Statistics and Its Interface Volume 17 (2024) 107–125

Asymptotic properties of relative error estimation for accelerated failure time model with divergent number of parameters

Fei Ye, Hongyi Zhou, and Ying Yang*

The paper considers the problem of parameter estimation in the accelerated failure time model with divergent number of parameters under fixed design. We propose an estimator based on the general relative error criterion. We show that the proposed estimator is consistent and asymptotically normal under mild regular conditions. We also propose a variable selection procedure and show its oracle property as well as the consistency of model selection. Numerical studies have been conducted to compare the performance of different general relative error based estimators.

AMS 2000 subject classifications: Primary 62F12; secondary 62J99.
Keywords and phrases: General relative error, Accelerated failure time model, Divergent number of parameters, Variable selection.

1. INTRODUCTION

Consider the following multiplicative model or accelerated failure time (AFT) model

$$y_k = \exp\big(x_k^T \beta\big) \cdot \varepsilon_k, \quad k = 1, \dots, n,$$

where x_k is a p-dimensional fixed designed covariate, y_k is the corresponding response variable, ε_k is the unobservable independent and identically distributed random error, β is the unknown regression coefficient.

Obviously, the AFT model can be changed into linear model after taking logarithm transformation. Traditionally, M-estimator, including least square estimator and least absolute estimator, is widely used to estimate the unknown parameters in the linear model. The asymptotic theory including weak and strong consistency and asymptotic normality of M-estimators in linear model has been comprehensively and systematically studied in the monograph [15]. [19] further developed the anaylysis of variance type methods based on least absolute deviation. Many other estimation methods have also been developed to handle different scenarios, including high-dimensional cases or situations involving censoring, see [25, 26, 27, 28] and references therein.

The above-mentioned estimation method mostly based on criteria related to the scale of absolute error, including the least square (LS) and the least absolute deviation (LAD). However, in many practical applications, particularly when analyzing heteroscedastic data, absolute error based method is not appropriate. For instance, in stock price prediction, higher share price may require less accuracy in terms of absolute error. Similarly, in lifetime data analysis, the accuracy requirements for prediction in terms of absolute error vary across different lifespans. Therefore, relative error criterion may emerge as a more suitable alternative due to its free to scale and robust to outliers. There are a number of studies considering the relative errors in the literature. For instance, [22] proposed an absolute relative error based estimation in linear model. [23] proved the strong consistency of the estimators minimizing squared relative errors and absolute relative errors. [24] derived the closed form of the best mean squared relative error prediction.

Recently, [6] proposed the least absolute relative error (LARE) based on the sum of two different types of relative error. There are also studies of estimators based on relative error criterion like minimum relative errors (MRE), relative least squares (RLS) and least product relative error (LPRE) see [21]. [11] proposed the general relative error criterion (GREC), including LARE, RLS, LPRE as special cases. [5] proposed an estimator for quantile model based on general relative error. [12] and [14] extended the multiplication regression model to partially linear and single index model respectively.

In recent years, some researchers were interested in the AFT model, whose number of parameters tends to infinity (abbreviated as: divergent-dimensional AFT model). The divergent-dimensional AFT model with response variable y_k, covariates x_k, unknown parameter β_n^* and error ε_k is represented by

$$y_k = \exp\big(x_k^T \beta_n^*\big) \cdot \varepsilon_k, \quad k = 1, \dots, n, \tag{1}$$

where x_k is a p_n-dimensional covariate whose dimension may vary with the sample size n. [10] studied the large sample theory of absolute relative errors model under high-dimensional settings. [9] proposed a general relative error criterion based estimation using empirical likelihood under divergent dimensional setting. Their works mainly focused

*Corresponding author.

on the hypothesis testing procedure of the unknown parameter, instead of estimating of the parameters. [7] studied the asymptotic properties of general relative error based estimators under random design when the loss function is convex.

[11] proposed a general relative error criterion for estimating the unknown parameter in the AFT model and they revealed the connection between relative error estimators and the M-estimation in the linear model. Using this connection, the asymptotic properties of many relative error estimators can be established in a unified way by the well-developed M-estimation theories [15].

The major contributions of this paper are two-fold. First, we extend the general relative error criterion to the estimation of AFT model with divergent number of parameters under fixed design. Second, we propose a model selection procedure and study its theoretical properties.

We estimate the unknown parameter β_n^* in (1) by minimizing the following general relative errors proposed by [11]:

$$\psi_n(\beta) = \sum_{k=1}^{n} \rho_0\left(\left|\frac{y_k - \exp(x_k^T\beta)}{y_k}\right|, \left|\frac{y_k - \exp(x_k^T\beta)}{\exp(x_k^T\beta)}\right|\right),$$

where $\rho_0(\cdot,\cdot)$ is a nonnegative loss function, more concrete examples see [11]. Taking the logarithm of the AFT model (1) yields a linear model

$$\log y_k = x_k^T\beta_n^* + \epsilon_k, \quad k = 1, \ldots, n, \tag{2}$$

where $\epsilon_k = \log\varepsilon_k$, independent and identical distributed with $\epsilon = \log\varepsilon$. Thus, the objective function becomes

$$\psi_n(\beta) = \sum_{k=1}^{n} \rho\big(\log y_k - x_k^T\beta\big),$$

where $\rho(t) = \rho_0(1 - \exp(-t), \ \exp(t) - 1)$. Different from [7], our restrictions on the loss function are much weaker, for instance, we do not require convexity of the loss function. We establish the $\sqrt{n/p_n}$-consistency and asymptotic normality of the estimator in a more general case. Moreover, we propose a consistent estimation of covariance matrix.

Variable selection is a fundamental problem in statistic modeling. [10] studied the variable selection procedure of least absolute relative error model. [20] considered a least product relative error based procedure for variable selection with fixed or an increasing number of parameters in regression model. [16] proposed a nonconcave penalized M-estimation of the least absolute relative error to deal with sparse AFT model. [9] developed the inference of parameter by using GREC and empirical likelihood. In this paper, we develop an alternative variable selection procedure which is different from [9] by minimizing penalized general relative error, that is,

$$\begin{aligned} Q_n(\beta) &= \frac{1}{n}\sum_{k=1}^{n} \rho_0\left(\left|\frac{y_k - \exp(x_k^T\beta)}{y_k}\right|, \left|\frac{y_k - \exp(x_k^T\beta)}{\exp(x_k^T\beta)}\right|\right) \\ &\quad + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|) \\ &= \frac{1}{n}\sum_{k=1}^{n} \rho\big(\log y_k - x_k^T\beta\big) + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|), \end{aligned}$$

where $p_{\lambda_n}(\cdot)$ is a nonnegative penalty function which depends on sample size n. It is worth noting that our proposed methodology is general enough to cover many existing relative error based selection methods. For instance, if we take $\rho(t) = |e^t - e^{-t}|$ with a Lasso-type penalty in (5.1), then our method is equivalent to that in [10]. If we take $\rho(t) = |e^t + e^{-t} - 2|$ with an adaptive Lasso-type penalty in (5.1), then our method corresponds to that in [20]. We also provide theoretical supports for our proposed method including the oracle properties and consistency of model selection.

The rest of the paper is organized as follows. Section 2 states the regularity conditions of the model. In Section 3, asymptotic properties, including consistency and asymptotic normalities, of estimators based on GREC are developed. In Section 4, we provide estimators based on two relative error criteria. Simulation studies are reported in Section 5. In Section 6, we propose a variable selection method and develop its oracle properties and consistency. All the proofs are deferred to Appendix.

2. REGULARITY CONDITIONS

To obtain the large-sample properties of the parameter estimators, we require the following conditions on the dimension, covariates and loss function etc.

2.1 Regularity conditions on dimension and covariates

In the following, we always assume that

- The number of parameters $p_n \to \infty$;
- The matrix $S_n = \sum_{k=1}^n x_k x_k^T$ is invertible.

Let the spectral decomposition of S_n be $S_n = Q_n\Lambda_n Q_n^T$, and denote $S_n^{1/2} = \Lambda_n^{1/2}Q_n^T$, $S_n^{-1/2} = (S_n^{1/2})^{-1} = Q_n\Lambda_n^{-1/2}$,

$$d_n = \max_{1\le k\le n} x_k^T S_n^{-1} x_k.$$

Consider the following regularity conditions on p_n and d_n.

(A0) $p_n/n \to 0$;
(A1) $d_n \to 0$;
(A2) $d_n = o(1/p_n)$;
(A3) $d_n = o(1/p_n^4)$.

The relationship between (A1)–(A3) is obvious, and the implication of (A0) from (A1) is introduced in Lemma A.1 later.

2.2 Regularity conditions on loss function and random error

Deriving the asymptotic properties of the estimators requires the following regularity conditions on the loss function $\rho(\cdot)$ and random error ϵ.

(B1) $\rho(t) \in C^2(\mathbb{R})$ is second order continuously differentiable;

(B2) $E[\rho'(\epsilon)] = 0, \quad E[\rho'(\epsilon)]^2 = \sigma^2 < \infty$;

(B3) $E[\rho''(\epsilon)] = \tau > 0, \quad E[\rho''(\epsilon)]^2 < \infty$;

(B4) $E\{\sup_{s:|s|\le t} |\rho''(\epsilon+s) - \rho''(\epsilon)|^2\} = O(t^2)$.

(B5) $E|\rho'(\epsilon)|^4 < \infty$, and

$$E\Big\{\sup_{s:|s|\le t} \big(\big|\rho'(\epsilon+s)\big|^2 - \big|\rho'(\epsilon)\big|^2\big)^2\Big\} = O\big(t^2\big).$$

(B1) is the requirement for the smoothness of the loss function, which is weaker than the one of [17] for the third-order differentiability of the likelihood function, and does not contain the convexity required by [7] and [8]. (B2) and (B3) ensure the identifiability of the estimates. (B4), on the other hand, requires some uniform continuity in its second-order differentiation, for example, (B4) naturally holds when $\rho''(\cdot)$ is a Lipschitz continuous function, and as we will see later, (B4) is mild for the loss functions induced from the relative error criteria. (B5) ensures the consistency of covariance estimator.

3. ASYMPTOTIC PROPERTIES OF THE PARAMETER ESTIMATORS

Unless otherwise mentioned, all limits as n goes to infinity.

The main conclusions about the asymptotic properties of the regression estimators and covariance estimators are as follows.

Theorem 3.1. *For the divergent-dimensional AFT model (1), if the conditions (A3) on p_n and d_n, (B1)–(B4) on $\rho(\cdot)$ and ϵ hold, then*

(i) *(Consistency) $\hat\beta_n$, the local minimal point of $\psi_n(\beta)$, exists and satisfies*

$$\big\|S_n^{1/2}\big(\hat\beta_n - \beta_n^*\big)\big\| = O_p(\sqrt{p_n}),$$

where $\|x\| = (\sum_j x_j^2)^{\frac{1}{2}}$ and β_n^* is the true value of the parameter.

(ii) *(Asymptotic normality) For a known numerical matrix A_n with order $m \times p_n$, if $A_n A_n^T$ converges to a positive definite symmetric matrix G with order $m \times m$, i.e. $A_n A_n^T \to G > 0$, then*

$$A_n S_n^{1/2}\big(\hat\beta_n - \beta_n^*\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \tau^{-2}\sigma^2 G\big).$$

For statistical inference needs, the asymptotic covariance matrix needs to be estimated too, which leads to the estimation of τ and σ^2. Denoted $e_k = \log y_k - x_k^T\hat\beta_n$ as residuals, the classical simple estimators of τ and σ^2 are given by [18] $\hat\tau = \frac{1}{n}\sum_{k=1}^n \rho''(e_k)$ and $\hat\sigma^2 = \frac{1}{n-p_n}\sum_{k=1}^n |\rho'(e_k)|^2$, respectively, but their asymptotic properties are unclear. To this end, we propose two new reweighted estimates

$$\hat\tau_n = \frac{1}{p_n}\sum_{k=1}^n \rho''(e_k) \cdot x_k^T S_n^{-1} x_k,$$

$$\hat\sigma_n^2 = \frac{1}{p_n}\sum_{k=1}^n \big|\rho'(e_k)\big|^2 \cdot x_k^T S_n^{-1} x_k.$$

For notation simplicity, we drop the subscript n of $\hat\tau_n$ and $\hat\sigma_n^2$. Their consistencies are as follows.

Theorem 3.2. *Under the conditions of Theorem 3.1,*

(i) $\hat\tau \xrightarrow{p} \tau$,

(ii) *Furthermore, if condition (B5) holds, then $\hat\sigma^2 \xrightarrow{p} \sigma^2$.*

4. ESTIMATORS BASED ON TWO RELATIVE ERROR CRITERIA

In this section, we examine estimators based on two relative error criteria: the Relative Least Squares (RLS) estimator and the Least Product Relative Errors (LPRE) estimator [13]. Their pre-transformed loss functions are $(\frac{y-\exp(x^T\beta)}{y})^2$ for RLS, and $|\frac{y-\exp(x^T\beta)}{y}| \times |\frac{y-\exp(x^T\beta)}{\exp(x^T\beta)}|$ for LPRE. The loss functions after transformation are $\rho(t) = (1-e^{-t})^2$ for RLS, and $\rho(t) = e^t + e^{-t} - 2$ for LPRE respectively. Denote $M(t) = Ee^{t\epsilon} = E\varepsilon^t$, which is the moment-generating function of ϵ. Using Theorem 3.1 and Theorem 3.2, we have the following two corollaries.

Corollary 4.1. *For the RLS estimator under the relative least squares criterion for the divergent-dimensional AFT model (1), if the condition (A3) holds, $M(-4) < +\infty$ and $M(-1) = M(-2)$, then the conclusions of Theorem 3.1 hold, where $\tau = 2M(-1)$ and $\sigma^2 = 4[M(-4) - 2M(-3) + M(-2)]$. Furthermore, if $M(-8) < +\infty$, then conclusions of Theorem 3.2 also hold.*

Corollary 4.2. *For the LPRE estimator under the relative least product absolute relative errors criterion for the dimension-changeable AFT model (1), if the condition (A3) holds, $M(\pm 2) < +\infty$ and $M(1) = M(-1)$, then the conclusions of Theorem 3.1 hold, where $\tau = 2M(1)$ and $\sigma^2 = M(2) + M(-2) - 2$. Furthermore, if $M(\pm 4) < +\infty$, then conclusions of Theorem 3.2 also hold.*

5. VARIABLE SELECTION UNDER GENERALIZED RELATIVE ERROR CRITERION

In this section, we study the penalized variable selection method of AFT model under general relative error criterion. We introduce regularity conditions, prove the oracle properties of parameter estimators and give an estimation of the covariance matrix.

5.1 Model and parameter estimation

Consider the AFT model (1) or its linear form (2). Assume the p_n-dimensional covariate β_n^* only has first q_n elements which are nonzero. Variables are selected by minimizing the following objective function:

$$Q_n(\beta) = \frac{1}{n}\sum_{k=1}^{n} \rho\big(\log y_k - x_k^T\beta\big) + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|).$$

The first term is the average loss and the second term is the penalty function controlling the shrinkage amount of estimators. λ_n is a preset threshold which represents the strength of penalty.

