_i) :

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ unknown parameter vector, and the error terms $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed with an unknown distribution $F(\cdot/\sigma)$ for a scale parameter $\sigma > 0$. Assume ϵ_i has a symmetric distribution with mean zero. We also assume that ϵ_i and \mathbf{X}_i are independent. Note that the first element of \mathbf{X}_i is set to be 1 if the model includes the intercept term.

Wang *et al.* [16] proposed a robust variable selection method with an exponential squared loss plus the adaptive lasso penalty [21]:

$$\sum_{i=1}^n \left[1 - \exp \left\{ -\frac{(Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2}{\gamma_n} \right\} \right] + \sum_{j=1}^p \frac{\log(n)}{|\tilde{\beta}_j|} |\beta_j|,$$

where γ_n is a tuning parameter depending on the sample size n and controlling the degree of robustness, and $(\tilde{\beta}_1, \dots, \tilde{\beta}_p)^T$ is an initial \sqrt{n} -consistent robust estimator of $\boldsymbol{\beta}$.

The modified Huber function proposed by Jiang *et al.* [7] takes the following form,

$$\rho_{\tau,\gamma}(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \leq \tau \\ \frac{\gamma + \tau^2}{2} \left[1 - \frac{\gamma}{\gamma + \tau^2} \exp(-(u^2 - \tau^2)/\gamma) \right] & \text{if } |u| > \tau, \end{cases}$$

which has been shown to perform very well for parameter estimation. The parameters τ and γ in $\rho_{\tau,\gamma}(u)$ control the robustness and efficiency of parameter estimators. τ regulates how many data are treated as outliers, and γ is used to determine how best to adapt these outliers to achieve high efficiency. When γ is fixed and τ equals infinity, $\rho_{\tau,\gamma}(u)$ is corresponding to the squared loss function. When τ equals zero, $\rho_{\tau,\gamma}(u)$ is corresponding to the exponential square loss function.

To achieve a more efficient variable selection, we now extend this loss function by adding a penalty term and propose the following penalized robust objective function:

$$\sum_{i=1}^n \rho_{\tau,\gamma} \left(\frac{Y_i - \mathbf{X}_i^T \boldsymbol{\beta}}{S_n} \right) + n\lambda_n \sum_{j=1}^p \hat{w}_j |\beta_j|, \quad (1)$$

where $S_n > 0$ is an estimator of the scale parameter σ , and \hat{w}_j is $1/|\tilde{\beta}_j|$, and λ_n is a tuning parameter controlling the sparseness of regression parameters $\boldsymbol{\beta}$.

Let $\hat{\boldsymbol{\beta}}_n$ be the resulting estimator from (1). Then, we can prove that $\hat{\boldsymbol{\beta}}_n$ has oracle properties. We first introduce some notations before proving the oracle properties. Let \mathcal{A} be a set of indices of important variables and \mathcal{A}_n be a set of indices of nonzero elements of $\hat{\boldsymbol{\beta}}_n$. Without loss of generality, we assume that the first d elements of the true value $\boldsymbol{\beta}_0$ are nonzero. Additionally, we make some necessary assumptions:

- C1. $\frac{1}{n} \mathbf{X}^T \mathbf{X} \rightarrow \mathbf{C}$, where $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_n^T)^T$, and

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}$$

with a $d \times d$ positive definite matrix \mathbf{C}_{11} .

- C2. $\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_2^2 / \sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$.
- C3. $\sqrt{n}(S_n - \sigma)$ is bounded in probability.

Theorem 2.1: Under conditions C1–C3, further if $\sqrt{n}\lambda_n \rightarrow 0$ and $n\lambda_n \rightarrow \infty$ as $n \rightarrow \infty$, the adaptive lasso estimator with modified Huber's loss satisfies the following properties:

- (1) Asymptotic normality:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{\mathcal{A}} - \boldsymbol{\beta}_0^{\mathcal{A}}) \xrightarrow{d} N \left(\mathbf{0}, \sigma^2 \frac{\mathbb{E} \psi_{\tau,\gamma}^2(\epsilon_i/\sigma)}{(\mathbb{E} \psi'_{\tau,\gamma}(\epsilon_i/\sigma))^2} \mathbf{C}_{11}^{-1} \right),$$

where $\psi_{\tau,\gamma}(\cdot)$ and $\psi'_{\tau,\gamma}(\cdot)$ are the first order and second order derivatives of $\rho_{\tau,\gamma}(\cdot)$, respectively.

- (2) Consistency: $\lim_{n \rightarrow \infty} P(\mathcal{A}_n = \mathcal{A}) = 1$.

The proof of Theorem 2.1 is shown in the Appendix. To obtain the parameter estimator $\hat{\beta}_n$ and select the important variables, we need to specify S_n and the tuning parameters τ , γ , and λ_n . We will give criteria to select accurate tuning parameters in the next section.

3. Selection of tuning parameters and algorithm

3.1. Selection of tuning parameters

Given a consistent estimator $\tilde{\beta}$ of β , the residuals $\{r_i(\tilde{\beta}) = Y_i - X_i^T \tilde{\beta}\}_{i=1}^n$ can be calculated. Then, a normalized median absolute deviation (MAD) estimator proposed by [18] for σ is used,

$$S_n(\tilde{\beta}) = 1.4826 \times \text{median}_i(|r_i(\tilde{\beta}) - \text{median}_j(r_j(\tilde{\beta}))|), \quad (2)$$

which is also used by Jiang *et al.* [7].

For tuning parameters τ and γ , which control the robustness and efficiency of the parameter estimators, we adopt a data-driven method proposed by Jiang *et al.* [7] based on the standardized residuals $\{u_i(\tilde{\beta}) = r_i(\tilde{\beta})/S_n(\tilde{\beta})\}_{i=1}^n$ [17]. For a given set of τ from 0.1 to 5 by 0.2 and γ from 0.5 to 5 by 0.2, we calculate

$$\hat{\text{eff}}(\tau, \gamma) = \frac{[\sum_{i=1}^n (I|u_i| \leq \tau) + \sum_{i=1}^n (I|u_i| > \tau)(1 - 2u_i^2/\gamma) \exp\{-(u_i^2 - \tau^2)/\gamma\}]^2}{n[\sum_{i=1}^n (I|u_i| \leq \tau)u_i^2 + \sum_{i=1}^n (I|u_i| > \tau)u_i^2 \exp\{-2(u_i^2 - \tau^2)/\gamma\}]}, \quad (3)$$

where $u_i = u_i(\tilde{\beta})$. The good choice of (τ, γ) is the pair maximizing $\hat{\text{eff}}(\tau, \gamma)$. Note that $\hat{\text{eff}}(\tau, \gamma)$ is an estimator of $\frac{(\text{E}\psi_{\tau, \gamma}^2(\epsilon_i/\sigma))^2}{\text{E}\psi_{\tau, \gamma}^2(\epsilon_i/\sigma)}$ in the covariance matrix in Theorem 2.1. For more details on justification of Equation (3), see Jiang *et al.* [7].

