JUSTC

Article

http://justc.ustc.edu.cn

Received: July 07, 2022; Accepted: October 08, 2022

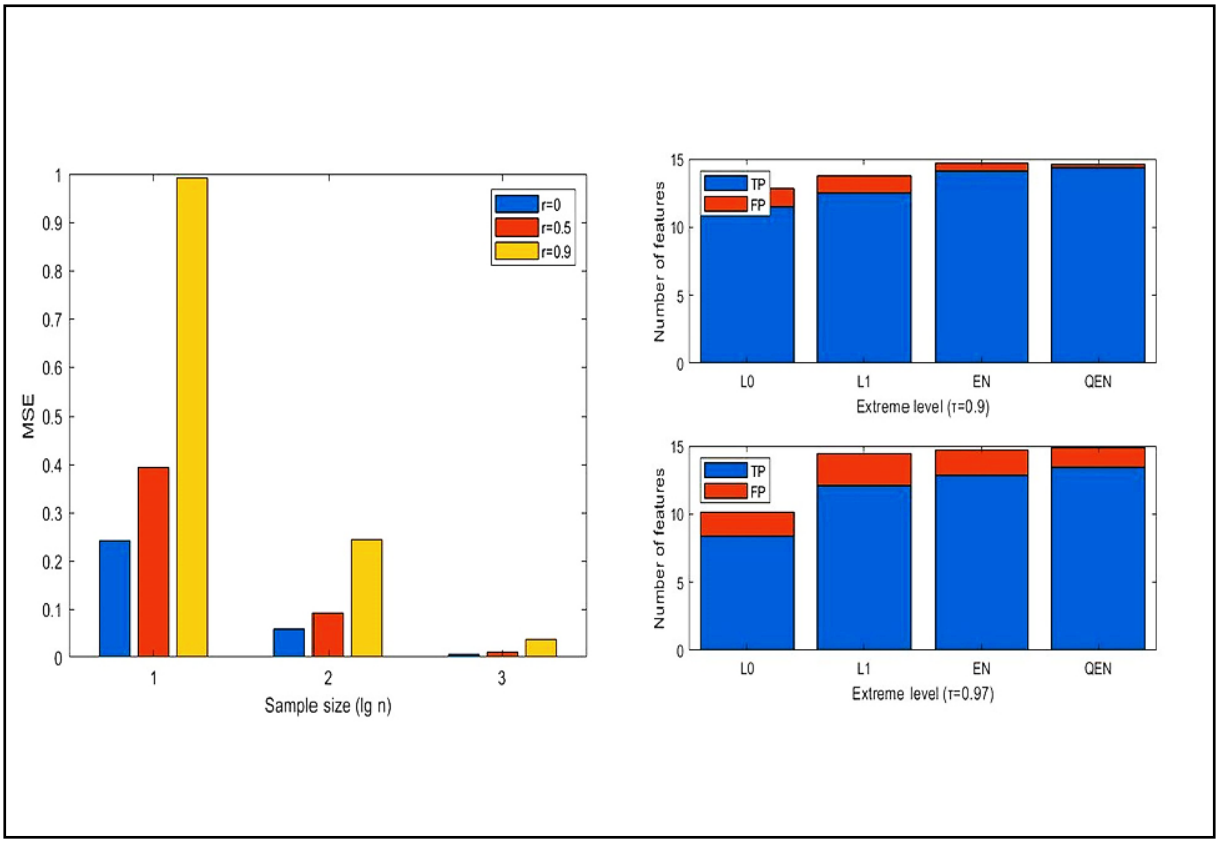
Variable selection in high-dimensional extremile regression via the quasi elastic net

Yimin Xiong, Zhi Zheng, and Weiping Zhang ✉

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Weiping Zhang, E-mail: zwp@ustc.edu.cn

Graphical abstract



Relationship between the MSE of estimators in QEN penalized extremile regression and sample size n with τ = 0.5 (left) and TP and FP in different penalized extremile regressions with high-dimensional and grouped data (right).

Public summary

- We propose a quasi elastic net penalized linear extremile regression to deal with high-dimensional data, which leads to a sparse solution as well as being suitable for strongly collinear situations.
- We adopt an EM algorithm to solve the L_0 approximation problem efficiently, and further solve the quasi elastic net penalized optimization problem.
- We prove that the proposed quasi elastic net penalized linear extremile regression model is effective through numerical studies.

Citation: Xiong Y M, Zheng Z, Zhang W P. Variable selection in high-dimensional extremile regression via the quasi elastic net. *JUSTC*, **2023**, 53(2): 1. DOI: 10.52396/JUSTC-2022-0099

JUSTC

Article

http://justc.ustc.edu.cn

Received: July 07, 2022; Accepted: October 08, 2022

Variable selection in high-dimensional extremile regression via the quasi elastic net

Yimin Xiong, Zhi Zheng, and Weiping Zhang ✉

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Weiping Zhang, E-mail: zwp@ustc.edu.cn

Cite This: *JUSTC*, **2023**, 53(2): 1 (10pp)

Read Online

Abstract: Extremile regression proposed in recent years not only retains the advantage of quantile regression that can fully show the information of sample data by setting different quantiles, but also has its own superiority compared with quantile regression and expectile regression, due to its explicit expression and conservativeness in estimating. Here, we propose a linear extremile regression model and introduce a variable selection method using a penalty called a quasi elastic net (QEN) to solve high-dimensional problems. Moreover, we propose an EM algorithm and establish corresponding theoretical properties under some mild conditions. In numerical studies, we compare the QEN penalty with the L_0, L_1, L_2 and elastic net penalties, and the results show that the proposed method is effective and has certain advantages in analysis.

Keywords: extremile regression; quasi elastic net; grouping effect; high-dimensional data; variable selection

CLC number: O212.4 **Document code:** A

2020 Mathematics Subject Classification: 62H25

1 Introduction

Although ordinary least squares (OLS) regression has been one of the most important methods in regression analysis by estimating the mean value, it sometimes shows poor robustness to outliers. Koenker and Bassett[1] proposed least absolute deviation regression (LADR) to estimate quantiles and can effectively address this problem and analyze sample data by setting different quantiles. Due to the nondifferentiability of the absolute value function, it is not convenient to solve the optimization problems. Then Newey and Powell[2] proposed asymmetric least squares (ALS) regression leading to expectiles. Considering that expectile regression lacks an explicit solution, Daouia et al.[3] proposed a weighted least square regression and defined extremiles. From their discussion, extremiles show their conceptual simplicity, convenient calculation and good properties, and they are more appropriate to be a risk protection method in tail analysis than quantiles and expectiles.

Consider a response Y satisfying $E|Y|<\infty$ and its cumulative distribution function F. For $\tau\in(0,1)$, the unconditional τth extremile of Y is defined as

$$\xi_\tau=\operatorname{argmin}_\theta E[J_\tau(F(Y))\cdot(Y-\theta)^2], \tag{1}$$

where $J_\tau(\cdot)=K'_\tau(\cdot)$ and

$$K_\tau(t)=\begin{cases}1-(1-t)^{s(\tau)}, & 0<\tau\leqslant\dfrac{1}{2};\\ t^{s(1-\tau)}, & \dfrac{1}{2}<\tau<1,\end{cases} \tag{2}$$

with $s(\tau)=\dfrac{\log(1/2)}{\log(1-\tau)}$. Under this definition, the unconditional τth quantile of Y can be obtained from

$$q_\tau\in\operatorname{argmin}_\theta E[J_\tau(F(Y))\cdot|Y-\theta|]. \tag{3}$$

For (3), we can consider its integral form

$$q_\tau\in\operatorname{argmin}_\theta\int_{-\infty}^{\infty}J_\tau(F(Y))\cdot|Y-\theta|\mathrm{d}F(Y)=\operatorname{argmin}_\theta\Big[\int_{-\infty}^{\theta}J_\tau(F(Y))\cdot(\theta-Y)\mathrm{d}F(Y)+\int_{\theta}^{\infty}J_\tau(F(Y))\cdot(Y-\theta)\mathrm{d}F(Y)\Big], \tag{4}$$

which means q_τ shall satisfy

$$-\int_{-\infty}^{q_\tau}J_\tau(F(Y))\mathrm{d}F(Y)+\int_{q_\tau}^{\infty}J_\tau(F(Y))\mathrm{d}F(Y)=0\Rightarrow K_\tau(F(q_\tau))=\frac{1}{2}(K_\tau(F(-\infty))+K_\tau(F(\infty)))=\frac{1}{2}(K_\tau(0)+K_\tau(1))=\frac{1}{2}, \tag{5}$$

when $\dfrac{1}{2}<\tau<1$, that is,

$$(F(q_\tau))^{s(1-\tau)}=(F(q_\tau))^{\frac{\log(1/2)}{\log\tau}}=\frac{1}{2}\Rightarrow F(q_\tau)=\tau. \tag{6}$$

We can obtain the same result when $0<\tau\leqslant\dfrac{1}{2}$, which means (3) is equivalent to the quantile under the definition of $K'_\tau(\cdot)$.

As we know, for the random variable Y, the special case $\tau=\dfrac{1}{2}$ of quantile and expectile leads to its median and mean, respectively. Then if we define a random variable Z with its cumulative distribution function $F_Z=K_\tau(F)$, it can be seen that the τth quantile and τth extremile represent the median and mean of Z, respectively, similar to the central behavior of quantiles and expectiles.