5.2 Regularity conditions

5.2.1 Regularity conditions on dimension and covariates

In this section, we further assume:

(A4) There are two positive constants r and R such that

$$0 < r < \Omega_{\min}(S_n/n) < \Omega_{\max}(S_n/n) < R < \infty,$$

where $\Omega_{\min}(S_n/n)$ and $\Omega_{\max}(S_n/n)$ are the largest and smallest eigenvalue of matrix S_n/n.

(A5) $p_n^2/n \to 0$.

5.2.2 Regular conditions on penalty function and threshold

Consider the following regularity conditions: Denote

$$\begin{aligned} a_n &= \max_{1\le j\le p_n} \big\{\big|p'_{\lambda_n}\big(|\beta_n^{*(j)}|\big)\big|,\ \beta_n^{*(j)} \neq 0\big\},\\ b_n &= \max_{1\le j\le p_n} \big\{\big|p''_{\lambda_n}\big(|\beta_n^{*(j)}|\big)\big|,\ \beta_n^{*(j)} \neq 0\big\},\\ c_n &= \min_{1\le j\le p_n} \big\{|\beta_n^{*(j)}|,\ \beta_n^{*(j)} \neq 0\big\}, \end{aligned}$$

and denote by $r_n = q_n/p_n$ be the proportion of nonzero elements in β_n^*. Consider the following regularity conditions:

(C1) $a_n = O(1/\sqrt{nr_n})$;
(C2) $b_n \to 0$;
(C3) $\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$;
(C4) $\sqrt{p_n/n} = o(\lambda_n)$;
(C5) $\lambda_n = o(c_n)$;
(C6) There exists two positive constants C and D such that if $\theta_1, \theta_2 > C\lambda_n$, then

$$\big|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)\big| \le D|\theta_1 - \theta_2|;$$

(C7) $a_n = o(1/\sqrt{nq_n})$;
(C8) $b_n = o(1/\sqrt{p_n})$.

Conditions (C1) and (C2) ensure the existence and $\sqrt{n/p_n}$-consistency of the estimator. Condition (C3) requires the singularity of penalty function at zero. (C3) and (C4) ensure the model is able to select the true model with probability tends to 1. Condition (C4) requries the rate of convergence of λ_n to zero not faster than $\sqrt{p_n/n}$ and condition (C5) requires the convergence rate not slower than c_n. Condition (C6) imposes restrictions on the smoothness of penalty function, together with condition (C5) ensure the asymptotic normality of the estimator. If we further assume (C7) and (C8), then the bias and variance caused by penalty function can be neglected, and the efficiency of the estimators follows, that is, the estimator has the same asymptotic distribution as it is under the true model.

5.3 Oracle properties of parameter estimation

Let $\bar{x}_k = (x_k^{(1)}, \dots, x_k^{(q_n)})^T$ be the first q_n elements of x_k. Denote

$$\begin{aligned} S_{n1} &= \sum_{k=1}^{n} \bar{x}_k \bar{x}_k^T,\\ \Sigma_{\lambda_n} &= \operatorname{diag}\big\{p''_{\lambda_n}\big(|\beta_n^{*(1)}|\big), \dots, p''_{\lambda_n}\big(|\beta_n^{*(q_n)}|\big)\big\},\\ \mathbf{b}_{\lambda_n} &= \big\{p'_{\lambda_n}\big(|\beta_n^{*(1)}|\big)\operatorname{sgn}\big(\beta_n^{*(1)}\big), \dots,\\ &\qquad p'_{\lambda_n}\big(|\beta_n^{*(q_n)}|\big)\operatorname{sgn}\big(\beta_n^{*(q_n)}\big)\big\}^T, \end{aligned}$$

where $\operatorname{sgn}(\cdot)$ represents the sign function. The main theorem of oracle properties are as follows.

Theorem 5.1. *Assume $\beta_n^* = \binom{\beta_{n1}^*}{0}$ has only q_n non-zero elements. Then, under conditions (A3)–(A4), (B1)–(B4) and (C1)–(C2), we have*

(i) *(Consistency of estimation) $Q_n(\beta)$ has a local minimum $\hat{\beta}_n$ satisfying*

$$\|\hat{\beta}_n - \beta_n^*\| = O_p\bigg(\sqrt{\frac{p_n}{n}}\bigg).$$

(ii) *(Consistency of model selction) If we further assume (C3)–(C4), then the above $\sqrt{n/p_n}$-consistent estimator $\hat{\beta}_n = \binom{\hat{\beta}_{n1}}{\hat{\beta}_{n2}}$ select the true model with probability tends to 1, that is,*

$$\mathbb{P}\{\hat{\beta}_{n2} = 0\} \to 1.$$

(iii) *(Asymptotic normality) If we further assume conditions (A5) and (C5)–(C6) hold and $B_nB_n^T$ converges to a $m \times m$-dimensional positive-definite matrix G, that is, $B_nB_n^T \to G > 0$ for $m \times q_n$-dimensional numerical matrix B_n, then*

$$B_n\Sigma_n S_{n1}^{1/2}\big(\hat{\beta}_{n1} - \beta_{n1}^* + \delta_n^*\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0,\, \tau^{-2}\sigma^2 G\big),$$

where

$$\begin{aligned} \Sigma_n &= I_{q_n} + n\tau^{-1}\big(S_{n1}^{-1/2}\big)^T\Sigma_{\lambda_n}\big(S_{n1}^{-1/2}\big),\\ \delta_n^* &= [\tau S_{n1}/n + \Sigma_{\lambda_n}]^{-1}\mathbf{b}_{\lambda_n}, \end{aligned}$$

are the additional variance and bias caused by the penalty term respectively. Furthermore, if conditions

(C7)–(C8) holds, then the estimation is efficient, which means:

$$B_n S_{n1}^{1/2}\big(\hat{\beta}_{n1}-\beta_{n1}^*\big)\xrightarrow{\mathcal{D}}\mathcal{N}\big(0,\tau^{-2}\sigma^2 G\big).$$

5.4 Estimation of covariance matrix and bias

Denote the covariance matrix and bias of $\hat{\beta}_{n1}$ by

$$\begin{aligned}\mathrm{Cov}_n &= \tau^{-2}\sigma^2\big(S_{n1}^{-1/2}\big)^T\Sigma_n^{-2}\big(S_{n1}^{-1/2}\big),\\ \delta_n^* &= [\tau S_{n1}/n+\Sigma_{\lambda_n}]^{-1}\mathbf{b}_{\lambda_n}\\ &= n\tau^{-1}\big(S_{n1}^{-1/2}\big)\Sigma_n^{-1}\big(S_{n1}^{-1/2}\big)^T\mathbf{b}_{\lambda_n},\end{aligned}$$

and their corresponding estimators by

$$\begin{aligned}\widehat{\mathrm{Cov}}_n &= \tau^{-2}\sigma^2\big(S_{n1}^{-1/2}\big)^T\widehat{\Sigma}_n^{-2}\big(S_{n1}^{-1/2}\big),\\ \hat{\delta}_n^* &= n\tau^{-1}\big(S_{n1}^{-1/2}\big)\widehat{\Sigma}_n^{-1}\big(S_{n1}^{-1/2}\big)^T\hat{\mathbf{b}}_{\lambda_n},\end{aligned}$$

where

$$\begin{aligned}\widehat{\Sigma}_n &= I_{q_n}+n\tau^{-1}\big(S_{n1}^{-1/2}\big)^T\widehat{\Sigma}_{\lambda_n}\big(S_{n1}^{-1/2}\big),\\ \widehat{\Sigma}_{\lambda_n} &= \mathrm{diag}\big\{p''_{\lambda_n}\big(|\hat{\beta}_n^{(1)}|\big),\ldots,p''_{\lambda_n}\big(|\hat{\beta}_n^{(q_n)}|\big)\big\},\\ \hat{\mathbf{b}}_{\lambda_n} &= \big\{p'_{\lambda_n}\big(|\hat{\beta}_n^{(1)}|\big)\,\mathrm{sgn}\big(\hat{\beta}_n^{(1)}\big),\ldots,p'_{\lambda_n}\big(|\hat{\beta}_n^{(q_n)}|\big)\,\mathrm{sgn}\big(\hat{\beta}_n^{(q_n)}\big)\big\}^T.\end{aligned}$$

We first point out that $\|\widehat{\mathrm{Cov}}_n\| = \|\mathrm{Cov}_n\| = O_p(\frac{1}{n})$, $\|\hat{\delta}_n^*\| = \|\delta_n^*\| = O_p(\sqrt{\frac{p_n}{n}})$, which can be obtained by A.5. Furthermore, we have

Theorem 5.2. *Under condition (A3)–(A5), (B1)–(B4) and (C1)–(C6), the estimation of covariance matrix and bias has the following consistency:*

(i) $\|Cov_n-\widehat{Cov}_n\| = o_p(\frac{1}{n})$;
(ii) $\|\delta_n^*-\hat{\delta}_n^*\| = o_p(\sqrt{\frac{p_n}{n}})$.

The proof will be given in the Appendix A. The above result mainly focused on the asymptotic properties of the proposed estimator. As suggested by an anonymous reviewer, we also derive the oracle inequality of the estimator under mild conditions. We present the technical conditions and main results in Appendix B.

6. SIMULATION

In this section, we evaluate the performance of the three estimators LS, RLS and LPRE through random simulations. The simulations are set as follows: the sample sizes n are taken to be 100, 200, 400 and 800, and the dimensions of the covariates p_n are 7, 8, 10 and 12, respectively. The covariates are designed as $x_k^T = (1, k/n, z_k^T)$, where z_k is a simple random sample from a p_n-2 dimensional normal distribution $\mathcal{N}(0,\Sigma)$, where $\Sigma(i,j) = 0.5^{|i-j|}$ and the regression coefficients of the covariates are set as $\beta_j = \sqrt{j}-1$. Then we consider three types of error distributions:

- $\epsilon\sim\mathcal{N}(0,1)$, the standard normal distribution;
- $\epsilon\sim U(-2,2)$, a uniform distribution over the interval $[-2,2]$;
- $\epsilon\sim f^*(t)\propto\exp(-\{e^t+e^{-t}\})$, which makes LPRE efficient as MLE.

Finally we generate the response variables through the model and use three methods to estimate parameters and their covariance matrix. We repeat this simulation 500 times and part of the results are summarized in the following Table 1 and more simulation results are deferred to Appendix C in Tables 2 and 3. The simulation results show that, except for the two RLS estimates under normal and uniform error distributions, as the sample size n and parameter dimension p_n increase, we have (1) $\|\hat{\beta}_n-\beta_n^*\|^2$ and its standard deviation both approach to zero, which supports the consistency of estimators; (2) the estimates of τ, σ^2 and $\tau^{-2}\sigma^2$ are all becoming more accurate and their standard deviations are all becoming smaller as sample size n increases, which supports the consistency of $\hat{\tau}$, $\hat{\sigma}^2$ and $\hat{\tau}^{-2}\hat{\sigma}^2$ (3) the estimates bias for each component β_j of the regression coefficient, biases are all around 0, indicating that the parameter estimates are asymptotically unbiased; estimated empirical standard deviation, the mean of the estimated standard deviation and theoretical asymptotic standard deviation are very close to each other, indicating that the variance estimates are correctly available.

While the RLS estimation is more complicated, we find (1) the estimates of regression parameters and variances are very unstable under normal error distribution, (2) under the uniform error distribution, the estimates of the regression coefficients are still unstable, but the estimates of the variances are shown to be accurate, (3) under the f^* error distribution, the estimation of both its regression coefficient and variance are accurate. This indicates that the RLS estimation, despite its theoretical large sample properties, is very sensitive to the error distribution in the case of limited sample size; for those inappropriate error distributions, a very large sample size may be required to ensure a better estimate. The LS and LPRE estimation, on the other hand, are more stable in the simulations, and they are optimal for each white effective error distribution, but the LPRE estimates perform better under uniform error distributions.

Simulations are also conducted to compare the performance of different variable selection methods. We examines three kind of loss function:

1. $\rho(t)=t^2$, denoted as "LS",
2. $\rho(t)=e^t+e^{-t}-2$, denoted as "LPRE",
3. $\rho(t)=e^{0.4t}+e^{-0.4t}-2$, denoted as "GREC".

The tuning parameter λ is chosen according to a fivefold cross-validation (denoted as "CV") procedure or minimizing the Bayesian information criterion (denoted as "BIC"). The CV approach is as follows: Denote the full dataset by $\mathcal{D}$. We randomly divide $\mathcal{D}$ into five approximately equal-sized subsets $\mathcal{D}_1,\ldots,\mathcal{D}_5$ and denote the training and validation sets by $\mathcal{D}-\mathcal{D}_v$ and $\mathcal{D}_v$ respectively, where $v=1,2,\ldots,5$.

