Ann Inst Stat Math (2019) 71:409–438
https://doi.org/10.1007/s10463-018-0645-1

Penalized expectile regression: an alternative to penalized quantile regression

Lina Liao[1] · **Cheolwoo Park**[1] · **Hosik Choi**[2]

Received: 18 March 2017 / Revised: 1 January 2018 / Published online: 19 February 2018

Abstract This paper concerns the study of the entire conditional distribution of a response given predictors in a heterogeneous regression setting. A common approach to address heterogeneous data is quantile regression, which utilizes the minimization of the L_1 norm. As an alternative to quantile regression, we consider expectile regression, which relies on the minimization of the asymmetric L_2 norm and detects heteroscedasticity effectively. We assume that only a small set of predictors is relevant to the response and develop penalized expectile regression with SCAD and adaptive LASSO penalties. With properly chosen tuning parameters, we show that the proposed estimators display oracle properties. A numerical study using simulated and real examples demonstrates the competitive performance of the proposed penalized expectile regression, and its combined use with penalized quantile regression would be helpful and recommended for practitioners.

Keywords Asymptotics · Expectile regression · Heteroscedasticity · Penalized regression · Variable selection

✉ Hosik Choi
choi.hosik@gmail.com

Lina Liao
linaliaostat@gmail.com

Cheolwoo Park
cpark@uga.edu

[1] Department of Statistics, University of Georgia, Athens, GA 30602, USA

[2] Department of Applied Statistics, Kyonggi University, Suwon, Kyonggi-do 16227, Korea

1 Introduction

Quantiles and expectiles, which contain information about the full distribution for a random variable, are extensions of median and mean, respectively. Quantiles are percentiles of the cumulative distribution function of a random variable. For instance, if x is the αth quantile of X, $P(X \leq x) = \alpha$. Unlike quantiles, which is the minimal value of the tail, expectiles incorporate information about the expectation of X, conditional on X being in a tail of its distribution (Newey and Powell 1987).

In financial time series, expectiles emerge as an alternative to popular risk measures such as value at risk (VaR) and expected shortfall (ES), as they have desirable properties (Ziegel 2014). A risk measure is an estimated amount of capital to be reserved at a given risk level to prevent substantial losses. VaR at α, the 100αth quantile of the return distribution, can be interpreted as the minimum potential loss at the $100\alpha\%$ level. While VaR is the most widely used risk measure, it does not provide information regarding the potential magnitude of losses because VaR only depends on the tail probability. ES is an alternative risk measure that considers the magnitude of the potential losses in the lower tail. However, it is also known that ES can be too conservative, which could be a major disadvantage to commercial and investment banks. In contrast to VaR and ES, which only concern the lower tail, the expectile relies on both tails of the distribution to measure risk. The squared loss function makes the expectile more sensitive to extreme values and to the shape of the distribution than VaR, which can be beneficial when measuring potential losses because one wants a risk measure to be sensitive to extreme tail losses.

When dealing with heterogeneous data in regression, we frequently see that targeting only a mean function is not sufficient to capture a complete picture of the relationship between the response variable and predictors. In such a case, quantile regression (Koenker and Bassett 1978) based on an asymmetric L_1 norm could be a more appropriate tool as it allows one to study the quantile structure of the conditional distribution. Quantile regression has been applied in various fields such as economics, survival analysis, and microarray studies. While quantile regression has a strong intuitive appeal, Newey and Powell (1987) point out three drawbacks: nondifferentiability, inefficiency for Gaussian-like error distributions, and difficulty of calculation of a covariance matrix. They propose expectile regression based on an asymmetric L_2 norm as an alternative way to analyze the complete conditional distribution of the response. Expectile regression generalizes ordinary mean regression, which is known to be efficient when typical assumptions, including homogeneity of errors, are met. It is also closely related to quantile regression, which is robust to outliers. Expectile regression has gained attention in several fields. Aigner et al. (1976) construct expectiles to estimate production frontiers. Sobotka et al. (2013a) investigate the relationship between women's education and fertility in Botswana via semiparametric expectile regression. Sobotka et al. (2013b) study statistical inference of semiparametric expectile regression. Schnabel and Eilers (2009) also demonstrate efficiency of expectiles over quantiles. Although expectile regression has found applications in various fields, to our knowledge, there has been little work with a penalized version of expectile regression. In this work, we fill this gap by investigating expectile regression with SCAD (Fan and Li 2001) and adaptive LASSO (Zou 2006) penalties.

A regularization approach has been extended to quantile regression because sets of important predictors might differ from quantile to quantile. Penalized quantile regression is capable of investigating the complete conditional distribution of the response variable and the sparsity pattern. Li and Zhu (2008) propose L_1 norm quantile regression, which selects variables automatically and controls the variance of the fitted coefficients simultaneously. Wu and Liu (2009) demonstrate the oracle properties of quantile regression with SCAD and adaptive LASSO penalties when the number of variables, p, is fixed. Wang et al. (2012) study the behavior of quantile regression in ultra-high-dimensional data when p is much larger than n. Belloni and Chernozhukov (2011) study theory of quantile regression with lasso penalty in an ultra-high-dimensional setting, and Belloni et al. (2015) develop post-selection inference methods for quantile regression with lasso penalty in an ultra-high- dimensional setting.

The main objective of this paper is to develop penalized expectile regression when p is fixed as in Wu and Liu (2009). We assume that only a small number of predictors influence the conditional distribution of the response variable. We note that these sets of relevant predictors may vary for different segments of the conditional distribution. Therefore, consideration of different expectiles would enable us to explore the entire conditional distribution of the response variable and its sparsity pattern. The main contribution of the proposed work is threefold: (i) computation of expectile regression is straightforward and simple, (ii) theoretical development of expectile regression is more manageable with L_2 norm and the estimation is more efficient because it uses the entire distribution information, and (iii) expectile regression is more sensitive to extreme values than quantile regression, which results in better detection of heteroscedasticity in the data when it is present. In the case of heteroscedasticity, penalized expectile regression yields superior performance in estimation and variable selection, as demonstrated in our simulation study. On the other hand, quantile regression is more interpretable and robust to outliers in a homogeneous setting than expectile regression. Therefore, we propose penalized expectile regression to complement penalized quantile regression; together they provide a more complete picture of the entire conditional distribution of a response given predictors for heterogeneous data.