DOI: 10.52396/JUSTC-2022-0099

JUSTC, **2023**, 53(2): 1

Simultaneously, with the development of the internet and related industries, high-dimensional data are becoming increasingly common with an explosion of computations in data analysis, so it is supposed to find effective methods to reduce the computation if we apply extremile regression under this background. Variable selection is one of the effective measures, and there have been different methods for selecting variables.

First, there are methods based on the criteria, such as the PRESS Criterion[4], C_p Criterion[5], Akaike Information Criterion (AIC)[6] and Bayesian Information Criterion (BIC)[7]. Those methods select variables through Selection/Estimation steps and the deviation of the Selection step will affect the result of the Estimation step due to the unknowns of the correct variables. Then, Geisser and Eddy[8] proposed the cross validation (CV) method to select variables without assumptions in parameter estimation, which has spawned hold-out validation[9], 5×2 cross validation[10] and other methods. Moreover, Candes and Tao[11] proposed the Dantzig selector (DS) method. Although the DS method does not rely on a specific function, it needs to select variables based on the assumption of sparsity of unknown coefficients. Because the DS method has the same compression for each unknown coefficient, it may lead to the overcompression of important coefficients. Dicker and Lin[12] proposed the ADS method with different weights. Later, James et al.[13] proposed the DS method of generalized linear models, and Antoniadis et al.[14] extended the DS method to Cox models. Furthermore, Fan and Lv[15] proposed a sure independence screening (SIS) method to deal with ultrahigh dimensional data, while Fan and Li[16] proposed an improved iterative sure independence screening (ISIS) method to deal with the collinearity between predictors. These methods can ensure selecting important predictors and independently measure the correlations between each predictor and the response. They can apply the marginal regression method to variable coefficient models and additive models. Finally, the variable selection method based on penalties has been widely used, and it solves corresponding optimization problems to achieve the goal. It can improve the calculation efficiency and increase the accuracy of estimation. The commonly used penalties include $L_p, p\in(0,1)$ [17, 18], lasso[19], SCAD[20], MCP[21] and elastic net[22]. Generally, the L_0 penalty is the most essential sparsity measure that penalizes the number of nonzero parameters directly, but it is NP-hard to solve an optimization problem with the L_0 penalty. Although L_1 is one of the common replacements for its convexity in the vast majority of cases, there exist some limitations, such as it might sometimes choose the wrong model[23].

In this paper, we use the quasi elastic net, which combines L_0 and L_2 penalties, to select variables in linear extremile regression and propose an efficient EM algorithm based on Liu and Li[24], which can approximately solve the optimization problem with the L_0 penalty. Significantly, if the predictors are highly correlated, the L_0 penalty would lead to poor performance; thus, we introduce ridge regression[25]. With the properties of L_2, the quasi elastic net is able to solve those highly correlated predictors and encourages a grouping effect. Moreover, we also establish several theoretical properties under some mild conditions.

The major contributions in this paper are as follows. First, we apply the proposed model to analyze high-dimensional data and retain its conceptual simplicity and convenient calculation. Second, we propose the quasi elastic net combining L_0 and L_2 penalties, which can both lead to a sparse solution and deal with highly correlated predictors. Third, we propose an efficient EM algorithm to solve the optimization problem with the L_0 penalty approximately and establish the theoretical properties.

The remainder of this paper is organized as follows. Section 2 shows our model and explains the method. Section 3 provides the theoretical properties of this method. Sections 4 and 5 present the comparisons between different methods in simulations and real data. Section 6 presents the conclusion. The appendix provides all proofs of theorems and lemmas.

2 Linear extremile regression with QEN

Considering an $n\times1$ response $\boldsymbol{y}=(y_1,\cdots,y_n)^{\mathrm{T}}$ and $n\times p$ predictor matrix $X=(\boldsymbol{x}_1^{\mathrm{T}},\cdots,\boldsymbol{x}_n^{\mathrm{T}})^{\mathrm{T}}$, the linear τth extremile regression with a quasi elastic net penalty is defined as

$$\hat{\boldsymbol{\beta}}_\tau=\mathrm{argmin}_{\beta_\tau}L(\boldsymbol{\beta}_\tau)=$$
$$\mathrm{argmin}_{\beta_\tau}\frac{1}{n}\sum_{i=1}^{n}J_\tau(\hat{F}(y_i|X))(y_i-\boldsymbol{x}_i\boldsymbol{\beta}_\tau)^2+\lambda_1\|\boldsymbol{\beta}_\tau\|_0+\lambda_2\|\boldsymbol{\beta}_\tau\|_2^2, \quad (7)$$

where $\hat{F}(\cdot|X)$ is an estimated distribution function of $\boldsymbol{y}$ depends on X, and we prefer nonparametric and semiparametric methods, especially a partial-linear single-index model[26], when dealing with high-dimensional data. $\boldsymbol{\beta}_\tau=(\beta_{\tau,1},\cdots,\beta_{\tau,p})^{\mathrm{T}}\in\mathbb{R}^{p\times1}$ are unknown coefficients, and $\lambda_1,\lambda_2>0$ are tuning parameters.

Let $R\subset\{1,\cdots,p\}$ be the subset index with $\beta_j\neq0$ and $O\subset\{1,\cdots,p\}$ be the subset index with $\beta_j=0$, then we have $R\cup O=\{1,\cdots,p\}$. Denoting $W=\mathrm{diag}(J_\tau(\hat{F}(y_1|X))/n,\cdots,J_\tau(\hat{F}(y_n|X))/n)$, $\boldsymbol{y}^*=(\boldsymbol{y}^{\mathrm{T}}W^{1/2},\boldsymbol{0})^{\mathrm{T}}\in\mathbb{R}^{(n+p)\times1}$, $X^*=(1+\lambda_2)^{-1/2}\cdot(X^{\mathrm{T}}W^{1/2},\lambda_2^{1/2}I)^T\in\mathbb{R}^{(n+p)\times p}$, $\boldsymbol{\beta}_\tau^*=(1+\lambda_2)^{1/2}\boldsymbol{\beta}_\tau$, the optimization problem can be rewritten as

$$\hat{\boldsymbol{\beta}}_\tau^*=\mathrm{argmin}_{\beta_\tau^*}\tilde{L}(\boldsymbol{\beta}_\tau^*)=$$
$$\mathrm{argmin}_{\beta_\tau^*}(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)^{\mathrm{T}}(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)+\lambda_1\|\boldsymbol{\beta}_\tau^*\|_0. \quad (8)$$

Note that the solution of (7) satisfies $\hat{\boldsymbol{\beta}}_\tau=(1+\lambda_2)^{-1/2}\hat{\boldsymbol{\beta}}_\tau^*$. Then, we focus on the optimization problem (8) and obtain the following two equations:

$$\tilde{L}(\boldsymbol{\beta}_\tau^*)=(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)^T(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)+\lambda_1\sum_{j\in R}\frac{\beta_{\tau,j}^{*2}}{\delta_j^2},$$
$$\boldsymbol{\delta}=\boldsymbol{\beta}_\tau^*. \quad (9)$$

For the first equation, $\tilde{L}(\boldsymbol{\beta}_\tau^*)$ is a quadratic function of $\boldsymbol{\beta}_\tau^*$ when $\boldsymbol{\delta}$ is known, so when $j\in R$ the first-order partial derivative is as follows:

$$\frac{\partial\tilde{L}(\boldsymbol{\beta}_\tau^*)}{\partial\beta_{\tau,j}^*}=-2\boldsymbol{x}_j^{*\mathrm{T}}(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)+2\lambda_1\frac{\beta_{\tau,j}^*}{\delta_j^2}=0, \quad (10)$$

where $\boldsymbol{x}_j^*$ is the jth column of X^*. We can rewrite (10) as

$$\delta_j^2\boldsymbol{x}_j^{*\mathrm{T}}(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)-\lambda_1\beta_{\tau,j}^*=0. \quad (11)$$

DOI: 10.52396/JUSTC-2022-0099
JUSTC, 2023, 53(2): 1

In addition, when $j \in O$, whick means $\beta^*_{\tau,j} = \delta_j = 0$, we also have

$$\delta_j^2 \boldsymbol{x}_j^{*\mathrm{T}}(\boldsymbol{y}^* - X^*\boldsymbol{\beta}^*_\tau) - \lambda_1 \beta^*_{\tau,j} = 0. \tag{12}$$

Hence, when we let $D = \mathrm{diag}(\delta_1^2, \cdots, \delta_p^2)$, the equations above are equivalent to the following matrix form:

$$DX^{*\mathrm{T}}(\boldsymbol{y}^* - X^*\boldsymbol{\beta}^*_\tau) - \lambda_1 \boldsymbol{\beta}^*_\tau = \boldsymbol{0}. \tag{13}$$

Finally, Eq. (13) leads to the solution

$$\begin{aligned} \boldsymbol{\beta}^*_\tau &= (DX^{*\mathrm{T}}X^* + \lambda_1 I)^{-1} DX^{*\mathrm{T}}\boldsymbol{y}^* = \\ &\quad (1+\lambda_2)^{\frac{1}{2}}(D(X^{\mathrm{T}}WX + \lambda_2 I) + \lambda_1(1+\lambda_2)I)^{-1}(DX^{\mathrm{T}}W\boldsymbol{y}), \\ \boldsymbol{\delta} &= \boldsymbol{\beta}^*_\tau. \end{aligned} \tag{14}$$

It is clear that these two equations can be treated as the M-step and E-step of the EM algorithm. Naturally, we can obtain the solution of (7) by $\hat{\boldsymbol{\beta}}_\tau = (1+\lambda_2)^{-1/2}\hat{\boldsymbol{\beta}}^*_\tau$. Unlike elastic net, our L_0 penalty will not be affected by $\hat{\boldsymbol{\beta}}_\tau = (1+\lambda_2)^{-1/2}\hat{\boldsymbol{\beta}}^*_\tau$ as L_1 penalty, which means $\hat{\boldsymbol{\beta}}_\tau$ will not incur a double amount of shrinkage, and there is no need to correct the estimator $\hat{\boldsymbol{\beta}}_\tau$. This is one of the reasons why we choose the L_0 penalty rather than L_1.