Table 1. Estimates of the main model statistics.

	n	p_n	$\|\|\hat{\beta}_n - \beta_n^*\|\|^2$		$\hat{\tau}$		$\hat{\sigma}^2$		$\hat{\tau}^{-2}\hat{\sigma}^2$	
error distribution			$\epsilon \sim \mathcal{N}(0,1)$							
LS					$\tau = 2.00$		$\sigma^2 = 4.00$		$\tau^{-2}\sigma^2 = 1.00$	
	100	7	0.24	(0.23)	–		3.65	(0.61)	0.91	(0.15)
	200	8	0.14	(0.13)	–		3.76	(0.43)	0.94	(0.11)
	400	10	0.07	(0.05)	–		3.87	(0.29)	0.97	(0.07)
	800	12	0.04	(0.03)	–		3.93	(0.21)	0.98	(0.05)
RLS					$\tau = 0.74$		$\sigma^2 = 23.03$		$\tau^{-2}\sigma^2 = 42.5$	
	100	7	5.31	(50.1)	1.05	(0.19)	0.97	(0.30)	0.89	(0.33)
	200	8	15.7	(191)	0.95	(0.16)	1.36	(0.37)	1.53	(0.56)
	400	10	8.14	(102)	0.90	(0.14)	1.86	(0.52)	2.37	(0.92)
	800	12	30.8	(326)	0.84	(0.12)	2.62	(0.68)	3.72	(1.33)
LPRE					$\tau = 3.30$		$\sigma^2 = 12.78$		$\tau^{-2}\sigma^2 = 1.18$	
	100	7	0.26	(0.24)	3.15	(0.24)	9.48	(3.82)	0.92	(0.22)
	200	8	0.15	(0.14)	3.19	(0.17)	10.30	(2.98)	0.99	(0.18)
	400	10	0.08	(0.06)	3.24	(0.12)	11.18	(2.41)	1.06	(0.16)
	800	12	0.05	(0.03)	3.27	(0.09)	11.95	(2.04)	1.11	(0.14)
error distribution			$\epsilon \sim U(-2,2)$							
LS					$\tau = 2.00$		$\sigma^2 = 5.33$		$\tau^{-2}\sigma^2 = 1.33$	
	100	7	0.35	(0.31)	–		4.79	(0.57)	1.20	(0.14)
	200	8	0.18	(0.15)	–		5.02	(0.40)	1.25	(0.10)
	400	10	0.09	(0.07)	–		5.14	(0.26)	1.28	(0.06)
	800	12	0.06	(0.04)	–		5.21	(0.19)	1.30	(0.05)
RLS					$\tau = 0.96$		$\sigma^2 = 0.60$		$\tau^{-2}\sigma^2 = 0.64$	
	100	7	20.3	(351)	1.02	(0.16)	0.55	(0.16)	0.52	(0.16)
	200	8	9.83	(115)	0.99	(0.12)	0.58	(0.13)	0.59	(0.14)
	400	10	27.5	(297)	0.97	(0.13)	0.59	(0.11)	0.61	(0.12)
	800	12	33.1	(356)	0.97	(0.12)	0.60	(0.09)	0.63	(0.10)
LPRE					$\tau = 3.63$		$\sigma^2 = 11.64$		$\tau^{-2}\sigma^2 = 0.89$	
	100	7	0.25	(0.23)	3.46	(0.19)	10.40	(1.73)	0.86	(0.06)
	200	8	0.13	(0.11)	3.53	(0.13)	10.94	(1.17)	0.87	(0.03)
	400	10	0.07	(0.05)	3.57	(0.08)	11.25	(0.75)	0.88	(0.02)
	800	12	0.04	(0.03)	3.59	(0.06)	11.38	(0.55)	0.88	(0.01)
error distribution			$\epsilon \sim f^*(t)$							
LS					$\tau = 2.00$		$\sigma^2 = 1.66$		$\tau^{-2}\sigma^2 = 0.41$	
	100	7	0.11	(0.10)	–		1.52	(0.23)	0.38	(0.06)
	200	8	0.06	(0.05)	–		1.58	(0.17)	0.39	(0.04)
	400	10	0.03	(0.02)	–		1.61	(0.12)	0.40	(0.03)
	800	12	0.02	(0.01)	–		1.63	(0.08)	0.41	(0.02)
RLS					$\tau = 1.35$		$\sigma^2 = 2.21$		$\tau^{-2}\sigma^2 = 1.21$	
	100	7	0.24	(0.22)	1.42	(0.15)	0.92	(0.29)	0.47	(0.18)
	200	8	0.13	(0.11)	1.40	(0.11)	1.19	(0.33)	0.62	(0.20)
	400	10	0.09	(0.07)	1.38	(0.08)	1.40	(0.31)	0.75	(0.19)
	800	12	0.04	(0.03)	1.37	(0.05)	1.60	(0.31)	0.86	(0.19)
LPRE					$\tau = 2.46$		$\sigma^2 = 2.46$		$\tau^{-2}\sigma^2 = 0.41$	
	100	7	0.11	(0.10)	2.42	(0.07)	2.19	(0.45)	0.37	(0.06)
	200	8	0.06	(0.05)	2.43	(0.05)	2.29	(0.33)	0.39	(0.04)
	400	10	0.03	(0.02)	2.44	(0.04)	2.36	(0.24)	0.40	(0.03)
	800	12	0.02	(0.01)	2.45	(0.02)	2.40	(0.17)	0.40	(0.02)

Note 1: Based on 500 simulations, means of estimators are outside the parentheses and standard deviations are inside the parentheses.

Note 2: Since $\tau = 2$ is not estimated for LS, it is indicated by "–".

Table 2. Random Simulation: Estimates of regression parameters and standard deviation: each error distribution and loss function.

$(\times 10^{-2})$		LS				RLS				LPRE			
n		100	200	400	800	100	200	400	800	100	200	400	800
p_n		7	8	10	12	7	8	10	12	7	8	10	12
error distribution $\epsilon \sim \mathcal{N}(0,1)$													
$\hat{\beta}_1$	Bias	-1	1	0	0	-1	-4	19	-37	0	1	0	0
	SE	10	14	20	7	238	322	173	474	11	15	21	8
	SEE	10	14	20	7	15	17	19	13	10	14	20	7
	t.SE	10	14	21	7	66	94	135	46	11	16	22	8
$\hat{\beta}_2$	Bias	1	-1	1	0	-11	-20	-3	-22	1	-2	1	1
	SE	17	25	35	12	124	178	110	230	18	26	36	13
	SEE	17	24	34	12	26	30	33	23	18	25	34	13
	t.SE	17	25	36	12	114	162	232	80	19	27	39	13
$\hat{\beta}_3$	Bias	0	0	0	0	1	4	-1	1	0	0	0	0
	SE	6	8	12	4	36	55	33	53	6	9	12	4
	SEE	6	8	11	4	9	10	11	8	6	8	11	4
	t.SE	6	8	12	4	38	54	78	27	6	9	13	4
$\hat{\beta}_4$	Bias	1	0	0	0	2	4	1	1	0	0	-1	0
	SE	7	10	14	4	31	55	69	51	7	10	14	5
	SEE	6	9	13	5	10	11	12	9	7	9	13	5
	t.SE	7	9	13	5	43	61	87	30	7	10	15	5
$\hat{\beta}_5$	Bias	0	-1	0	0	0	-4	2	0	0	-1	1	0
	SE	6	10	13	5	27	63	42	50	7	10	14	5
	SEE	6	9	13	5	10	11	12	9	7	9	13	5
	t.SE	7	9	13	5	43	61	88	30	7	10	15	5
$\hat{\beta}_6$	Bias	0	0	-1	0	2	1	0	1	0	0	-1	0
	SE	7	9	13	4	39	55	38	40	7	10	14	5
	SEE	6	9	13	5	10	11	12	9	7	9	13	5
	t.SE	7	9	13	5	43	61	88	30	7	10	15	5
$\hat{\beta}_7$	Bias	0	0	0	0	-1	4	1	4	0	0	-1	0
	SE	7	10	12	5	26	68	40	75	7	10	12	5
	SEE	6	9	11	5	10	11	11	9	7	9	11	5
	t.SE	7	9	12	5	43	61	78	30	7	10	13	5
$\hat{\beta}_8$	Bias	–	0	0	0	–	2	2	-3	–	0	0	0
	SE	–	7	8	4	–	32	63	45	–	7	9	5
	SEE	–	6	8	5	–	10	10	9	–	7	8	5
	t.SE	–	7	8	5	–	43	54	30	–	7	9	5
$\hat{\beta}_9$	Bias	–	–	0	0	–	–	2	-2	–	–	0	0
	SE	–	–	6	5	–	–	51	47	–	–	7	5
	SEE	–	–	6	5	–	–	10	9	–	–	7	5
	t.SE	–	–	7	5	–	–	43	30	–	–	7	5
$\hat{\beta}_{10}$	Bias	–	–	0	0	–	–	0	4	–	–	0	0
	SE	–	–	6	5	–	–	25	71	–	–	6	5
	SEE	–	–	6	5	–	–	9	9	–	–	6	5
	t.SE	–	–	6	5	–	–	38	30	–	–	6	5
$\hat{\beta}_{11}$	Bias	–	–	–	0	–	–	–	1	–	–	–	0
	SE	–	–	–	4	–	–	–	57	–	–	–	5
	SEE	–	–	–	5	–	–	–	9	–	–	–	5
	t.SE	–	–	–	5	–	–	–	30	–	–	–	5
$\hat{\beta}_{12}$	Bias	–	–	–	0	–	–	–	1	–	–	–	0
	SE	–	–	–	4	–	–	–	41	–	–	–	5
	SEE	–	–	–	4	–	–	–	8	–	–	–	4
	t.SE	–	–	–	4	–	–	–	27	–	–	–	4

Table 2. (Continued.)

$(\times 10^{-2})$		LS				RLS				LPRE			
n		100	200	400	800	100	200	400	800	100	200	400	800
p_n		7	8	10	12	7	8	10	12	7	8	10	12
error distribution $\epsilon \sim U(-2,2)$													
$\hat{\beta}_1$	Bias	−2	0	−2	0	−47	−19	−22	−53	−2	0	−1	0
	SE	11	17	24	8	450	268	381	500	9	14	20	7
	SEE	11	16	23	8	8	11	15	6	10	13	19	7
	t.SE	12	17	24	8	8	12	17	6	10	14	19	7
$\hat{\beta}_2$	Bias	1	−2	−1	−1	−24	−11	−16	−26	1	−2	−2	−1
	SE	19	28	41	14	220	130	175	238	16	24	35	12
	SEE	20	28	39	14	14	19	25	10	16	23	33	12
	t.SE	20	29	41	14	14	20	29	10	16	23	33	12
$\hat{\beta}_3$	Bias	0	1	−1	0	−4	1	0	0	0	1	−1	0
	SE	7	10	14	5	59	48	44	78	6	9	12	4
	SEE	7	9	13	5	5	6	9	3	5	8	11	4
	t.SE	7	10	14	5	5	7	10	3	5	8	11	4
$\hat{\beta}_4$	Bias	0	−1	0	0	−3	−1	−3	3	0	−1	0	0
	SE	8	12	16	5	42	32	72	56	7	10	14	4
	SEE	7	10	15	5	5	7	10	4	6	9	12	4
	t.SE	8	11	15	5	5	8	11	4	6	9	13	4
$\hat{\beta}_5$	Bias	0	0	1	0	−1	3	0	0	0	0	1	0
	SE	8	11	16	6	69	33	47	50	6	9	14	5
	SEE	7	10	15	5	5	7	10	4	6	9	12	4
	t.SE	8	11	16	5	5	7	11	4	6	9	13	4
$\hat{\beta}_6$	Bias	0	0	−1	0	0	1	−4	3	0	0	−1	0
	SE	8	11	16	5	58	39	119	31	7	9	13	4
	SEE	7	10	15	5	5	7	10	4	6	9	12	4
	t.SE	8	11	16	5	5	7	11	4	6	9	13	4
$\hat{\beta}_7$	Bias	0	0	1	0	−1	0	0	2	0	0	1	0
	SE	8	12	14	6	56	43	60	26	7	10	12	5
	SEE	7	10	13	5	5	7	9	4	6	9	11	4
	t.SE	8	11	14	5	5	7	10	4	6	9	11	4
$\hat{\beta}_8$	Bias	–	0	0	0	–	−3	1	3	–	0	0	0
	SE	–	8	10	6	–	41	41	47	–	6	8	5
	SEE	–	7	9	5	–	5	6	4	–	6	8	4
	t.SE	–	8	10	5	–	5	7	4	–	6	8	4
$\hat{\beta}_9$	Bias	–	–	0	0	–	–	−1	0	–	–	0	0
	SE	–	–	7	6	–	–	41	32	–	–	6	5
	SEE	–	–	7	5	–	–	5	4	–	–	6	4
	t.SE	–	–	8	5	–	–	5	4	–	–	6	4
$\hat{\beta}_{10}$	Bias	–	–	0	0	–	–	0	−1	–	–	0	0
	SE	–	–	7	6	–	–	40	26	–	–	6	5
	SEE	–	–	7	5	–	–	5	4	–	–	5	4
	t.SE	–	–	7	5	–	–	5	4	–	–	6	4
$\hat{\beta}_{11}$	Bias	–	–	–	−1	–	–	–	−1	–	–	–	0
	SE	–	–	–	5	–	–	–	37	–	–	–	4
	SEE	–	–	–	5	–	–	–	4	–	–	–	4
	t.SE	–	–	–	5	–	–	–	4	–	–	–	4
$\hat{\beta}_{12}$	Bias	–	–	–	0	–	–	–	−2	–	–	–	0
	SE	–	–	–	5	–	–	–	45	–	–	–	4
	SEE	–	–	–	5	–	–	–	3	–	–	–	4
	t.SE	–	–	–	5	–	–	–	3	–	–	–	4

Table 2. (Continued.)

$(\times10^{-2})$		LS				RLS				LPRE			
n		100	200	400	800	100	200	400	800	100	200	400	800
p_n		7	8	10	12	7	8	10	12	7	8	10	12
error distribution $\epsilon \sim f^*(t)$													
$\hat{\beta}_1$	Bias	-1	0	0	0	2	4	8	2	-1	0	0	0
	SE	7	10	13	5	11	15	20	7	7	10	13	5
	SEE	6	9	13	5	9	11	14	7	6	9	13	5
	t.SE	7	9	13	5	11	16	23	8	6	9	13	5
$\hat{\beta}_2$	Bias	1	0	-1	0	0	0	-2	-1	1	0	-1	0
	SE	11	16	24	8	19	24	34	13	11	16	24	8
	SEE	11	16	22	8	15	19	24	11	11	15	22	8
	t.SE	11	16	23	8	19	27	39	14	11	16	23	8
$\hat{\beta}_3$	Bias	0	0	0	0	0	0	-1	0	0	0	0	0
	SE	4	6	8	3	6	8	12	4	4	6	7	3
	SEE	4	5	7	3	5	6	8	4	4	5	7	3
	t.SE	4	5	8	3	6	9	13	5	4	5	8	3
$\hat{\beta}_4$	Bias	0	0	0	0	0	0	0	0	0	0	0	0
	SE	4	6	8	3	7	9	13	5	4	6	8	3
	SEE	4	6	8	3	6	7	9	4	4	6	8	3
	t.SE	4	6	9	3	7	10	15	5	4	6	9	3
$\hat{\beta}_5$	Bias	0	0	0	0	0	0	-1	0	0	0	0	0
	SE	4	6	9	3	7	10	13	5	4	6	9	3
	SEE	4	6	8	3	6	7	9	4	4	6	8	3
	t.SE	4	6	9	3	7	10	15	5	4	6	9	3
$\hat{\beta}_6$	Bias	0	0	0	0	0	0	1	0	0	0	0	0
	SE	4	6	9	3	7	9	13	5	4	6	9	3
	SEE	4	6	8	3	6	7	9	4	4	6	8	3
	t.SE	4	6	9	3	7	10	15	5	4	6	9	3
$\hat{\beta}_7$	Bias	0	0	0	0	0	0	0	0	0	0	0	0
	SE	4	6	8	3	6	10	12	5	4	6	8	3
	SEE	4	6	7	3	6	7	8	4	4	6	7	3
	t.SE	4	6	8	3	7	10	13	5	4	6	8	3
$\hat{\beta}_8$	Bias	–	0	0	0	–	0	0	0	–	0	0	0
	SE	–	4	5	3	–	7	8	5	–	4	5	3
	SEE	–	4	5	3	–	6	6	4	–	4	5	3
	t.SE	–	4	5	3	–	7	9	5	–	4	5	3
$\hat{\beta}_9$	Bias	–	–	0	0	–	–	0	0	–	–	0	0
	SE	–	–	5	3	–	–	7	5	–	–	4	3
	SEE	–	–	4	3	–	–	6	4	–	–	4	3
	t.SE	–	–	4	3	–	–	7	5	–	–	4	3
$\hat{\beta}_{10}$	Bias	–	–	0	0	–	–	0	0	–	–	0	0
	SE	–	–	4	3	–	–	6	4	–	–	4	3
	SEE	–	–	4	3	–	–	5	4	–	–	4	3
	t.SE	–	–	4	3	–	–	6	5	–	–	4	3
$\hat{\beta}_{11}$	Bias	–	–	–	0	–	–	–	0	–	–	–	0
	SE	–	–	–	3	–	–	–	5	–	–	–	3
	SEE	–	–	–	3	–	–	–	4	–	–	–	3
	t.SE	–	–	–	3	–	–	–	5	–	–	–	3
$\hat{\beta}_{12}$	Bias	–	–	–	0	–	–	–	0	–	–	–	0
	SE	–	–	–	3	–	–	–	4	–	–	–	3
	SEE	–	–	–	3	–	–	–	4	–	–	–	3
	t.SE	–	–	–	3	–	–	–	5	–	–	–	3

Note: Based on 500 simulations and "–" represents "not estimated".