Recently, Gu and Zou (2016) develop penalized expectile regression in a high-dimensional setting when p is larger than n and propose a way of detecting heteroscedasticity using a two-step procedure. While their theoretical work considers a more generalized setting than ours by letting p grow to infinity, the proposed work provides the following two additions to the regularized regression literature that is not covered in Gu and Zou (2016): (i) we discuss pros and cons of penalized quantile and expectile regression and conduct a thorough simulation study to compare the finite sample performance of both approaches, and (ii) we provide asymptotic distributions of penalized expectile regression with SCAD and adaptive LASSO penalties for both $i.i.d.$ and non-$i.i.d.$ random errors.

The remainder of the paper is organized as follows: In Sect. 2, we review expectiles and describe the relationship with quantiles. Section 3 describes the proposed penalized expectile regression using SCAD and adaptive LASSO penalties, our tuning selection method, and presents its theoretical properties. In Sect. 4, we compare penalized expectile and quantile regression via simulated examples. Section 5 analyzes a real

example and compares prediction and variable selection results with penalized quantile regression. We provide concluding remarks and future directions in Sect. 6. Appendix details the proofs of theorems presented in Sect. 3.

2 Expectile

The τth, $\tau \in (0, 1)$, expectile is defined as the value of m which minimizes

$$E[|\tau - I(Y < m)|\,(Y - m)^2]$$

for a random variable Y. Jones (1994) shows that expectiles are quantiles of a distribution G with an explicit relation to the original distribution F:

$$G(y) = \frac{P(y) - yF(y)}{2(P(y) - yF(y)) + (y - \mu)},$$

where $P(y) = \int_{-\infty}^{y} xf(x)\mathrm{d}x$ and $\mu = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x$. Table 1 (Kuan et al. 2009) contains the corresponding quantile α to the expectile $\tau = 0.01, 0.03, 0.05, 0.1, 0.25$, respectively, for the Uniform($-a, a$), $N(0, 1)$, $t(30)$, $t(10)$, $t(5)$ and $t(3)$ distributions where $t(d)$ indicates the distribution with d degrees of freedom, illustrating the relationship above. Note that for a given expectile, the corresponding quantile depends on the distribution.

Quantile regression and expectile regression, which generalize median regression and mean regression, respectively, are of great interest mainly for two reasons. First, data analysts often want to obtain a more complete relationship among the variables in a regression setting rather than simple mean or median information. Second, a mean regression function could be an incomplete summary for the relationship when the assumption of a common normal distribution fails to hold, or when a heterogeneous variation is present.

Suppose, from some unknown distribution, we draw a random sample $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, where $\mathbf{x}_i$ and y_i denote the p-dimensional predictor and the response variable, respectively. Denote the vector of parameters as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$. The quantile

Table 1 τ and corresponding α for different distributions

τ	$U(-a, a)$	$N(0, 1)$	$t(30)$	$t(10)$	$t(5)$
0.01	0.092	0.043	0.040	0.035	0.030
0.05	0.186	0.126	0.123	0.115	0.100
0.10	0.250	0.195	0.190	0.183	0.166
0.25	0.366	0.332	0.328	0.322	0.319
0.75	0.634	0.669	0.671	0.678	0.689
0.90	0.750	0.806	0.810	0.819	0.835
0.95	0.813	0.873	0.877	0.886	0.901
0.99	0.909	0.957	0.960	0.965	0.973

regression estimators, proposed by Koenker and Bassett (1978), are defined as those vectors which minimize the function

$$Q_n(\boldsymbol{\beta}; \alpha) = \sum_{i=1}^{n} r_\alpha(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$$

over $\boldsymbol{\beta}$ in $\mathbb{R}^p$ for fixed values of $\alpha \in (0, 1)$, where $r_\alpha(\cdot)$ is a convex loss function of the form

$$r_\alpha(\epsilon) = |\alpha - I(\epsilon < 0)| \cdot |\epsilon|\,. \tag{1}$$

Similarly, the estimator of expectile regression can be obtained in the form of a vector that minimizes the following asymmetric least squares loss function:

$$R_n(\boldsymbol{\beta}; \tau) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

over $\boldsymbol{\beta}$ in $\mathbb{R}^p$ for fixed values of $\tau \in (0, 1)$, where $\rho_\tau(\cdot)$ is a convex loss function of the form

$$\rho_\tau(\epsilon) = |\tau - I(\epsilon < 0)| \cdot \epsilon^2. \tag{2}$$

We note that computation is no longer an issue for quantile regression due to recent advanced optimization techniques, but it is still true that computation in the expectile regression is more straightforward because it is based on the squared loss function in the form of (2). Thus, the optimization problem of expectile regression can be easily solved by many iteratively updated type algorithms, such as iteratively reweighted least squares (IRLS). Another attractive feature of expectile regression is that the expectile regression estimator depends on the shape of the entire distribution, while quantile regression estimator only relies on the percentiles of the estimated tail distribution. Hence, the expectile regression estimator contains additional information about the magnitude of the tail distribution and reflects the real value more accurately, especially for heavy-tailed distributions mentioned in Sect. 1. In terms of robustness, quantile regression is more resistant to outliers than expectile regression because quantile regression utilizes the L_1 norm. However, sensitivity to extreme values can be beneficial if detecting heteroscedasticity in the data is of the main interest. We demonstrate the utility of expectile regression in detecting heteroscedasticity of the distribution in our simulation study in Sect. 4.

3 Penalized expectile regression

With large-sized data, selecting relevant variables and obtaining an interpretable model are important in regression analysis. To achieve these two goals for expectile regression, we apply regularization approaches using two different types of penalty functions.

We consider the same random sample in Sect. 2. Denote the true vector of parameters as $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{p0})^{\mathrm{T}}$. Without loss of generality, we assume the first q elements of $\boldsymbol{\beta}_0$ are nonzero and the last $p - q$ elements are zero. That is, $\boldsymbol{\beta}_0$ can be written

as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{10}$ is a q-dimensional vector of nonzero elements and $\boldsymbol{\beta}_{20} = \mathbf{0}$, a $(p-q)$-dimensional vector of zero. We decompose $\boldsymbol{\beta}$ and $\mathbf{x}_i$ accordingly and write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{x}_i = (\mathbf{x}_{i1}^{\mathrm{T}}, \mathbf{x}_{i2}^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, \ldots, n$. We focus on penalized linear expectile regression and assume that y depends on $\mathbf{x}$ in a linear fashion. Namely, in our sample, we have the following linear model:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i = \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{x}_{i2}^{\mathrm{T}}\boldsymbol{\beta}_2 + \epsilon_i, \qquad i = 1, \ldots, n. \tag{3}$$

For some predetermined $\tau \in (0, 1)$, the τth expectile of the random error ϵ_i is zero.