According to the content described above, it indicates that the last two equations can be processed by the EM algorithm. Based on this, the EM algorithm is summarized as Algorithm 1.

Note that the initialization of $\boldsymbol{\beta}$ is not the only one specified, and we tend to choose the solution of Eq. (7) with $\lambda_1 = 0$ which provides a faster convergence rate. However, remarkably, there also exists a trivial solution $\boldsymbol{\beta} = \boldsymbol{0}$ of Eq. (14).

Determination of λ_1 and λ_2. To implement the algorithm, the tuning parameters λ_1 and λ_2 must be chosen. Because of the huge advantage of L_0, we can directly determine λ_1 by the variable selection criteria AIC ($n > p$) or BIC ($n \ll p$) with $\lambda_1 = 2$ or $\log n$ under the optimization problem (8). Then, we pick a grid of values for λ_2, such as a logspace sequence $(0.01, 0.1, 1, 10, 100)$, and compare those (λ_1, λ_2) by k-fold (usually tenfold) cross validation under the optimization problem (7).

3 Theoretical properties

In this section, we establish theoretical properties of the linear extremile regression with QEN, including the convergency of the proposed EM algorithm, grouping effect and consistency of the estimator.

Algorithm 1 The EM algorithm

Given λ_1, λ_2, small number ε, and training data $\{X, \boldsymbol{y}\}$

Initialize $\boldsymbol{\beta} = (1+\lambda_2)^{1/2}(X^{\mathrm{T}}WX + \lambda_2 I)^{-1}(X^{\mathrm{T}}W\boldsymbol{y}), \boldsymbol{\delta} = \boldsymbol{0}$

while $\|\boldsymbol{\beta} - \boldsymbol{\delta}\|_2 \geqslant \varepsilon$ **do**

E-step: $\boldsymbol{\delta} = \boldsymbol{\beta}$

$D = \mathrm{diag}(\delta_1^2, \cdots, \delta_p^2)$

M-step: $\boldsymbol{\beta} = (1+\lambda_2)^{1/2}(D(X^{\mathrm{T}}WX + \lambda_2 I) + \lambda_1(1+\lambda_2)I)^{-1}(DX^{\mathrm{T}}W\boldsymbol{y})$

end while

return $\boldsymbol{\beta} = (1+\lambda_2)^{-1/2}\boldsymbol{\beta}$

3.1 Convergency of the algorithm

This subsection pays attention to the conditions the proposed EM algorithm should satisfy, and we have the following Theorem 3.1 and Lemma 3.1 to ensure its convergency.

Theorem 3.1. Assume the response $\boldsymbol{y}$, the predictors X, initialize solution $\boldsymbol{\beta}^{*(0)}_\tau = (\beta^{*(0)}_{\tau,1}, \cdots, \beta^{*(0)}_{\tau,p})$. Let $D = \mathrm{diag}(\beta^{*(0)}_{\tau,1}, \cdots, \beta^{*(0)}_{\tau,p})$. When the tuning parameters λ_1 and λ_2 satisfy

$$2\lambda_1(1+\lambda_2)^{\frac{3}{2}}\|(D(X^{\mathrm{T}}WX + \lambda_2 I) + \lambda_1(1+\lambda_2)I)^{-2}\|_\infty \|D^{\frac{1}{2}}X^{\mathrm{T}}W\boldsymbol{y}\|_\infty < 1, \tag{15}$$

the estimator in (9) will converge to an optimal solution.

Lemma 3.1. Assume $\tilde{\boldsymbol{x}}_i^{\mathrm{T}} W \tilde{\boldsymbol{x}}_j = 0, \forall i \neq j \in \{1, \cdots, p\}$, the maximum value of λ_1 will meet the condition

$$\lambda_{1_{\max}} = \min_j \left\{ \frac{(\tilde{\boldsymbol{x}}_j^{\mathrm{T}} W \boldsymbol{y})^2}{4(\tilde{\boldsymbol{x}}_j^{\mathrm{T}} W \tilde{\boldsymbol{x}}_j + \lambda_2)} \right\}. \tag{16}$$

Note that both Theorem 3.1 and Lemma 3.1 provide mild restrictions of λ_1 and λ_2. Moreover, Theorem 3.1 indicates that the EM algorithm will lead the estimator to converge to an optimal solution under proper λ_1, λ_2 and initial solution $\boldsymbol{\beta}^{*(0)}_\tau$. Note that the initial solution should avoid the trivial solution $\boldsymbol{0}$, and we generally choose the solution of extremile regression with the L_2 penalty as an initial solution to accelerate convergence.

Theorem 3.2 shows the relationship between the final solution of the proposed EM algorithm and the optimization problem with the L_0 penalty.

Theorem 3.2. Given proper $\boldsymbol{\beta}^{*(0)}_\tau$, the final solution of the proposed EM algorithm is an optimal solution to the L_0 approximation problem that minimizes (8).

3.2 Grouping effect

In $n \ll p$ problems[27], the grouped variables are sometimes important, which has been discussed by Zou and Hastie[22]. There have been many methods in the literature, for instance, using principal component analysis to find a set of highly correlated genes[28], using supervised learning methods to select groups of predictive genes[29], and using regularized regression to find the grouped genes[30]. The proposed QEN encourages a grouping effect as elastic net does under certain conditions, which means the regression coefficients tend to be equal if the corresponding observation predictors are highly correlated. There is the following lemma.

Lemma 3.2. Consider weighted optimization problem as

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} w_i |y_i - \boldsymbol{x}_i\boldsymbol{\beta}|^2 + \lambda f(\boldsymbol{\beta}), \tag{17}$$

where $w_i > 0, \forall i = 1, \cdots, m$, are the weight values, $\lambda > 0$ is a tuning parameter, $f(\cdot)$ is a positive function for $\boldsymbol{\beta} \neq \boldsymbol{0}$.

Assuming $\tilde{\boldsymbol{x}}_i = \tilde{\boldsymbol{x}}_j$, $i \neq j \in \{1, \cdots, p\}$, where $\tilde{\boldsymbol{x}}_i$ is the ith column of $X = (\boldsymbol{x}_1^{\mathrm{T}}, \cdots, \boldsymbol{x}_n^{\mathrm{T}})^{\mathrm{T}}$, we have these two inferences:

(a) If $f(\cdot)$ is strictly convex, then $\hat{\beta}_i = \hat{\beta}_j$, $\forall \lambda > 0$.

(b) If $f(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_0$, then $\hat{\beta}_i\hat{\beta}_j \geqslant 0$ and the solution of (17) will be unique if and only if $\hat{\beta}_i = \hat{\beta}_j = 0$. Furthermore, if $\hat{\beta}_i\hat{\beta}_j > 0$, $\hat{\boldsymbol{\beta}}^*$ is another minimizer, where

DOI: 10.52396/JUSTC-2022-0099
JUSTC, **2023**, 53(2): 1

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k, & k \neq i \text{ and } k \neq j; \\ s\hat{\beta}_i + (1-s)\hat{\beta}_j, & k = i; \\ (1-s)\hat{\beta}_i + s\hat{\beta}_j, & k = j; \end{cases} \tag{18}$$

for any $s \in (0,1)$ and if there is one, and only one 0 in $\hat{\beta}_i$ and $\hat{\beta}_j$, we can set s to 0 in (18) to obtain another minimizer.

(c) If $f(\cdot)$ is the QEN penalty, then $\hat{\beta}_i\hat{\beta}_j \geqslant 0$. Furthermore, if $\hat{\beta}_i\hat{\beta}_j > 0$, then $\hat{\beta}_i = \hat{\beta}_j$.

This lemma shows the motivation to choose the QEN penalty rather than L_0, as L_0 penalty may not lead to a unique solution. Although the proposed QEN penalty is not strictly convex, it can guarantee the grouping effect if the unknown coefficients are nonzero. Then, Theorem 3.3 indicates the relationship between unknown coefficients.