Table 3. Simulation result of variable selection procedure.

Method	$\epsilon \sim \mathcal{N}(0,1)$				$\epsilon \sim U(-2,2)$				$\epsilon \sim f^*(t)$			
	CF	FPR	FNR	ME	CF	FPR	FNR	ME	CF	FPR	FNR	ME
$\mathbf{n = 100}$, $\mathbf{p_n = 7}$, $\mathbf{q_n = 2}$												
LS	0.768	0.117	0	0.032	0.748	0.125	0	0.043	0.756	0.120	0	0.012
LPRE-CV	0.560	0.207	0	0.040	0.602	0.182	0	0.031	0.718	0.135	0	0.014
LPRE-BIC	0.722	0.118	0	0.035	0.812	0.074	0	0.028	0.886	0.044	0	0.010
GREC-CV	0.772	0.105	0	0.085	0.778	0.107	0	0.106	0.804	0.091	0	0.116
GREC-BIC	0.838	0.060	0	0.031	0.890	0.040	0	0.036	0.854	0.052	0	0.012
$\mathbf{n = 200}$, $\mathbf{p_n = 8}$, $\mathbf{q_n = 3}$												
LS	0.742	0.103	0	0.020	0.716	0.112	0	0.030	0.760	0.100	0	0.009
LPRE-CV	0.590	0.151	0	0.024	0.586	0.155	0	0.021	0.772	0.097	0	0.009
LPRE-BIC	0.768	0.070	0	0.023	0.84	0.044	0	0.019	0.914	0.027	0	0.007
GREC-CV	0.814	0.070	0	0.144	0.772	0.09	0	0.101	0.810	0.077	0	0.066
GREC-BIC	0.910	0.025	0	0.022	0.922	0.022	0	0.023	0.908	0.026	0	0.008
$\mathbf{n = 400}$, $\mathbf{p_n = 10}$, $\mathbf{q_n = 4}$												
LS	0.714	0.100	0	0.014	0.756	0.083	0	0.017	0.756	0.084	0	0.005
LPRE-CV	0.682	0.100	0	0.016	0.764	0.074	0	0.012	0.760	0.074	0	0.005
LPRE-BIC	0.830	0.040	0	0.014	0.912	0.020	0	0.015	0.936	0.016	0	0.004
GREC-CV	0.778	0.077	0	0.070	0.802	0.063	0	0.108	0.814	0.059	0	0.086
GREC-BIC	0.920	0.018	0	0.012	0.926	0.016	0	0.019	0.936	0.014	0	0.005
$\mathbf{n = 800}$, $\mathbf{p_n = 12}$, $\mathbf{q_n = 5}$												
LS	0.744	0.072	0	0.007	0.714	0.084	0	0.010	0.786	0.061	0	0.003
LPRE-CV	0.714	0.067	0	0.009	0.708	0.072	0	0.007	0.770	0.057	0	0.003
LPRE-BIC	0.848	0.033	0	0.008	0.934	0.012	0	0.007	0.954	0.008	0	0.003
GREC-CV	0.838	0.041	0	0.020	0.832	0.047	0	0.026	0.862	0.036	0	0.006
GREC-BIC	0.940	0.011	0	0.007	0.958	0.007	0	0.009	0.956	0.008	0	0.003

For each λ and v, we find the estimator $\hat{\beta}_{\lambda,n}$ of β_n using the training set $\mathcal{D} - \mathcal{D}_v$. Then we minimize the following CV criterion to choose λ:

$$CV(\lambda) = \sum_{v=1}^{5} \sum_{(y_k, x_k) \in \mathcal{D}_v} \rho\big(\log y_k - x_k^T \hat{\beta}_{\lambda,n}\big),$$

or alternatively, minimize the following BIC criterion to choose λ:

$$BIC(\lambda) = \log\left[\frac{1}{n}\sum_{k=1}^{n} \rho\big(\log y_k - x_k^T \hat{\beta}_{\lambda,n}\big)\right] + c_n df_\lambda \frac{\log n}{n},$$

where df_λ represents the number of non-zero elements in $\hat{\beta}_{\lambda,n}$ and c_n satisfies $c_n q_n \log(n)/n \to 0$ as $n \to \infty$. For simplicity, we choose $c_n = 1$ in our simulation. As shown in Table 3, the accuracy of model selection shows an increasing trend as $\frac{n}{p_n}$ grows in all settings, which agree with the p_n/n-consistency in our theory. The sensitivity to different choices for the tuning parameters are shown in Table 3. Our simulation results in Table 3 show that BIC outperforms the CV criterion under general relative error settings. One possible explanation for this phenomenon is that the CV criterion primarily emphasizes prediction performance on the validation set, while BIC focuses on constraining model complexity. The LPRE based selection methods is comparable to least square loss and GREC based selection methods outperforms the other two methods in all cases.

7. CONCLUSIONS

In this paper, we propose a general relative error criterion based estimation for the unknown parameter in the divergent-dimensional AFT model. The consistency and asymptotic normality of the estimators are established. Two special cases of general relative error estimators, RLS estimator and LPRE estimator, are studied. We also propose a corresponding variable selection procedure based on penalized general relative error. The oracle properties and consistency of model selection are proved. The simulation results indicate that our estimation methods are feasible.

APPENDIX A. PROOFS

Proof of Theorem 3.1. Proof of Theorem 3.1 is based on the Taylor expansion of the objective function $\psi_n(\beta)$ at the minimal point $\beta = \hat{\beta}_n^*$:

$$\begin{aligned} &\psi_n(\beta_n^* + \gamma_n) - \psi_n(\beta_n^*) \\ =&\big[\dot{\psi}_n(\beta_n^*)\big]^T \gamma_n + \frac{1}{2}\gamma_n^T \big[\ddot{\psi}_n(\beta_n^* + t\gamma_n)\big]\gamma_n, \end{aligned} \tag{3}$$

where $t \in [0,1]$ and

$$\dot{\psi}_n(\beta) = \frac{\partial}{\partial \beta}\psi_n(\beta) = -\sum_{k=1}^{n} \rho'\big(\epsilon_k - x_k^T(\beta - \beta_n^*)\big)x_k,$$

$$\ddot{\psi}_n(\beta) = \frac{\partial^2}{\partial \beta \partial \beta'}\psi_n(\beta) = \sum_{k=1}^{n} \rho''\big(\epsilon_k - x_k^T(\beta - \beta_n^*)\big)x_k x_k^T.$$

We first point out the following facts about the covariates.

Lemma A.1. *Denote $z_k = (S_n^{-1/2})^T x_k$, we have*

(i) $\|z_k z_k^T\| \le \|z_k\|^2 \le d_n$;
(ii) $\sum_{k=1}^n z_k z_k^T = I_{p_n}$;
(iii) $\sum_{k=1}^n \|z_k\|^2 = p_n$. □

Proof of Lemma A.1. The proof is straightforward, the details are omitted.

Now we show that terms in the Taylor expansion (3) have the following properties.

Lemma A.2. *Denote $W_n(\beta) = (S_n^{-1/2})^T \dot{\psi}_n(\beta)$ and $V_n(\beta) = (S_n^{-1/2})^T \ddot{\psi}_n(\beta) S_n^{-1/2}$,*

(i) *$\|W_n(\beta_n^*)\| = O_p(\sqrt{p_n})$ under conditions (B1) and (B2);*
(ii) *If $A_n A_n^T \to G > 0$, then $A_n W_n(\beta_n^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2 G)$ under conditions (A1), (B1) and (B2);*
(iii) *If $\alpha_n^2 p_n d_n \to 0$, then $\|V_n(\beta_n^*) - \tau I_{p_n}\| = o_p(\alpha_n^{-1})$ under conditions (B1) and (B3);*
(iv) *If $\|S_n^{1/2}\gamma_n\| = O_p(\sqrt{p_n})$, $\alpha_n^2 p_n^3 d_n \to 0$ and $\alpha_n p_n \to +\infty$, then $\sup_{t\in[0,1]} \|V_n(\beta_n^* + t\gamma_n) - V_n(\beta_n^*)\| = o_p(\alpha_n^{-1})$ under conditions (B1) and (B4).* □

Proof of Lemma A.2(i). For notation simplicity, denote $W_n = W_n(\beta_n^*)$, then it is straightforward to calculate

$$W_n = \left(S_n^{-1/2}\right)^T \left[\dot{\psi}_n(\beta_n^*)\right] = \left(S_n^{-1/2}\right)^T \sum_{k=1}^n \rho'(\epsilon_k) x_k$$
$$= \sum_{k=1}^n \rho'(\epsilon_k) z_k = \mathbb{Z}e,$$

where $\mathbb{Z} = (z_1, \ldots, z_n)$ is a $p_n \times n$-dimensional matrix, $e = (\rho'(\epsilon_1), \ldots, \rho'(\epsilon_n))^T$ is an n-dimensional vector, thus,

$$\begin{aligned}
&\mathbb{P}\{\|W_n\| > C\} \\
&\le C^{-2} \cdot E\|W_n\|^2 = C^{-2} \cdot E\,\mathrm{tr}\left(e^T \mathbb{Z}^T \mathbb{Z} e\right) \\
&= C^{-2} \cdot E\,\mathrm{tr}\left(\mathbb{Z}^T \mathbb{Z} e e^T\right) = C^{-2} \cdot \mathrm{tr}\left(\mathbb{Z}^T \mathbb{Z}\, E\, e e^T\right) \\
&= C^{-2} \cdot \mathrm{tr}\left(\mathbb{Z}^T \mathbb{Z}\right) \cdot \sigma^2 = C^{-2} \cdot \mathrm{tr}\left(\mathbb{Z}\mathbb{Z}^T\right) \cdot \sigma^2 \\
&= C^{-2} \cdot p_n \cdot \sigma^2,
\end{aligned}$$

and $\|W_n\| = O_p(\sqrt{p_n})$ follows. □

Proof of Lemma A.2(ii). It follows from Cramer-Wold device that *(ii)* holds if and only if for all $s \in \mathbb{R}^m$,

$$s^T A_n W_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\, \sigma^2 s^T G s\right).$$

Denote $Z_k = \rho'(\epsilon_k) s^T A_n z_k$, then $s^T A_n W_n = \sum_{k=1}^n Z_k$ and Z_k satisfies

$$\begin{aligned}
E\,Z_k &= 0, \\
\sum_{k=1}^n E\left(Z_k^2\right) &= \sum_{k=1}^n E\left[\rho'(\epsilon_k)\right]^2 |s^T A_n z_k|^2 \\
&= \sigma^2 s^T A_n \left(\sum_{k=1}^n z_k z_k^T\right) A_n^T s \\
&= \sigma^2 s^T \left(A_n A_n^T\right) s \to \sigma^2 s^T G s.
\end{aligned}$$

For any $\epsilon > 0$,

$$\begin{aligned}
&\sum_{k=1}^n E\left(|Z_k|^2;\, |Z_k|^2 > \varepsilon\right) \\
&= \sum_{k=1}^n |s^T A_n z_k|^2 \cdot E\left(|\rho'(\epsilon_k)|^2;\, \left|\rho'(\epsilon_k)\right|^2 > \frac{\varepsilon}{|s^T A_n z_k|^2}\right) \\
&\le \sum_{k=1}^n |s^T A_n z_k|^2 \cdot E\left(|\rho'(\epsilon_k)|^2;\, \left|\rho'(\epsilon_k)\right|^2 > \frac{\varepsilon}{\|s\|^2 \cdot \|A_n\|^2 \cdot \|z_k\|^2}\right) \\
&\le s^T \left(A_n A_n^T\right) s \cdot E\left(|\rho'(\epsilon)|^2;\, \left|\rho'(\epsilon)\right|^2 > \frac{\varepsilon}{\|s\|^2 \cdot \|A_n\|^2 \cdot d_n}\right) \\
&\to 0.
\end{aligned}$$

The last equality holds because $\|A_n\| = O(1)$, which follows from

$$\|A_n\|^2 \le \|A_n\|_F^2 = \mathrm{tr}\left(A_n A_n^T\right) \to \mathrm{tr}(G) > 0.$$

The conclusion can be verified by Lindeberg-Feller central limit theorem. □

Proof of Lemma A.2(iii). For notation simplicity, denote $V_n = V_n(\beta_n^*)$, then

$$\begin{aligned}
V_n &= \left(S_n^{-1/2}\right)^T \left[\ddot{\psi}_n(\beta_n^*)\right] S_n^{-1/2} \\
&= \left(S_n^{-1/2}\right)^T \left[\sum_{k=1}^n \rho''(\epsilon_k) x_k x_k^T\right] S_n^{-1/2} \\
&= \sum_{k=1}^n \rho''(\epsilon_k) z_k z_k^T,
\end{aligned}$$

then $E(V_n) = \tau I_{p_n}$ and for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\alpha_n \|V_n - \tau I_{p_n}\| > \varepsilon\right\} \\
&\le \frac{\alpha_n^2}{\varepsilon^2} \cdot E\|V_n - \tau I_{p_n}\|^2 \\
&= \frac{\alpha_n^2}{\varepsilon^2} \cdot E\left\|\sum_{k=1}^n \left(\rho''(\epsilon_k) - \tau\right) \cdot z_k z_k^T\right\|^2 \\
&\le \frac{\alpha_n^2}{\varepsilon^2} \cdot E\left\|\sum_{k=1}^n \left(\rho''(\epsilon_k) - \tau\right) \cdot z_k z_k^T\right\|_F^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\alpha_n^2}{\varepsilon^2} \cdot E \sum_{i=1}^{p_n}\sum_{j=1}^{p_n}\left(\sum_{k=1}^{n}\big(\rho''(\epsilon_k)-\tau\big)\cdot z_k^{(i)} z_k^{(j)}\right)^2 \\
&= \frac{\alpha_n^2}{\varepsilon^2} \cdot \sum_{i=1}^{p_n}\sum_{j=1}^{p_n}\sum_{k=1}^{n}\big(\mathrm{var}\big(\rho''(\epsilon_k)\big)\cdot |z_k^{(i)} z_k^{(j)}|^2\big) \\
&= \frac{\mathrm{var}(\rho''(\epsilon))}{\varepsilon^2}\cdot \alpha_n^2 \cdot \sum_{k=1}^{n}\|z_k\|^4 \\
&\le \frac{\mathrm{var}(\rho''(\epsilon))}{\varepsilon^2}\cdot \alpha_n^2 \cdot p_n \cdot d_n \to 0,
\end{aligned}
$$

