axioms

MDPI

Article

Sparse Robust Weighted Expectile Screening for Ultra-High-Dimensional Data

Xianjun Wu [1,2], Pingping Han [1] and Mingqiu Wang [1,*]

[1] School of Statistics and Data Science, Qufu Normal University, Qufu 273100, China; wuxjun152@163.com (X.W.); hp97tjx@163.com (P.H.)
[2] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China
* Correspondence: mqwang@qfnu.edu.cn

Abstract: This paper investigates robust feature screening for ultra-high dimensional data in the presence of outliers and heterogeneity. Considering the susceptibility of likelihood methods to outliers, we propose a Sparse Robust Weighted Expectile Regression (SRoWER) method that combines the L_2E criterion with expectile regression. By utilizing the IHT algorithm, our method effectively incorporates correlations of covariates and enables joint feature screening. The proposed approach demonstrates robustness against heavy-tailed errors and outliers in data. Simulation studies and a real data analysis are provided to demonstrate the superior performance of the SRoWER method when dealing with outlier-contaminated explanatory variables and/or heavy-tailed error distributions.

Keywords: asymmetric least squares; feature screening; heteroscedasticity; robust regression; ultra-high dimensional data

MSC: 62J05; 62J07; 62F12

1. Introduction

With the exponential growth of data sets in various fields over the past two decades, numerous exhaustive methods have been proposed to address the issue of coefficient sparsity in high-dimensional statistical models, such as bridge regression [1], LASSO [2], SCAD and other folded-concave penalties [3], and the Dantzig selector [4]. While these methods have demonstrated their effectiveness both theoretically and practically, real-world scenarios present new challenges such as identifying disease-causing genes among millions of other genes or analyzing key factors contributing to stock price fluctuations from vast amounts of business data. To tackle ultra-high dimensional data, a range of techniques has emerged. One notable technique is Sure Independent Screening (SIS), initially developed by Fan and Lv [5] for screening out irrelevant factors before conducting variable selection in ultra-high dimensional linear models. There are numerous further developments based on SIS [6–9]. However, these methods overlook the correlations of covariates, despite their computational efficiency. Consequently, additional procedures have been proposed to address this limitation, including ISIS [6], FR [10], SMLE [11].

The aforementioned approaches, which are all based on the maximum likelihood function or Pearson's correlation, become invalid in the presence of outliers. Therefore, robust methods have been extensively studied in the literature. Although quantile regression [12] is effective in handling heterogeneous data, the significantly higher computational cost compared to least squares error necessitates an investigation of the asymmetric least squares (ALS) regression (i.e., expectile regression [13–16]). ALS regression provides a

check for updates

Academic Editor: Hari Mohan Srivastava

Received: 2 April 2025
Revised: 21 April 2025
Accepted: 25 April 2025
Published: 28 April 2025

Citation: Wu, X.; Han, P.; Wang, M. Sparse Robust Weighted Expectile Screening for Ultra-High-Dimensional Data. *Axioms* **2025**, *14*, 340. https://doi.org/10.3390/axioms14050340

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

more comprehensive interpretation of the conditional distribution than the ordinary least squares (OLS) method by allocating different squared error losses to positive and negative residuals, respectively. Moreover, its smooth differentiability greatly reduces computational costs and facilitates theoretical research. Building upon ALS and quantile regression, numerous methods have been proposed to address heterogeneous data with high-dimensionality, such as [17,18] for variable selection and [19–24] for feature screening. The study of [25] proposed an expectile partial correlation screening (EPCS) procedure to sequentially identify important variables for expectile regressions in ultra-high dimensions, and proved that this procedure can lead to a sure screening set. Another robust parametric technique called DPD-SIS [26,27] has been developed for ultra-high-dimensional linear regression models and generalized linear models. This approach is based on the robust minimum density power divergence estimator [28], but it is still limited to addressing marginal aspects without accounting for the correlations between features. In addition, the DPD-SIS cannot handle heterogeneity, which is often a feature of ultra-high-dimensional data.

In the context of heterogeneity and outliers in the data, we propose a new method called Robust Weighted Expectile Regression (RoWER), which combines the L_2 error criterion with expectile regression to achieve the robustness and address the heterogeneity. Furthermore, we developed a sparse restricted RoWER (SRoWER) approach to achieve feature screening. Under general assumptions, we show that the SRoWER can be used for sure screening property. Numerical studies validate the robustness and efficacy of SRoWER. There are three advantages of our SRoWER method, including the following: (1) The SRoWER can provide more reliable screening results, particularly in the presence of outliers in both covariates and the response; (2) In the case of heteroscedasticity, the SRoWER yields superior performance in estimation and feature screening as demonstrated in simulation studies; (3) The SRoWER can be efficiently solved by an iterative hard-thresholding-based algorithm.

The remaining sections of this article are organized as follows. Section 2 introduces the model and the RoWER method. In Section 3, we present the SRoWER method for feature screening and establish sure independent screening property. Section 4 describes simulation studies and a real data analysis that evaluates the finite sample performance of the SRoWER method. Concluding remarks are provided in Section 5. The proofs of the main results can be found in the Appendix A.

2. Model and Method

2.1. L_2E Criterion for Asymmetric Normal Distribution

To address the problem that likelihood methods are sensitive to outliers, Scott [29] proposed a L_2E method, for which the objective function is

$$\int_{-\infty}^{+\infty} f(v|\theta)^2 dv - \frac{2}{n}\sum_{i=1}^{n} f(v_i|\theta),$$

where $f(v|\theta)$ is a parametric probability density function (pdf) of a random variable V and $v_i, i = 1, \ldots, n$ is a given random sample.

Here, we assume that V follows the asymmetric normal distribution, i.e., $V \sim AN(v; \mu_\tau, \sigma^2, \tau)$. The corresponding pdf is

$$f(v; \mu_\tau, \sigma^2, \tau) = \frac{2}{\sqrt{\pi\sigma^2}} \frac{\sqrt{\tau(1-\tau)}}{\sqrt{\tau} + \sqrt{1-\tau}} \exp\left\{-\rho_\tau\left(\frac{v-\mu_\tau}{\sigma}\right)\right\},$$

where $\rho_\tau(t) = \omega_\tau(t)t^2$ is asymmetric squared error loss [13], $\omega_\tau(t) = |\tau - I(t \le 0)|$ with $I(\cdot)$ being the indicator function. Moreover, μ_τ, σ and τ are the location, scale and

asymmetric parameters, respectively. The following proposition gives the L_2E criterion of asymmetric normal distribution.

Proposition 1. *Suppose* $V \sim AN(v; \mu_\tau, \sigma^2, \tau)$, *then the* L_2E *criterion is*

$$\frac{1}{\sqrt{\pi\sigma^2}} \frac{\sqrt{2\tau(1-\tau)}}{\sqrt{\tau}+\sqrt{1-\tau}} \left\{ 1 - 2\sqrt{2} \frac{1}{n} \sum_{i=1}^{n} \exp\left[-\rho_\tau \left(\frac{v_i - \mu_\tau}{\sigma} \right) \right] \right\}. \tag{1}$$

2.2. RoWER

We consider τ-mean [13] of the random variable $Z \in \mathbb{R}$,

$$\mathcal{E}^\tau(Z) = \arg\min_{a \in \mathbb{R}} \mathbb{E}\{\rho_\tau(Z-a)\}, \tau \in (0,1).$$

In fact, the τ-mean corresponds to Efron's w-mean [30], where $w = \tau/(1-\tau)$. In economics, the τ-mean is also called the τ-expectile. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be the n-dimensional response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ be the $n \times p$ design matrix with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T, i = 1, \ldots, n$. The ALS regression is carried out using the following

$$\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}),$$

which degenerates to the OLS regression when $\tau = 0.5$.

Consider the following linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\tau + \boldsymbol{\varepsilon}^\tau, \tag{2}$$

where $\boldsymbol{\beta}^\tau = (\beta_1(\tau), \ldots, \beta_p(\tau))^T$ is a p-dimensional parameter vector and $\boldsymbol{\varepsilon}^\tau = (\varepsilon_1^\tau, \ldots, \varepsilon_n^\tau)^T$ is the vector of n independent errors satisfying $\mathcal{E}^\tau(\varepsilon_i^\tau | \boldsymbol{x}_i) = 0, i = 1, \ldots, n$ for some $\tau \in (0,1)$. We adopt the sparsity assumption on $\boldsymbol{\beta}^\tau$, that is, the regression coefficient vector $\boldsymbol{\beta}^\tau$ has many zero components. In model (2), it is crucial to understand that varying τ allows for variations in the coefficient vector $\boldsymbol{\beta}^\tau$, so we can model different locations of the conditional distribution. For convenience, the superscript τ of $\boldsymbol{\beta}^\tau$ and $\boldsymbol{\varepsilon}^\tau$ is omitted in the following so that no confusion arises.

By substituting $\mu_\tau = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\sigma = 1$ into (1), we can obtain the following loss function by disregarding the terms that are independent of $\boldsymbol{\beta}$,

$$L_n(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \exp\{-\rho_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})\}. \tag{3}$$

However, (3) may not be strictly convex, so we propose a new loss function in the following Proposition 2 by Taylor's expansion and logarithmic transformation.

Proposition 2. *Given a consistent estimator* $\tilde{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$, *minimizing (3) is transformed into minimizing the following loss*

$$D_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \pi_i(\tilde{\boldsymbol{\beta}}) \rho_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{4}$$

where

$$\pi_i(\tilde{\boldsymbol{\beta}}) = \frac{\exp\{-\rho_\tau(y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}})\}}{\sum_{l=1}^{n} \exp\{-\rho_\tau(y_l - \boldsymbol{x}_l^T \tilde{\boldsymbol{\beta}})\}}, \quad i = 1, \ldots, n,$$

which is abbreviated as π_i.