Theorem 3.3. Assume the response $\boldsymbol{y}$ satisfies $\boldsymbol{y}^{\mathrm{T}}W\boldsymbol{y} = 1$ and the observation variables matrix X satisfies

$$\tilde{\boldsymbol{x}}_i^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j = \begin{cases} 1, & i = j; \\ \rho_{ij}, & i \neq j, \end{cases} \tag{19}$$

where $\rho_{ij} \in (-1,1)$. Supposing $\hat{\boldsymbol{\beta}}_\tau = (\hat{\beta}_{\tau,1}, \cdots, \hat{\beta}_{\tau,p})^{\mathrm{T}}$ is the minimizer of the optimization problem (7), the following inequality can be obtained when $\hat{\beta}_{\tau,i}\hat{\beta}_{\tau,j} \neq 0, i \neq j \in \{1, \cdots, p\}$, with probability tending to 1,

$$D(i,j) \equiv \frac{|\hat{\beta}_{\tau,i} - \hat{\beta}_{\tau,j}|}{\sqrt{\boldsymbol{y}^{\mathrm{T}}W\boldsymbol{y}}} \leqslant \frac{\sqrt{2(1-\rho_{ij})}}{\lambda_2}. \tag{20}$$

The distance function $D(i,j)$ indicates the difference between $\hat{\beta}_{\tau,i}$ and $\hat{\beta}_{\tau,j}$, especially when ρ_{ij} approaches 1, the difference between $\hat{\beta}_{\tau,i}$ and $\hat{\beta}_{\tau,j}$ is almost 0, and $\hat{\beta}_{\tau,i}$ and $-\hat{\beta}_{\tau,j}$ are almost equal when ρ_{ij} is close to -1. It is clear that only tuning parameter λ_2 of $\|\boldsymbol{\beta}_\tau\|_2^2$ can affect the upper bound of $D(i,j)$, which indicates the grouping effect of the L_2 penalty.

It is necessary to note that $\tilde{\boldsymbol{x}}_i^{\mathrm{T}}W(-\tilde{\boldsymbol{x}}_j) = -\rho_{ij}$ is close to 1 when ρ_{ij} is close to -1 and the corresponding coefficient of $-\tilde{\boldsymbol{x}}_j$ should be $-\hat{\beta}_{\tau,j}$ by

$$\begin{aligned} \boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau =& \boldsymbol{y} - \cdots - \tilde{\boldsymbol{x}}_i\hat{\beta}_{\tau,i} - \cdots - \tilde{\boldsymbol{x}}_j\hat{\beta}_{\tau,j} - \cdots = \\ & \boldsymbol{y} - \cdots - \tilde{\boldsymbol{x}}_i\hat{\beta}_{\tau,i} - \cdots - (-\tilde{\boldsymbol{x}}_j)(-\hat{\beta}_{\tau,j}) - \cdots. \end{aligned}$$

Then, according to (20), we have

$$\frac{|\hat{\beta}_{\tau,i} - (-\hat{\beta}_{\tau,j})|}{\sqrt{\boldsymbol{y}^{\mathrm{T}}W\boldsymbol{y}}} = \frac{|\hat{\beta}_{\tau,i} + \hat{\beta}_{\tau,j}|}{\sqrt{\boldsymbol{y}^{\mathrm{T}}W\boldsymbol{y}}} \leqslant \frac{\sqrt{2(1-(-\rho_{ij}))}}{\lambda_2} = \frac{\sqrt{2(1+\rho_{ij})}}{\lambda_2}.$$

This means that $\hat{\beta}_{\tau,i}$ and $-\hat{\beta}_{\tau,j}$ are almost equal, as $|\hat{\beta}_{\tau,i} + \hat{\beta}_{\tau,j}|$ is close to 0 when ρ_{ij} is close to -1.

3.3 Properties of the estimator

In this subsection, we propose the properties of the estimator. In the optimization problem (8), the sample size is consistently larger than the number of predictors, as $n + p > p$ holds. This means that (8) will always be a low-dimensional problem even if (7) is high-dimensional, and it helps reduce the regularization conditions.

Let $\boldsymbol{\beta}_{\tau 0}$ be the true value, $O = \{1 \leqslant j \leqslant p : \beta_{\tau 0,j} = 0\} \subset \{1, \cdots, p\}$. The following conditions are essential for theoretical properties.

(C1) $\log p = o(n), n \to \infty$.

(C2) Given τ, $\epsilon_i = y_i - \boldsymbol{x}_i\boldsymbol{\beta}_{\tau 0}, i = 1, \cdots, n$, are independent identically distributed random variables with mean 0 and variance $\sigma^2 < \infty$.

(C3) There exists a constant $K > 0$ such that $\lambda_{\max}\left(\frac{X^{\mathrm{T}}WX + \lambda_2 I}{(1+\lambda_2)(n+p)}\right) \leqslant K < \infty$, where $\lambda_{\max}(A)$ represents the largest eigenvalue of matrix A.

(C4) $\frac{\max_{1\leqslant j\leqslant p}\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j + \lambda_2}{\sqrt{(1+\lambda_2)(n+p)}} = O(\sqrt{\log p(n+p)})$ or $O(1)$ as $n, p \to \infty$.

(C5) $\|\boldsymbol{\beta}_{\tau 0}\|_0 = O(1)$.

These conditions are relatively mild. (C1) shows the relationship between sample size n and characteristic number p, and the proposed method applies to ultrahigh dimensional cases such as $p = \exp n^\alpha$ for $0 < \alpha < 1$. (C2) restricts the distribution of errors. (C3) is a standard linear regression condition. (C4) is a condition following Liu and Li[24]. (C5) indicates that the model is sparse. Under those conditions, we establish the consistency of the proposed estimator $\hat{\boldsymbol{\beta}}_\tau$ as follows:

Theorem 3.4. Assume that conditions (C1)–(C5) hold. Given proper λ_2. Let

$$\mu(X) = \max_{1\leqslant i<j\leqslant p} \frac{\tilde{\boldsymbol{x}}_i^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j}{(\tilde{\boldsymbol{x}}_i^{\mathrm{T}}W\tilde{\boldsymbol{x}}_i + \lambda_2)^{\frac{1}{2}}(\tilde{\boldsymbol{x}}_j^{T}W\tilde{\boldsymbol{x}}_j + \lambda_2)^{\frac{1}{2}}}. \tag{21}$$

For some $0 < \nu < 1$ and any $0 < q < 1/2$, let $n(\nu) = (1-\nu)(1 + 1/\mu(X))$ and

$$\lambda_1 = \frac{3\log(p/q)}{\nu(1+\mu(X))} \cdot \frac{\max_j \tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j + \lambda_2}{\min_j \tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j + \lambda_2}, \tag{22}$$

and $\hat{\boldsymbol{\beta}}_\tau$ is the minimizer of (7) under the restriction $\|\boldsymbol{\beta}_\tau\|_0 \leqslant n(\nu)$ with λ_1. Then,

(a)

$$\|\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_{\tau 0}\|_2 = O_p\left(\sqrt{\frac{\log p(n+p)}{n+p}}\right). \tag{23}$$

(b) With probability tending to 1, $\hat{\boldsymbol{\beta}}_{\tau,j} = 0, \forall j \in O$.

Theorem 3.3 indicates that under mild conditions (C1)–(C5), the proposed estimator has consistency and guarantees that the components are zeros if the true values of the corresponding components are zeros.

4 Simulation study

In this section, we first use one example to test the efficiency of the proposed algorithm for penalized linear extremile regression, while the next one shows the relationship between the MSE of the proposed QEN penalty and the sample size n, and the last two show the advantages of the QEN. It should also be noted that the estimated coefficients will be treated as zero if their absolute value is smaller than 10^{-6}. We simulate data from the true model

$$y = \boldsymbol{x}\boldsymbol{\beta} + \epsilon. \tag{24}$$

The components of $\boldsymbol{x}$ were standard normal in Examples 4.1, 4.2 and 4.3. Within each example, our simulated data consist of a training set, an independent tuning data set and a testing set. Denote the sample size of the training data set by n, while

DOI: 10.52396/JUSTC-2022-0099
JUSTC, 2023, 53(2): 1

an independent tuning data set and testing data set of size n and $100n$, respectively, were generated in the same way, and we repeated the process 100 times to compute the average value. Here are the details of the three scenarios.

Example 4.1. We generated data from the model (24) with $n = 20$. We set $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\epsilon \sim N(0, 1)$. The pairwise correlation between $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ was set to $r^{|i-j|}$, $r = 0, 0.5, 0.9$ and let $\tau = 0.5$.

Example 4.2. We generated data from the model (24) with $n = 10, 100, 1000$. We set $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\epsilon \sim N(0, 1)$. The pairwise correlation between $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ was set to $r^{|i-j|}$, $r = 0, 0.5, 0.9$ and let $\tau = 0.5$.

Example 4.3. We generated data from the model (24) with $n = 20$. We set

$$\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, \underbrace{0, \cdots, 0}_{45})$$

and $\epsilon \sim t(3)$. The pairwise correlation between $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ was set to $r^{|i-j|}$, $r = 0.5, 0.9$ and let $\tau = 0.9, 0.97$.

Example 4.4. We generated data from the model (24) with $n = 20$. We set

$$\boldsymbol{\beta} = (\underbrace{5, \cdots, 5}_{15}, \underbrace{0, \cdots, 0}_{985})$$

and $\epsilon \sim 0.5W(1, 1/2) + 0.5N(-2, 3)$. Predictors X were generated as follows:

$$\begin{aligned} &x_i = Z_1 + \epsilon_i^x, Z_1 \sim N(0, 1),\ i = 1, 2, 3, 4, 5; \\ &x_i = Z_2 + \epsilon_i^x, Z_2 \sim N(0, 1),\ i = 6, 7, 8, 9, 10; \\ &x_i = Z_3 + \epsilon_i^x, Z_3 \sim N(0, 1),\ i = 11, 12, 13, 14, 15; \\ &x_i \overset{\text{i.i.d}}{\sim} N(0, 1),\ i = 16, \cdots, 1000; \end{aligned} \tag{25}$$

where $\epsilon_i^x \overset{\text{i.i.d}}{\sim} N(0, 0.01)$, $i = 1, \cdots, 15$. Let $\tau = 0.9, 0.97$. In this model, we have three equally important groups with 5 predictors and 985 pure noise. An ideal method would select only the 15 true features and set the coefficients of the 985 noise to 0.