and the conclusion follows. □

Proof of Lemma A.2(iv). Note that

$$
\begin{aligned}
&B_n(\gamma_n) \\
&:= \sup_{t\in[0,1]} \big\| \big(S_n^{-1/2}\big)^T \big[\ddot{\psi}_n(\beta_n^* + t\gamma_n) - \ddot{\psi}_n(\beta_n^*)\big] S_n^{-1/2} \big\| \\
&= \sup_{t\in[0,1]} \Bigg\| \big(S_n^{-1/2}\big)^T \Bigg[\sum_{k=1}^{n} \big[\rho''\big(\epsilon_k - t x_k^T \gamma_n\big) \\
&\quad - \rho''(\epsilon_k)\big] x_k x_k^T \Bigg] S_n^{-1/2} \Bigg\| \\
&= \sup_{t\in[0,1]} \Bigg\| \sum_{k=1}^{n} \big[\rho''\big(\epsilon_k - t z_k^T v_n\big) - \rho''(\epsilon_k)\big] z_k z_k^T \Bigg\|,
\end{aligned}
$$

where $v_n = S_n^{1/2}\gamma_n$ and $\|v_n\| = \|S_n^{1/2}\gamma_n\| = O_p(\sqrt{p_n})$. Let

$$\omega(u,\delta) = \sup_{t:|t|\le\delta} \big|\rho''(u+t) - \rho''(u)\big|$$

be the oscillation of $\rho''(\cdot)$ in the δ-neighborhood $(u-\delta, u+\delta)$ with respect to u. We first deal with the case when $\|v_n\| = O(\sqrt{p_n})$. For any $\epsilon > 0$,

$$
\begin{aligned}
&\mathbb{P}\big\{\alpha_n B_n(\gamma_n) > \varepsilon\big\} \\
&\le \frac{\alpha_n^2}{\varepsilon^2}\cdot E\left\{ \sup_{t\in[0,1]} \left\| \sum_{k=1}^{n} \big[\rho''\big(\epsilon_k - t z_k^T v_n\big) - \rho''(\epsilon_k)\big]\cdot z_k z_k^T \right\|^2 \right\} \\
&\le \frac{\alpha_n^2}{\varepsilon^2}\cdot E\left\{ \sup_{t\in[0,1]} \sum_{k=1}^{n} \big|\rho''\big(\epsilon_k - t z_k^T v_n\big) - \rho''(\epsilon_k)\big| \cdot \big\|z_k z_k^T\big\| \right\}^2 \\
&\le \frac{\alpha_n^2}{\varepsilon^2}\cdot E\left\{ \sum_{k=1}^{n} \omega\big(\epsilon_k, |z_k^T v_n|\big)\cdot \big\|z_k z_k^T\big\| \right\}^2 \\
&\le \frac{\alpha_n^2}{\varepsilon^2}\cdot \left\{ \sum_{k=1}^{n} E\omega^2\big(\epsilon_k, |z_k^T v_n|\big) \right\} \left\{ \sum_{k=1}^{n} \big\|z_k z_k^T\big\|^2 \right\}.
\end{aligned}
$$

Since $|z_k^T v_n| \le \|z_k\|\|v_n\| \le \sqrt{d_n p_n}$ and $\alpha_n^2 p_n^3 d_n \to 0$ with $\alpha_n p_n \to +\infty$ implies condition (A2), thus follows from condition (B4),

$$E\omega^2\big(\epsilon_k, |z_k^T v_n|\big) = O(1)\cdot |z_k^T v_n|^2,$$

holds uniformly in k, therefore,

$$
\begin{aligned}
&\mathbb{P}\big\{\alpha_n B_n(\gamma_n) > \varepsilon\big\} \\
&\le \frac{\alpha_n^2}{\varepsilon^2}\cdot \left\{ \sum_{k=1}^{n} E\omega^2\big(\epsilon_k, |z_k^T v_n|\big) \right\} \left\{ \sum_{k=1}^{n} \big\|z_k z_k^T\big\|^2 \right\} \\
&\le \frac{\alpha_n^2}{\varepsilon^2}\cdot \left\{ O(1) \sum_{k=1}^{n} |z_k^T v_n|^2 \right\} \left\{ \sum_{k=1}^{n} \|z_k\|^4 \right\} \\
&\le \frac{\alpha_n^2}{\varepsilon^2}\cdot \left\{ O(1) \left(\sum_{k=1}^{n} \|z_k\|^2 \right) \|v_n\|^2 \right\} \cdot p_n d_n \\
&= \frac{O(1)}{\varepsilon^2}\cdot \alpha_n^2 \cdot p_n^3 \cdot d_n \to 0.
\end{aligned}
$$

Now we deal with the general case $\|v_n\| = O_p(\sqrt{p_n})$. For any given $\varepsilon > 0$, $M > 0$,

$$
\begin{aligned}
\mathbb{P}\big\{\alpha_n B_n(\gamma_n) > \varepsilon\big\} &\le \mathbb{P}\big\{\alpha_n B_n(\gamma_n) > \varepsilon;\ \|v_n\| \le M\big\} \\
&\quad + \mathbb{P}\big\{\|v_n\| > M\big\}.
\end{aligned}
$$

The conclusion follows from choosing a sufficiently large M so that the second term can be sufficiently small and the first term converges to zero for any fixed M. □

Proof of Theorem 3.1(i): Consistency. The local minimum $\hat{\beta}_n$ of $\psi_n(\beta)$ satisfying $\|S_n^{1/2}(\hat{\beta}_n - \beta_n^*)\| = O_p(\sqrt{p_n})$ is equivalent to: for any $\delta > 0$, there exists $C > 0$,

$$\mathbb{P}\Big\{ \inf_{u:\|u\|=C} \psi_n\big(\beta_n^* + \sqrt{p_n} S_n^{-1/2} u\big) > \psi_n\big(\beta_n^*\big) \Big\} \ge 1 - \delta,$$

for sufficiently large n, which means there exists a local minimum $\hat{\beta}_n$ in the ball $\{\beta_n^* + \sqrt{p_n} S_n^{-1/2} u : \|u\| \le C\}$ with probability tending to 1. Let $\gamma_n = \sqrt{p_n} S_n^{-1/2} u$, using the Taylor expansion in (3), we have

$$
\begin{aligned}
&\psi_n\big(\beta_n^* + \sqrt{p_n} S_n^{-1/2} u\big) - \psi_n\big(\beta_n^*\big) \\
&= \dot{\psi}_n\big(\beta_n^*\big)^T \big(\sqrt{p_n} S_n^{-1/2} u\big) \\
&\quad + \frac{1}{2}\big(\sqrt{p_n} S_n^{-1/2} u\big)^T \big[\ddot{\psi}_n\big(\beta_n^* + t\gamma_n\big)\big] \big(\sqrt{p_n} S_n^{-1/2} u\big) \\
&= \sqrt{p_n} W_n^T u + \frac{1}{2} p_n u^T \{\tau I_{p_n} + J_{1n} + J_{2n}\} u,
\end{aligned}
$$

where

$$
\begin{aligned}
J_{1n} &= \big(S_n^{-1/2}\big)^T \big[\ddot{\psi}_n\big(\beta_n^*\big)\big] S_n^{-1/2} - \tau I_{p_n}, \\
J_{2n} &= \big(S_n^{-1/2}\big)^T \big[\ddot{\psi}_n\big(\beta_n^* + t\gamma_n\big) - \ddot{\psi}_n\big(\beta_n^*\big)\big] S_n^{-1/2}.
\end{aligned}
$$

A.2(i) shows $\|W_n\| = O_p(\sqrt{p_n})$. Take $\alpha_n = 1$ in A.2(iii) and A.2(iv), then condition (A3) implies $\|J_{1n}\| = o_p(1/p_n)$ and $\|J_{2n}\| = o_p(1)$. Hence,

$$
\begin{aligned}
&\mathbb{P}\Big\{ \inf_{u:\|u\|=C} \psi_n\big(\beta_n^* + \sqrt{p_n} S_n^{-1/2} u\big) > \psi_n\big(\beta_n^*\big) \Big\} \\
&= \mathbb{P}\Big\{ \inf_{u:\|u\|=C} \sqrt{p_n} W_n^T u + \frac{p_n}{2} u^T \{\tau I_{p_n} + J_{1n} + J_{2n}\} u > 0 \Big\}
\end{aligned}
$$

$$
\begin{aligned}
&\geq \mathbb{P}\Big\{-\sqrt{p_n}|O_p(\sqrt{p_n})|C + p_n\Big(\frac{\tau}{2} - |o_p(1)|\Big)C^2 > 0\Big\}\\
&= \mathbb{P}\Big\{|O_p(1)| < \Big(\frac{\tau}{2} - |o_p(1)|\Big)\cdot C\Big\}\\
&\geq \mathbb{P}\Big\{|O_p(1)| < \Big(\frac{\tau}{2} - |o_p(1)|\Big)\cdot C,\, |o_p(1)| < \frac{\tau}{4}\Big\}\\
&\geq \mathbb{P}\Big\{|O_p(1)| < \frac{\tau}{4}\cdot C,\, |o_p(1)| < \frac{\tau}{4}\Big\}\\
&\geq \mathbb{P}\Big\{|O_p(1)| < \frac{\tau}{4}\cdot C\Big\} - \mathbb{P}\Big\{|o_p(1)| \geq \frac{\tau}{4}\Big\}.
\end{aligned}
$$

The above probability is greater than $1-\delta$ by choosing a sufficiently large C. □

Proof of Theorem 3.1(ii): Asymptotic normality. Using mean value theorem of vector-valued function, we have

$$
\dot{\psi}_n(\hat{\beta}_n) - \dot{\psi}_n(\beta_n^*) = \Big[\int_0^1 \ddot{\psi}_n\big(\beta_n^* + t(\hat{\beta}_n - \beta_n^*)\big)dt\Big](\hat{\beta}_n - \beta_n^*).
$$

Pre-multiplying $(S_n^{-1/2})^T$ on both sides, the right hand side becomes

$$
D_n + W_n = (\tau I_{p_n} + C_n)\cdot S_n^{1/2}(\hat{\beta}_n - \beta_n^*),
$$

where

$$
\begin{aligned}
D_n &= \big(S_n^{-1/2}\big)^T\dot{\psi}_n(\hat{\beta}_n),\\
C_n &= \int_0^1 \big[\big(S_n^{-1/2}\big)^T\ddot{\psi}_n\big(\beta_n^* + t(\hat{\beta}_n - \beta_n^*)\big)S_n^{-1/2} - \tau I_{p_n}\big]dt.
\end{aligned}
$$

Pre-multiplying $m \times p_n$-dimensional matrix A_n, we obtain

$$
\begin{aligned}
&A_nD_n + A_nW_n\\
&= A_n(\tau I_{p_n} + C_n)\cdot S_n^{1/2}(\hat{\beta}_n - \beta_n^*)\\
&= \tau A_nS_n^{1/2}(\hat{\beta}_n - \beta_n^*) + A_nC_nS_n^{1/2}(\hat{\beta}_n - \beta_n^*).
\end{aligned}
$$

Noted that

$$
\|A_n\| \leq \|A_n\|_F = \sqrt{\operatorname{tr}\big(A_nA_n^T\big)} \to \sqrt{\operatorname{tr}(G)},
$$

which implies that $\|A_n\| = O(1)$. Since $\hat{\beta}_n$ is a local minimum, hence,

$$
P\big\{\|D_n\| \leq \varepsilon\big\} \geq P\{D_n = 0\} = P\big\{\dot{\psi}_n(\hat{\beta}_n) = 0\big\} \to 1,
$$

which implies that $\|B_n\| = o_p(1)$. Let $\gamma_n = \hat{\beta}_n - \beta_n^*$ in A.2(iii) and A.2(iv), the consistency $\|S_n^{1/2}\gamma_n\| = O_p(\sqrt{p_n})$ with $\alpha_n = \sqrt{p_n}$ and condition (A3) implies $\|C_n\| = o_p(1/\sqrt{p_n})$. Consequently,

$$
\begin{aligned}
\|A_nD_n\| \leq \|A_n\|\|D_n\| &= O(1)o_p(1) = o_p(1),\\
\big\|A_nC_nS_n^{1/2}(\hat{\beta}_n - \beta_n^*)\big\| &\leq \|A_n\|\|C_n\|\big\|S_n^{1/2}(\hat{\beta}_n - \beta_n^*)\big\|\\
&= O(1)o_p(1/\sqrt{p_n})O_p(\sqrt{p_n}) = o_p(1).
\end{aligned}
$$

Thus,

$$
A_nS_n^{1/2}(\hat{\beta}_n - \beta_n^*) = \tau^{-1}A_nW_n + o_p(1).
$$

The asymptotic normality follows from A.2(ii). □

Proof of Theorem 3.2(i). The proof involves two steps. We first prove

$$
\hat{\tau} = \arg\min_t\big\|\big(S_n^{-1/2}\big)^T\big[\ddot{\psi}_n(\hat{\beta}_n)\big]S_n^{-1/2} - t\cdot I_{p_n}\big\|_F^2.
$$

Noted that only the diagonal of $(S_n^{-1/2})^T[\ddot{\psi}_n(\hat{\beta}_n)]S_n^{-1/2} - t\cdot I_{p_n}$ depends on t, we have

$$
\begin{aligned}
&\big\|\big(S_n^{-1/2}\big)^T\big[\ddot{\psi}_n(\hat{\beta}_n)\big]S_n^{-1/2} - t\cdot I_{p_n}\big\|_F^2\\
&= \sum_{i=1}^{p_n}\Big\{\Big(\sum_{k=1}^n \rho''(e_k)|z_k^{(i)}|^2\Big) - t\Big\}^2 + C_n,
\end{aligned}
$$

by direct calculation. where C_n is independent of t. $\hat{\tau}$ can be solved by equating the derivative to zero:

$$
\begin{aligned}
\hat{\tau} &= \frac{1}{p_n}\sum_{i=1}^{p_n}\sum_{k=1}^n \rho''(e_k)\cdot|z_k^{(i)}|^2 = \frac{1}{p_n}\sum_{k=1}^n \rho''(e_k)\cdot\|z_k\|^2\\
&= \frac{1}{p_n}\sum_{k=1}^n \rho''(e_k)\cdot x_k^TS_n^{-1}x_k.
\end{aligned}
$$

Then, using the equivalence of norm A.2(iii), A.2(iv) and the result above, we obtain

$$
\begin{aligned}
|\hat{\tau} - \tau| = \|\hat{\tau}I_{p_n} - \tau I_{p_n}\| &\leq \|\hat{\tau}I_{p_n} - \tau I_{p_n}\|_F\\
&\leq \big\|\big(S_n^{-1/2}\big)^T\big[\ddot{\psi}_n(\hat{\beta}_n)\big]S_n^{-1/2} - \hat{\tau}I_{p_n}\big\|_F\\
&+ \big\|\big(S_n^{-1/2}\big)^T\big[\ddot{\psi}_n(\hat{\beta}_n)\big]S_n^{-1/2} - \tau I_{p_n}\big\|_F\\
&\leq 2\big\|\big(S_n^{-1/2}\big)^T\big[\ddot{\psi}_n(\hat{\beta}_n)\big]S_n^{-1/2} - \tau I_{p_n}\big\|_F\\
&\leq 2\sqrt{p_n}\big\|\big(S_n^{-1/2}\big)^T\big[\ddot{\psi}_n(\hat{\beta}_n)\big]S_n^{-1/2} - \tau I_{p_n}\big\|\\
&= 2\sqrt{p_n}o_p\Big(\frac{1}{\sqrt{p_n}}\Big) = o_p(1),
\end{aligned}
$$

and the consistency is proved. □

Proof of Theorem 3.2(ii). The following fact is needed to prove 3.2(ii).