Here π_i's can be treated as the weights of the asymmetric least squares loss, and the loss (4) is referred to as the RoWER. When $\tau = 0.5$, the RoWER degenerates to the weighted least squares regression. This paper chooses the consistent estimator $\tilde{\boldsymbol{\beta}}$ as $\boldsymbol{\beta}_L$ based on Lemma A5. We assume that π_i's are lower bounded.

3. The SRoWER and Sure Screening Property

Let $\mathcal{M}$ be any subset of $\{1, \ldots, p\}$, which corresponds to a submodel with the relevant regression coefficient vector $\boldsymbol{\beta}_{\mathcal{M}} = (\beta_j, j \in \mathcal{M})^T$ and the design matrix $\boldsymbol{X}_{\mathcal{M}} = (\boldsymbol{x}_{1\mathcal{M}}, \ldots, \boldsymbol{x}_{n\mathcal{M}})^T$, $\boldsymbol{x}_{i\mathcal{M}} = (x_{ij}, j \in \mathcal{M})^T$. In addition, let $\|\cdot\|_2$ be the L_2-norm, and $\|\cdot\|_0$ be the L_0-norm, which denotes the number of non-zero components of a vector. The size of model $\mathcal{M}$ is denoted as $s(\mathcal{M})$. The true model is represented by $\mathcal{M}^* = \{j : \beta_j^* \neq 0\}$, with $\boldsymbol{\beta}^*$ being the true regression coefficient vector, and $s(\mathcal{M}^*) = \|\boldsymbol{\beta}^*\|_0 = k_0$.

3.1. The IHT Algorithm

For the objective function $D_n(\boldsymbol{\beta})$, assuming that $\boldsymbol{\beta}$ is sparse with $s(\mathcal{M}^*) = k_0 \leqslant k$ for some known k, the RoWER method with sparsity restriction (SRoWER) yields an estimator of $\boldsymbol{\beta}$ defined as

$$\hat{\boldsymbol{\beta}}_{[k]} = \arg\min_{\boldsymbol{\beta}} D_n(\boldsymbol{\beta}), \quad \|\boldsymbol{\beta}\|_0 \leqslant k, \tag{5}$$

and $\widehat{\mathcal{M}} = \{j : \hat{\beta}_{[k]j} \neq 0\}$ stands for the set of subscripts of the non-zero components of $\hat{\boldsymbol{\beta}}_{[k]}$.

For feature screening, the goal is to retain a relatively small number of features from p features. Currently, many studies have proposed methods to solve such problems. For example, Mallat and Zhang [31] proposed the matching pursuit algorithm. Moreover, the hard thresholding method proposed by Blumensath and Davies [32] is particularly effective for linear models. We now follow the idea of an iterative hard thresholding (IHT) algorithm to compute the SRoWER estimate. For $\boldsymbol{\gamma}$ within the neighborhood of a given $\boldsymbol{\beta}$, the IHT uses the approximation of $D_n(\cdot)$,

$$Q_n(\boldsymbol{\gamma}; \boldsymbol{\beta}) = D_n(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \boldsymbol{T}_n(\boldsymbol{\beta}) + (u/2)\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2, \tag{6}$$

where

$$\boldsymbol{T}_n(\boldsymbol{\beta}) = \frac{\partial D_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{n} 2\pi_i \omega_{\tau,i}(\boldsymbol{\beta})(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})\boldsymbol{x}_i,$$

$\omega_{\tau,i}(\boldsymbol{\beta}) = |\tau - I(y_i \leq \boldsymbol{x}_i^T \boldsymbol{\beta})|$, and $u > 0$ is a scale parameter. Denote $\boldsymbol{T}_n(\boldsymbol{\beta}) = (T_{nj}(\boldsymbol{\beta}), j = 1, \ldots, p)^T$.

By (6), the approximate solution of (5) can be obtained by the following iterative procedure

$$\boldsymbol{\beta}^{(t+1)} = \arg\min_{\boldsymbol{\gamma}} Q_n(\boldsymbol{\gamma}; \boldsymbol{\beta}^{(t)}), \quad \|\boldsymbol{\gamma}\|_0 \leq k. \tag{7}$$

The optimization of (7) is equivalent to

$$\boldsymbol{\beta}^{(t+1)} = \arg\min_{\boldsymbol{\gamma}} \|\boldsymbol{\gamma} - u^{-1}\{u\boldsymbol{\beta}^{(t)} - \boldsymbol{T}_n(\boldsymbol{\beta}^{(t)})\}\|_2^2, \quad \|\boldsymbol{\gamma}\|_0 \leq k. \tag{8}$$

If there is no constraint $\|\boldsymbol{\gamma}\|_0 \leq k$, the analytic solution of (8) is $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta}^{(t)} - u^{-1}\boldsymbol{T}_n(\boldsymbol{\beta}^{(t)})$. However, due to the sparsity restriction, $\boldsymbol{\beta}^{(t+1)}$ can be obtained by selecting the component of $\tilde{\boldsymbol{\gamma}}$ with the largest absolute value before k, i.e.,

$$\boldsymbol{V}(\tilde{\boldsymbol{\gamma}}; k) = (V(\tilde{\gamma}_1; r), \ldots, V(\tilde{\gamma}_p; r))^T,$$

where r is the k-th largest component of $(|\tilde{\gamma}|_1, \ldots, |\tilde{\gamma}|_p)^T$, and $V(\gamma; r) = \gamma I(|\gamma| \geq r)$ is a hard thresholding function. Given the sparse solution $\boldsymbol{\beta}^{(t)}$ obtained at the t-th iteration, iterating (8) is equivalent to iterating the following expression

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{V}(\boldsymbol{\beta}^{(t)} - u^{-1}\boldsymbol{T}_n(\boldsymbol{\beta}^{(t)}); k).$$

The ultra-high dimensional case is often faced with a huge amount of computational tasks including matrix operations. However, the use of thresholding functions can eliminate this issue. Moreover, it naturally incorporates information on the correlations between predictors. Theorem 1 shows that the value of $D_n(\cdot)$ decreases as the number of iterations increases.

Theorem 1. *Let* $\{\boldsymbol{\beta}^{(t)}\}$ *be the sequence obtained by* (7), $\lambda_{\max}$ *be the maximum eigenvalue of* $\boldsymbol{X}^T\boldsymbol{X}$. *If* $u/2 - \bar{c}\lambda_{\max} \geq 0$ *with* $\bar{c} = \tau \vee (1-\tau)$, *the value of* $D_n(\cdot)$ *decreases as the number of iterations increases, i.e.,* $D_n(\boldsymbol{\beta}^{(t+1)}) \leq D_n(\boldsymbol{\beta}^{(t)})$.

3.2. Sure Screening Property

This subsection will prove the sure screening property of feature screening based on the SRoWER method. Define

$$\mathcal{M}_+^k = \{\mathcal{M} : \mathcal{M}^* \subset \mathcal{M}; \|\mathcal{M}\|_0 \leq k\},$$

$$\mathcal{M}_-^k = \{\mathcal{M} : \mathcal{M}^* \not\subset \mathcal{M}; \|\mathcal{M}\|_0 \leq k\}$$

as the collections of the over-fitted models and the under-fitted models, respectively. When p, k_0, k and $\boldsymbol{\beta}^*$ vary along with the sample size n, we provide the asymptotic property of $\hat{\boldsymbol{\beta}}_{[k]}$. Additionally, we make the following assumptions, some of which are completely technical and only help us comprehend the SRoWER method theoretically.

(A1) $\log p = O(n^\alpha)$ for some $0 \leq \alpha < 1$.

(A2) There exist $w_1, w_2 > 0$ and some non-negative constants η_1, η_2, such that

$$\min_{j \in \mathcal{M}^*} |\beta_j^*| \geq w_1 n^{-\eta_1}$$

and

$$k_0 \leq k \leq w_2 n^{\eta_2}.$$

(A3) There exists a constant $\eta_3 > 0$, such that $\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |x_{ij}| \leq \eta_3$.

(A4) Suppose that the random errors ε_i are i.i.d. sub-Gaussian random variables satisfying $\mathcal{E}^\tau(\varepsilon_i | \boldsymbol{x}_i) = 0, i = 1, \ldots, n$.

(A5) Let $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and

$$1/2 R_n = D_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta}^*) - \boldsymbol{\delta}^T \boldsymbol{T}_n(\boldsymbol{\beta}^*).$$

There exists a constant $v > 0$, such that, for sufficiently large n,

$$R_n \geq vn \|\boldsymbol{\delta}_{\mathcal{M}^*}\|_2^2$$

for $\|\boldsymbol{\delta}_{\mathcal{M}_c^*}\|_1 \leq 3\|\boldsymbol{\delta}_{\mathcal{M}^*}\|_1$ with $\mathcal{M}_c^*$ being the complement of $\mathcal{M}^*$.

Condition (A1) shows that p diverges exponentially with n, which is a common setting in the ultra-high dimension. The two requirements in Condition (A2) are crucial for establishing the sure screening property. The former one implies that the signals of the true model are stronger than the random errors, so they are detectable. The latter one implies that the sparsity of $\boldsymbol{\beta}$ makes sure screening possible with $s(\widehat{\mathcal{M}}) = k \geq k_0$. Condition (A3)

is a regular condition for the theoretical derivation. Condition (A4) is the same as the assumption of [17]. Condition (A5) is similar to [11].