In those examples, we consider the following indicators: #SF, the average value of the number of selected predictors; P, the percentage of selecting the true model; MSE, the average mean squared error; CPU-time, the average CPU time in seconds; TP, the average value of the number of estimated nonzero components while those components are nonzero in true value; FP, the average value of the number of estimated nonzero components while those components are zeros in true value.

Table 1 shows the performance of extremile regression without penalty and with L_0, L_1, L_2, EN, QEN penalties under $\tau = 0.5$, which means OLS regression with different penalties, and indicates that the proposed algorithm is efficient. It is obvious that L_0 and QEN have advantages in CPU time, and QEN has a better performance in correction rate than EN. Moreover, L_0 might be superior to QEN because r is small, but QNE is more robust with the increase of r because of the L_2 penalty.

Table 2 shows the performance of extremile regression with the QEN penalty under different sample sizes and $\tau = 0.5$. Although the result obtained by the QEN penalty is better when r is small, #SF and frequencies of selecting the true model will increase with increasing n, and the corresponding MSE will be significantly reduced when r is fixed.

Tables 3 and 4 show the performance of extremile regression without penalty and with L_0, L_1, L_2, EN, and QEN penalties in general and high-dimensional situations with $\tau = 0.9, 0.97$. Although L_1 and EN are able to select more variables than L_0 and QEN, the value of QEN's TP/FP is much larger, which means that the QEN is able to select correct variables more effectively. Simultaneously, EN and QEN are more robust as τ increases, and they also perform better when r becomes larger. In the grouped variable situation, EN and QEN perform better, similarly affirming the grouping effect in theoretical properties.

Table 1. Results in extremile regression without penalty and with L_0, L_1, L_2, EN, and QEN penalties with different r settings, $\tau = 0.5$, 100 repetitions (standard deviations are shown in parentheses).

r	Method	#SF	P	MSE	CPU-time
0	Extremile	8.00(0.00)	0.00	0.020	–
	L_0	2.99(0.54)	0.99	0.015	0.0006(0.0031)
	L_1	2.93(0.55)	0.92	0.635	0.0043(0.0099)
	L_2	8.00(0.00)	0.00	0.019	–
	EN	2.93(0.58)	0.91	0.622	0.0040(0.0090)
	QEN	2.98(0.62)	0.99	0.019	0.0004(0.0026)
0.5	Extremile	8.00(0.00)	0.00	0.031	–
	L_0	2.89(0.69)	0.93	0.047	0.0009(0.0037)
	L_1	2.66(0.72)	0.83	0.796	0.0084(0.0161)
	L_2	8.00(0.00)	0.00	0.035	–
	EN	2.61(0.72)	0.86	0.758	0.0080(0.0136)
	QEN	2.9(0.77)	0.94	0.064	0.0007(0.0034)
0.9	Extremile	8.00(0.00)	0.00	0.160	–
	L_0	2.20(0.74)	0.17	0.401	0.0012(0.0035)
	L_1	2.06(0.76)	0.12	0.668	0.0357(0.0518)
	L_2	8.00(0.00)	0.00	0.237	–
	EN	2.46(0.77)	0.78	0.594	0.0268(0.0352)
	QEN	2.78(0.80)	0.92	0.104	0.0009(0.0037)

DOI: 10.52396/JUSTC-2022-0099
JUSTC, **2023**, 53(2): 1

Table 2. Results in extremile regression with QEN penalty with different r, n settings, $\tau = 0.5$, 100 repetitions (standard deviations are shown in parentheses).

r	n	#SF	P	MSE
0	10	2.98(0.47)	0.98	0.242
	100	3.00(0.00)	1.00	0.058
	1000	3.00(0.00)	1.00	0.007
0.5	10	2.92(0.68)	0.94	0.393
	100	2.93(0.63)	0.96	0.092
	1000	2.95(0.60)	0.97	0.011
0.9	10	2.83(0.77)	0.87	0.991
	100	2.86(0.74)	0.90	0.243
	1000	2.87(0.69)	0.92	0.036

Table 3. Results in extremile regression without penalty and with L_0, L_1, L_2, EN, and QEN penalties with different r, τ settings, 100 repetitions (standard deviations are shown in parentheses).

τ	Method	$r = 0.5$			$r = 0.9$		
		#SF	TP	FP	#SF	TP	FP
0.9	Extremile	50.00(0.00)	3.00(0.00)	47.00(0.00)	50.00(0.00)	3.00(0.00)	47.00(0.00)
	L_0	2.86(0.83)	2.03(0.89)	0.83(1.11)	2.82(1.07)	1.52(0.75)	1.30(1.34)
	L_1	2.89(0.94)	2.04(0.76)	0.85(1.03)	2.78(0.99)	1.51(0.85)	1.27(1.10)
	L_2	50.00(0.00)	3.00(0.00)	47.00(0.00)	50.00(0.00)	3.00(0.00)	47.00(0.00)
	EN	2.93(1.13)	2.37(0.78)	0.56(1.08)	2.95(1.14)	2.18(0.98)	0.77(1.00)
	QEN	2.87(0.91)	2.56(0.96)	0.31(1.03)	2.96(1.08)	2.37(0.81)	0.59(1.18)
0.97	Extremile	50(0.00)	3.00(0.00)	47.00(0.00)	50.00(0.00)	3.00(0.00)	47.00(0.00)
	L_0	2.72(0.65)	1.71(0.81)	1.01(0.98)	2.56(0.89)	1.41(0.72)	1.15(1.13)
	L_1	2.98(0.71)	1.92(0.79)	1.06(0.94)	2.74(0.78)	1.50(0.80)	1.24(1.00)
	L_2	50.00(0.00)	3.00(0.00)	47.00(0.00)	50.00(0.00)	3.00(0.00)	47.00(0.00)
	EN	2.91(0.77)	2.24(0.77)	0.67(0.89)	2.87(0.84)	2.20(0.92)	0.67(0.94)
	QEN	2.89(0.84)	2.30(0.89)	0.59(0.92)	2.86(0.89)	2.26(0.85)	0.60(0.98)

Table 4. Results in extremile regression without penalty and with L_0, L_1, L_2, EN, and QEN penalties with different τ settings, 100 repetitions (standard deviations are shown in parentheses).

τ	Method	#SF	TP	FP
0.9	Extremile	1000.00(0.00)	15.00(0.00)	985.00(0.00)
	L_0	12.86(2.02)	11.47(1.99)	1.39(2.58)
	L_1	13.74(1.56)	12.52(1.47)	1.22(0.46)
	L_2	1000.00(0.00)	15.00(0.00)	985.00(0.00)
	EN	14.67(2.31)	14.15(2.23)	0.52(0.73)
	QEN	14.65(1.45)	14.39(1.51)	0.26(0.54)
0.97	Extremile	1000.00(0.00)	15.00(0.00)	985.00(0.00)
	L_0	10.19(1.69)	8.42(2.00)	1.77(2.27)
	L_1	14.49(2.53)	12.12(2.49)	2.37(1.62)
	L_2	1000.00(0.00)	15.00(0.00)	985.00(0.00)
	EN	14.52(2.71)	12.65(2.62)	1.87(1.21)
	QEN	14.88(1.70)	13.44(2.19)	1.44(1.36)

These simulations all indicate that our QEN penalty is competent in variable selection because of its L_0 penalty, so this is an advantage of the proposed method when we face high-dimensional problems. On the other hand, although L_0 performs no worse than QEN when the correlations of the predictors are weak, the proposed QEN performs better as the correlations increase and its estimator always has lower MSE. Moreover, the proposed QEN also tends to choose more predictors than the L_0 penalty, especially in tail analysis, which means that we can obtain more information.

5 Real data analysis

Our real data about communities and crime are downloaded from UCI①. The variables in the dataset involving various

① http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.

DOI: 10.52396/JUSTC-2022-0099
JUSTC, **2023**, 53(2): 1

Table 5. Results of the Communities and Crime Data in extremile regression with L_0, L_1, EN, and QEN penalties with different τ settings, 100 repetitions (standard deviations are shown in parentheses).