The tuning parameter λ_n should be appropriately selected because it influences the sparseness of resulting regression estimators. There are several methods to select λ_n , such as cross validation, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). To reduce intensive computation and guarantee consistent variable selection, we propose to select the regularization parameter by minimizing the following objective function, which is inspired by Chen and Chen [4],

$$\log \left\{ \sum_{i=1}^n \rho_{\hat{\tau}, \hat{\gamma}} \left(\frac{Y_i - \mathbf{X}_i^T \hat{\beta}}{S_n} \right) \right\} + \left(\frac{\log(n)}{n} + \frac{2\xi \log(p)}{n} \right) \sum_{j=1}^p I\{\hat{\beta}_j \neq 0\}, \quad (4)$$

where $0 \leq \xi \leq 1$ and $I(\cdot)$ is an indicator function. Throughout the paper, we set $\xi = 0$ when $p \leq n$. When $p > n$, we set $\xi = 0.5$.

3.2. Algorithm

Algorithms for lasso problems can be applied to corresponding adaptive lasso problems. For common lasso problems, a variety of efficient algorithms have been proposed to solve the penalized convex loss functions. For example, Tseng and Yun [12] proposed the block coordinate gradient descent (BGCD) algorithm, and Beck and Teboulle [1] proposed the fast iterative soft thresholding algorithm (FISTA). FISTA can be seen as an extended gradient algorithm. Gradient algorithms can be used to solve nonconvex optimization problem

and perhaps converge to some local minima instead of global minima. Although the proposed loss function in this paper is nonconvex, such algorithms could also perform well if a good initial estimate of β is given, such as the MM estimator [19] for $p \leq n$ and the MM-Ridge estimator [10] for $p > n$. The estimators obtained via the MM or MM-Ridge for the regression coefficients are consistent. Note that the tolerance of convergence and the maximum iteration number of FISTA are 10^{-5} and 10^4 in simulation studies, respectively.

The algorithm for the proposed method via the FISTA is shown as follows:

- Step 1. Obtain an initial estimate $\tilde{\beta}_n$ of β using the MM method ($p \leq n$) or the MM-Ridge method ($p > n$);
- Step 2. Compute $S_n(\tilde{\beta}_n)$ by (2) and the standardized residuals $u_i(\tilde{\beta}_n)$ for $i = 1, \dots, n$;
- Step 3. Select (τ, γ) by maximizing (3);
- Step 4. Choose λ_n using the extended BIC criterion given by (4);
- Step 5. Minimize (1) in terms of β via FISTA;
- Step 6. Repeat steps 2–5 until convergence.

4. Simulation studies

In this section, we carry out simulation studies to evaluate the performance of the proposed method (denoted as HESL). Meanwhile, the exponential square loss method [16] (denoted as ESL) and the Tukey-lasso method [3] (denoted as Biweight) are also included. Note that the MM estimators are calculated using the Matlab function *robustfit*. The data are generated from the following model:

$$Y_i = X_{i1}\beta_1 + \dots + X_{ip}\beta_p + \epsilon_i, \quad i = 1, \dots, n.$$

Two sample sizes are considered, $n = 100$ and 200 . Note that the design matrix is fixed for each case. We conduct a simulation study with 500 independent realizations when $p < n$ and 100 independent realizations when $p > n$ as the training data. In order to evaluate the prediction accuracy, we generate n testing samples, which are independent of the training data and without outliers in the response and covariates based on the same design matrix as that of the training data. The average number of four true zero coefficients that are properly estimated to be zero (CN), the average number of four nonzero coefficients improperly estimated to be zero (IC), and the percentage of correctly fitted models (CF) are reported. In addition, we also report the average, median, and standard error of the mean squared prediction error (MSPE) $n^{-1} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$ over the testing data, denoted by AM, MM, and SDM, respectively. When the error does not follow a normal distribution, the first moment is infinite, and thus we replace MSPE with the median of absolute prediction error (MAPE), $\text{median}\{|\hat{Y}_i - Y_i|, i = 1, \dots, n\}$ over the testing data.

4.1. Simulation studies for $p < n$

In this subsection, we set $p = 8$. The covariate $\mathbf{X}_i = (X_{i1}, \dots, X_{i8})^T$ is sampled from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the (i, j) th element of Σ is $0.5^{|i-j|}$. The regression coefficient vector is $\beta = (0, 1, 1.5, 2, 1, 0, 0, 0)^T$. The random error terms $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed. We consider the following three cases:

Table 1. Simulations results for $p < n$ in Subsection 4.1.

Sample size	Method	CN	IC	CF	AM	MM	SDM
Case 1 $\epsilon_i \sim N(0, 1)$							
100	HESL ₀	4.000	0.138	0.866	1.259	1.247	0.068
	ESL	3.970	0.002	0.968	1.485	1.437	0.207
	Biweight	4.000	0.108	0.902	1.246	1.239	0.054
200	HESL ₀	4.000	0.054	0.948	1.099	1.093	0.027
	ESL	4.000	0.000	1.000	1.203	1.194	0.074
	Biweight	4.000	0.048	0.954	1.097	1.091	0.025
Case 2 $\epsilon_i \sim t(1)$							
100	HESL ₀	3.944	0.080	0.878	1.076	1.064	0.116
	ESL	3.646	0.012	0.688	1.332	1.283	0.242
	Biweight	3.972	0.172	0.830	1.079	1.070	0.106
200	HESL ₀	4.000	0.020	0.982	1.096	1.092	0.048
	ESL	3.998	0.004	0.994	1.167	1.157	0.079
	Biweight	4.000	0.078	0.930	1.099	1.096	0.050
Case 3 $\epsilon_i \sim 0.8N(0, 1) + 0.2t(1)$							
100	HESL ₀	3.736	0.308	0.522	0.792	0.792	0.089
	ESL	3.202	0.028	0.198	0.869	0.868	0.098
	Biweight	3.786	0.380	0.522	0.789	0.792	0.080
200	HESL ₀	3.992	0.040	0.956	0.859	0.816	0.089
	ESL	3.736	0.000	0.744	0.956	0.929	0.130
	Biweight	3.990	0.064	0.930	0.882	0.849	0.096

Case 1: No outliers. $\epsilon_i \sim N(0, 1)$.