If conditions (B1)–(B5) hold, then

$$
\sup_{t\in[0,1]}\Big\|\sum_{k=1}^n |\rho''(e_k)|^2z_kz_k^T - \sigma^2I_{p_n}\Big\| = o_p\Big(\frac{1}{\sqrt{p_n}}\Big),
$$

where $z_k = (S_n^{-1/2})^Tx_k$ is defined by A.1.

The proof can be obtained by the same argument as the proof of A.2(ii) and 3.1(ii). □

Proof of Corollary 4.1. Since $\rho(t) = (1 - e^{-t})^2$, by direct calculation,

$$\begin{aligned}
\rho'(t) &= 2\big(e^{-t} - e^{-2t}\big),\\
E\big[\rho'(\epsilon)\big] &= 2\big[M(-1) - M(-2)\big] = 0,\\
E\big[\rho'(\epsilon)\big]^2 &= 4\big[M(-4) - 2M(-3) + M(-2)\big] = \sigma^2,\\
\rho''(t) &= 4e^{-2t} - 2e^{-t},\\
E\big[\rho''(\epsilon)\big] &= 4\big[M(-2) - 2M(-1)\big] = 2M(-1) = \tau,\\
E\big[\rho''(\epsilon)\big]^2 &= 4M(-2) - 16M(-3) + 16M(-4) < +\infty.
\end{aligned}$$

Condition (B1)–(B3) are satisfied. As $t \to 0+$,

$$\begin{aligned}
&E\Big\{\sup_{s:|s|\le t}\big|\rho''(\epsilon+s) - \rho''(\epsilon)\big|^2\Big\}\\
&= E\Big\{\sup_{s:|s|\le t}\big|\big(4e^{-2(\epsilon+s)} - 2e^{-(\epsilon+s)}\big) - \big(4e^{-2\epsilon} - 2e^{-\epsilon}\big)\big|^2\Big\}\\
&= E\Big\{\sup_{s:|s|\le t}\big|4e^{-2\epsilon}\big(e^{-2s} - 1\big) + 2e^{-\epsilon}\big(1 - e^{-s}\big)\big|^2\Big\}\\
&\le E\big\{4e^{-2\epsilon}\big(e^{2t} - 1\big) + 2e^{-\epsilon}\big(e^{t} - 1\big)\big\}^2\\
&\le E\big\{4e^{-2\epsilon}(3t) + 2e^{-\epsilon}(2t)\big\}^2\\
&= \big\{E\big(12e^{-2\epsilon} + 4e^{-\epsilon}\big)^2\big\}t^2.
\end{aligned}$$

Condition (A4) is satisfied.

$$\begin{aligned}
&E\Big\{\sup_{s:|s|\le t}\big||\rho'(\epsilon+s)|^2 - |\rho'(\epsilon)|^2\big|^2\Big\}\\
&= E\Big\{\sup_{s:|s|\le t} 4\big|e^{-2\epsilon}\big(e^{-2s} - 1\big) + e^{-4\epsilon}\big(e^{-4s} - 1\big)\\
&\qquad + 2e^{-3\epsilon}\big(1 - e^{-3s}\big)\big|^2\Big\}\\
&\le E\big\{4\big|e^{-2\epsilon}(3t) + e^{-4\epsilon}(4t) + 2e^{-3\epsilon}(4t)\big|^2\big\}\\
&= \big\{E\big(6e^{-2\epsilon} + 8e^{-4\epsilon} + 16e^{-3\epsilon}\big)^2\big\}t^2,
\end{aligned}$$

which implies condition (A5) is satisfied when $M(-8) < +\infty$, and the corollary follows. □

Proof of Corollary 4.2. Since $\rho(t) = e^t + e^{-t} - 2$, by direct calculation,

$$\begin{aligned}
\rho'(t) &= e^{t} - e^{-t},\\
E\big[\rho'(\epsilon)\big] &= M(1) - M(-1) = 0,\\
E\big[\rho'(\epsilon)\big]^2 &= \big[M(2) + M(-2) - 2\big] = \sigma^2,\\
\rho''(t) &= e^{t} + e^{-t},\\
E\big[\rho''(\epsilon)\big] &= M(1) + M(-1) = 2M(1) = \tau,\\
E\big[\rho''(\epsilon)\big]^2 &= \big[M(2) + M(-2) + 2\big] < +\infty,
\end{aligned}$$

hence, conditions (B1-B3) are satisfied. As $t \to 0+$,

$$\begin{aligned}
&E\Big\{\sup_{s:|s|\le t}\big|\rho''(\epsilon+s) - \rho''(\epsilon)\big|^2\Big\}\\
&= E\Big\{\sup_{s:|s|\le t}\big|\big(e^{(\epsilon+s)} + e^{-(\epsilon+s)}\big) - \big(e^{\epsilon} + e^{-\epsilon}\big)\big|^2\Big\}\\
&\le \big\{E\big(2e^{\epsilon} + 2e^{-\epsilon}\big)^2\big\}t^2.
\end{aligned}$$

Condition (A4) is satisfied.

$$\begin{aligned}
&E\Big\{\sup_{s:|s|\le t}\big||\rho'(\epsilon+s)|^2 - |\rho'(\epsilon)|^2\big|^2\Big\}\\
&= E\Big\{\sup_{s:|s|\le t}\big|\big(e^{2(\epsilon+s)} + e^{-2(\epsilon+s)-2}\big) - \big(e^{2\epsilon} + e^{-2\epsilon} - 2\big)\big|^2\Big\}\\
&\le \big\{E\big(3e^{2\epsilon} + 3e^{-2\epsilon}\big)^2\big\}t^2,
\end{aligned}$$

which implies condition (A5) when $M(\pm 4) < +\infty$, and the corollary follows. □

Proof of Theorem 5.1(i): Consistency of estimation. According to condition (A4), the theorem is equivalent to

$$\big\|S_n^{1/2}\big(\hat\beta_n - \beta_n^*\big)\big\| = O_p(\sqrt{p_n}).$$

Then, we prove for any given $\delta > 0$, there exists $C > 0$ such that

$$\mathbb{P}\Big\{\inf_{u:\|u\|=C} Q_n\big(\beta_n^* + \sqrt{p_n}S_n^{-1/2}u\big) > Q_n\big(\beta_n^*\big)\Big\} \ge 1 - \delta,$$

holds for sufficiently large n. That is, there exists a local minimum $\hat\beta_n$ inside the ball $\{\beta_n^* + \sqrt{p_n}S_n^{-1/2}u : \|u\| \le C\}$ with probability tends to 1. Denote

$$D_n(\gamma_n) = Q_n\big(\beta_n^* + \gamma_n\big) - Q_n\big(\beta_n^*\big).$$

Take $\gamma_n = \sqrt{p_n}S_n^{-1/2}u$, $\|u\| = C$, then follow the proof of 3.1(i), we have

$$\begin{aligned}
D_n(\gamma_n) &= \frac{1}{n}\big\{\psi_n\big(\beta_n^* + \sqrt{p_n}S_n^{-1/2}u\big) - \psi_n\big(\beta_n^*\big)\big\}\\
&\quad + \sum_{j=1}^{p_n}\big\{p_{\lambda_n}\big(|\beta_n^{*(j)} + \gamma_n^{(j)}|\big) - p_{\lambda_n}\big(|\beta_n^{*(j)}|\big)\big\}\\
&\ge \frac{p_n}{n}\bigg\{-\big|O_p(1)\big|C + \bigg(\frac{\tau}{2} - \big|o_p(1)\big|\bigg)C^2\bigg\}\\
&\quad + \sum_{j=1}^{q_n}\big\{p_{\lambda_n}\big(|\beta_n^{*(j)} + \gamma_n^{(j)}|\big) - p_{\lambda_n}\big(|\beta_n^{*(j)}|\big)\big\}\\
&= J_{3n} + J_{4n}.
\end{aligned}$$

Condition (A2) and (A4) indicates

$$\|\gamma_n\|^2 = p_n u^T S_n^{-1} u \le \frac{p_n}{\Omega_{\min}(S_n)}C^2 < \frac{p_n}{nr}C^2 \to 0.$$

Hence, using the Taylor expansion, we have

$$J_{4n} = \sum_{j=1}^{q_n}\big\{p_{\lambda_n}\big(|\beta_n^{*(j)} + \gamma_n^{(j)}|\big) - p_{\lambda_n}\big(|\beta_n^{*(j)}|\big)\big\}$$

$$
\begin{aligned}
&= \sum_{j=1}^{q_n} \bigl\{ p'_{\lambda_n}\bigl(|\beta_n^{*(j)}|\bigr)\operatorname{sgn}\bigl(\beta_n^{*(j)}\bigr)\gamma_n^{(j)} \\
&\quad + p''_{\lambda_n}\bigl(|\beta_n^{*(j)}|\bigr)\bigl(\gamma_n^{(j)}\bigr)^2\bigl[1+o(1)\bigr]/2 \bigr\} \\
&\geq -a_n \cdot \sqrt{q_n}\|\gamma_n\| - b_n \cdot \|\gamma_n\|^2 \\
&\geq -a_n \cdot \sqrt{(p_n q_n)/(nr)} \cdot C - b_n \cdot p_n/(nr) \cdot C^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
D_n(\gamma_n) &\geq J_{3n} + J_{4n} \\
&\geq -\frac{p_n}{n}\bigl(|O_p(1)| + a_n \cdot \sqrt{nr_n/r}\bigr)C \\
&\quad + \frac{p_n}{n}\left(\frac{\tau}{2} - |o_p(1)| - \frac{b_n}{r}\right)C^2.
\end{aligned}
$$

Using regularity conditions (C1), (C2) and the same argument as the proof of 3.1(i), the existence and consistency of the estimator follows. □

Proof of Theorem 5.1(ii): Consistency of model selection. We prove that if $\|\beta_n - \beta_n^*\| = O_p(\sqrt{p_n/n})$, then for any constant C,

$$
\mathbb{P}\left\{ 0 = \arg \min_{\beta_{n2}:\|\beta_n - \beta_n^*\| \leq C\sqrt{p_n/n}} Q_n \begin{pmatrix} \beta_{n1} \\ \beta_{n2} \end{pmatrix} \right\} \to 1.
$$

Denote

$$
\frac{\partial}{\partial \beta} Q_n(\beta) = \frac{1}{n}\dot{\psi}_n(\beta) + \mathbf{b}_n(\beta),
$$

where

$$
\mathbf{b}_n^{(j)}(\beta) = p'_{\lambda_n}\bigl(|\beta^{(j)}|\bigr)\operatorname{sgn}\bigl(\beta^{(j)}\bigr).
$$

By the mean value theorem of vector-valued function, we have

$$
\dot{\psi}_n(\beta_n) = \dot{\psi}_n(\beta_n^*) + \left[\int_0^1 \ddot{\psi}_n\bigl(\beta_n^* + t(\beta - \beta_n^*)\bigr)dt\right](\beta - \beta_n^*).
$$

Since $\|\beta_n - \beta_n^*\| = O_p(\sqrt{p_n/n})$, follow the proof of Theorem 3.1, we have

$$
\dot{\psi}_n(\beta_n) = \bigl(S_n^{1/2}\bigr)^T\bigl\{W_n + (\tau I_{p_n} + C_n)\cdot S_n^{1/2}(\beta_n - \beta_n^*)\bigr\},
$$

where

$$
\begin{aligned}
W_n &= \bigl(S_n^{-1/2}\bigr)^T \dot{\psi}_n(\beta_n^*), \\
C_n &= \int_0^1 \bigl[\bigl(S_n^{-1/2}\bigr)^T \ddot{\psi}_n\bigl(\beta_n^* + t(\beta_n - \beta_n^*)\bigr)S_n^{-1/2} - \tau I_{p_n}\bigr]dt.
\end{aligned}
$$

Take $\gamma_n = \beta_n - \beta_n^*, \alpha_n = 1$ in A.2(iii) and A.2(iv), we have $\|C_n\| = o_p(1)$. Therefore,

$$
\begin{aligned}
&\left\|\frac{1}{n}\dot{\psi}_n(\beta_n)\right\| \\
&\leq \frac{1}{n}\bigl\|\bigl(S_n^{1/2}\bigr)^T\bigr\|\bigl\{\|W_n\| + \bigl(\|\tau I_{p_n}\| + \|C_n\|\bigr)\|S_n^{1/2}\|\|\beta_n - \beta_n^*\|\bigr\} \\
&= n^{-1}O(\sqrt{n})\bigl\{O_p(\sqrt{p_n}) + \bigl(\tau + o_p(1)\bigr)O(\sqrt{n})O_p(\sqrt{p_n/n})\bigr\} \\
&= O_p(\sqrt{p_n/n}).
\end{aligned}
$$

Then under regularity condition (C4)

$$
\begin{aligned}
\frac{\partial}{\partial \beta} Q_n(\beta_n) &= O_p(\sqrt{p_n/n}) + \mathbf{b}_n(\beta_n) \\
&= \lambda_n\bigl\{o_p(1) + \lambda_n^{-1}\mathbf{b}_n(\beta_n)\bigr\},
\end{aligned}
$$

whose j-th element is

$$
\frac{\partial}{\partial \beta^{(j)}} Q_n(\beta_n) = \lambda_n\left\{o_p(1) + \frac{p'_{\lambda_n}(|\beta_n^{(j)}|)}{\lambda_n}\operatorname{sgn}\bigl(\beta_n^{(j)}\bigr)\right\}.
$$

$\beta_n^{(j)} \xrightarrow{p} 0$ if $j > q_n$, hence condition (C3) indicates that the sign of left hand side is determined by $\operatorname{sgn}(\beta_n^{(j)})$. Therefore, with β_{n1} fixed, $Q_n(\beta_n)$ achieves its minimum at $\beta_{n2} = 0$ with probability tends to 1, and the consistency of model selection follows.