τ	Method	Test error	#SF
0.9	L_0	0.7362(0.0085)	23.49(14.61)
	L_1	0.7364(0.0086)	27.56(11.51)
	EN	0.7346(0.0083)	26.56(12.31)
	QEN	0.7351(0.0090)	25.48(9.02)
0.95	L_0	0.9010(0.0073)	20.63(16.28)
	L_0	0.9021(0.0068)	24.35(12.57)
	EN	0.9004(0.0065)	23.87(10.36)
	QEN	0.9006(0.0068)	22.39(8.24)
0.97	L_0	0.9892(0.0048)	19.48(17.71)
	L_0	0.9897(0.0045)	22.66(14.23)
	EN	0.9884(0.0040)	22.23(9.84)
	QEN	0.9886(0.0042)	21.60(7.69)

community data were picked if there was any plausible connection to crime ($p = 122$), with the goal attribute PerCapitaViolentCrimes (total number of violent crimes per 100k population), and all numeric data were normalized into the decimal range 0.00–1.00 using an unsupervised equal-interval binning method. Redmond and Baveja[31] adopted this dataset to present an artificial-intelligence software crime similarity system (CSS) to help police departments develop a strategic viewpoint toward decision-making and utilize the socioeconomic, crime and enforcement profiles of cities to generate a list of communities that are the best candidates to cooperate and share experiences. We aim to apply the QEN penalized linear extremile regression to select variables that are strongly correlated to PerCapitaViolentCrimes under different extreme τ.

In each repetition, we randomly choose 20 samples from 319 complete samples as a tuning set and 20 samples as a training set, while the remaining 279 groups are used as the testing set to compute the test error. The performance of 100 repetitions of penalized extremile regression with different penalties and different τ is summarized in Table 5, where test error represents the error of estimation under the testing set and #SF means the average value of the number of selected predictors.

Table 5 shows the performance of extremile regression without penalty and with L_0, L_1, L_2, EN, and QEN penalties 100 repetitions with $\tau = 0.9, 0.95, 0.97$ under the Communities and Crime Data. This indicates that the test errors of different penalties have little difference, but QEN selects fewer variables with smaller standard deviations than L_0 and L_1 and has more robustness.

6 Conclusions

In this paper, we propose a linear extremile regression model with the QEN penalty to select variables in high-dimensional problems. The proposed QEN penalty retains the advantages of L_0 and L_2 penalties, which implies that it is able to obtain sparse solutions and deal with highly correlated predictors. Simultaneously, we propose an EM algorithm to solve optimization problems with the L_0 penalty approximately and establish corresponding theoretical properties under mild conditions. To summarize, the proposed method has important implications for extending the traditional extremile regression theory.

Future work can be extended to the following aspects. First, the extremile regression with penalty can be extended to nonparametric and semiparametric models. Second, the proposed EM algorithm can be applied to the L_p penalty, $p \in (0, 2)$, so we can consider optimization problems with different penalties. Finally, we seek methods to determine the true value of unknown coefficients under different τ to improve the evaluation standards of variable selection methods in extremile regression.

Acknowledgements

We thank the reviewers for their perceptive comments and suggestions. This work was supported by the National Natural Science Foundation of China (12171450).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Yimin Xiong is currently a master's student under the supervision of Professor Weiping Zhang at the University of Science and Technology of China. His research is focused on variable selection.

Weiping Zhang received his Ph.D. degree from the University of Science and Technology of China (USTC). He is currently a Professor at the USTC. His research interests mainly focus on longitudinal data analysis and Bayesian analysis.

References

[1] Koenker R, Bassett G. Regression quantiles. *Econometrica,* **1978**, *46*: 33–50.
[2] Newey W K, Powell J L. Asymmetric least squares estimation and testing. *Econometrica,* **1987**, *55*: 819–847.
[3] Daouia A, Gijbels I, Stupfler G. Extremiles: A new perspective on asymmetric least squares. *Journal of the American Statistical Association,* **2019**, *114* (527): 1366–1381.
[4] Allen D M. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics,* **1974**, *16*

DOI: 10.52396/JUSTC-2022-0099
JUSTC, **2023**, 53(2): 1

(1): 125–127.
[5] Mallows C L. Some comments on C_p. *Technometrics,* **2000**, *42* (1): 87–94.
[6] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* **1974**, *19* (6): 716–723.
[7] Schwarz G. Estimating the dimension of a model. *The Annals of Statistics,* **1978**, *6* (2): 461–464.
[8] Geisser S, Eddy W F. A predictive approach to model selection. *Journal of the American Statistical Association,* **1979**, *74* (365): 153–160.
[9] Devroye L, Wagner T. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory,* **1979**, *25* (5): 601–604.
[10] Dietterich T G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation,* **1998**, *10* (7): 1895–1923.
[11] Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics,* **2007**, *35* (6): 2313–2351.
[12] Dicker L, Lin X. Parallelism, uniqueness, and large-sample asymptotics for the Dantzig selector. *Canadian Journal of Statistics,* **2013**, *41* (1): 23–35.
[13] James G M, Radchenko P, Lv J. DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **2009**, *71* (1): 127–142.
[14] Antoniadis A, Fryzlewicz P, Letué F. The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics,* **2010**, *37* (4): 531–552.
[15] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **2008**, *70* (5): 849–911.
[16] Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association,* **2011**, *106* (494): 544–557.
[17] Liu Z, Lin S, Tan M. Sparse support vector machines with L_p penalty for biomarker identification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* **2008**, *7* (1): 100–107.
[18] Mazumder R, Friedman J H, Hastie T. SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association,* **2011**, *106* (495): 1125–1138.
[19] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* **1996**, *58* (1): 267–288.
[20] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association,* **2001**, *96* (456): 1348–1360.
[21] Zhang C H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics,* **2010**, *38* (2): 894–942.
[22] Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **2005**, *67* (2): 301–320.
[23] Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association,* **2006**, *101* (476): 1418–1429.
[24] Liu Z, Li G. Efficient regularized regression with penalty for variable selection and network construction. *Computational and Mathematical Methods in Medicine,* **2016**, *2016*: 3456153.
[25] Tihonov A N. Solution of incorrectly formulated problems and the regularization method. *Soviet Math.,* **1963**, *4*: 1035–1038.
[26] Wang J, Xue L, Zhu L, et al. Estimation for a partial-linear single-index model. *The Annals of Statistics,* **2010**, *38* (1): 246–274.
[27] West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences,* **2001**, *98* (20): 11462–11467.
[28] Hastie T, Tibshirani R, Eisen M B, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology,* **2000**, *1*: research0003.1.
[29] Hastie T, Tibshirani R, Botstein D, et al. Supervised harvesting of expression trees. *Genome Biology,* **2001**, *2*: research0003.1.
[30] Segal M S, Dahlquist K D, Conklin B R. Regression approaches for microarray data analysis. *Journal of Computational Biology,* **2003**, *10* (6): 961–980.
[31] Redmond M, Baveja A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research,* **2002**, *141* (3): 660–678.

Appendix

Proof of Theorem 3.1.

$$\boldsymbol{\beta}_\tau^* = (1+\lambda_2)^{\frac{1}{2}}(D(X^{\mathrm{T}}WX+\lambda_2 I)+\lambda_1(1+\lambda_2)I)^{-1}(DX^{\mathrm{T}}W\boldsymbol{y}), \tag{A.1}$$

where $D=\mathrm{diag}(\beta_{\tau,1}^{*2},\cdots,\beta_{\tau,p}^{*2})$.

Let $G(\boldsymbol{\beta}_\tau^*)=(1+\lambda_2)^{\frac{1}{2}}(D(X^{\mathrm{T}}WX+\lambda_2 I)+\lambda_1(1+\lambda_2)I)^{-1}(DX^{\mathrm{T}}W\boldsymbol{y})$, then $G(\boldsymbol{\beta}_\tau^*)$ is Lipschitz continuous for $\boldsymbol{\beta}_\tau^*\in\mathbb{R}^p$, and

$$\begin{aligned}\nabla G(\boldsymbol{\beta}_\tau^*) =&(1+\lambda_2)^{\frac{1}{2}}(D(X^{\mathrm{T}}WX+\lambda_2 I)+\lambda_1(1+\lambda_2)I)^{-2}((D(X^{\mathrm{T}}WX+\lambda_2 I)+\lambda_1(1+\lambda_2)I)(2D^{\frac{1}{2}}X^{\mathrm{T}}W\boldsymbol{y})-2D^{\frac{1}{2}}(X^{\mathrm{T}}WX+\lambda_2 I)(DX^{\mathrm{T}}W\boldsymbol{y}))=\\ &2\lambda_1(1+\lambda_2)^{\frac{3}{2}}(D(X^{\mathrm{T}}WX+\lambda_2 I)+\lambda_1(1+\lambda_2)I)^{-2}(D^{\frac{1}{2}}X^{\mathrm{T}}W\boldsymbol{y}).\end{aligned} \tag{A.2}$$

Because $2\lambda_1(1+\lambda_2)^{3/2}\|(D(X^{\mathrm{T}}WX+\lambda_2 I)+\lambda_1(1+\lambda_2)I)^{-2}\|_\infty\|D^{1/2}X^{\mathrm{T}}W\boldsymbol{y}\|_\infty<1$, there exists a Lipschitz constant that satisfies

$$\begin{aligned}&0<\gamma=\|\nabla G(\boldsymbol{\beta}_\tau^*)\|_\infty=2\lambda_1(1+\lambda_2)^{\frac{3}{2}}\|(D(X^{\mathrm{T}}WX+\lambda_2 I)+\lambda_1(1+\lambda_2)I)^{-2}(D^{\frac{1}{2}}X^{\mathrm{T}}W\boldsymbol{y})\|_\infty\leqslant\\ &2\lambda_1(1+\lambda_2)^{\frac{3}{2}}\|(D(X^{\mathrm{T}}WX+\lambda_2 I)+\lambda_1(1+\lambda_2)I)^{-2}\|_\infty\|(D^{\frac{1}{2}}X^{\mathrm{T}}W\boldsymbol{y})\|_\infty<1.\end{aligned} \tag{A.3}$$