We consider the following objective function of the penalized expectile regression model:

$$R_n(\boldsymbol{\beta}; \tau) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|), \tag{4}$$

where $\rho_\tau(\cdot)$ is defined in (2) and $p_{\lambda_n}(\cdot)$ is a penalty function with tuning parameter λ_n. The estimator of the penalized expectile regression minimizes (4). Among various penalty functions, we consider SCAD and adaptive LASSO because of their properties of unbiasedness, sparsity, and continuity. For SCAD,

$$p_{\lambda_n}(|\beta_j|) = \begin{cases} \lambda_n|\beta_j| & \text{if } |\beta_j| \le \lambda_n; \\ -\dfrac{|\beta_j|^2 - 2a\lambda_n|\beta_j| + \lambda_n^2}{2(a-1)} & \text{if } \lambda_n < |\beta_j| \le a\lambda_n; \\ \dfrac{(a+1)\lambda_n^2}{2} & \text{if } |\beta_j| > a\lambda_n \end{cases} \tag{5}$$

for some $a > 2$. For adaptive LASSO,

$$p_{\lambda_n}(|\beta_j|) = \lambda_n w_j|\beta_j|,$$

where w_j is a prespecified weight. In the classical linear regression setting, Zou (2006) suggests to construct the adaptive weights using the least squares estimates $\hat{\boldsymbol{\beta}}(\mathrm{ols})$, i.e., $w_j = 1/|\hat{\beta}_j(\mathrm{ols})|^\gamma$ for some chosen $\gamma > 0$. We use the expectile regression estimates without any penalty to construct our weights in the same fashion.

3.1 Asymptotic properties

We study the theoretical properties of the expectile regression with SCAD and adaptive LASSO penalties in a similar setting in Wu and Liu (2009). The following conditions are needed to facilitate the theoretical proofs.

Condition 1: For any given $\tau \in (0, 1)$, the errors $\{\epsilon_i, i = 1, 2, \ldots, n\}$ are independent and identically distributed, with τth expectile zero and a continuous, positive density $f(\cdot)$ in a neighborhood of zero. We further assume $\mathrm{E}(\epsilon_i^4) < \infty$.

Condition 2: The row vectors of the design matrix $\boldsymbol{X}$, $\{\mathbf{x}_i, i = 1, 2, \ldots, n\}$, are a deterministic sequence. We assume that there exists a positive definite matrix Σ

such that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \Sigma$. In addition, we assume $\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i\right)^2 = O(1)$. Denote the top-left q-by-q submatrix of Σ by Σ_{11} and the right-bottom $(p-q)$-by-$(p-q)$ submatrix of Σ by Σ_{22}.

For SCAD expectile regression, we show consistency and oracle properties.

Theorem 1 (Consistency). *Assume that Conditions 1 and 2 are satisfied. If $\lambda_n \to 0$ as $n \to \infty$, there exists a local minimizer of $R_n(\boldsymbol{\beta};\tau)$ in (4), $\hat{\boldsymbol{\beta}}^{(\mathrm{SCAD})}$, such that*

$$\|\hat{\boldsymbol{\beta}}^{(\mathrm{SCAD})} - \boldsymbol{\beta}_0\| = O_p(n^{-\frac{1}{2}}).$$

Denote $g_\tau(\epsilon_i) = \rho_\tau^{'}(\epsilon_i - t)\,|_{t=0} = -2\tau v I(v \geq 0) - 2(1-\tau)vI(v<0)$, $h_\tau(\epsilon_i) = \rho_\tau^{''}(\epsilon_i - t)\,|_{t=0} = 2\tau I(v \geq 0) + 2(1-\tau)I(v<0), i = 1,\ldots,n$, and $\sigma_{g_\tau}^2 = \mathrm{Var}(g_\tau(\epsilon_i))$, $\mu_{h_\tau} = \mathrm{E}(h_\tau(\epsilon_i))$.

Theorem 2 (Oracle property) *Assume that Conditions 1 and 2 are satisfied. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to one, the root-n consistent local minimizer $\hat{\boldsymbol{\beta}}^{(\mathrm{SCAD})} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1^{(\mathrm{SCAD})} \\ \hat{\boldsymbol{\beta}}_2^{(\mathrm{SCAD})} \end{pmatrix}$ in Theorem 1 satisfies*

(a) Sparsity: $\hat{\boldsymbol{\beta}}_2^{(\mathrm{SCAD})} = \mathbf{0}$;

(b) Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{(\mathrm{SCAD})} - \boldsymbol{\beta}_{10}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \sigma_{g_\tau}^2/\mu_{h_\tau}^2\,\Sigma_{11}^{-1})$. Here $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution and Σ_{11} is defined in Condition 2.

The oracle property also holds for adaptive LASSO expectile regression.

Theorem 3 (Oracle property). *Assume that Conditions 1 and 2 are satisfied. If $\sqrt{n}\lambda_n \to 0$ and $n^{(\gamma+1)/2}\lambda_n \to \infty$ as $n \to \infty$, then the adaptive LASSO expectile regression estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{AL})} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1^{(\mathrm{AL})} \\ \hat{\boldsymbol{\beta}}_2^{(\mathrm{AL})} \end{pmatrix}$, which minimizes (4), satisfies*

(a) Sparsity: $\hat{\boldsymbol{\beta}}_2^{(\mathrm{AL})} = \mathbf{0}$;

(b) Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{(\mathrm{AL})} - \boldsymbol{\beta}_{10}) \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \sigma_{g_\tau}^2/\mu_{h_\tau}^2\,\Sigma_{11}^{-1}\right)$.

Remark 1 Notice that the asymptotic covariance matrix of the penalized expectile regression estimator requires the fourth moment of the errors while penalized quantile estimator in Wu and Liu (2009) requires the second moment. However, the asymptotic covariance matrix of the penalized quantile regression estimator relies on the density function of the errors at the origin, which is usually unknown and difficult to estimate.