Theorem 2. *Suppose that Conditions (A1)–(A5) are satisfied with* $2\eta_1 + \eta_2 < 1 - \alpha$. *Let* $\widehat{\mathcal{M}}$ *be the estimated model obtained by the* SRoWER *with size k; then, we have*

$$\mathbb{P}\{\mathcal{M}^* \subset \widehat{\mathcal{M}}\} \to 1, \text{as } n \to \infty.$$

By using feature screening, important features that are highly correlated with the response variable can be kept in $\widehat{\mathcal{M}}$. However, it is necessary to note that there is no explicit choice for k, because it depends on the different dimensions. Note that the IHT algorithm needs a initial estimate $\boldsymbol{\beta}^{(0)}$. To further enhance computational efficiency, the LASSO estimate is chosen as the initial value of the iterations. The following theorem shows that with the initial value $\boldsymbol{\beta}^{(0)}$ obtained using LASSO, the IHT-implemented SRoWER can satisfy the property of sure independent screening within a finite iteration.

Theorem 3. *Let* $\boldsymbol{\beta}^{(t)}$ *be the t-th update of the IHT procedure. The scale parameter* $u \geq \xi kn$ *for some* $\xi > 0$, *and let* $\mathcal{M}^{(t)} = \{j : \beta_j^{(t)} \neq 0\}$ *be the screening features. The initial value of iteration is*

$$\boldsymbol{\beta}^{(0)} = \arg\min_{\boldsymbol{\beta}} \{D_n(\boldsymbol{\beta}) + n\lambda \|\boldsymbol{\beta}\|_1\},$$

where λ *satisfies* $\lambda n^{(1-\alpha)/2} \to \infty$ *and* $\lambda = o(n^{-(\eta_1+\eta_2)})$. *Then, under Conditions (A1)–(A5), for any finite* $t \geq 1$, *we have*

$$\mathbb{P}\{\mathcal{M}^* \subset \mathcal{M}^{(t)}\} \to 1, \text{as } n \to \infty.$$

3.3. The Choice of k

For the SRoWER method, we need prespecified k, such as $\log(n)n^{1/3}$ [4,6,11]. Here, we treat k as a tuning parameter to control model complexity, and determine k by minimizing the following EBIC score:

$$EBIC(k) = \log\left\{\frac{1}{n}\sum_{i=1}^{n} \pi_i(\tilde{\boldsymbol{\beta}})\rho_\tau(y_i - \boldsymbol{x}_{i\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}})\right\} + k\frac{\log(n)\log(p)}{2n},$$

where $k = s(\mathcal{M})$. The study of [33] proposed the EBIC for model selection for large model spaces. Here, we use it to determine k for comparing the SRoWER with the EPCS proposed by [25], which also used the EBIC for model selection.

Note that the EBIC selector for determining k requires searching over $k = 1, 2, \ldots, p$. To balance the computation and model selection accuracy in practice, we minimize $EBIC(k)$ for $k = 1, \ldots, \log(n)n^{1/3}$.

4. Numerical Studies

4.1. Simulation Studies

In this subsection, the finite sample performance of SRoWER is evaluated using simulation studies and compared with EPCS [25] and SMLE [11] based on expectile regression, i.e., $\pi_i = 1, i = 1, \ldots, n$ in the SRoWER. The IHT algorithm is used to carry out feature screening based on SRoWER, and the iteration is stopped when $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2 \leq 10^{-3}$.

We take $p = 500$, $n = 100, 200$, and expectile level $\tau = 0.5, 0.05$, which correspond to the mean regression and an extreme expectile regression, respectively. All simulation results are based on 200 replications (with standard deviations in parentheses). To evaluate the performance of the screening approach, we use three criteria: the number of true positive

variables (TP), the percentage of correctly fitted models (CF), and the root mean-squared error (RMSE) $\|\hat{\boldsymbol{\beta}}_{[k]} - \boldsymbol{\beta}^*\|_2$.

Example 1. *Consider the linear model*

$$y_i = 3x_{i6} + 1.5x_{i12} + x_{i15} + x_{i20} + [\varepsilon_i - \mathcal{E}^\tau(\varepsilon_i)], \quad i = 1, \ldots, n,$$

where the candidate features $\boldsymbol{x}_i$ *are i.i.d. generated from multivariate normal distribution* $N_p(\mathbf{0}, \Sigma)$ *with* $\Sigma = (\sigma_{jk})$*, and* $\sigma_{jk} = \rho^{|j-k|}$*. We set* $\rho = 0$ *and* 0.5*. The true model is* $\mathcal{M}^* = \{6, 12, 15, 20\}$*.* ε_i *is generated from the standard gumbel distribution (Gumbel), the standard normal distribution (Normal), and the t distribution with three degrees of freedom (T), respectively.*

Example 1 considers a relatively simple case. The simulation results are given in Tables 1 and 2. For $\tau = 0.5$, we can see that all three considered methods can almost screen all important features for three different error distributions. No one method can control the other two methods in all cases, but the SRoWER performs better than the SMLE and EPCS in most instances. However, at extreme expectiles $(\tau = 0.05)$, the SRoWER method performs much better than the SMLE and EPCS in terms of RMSE except for the case of Gumbel with $n = 200$. In addition, all the results become better when the sample size increases from 100 to 200.

Example 2. *For the linear model in Example 1, to examine the robustness of the SRoWER, we consider the case where there are outliers in the covariates. We first generate data as those in Example 1. Next, we artificially add outliers from* $N(0, 5^2)$ *to random 50 covariates of 10% of the observations. The other settings are the same as those in Example 1.*

Table 1. Simulation results of Example 1 for $\tau = 0.5$.

ρ	ε	n	Method	TP	CF	RMSE
0	Normal	100	SRoWER	4 (0)	0.880 (0.326)	0.985 (0.427)
			SMLE	4 (0)	0.910 (0.287)	1.092 (0.418)
			EPCS	4 (0)	0.935 (0.247)	0.951 (0.424)
		200	SRoWER	4 (0)	0.95 (0.218)	0.514 (0.077)
			SMLE	4 (0)	0.95 (0.218)	0.693 (0.162)
			EPCS	4 (0)	0.95 (0.218)	0.696 (0.162)
	Gumbel	100	SRoWER	3.995 (0.071)	0.880 (0.326)	1.153 (0.349)
			SMLE	4.000 (0.000)	0.905 (0.294)	1.518 (0.514)
			EPCS	3.990 (0.100)	0.920 (0.272)	1.390 (0.416)
		200	SRoWER	4 (0)	0.900 (0.301)	0.756 (0.156)
			SMLE	4 (0)	0.965 (0.184)	0.877 (0.216)
			EPCS	4 (0)	0.965 (0.184)	0.922 (0.244)
	T	100	SRoWER	3.900 (0.375)	0.765 (0.425)	1.665 (0.587)
			SMLE	3.845 (0.471)	0.795 (0.405)	2.351 (0.914)
			EPCS	3.795 (0.578)	0.765 (0.425)	2.450 (1.143)
		200	SRoWER	3.980 (0.199)	0.91 (0.287)	0.854 (0.241)
			SMLE	3.960 (0.262)	0.92 (0.272)	1.260 (0.451)
			EPCS	3.945 (0.320)	0.92 (0.272)	1.286 (0.457)

Table 1. *Cont.*

ρ	ε	n	Method	TP	CF	RMSE
0.5	Normal	100	SRoWER	4 (0)	0.935 (0.247)	0.714 (0.245)
			SMLE	4 (0)	0.945 (0.229)	0.834 (0.330)
			EPCS	4 (0)	0.955 (0.208)	0.741 (0.327)
		200	SRoWER	4 (0)	0.905 (0.294)	0.594 (0.094)
			SMLE	4 (0)	0.975 (0.157)	0.322 (0.214)
			EPCS	4 (0)	0.985 (0.122)	0.290 (0.178)
	Gumbel	100	SRoWER	3.995 (0.071)	0.90 (0.301)	0.943 (0.379)
			SMLE	3.995 (0.071)	0.91 (0.287)	1.175 (0.501)
			EPCS	3.975 (0.186)	0.91 (0.287)	1.187 (0.519)
		200	SRoWER	4 (0)	0.905 (0.294)	0.695 (0.169)
			SMLE	4 (0)	0.975 (0.157)	0.570 (0.289)
			EPCS	4 (0)	0.990 (0.100)	0.397 (0.167)
	T	100	SRoWER	3.870 (0.429)	0.775 (0.419)	1.924 (0.580)
			SMLE	3.780 (0.551)	0.770 (0.422)	2.724 (1.067)
			EPCS	3.525 (0.795)	0.550 (0.499)	2.711 (0.907)
		200	SRoWER	3.970 (0.222)	0.845 (0.363)	0.986 (0.277)
			SMLE	3.955 (0.252)	0.950 (0.218)	0.665 (0.234)
			EPCS	3.940 (0.327)	0.935 (0.247)	0.729 (0.223)

Example 2 considers the case where both covariates and response variables have outliers. The simulation results are shown in Tables 3 and 4. We can see that the SRoWER has the smallest RMSE compared with SMLE and EPCS. Three considered methods have similar performance in variable selection, except for the case of T. Both SRoWER and SMLE perform better than EPCS in terms of CF for the case of T.

Table 2. Simulation results of Example 1 for $\tau = 0.05$.