Case 2: Heavy-tailed distribution. $\epsilon_i \sim t(1)$.

Case 3: Outliers in both response and covariates. There are some 10%–outliers from $N(10, 1)$ in covariates, and $\epsilon_i \sim 0.8N(0, 1) + 0.2t(1)$.

The simulation results are presented in Table 1. The results show that, for three cases, the number of correctly fitted models increases as the sample size increases, which reveals their consistency. The ESL method tends to select a small number of covariates so that IC of ESL is close to zero, and sometimes CN is smaller than the number of important variables, especially when the sample size is not large enough. When there are no outliers in the data, our method is comparable to Biweight in terms of correctness, and our method and Biweight perform better than ESL in terms of prediction accuracy. Additionally, if there are some outliers in response and/or covariates, whether the sample size is large or small, our method has the highest correctness among these three methods and nearly holds the smallest prediction error. In Case 3, when $n = 100$, Biweight is slightly better than HESL₀ in the light of the CN and AM, that is because HESL₀ has more tuning parameters to estimate and suitable estimates for tuning parameters is difficult when the sample size is relatively small.

4.2. Simulation studies for $p > n$

In this subsection, we set $p = 300$. The covariates $(X_{i1}, \dots, X_{i300})^T$ are sampled from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the (i, j) th element of Σ is $0.5^{|i-j|}$. The regression coefficient vector is $\beta = (0, 1, 1, 1, 1, 1, 0, \dots, 0)^T$. The random error terms $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed. We consider the following three cases:

Table 2. Simulations results for $p > n$ in Subsection 4.2.

Sample size	Method	CN	IC	CF	AM	MM	SDM
Case 1 $\epsilon_i \sim N(0, 1)$							
100	HESL _{0.5}	5.000	0.000	1.000	1.154	1.130	0.090
	HESL _{0.75}	5.000	0.000	1.000	1.156	1.133	0.092
	Biweight	5.000	0.050	0.970	1.151	1.136	0.075
200	HESL _{0.5}	5.000	0.010	0.990	0.810	0.804	0.042
	HESL _{0.75}	5.000	0.010	0.990	0.810	0.804	0.042
	Biweight	5.000	0.170	0.910	0.806	0.802	0.034
Case 2 $\epsilon_i \sim 0.9N(0, 1) + 0.1N(10, 1)$							
100	HESL _{0.5}	4.920	0.000	0.960	1.194	1.029	1.314
	HESL _{0.75}	4.810	0.000	0.940	1.412	1.032	1.921
	Biweight	4.980	0.030	0.960	1.036	1.012	0.139
200	HESL _{0.5}	5.000	0.000	1.000	1.227	1.207	0.068
	HESL _{0.75}	5.000	0.000	1.000	1.227	1.207	0.068
	Biweight	5.000	0.020	0.980	1.218	1.210	0.053
Case 3 $\epsilon_i \sim 0.9N(0, 1) + 0.1N(10, 1)$							
100	HESL _{0.5}	4.960	0.030	0.930	1.049	1.005	0.146
	HESL _{0.75}	4.790	0.020	0.890	1.353	1.012	1.710
	Biweight	4.980	0.100	0.900	1.017	0.991	0.096
200	HESL _{0.5}	5.000	0.000	1.000	1.152	1.143	0.051
	HESL _{0.75}	5.000	0.000	1.000	1.153	1.143	0.051
	Biweight	5.000	0.030	0.970	1.147	1.139	0.043

Case 1: No outliers. $\epsilon_i \sim N(0, 1)$.
Case 2: Outliers only in response. $\epsilon_i \sim 0.9N(0, 1) + 0.1N(10, 1)$.
Case 3: Outliers in both response and covariates. There are some 10%–outliers from $N(10, 1)$ in covariates, and $\epsilon_i \sim 0.9N(0, 1) + 0.1N(10, 1)$.

Note that when p is larger than n , ESL does not work. Hence, we only demonstrate the results of HESL and Biweight. The simulation results are presented in Table 2. Since Wang and Zhu [15] mentioned that the extended BIC for the squared loss with $\xi = 0.5$ might perform worse than that with $0.75 \leq \xi \leq 1$ when $p > n$, we report the results of HESL based on the extended BIC with different values of ξ , 0.5 and 0.75. When n equals to 100, HESL_{0.5} performs better than HESL_{0.75}. If n is large, the performance of two methods is the same. Thus, we recommend the extended BIC with $\xi = 0.5$ when p is larger than n . The conclusion in Table 2 is similar to that in Table 1 when $p < n$. The number of models correctly fitted by HESL is larger than that by Biweight, and HESL is comparable to Biweight in terms of prediction accuracy.

Following the suggestion of the associate editor, we also consider the cases that the data generation models are the same as those in Subsection 4.1 for $p > n$. The results are listed in Table 3. For Case 1 and Case 3, Biweight and HESL are comparable. For Case 2, the distribution is with heavy tails, the CF of Biweight and HESL are very low, which may be because the data is overdispersion, and many data are treated as ‘bad’ data.

4.3. Simulation studies for an equi-correlation structure

Following the advice of the associate editor, we also conduct the simulations for covariates with an equi-correlation structure, that is $R(\rho) = (1 - \rho)I + \rho J$, where I is an identity matrix, and J is a matrix of all ones. We set $\rho = 0.5$, and the results for $p < n$ and $p > n$

Table 3. Simulations results for $p > n$ with the same setting as $p < n$ in Subsection 4.1.

Sample size	Method	CN	IC	CF	AM	MM	SDM
Case 1 $\epsilon_j \sim N(0, 1)$							
100	HESL _{0.5}	4.000	0.000	1.000	0.792	0.785	0.064
	HESL _{0.75}	4.000	0.000	1.000	0.792	0.785	0.064
	Biweight	4.000	0.000	1.000	0.719	0.722	0.043
200	HESL _{0.5}	4.000	0.010	0.990	0.715	0.712	0.032
	HESL _{0.75}	4.000	0.000	1.000	0.717	0.715	0.034
	Biweight	4.000	0.000	1.000	0.720	0.713	0.038
Case 2 $\epsilon_j \sim t(1)$							
100	HESL _{0.5}	2.750	0.030	0.010	1.366	1.241	0.491
	HESL _{0.75}	2.700	0.020	0.010	1.396	1.241	0.537
	Biweight	3.030	0.120	0.030	1.192	1.147	0.232
200	HESL _{0.5}	3.350	0.000	0.350	1.434	1.449	0.116
	HESL _{0.75}	3.250	0.000	0.250	1.450	1.459	0.109
	Biweight	3.630	0.040	0.590	1.385	1.379	0.097
Case 3 $\epsilon_j \sim 0.8N(0, 1) + 0.2t(1)$							
100	HESL _{0.5}	3.480	0.020	0.530	0.995	0.955	0.331
	HESL _{0.75}	3.420	0.000	0.490	1.000	0.964	0.330
	Biweight	3.720	0.060	0.660	0.887	0.870	0.106
200	HESL _{0.5}	4.000	0.010	0.990	0.764	0.759	0.037
	HESL _{0.75}	4.000	0.010	0.990	0.765	0.759	0.038
	Biweight	4.000	0.000	1.000	0.754	0.751	0.035

Table 4. Simulations results for $p < n$ with an equi-correlation structure.