The following corollary extends Theorems 2 and 3 to non-$i.i.d.$ errors under the following assumptions:

Condition 3: As $n \to \infty$, $\frac{1}{n}\max_i \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i \mathrm{Var}(g_\tau(\epsilon_i)) \to 0$.

Condition 4: We assume that there exist positive definite matrices Σ^{g_τ} and Σ^{h_τ} such that $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathrm{Var}(g_\tau(\epsilon_i)) = \Sigma^{g_\tau}$ and $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathrm{E}(h_\tau(\epsilon_i)) = \Sigma^{h_\tau}$. Denote the top-left q-by-q submatrix of Σ^{g_τ} by $\Sigma_{11}^{g_\tau}$ and Σ^{h_τ} by $\Sigma_{11}^{h_\tau}$, respectively.

Corollary 1 *For non-i.i.d. random errors that satisfy Conditions 3 and 4, Theorems 2 and 3 hold with the limiting distribution*

$$N(\mathbf{0}, (\Sigma_{11}^{h_\tau})^{-1} \Sigma_{11}^{g_\tau} (\Sigma_{11}^{h_\tau})^{-1}).$$

The proofs of Theorems 1–3 and Corollary 1 are provided in Appendix.

Remark 2 Gu and Zou (2016) consider heteroscedasticity in the sense that they allow a different model depending on a different expectile of the conditional distribution. However, they assume i.i.d. for the regression error ϵ_i in their theory, while we consider both i.i.d. and non-i.i.d. cases in this paper. Our theory for non-i.i.d. can be applied to error distributions whose fourth moment exists as in Condition 1.

3.2 Computational algorithms

This subsection concerns the computational aspect of the proposed penalized expectile regression. Unlike quantile regression, which minimizes an asymmetrically weighted sum of absolute deviations of residuals, expectile regression solves a minimization problem of an asymmetrically weighted sum of squared residuals.

3.2.1 Adaptive LASSO penalty

In the case of the adaptive LASSO, the optimization problem becomes

$$\arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) + n\lambda_n \sum_{j=1}^{p} w_j |\beta_j| \right\}. \tag{6}$$

Problem (6) can be formulated as the standard form of quadratic programming. We take positive slack variables $\xi = (\xi_i)_{i=1,\ldots,n}$, $\zeta = (\zeta_i)_{i=1,\ldots,n}$ and $\mathbf{v} = (v_j)_{j=1,\ldots,p}$, and then, the equivalent problem of (6) is given as

$$\min_{\xi,\zeta,\mathbf{v},\beta} \left\{ \tau \sum_{i=1}^{n} \xi_i^2 + (1-\tau) \sum_{i=1}^{n} \zeta_i^2 + n\lambda_n \sum_{j=1}^{p} w_j v_j \right\},$$

subject to $\xi_i - \zeta_i = y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}, i = 1, \ldots, n$, and $-v_j \le \beta_j \le v_j, j = 1, \ldots, p$. To solve this, we use the Rmosek (Friberg 2014) interface to MOSEK (MOSEK ApS 2011), which is known to provide a flexible and reliable platform for convex programming (Koenker and Mizera 2014).

3.2.2 SCAD penalty

For the SCAD penalty with (5), because the objective function $R_n(\boldsymbol{\beta};\tau)$ in (4) is not convex, we apply the concave–convex procedure (CCCP) (Yuille and Rangarajan 2003; Zou and Li 2008; Kim et al. 2008). The CCCP searches for a local minimizer by successively minimizing the locally upper-tight convex function of the objective function. When the objective function can be decomposed into a sum of convex and concave functions, the locally upper-tight convex function can be obtained by locally linearizing the concave function. We name it as local linear approximation (LLA) algorithm (Zou and Li 2008).

For $R_n(\boldsymbol{\beta};\tau)$ in (4), note that $R_n(\boldsymbol{\beta};\tau) = R_n^{(\mathrm{conv})}(\boldsymbol{\beta};\tau) + R_n^{(\mathrm{conc})}(\boldsymbol{\beta};\tau)$, where $R_n^{(\mathrm{conv})}(\boldsymbol{\beta};\tau)$ is a convex function and $R_n^{(\mathrm{conc})}(\boldsymbol{\beta};\tau)$ is a differentiable concave function given by

$$R_n^{(\mathrm{conv})}(\boldsymbol{\beta};\tau) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + n\lambda_n \sum_{j=1}^{p} |\beta_j|,$$

$$R_n^{(\mathrm{conc})}(\boldsymbol{\beta};\tau) = \sum_{j=1}^{p} \big(np_{\lambda_n}(|\beta_j|) - n\lambda_n|\beta_j|\big).$$

Given an initial solution $\hat{\boldsymbol{\beta}}^{(1)}$, the locally upper-tight convex function of $R_n(\boldsymbol{\beta};\tau)$ becomes

$$R_n^{(\mathrm{conv})}(\boldsymbol{\beta};\tau) + \{\partial R_n^{(\mathrm{conc})}(\hat{\boldsymbol{\beta}}^{(1)};\tau)/\partial\boldsymbol{\beta}\}^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(1)}) + R_n^{(\mathrm{conc})}(\hat{\boldsymbol{\beta}}^{(1)};\tau).$$

Then it iteratively updates $\hat{\boldsymbol{\beta}}^{(1)}$ with the minimizer of the locally upper-tight convex function, which leads $\hat{\boldsymbol{\beta}}^{(1)}$ to converge eventually to a local minimizer, and the LLA algorithm is summarized as follows:

- Initialize $\hat{\boldsymbol{\beta}}^{(1)}$ and $\lambda_n > 0$.
- For $m = 1, 2, \ldots,$ update the following equation until convergence:

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \mathrm{argmin}_{\boldsymbol{\beta}} R_n^{(\mathrm{conv})}(\boldsymbol{\beta};\tau) + \{\partial R_n^{(\mathrm{conc})}(\hat{\boldsymbol{\beta}}^{(m)};\tau)/\partial\boldsymbol{\beta}\}^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m)}) + R_n^{(\mathrm{conc})}(\hat{\boldsymbol{\beta}}^{(m)};\tau).$$