Therefore, given the initial value $\boldsymbol{\beta}_\tau^{*(0)}$, the sequence $\{\boldsymbol{\beta}_\tau^{*(r)}\}$ is bounded. That is, $\forall i\in\mathbb{N}^*$, $\exists\alpha\in(0,1)$ and $\boldsymbol{\xi}=\alpha\boldsymbol{\beta}_\tau^{*(i)}+(1-\alpha)\boldsymbol{\beta}_\tau^{*(i-1)}$ satisfies

$$\begin{aligned}&\|\boldsymbol{\beta}_\tau^{*(i+1)}-\boldsymbol{\beta}_\tau^{*(i)}\|_\infty=\|G(\boldsymbol{\beta}_\tau^{*(i)})-G(\boldsymbol{\beta}_\tau^{*(i-1)})\|_\infty=\|\nabla G(\boldsymbol{\xi})(\boldsymbol{\beta}_\tau^{*(i)}-\boldsymbol{\beta}_\tau^{*(i-1)})\|_\infty\leqslant\\ &\|\nabla G(\boldsymbol{\xi})\|_\infty\|(\boldsymbol{\beta}_\tau^{*(i)}-\boldsymbol{\beta}_\tau^{*(i-1)})\|_\infty<\gamma\|(\boldsymbol{\beta}_\tau^{*(i)}-\boldsymbol{\beta}_\tau^{*(i-1)})\|_\infty<\cdots<\gamma^i\|(\boldsymbol{\beta}_\tau^{*(1)}-\boldsymbol{\beta}_\tau^{*(0)})\|_\infty.\end{aligned} \tag{A.4}$$

Therefore,$\forall k\in\mathbb{N}^*$, we have

DOI: 10.52396/JUSTC-2022-0099
JUSTC, **2023**, 53(2): 1

$$\|\boldsymbol{\beta}_\tau^{*(i+k)}-\boldsymbol{\beta}_\tau^{*(i)}\|_\infty=\|(\boldsymbol{\beta}_\tau^{*(i+k)}-\boldsymbol{\beta}_\tau^{*(i+k-1)})+\cdots+(\boldsymbol{\beta}_\tau^{*(i+1)}-\boldsymbol{\beta}_\tau^{*(i)})\|_\infty\leqslant\|(\boldsymbol{\beta}_\tau^{*(i+k)}-\boldsymbol{\beta}_\tau^{*(i+k-1)})\|_\infty+\cdots+\|(\boldsymbol{\beta}_\tau^{*(i+1)}-\boldsymbol{\beta}_\tau^{*(i)})\|_\infty<$$
$$(\gamma^{i+k-1}+\cdots+\gamma^i)\|\|(\boldsymbol{\beta}_\tau^{*(1)}-\boldsymbol{\beta}_\tau^{*(0)})\|_\infty=\frac{\gamma^i(1-\gamma^k)}{1-\gamma}\|(\boldsymbol{\beta}_\tau^{*(1)}-\boldsymbol{\beta}_\tau^{*(0)})\|_\infty. \tag{A.5}$$

Since $\|(\boldsymbol{\beta}_\tau^{*(1)}-\boldsymbol{\beta}_\tau^{*(0)})\|_\infty$ is bounded, then

$$\lim_{i,k\to\infty}\|\boldsymbol{\beta}_\tau^{*(i+k)}-\boldsymbol{\beta}_\tau^{*(i)}\|_\infty=0. \tag{A.6}$$

Therefore $\{\boldsymbol{\beta}_\tau^{*(r)}\}$ is a Cauchy sequence with a limit $\boldsymbol{\beta}_\tau^{*0}$.

Note that $G(\boldsymbol{\beta}_\tau^*)$ is nonconvex, and there might exist multiple local optimal solutions. However, the proposed EM algorithm always reaches the same optimal solution closest to the initial value $\boldsymbol{\beta}_\tau^{*(0)}$. Assume that there were two different solutions $\boldsymbol{\beta}_\tau^{*0}$ and $\boldsymbol{\beta}_\tau^{*1}$ with the initial value $\boldsymbol{\beta}_\tau^{*(0)}$, then

$$\|\boldsymbol{\beta}_\tau^{*0}-\boldsymbol{\beta}_\tau^{*1}\|_\infty=\|G(\boldsymbol{\beta}_\tau^{*0})-G(\boldsymbol{\beta}_\tau^{*1})\|_\infty\leqslant\gamma\|\boldsymbol{\beta}_\tau^{*0}-\boldsymbol{\beta}_\tau^{*1}\|_\infty. \tag{A.7}$$

As $\gamma<1$, the above inequality holds if and only if $\|\boldsymbol{\beta}_\tau^{*0}-\boldsymbol{\beta}_\tau^{*1}\|_\infty=0$, that is $\boldsymbol{\beta}_\tau^{*0}=\boldsymbol{\beta}_\tau^{*1}$.

Moreover,

$$\|\boldsymbol{\beta}_\tau^{*(r+1)}-\boldsymbol{\beta}_\tau^{*0}\|_\infty=\|G(\boldsymbol{\beta}_\tau^{*(r)})-G(\boldsymbol{\beta}_\tau^{*0})\|_\infty\leqslant\gamma\|\boldsymbol{\beta}_\tau^{*(r)}-\boldsymbol{\beta}_\tau^{*0}\|_\infty. \tag{A.8}$$

Therefore, $\{\boldsymbol{\beta}_\tau^{*(r)}\}$ converges.

Proof of Lemma 3.1. According to the given conditions, when $\delta_j\neq0$, then

$$\begin{gathered}\delta_j^2\frac{\partial\tilde{L}(\boldsymbol{\beta}_\tau^*)}{\partial\beta_{\tau,j}^*}=-2\delta_j^2\boldsymbol{x}_j^{*T}(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)+2\lambda_1\beta_{\tau,j}^*=0,\\ \delta_j=\beta_{\tau,j}^*,\forall j\in\{1,\cdots,p\}.\end{gathered} \tag{A.9}$$

The above equations are equivalent to

$$\delta_j^2\boldsymbol{x}_j^{*\mathrm{T}}(\boldsymbol{y}^*-X^*\boldsymbol{\delta}^*)-\lambda_1\delta_j=0. \tag{A.10}$$

As $\tilde{\boldsymbol{x}}_i^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j=0,\forall i\neq j\in\{1,\cdots,p\}$, Eq. (A.10) becomes the following quadratic equation of δ_j:

$$(\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j+\lambda_2)\delta_j^2-(1+\lambda_2)^{\frac{1}{2}}(\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\boldsymbol{y})\delta_j+\lambda_1(1+\lambda_2)=0. \tag{A.11}$$

From the discriminant of the quadratic equation, we have

$$(1+\lambda_2)(\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\boldsymbol{y})^2-4\lambda_1(1+\lambda_2)(\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j+\lambda_2)\geqslant0, \tag{A.12}$$

that is,

$$\lambda_1\leqslant\frac{(\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\boldsymbol{y})^2}{4(\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j+\lambda_2)},\forall j\in\{1,\cdots,p\}. \tag{A.13}$$

Therefore, the maximal λ_1 satisfies

$$\lambda_{1_{\max}}=\min_j\left\{\frac{(\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\boldsymbol{y})^2}{4(\tilde{\boldsymbol{x}}_j^{\mathrm{T}}W\tilde{\boldsymbol{x}}_j+\lambda_2)}\right\}. \tag{A.14}$$

Proof of Theorem 3.2. First, we show that the proposed algorithm is an L_0 approximation. Let $\boldsymbol{\beta}_{\tau0}^*=(\boldsymbol{\beta}_{\tau01}^{*\mathrm{T}},\boldsymbol{\beta}_{\tau02}^{*\mathrm{T}})^{\mathrm{T}}$ be the true value, where $\boldsymbol{\beta}_{\tau01}^*\in\mathbb{R}^q,q\leqslant p$ are the nonzero components, $\boldsymbol{\beta}_{\tau02}^*=\mathbf{0}\in\mathbb{R}^{p-q}$. Assume $S=\mathrm{diag}(1,\cdots,1,0,\cdots,0)\in\mathbb{R}^{p\times p}$, where p is the number of 1s. If p is known, we have $(X^*S)^{\mathrm{T}}X^*S=S(X^{\mathrm{T}}WX+\lambda_2I)S=S(X^{\mathrm{T}}WX+\lambda_2I)$ and can estimate $\boldsymbol{\beta}_\tau^*$ with λ' of a ridge regression given X^*S and $\boldsymbol{y}^*$:

$$\begin{gathered}\hat{\boldsymbol{\beta}}_\tau^*=((X^*S)^{\mathrm{T}}X^*S+\lambda'I)^{-1}(X^*S)^{\mathrm{T}}\boldsymbol{y}^*=((1+\lambda_2)^{-1}S(X^{\mathrm{T}}WX+\lambda_2I)S+\lambda'I)^{-1}((1+\lambda_2)^{-\frac{1}{2}}SX^{\mathrm{T}}W\boldsymbol{y})=\\ (1+\lambda_2)^{\frac{1}{2}}(S(X^{\mathrm{T}}WX+\lambda_2I)+\lambda'(1+\lambda_2)I)^{-1}SX^{\mathrm{T}}W\boldsymbol{y}.\end{gathered} \tag{A.15}$$

It is guaranteed that the corresponding components of $\hat{\boldsymbol{\beta}}_\tau^*$ are zeros, but we generally do not know S in reality, which means estimating S is an NP-hard problem and affects the difficulty of estimating $\hat{\boldsymbol{\beta}}_\tau^*$. Compared with the first equation of (14), it can be found that the only difference between these two equations is S and D, where the diagonal elements in S are replaced by δ_j^2 in D. Moreover, if $\boldsymbol{\beta}_{\tau,j}^*=0$, then $\delta_j^2=\delta_j=0$, so the proposed algorithm is an L_0 approximation.