Sample size	Method	CN	IC	CF	AM	MM	SDM
Case 1 $\epsilon_j \sim N(0, 1)$							
100	HESL ₀	4.000	0.186	0.844	0.604	0.600	0.041
	ESL	3.936	0.000	0.936	0.750	0.746	0.101
	Biweight	4.000	0.122	0.892	0.598	0.596	0.038
200	HESL ₀	4.000	0.038	0.968	0.664	0.662	0.022
	ESL	4.000	0.000	1.000	0.687	0.684	0.031
	Biweight	4.000	0.044	0.960	0.663	0.662	0.021
Case 2 $\epsilon_j \sim t(1)$							
100	HESL ₀	3.930	0.078	0.866	0.887	0.854	0.116
	ESL	3.604	0.010	0.652	1.098	1.085	0.225
	Biweight	3.936	0.182	0.790	0.893	0.867	0.113
200	HESL ₀	4.000	0.020	0.982	1.001	0.994	0.045
	ESL	3.996	0.002	0.994	1.036	1.016	0.076
	Biweight	4.000	0.052	0.950	1.007	0.998	0.047
Case 3 $\epsilon_j \sim 0.8N(0, 1) + 0.2t(1)$							
100	HESL ₀	3.604	0.236	0.466	0.784	0.783	0.075
	ESL	3.158	0.018	0.148	0.821	0.823	0.073
	Biweight	3.674	0.314	0.484	0.779	0.777	0.071
200	HESL ₀	3.988	0.090	0.900	0.880	0.857	0.112
	ESL	3.772	0.008	0.776	0.994	1.020	0.156
	Biweight	3.992	0.118	0.880	0.908	0.906	0.122

are presented in Tables 4 and 5, respectively. When $p < n$, the results have the same pattern as those for the AR(1) structure shown in Table 1. However, when $p > n$ and $n = 100$, both HESL and Biweight do not perform well. When $p > n$ and $n = 200$, both HESL and Biweight have improvement. This may be because the covariates have the same relatively strong correlation, and thus the methods are difficult to identify the important covariates when sample size is small.

Table 5. Simulations results for $p > n$ with an equi-correlation structures.

Sample size	Method	CN	IC	CF	AM	MM	SDM
Case 1 $\epsilon_i \sim N(0, 1)$							
100	HESL _{0.5}	4.770	0.660	0.440	1.862	1.226	2.754
	HESL _{0.75}	4.230	0.510	0.440	3.340	1.236	5.002
	Biweight	5.000	2.320	0.240	1.190	1.177	0.141
200	HESL _{0.5}	5.000	0.130	0.870	0.937	0.916	0.069
	HESL _{0.75}	5.000	0.090	0.910	0.938	0.918	0.069
	Biweight	5.000	1.510	0.380	0.919	0.915	0.050
Case 2 $\epsilon_i \sim 0.9N(0, 1) + 0.1N(10, 1)$							
100	HESL _{0.5}	0.640	0.390	0.000	12.940	14.960	4.693
	HESL _{0.75}	0.430	0.210	0.000	13.496	14.960	3.943
	Biweight	4.380	6.560	0.000	1.706	1.424	0.888
200	HESL _{0.5}	4.940	0.650	0.550	1.541	1.338	1.695
	HESL _{0.75}	4.940	0.560	0.580	1.561	1.344	1.694
	Biweight	5.000	0.600	0.640	1.199	1.184	0.096
Case 3 $\epsilon_i \sim 0.9N(0, 1) + 0.1N(10, 1)$							
100	HESL _{0.5}	0.820	0.440	0.000	10.843	13.877	4.752
	HESL _{0.75}	0.460	0.160	0.000	11.797	13.877	4.030
	Biweight	4.360	6.080	0.020	1.673	1.518	0.587
200	HESL _{0.5}	4.944	0.394	0.680	1.465	1.259	1.680
	HESL _{0.75}	4.990	0.320	0.750	1.336	1.289	0.241
	Biweight	5.000	0.440	0.700	1.129	1.101	0.084

5. Air pollution data

In this section, we analyze the air pollution data collected on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's), which was analyzed by Gijbels and Vrinssen [6]. The dataset includes 14 covariates: mean January temperature (JanT), mean July temperature (JulT), relative humidity (RH), annual rainfall (Rain), median education (Edu), population density (PD), percentage of non whites (%NW), percentage of white collar workers (%WC), population (Pop), population per household (P/H), median income (Income), hydrocarbon pollution potential (HP), nitrous oxide pollution potential (NP), and sulfur dioxide pollution potential(SP). The response variable is age-adjusted mortality. Note that we assume that the data are missing completely at random, and thus we remove the 21st observation as a result of its two missing values. There is high skewness of HP, NP, and SP so that a logarithm transformation has been taken for them. Such transformed covariates are denoted by LogHP, LogNP, and LogSP, respectively.

Figures 1 and 2 indicate that there could be some underlying outliers in the response and covariates. Therefore, robust methods are more desirable. Before analyzing the data, we normalize the response and all the covariates by their median and mean absolute deviation estimations. We use the proposed method (HESL), the ESL method, the Biweight method, and the MM method to analyze the data. The selected variables and prediction errors (denoted by PMSE) via the four different methods are shown in Table 6. The proposed HESL method and the Biweight method select the same variables. However, the ESL method selects fewer variables than the HESL and Biweight methods, which is consistent with the conclusion of the simulation studies. The variables selected by both HESL and Biweight are mean January temperature, annual rainfall, percentage of non whites, percentage of white collar workers, hydrocarbon pollution potential, and sulfur dioxide pollution potential. The proposed method and the Biweight method have similar prediction errors.