Let $\mathcal{V}_1^{(m)} = \{j : |\hat{\beta}_j^{(m)}| \le \lambda_n\}$, $\mathcal{V}_2^{(m)} = \{j : \lambda_n < |\hat{\beta}_j^{(m)}| \le a\lambda_n\}$, and $\mathcal{V}_3^{(m)} = \{j : |\hat{\beta}_j^{(m)}| > a\lambda_n\}$, where $\hat{\beta}^{(m)}$ is the solution of the mth iteration step in the LLA. The mth problem excluding irrelevant parameters is equivalent to minimizing

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + n\lambda_n \sum_{j=1}^{p} w_j^{(m)}|\beta_j|, \tag{7}$$

where the weights $w_j^{(m)}$ are defined as follows: $w_j^{(m)} = 1$, $j \in \mathcal{V}_1^{(m)}$, $w_j^{(m)} = 1 - (|\hat{\beta}_j^{(m)}| - \lambda_n)/(a-1)\lambda_n$, $j \in \mathcal{V}_2^{(m)}$, and $w_j^{(m)} = 0$, $j \in \mathcal{V}_3^{(m)}$. Then, problem (7) has

the same form as (6). In our numerical study, we set $\hat{\boldsymbol{\beta}}^{(1)} = 0$ in the LLA algorithm, which leads to solving the LASSO problem at the initial stage.

We choose the tuning parameter λ to minimize the validation error

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}})$$

for the computed estimate $\hat{\boldsymbol{\beta}}$ as in Gu and Zou (2016) to make a fair comparison in our numerical study.

4 Simulation

In this section, we compare our proposed penalized expectile regression with the penalized quantile regression (Wu and Liu 2009) and the SALES (Gu and Zou 2016), using the SCAD and adaptive LASSO penalties. The SALES solve the optimization problem in (4) using the cyclic coordinate descent and proximal gradient algorithms. We consider two simulation settings.

Setting 1: We generate our data from the linear model with non-$i.i.d.$ random errors studied by Kocherginsky et al. (2005) and Wu and Liu (2009),

$$Y = 1 + X_1 + X_2 + X_3 + (1 + X_3)\epsilon, \tag{8}$$

where $X_2 = X_1 + X_3 + Z$, with both X_1 and Z generated from the standard normal distribution and X_3 generated from the uniform distribution on [0, 1]. The variables X_1, X_3, Z and ϵ are mutually independent.

Setting 2: We use a similar example in Wang et al. (2012) and Gu and Zou (2016):

$$Y = X_6 + X_{12} + X_{15} + X_{20} + (1 + 0.7X_1)\epsilon. \tag{9}$$

We first generate $(Z_1, \ldots, Z_p)^{\mathrm{T}}$ from the multivariate normal distribution with zero mean and the (i, j)th element of the covariance matrix $0.5^{|i-j|}$. Then, we set $X_1 = \Phi(Z_1)$ and $X_j = Z_j$ for $j = 2, 3, \ldots, p$, where Φ is the standard normal cumulative distribution function. Note that X_1 does not have an impact on the mean but only on the variance.

For each case, we generate independent training and validation data sets of size $n = 100$ and 200, and testing data set of size $n = 10,000$. The training data set is used to obtain the penalized expectile regression estimate given a fixed expectile value τ and a tuning parameter λ_n. Let λ_n range from 0.01 to 10 with the gap, $\log(\lambda_{i+1}) - \log(\lambda_i) = 0.1$, and it is chosen by the validation error described in Sect. 3.2. Then, a test error whose definition is given below is computed on the testing data set. We repeat the whole procedure 100 times and evaluate the performance of the penalized expectile and quantile regression. For penalized expectile regression, we consider five different expectiles $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$. Notice that expectiles and quantiles have one-to-one correspondence given a specific distribution (Table 1). Therefore, it is

fair to compare the test error of penalized τth expectile regression estimate and the corresponding penalized αth quantile regression estimate. We apply the LLA algorithm to both penalized expectile and quantile regressions.

For a comprehensive comparison among penalized expectile/quantile regressions and SALES, we consider the standard normal and a heavy- tailed distribution t(10) for ϵ in Setting 1, and the standard normal distribution only in Setting 2. For each distribution of ϵ, we study two different dimensions of the predictors, $p = 10$ and 50, by adding independent noise variables following the standard normal distribution, $N(0, 1)$ to the original X_1, X_2, and X_3 in Setting 1 and X_1, X_6, X_{12}, X_{15}, and X_{20} in Setting 2. For fair comparison, we calculate the corresponding quantile for each expectile for a given distribution. Since test errors use different loss functions, we report the absolute errors (AE) defined as the absolute distance between the estimated and the true parameters, i.e., $\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j|$. In the results, `ET` denotes the proposed expectile method, `ES` the SALES, and `QT` the quantile regression.

Table 2 reports the absolute errors for model (8) with ϵ generated from the standard normal and $t(10)$ distribution with $p = 10, 50$ and $n = 100, 200$. As expected, all methods produce smaller errors for lower dimension and larger sample size. For both distributions and both small and large predictors, it can be seen that there is no single estimator that dominates the performance in terms of the absolute errors. In comparison between `ALASSO ET` and `SCAD ET`, `ALASSO ET` tends to yield smaller errors for high $\tau = 0.75, 0.9$ while `SCAD ET` for low $\tau = 0.1, 0.25, 0.5$. They both produce smaller absolute errors than the SALES methods (`ALASSO ES` and `SCAD ES`) in many cases. We see that the performance of penalized expectile and quantile regression is competitive. Penalized expectile regression tends to perform better for p=10, and penalized quantile regression tends to perform better for $p = 50$ for both sample sizes.

Tables 3 and 4 report the variable selection results for Setting 1. Again, all methods show better selection for lower dimension and larger sample size. It can be seen that every method successfully selects X_2 with high probability (P_2) for all of the expectiles or quantiles. For X_1, `ALASSO ET` chooses it with the highest probability (P_1) for each level of expectile or quantile across different dimensional cases (0.91–1.00) followed by `SCAD ET` (0.86–1.00). Both `ES` and `QT` produce similar but slightly lower probabilities (0.74–1.00 and 0.79–1.00, respectively) compared to `ET` for both penalties. On the other hand, selecting X_3 is a challenging task for the six methods in the sense that X_3 can be easily dominated by $(1 + X_3)\epsilon$ in (8), because $X_3 \sim U(0, 1)$ and $\epsilon \sim N(0, 1)$. The proportions including X_3 (P_3) for `ET` are higher than those of `ES` in most cases and are similar to those of `QT` for both penalties. In conclusion, the proposed penalized expectile regression (`ET`) tends to produce smaller absolute errors, performs better variable selection for different distributions and dimensions across various τ values compared to the SALES (`ES`), and shows competitive performance compared to the penalized quantile regression in terms of absolute errors and variable selection.