ρ	ε	n	Method	TP	CF	RMSE
0	Normal	100	SRoWER	3.995 (0.071)	0.725 (0.448)	1.565 (0.422)
			SMLE	4.000 (0.000)	0.760 (0.428)	2.187 (0.661)
			EPCS	3.995 (0.071)	0.510 (0.501)	3.067 (0.951)
		200	SRoWER	4 (0)	0.725 (0.448)	0.891 (0.229)
			SMLE	4 (0)	0.690 (0.464)	1.635 (0.589)
			EPCS	4 (0)	0.540 (0.500)	2.031 (0.697)
	Gumbel	100	SRoWER	3.980 (0.140)	0.875 (0.332)	1.050 (0.192)
			SMLE	3.995 (0.071)	0.950 (0.218)	0.932 (0.291)
			EPCS	3.995 (0.071)	0.830 (0.377)	1.391 (0.393)
		200	SRoWER	4 (0)	0.885 (0.320)	0.601 (0.111)
			SMLE	4 (0)	0.985 (0.122)	0.422 (0.131)
			EPCS	4 (0)	0.960 (0.196)	0.619 (0.184)
	T	100	SRoWER	3.120 (1.049)	0.295 (0.457)	4.441 (1.483)
			SMLE	3.325 (0.907)	0.235 (0.425)	7.941 (3.112)
			EPCS	3.120 (1.025)	0.065 (0.247)	9.146 (3.500)
		200	SRoWER	3.765 (0.558)	0.400 (0.491)	2.918 (1.069)
			SMLE	3.800 (0.540)	0.180 (0.385)	7.217 (2.362)
			EPCS	3.800 (0.530)	0.125 (0.332)	7.767 (2.407)

Table 2. *Cont.*

ρ	ε	n	Method	TP	CF	RMSE
0.5	Normal	100	SRoWER	3.985 (0.122)	0.79 (0.408)	1.355 (0.364)
			SMLE	3.990 (0.141)	0.83 (0.377)	1.746 (0.645)
			EPCS	3.880 (0.487)	0.50 (0.501)	2.710 (0.887)
		200	SRoWER	4 (0)	0.665 (0.473)	0.962 (0.251)
			SMLE	4 (0)	0.725 (0.448)	1.752 (0.515)
			EPCS	4 (0)	0.705 (0.457)	1.733 (0.499)
	Gumbel	100	SRoWER	4.00 (0.000)	0.890 (0.314)	0.979 (0.246)
			SMLE	4.00 (0.000)	0.935 (0.247)	1.035 (0.253)
			EPCS	3.92 (0.405)	0.635 (0.483)	1.563 (0.618)
		200	SRoWER	4 (0)	0.765 (0.425)	0.774 (0.120)
			SMLE	4 (0)	0.935 (0.247)	0.670 (0.301)
			EPCS	4 (0)	0.935 (0.247)	0.541 (0.372)
	T	100	SRoWER	3.260 (0.926)	0.270 (0.445)	4.738 (1.573)
			SMLE	3.415 (0.852)	0.265 (0.442)	7.569 (2.619)
			EPCS	2.615 (1.159)	0.065 (0.247)	8.153 (2.855)
		200	SRoWER	3.725 (0.593)	0.425 (0.496)	2.296 (0.635)
			SMLE	3.770 (0.518)	0.220 (0.415)	7.199 (2.344)
			EPCS	3.460 (0.801)	0.145 (0.353)	6.908 (2.204)

Example 3. *Here we consider a heterogeneous model. We first generate* $(z_1, \ldots, z_p)^T$ *from multivariate normal distribution* $N_p(\mathbf{0}, \Sigma)$ *with* $\Sigma = (\sigma_{jk})$, *and* $\sigma_{jk} = \rho^{|j-k|}$. *We set* $\rho = 0$ *and* 0.5. *Let* $x_1 = \Phi(z_1)$ *and* $x_j = z_j$, $j = 2, 3, \ldots, p$, *where* $\Phi(\cdot)$ *is the cumulative distribution function of standard normal distribution. The response is then simulated from the following normal linear heteroscedastic model*

$$y_i = 3x_{i6} + 1.5x_{i12} + x_{i15} + x_{i20} + (0.7x_{i1})[\varepsilon_i - \mathcal{E}^{\tau}(\varepsilon_i)],$$

where $\varepsilon_i \sim N(0, 1)$. *Meanwhile, the other settings are the same as those in Example 2.*

Table 3. Simulation results of Example 2 for $\tau = 0.5$.

ρ	ε	n	Method	TP	CF	RMSE
0	Normal	100	SRoWER	3.685 (0.713)	0.645 (0.480)	1.695 (0.569)
			SMLE	3.770 (0.537)	0.685 (0.466)	2.771 (1.039)
			EPCS	3.700 (0.716)	0.690 (0.464)	2.880 (0.987)
		200	SRoWER	3.940 (0.277)	0.755 (0.431)	1.492 (0.503)
			SMLE	3.935 (0.285)	0.820 (0.385)	2.216 (0.806)
			EPCS	3.915 (0.372)	0.820 (0.385)	2.150 (0.711)
	Gumbel	100	SRoWER	3.655 (0.720)	0.635 (0.483)	1.553 (0.521)
			SMLE	3.680 (0.663)	0.695 (0.462)	2.986 (1.355)
			EPCS	3.605 (0.782)	0.675 (0.470)	2.806 (0.871)
		200	SRoWER	3.825 (0.464)	0.725 (0.448)	1.296 (0.551)
			SMLE	3.820 (0.478)	0.780 (0.415)	2.050 (0.688)
			EPCS	3.805 (0.528)	0.805 (0.397)	1.456 (0.585)
	T	100	SRoWER	3.495 (0.821)	0.560 (0.498)	2.030 (0.712)
			SMLE	3.445 (0.819)	0.585 (0.494)	3.532 (1.271)
			EPCS	3.380 (0.927)	0.540 (0.500)	3.123 (1.117)
		200	SRoWER	3.895 (0.380)	0.675 (0.470)	1.611 (0.518)
			SMLE	3.805 (0.508)	0.785 (0.412)	1.831 (0.558)
			EPCS	3.830 (0.482)	0.810 (0.393)	2.013 (0.693)

Table 3. *Cont.*

ρ	ε	n	Method	TP	CF	RMSE
0.5	Normal	100	SRoWER	3.735 (0.646)	0.660 (0.475)	1.740 (0.586)
			SMLE	3.760 (0.587)	0.715 (0.453)	3.033 (0.881)
			EPCS	3.660 (0.760)	0.690 (0.464)	3.292 (1.296)
		200	SRoWER	3.950 (0.240)	0.645 (0.480)	1.588 (0.421)
			SMLE	3.860 (0.437)	0.690 (0.464)	2.469 (0.457)
			EPCS	3.855 (0.525)	0.745 (0.437)	2.582 (0.500)
	Gumbel	100	SRoWER	3.735 (0.622)	0.655 (0.477)	1.839 (0.567)
			SMLE	3.725 (0.601)	0.690 (0.464)	2.801 (1.184)
			EPCS	3.685 (0.615)	0.670 (0.471)	2.463 (1.101)
		200	SRoWER	3.885 (0.377)	0.640 (0.481)	1.617 (0.487)
			SMLE	3.840 (0.394)	0.675 (0.470)	2.523 (0.661)
			EPCS	3.820 (0.468)	0.715 (0.453)	2.624 (0.710)
	T	100	SRoWER	3.455 (0.801)	0.445 (0.498)	2.114 (0.549)
			SMLE	3.425 (0.792)	0.475 (0.501)	3.765 (1.072)
			EPCS	3.100 (1.070)	0.405 (0.492)	3.186 (1.153)
		200	SRoWER	3.910 (0.304)	0.520 (0.501)	1.949 (0.523)
			SMLE	3.775 (0.525)	0.685 (0.466)	2.756 (0.671)
			EPCS	3.670 (0.703)	0.660 (0.475)	2.877 (0.735)

From Tables 5 and 6, we can see that the conclusions are similar to those of Examples 1 and 2. Hence, the SRoWER performs well even in the case that there are outliers in the heterogeneous model.

Table 4. Simulation results of Example 2 for $\tau = 0.05$.

ρ	ε	n	Method	TP	CF	RMSE
0	Normal	100	SRoWER	3.530 (0.918)	0.385 (0.488)	2.090 (0.576)
			SMLE	3.755 (0.638)	0.535 (0.500)	4.484 (1.384)
			EPCS	3.575 (0.865)	0.325 (0.470)	4.070 (1.368)
		200	SRoWER	3.75 (0.671)	0.550 (0.499)	1.406 (0.587)
			SMLE	3.98 (0.140)	0.490 (0.501)	4.210 (1.145)
			EPCS	3.87 (0.463)	0.495 (0.501)	3.615 (1.053)
	Gumbel	100	SRoWER	3.490 (0.930)	0.395 (0.490)	1.649 (0.537)
			SMLE	3.745 (0.626)	0.630 (0.484)	4.217 (1.552)
			EPCS	3.535 (0.856)	0.475 (0.501)	3.409 (1.670)
		200	SRoWER	3.755 (0.606)	0.600 (0.491)	1.370 (0.578)
			SMLE	3.885 (0.391)	0.655 (0.477)	3.338 (1.196)
			EPCS	3.785 (0.592)	0.675 (0.470)	3.088 (1.085)
	T	100	SRoWER	2.795 (1.127)	0.165 (0.372)	3.357 (0.890)
			SMLE	2.985 (1.082)	0.185 (0.389)	8.103 (2.779)
			EPCS	2.735 (1.184)	0.070 (0.256)	8.865 (2.696)
		200	SRoWER	3.44 (0.900)	0.34 (0.475)	2.686 (0.759)
			SMLE	3.67 (0.635)	0.17 (0.377)	7.480 (2.210)
			EPCS	3.56 (0.692)	0.09 (0.287)	7.741 (2.436)

Table 4. *Cont.*

ρ	ε	n	Method	TP	CF	RMSE
0.5	Normal	100	SRoWER	3.610 (0.966)	0.40 (0.491)	1.966 (0.582)
			SMLE	3.795 (0.524)	0.54 (0.500)	4.450 (1.279)
			EPCS	3.365 (1.170)	0.30 (0.459)	4.641 (1.729)
		200	SRoWER	3.835 (0.547)	0.480 (0.501)	1.556 (0.400)
			SMLE	3.925 (0.346)	0.515 (0.501)	3.623 (0.840)
			EPCS	3.760 (0.628)	0.545 (0.499)	3.360 (0.939)
	Gumbel	100	SRoWER	3.700 (0.702)	0.435 (0.497)	1.739 (0.621)
			SMLE	3.785 (0.529)	0.685 (0.466)	3.233 (1.333)
			EPCS	3.560 (0.895)	0.505 (0.501)	3.989 (1.709)
		200	SRoWER	3.835 (0.509)	0.530 (0.500)	1.620 (0.572)
			SMLE	3.910 (0.335)	0.620 (0.487)	3.390 (1.064)
			EPCS	3.770 (0.607)	0.665 (0.473)	3.121 (1.015)
	T	100	SRoWER	2.815 (1.099)	0.115 (0.320)	3.167 (0.756)
			SMLE	2.960 (0.992)	0.155 (0.363)	8.137 (2.316)
			EPCS	2.265 (1.068)	0.025 (0.157)	8.689 (2.627)
		200	SRoWER	3.445 (0.843)	0.305 (0.462)	2.703 (0.689)
			SMLE	3.620 (0.662)	0.155 (0.363)	7.437 (2.066)
			EPCS	3.115 (0.998)	0.145 (0.353)	8.051 (1.960)

Table 5. Simulation results of Example 3 for $\tau = 0.5$.