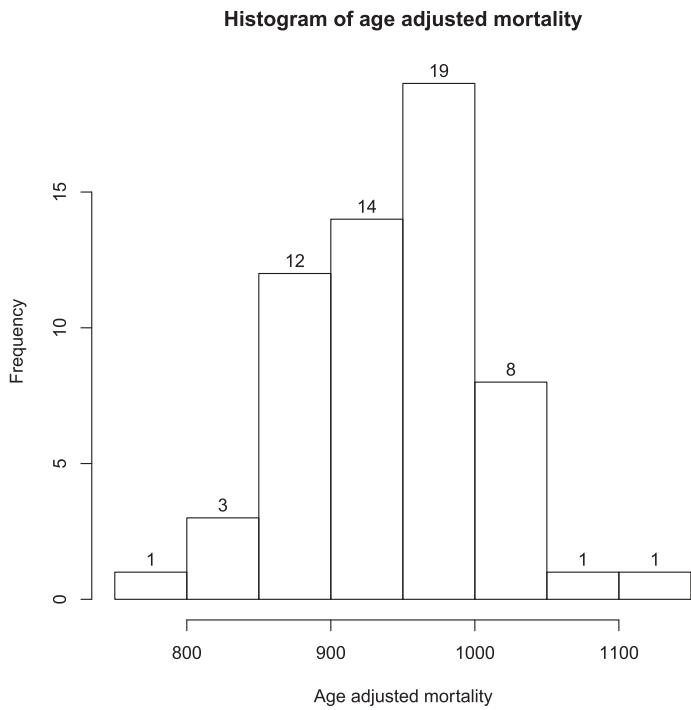


Figure 1. Histogram of age adjusted mortality.

Table 6. Selected covariates of the air pollution data.

	HESL ₀	ESL	Biweight	MM
Intercept	−0.0870		−0.0845	−0.1824
JanT	−0.0470		−0.0566	−0.1295
JulT				−0.0536
RH				0.0237
Rain	0.1616		0.1448	0.1575
Edu				0.0374
PD				0.0945
%NW	0.3931	0.2713	0.4019	0.5062
%WC	−0.1472		−0.1307	−0.1797
Pop				0.0048
P/H				−0.0171
Income				−0.0060
LogHP	−0.0821		−0.1100	−0.3585
LogNP				0.2939
LogSP	0.4080	0.0292	0.4300	0.3067
PMSE	0.0053	0.0080	0.0055	0.0052

MM has the smallest prediction error, because MM selects all the variables and maybe overfits.

6. Conclusions and discussions

In order to ensure robust and highly efficient parameter estimators and simultaneously enhance the correctness, we propose a robust variable selection method based on the

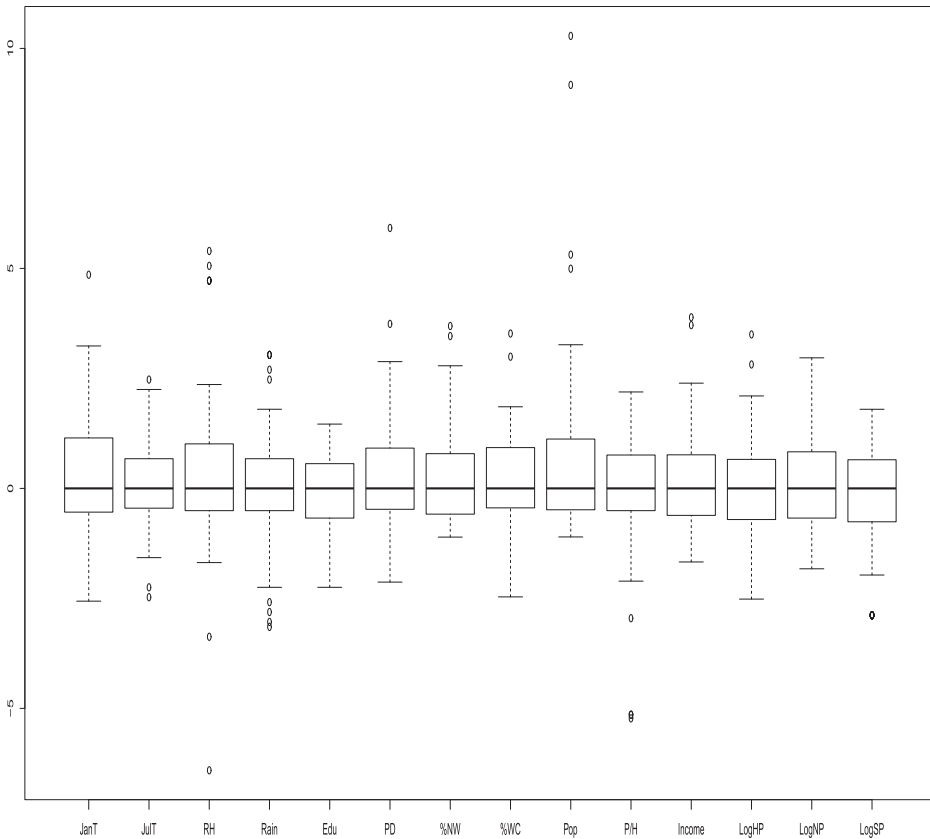


Figure 2. Boxplots of fourteen normalized covariates.

modified Huber's function with a squared exponential loss. The proposed method performs better than the estimators based on the exponential squared loss [16] and the Tukey's Biweight method [3] when there are outliers in the response and/or covariates in terms of correctness. The proposed method is asymptotically normal and has the oracle property. In this paper, we assume that the random errors follow a symmetric distribution with zero mean so that the estimator of the intercept is unbiased. This assumption can be relaxed to an asymmetric distribution, and the estimators of the regression coefficients are still consistent and unbiased except for the intercept [2] we only focus on the linear regression model. The idea could be extended to other models, such as partial linear regression models and varying coefficients models, which will be studied in a future work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the National Natural Science Foundation of China (No. 11871390), Australian Research Council Discovery Project (DP160104292), the Fundamental Research Funds

for the Central Universities (No. xjj2017180), the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2018JQ1006) and the Natural Science Foundation of Guangdong (Nos. 2018A030313171, 2019A1515011830).