Table 5 reports the absolute errors for model (9) with ϵ generated from the standard normal distribution with $p = 20, 50$ and $n = 100, 200$. Again, there is no dominant winner in terms of the absolute errors. It can be seen that `ET` and `ES` perform similarly and tend to yield smaller errors than `QT` in most cases.

Table 2 Absolute errors for Setting 1

n	p	τ		ALASSO			SCAD		
				ET	ES	QT	ET	ES	QT
100	10	0.10	Normal	$1.48_{(0.04)}$	$1.48_{(0.04)}$	$1.57_{(0.05)}$	$1.37_{(0.06)}$	$1.52_{(0.03)}$	$1.48_{(0.06)}$
			$t(10)$	$1.75_{(0.06)}$	$1.76_{(0.06)}$	$1.75_{(0.05)}$	$1.69_{(0.07)}$	$1.80_{(0.06)}$	$1.74_{(0.08)}$
		0.25	Normal	$1.31_{(0.04)}$	$1.30_{(0.04)}$	$1.39_{(0.05)}$	$1.16_{(0.05)}$	$1.43_{(0.03)}$	$1.22_{(0.05)}$
			$t(10)$	$1.46_{(0.05)}$	$1.45_{(0.05)}$	$1.46_{(0.04)}$	$1.32_{(0.06)}$	$1.61_{(0.05)}$	$1.34_{(0.06)}$
		0.50	Normal	$1.11_{(0.05)}$	$1.11_{(0.05)}$	$1.28_{(0.05)}$	$1.09_{(0.05)}$	$1.39_{(0.04)}$	$1.15_{(0.06)}$
			$t(10)$	$1.21_{(0.05)}$	$1.22_{(0.05)}$	$1.28_{(0.05)}$	$1.18_{(0.05)}$	$1.56_{(0.04)}$	$1.23_{(0.06)}$
		0.75	Normal	$1.13_{(0.05)}$	$1.14_{(0.05)}$	$1.23_{(0.06)}$	$1.13_{(0.05)}$	$1.50_{(0.04)}$	$1.25_{(0.07)}$
			$t(10)$	$1.27_{(0.06)}$	$1.28_{(0.06)}$	$1.33_{(0.06)}$	$1.34_{(0.06)}$	$1.74_{(0.05)}$	$1.46_{(0.08)}$
		0.90	Normal	$1.40_{(0.06)}$	$1.39_{(0.06)}$	$1.43_{(0.06)}$	$1.48_{(0.07)}$	$1.79_{(0.05)}$	$1.55_{(0.08)}$
			$t(10)$	$1.59_{(0.07)}$	$1.61_{(0.08)}$	$1.75_{(0.08)}$	$1.86_{(0.09)}$	$2.12_{(0.07)}$	$1.93_{(0.09)}$
	50	0.10	Normal	$1.80_{(0.06)}$	$1.78_{(0.06)}$	$1.93_{(0.07)}$	$1.67_{(0.08)}$	$1.77_{(0.06)}$	$1.65_{(0.07)}$
			$t(10)$	$1.93_{(0.07)}$	$1.93_{(0.07)}$	$2.08_{(0.07)}$	$1.78_{(0.08)}$	$1.99_{(0.06)}$	$1.88_{(0.08)}$
		0.25	Normal	$1.67_{(0.06)}$	$1.68_{(0.06)}$	$1.81_{(0.06)}$	$1.51_{(0.06)}$	$1.78_{(0.05)}$	$1.40_{(0.07)}$
			$t(10)$	$1.72_{(0.06)}$	$1.73_{(0.06)}$	$1.92_{(0.07)}$	$1.66_{(0.06)}$	$1.92_{(0.05)}$	$1.48_{(0.07)}$
		0.50	Normal	$1.65_{(0.06)}$	$1.62_{(0.06)}$	$1.82_{(0.07)}$	$1.47_{(0.06)}$	$1.92_{(0.05)}$	$1.34_{(0.08)}$
			$t(10)$	$1.62_{(0.06)}$	$1.61_{(0.06)}$	$1.72_{(0.07)}$	$1.61_{(0.06)}$	$2.01_{(0.05)}$	$1.28_{(0.06)}$
		0.75	Normal	$1.58_{(0.06)}$	$1.60_{(0.06)}$	$1.81_{(0.07)}$	$1.65_{(0.07)}$	$2.04_{(0.05)}$	$1.44_{(0.07)}$
			$t(10)$	$1.63_{(0.06)}$	$1.62_{(0.06)}$	$1.66_{(0.06)}$	$1.71_{(0.07)}$	$2.10_{(0.05)}$	$1.40_{(0.07)}$
		0.90	Normal	$1.82_{(0.07)}$	$1.82_{(0.07)}$	$1.97_{(0.07)}$	$1.79_{(0.08)}$	$2.09_{(0.06)}$	$1.85_{(0.09)}$
			$t(10)$	$1.81_{(0.08)}$	$1.80_{(0.08)}$	$2.11_{(0.08)}$	$2.03_{(0.09)}$	$2.20_{(0.05)}$	$2.11_{(0.11)}$