ρ	n	Method	TP	CF	RMSE
0	100	SRoWER	3.81 (0.579)	0.690 (0.464)	1.600 (0.642)
		SMLE	3.86 (0.471)	0.665 (0.473)	3.671 (1.112)
		EPCS	3.83 (0.532)	0.670 (0.471)	3.186 (1.061)
	200	SRoWER	3.945 (0.25)	0.785 (0.412)	0.968 (0.513)
		SMLE	3.965 (0.21)	0.765 (0.425)	1.956 (0.784)
		EPCS	3.965 (0.21)	0.815 (0.389)	1.762 (0.859)
0.5	100	SRoWER	3.895 (0.495)	0.65 (0.478)	1.432 (0.566)
		SMLE	3.915 (0.372)	0.64 (0.481)	2.811 (0.826)
		EPCS	3.845 (0.559)	0.70 (0.459)	2.841 (0.937)
	200	SRoWER	3.970 (0.198)	0.675 (0.470)	1.325 (0.557)
		SMLE	3.980 (0.140)	0.665 (0.473)	1.968 (0.585)
		EPCS	3.965 (0.184)	0.675 (0.470)	2.221 (0.641)

Table 6. Simulation results of Example 3 for $\tau = 0.05$.

ρ	n	Method	TP	CF	RMSE
0	100	SRoWER	3.610 (0.907)	0.545 (0.499)	1.189 (0.659)
		SMLE	3.845 (0.531)	0.530 (0.500)	4.331 (1.956)
		EPCS	3.725 (0.783)	0.515 (0.501)	4.003 (1.500)
	200	SRoWER	3.750 (0.735)	0.600 (0.491)	0.947 (0.715)
		SMLE	3.945 (0.250)	0.535 (0.500)	4.572 (1.206)
		EPCS	3.900 (0.401)	0.560 (0.498)	3.445 (1.388)
0.5	100	SRoWER	3.73 (0.735)	0.435 (0.497)	1.602 (0.523)
		SMLE	3.92 (0.323)	0.550 (0.499)	3.820 (1.434)
		EPCS	3.73 (0.692)	0.520 (0.501)	3.843 (1.253)
	200	SRoWER	3.950 (0.240)	0.49 (0.501)	1.427 (0.527)
		SMLE	3.985 (0.122)	0.53 (0.500)	3.478 (1.163)
		EPCS	3.905 (0.383)	0.54 (0.500)	3.212 (0.929)

4.2. Real Data Example

This subsection applies the SRoWER method for feature screening to the Mid-Atlantic wage data with 3000 observations and 8 predictors from [34] that are available in the 'ISLR' package in R. A total of eight predictors (two continuous and six categorical) are considered. The continuous variables are year that wage information was recorded (year) and age of worker (age). The categorical factors include a factor with the following levels: 1. Never Married, 2. Married, 3. Widowed, 4. Divorced, and 5. Separated, which indicate marital status (marital). Another factor (race) contains the following levels: 1. White, 2. Black, 3. Asian and 4. Other. Another factor (education) contains the following levels 1. <HS Grad, 2. HS Grad, 3. Some College, 4. College Grad and 5. Advanced Degree. Another factor (jobclass) contains the following levels: 1. Industrial and 2. Information indicating type of job. Another factor (health) contains the following levels, 1. <= Good and 2. >= Very Good, indicating health level of worker. Another factor contains the following levels, 1. Yes and 2. No, indicating whether worker has health insurance (health_ins). We use the dummy variables to represent six categorical variables. Therefore, there are 16 covariates, and the response is the logarithm of wage. Following the set up of [35], to demonstrate the application in high dimension, we extend the data by introducing the following artificial covariates:

$$X_j = \frac{Z_j + 2W}{3}, j = 17, \ldots, 500,$$

where Z_j is the standard normal random variables and W follows the standard uniform distribution.

To test the prediction performance of SRoWER, EPCS, and SMLE, we randomly generatde 100 partitions of the full data, and divided the data into two parts, where $n_{tr} = 1500$ samples are treated as training data and the remaining $n_{te} = 1500$ samples are treated as testing data. The average of model size (Size), the number of selected noise variables (SNV) and expectile prediction error (EPE) at $\tau = 0.5$ and 0.05, where EPE is computed by the test data

$$EPE_\tau = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \pi_i(\tilde{\boldsymbol{\beta}}) \rho_\tau(y_i - \hat{y}_i),$$

where $\hat{y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_\tau$ is the τ expectile estimate of the i-th observation with $\hat{\boldsymbol{\beta}}_\tau$ being calculated based on the training data set, and y_i is the i-th observation in the test set.

The results are reported in Table 7. For $\tau = 0.5$, both SRoWER and SMLE perform similarly in terms of EPE and SNV, while the SRoWER includes more than one variable compared with the SMLE. Although the model size of the SRoWER and EPCS are similar, the EPE of the EPCS is the largest among three methods. For $\tau = 0.05$, the SRoWER performs best, while the EPCS performs worst. The selected model sizes vary for different τ, it indicates the heteroscedasticity of the model. This conclusion agrees with the results of [36].

Table 7. Expectile prediction error (EPE), model size (Size), and selected noise variables (SNV) over 100 repetitions and their standard errors (in parentheses) for wage data.

τ	Method	EPE	Size	SNV
0.5	SRoWER	0.041 (0.002)	6.63 (1.107)	0.06 (0.239)
	SMLE	0.042 (0.002)	5.44 (0.978)	0.04 (0.243)
	EPCS	0.048 (0.005)	6.45 (0.687)	0.00 (0.000)
0.05	SRoWER	0.017 (0.001)	4.60 (1.511)	0.33 (0.620)
	SMLE	0.030 (0.014)	6.03 (2.422)	1.12 (1.565)
	EPCS	0.275 (0.183)	10.11 (3.127)	3.62 (2.107)

5. Conclusions

To deal with the heterogeneity and the outliers in the covariates and/or the response, this paper proposes the RoWER method, which is further applied to screen the features in ultra-high dimensional data. We have also proposed an iterative hard-thresholding algorithm for implementing the feature screening procedure, and establish the sure screening property for the SRoWER method. Simulation studies and a real data analysis verify that the SRoWER method not only reduces the huge computational effort faced by ultra-high dimensional data, but also shows excellent robustness in heterogeneous data. Compared with ISIS [6], the SRoWER naturally accounts for the joint effects between features, and benefits from the advantage of the SMLE in terms of computational efficiency. Based on the proposed method, the problem of robust feature screening for classification data also presents a promising direction for future research.

Author Contributions: Conceptualization, M.W.; methodology, X.W., P.H. and M.W.; software, X.W.; formal analysis, X.W. and P.H.; writing—original draft preparation, X.W.; writing—review and editing, P.H. and M.W.; supervision, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (12271294), and the Natural Science Foundation of Shandong Province (ZR2024MA089).

Data Availability Statement: Data sets were provided in the 'ISLR' package in R.

Acknowledgments: The authors are grateful to the editor and reviewers for their valuable comments and suggestions. We also sincerely thank Yundong Tu for providing their codes for us.

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

Appendix A

Proof of Proposition 1. It is seen that

$$\begin{aligned}\int_{-\infty}^{+\infty} f^2 du &= \int_{-\infty}^{+\infty} \frac{4\tau(1-\tau)}{\pi\sigma^2(1+2\sqrt{\tau(1-\tau)})} \exp\left\{-2\rho_\tau\left(\frac{u-\mu_\tau}{\sigma}\right)\right\} du \\ &= \frac{4\tau(1-\tau)}{\pi\sigma^2(1+2\sqrt{\tau(1-\tau)})} \int_{-\infty}^{+\infty} \exp\left\{-2|\tau - I(u \le \mu_\tau)| \frac{(u-\mu_\tau)^2}{\sigma^2}\right\} du \\ &= \frac{4\tau(1-\tau)}{\pi\sigma^2(1+2\sqrt{\tau(1-\tau)})} \times \\ &\quad \left[\int_{\mu_\tau}^{\infty} \exp\left\{-2\tau\frac{(u-\mu_\tau)^2}{\sigma^2}\right\} du + \int_{-\infty}^{\mu_\tau} \exp\left\{-2(1-\tau)\frac{(u-\mu_\tau)^2}{\sigma^2}\right\} du\right].\end{aligned}$$

Note that

$$\int_{\mu_\tau}^{+\infty} \exp\left\{-2\tau\frac{(u-\mu_\tau)^2}{\sigma^2}\right\} du = \frac{\sigma\sqrt{\pi}}{2\sqrt{2\tau}},$$

$$\int_{-\infty}^{\mu_\tau} \exp\left\{-2(1-\tau)\frac{(u-\mu_\tau)^2}{\sigma^2}\right\} du = \frac{\sigma\sqrt{\pi}}{2\sqrt{2(1-\tau)}}.$$