References

- [1] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM. J. Imaging. Sci. 2 (2009), pp. 183–202.
- [2] R.J. Carroll, *On estimating variances of robust estimators when the errors are asymmetric*, J. Am. Stat. Assoc. 74 (1979), pp. 674–679.
- [3] L. Chang, S. Roberts, and A. Welsh, *Robust lasso regression using Tukey's biweight criterion*, Technometrics 60 (2018), pp. 36–47.
- [4] J. Chen and Z. Chen, *Extended Bayesian information criteria for model selection with large model spaces*, Biometrika 95 (2008), pp. 759–771.
- [5] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc. 96 (2001), pp. 1348–1360.
- [6] I. Gijbels and I. Vrinssen, *Robust nonnegative garrote variable selection in linear regression*, Comput. Statist. Data Anal. 85 (2015), pp. 1–22.
- [7] Y. Jiang, Y.-G. Wang, L. Fu, and X. Wang, *Robust estimation using modified Huber's functions with new tails*, Technometrics 61 (2018), pp. 111–122.
- [8] B.A. Johnson and L. Peng, *Rank-based variable selection*, J. Nonparametr. Stat. 20 (2008), pp. 241–252.
- [9] C. Leng, *Variable selection and coefficient estimation via regularized rank regression*, Stat. Sin. 20 (2010), pp. 167–181.
- [10] R.A. Maronna, *Robust ridge regression for high-dimensional data*, Technometrics 53 (2011), pp. 44–53.
- [11] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B 58 (1996), pp. 267–288.
- [12] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program. 117 (2009), pp. 387–423.
- [13] H. Wang, G. Li, and G. Jiang, *Robust regression shrinkage and consistent variable selection through the LAD-Lasso*, J. Bus. Econ. Stat. 25 (2007), pp. 347–355.
- [14] L. Wang and R. Li, *Weighted Wilcoxon-type smoothly clipped absolute deviation method*, Biometrics 65 (2009), pp. 564–571.
- [15] T. Wang and L. Zhu, *Consistent tuning parameter selection in high dimensional sparse linear regression*, J. Multivar. Anal. 102 (2011), pp. 1141–1151.
- [16] X. Wang, Y. Jiang, M. Huang, and H. Zhang, *Robust variable selection with exponential squared loss*, J. Am. Stat. Assoc. 108 (2013), pp. 632–643.
- [17] Y.-G. Wang, X. Lin, M. Zhu, and Z. Bai, *Robust estimation using the Huber function with a data-dependent tuning constant*, J. Comput. Graph. Stat. 16 (2007), pp. 468–481.
- [18] A.H. Welsh, *Bahadur representations for robust scale estimators based on regression residuals*, Ann. Statist. 14 (1986), pp. 1246–1251.
- [19] V.J. Yohai, *High breakdown-point and high efficiency robust estimates for regression*, Ann. Statist. 15 (1987), pp. 642–656.
- [20] C. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Ann. Statist. 38 (2010), pp. 894–942.
- [21] H. Zou, *The adaptive lasso and its oracle properties*, J. Am. Stat. Assoc. 101 (2006), pp. 1418–1429.
- [22] C. Geyer, *On the Asymptotics of constrained M-estimation*, Ann. Statist. (1994), 22, pp. 1993–2010.
- [23] K. Knight and W. Fu, *Asymptotics for lasso-type estimators*, Ann. Statist. (2000), 28, pp. 1356–1378.

Appendix. Proof of Theorem 2.1

Proof: The first part is to prove the asymptotic normality. Recall that β_0 and σ are the true values of parameters. Let $\beta = \beta_0 + \frac{\mathbf{v}}{\sqrt{n}}$, $S = \sigma + \frac{s}{\sqrt{n}}$, and

$$\Psi_n(\mathbf{v}, s) = \sum_{i=1}^n \rho_{\tau, \gamma} \left(\frac{Y_i - \mathbf{X}_i^T \left(\beta_0 + \frac{\mathbf{v}}{\sqrt{n}} \right)}{\sigma + \frac{s}{\sqrt{n}}} \right) + n\lambda_n \sum_{j=1}^p \hat{w}_j |\beta_{0j} + \frac{v_j}{\sqrt{n}}|.$$

Let $\hat{\mathbf{v}}^{(n)} = \arg \min \Psi_n(\mathbf{v}, s)$, and then $\hat{\beta}_n = \beta_0 + \frac{\hat{\mathbf{v}}^{(n)}}{\sqrt{n}}$, or $\hat{\mathbf{v}}^{(n)} = \sqrt{n}(\hat{\beta}_n - \beta_0)$. Note that $\Psi_n(\mathbf{v}, s) - \Psi_n(\mathbf{0}, s) = V^{(n)}(\mathbf{v}, s)$, where

$$\begin{aligned} V^{(n)}(\mathbf{v}, s) &= \sum_{i=1}^n \left\{ \rho_{\tau, \gamma} \left(\frac{Y_i - \mathbf{X}_i^T \left(\beta_0 + \frac{\mathbf{v}}{\sqrt{n}} \right)}{\sigma + \frac{s}{\sqrt{n}}} \right) - \rho_{\tau, \gamma} \left(\frac{Y_i - \mathbf{X}_i^T \beta_0}{\sigma + \frac{s}{\sqrt{n}}} \right) \right\} \\ &\quad + \sqrt{n}\lambda_n \sum_{j=1}^p \hat{w}_j \sqrt{n} \left(|\beta_{0j} + \frac{v_j}{\sqrt{n}}| - |\beta_{0j}| \right) \\ &= A^{(n)}(\mathbf{v}, s) + B^{(n)}(\mathbf{v}, s) \end{aligned}$$

Taking a second-order Taylor series expansion leads to

$$\begin{aligned} &\sum_{i=1}^n \rho_{\tau, \gamma} \left(\frac{Y_i - \mathbf{X}_i^T \left(\beta_0 + \frac{\mathbf{v}}{\sqrt{n}} \right)}{\sigma + \frac{s}{\sqrt{n}}} \right) \\ &= \sum_{i=1}^n \left\{ \rho_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) - \frac{1}{\sigma} \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}} - \frac{\epsilon_i}{\sigma^2} \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \frac{s}{\sqrt{n}} \right. \\ &\quad + \frac{1}{2\sigma^2} \psi'_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \left(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}} \right)^2 + \left(\frac{1}{\sigma^2} \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) + \frac{\epsilon_i}{\sigma^3} \psi'_d \left(\frac{\epsilon_i}{\sigma} \right) \right) \frac{\mathbf{X}_i^T \mathbf{v} s}{n} \\ &\quad \left. + \frac{1}{2} \left(\frac{2\epsilon_i}{\sigma^3} \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) + \frac{\epsilon_i^2}{\sigma^4} \psi'_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \right) \left(\frac{s}{\sqrt{n}} \right)^2 + \Delta \left(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}}, \frac{s}{\sqrt{n}} \right) \right\} \end{aligned}$$