Table 2 continued

n	p	τ		ALASSO			SCAD		
				ET	ES	QT	ET	ES	QT
200	10	0.10	Normal	$1.33_{(0.03)}$	$1.34_{(0.03)}$	$1.37_{(0.03)}$	$1.33_{(0.04)}$	$1.37_{(0.03)}$	$1.26_{(0.04)}$
			$t(10)$	$1.51_{(0.04)}$	$1.52_{(0.05)}$	$1.53_{(0.04)}$	$1.49_{(0.05)}$	$1.51_{(0.04)}$	$1.54_{(0.05)}$
		0.25	Normal	$1.12_{(0.03)}$	$1.13_{(0.03)}$	$1.18_{(0.03)}$	$1.01_{(0.04)}$	$1.24_{(0.03)}$	$0.97_{(0.04)}$
			$t(10)$	$1.22_{(0.04)}$	$1.24_{(0.04)}$	$1.26_{(0.04)}$	$1.08_{(0.04)}$	$1.35_{(0.03)}$	$1.06_{(0.05)}$
		0.50	Normal	$0.76_{(0.03)}$	$0.77_{(0.03)}$	$0.96_{(0.04)}$	$0.76_{(0.04)}$	$0.99_{(0.04)}$	$0.76_{(0.04)}$
			$t(10)$	$0.90_{(0.04)}$	$0.89_{(0.04)}$	$0.92_{(0.04)}$	$0.83_{(0.04)}$	$1.18_{(0.04)}$	$0.81_{(0.04)}$
		0.75	Normal	$0.68_{(0.03)}$	$0.68_{(0.03)}$	$0.82_{(0.04)}$	$0.82_{(0.04)}$	$1.04_{(0.04)}$	$0.92_{(0.05)}$
			$t(10)$	$0.86_{(0.04)}$	$0.87_{(0.04)}$	$0.88_{(0.05)}$	$0.96_{(0.05)}$	$1.27_{(0.04)}$	$0.96_{(0.05)}$
		0.90	Normal	$0.93_{(0.05)}$	$0.94_{(0.05)}$	$1.08_{(0.05)}$	$1.21_{(0.05)}$	$1.39_{(0.05)}$	$1.35_{(0.06)}$
			$t(10)$	$1.25_{(0.05)}$	$1.26_{(0.06)}$	$1.21_{(0.06)}$	$1.46_{(0.07)}$	$1.74_{(0.05)}$	$1.45_{(0.05)}$
	50	0.10	Normal	$1.44_{(0.03)}$	$1.44_{(0.03)}$	$1.54_{(0.04)}$	$1.25_{(0.04)}$	$1.38_{(0.03)}$	$1.30_{(0.04)}$
			$t(10)$	$1.65_{(0.05)}$	$1.64_{(0.05)}$	$1.63_{(0.05)}$	$1.50_{(0.06)}$	$1.62_{(0.05)}$	$1.47_{(0.06)}$
		0.25	Normal	$1.32_{(0.04)}$	$1.32_{(0.03)}$	$1.38_{(0.03)}$	$1.09_{(0.05)}$	$1.41_{(0.03)}$	$1.01_{(0.04)}$
			$t(10)$	$1.43_{(0.05)}$	$1.43_{(0.05)}$	$1.42_{(0.04)}$	$1.21_{(0.04)}$	$1.55_{(0.04)}$	$1.10_{(0.05)}$
		0.50	Normal	$1.08_{(0.04)}$	$1.09_{(0.04)}$	$1.18_{(0.05)}$	$0.84_{(0.05)}$	$1.49_{(0.03)}$	$0.78_{(0.04)}$
			$t(10)$	$1.21_{(0.06)}$	$1.20_{(0.05)}$	$1.23_{(0.05)}$	$1.06_{(0.05)}$	$1.62_{(0.04)}$	$0.87_{(0.05)}$
		0.75	Normal	$0.93_{(0.05)}$	$0.92_{(0.05)}$	$1.04_{(0.05)}$	$0.92_{(0.05)}$	$1.62_{(0.04)}$	$0.87_{(0.05)}$
			$t(10)$	$1.09_{(0.05)}$	$1.10_{(0.05)}$	$1.16_{(0.06)}$	$1.10_{(0.05)}$	$1.79_{(0.05)}$	$1.00_{(0.05)}$
		0.90	Normal	$1.08_{(0.05)}$	$1.07_{(0.05)}$	$1.27_{(0.06)}$	$1.23_{(0.05)}$	$1.73_{(0.05)}$	$1.32_{(0.06)}$
			$t(10)$	$1.36_{(0.06)}$	$1.36_{(0.06)}$	$1.44_{(0.07)}$	$1.48_{(0.06)}$	$2.04_{(0.06)}$	$1.46_{(0.07)}$

Table 3 Variable selection results with $n = 100$ in Setting 1

n	p	τ		ALASSO			SCAD		
				ET	ES	QT	ET	ES	QT
100	10	0.10	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.98_{(0.01)}$	$0.98_{(0.01)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.25_{(0.04)}$	$0.24_{(0.04)}$	$0.22_{(0.04)}$	$0.38_{(0.05)}$	$0.04_{(0.02)}$	$0.37_{(0.05)}$
		0.25	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.98_{(0.01)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.43_{(0.05)}$	$0.44_{(0.05)}$	$0.36_{(0.05)}$	$0.44_{(0.05)}$	$0.13_{(0.03)}$	$0.53_{(0.05)}$
		0.50	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.64_{(0.05)}$	$0.64_{(0.05)}$	$0.56_{(0.05)}$	$0.68_{(0.05)}$	$0.29_{(0.05)}$	$0.73_{(0.04)}$
		0.75	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.82_{(0.04)}$	$0.82_{(0.04)}$	$0.75_{(0.04)}$	$0.89_{(0.03)}$	$0.49_{(0.05)}$	$0.88_{(0.03)}$
		0.90	P_1	$0.97_{(0.02)}$	$0.98_{(0.01)}$	$0.94_{(0.02)}$	$0.97_{(0.02)}$	$0.91_{(0.03)}$	$0.93_{(0.03)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_3	$0.82_{(0.04)}$	$0.83_{(0.04)}$	$0.79_{(0.04)}$	$0.94_{(0.02)}$	$0.52_{(0.05)}$	$0.92_{(0.03)}$
	50	0.10	P_1	$0.93_{(0.03)}$	$0.93_{(0.03)}$	$0.93_{(0.03)}$	$0.89_{(0.03)}$	$0.90_{(0.03)}$	$0.85_{(0.04)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.15_{(0.04)}$	$0.14_{(0.03)}$	$0.08_{(0.03)}$	$0.24_{(0.04)}$	$0.00_{(0.00)}$	$0.26_{(0.04)}$
		0.25	P_1	$0.95_{(0.02)}$	$0.95_{(0.02)}$	$0.89_{(0.03)}$	$0.91_{(0.03)}$	$0.92_{(0.03)}$	$0.88_{(0.03)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.18_{(0.04)}$	$0.20_{(0.04)}$	$0.22_{(0.04)}$	$0.33_{(0.05)}$	$0.00_{(0.00)}$	$0.33_{(0.05)}$
		0.50	P_1	$0.95_{(0.02)}$	$0.94_{(0.02)}$	$0.93_{(0.03)}$	$0.94_{(0.02)}$	$0.92_{(0.03)}$	$0.91_{(0.03)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.41_{(0.05)}$	$0.41_{(0.05)}$	$0.36_{(0.05)}$	$0.52_{(0.05)}$	$0.00_{(0.00)}$	$0.53_{(0.05)}$
		0.75	P_1	$0.95_{(0.02)}$	$0.95_{(0.02)}$	$0.93_{(0.03)}$	$0.91_{(0.03)}$	$0.87_{(0.03)}$	$0.89_{(0.03)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.98_{(0.01)}$
			P_3	$0.53_{(0.05)}$	$0.53_{(0.05)}$	$0.39_{(0.05)}$	$0.62_{(0.05)}$	$0.02_{(0.01)}$	$0.66_{(0.05)}$
		0.90	P_1	$0.91_{(0.03)}$	$0.91_{(0.03)}$	$0.85_{(0.04)}$	$0.86_{(0.03)}$	$0.72_{(0.05)}$	$0.79_{(0.04)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.98_{(0.01)}$
			P_3	$0.53_{(0.05)}$	$0.54_{(0.05)}$	$0.46_{(0.05)}$	$0.68_{(0.05)}$	$0.00_{(0.00)}$	$0.68_{(0.05)}$