Before we prove Theorem 5.1(iii), we first state some notations and lemmas. Let $\beta_1 \in \mathbb{R}^{q_n}$,

$$
\begin{aligned}
\dot{\psi}_{n1}(\beta_1) &= -\sum_{k=1}^n \rho'\bigl(\epsilon_k - \bar{x}_k^T(\beta_1 - \beta_{n1}^*)\bigr)\bar{x}_k, \\
\ddot{\psi}_{n1}(\beta_1) &= \sum_{k=1}^n \rho''\bigl(\epsilon_k - \bar{x}_k^T(\beta_1 - \beta_{n1}^*)\bigr)\bar{x}_k\bar{x}_k^T,
\end{aligned}
$$

and

$$
Q_{n1}(\beta_1) = Q_n\begin{pmatrix} \beta_1 \\ 0 \end{pmatrix}.
$$

Its partial derivatives equal to

$$
\frac{\partial}{\partial \beta_1} Q_{n1}(\beta_1) = \frac{1}{n}\dot{\psi}_{n1}(\beta_1) + \mathbf{b}_{n1}(\beta_1),
$$

where $\mathbf{b}_{n1}^{(j)}(\beta_1) = p'_{\lambda_n}(|\beta_1^{(j)}|)\operatorname{sgn}(\beta_1^{(j)})$. Since we have already proved the existence and consistency of the estimator and the consistency of model selection, we may asuume $\hat{\beta}_{n2} = 0$, $\|\hat{\beta}_{n1}\| \leq \|\hat{\beta}_n\| = O_p(\sqrt{p_n/n})$. Since $\hat{\beta}_n$ is a local minimum, hence, $\hat{\beta}_{n1}$ satisfies

$$
0 = \frac{\partial}{\partial \beta_1} Q_{n1}(\hat{\beta}_{n1}) = \frac{1}{n}\dot{\psi}_{n1}(\hat{\beta}_{n1}) + \mathbf{b}_{n1}(\hat{\beta}_{n1}). \tag{4}
$$

Consider the first term at the right hand side. Using mean value theorem of vector-valued functions, we have:

$$
\begin{aligned}
&\dot{\psi}_{n1}(\hat{\beta}_{n1}) \\
&= \dot{\psi}_{n1}(\beta_{n1}^*) \\
&\quad + \left[\int_0^1 \ddot{\psi}_{n1}\bigl(\beta_{n1}^* + t(\hat{\beta}_{n1} - \beta_{n1}^*)\bigr)dt\right](\hat{\beta}_{n1} - \beta_{n1}^*)
\end{aligned}
$$

$$= \big(S_{n1}^{1/2}\big)^T \big\{W_{n1} + (\tau I_{q_n} + C_{n1}) \cdot S_{n1}^{1/2}\big(\hat\beta_{n1} - \beta_{n1}^*\big)\big\},$$

where

$$W_{n1} = \big(S_{n1}^{-1/2}\big)^T \dot\psi_{n1}(\beta_{n1}^*),$$
$$C_{n1} = \int_0^1 \big[\big(S_{n1}^{-1/2}\big)^T \ddot\psi_{n1}\big(\beta_{n1}^* + t\big(\hat\beta_{n1} - \beta_{n1}^*\big)\big) S_{n1}^{-1/2} - \tau I_{q_n}\big]\, dt.$$

Lemma A.3. *Under conditions (A1), (A4), (B1) and (B2), for a $m \times q_n$-dimensional matrix A_n, if $A_n A_n^T$ converges to $m \times m$-dimensional positive-definite matrix G, that is $A_n A_n^T \to G > 0$, then $A_n W_{n1} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2 G)$.*

Lemma A.4. *Under conditions (A3), (A4), (B1) and (B3), we have $\|C_{n1}\| = o_p(\frac{1}{\sqrt{p_n}})$.* □

Proof of Lemma A.3 and A.4. Since $q_n < p_n$ have already been proved in the proof of A.2, we only need to show that

$$d_{n1} = O(d_n), \quad \text{where} \quad d_{n1} = \max_{1 \le k \le n} \bar x_k^T S_{n1}^{-1} \bar x_k.$$

Noted that S_{n1} is a $q_n \times q_n$-dimensional submatrix of S_n on the left upper corner, thus

$$\bar x_k^T S_{n1}^{-1} \bar x_k \le \frac{\|\bar x_k\|^2}{\Omega_{\min}(S_{n1})} \le \frac{\|x_k\|^2}{\Omega_{\min}(S_{n1})},$$
$$\Omega_{\min}(S_{n1}) = \inf_{u \ne 0} \frac{u^T S_{n1} u}{\|u\|^2} = \inf_{u \ne 0} \frac{(u^T, 0) S_n (u^T, 0)^T}{\|(u^T, 0)\|^2}$$
$$\ge \inf_{v \ne 0} \frac{v^T S_n v}{\|v\|^2} = \Omega_{\min}(S_n) > nr,$$

therefore, $d_{n1} < d_n^*/r$, where $d_n^* = \max_{1\le k \le n} \frac{\|x_k\|^2}{n}$. Since $\frac{\|x_k\|^2}{nR} \le x_k^T S_n^{-1} x_k \le \frac{\|x_k\|^2}{nr}$, d_n and d_n^* are of same order, that is, $d_n \sim d_n^*$. If $\max_{1 \le k \le n} \|x_k\|^2 = o(n/p_n^4)$, then conditions (A4) and (A5) implies (A0-A3). It follows from condition (A4) that $d_n^* \sim d_n$, hence, $d_{n1} = O(d_n)$ and the lemma is proved. □

Proof of Theorem 5.1(iii): Asymptotic normality. Using (A.4), (4) becomes

(5)
$$-\tau^{-1} W_{n1} = S_{n1}^{1/2}\big(\hat\beta_{n1} - \beta_{n1}^*\big) + o_p(1) + n\tau^{-1}\big(S_{n1}^{-1/2}\big)^T \mathbf{b}_{n1}(\hat\beta_{n1}).$$

Now we consider the term $\mathbf{b}_{n1}(\hat\beta_{n1})$. Denote

$$g_n(t) = p'_{\lambda_n}(|t|)\, \mathrm{sgn}(t),$$
$$g'_n(t) = p''_{\lambda_n}(|t|), \quad \forall t \ne 0.$$

From condition (C6), if $|s|, |t| > C\lambda_n$, then

$$g_n(s) - g_n(t) = g'_n\big(s^*\big)(s - t), \quad s^* \in [s, t]$$
$$= g'_n(t)(s - t) + \big(g'_n\big(s^*\big) - g'_n(t)\big)(s - t)$$
$$= g'_n(t)(s - t) + R_n(s, t),$$

where the residual $|R_n(s,t)| \le D(s - t)^2$. For any nonzero parameter $\beta_n^{*(j)}$, we know from condition (C5) that $|\beta_n^{*(j)}|/\lambda_n \ge c_n/\lambda_n \to +\infty$, and from condition (C4) that

$$|\hat\beta_{n1}^{(j)} - \beta_n^{*(j)}|/\lambda_n \le \big\|\hat\beta_{n1} - \beta_{n1}^*\big\|/\lambda_n \to 0.$$

Hence, $|\hat\beta_{n1}^{(j)}|/\lambda_n \to +\infty$, and

$$p'_{\lambda_n}\big(|\hat\beta_{n1}^{(j)}|\big)\, \mathrm{sgn}\big(\hat\beta_{n1}^{(j)}\big)$$
$$= p'_{\lambda_n}\big(|\beta_{n1}^{*(j)}|\big)\, \mathrm{sgn}\big(\beta_{n1}^{*(j)}\big)$$
$$+ p''_{\lambda_n}\big(|\beta_{n1}^{*(j)}|\big)\big(\hat\beta_{n1} - \beta_{n1}^*\big)$$
$$+ R_n\big(\hat\beta_{n1}, \beta_{n1}^*\big).$$

written in vector form, that is

$$\mathbf{b}_{n1}(\hat\beta_{n1}) = \mathbf{b}_{\lambda_n} + \Sigma_{\lambda_n}\big(\hat\beta_{n1} - \beta_{n1}^*\big) + \mathbf{R}_n, \tag{6}$$

where

$$\|\mathbf{b}_{\lambda_n}\| = \left(\sum_{j=1}^{q_n} |p'_{\lambda_n}\big(|\beta_{n1}^{*(j)}|\big)|^2\right)^{1/2}$$
$$\le |a_n|\sqrt{q_n},$$
$$\big\|\Sigma_{\lambda_n}\big(\hat\beta_{n1} - \beta_{n1}^*\big)\big\| \le \|\Sigma_{\lambda_n}\|\big\|\big(\hat\beta_{n1} - \beta_{n1}^*\big)\big\|$$
$$= |b_n| O_p\big(\sqrt{p_n/n}\big),$$
$$\|\mathbf{R}_n\| \le D\left(\sum_{j=1}^{q_n} |\hat\beta_{n1}^{(j)} - \beta_n^{*(j)}|^4\right)^{1/2}$$
$$\le D\big\|\hat\beta_{n1} - \beta_{n1}^*\big\|^2 = O_p\left(\frac{p_n}{n}\right).$$

In addition, from condition (A5),

$$\big\|n\tau^{-1}\big(S_{n1}^{-1/2}\big)^T \mathbf{R}_n\big\|$$
$$\le nO\left(\frac{1}{\sqrt n}\right) O_p\left(\frac{p_n}{n}\right) = O_p\left(\sqrt{\frac{p_n^2}{n}}\right) = o_p(1).$$

Combine with (5)

$$-\tau^{-1} W_{n1} = \Sigma_n S_{n1}^{1/2}\big(\hat\beta_{n1} - \beta_{n1}^* + \delta_n^*\big) + o_p(1). \tag{7}$$

If condition (C7) and (C8) also holds, then

$$\big\|n\tau^{-1}\big(S_{n1}^{-1/2}\big)^T \mathbf{b}_{n1}(\hat\beta_{n1})\big\|$$
$$\le |a_n|\sqrt{nq_n} + |b_n| O_p(\sqrt{p_n}) + o_p(1) = o_p(1).$$

Combine with (5), we have

$$-\tau^{-1} W_{n1} = S_{n1}^{1/2}\big(\hat\beta_{n1} - \beta_{n1}^*\big) + o_p(1). \tag{8}$$

Together with (7), (8) and lemma A.3, the asymptotic normality and efficiency are proved.

Before we prove Theorem 5.2, we first point out the following facts.

Lemma A.5. *Under condition (A3–A5), (B1–B4) and (C1–C6), we have*

(i) $\|\Sigma_{\lambda_n}\| = o(1)$, $\|\Sigma_{\lambda_n} - \widehat{\Sigma}_{\lambda_n}\| = O_p(\sqrt{\frac{p_n}{n}})$;
(ii) $\|\mathbf{b}_{\lambda_n}\| = O(\sqrt{\frac{p_n}{n}})$, $\|\mathbf{b}_{\lambda_n} - \hat{\mathbf{b}}_{\lambda_n}\| = o_p(\sqrt{\frac{p_n}{n}})$;
(iii) $\|\Sigma_n\| = O(1), \|\Sigma_n^{-1}\| = O(1)$, $\|\widehat{\Sigma}_n\| = O_p(1), \|\widehat{\Sigma}_n^{-1}\| = O_p(1)$. □

Proof of Lemma A.5. First we consider (i). Note that Σ_{λ_n} is a diagonal matrix, from condition (C2), we have $\|\Sigma_{\lambda_n}\| = b_n = o(1)$. Similarly, $(\Sigma_{\lambda_n} - \widehat{\Sigma}_{\lambda_n})$ is also a diagonal matrix, from condition (C6), we have

$$\|\Sigma_{\lambda_n} - \widehat{\Sigma}_{\lambda_n}\| \leq D\|\hat{\beta}_{n1} - \beta^*_{n1}\| = O_p\left(\sqrt{\frac{p_n}{n}}\right).$$

Then we consider (ii). From condition (C1):

$$\|\mathbf{b}_{\lambda_n}\| \leq \sqrt{q_n} a_n = O\left(\sqrt{\frac{p_n}{n}}\right),$$

and we know from (6) that

$$\begin{aligned}\|\mathbf{b}_{\lambda_n} - \hat{\mathbf{b}}_{\lambda_n}\| &\leq \|\Sigma_{\lambda_n}(\hat{\beta}_{n1} - \beta^*_{n1})\| + \|\mathbf{R}_n\| \\ &\leq o(1)O_p\left(\sqrt{\frac{p_n}{n}}\right) + O\left(\frac{p_n}{n}\right) \\ &= o_p\left(\sqrt{\frac{p_n}{n}}\right).\end{aligned}$$

Finally we consider (iii). The property of matrix norm indicates

$$\begin{aligned}\|\Sigma_n\| &\leq \|I_{q_n}\| + n\tau^{-1}\|(S_{n1}^{-1/2})^T\|\|\Sigma_{\lambda_n}\|\|(S_{n1}^{-1/2})\| \\ &= 1 + o(1) = O(1), \\ \|\Sigma_n^{-1}\| &\leq \frac{1}{1 - n\tau^{-1}\|(S_{n1}^{-1/2})^T\|\|\Sigma_{\lambda_n}\|\|(S_{n1}^{-1/2})\|} \\ &= \frac{1}{1 - o(1)} = O(1),\end{aligned}$$

and the conclusion of (i) indicates

$$\|\widehat{\Sigma}_{\lambda_n}\| \leq \|\Sigma_{\lambda_n}\| + \|\Sigma_{\lambda_n} - \widehat{\Sigma}_{\lambda_n}\| = o_p(1).$$

Similarly, we may prove $\|\widehat{\Sigma}_n\| = O_p(1), \|\widehat{\Sigma}_n^{-1}\| = O_p(1)$.