Then, we have

$$\begin{aligned}\int_{-\infty}^{+\infty} f^2 du &= \frac{4\tau(1-\tau)}{\pi\sigma^2(1+2\sqrt{\tau(1-\tau)})}\left[\frac{\sigma\sqrt{\pi}}{2\sqrt{2\tau}}+\frac{\sigma\sqrt{\pi}}{2\sqrt{2(1-\tau)}}\right]\\ &= \frac{(\sqrt{\tau}+\sqrt{1-\tau})\sqrt{2\tau(1-\tau)}}{\sigma\sqrt{\pi}(1+2\sqrt{\tau(1-\tau)})}\\ &= \frac{1}{\sqrt{\pi\sigma^2}}\frac{\sqrt{2\tau(1-\tau)}}{\sqrt{\tau}+\sqrt{1-\tau}}.\end{aligned}$$

Hence, the L_2E criterion for asymmetric normal distribution is

$$\frac{1}{\sqrt{\pi\sigma^2}}\frac{\sqrt{2\tau(1-\tau)}}{\sqrt{\tau}+\sqrt{1-\tau}}\left\{1-2\sqrt{2}\frac{1}{n}\sum_{i=1}^{n}\exp\left[-\rho_\tau\left(\frac{v_i-\mu_\tau}{\sigma}\right)\right]\right\}.$$

The proposition is proved. □

Proof of Proposition 2. By Taylor's expansion, we have

$$\begin{aligned}&\log(-L_n(\boldsymbol{\beta}))\\ \approx&\log(-L_n(\tilde{\boldsymbol{\beta}}))-\sum\pi_i(\tilde{\boldsymbol{\beta}})\left[\rho_\tau(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})-\rho_\tau(y_i-\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})\right],\end{aligned}$$

for $\boldsymbol{\beta}\approx\tilde{\boldsymbol{\beta}}$, where

$$\pi_i(\tilde{\boldsymbol{\beta}})=\frac{\exp\{-\rho_\tau(y_i-\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})\}}{\sum_{l=1}^{n}\exp\{-\rho_\tau(y_l-\boldsymbol{x}_l^T\tilde{\boldsymbol{\beta}})\}}.$$

Therefore, the minimization of $L_n(\boldsymbol{\beta})$ is transformed into the minimization of the following objective function

$$D_n(\boldsymbol{\beta})=\sum_{i=1}^{n}\pi_i(\tilde{\boldsymbol{\beta}})\rho_\tau(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}).$$

The proposition is then proved. □

Before proving Theorem 1, we give the following lemmas.

Lemma A1 ([17]). *The asymmetric least squared loss $\rho_\tau(\cdot)$ is continuously differentiable, but is not twice differentiable at zero when $\tau\neq 0.5$. Moreover, for any $t,t_0\in\mathbb{R}$ and $\tau\in(0,1)$, we have*

$$\underline{c}(t-t_0)^2\leq\rho_\tau(t)-\rho_\tau(t_0)-\rho'_\tau(t_0)(t-t_0)\leq\bar{c}(t-t_0)^2,$$

where $\underline{c}=\tau\wedge(1-\tau)$, $\bar{c}=\tau\vee(1-\tau)$, which confirms that ρ_τ is strongly convex.

Lemma A2 ([17]). *For any $t,t_0\in\mathbb{R}$ and $\tau\in(0,1)$, we have*

$$2\underline{c}|t-t_0|\leq|\rho'_\tau(t)-\rho'_\tau(t_0)|\leq 2\bar{c}|t-t_0|.$$

The lemma follows that $\rho'_\tau(\cdot)$ is Lipschitz continuous.

Lemma A3 ([17]). *Let $Z_1,\ldots,Z_n\in\mathbb{R}$ be i.i.d. sub-Gaussian random variables, and $K=\max_i\|Z_i\|_{\mathrm{SG}}$ be sub-Gaussian norm, $i=1,\ldots,n$, where $\|Z_i\|_{\mathrm{SG}}=\sup_{s\geq1}s^{-1/2}(E|Z_i|^s)^{1/s}$. Then for any $\boldsymbol{a}=(a_1,\ldots,a_n)^T\in\mathbb{R}^n$ and $h\geq0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n}a_iZ_i\right|\geq h\right\}\leq 2\exp\left\{-\frac{Ch^2}{K^2\|\boldsymbol{a}\|_2^2}\right\}.$$

where $C > 0$ is a constant.

Lemma A4 ([17]). *Let Z be sub-Gaussian random variable, $Z^+ = \max(Z, 0)$, $Z^- = \max(-Z, 0)$; then, random variables Z^+ and Z^- are also sub-Gaussian. For any $b_1,\ b_2 \in \mathbb{R}$, $b_1 Z^+ + b_2 Z^-$ is sub-Gaussian.*

Proof of Theorem 1. The definition of the IHT algorithm based on $Q_n(\cdot)$ is

$$Q_n(\gamma; \boldsymbol{\beta}) = D_n(\boldsymbol{\beta}) + (\gamma - \boldsymbol{\beta})^T \boldsymbol{T}_n(\boldsymbol{\beta}) + (u/2)\|\gamma - \boldsymbol{\beta}\|_2^2,$$

$$\boldsymbol{\beta}^{(t+1)} = \text{argmin}_\gamma Q_n(\gamma; \boldsymbol{\beta}^{(t)}), \quad \|\gamma\|_0 \le k.$$

By Lemma A1 and the assumption $u/2 \ge \bar{c}\lambda_{\max}$, we have

$$\begin{aligned} D_n(\boldsymbol{\beta}^{(t)}) &= Q_n(\boldsymbol{\beta}^{(t)}; \boldsymbol{\beta}^{(t)}) \\ &\ge Q_n(\boldsymbol{\beta}^{(t+1)}; \boldsymbol{\beta}^{(t)}) \\ &= D_n(\boldsymbol{\beta}^{(t+1)}) - [D_n(\boldsymbol{\beta}^{(t+1)}) - D_n(\boldsymbol{\beta}^{(t)}) - (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})^T \boldsymbol{T}_n(\boldsymbol{\beta}^{(t)})] \\ &\quad + (u/2)\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 \\ &\ge D_n(\boldsymbol{\beta}^{(t+1)}) - \bar{c}\|\boldsymbol{X}(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})\|_2^2 + u/2\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 \\ &\ge D_n(\boldsymbol{\beta}^{(t+1)}) + (u/2 - \bar{c}\lambda_{\max})\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2. \end{aligned}$$

This proves that $D_n(\boldsymbol{\beta}^{(t)})$ decreases after each iteration. □

Lemma A5. *For some $\lambda > 0$, denote $\boldsymbol{\beta}_L = \text{argmin}_{\boldsymbol{\beta}}\{D_n(\boldsymbol{\beta}) + n\lambda\|\boldsymbol{\beta}\|_1\}$. Under Conditions of Theorem 3, we have*

$$\mathbb{P}\{\|\boldsymbol{\beta}_L - \boldsymbol{\beta}^*\|_1 \le 16v^{-1}\lambda k_0\} \to 1, n \to \infty,$$

where v is the same as defined in Condition (A5).

Proof of Lemma A5. Based on the definition of $\boldsymbol{\beta}_L$, we have

$$D_n(\boldsymbol{\beta}_L) + n\lambda\|\boldsymbol{\beta}_L\|_1 \le D_n(\boldsymbol{\beta}^*) + n\lambda\|\boldsymbol{\beta}^*\|_1,$$

that is,

$$D_n(\boldsymbol{\beta}_L) - D_n(\boldsymbol{\beta}^*) \le n\lambda\|\boldsymbol{\beta}^*\|_1 - n\lambda\|\boldsymbol{\beta}_L\|_1.$$

Denote

$$1/2R_n = D_n(\boldsymbol{\beta}_L) - D_n(\boldsymbol{\beta}^*) - (\boldsymbol{\beta}_L - \boldsymbol{\beta}^*)^T \boldsymbol{T}_n(\boldsymbol{\beta}^*)$$

and $\delta^L = \boldsymbol{\beta}_L - \boldsymbol{\beta}^*$, it follows that

$$\begin{aligned} n^{-1}R_n &\le 2n^{-1}\sum_{i=1}^n 2\pi_i\omega_{\tau,i}(\boldsymbol{\beta}^*)(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*)\boldsymbol{x}_i^T\delta^L + 2\lambda\|\boldsymbol{\beta}^*\|_1 - 2\lambda\|\boldsymbol{\beta}_L\|_1 \\ &\le 2n^{-1}\max_{1\le j\le p}|T_{nj}(\boldsymbol{\beta}^*)| \cdot \|\delta^L\|_1 + 2\lambda\|\boldsymbol{\beta}^*\|_1 - 2\lambda\|\boldsymbol{\beta}_L\|_1. \end{aligned}$$

We now derive a bound on $\boldsymbol{T}_n(\boldsymbol{\beta}^*)$. Define

$$\mathcal{A} = \{\max_{1\le j\le p}|T_{nj}(\boldsymbol{\beta}^*)| \le n\lambda/2\}.$$