Next, we consider an adaptive ridge regression

$$\boldsymbol{\beta}_\tau^*=\mathrm{argmin}_{\beta_\tau^*}(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)^{\mathrm{T}}(\boldsymbol{y}^*-X^*\boldsymbol{\beta}_\tau^*)+\sum_{j=1}^p\lambda_j'\beta_{\tau,j}^{*2}, \tag{A.16}$$

where

DOI: 10.52396/JUSTC-2022-0099
JUSTC, **2023**, 53(2): 1

$$\lambda_j' = \begin{cases} \dfrac{\lambda_1}{\delta_j^2}, & \delta_j \neq 0; \\ \infty, & \delta_j = 0. \end{cases} \tag{A.17}$$

If the component of true value $\boldsymbol{\beta}_{\tau 0}^*$ satisfies $\beta_{\tau 0,j}^* = 0$, as the proposed algorithm is an L_0 approximation, we have $\delta_j = 0$ and $\lambda_j' = \infty$ and the regression is equivalent to

$$\boldsymbol{\beta}_\tau^* = \text{argmin}_{\boldsymbol{\beta}_\tau^*}(\boldsymbol{y}^* - X^*\boldsymbol{\beta}_\tau^*)^{\mathrm{T}}(\boldsymbol{y}^* - X^*\boldsymbol{\beta}_\tau^*) + \lambda_1 \sum_{j \in R} \frac{\boldsymbol{\beta}_{\tau,j}^{*2}}{\delta_j^2}, \tag{A.18}$$

that is (9). Therefore, the final solution of the proposed EM algorithm is an optimal solution for the L_0 approximation problem that minimizes (8).

Proof of Lemma 3.2. (a) Note that it is a weighted least square model and a similar conclusion in the ordinary least squares model has been proven in Ref. [22, Lemma 2].

(b) If $\hat{\beta}_i\hat{\beta}_j < 0$, consider $\hat{\boldsymbol{\beta}}^*$ as

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k, & k \neq i \text{ and } j; \\ 0, & k = i; \\ \hat{\beta}_i + \hat{\beta}_j, & k = j. \end{cases} \tag{A.19}$$

Then, we have $\|\hat{\boldsymbol{\beta}}^*\|_0 < \|\hat{\boldsymbol{\beta}}\|_0$. Therefore, $\hat{\boldsymbol{\beta}}$ cannot be a minimizer, which is a contradiction. If there is one, and only one 0 in $\hat{\beta}_i$ and $\hat{\beta}_j$, we can consider a similar $\hat{\beta}_k^*$.

(c) If $f(\cdot)$ is the QEN penalty, the proof of $\hat{\beta}_i\hat{\beta}_j \geq 0$ is the same as (b). When $\hat{\beta}_i\hat{\beta}_j > 0$, let $\hat{\beta}_i \neq \hat{\beta}_j$ and we can consider $\hat{\boldsymbol{\beta}}^*$ as

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k; \\ \dfrac{\hat{\beta}_i + \hat{\beta}_j}{2}, & k \neq i \text{ and } j; \\ & k = i \text{ or } j. \end{cases} \tag{A.20}$$

Because the L_2 penalty is strictly convex, we have $\|\hat{\boldsymbol{\beta}}^*\|_0 = \|\hat{\boldsymbol{\beta}}\|_0$ and $\|\hat{\boldsymbol{\beta}}^*\|_2^2 < \|\hat{\boldsymbol{\beta}}\|_2^2$, which is a contradiction. Then, we have $\hat{\beta}_i = \hat{\beta}_j$.

Proof of Theorem 3.3. As $\hat{\beta}_{\tau,i}\hat{\beta}_{\tau,j} > 0$, $\hat{\beta}_{\tau,i}$ and $\hat{\beta}_{\tau,j}$ are nonzero, and if $\hat{\beta}_{\tau,k} \neq 0$, then

$$\frac{\partial L(\hat{\boldsymbol{\beta}}_\tau)}{\partial \beta_{\tau,k}} = -2\tilde{\boldsymbol{x}}_k^{\mathrm{T}} W(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau) + 2\lambda_2\beta_{\tau,k} = 0. \tag{A.21}$$

Substituting $\hat{\beta}_{\tau,i}$ and $\hat{\beta}_{\tau,j}$ into the above equation, we can obtain the following equations:

$$\begin{aligned} -2\tilde{\boldsymbol{x}}_i^{\mathrm{T}} W(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau) + 2\lambda_2\beta_{\tau,i} = 0, \\ -2\tilde{\boldsymbol{x}}_j^{\mathrm{T}} W(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau) + 2\lambda_2\beta_{\tau,j} = 0. \end{aligned} \tag{A.22}$$

Subtracting the two equations, we can obtain the following equation:

$$(\tilde{\boldsymbol{x}}_i^{\mathrm{T}} - \tilde{\boldsymbol{x}}_j^{\mathrm{T}}) W(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau) - \lambda_2(\beta_{\tau,i} - \beta_{\tau,j}) = 0. \tag{A.23}$$

As $\hat{\boldsymbol{\beta}}_\tau$ is an optimal solution, we have $L(\hat{\boldsymbol{\beta}}_\tau) \leqslant L(\boldsymbol{0})$, that is,

$$(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau)^{\mathrm{T}} W(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau) + \lambda_1\|\hat{\boldsymbol{\beta}}_\tau\|_0 + \lambda_2\|\hat{\boldsymbol{\beta}}_\tau\|_2^2 \leqslant \boldsymbol{y}^{\mathrm{T}} W \boldsymbol{y} = 1. \tag{A.24}$$

Since $\tilde{\boldsymbol{x}}_i^{\mathrm{T}} W \tilde{\boldsymbol{x}}_i = 1$, $\tilde{\boldsymbol{x}}_i^{\mathrm{T}} W \tilde{\boldsymbol{x}}_j = \rho_{ij}$, we have $(\tilde{\boldsymbol{x}}_i^{\mathrm{T}} - \tilde{\boldsymbol{x}}_j^{\mathrm{T}}) W(\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j) = 2(1 - \rho_{ij})$, so

$$|\beta_{\tau,i} - \beta_{\tau,j}| = \left| \frac{(\tilde{\boldsymbol{x}}_i^{\mathrm{T}} - \tilde{\boldsymbol{x}}_j^{\mathrm{T}}) W(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau)}{\lambda_2} \right| \leqslant \frac{\|(\tilde{\boldsymbol{x}}_i^{\mathrm{T}} - \tilde{\boldsymbol{x}}_j^{\mathrm{T}}) W^{\frac{1}{2}}\|_2 \|W^{\frac{1}{2}}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\tau)\|_2}{\lambda_2} \leqslant \frac{\sqrt{2(1-\rho_{ij})}}{\lambda_2}. \tag{A.25}$$

Proof of Theorem 3.4. (a) Let $\boldsymbol{\beta}_{\tau 0}^* = (1 + \lambda_2)^{1/2}\boldsymbol{\beta}_{\tau 0}$, then $\boldsymbol{\beta}_{\tau 0}^*$ is the true value of Eq. (8). From Liu and Li[24], we have

$$\|\hat{\boldsymbol{\beta}}_\tau^* - \boldsymbol{\beta}_{\tau 0}^*\|_2 = O_p\left(\sqrt{\frac{\log p(n+p)}{n+p}} \right). \tag{A.26}$$

Note that $\hat{\boldsymbol{\beta}}_\tau^* = (1 + \lambda_2)^{1/2}\hat{\boldsymbol{\beta}}_\tau$, and λ_2 is a constant, so

$$\|\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_{\tau 0}\|_2 = O_p\left(\sqrt{\frac{\log p(n+p)}{n+p}} \right). \tag{A.27}$$

(b) As $\boldsymbol{\beta}_{\tau 0}^* = (1 + \lambda_2)^{1/2}\boldsymbol{\beta}_{\tau 0}$, we have $\beta_{\tau 0,j}^* = \beta_{\tau 0,j} = 0$ while $j \in O$, that is, $O = \{1 \leqslant j \leqslant p : \beta_{\tau 0,j}^* = 0\} = \{1 \leqslant j \leqslant p : \beta_{\tau 0,j} = 0\}$.

From Liu and Li[24], we have $\hat{\boldsymbol{\beta}}_{\tau,j}^* = 0, \forall j \in O$, with probability tending to 1.

Note that $\hat{\boldsymbol{\beta}}_\tau^* = (1 + \lambda_2)^{1/2}\hat{\boldsymbol{\beta}}_\tau$, so with probability tending to 1, $\hat{\boldsymbol{\beta}}_{\tau,j} = 0, \forall j \in O$.

DOI: 10.52396/JUSTC-2022-0099
JUSTC, **2023**, 53(2): 1