Return to the proof of Theorem 5.2. □

Proof of Theorem 5.2. We first prove the consistency of the covariance matrix estimation. Using A.5, we have from direct calculation that:

$$\begin{aligned}&\|\mathrm{Cov}_n - \widehat{\mathrm{Cov}}_n\| \\ &= \|\tau^{-2}\sigma^2(S_{n1}^{-1/2})^T(\Sigma_n^{-2} - \widehat{\Sigma}_n^{-2})(S_{n1}^{-1/2})\| \\ &\leq O\left(\frac{1}{n}\right)\|\Sigma_n^{-2} - \widehat{\Sigma}_n^{-2}\| \\ &= O\left(\frac{1}{n}\right)\|\Sigma_n^{-2}(\widehat{\Sigma}_n - \Sigma_n)(\widehat{\Sigma}_n + \Sigma_n)\widehat{\Sigma}_n^{-2}\| \\ &\leq O\left(\frac{1}{n}\right)\|\Sigma_n^{-1}\|^2(\|\widehat{\Sigma}_n - \Sigma_n\|)(\|\widehat{\Sigma}_n\| + \|\Sigma_n\|)\|\widehat{\Sigma}_n^{-1}\|^2 \\ &= O_p\left(\frac{1}{n}\right)\|\widehat{\Sigma}_n - \Sigma_n\| \\ &= O_p\left(\frac{1}{n}\right)\|n\tau^{-1}(S_{n1}^{-1/2})^T(\widehat{\Sigma}_{\lambda_n} - \Sigma_{\lambda_n})(S_{n1}^{-1/2})\| \\ &\leq O_p\left(\frac{1}{n}\right)O_p\left(\sqrt{\frac{p_n}{n}}\right) = o_p\left(\frac{1}{n}\right).\end{aligned}$$

Then we prove the consistency of bias estimation. Similarly, we have from direct calculation that

$$\begin{aligned}&\|\delta^*_n - \hat{\delta}^*_n\| \\ &= \|n\tau^{-1}(S_{n1}^{-1/2})\{\Sigma_n^{-1}(S_{n1}^{-1/2})^T\mathbf{b}_{\lambda_n} - \widehat{\Sigma}_n^{-1}(S_{n1}^{-1/2})^T\hat{\mathbf{b}}_{\lambda_n}\}\| \\ &\leq O(\sqrt{n})\|\Sigma_n^{-1}(S_{n1}^{-1/2})^T\mathbf{b}_{\lambda_n} - \widehat{\Sigma}_n^{-1}(S_{n1}^{-1/2})^T\hat{\mathbf{b}}_{\lambda_n}\| \\ &= O(\sqrt{n})\|(\Sigma_n^{-1} - \widehat{\Sigma}_n^{-1})(S_{n1}^{-1/2})^T\mathbf{b}_{\lambda_n} \\ &\quad + \widehat{\Sigma}_n^{-1}(S_{n1}^{-1/2})^T(\mathbf{b}_{\lambda_n} - \hat{\mathbf{b}}_{\lambda_n})\| \\ &\leq O(\sqrt{n})\left\{\|\Sigma_n^{-1} - \widehat{\Sigma}_n^{-1}\|O\left(\frac{1}{\sqrt{n}}\right)O_p\left(\sqrt{\frac{p_n}{n}}\right)\right. \\ &\quad \left. + O\left(\frac{1}{\sqrt{n}}\right)o_p\left(\sqrt{\frac{p_n}{n}}\right)\right\} \\ &= O_p\left(\sqrt{\frac{p_n}{n}}\right)\{\|\Sigma_n^{-1} - \widehat{\Sigma}_n^{-1}\| + o_p(1)\} \\ &= O_p\left(\sqrt{\frac{p_n}{n}}\right)\{\|\Sigma_n^{-1}(\widehat{\Sigma}_n - \Sigma_n)\widehat{\Sigma}_n^{-1}\| + o_p(1)\} \\ &\leq O_p\left(\sqrt{\frac{p_n}{n}}\right)\left\{O_p\left(\sqrt{\frac{p_n}{n}}\right) + o_p(1)\right\} = o_p\left(\sqrt{\frac{p_n}{n}}\right),\end{aligned}$$

and the theorem is proved. □

APPENDIX B

This section provides the oracle inequalities for our proposed estimators. We need the following regularity conditions:

(D1) $\rho(\cdot) \in C^2(\mathbb{R})$ is second order continuously differentiable with $\rho''(\cdot) \geq \rho_0$ for some $\rho_0 > 0$.
(D2) There exists two positive constant c_1, c_2 such that $c_1 \leq \frac{p'_{\lambda_n}(t)}{\lambda_n} \leq c_2$ holds for all $t > 0$.
(D3) $p_{\lambda_n}(0) = 0$ and $p''_{\lambda_n}(\cdot) \geq 0$.
(D4) $\rho'(\epsilon)$ is sub-Gaussian with parameter μ.
(D5) $\max_{1 \leq i \leq p_n} \frac{1}{\sqrt{n}}\|X_i\|_2 \leq C$ for some constant $C > 0$.

The main theorem of oracle inequalities are as follows.

Theorem B.1. *Let $e = (\rho'(\epsilon_1), \ldots, \rho'(\epsilon_n))^T$. Then, under conditions (A4), (D1)–(D3), we have*

(i) *$Q_n(\beta)$ has a unique minimizer $\hat{\beta}_n$. If $\|\frac{Xe}{n}\|_\infty \le c_1\lambda_n$, then $\hat{\beta}_n$ satisfies*

$$\|\hat{\beta}_n - \beta_n^*\|_2 \le \frac{4c_2}{r\rho_0}\sqrt{q_n}\lambda_n.$$

(ii) *If we further assume conditions (D4)–(D5) hold, and we take $\lambda_n = 2\mu C\{\sqrt{\frac{\log p_n}{n}} + \delta\}$ for some $\delta > 0$, then*

$$\|\hat{\beta}_n - \beta_n^*\|_2 \le \frac{4c_2}{r\rho_0}\sqrt{q_n}\lambda_n$$

holds with probability at least $1 - e^{-2n\delta^2}$.

Proof of Theorem B.1(i). Under conditions (D1)–(D3), Q_n is a convex function. Therefore, there Q_n has a unique minimizer $\hat{\beta}_n$. By Taylor expansion, we have

(9)

$$\begin{aligned}
0 &\ge Q_n(\hat{\beta}_n) - Q_n(\beta_n^*)\\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\rho\big(\epsilon_k + x_k^T(\beta_n^* - \hat{\beta}_n)\big) - \rho(\epsilon_k)\right]\\
&\quad - \sum_{j=1}^{q_n} p_{\lambda_n}(|\beta_{nj}^*|) + \sum_{j=1}^{q_n} p_{\lambda_n}(|\hat{\beta}_{nj}|) + \sum_{j=q_n+1}^{p_n} p_{\lambda_n}(|\hat{\beta}_{nj}|)\\
&= \frac{1}{n}\sum_{k=1}^{n}\rho'(\epsilon_k)x_k^T(\beta_n^* - \hat{\beta}_n)\\
&\quad + \frac{1}{2n}\sum_{k=1}^{n}\rho''(\epsilon_k + t\gamma_n)(\beta_n^* - \hat{\beta}_n)^T x_k x_k^T(\beta_n^* - \hat{\beta}_n)\\
&\quad - \sum_{j=1}^{q_n} p_{\lambda_n}(|\beta_{nj}^*|) + \sum_{j=1}^{q_n} p_{\lambda_n}(|\hat{\beta}_{nj}|) + \sum_{j=q_n+1}^{p_n} p_{\lambda_n}(|\hat{\beta}_{nj}|).
\end{aligned}$$

Under conditions (A4) and (D1), we have

$$\begin{aligned}
&\frac{1}{2n}\sum_{k=1}^{n}\rho''(\epsilon_k + t\gamma_n)(\beta_n^* - \hat{\beta}_n)^T x_k x_k^T(\beta_n^* - \hat{\beta}_n)\\
&\ge \frac{\rho_0 r}{2}\|\beta_n^* - \hat{\beta}_n\|_2^2,
\end{aligned}$$

and under the assumption of Theorem B.1, we have

$$\begin{aligned}
\left|\frac{1}{n}\sum_{k=1}^{n}\rho'(\epsilon_k)x_k^T(\beta_n^* - \hat{\beta}_n)\right| &\le \left\|\frac{Xe}{n}\right\|_\infty \|\hat{\beta}_n - \beta_n^*\|_1\\
&\le c_1\lambda_n\|\hat{\beta}_n - \beta_n^*\|_1.
\end{aligned}$$

Plugging the above inequalities into (9), together with condition (D3) yields

$$\begin{aligned}
&\frac{1}{2}\rho_0 r\|\beta_n^* - \hat{\beta}_n\|_2^2\\
&\le c_1\lambda_n\sum_{j=1}^{q_n}|\beta_{nj}^* - \hat{\beta}_{nj}| + c_1\lambda_n\sum_{j=q_n+1}^{p_n}|\hat{\beta}_{nj}|\\
&\quad - \sum_{j=1}^{q_n} p_{\lambda_n}(|\beta_{nj}^*|) + \sum_{j=1}^{q_n} p_{\lambda_n}(|\hat{\beta}_{nj}|) + \sum_{j=q_n+1}^{p_n} p_{\lambda_n}(|\hat{\beta}_{nj}|)\\
&= c_1\lambda_n\sum_{j=1}^{q_n}|\beta_{nj}^* - \hat{\beta}_{nj}| + c_1\lambda_n\sum_{j=q_n+1}^{p_n}|\hat{\beta}_{nj}|\\
&\quad - \sum_{j=1}^{q_n} p'_{\lambda_n}(\theta_{j1})(|\beta_{nj}^*| - |\hat{\beta}_{nj}|) + \sum_{j=q_n+1}^{p_n} p'_{\lambda_n}(\theta_{j2})|\hat{\beta}_{nj}|,
\end{aligned}$$

where θ_{j1} is some point between $|\beta_{nj}^*|$ and $|\hat{\beta}_{nj}|$, θ_{j2} is some point in $[0, |\hat{\beta}_{nj}|]$. Using condition (D2), we further have

$$\begin{aligned}
\frac{1}{2}\rho_0 r\|\beta_n^* - \hat{\beta}_n\|_2^2 &\le (c_1 + c_2)\lambda_n\sum_{i=1}^{q_n}|\hat{\beta}_{nj} - \beta_{nj}^*|\\
&\le 2c_2\lambda_n\sqrt{q_n}\|\hat{\beta}_n - \beta_n^*\|_2,
\end{aligned}$$

where the last inequality is obtained by Cauchy-Schwarz inequality. Then the result of Theorem B.1(i) follows. □

Proof of Theorem B.1(ii). Condition (D4) implies $\rho'(\epsilon_k)$ is sub-Gaussian with parameter μ, therefore, $\frac{1}{n}\sum_{k=1}^{n}\rho'(\epsilon_k)x_{ki}$ is sub-Gaussian with parameter $\mu(\sum_{k=1}^{n}x_{ki}^2)^{1/2}$ for all $i \in \{1, \ldots, p_n\}$. By Hoeffding inequality and condition (D5), we have

$$\begin{aligned}
\mathbb{P}\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\rho'(\epsilon_k)x_{ki}\right| \ge \lambda_n\right\} &\le 2\exp\left\{-\frac{n^2t^2}{2\mu^2\sum_{k=1}^{n}x_{ki}^2}\right\}\\
&\le 2\exp\left\{-\frac{nt^2}{2\mu^2C^2}\right\}.
\end{aligned}$$

It follows that

$$\mathbb{P}\left\{\max_{1\le i\le p_n}\left|\frac{1}{n}\sum_{k=1}^{n}\rho'(\epsilon_k)x_{ki}\right| \ge \lambda_n\right\} \le 2p_n\exp\left\{-\frac{nt^2}{2\mu^2C^2}\right\}.$$

Take $\lambda_n = 2C\mu\{\sqrt{\frac{\log p_n}{n}} + \delta\}$ and Theorem B.1(ii) follows. □

APPENDIX C

More simulation results are shown in Table 2 and Table 3, where the meaning of the notations in Table 2 are as follows:

- Bias: the empirical bias of the estimate.
- SE: the empirical standard error.
- SEE: the estimated standard error.
- t.SE: theoretical asymptotic standard error.

Table 3 shows the simulation result of variable selection procedure, where the meaning of the notations in Table 3 are as follows:

- CF: the rate of correct fit.
- FPR: false positive rate.
- FNR: false negative rate.
- ME: model error.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 12271286 and No. 11931001).

Received 15 November 2022

REFERENCES

[1] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed. Wiley, New York. MR1700749

[2] Bourbaki, N. (1966). *General Topology* **1**. Addison–Wesley, Reading, MA.

[3] Ethier, S. N. and Kurtz, T. G. (1985). *Markov Processes: Characterization and Convergence*. Wiley, New York. MR838085

[4] Prokhorov, Yu. (1956). Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.* **1** 157–214. MR84896

[5] Hu, T. and Liang, B. (2021). A new class of estimators based on a general relative loss function. *Mathematics* **9** 1138.

[6] Chen, K., Guo, S., Lin, Y., and Ying, Z. (2010). Least absolute relative error estimation. *Journal of the American Statistical Association* **105** 1104–1112. MR2752606

[7] Ding, H., Wang, Z., and Wu, Y. (2018). A relative error-based estimation with an increasing number of parameters. *Communications in Statistics. Theory and Methods* **47** 196–209. MR3765067

[8] Li, G., Peng, H., and Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica* **21** 391–419. MR2796868

[9] Li, Z., Liu, Y., and Liu, Z. (2017). Empirical likelihood and general relative error criterion with divergent dimension. *Statistics* **51** 1006–1022. MR3698498

[10] Xia, X., Liu, Z., and Yang, H. (2016). Regularized estimation for the least absolute relative error models with a diverging number of covariates. *Computational Statistics and Data Analysis* **96** 104–119. MR3433254

[11] Yang, Y. and Ye, F. (2013). General relative error criterion and M-estimation. *Frontiers of Mathematics in China* **8** 695–715. MR3044675

[12] Zhang, Q. and Wang, Q. (2013). Local least absolute relative error estimating approach for partially linear multiplicative model. *Statistica Sinica* **23** 1091–1116. MR3114706

[13] Chen, Y., Liu, H., and Ma, J. (2022). Local least product relative error estimation for single-index varying-coefficient multiplicative model with positive responses. *Journal of Computational and Applied Mathematics* **415** Paper No. 114478. MR4441358

[14] Wang, Z., Chen, Z., and Wu, Y. (2017). A relative error estimation approach for multiplicative single index model. *Journal of Systems Science and Complexity* **30** 1160–1172. MR3679384

[15] Chen, X. and Zhao, L. (1996). *M-methods in Linear Model*. Shanghai Scientific and Technical Publishers, Shanghai. (In Chinese)

[16] Fan, R., Zhang, S., and Wu, Y. (2021). Nonconcave penalized M-estimation for the least absolute relative errors model. *Communications in Statistics. Theory and Methods* 1–18. MR4541774

[17] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961. MR2065194

[18] Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* **1** 799–821. MR0356373

[19] Chen, K., Ying, Z., Zhang, H., and Zhao, L. (2008). Analysis of least absolute deviation. *Biometrika* **95** 107–122. MR2394499

[20] Hao, M., Lin, Y., and Zhao, X. (2016). A relative error-based approach for variable selection. *Computational Statistics & Data Analysis* **103** 250–262. MR3522631

[21] Chen, K., Lin, Y., Wang, Z., and Ying, Z. (2016). Least product relative error estimation. *Journal of Multivariate Analysis* **144** 91–98. MR3434942

[22] Narula, S. C., and Wellington, J. F. (1977). Prediction, linear regression and the minimum sum of relative errors. *Technometrics* **19** 185–190.

[23] Khoshgoftaar, T. M., Bhattacharyya, B. B., and Richardson, G. D. (1992). Predicting software errors, during development, using nonlinear regression models: a comparative study. *IEEE Transactions on Reliability* **41** 390–395.

[24] Park, H., and Stefanski, L. A. (1998). Relative-error prediction. *Statistics & Probability Letters* **40** 227–236. MR1650001

[25] Huang, J., Ma, S., and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62**, 813–820. MR2247210

[26] Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society: Series B* **70**, 351–370. MR2424757

[27] Cai, T., Huang, J., and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65**, 394–404. MR2751463

[28] Xu, J., Leng, C., and Ying, Z. (2010). Rank-based variable selection with censored data. *Statistics and Computing* **20**, 165–176. MR2610770

Fei Ye
School of Statistics
Capital University of Economics and Business
Beijing 100070
China
E-mail address: yefei@cueb.edu.cn

Hongyi Zhou
Department of Mathematics
Tsinghua University
Beijing 100084
China
E-mail address: zhou-hy21@mails.tsinghua.edu.cn

Ying Yang
Department of Mathematics
Tsinghua University
Beijing 100084
China
E-mail address: yingyang@mail.tsinghua.edu.cn