with $\Delta \left(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}}, \frac{s}{\sqrt{n}} \right) / \left\| \frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}}, \frac{s}{\sqrt{n}} \right\|^2 \rightarrow 0$, and

$$\begin{aligned} \sum_{i=1}^n \rho_{\tau, \gamma} \left(\frac{Y_i - \mathbf{X}_i^T \beta_0}{\sigma + \frac{s}{\sqrt{n}}} \right) &= \sum_{i=1}^n \left\{ \rho_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) - \frac{\epsilon_i}{\sigma^2} \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \frac{s}{\sqrt{n}} \right. \\ &\quad \left. + \frac{1}{2} \left(\frac{2\epsilon_i}{\sigma^3} \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) + \frac{\epsilon_i^2}{\sigma^4} \psi'_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \right) \left(\frac{s}{\sqrt{n}} \right)^2 + \Delta \left(\frac{s}{\sqrt{n}} \right) \right\} \end{aligned}$$

with $\Delta(\frac{s}{\sqrt{n}})/(\frac{s}{\sqrt{n}})^2 \rightarrow 0$. Then, $A^{(n)}(\mathbf{v}, s)$ can be demonstrated as

$$\begin{aligned} A^{(n)}(\mathbf{v}, s) &= -\frac{1}{\sigma} \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{X}_i^T \right) \mathbf{v} + \frac{1}{2\sigma^2} \mathbf{v}^T \left(\frac{1}{n} \sum_{i=1}^n \psi'_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{X}_i \mathbf{X}_i^T \right) \mathbf{v} \\ &\quad + \frac{1}{\sigma^2} \left(\frac{1}{n} \sum_{i=1}^n \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{X}_i^T + \frac{1}{n} \sum_{i=1}^n \frac{\epsilon_i}{\sigma} \psi'_{\tau, \gamma} \left(\frac{r_i}{\sigma} \right) \mathbf{X}_i^T \right) \mathbf{v} s \\ &\quad + \sum_{i=1}^n \Delta \left(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}}, \frac{s}{\sqrt{n}} \right) - \sum_{i=1}^n \Delta \left(\frac{s}{\sqrt{n}} \right). \end{aligned}$$

It is easy to verify that $E\psi_{\tau, \gamma}(\epsilon_i/\sigma) = 0$ and $\text{Var}(\psi_{\tau, \gamma}(\epsilon_i/\sigma)) = E\psi_{\tau, \gamma}^2(\epsilon_i/\sigma)$, which yield

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \psi_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{X}_i^T \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, E\psi_{\tau, \gamma}^2 \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{C}) = \mathcal{N}(\mathbf{0}, \mathbf{W}).$$

Based on assumption C1 and the fact that $\text{Var}(\psi'_{\tau, \gamma}(\epsilon_i/\sigma))$ is finite, we have

$$\frac{1}{n} \sum_{i=1}^n \psi'_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{X}_i \mathbf{X}_i^T \xrightarrow{p} E\psi'_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{C}.$$

Similarly, $E\psi_{\tau, \gamma}(\epsilon_i/\sigma) = 0$ and the fact that $\text{Var}(\psi_{\tau, \gamma}(\epsilon_i/\sigma))$ is finite yield

$$\frac{1}{n} \sum_{i=1}^n \psi_d \left(\frac{r_i}{\sigma} \right) \mathbf{X}_i^T \xrightarrow{p} \mathbf{0}.$$

Since $\psi'_{\tau, \gamma}(\epsilon_i/\sigma)$ is an even function, then $E[\frac{\epsilon_i}{\sigma} \psi'_{\tau, \gamma}(\frac{\epsilon_i}{\sigma})] = 0$. Also, $E(\frac{\epsilon_i^2}{\sigma^2} \psi_{\tau, \gamma}^2(\frac{\epsilon_i}{\sigma}))$ is finite, and thus

$$\frac{1}{n} \sum_{i=1}^n \frac{\epsilon_i}{\sigma} \psi'_{\tau, \gamma} \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{X}_i^T \xrightarrow{p} \mathbf{0}.$$

Recall that $\Delta(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}}, \frac{s}{\sqrt{n}})/\|\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}}, \frac{s}{\sqrt{n}}\|^2 \rightarrow 0$. Then, by assumption C2, $\forall \xi > 0$, $\exists N(\xi) > 0$, $\forall n \geq N(\xi)$,

$$\sum_{i=1}^n \left| \Delta \left(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}}, \frac{s}{\sqrt{n}} \right) \right| \leq \sum_{i=1}^n \xi \left(\left(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}} \right)^2 + \frac{s^2}{n} \right) = \xi \left(\sum_{i=1}^n \left(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}} \right)^2 + s^2 \right),$$

which means that $\sum_{i=1}^n \Delta(\frac{\mathbf{X}_i^T \mathbf{v}}{\sqrt{n}}, \frac{s}{\sqrt{n}}) \rightarrow 0$ as $n \rightarrow \infty$. Similarly, $\sum_{i=1}^n \Delta(\frac{s}{\sqrt{n}}) \rightarrow 0$ as well. Therefore,

$$A^{(n)}(\mathbf{v}, s) \xrightarrow{d} -\frac{1}{\sigma} \mathbf{v}^T \mathbf{W} + \frac{E\psi'_{\tau, \gamma}(\epsilon_i/\sigma)}{2\sigma^2} \mathbf{v}^T \mathbf{C} \mathbf{v}.$$

In addition, based on the pattern in Zou [21], we analyze the limiting behavior of $B^{(n)}(\mathbf{v}, s)$. Recall that

$$B^{(n)}(\mathbf{v}, s) = \sqrt{n} \lambda_n \sum_{j=1}^p \hat{w}_j \sqrt{n} \left(|\beta_{0j} + \frac{v_j}{\sqrt{n}}| - |\beta_{0j}| \right).$$

Based on the consistency of MM-estimates [19], if $\beta_{0j} \neq 0$, then $\hat{w}_j = 1/|\hat{\beta}_j^{MM}| \xrightarrow{p} 1/|\beta_{0j}|$, and $\sqrt{n}(|\beta_{0j} + \frac{v_j}{\sqrt{n}}| - |\beta_{0j}|) \rightarrow v_j \text{sgn}(\beta_{0j})$. By Slutsky's theorem and the assumption that $\sqrt{n} \lambda_n \rightarrow 0$,

we have

$$\sqrt{n}\lambda_n\hat{w}_j\sqrt{n}\left(|\beta_{0j} + \frac{v_j}{\sqrt{n}}| - |\beta_{0j}|\right) \xrightarrow{p} 0.$$