By Lemma A3 and Condition (A3), for each $j \in \{1, \ldots, p\}$, we have

$$\begin{aligned}\mathbb{P}(|T_{nj}(\boldsymbol{\beta}^*)| > n\lambda/2) &= \mathbb{P}\left\{\left|\sum_{i=1}^{n} \pi_i x_{ij}\zeta_i\right| > n\lambda/2\right\} \\ &\leq 2\exp\left\{-\frac{Cn^2\lambda^2}{4K^2\sum_{i=1}^{n}(\pi_i x_{ij})^2}\right\} \\ &\leq 2\exp\left\{-\frac{Cn\lambda^2}{4K^2\eta_3^2}\right\},\end{aligned}$$

where $\zeta_i = 2\omega_{\tau,i}(\boldsymbol{\beta}^*)(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*)$, which is sub-Gaussian with Condition (A4) and Lemma A4. Hence, for any $\alpha \in (0,1)$, and some generic positive constants c, a,

$$\begin{aligned}\mathbb{P}(\mathcal{A}^c) &\leq \sum_{j=1}^{p}\mathbb{P}\{|T_{nj}(\boldsymbol{\beta}^*)| > n\lambda/2\} \\ &\leq 2p\exp\left\{-\frac{Cn\lambda^2}{4K^2\eta_3^2}\right\} \\ &\leq \exp\{an^\alpha - c\lambda^2 n\} \to 0, \end{aligned} \tag{A1}$$

as $n \to 0$. This implies $\mathbb{P}(\mathcal{A}) \to 1$. Therefore, under the event $\mathcal{A}$, we have

$$n^{-1}R_n \leq \lambda\|\boldsymbol{\delta}^L\|_1 + 2\lambda\|\boldsymbol{\beta}^*\|_1 - 2\lambda\|\boldsymbol{\beta}_L\|_1.$$

This further implies that

$$\begin{aligned}n^{-1}R_n + \lambda\|\boldsymbol{\delta}^L\|_1 &\leq 2\lambda\left(\sum_{i=1}^{p}|\beta_j - \beta_j^*| + |\beta_j^*| - |\beta_j|\right) \\ &= 2\lambda\left(\sum_{j\in\mathcal{M}^*}|\beta_j - \beta_j^*| + |\beta_j^*| - |\beta_j|\right) \\ &\leq 4\lambda\|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_1. \end{aligned} \tag{A2}$$

Since $R_n \geq 0$ by Lemma A1, we have $\|\boldsymbol{\delta}^L\|_1 \leq 4\|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_1$, which leads to $\|\boldsymbol{\delta}^L_{\mathcal{M}_c^*}\|_1 \leq 3\|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_1$. By Condition (A5), (A2) and Cauchy inequality, it follows that

$$\|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_1^2 \leq k_0\|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_2^2 \leq k_0 v^{-1}(n^{-1}R_n) \leq 4v^{-1}k_0\lambda\|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_1,$$

which gives rise to $\|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_1 \leq 4v^{-1}\lambda k_0$. Thus, under the event $\mathcal{A}$, we have

$$\|\boldsymbol{\delta}^L\|_1 = \|\boldsymbol{\delta}^L_{\mathcal{M}_c^*}\|_1 + \|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_1 \leq 4\|\boldsymbol{\delta}^L_{\mathcal{M}^*}\|_1 \leq 16v^{-1}\lambda k_0.$$

The lemma is proved. □

Proof of Theorem 2. Let $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ be the estimator of $\boldsymbol{\beta}$ obtained by SRoWER based on model $\mathcal{M}$. If $\mathbb{P}\{\widehat{\mathcal{M}} \in \mathcal{M}_+^k\} \to 1$, the theorem is then proved. This suffices to show that

$$\mathbb{P}\{\min\nolimits_{\mathcal{M}\in\mathcal{M}_-^k} D_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) \leq \max\nolimits_{\mathcal{M}\in\mathcal{M}_+^k} D_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\} \to 0,$$

as $n \to \infty$.

For any $\mathcal{M} \in \mathcal{M}_-^k$, define $\mathcal{M}' = \mathcal{M} \cup \mathcal{M}^* \in \mathcal{M}_+^{2k}$. Consider $\boldsymbol{\beta}_{\mathcal{M}'}$ close to $\boldsymbol{\beta}^*_{\mathcal{M}'}$ such that $\|\boldsymbol{\beta}_{\mathcal{M}'} - \boldsymbol{\beta}^*_{\mathcal{M}'}\|_2 = w_1 n^{-\eta_1}$, for some $w_1,\ \eta_1 > 0$. When n is sufficiently large, $\boldsymbol{\beta}_{\mathcal{M}'}$ falls into a small neighborhood of $\boldsymbol{\beta}^*_{\mathcal{M}'}$. By Lemma A1, we have

$$\begin{aligned}
&D_n(\boldsymbol{\beta}^*_{\mathcal{M}'}) - D_n(\boldsymbol{\beta}_{\mathcal{M}'}) \\
=&\sum_{i=1}^n \pi_i \left[\rho_\tau(y_i - \boldsymbol{x}_{i,\mathcal{M}'}^T \boldsymbol{\beta}^*_{\mathcal{M}'}) - \rho_\tau(y_i - \boldsymbol{x}_{i,\mathcal{M}'}^T \boldsymbol{\beta}_{\mathcal{M}'})\right] \\
\le&\sum_{i=1}^n \left[\pi_i (\boldsymbol{\beta}_{\mathcal{M}'} - \boldsymbol{\beta}^*_{\mathcal{M}'})^T \zeta_i \boldsymbol{x}_{i,\mathcal{M}'} - \underline{c}\pi_i (\boldsymbol{x}_{i,\mathcal{M}'}^T (\boldsymbol{\beta}_{\mathcal{M}'} - \boldsymbol{\beta}^*_{\mathcal{M}'}))^2\right] \\
\le&\left|\sum_{i=1}^n \pi_i (\boldsymbol{\beta}_{\mathcal{M}'} - \boldsymbol{\beta}^*_{\mathcal{M}'})^T \zeta_i \boldsymbol{x}_{i,\mathcal{M}'}\right| - \underline{c\pi}\lambda_{\min} w_1^2 n^{-2\eta_1},
\end{aligned}$$

where $\zeta_i = 2\omega_\tau(\boldsymbol{\beta}^*_{\mathcal{M}'})(y_i - \boldsymbol{x}_{i,\mathcal{M}'}^T \boldsymbol{\beta}^*_{\mathcal{M}'})$, $\lambda_{\min}$ is the smallest eigenvalue of $\boldsymbol{X}_{\mathcal{M}'}^T \boldsymbol{X}_{\mathcal{M}'}$, and $\underline{\pi}$ is the lower bound of $\{\pi_i, 1 \le i \le n\}$. By Lemma A3, we have

$$\begin{aligned}
&\mathbb{P}\{D_n(\boldsymbol{\beta}^*_{\mathcal{M}'}) - D_n(\boldsymbol{\beta}_{\mathcal{M}'}) \ge 0\} \\
\le&\mathbb{P}\left\{\left|\sum_{i=1}^n \pi_i (\boldsymbol{\beta}_{\mathcal{M}'} - \boldsymbol{\beta}^*_{\mathcal{M}'})^T \zeta_i \boldsymbol{x}_{i,\mathcal{M}'}\right| \ge \underline{\pi c}\lambda_{\min} w_1^2 n^{-2\eta_1}\right\} \\
\le&2\exp\left\{-\frac{\underline{c}^2 \underline{\pi}^2 \lambda_{\min}^2 w_1^4 n^{-4\eta_1}}{4K^2 \sum_{i=1}^n \pi_i^2 ((\boldsymbol{\beta}_{\mathcal{M}'} - \boldsymbol{\beta}^*_{\mathcal{M}'})^T \boldsymbol{x}_{i,\mathcal{M}'})^2}\right\} \\
\le&2\exp\left\{-\frac{\underline{c}^2 \underline{\pi}^2 \lambda_{\min}^2 w_1^2 n^{-4\eta_1}}{4K^2 \lambda_{\max} n^{-2\eta_1}}\right\},
\end{aligned}$$

which leads to

$$\begin{aligned}
&\mathbb{P}\{D_n(\boldsymbol{\beta}^*_{\mathcal{M}'}) \ge D_n(\boldsymbol{\beta}_{\mathcal{M}'})\} \\
\le&2\exp\left\{-\frac{\underline{c}^2 \underline{\pi}^2 \lambda_{\min}^2 w_1^2 n^{-4\eta_1}}{4K^2 \lambda_{\max} n^{-2\eta_1}}\right\}.
\end{aligned}$$

Thus, by Bonferroni inequality and Condition (A1),

$$\begin{aligned}
&\mathbb{P}\{D_n(\boldsymbol{\beta}^*_{\mathcal{M}'}) \ge \min\nolimits_{\mathcal{M}\in\mathcal{M}_-^k} D_n(\boldsymbol{\beta}_{\mathcal{M}'})\} \\
\le&\sum_{\mathcal{M}\in\mathcal{M}_-^k} \mathbb{P}\{D_n(\boldsymbol{\beta}^*_{\mathcal{M}'}) \ge D_n(\boldsymbol{\beta}_{\mathcal{M}'})\} \\
\le&2p^k \exp\{-bn^{1-2\eta_1}\} \\
\le&2\exp\{cw_2 n^{\eta_2} n^{\alpha} - bn^{1-2\eta_1}\} = o(1),
\end{aligned}$$

where b is some generic positive constant. Due to the convexity of $D_n(\boldsymbol{\beta}_{\mathcal{M}'})$ in $\boldsymbol{\beta}_{\mathcal{M}'}$, the above conclusion holds for any $\boldsymbol{\beta}_{\mathcal{M}'}$, such that $\|\boldsymbol{\beta}_{\mathcal{M}'} - \boldsymbol{\beta}^*_{\mathcal{M}'}\|_2 \ge w_1 n^{-\eta_1}$.

For any $\mathcal{M} \in \mathcal{M}_-^k$, let $\dot{\boldsymbol{\beta}}_{\mathcal{M}'}$ be $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ augmented with zero corresponding to the elements in $\mathcal{M}'/\mathcal{M}^*$. By Condition (A2), it is seen that

$$\|\dot{\boldsymbol{\beta}}_{\mathcal{M}'} - \boldsymbol{\beta}^*_{\mathcal{M}'}\|_2 \ge \|\boldsymbol{\beta}^*_{\mathcal{M}'/\mathcal{M}^*}\|_2 \ge w_1 n^{-\eta_1}.$$

Consequently,

$$\begin{aligned}
&\mathbb{P}\{\max\nolimits_{\mathcal{M}\in\mathcal{M}_+^k} D_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) \ge \min\nolimits_{\mathcal{M}\in\mathcal{M}_-^k} D_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\} \\
\le&\mathbb{P}\{D_n(\boldsymbol{\beta}^*_{\mathcal{M}'}) \ge \min\nolimits_{\mathcal{M}\in\mathcal{M}_-^k} D_n(\dot{\boldsymbol{\beta}}_{\mathcal{M}'})\} \\
=&o(1).
\end{aligned}$$

The theorem is proved. □

Proof of Theorem 3. Let $q = \min_{j\in\mathcal{M}^*}|\beta_j^*|$. By Condition (A2), $q = O(w_1 n^{-\eta_1})$, it suffices to prove for any $t \geq 0$

$$\mathbb{P}\Big\{\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty \leq q/2\Big\} \to 1.$$

It is clearly implied by

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_P(q). \tag{A3}$$

We use mathematical induction to prove (A3).

Step 1. When $t = 0$, the initial value for the iteration $\boldsymbol{\beta}^{(0)}$ is defined as the LASSO estimator of $\boldsymbol{\beta}$, that is $\boldsymbol{\beta}_L$. By Lemma A5, we have

$$\mathbb{P}\{\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_1 \leq 16v^{-1}\lambda k_0\} \to 1, \text{as } n \to \infty.$$

Under Condition (A2), we have $k_0 = O(n^{\eta_2})$, $q^{-1} = O(n^{\eta_1})$. By $\lambda = o(n^{-(\eta_1+\eta_2)})$, $\lambda k_0 = o(q)$. Hence, when $t = 0$, (A3) holds.

Step 2. Suppose (A3) is true for $t-1$, we show that (A3) is also true for t with $t \geq 1$. By the definition of $\boldsymbol{V}$, $\boldsymbol{\beta}^{(t)} = \boldsymbol{V}(\boldsymbol{\gamma}^{(t)}; k)$, where $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\beta}^{(t-1)} - u^{-1}\boldsymbol{T}_n(\boldsymbol{\beta}^{(t-1)})$, then

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty = \|\boldsymbol{V}(\boldsymbol{\gamma}^{(t)}; k) - \boldsymbol{\beta}^*\|_\infty \leq \|\boldsymbol{\gamma}^{(t)} - \boldsymbol{\beta}^*\|_\infty.$$

Thus, (A3) is proved by showing $\|\boldsymbol{\gamma}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_P(q)$. Note that

$$\|\boldsymbol{\gamma}^{(t)} - \boldsymbol{\beta}^*\|_\infty \leq \|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty + u^{-1}\|\boldsymbol{T}_n(\boldsymbol{\beta}^{(t-1)})\|_\infty.$$

Lemma A2 implies that for any $j \in \{1, \ldots, p\}$, we have

$$|T_{nj}(\boldsymbol{\beta}^{(t-1)}) - T_{nj}(\boldsymbol{\beta}^*)| \leq 2\bar{c}\sum_{i=1}^{n}|\boldsymbol{x}_i^T(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*)||\pi_i x_{ij}|,$$

which further implies

$$\begin{aligned}
&u^{-1}|T_{nj}(\boldsymbol{\beta}^{(t-1)})| \\
\leq &u^{-1}|T_{nj}(\boldsymbol{\beta}^*)| + 2u^{-1}\bar{c}\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty \max_{1\leq j\leq p}\left\{\sum_{i=1}^{n}\sum_{h\in\mathcal{M}_t}|x_{ij}x_{ih}|\right\} \\
\leq &u^{-1}\|T_n(\boldsymbol{\beta}^*)\|_\infty + 2u^{-1}nk\eta_3^2\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty,
\end{aligned}$$

where $\mathcal{M}_t = \mathcal{M}^{(t)} \cup \mathcal{M}^*$, $s(\mathcal{M}_t) \leq 2k$. Meanwhile, (A1) gives rise to $\|\boldsymbol{T}_n(\boldsymbol{\beta}^*)\|_\infty = O_P(n\lambda)$. By $\lambda n^{\eta_1+\eta_2} = o(1)$ and $u \geq \xi nk$, we get $u^{-1}\|\boldsymbol{T}_n(\boldsymbol{\beta}^*)\|_\infty = o_P(q)$, due to

$$2u^{-1}nk\eta_3^2\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty \leq 2(\xi nk)^{-1}nk\eta_3^2\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty = o_P(q),$$

so we have $\|\boldsymbol{\gamma}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_P(q)$. Thus the second step of the mathematical induction is completed. The theorem is therefore proved. □

References

1. Frank, L.E.; Friedman, J.H. A statistical view of some chemometrics regression tools. *Technometrics* **1993**, *35*, 109–135. [CrossRef]
2. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
3. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
4. Candes, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35*, 2313–2351.
5. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2008**, *70*, 849–911. [CrossRef]

6. Fan, J.; Samworth, R.; Wu, Y. Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **2009**, *10*, 2013–2038.
7. Zhu, L.P.; Li, L.; Li, R.; Zhu, L.X. Model-free feature screening for ultrahigh-dimensional data. *J. Am. Stat. Assoc.* **2011**, *106*, 1464–1475. [CrossRef]
8. Li, R.; Zhong, W.; Zhu, L. Feature screening via distance correlation learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1129–1139. [CrossRef]
9. Fan, J.; Song, R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Stat.* **2010**, *38*, 3567–3604. [CrossRef]
10. Wang, H. Forward regression for ultra-high dimensional variable screening. *J. Am. Stat. Assoc.* **2009**, *104*, 1512–1524. [CrossRef]
11. Xu, C.; Chen, J. The sparse MLE for ultrahigh-dimensional feature screening. *J. Am. Stat. Assoc.* **2014**, *109*, 1257–1269. [CrossRef] [PubMed]
12. Koenker, R. *Quantile Regression*; Cambridge University Press: New York, NY, USA, 2005.
13. Newey, W.K.; Powell, J.L. Asymmetric least squares estimation and testing. *Econom. J. Econom. Soc.* **1987**, *55*, 819–847. [CrossRef]
14. Zhao, J.; Chen, Y.; Zhang, Y. Expectile regression for analyzing heteroscedasticity in high dimension. *Stat. Probab. Lett.* **2018**, *137*, 304–311. [CrossRef]
15. Ciuperca, G. Variable selection in high-dimensional linear model with possibly asymmetric errors. *Comput. Stat. Data Anal.* **2021**, *155*, 107112. [CrossRef]
16. Song, S.; Lin, Y.; Zhou, Y. Linear expectile regression under massive data. *Fundam. Res.* **2021**, *1*, 574–585. [CrossRef]
17. Gu, Y.; Zou, H. High-dimensional generalizations of asymmetric least squares regression and their applications. *Ann. Stat.* **2016**, *44*, 2661–2694. [CrossRef]
18. Wang, L.; Wu, Y.; Li, R. Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Am. Stat. Assoc.* **2012**, *107*, 214–222. [CrossRef]
19. He, X.; Wang, L.; Hong, H.G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Stat.* **2013**, *41*, 342–369. [CrossRef]
20. Wu, Y.; Yin, G. Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **2015**, *102*, 65–76. [CrossRef]
21. Zhong, W.; Zhu, L.; Li, R.; Cui, H. Regularized quantile regression and robust feature screening for single index models. *Stat. Sin.* **2016**, *26*, 69–95. [CrossRef]
22. Ma, Y.; Li, Y.; Lin, H. Concordance measure-based feature screening and variable selection. *Stat. Sin.* **2017**, *27*, 1967–1985. [CrossRef]
23. Chen, L.P. A note of feature screening via a rank-based coefficient of correlation. *Biom. J.* **2023**, *65*, 2100373. [CrossRef]
24. Chen, L.P. Feature screening via concordance indices for left-truncated and right-censored survival data. *J. Stat. Plan. Inference* **2024**, *232*, 106153. [CrossRef]
25. Tu, Y.; Wang, S. Variable screening and model averaging for expectile regressions. *Oxf. Bull. Econ. Stat.* **2023**, *85*, 574–598. [CrossRef]
26. Ghosh, A.; Ponzi, E.; Sandanger, T.; Thoresen, M. Robust sure independence screening for nonpolynomial dimensional generalized linear models. *Scand. J. Stat.* **2023**, *50*, 1232–1262. [CrossRef]
27. Ghosh, A.; Thoresen, M. A robust variable screening procedure for ultra-high dimensional data. *Stat. Methods Med Res.* **2021**, *30*, 1816–1832. [CrossRef]
28. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559. [CrossRef]
29. Scott, D.W. Parametric statistical modeling by minimum integrated square error. *Technometrics* **2001**, *43*, 274–285. [CrossRef]
30. Efron, B. Regression percentiles using asymmetric squared error loss. *Stat. Sin.* **1991**, *1*, 93–125.
31. Mallat, S.G.; Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **1993**, *41*, 3397–3415. [CrossRef]
32. Blumensath, T.; Davies, M.E. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **2009**, *27*, 265–274. [CrossRef]
33. Chen, J.; Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **2008**, *95*, 759–771. [CrossRef]
34. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer-Verlag: New York, NY, USA, 2013.
35. Fan, J.; Ma, Y.; Dai, W. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Am. Stat. Assoc.* **2014**, *109*, 1270–1284. [CrossRef] [PubMed]
36. Wang, M.; Kang, X.; Liang, J.; Wang, K.; Wu, Y. Heteroscedasticity identification and variable selection via multiple quantile regression. *J. Stat. Comput. Simul.* **2024**, *94*, 297–314. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.