If $\beta_{0j} = 0$, then $\sqrt{n}(|\beta_{0j} + \frac{v_j}{\sqrt{n}}| - |\beta_{0j}|) = |v_j|$, and

$$\sqrt{n}\lambda_n\hat{w}_j\sqrt{n}\left(|\beta_{0j} + \frac{v_j}{\sqrt{n}}| - |\beta_{0j}|\right) = n\lambda_n(\sqrt{n}\hat{\beta}_j^{MM})^{-1}|v_j| \xrightarrow{p} \infty$$

since $\sqrt{n}\hat{\beta}_j^{MM} = O_p(1)$ and $n\lambda_n \rightarrow \infty$. By Slutsky's theorem, we obtain $V^{(n)}(\mathbf{v}, s) \xrightarrow{d} V(\mathbf{v})$ for every \mathbf{v} , where

$$V(\mathbf{v}) = \begin{cases} -\frac{1}{\sigma}\mathbf{v}_{\mathcal{A}}^T\mathbf{W}_{\mathcal{A}} + \frac{\mathbb{E}\psi'_{\tau,\gamma}(\epsilon_i/\sigma)}{2\sigma^2}\mathbf{v}_{\mathcal{A}}^T\mathbf{C}_{11}\mathbf{v}_{\mathcal{A}} & \text{if } v_j = 0 \forall j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

$V^{(n)}(\mathbf{v}, s)$ is convex, and the unique minimum of $V(\mathbf{v})$ is $(\frac{\sigma}{\mathbb{E}\psi'_{\tau,\gamma}(\epsilon_i/\sigma)}\mathbf{C}_{11}^{-1}\mathbf{W}_{\mathcal{A}}, 0)^T$. Then, following the epi-convergence results of Geyer [22] and Knight and Fu [23], we have

$$\hat{\mathbf{v}}_{\mathcal{A}}^{(n)} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2 \frac{\mathbb{E}\psi_{\tau,\gamma}^2(\epsilon_i/\sigma)}{(\mathbb{E}\psi'_{\tau,\gamma}(\epsilon_i/\sigma))^2} \mathbf{C}_{11}^{-1}\right) \quad \text{and} \quad \hat{\mathbf{v}}_{\mathcal{A}^c}^{(n)} \xrightarrow{d} \mathbf{0}.$$

The asymptotic normality part is finished.

Now, we prove the consistency part. Due to the asymptotic normality above, $\forall j \in \mathcal{A}$, $\hat{\beta}_{nj} \xrightarrow{p} \beta_{0j}$, which means that $P(j \in \mathcal{A}_n) \rightarrow 1$. Then, it suffices to show that $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n) \rightarrow 0$. Consider the event $j' \in \mathcal{A}_n$. Following the KKT optimality conditions, we know that

$$\frac{1}{S_n} \sum_{i=1}^n \psi_{\tau,\gamma} \left(\frac{Y_i - \mathbf{X}_i^T \hat{\beta}_n}{S_n} \right) x_{ij'} = n\lambda_n \hat{w}_{j'}.$$

Note that taking a first-order Taylor expansion,

$$\begin{aligned} \frac{1}{S_n\sqrt{n}} \sum_{i=1}^n \psi_{\tau,\gamma} \left(\frac{Y_i - \mathbf{X}_i^T \hat{\beta}_n}{S_n} \right) x_{ij'} &= \frac{1}{S_n\sqrt{n}} \sum_{i=1}^n \psi_{\tau,\gamma} \left(\frac{Y_i - \mathbf{X}_i^T \left(\beta_0 + \frac{\hat{\mathbf{v}}^{(n)}}{\sqrt{n}} \right)}{\sigma + \frac{\hat{\mathbf{v}}^{(n)}}{\sqrt{n}}} \right) x_{ij'} \\ &= \frac{1}{S_n\sqrt{n}} \sum_{i=1}^n \left\{ \psi_{\tau,\gamma} \left(\frac{\epsilon_i}{\sigma} \right) - \frac{1}{\sigma} \psi'_{\tau,\gamma} \left(\frac{\epsilon_i}{\sigma} \right) \frac{\mathbf{X}_i^T \hat{\mathbf{v}}^{(n)}}{\sqrt{n}} \right. \\ &\quad \left. - \frac{\epsilon_i}{\sigma^2} \psi'_{\tau,\gamma} \left(\frac{\epsilon_i}{\sigma} \right) \frac{\hat{\mathbf{v}}^{(n)}}{\sqrt{n}} + \Delta \left(\frac{\mathbf{X}_i^T \hat{\mathbf{v}}^{(n)}}{\sqrt{n}}, \frac{\hat{\mathbf{v}}^{(n)}}{\sqrt{n}} \right) \right\} x_{ij'}. \end{aligned}$$

Furthermore, we note that

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \psi_{\tau,\gamma} \left(\frac{\epsilon_i}{\sigma} \right) x_{ij'} \right) \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \frac{\mathbb{E}\psi_{\tau,\gamma}^2(\epsilon_i/\sigma)}{n} \sum_{i=1}^n x_{ij'}^2 \right),$$

and $\frac{1}{n} \sum_{i=1}^n \psi'_{\tau,\gamma} \left(\frac{\epsilon_i}{\sigma} \right) \mathbf{X}_i^T \hat{\mathbf{v}}^{(n)} x_{ij'}$ converges in distribution to a normal distribution with bounded variance as $\hat{\mathbf{v}}^{(n)} = \sqrt{n}(\hat{\beta}_n - \beta_0)$. Moreover, similar to the proof of the asymptotic normality,

$\frac{1}{n} \sum_{i=1}^n \frac{\epsilon_i}{\sigma} \psi'_{\tau, \gamma}(\frac{\epsilon_i}{\sigma}) x_{ij'} \rightarrow \mathbf{0}$ and $\sum_{i=1}^n \Delta(\frac{\mathbf{X}_i^T \hat{\mathbf{v}}^{(n)}}{\sqrt{n}}, \frac{\hat{s}^{(n)}}{\sqrt{n}}) \rightarrow \mathbf{0}$. However,

$$\sqrt{n} \lambda_n \hat{w}_{j'} = n \lambda_n (\sqrt{n} \hat{\beta}_{j'}^{MM})^{-1} \xrightarrow{P} \infty.$$

Thus, it is concluded that

$$P(j' \in \mathcal{A}_n) \leq P\left(\frac{1}{S_n} \sum_{i=1}^n \psi_{\tau, \gamma}\left(\frac{Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_n}{S_n}\right) x_{ij'} = \lambda_n \hat{w}_{j'}\right) \rightarrow 0.$$

■