Table 6 reports the variable selection results with $n = 100$ for Setting 2. We do not report the results for $n = 200$ because the results are similar. All methods successfully select the four variables in the mean (X_6, X_{12}, X_{15}, and X_{20}) with high probabilities. For X_1, all methods show low probabilities (P_1) because it is not part of the mean terms. Overall, `SCAD ES` has the lowest probabilities (0–0.03), and `ALASSO ES` performs similarly with `ALASSO ET` (0.05–0.26), followed by `ALASSO QT` (0.14–0.33). On the other hand, `SCAD QT` shows the highest probabilities (0.16–0.40). In conclusion, the proposed penalized expectile regression (`ET`) tends to produce similar

Table 4 Variable selection results with $n = 200$ for Setting 1

n	p	τ		ALASSO			SCAD		
				ET	ES	QT	ET	ES	QT
200	10	0.10	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.20_{(0.04)}$	$0.22_{(0.04)}$	$0.14_{(0.03)}$	$0.29_{(0.05)}$	$0.08_{(0.03)}$	$0.30_{(0.05)}$
		0.25	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.45_{(0.05)}$	$0.46_{(0.05)}$	$0.35_{(0.05)}$	$0.51_{(0.05)}$	$0.21_{(0.04)}$	$0.64_{(0.05)}$
		0.50	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.87_{(0.03)}$	$0.87_{(0.03)}$	$0.84_{(0.04)}$	$0.85_{(0.04)}$	$0.62_{(0.05)}$	$0.92_{(0.03)}$
		0.75	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.97_{(0.02)}$	$0.97_{(0.02)}$	$0.93_{(0.03)}$	$0.96_{(0.02)}$	$0.82_{(0.04)}$	$0.99_{(0.01)}$
		0.90	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$0.86_{(0.03)}$	$1.00_{(0.00)}$
	50	0.10	P_1	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.98_{(0.01)}$	$0.99_{(0.01)}$	$0.97_{(0.02)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.06_{(0.02)}$	$0.05_{(0.02)}$	$0.06_{(0.02)}$	$0.18_{(0.04)}$	$0.00_{(0.00)}$	$0.16_{(0.04)}$
		0.25	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$0.98_{(0.01)}$	$0.99_{(0.01)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.22_{(0.04)}$	$0.22_{(0.04)}$	$0.17_{(0.04)}$	$0.42_{(0.05)}$	$0.00_{(0.00)}$	$0.44_{(0.05)}$
		0.50	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.63_{(0.05)}$	$0.62_{(0.05)}$	$.51_{(0.05)}$	$0.79_{(0.04)}$	$0.05_{(0.02)}$	$0.80_{(0.04)}$
		0.75	P_1	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.87_{(0.03)}$	$0.87_{(0.03)}$	$0.84_{(0.04)}$	$0.94_{(0.02)}$	$0.21_{(0.04)}$	$0.96_{(0.02)}$
		0.90	P_1	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.97_{(0.02)}$	$0.99_{(0.01)}$
			P_2	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_3	$0.89_{(0.03)}$	$0.89_{(0.03)}$	$0.93_{(0.03)}$	$1.00_{(0.00)}$	$0.17_{(0.04)}$	$0.99_{(0.01)}$

absolute errors to those of ES and smaller errors than those of QT in this simulation setting. In terms of variable selection, ES selects X_1 the fewest times, followed by ET. QT often selects X_1, which might yield higher absolute errors as a result.

5 Real data analysis

The data set comes from a study that investigates different methodological problems associated with clean air using housing market data by Harrison and Rubinfeld (1978).

Table 5 Absolute errors for Setting 2

n	p	τ	ALASSO			SCAD		
			ET	ES	QT	ET	ES	QT
100	20	0.10	$1.83_{(0.06)}$	$1.84_{(0.06)}$	$2.14_{(0.10)}$	$1.84_{(0.08)}$	$1.56_{(0.05)}$	$2.01_{(0.11)}$
		0.25	$1.58_{(0.05)}$	$1.56_{(0.05)}$	$1.76_{(0.06)}$	$1.49_{(0.05)}$	$1.41_{(0.04)}$	$1.77_{(0.11)}$
		0.50	$1.51_{(0.04)}$	$1.51_{(0.04)}$	$1.74_{(0.07)}$	$1.37_{(0.05)}$	$1.38_{(0.04)}$	$1.54_{(0.07)}$
		0.75	$1.54_{(0.05)}$	$1.53_{(0.05)}$	$1.73_{(0.06)}$	$1.30_{(0.04)}$	$1.40_{(0.04)}$	$1.41_{(0.06)}$
		0.90	$1.67_{(0.06)}$	$1.68_{(0.06)}$	$1.91_{(0.08)}$	$1.44_{(0.06)}$	$1.53_{(0.05)}$	$1.62_{(0.08)}$
	50	0.10	$2.20_{(0.09)}$	$2.19_{(0.09)}$	$2.60_{(0.11)}$	$1.80_{(0.09)}$	$1.72_{(0.09)}$	$2.14_{(0.13)}$
		0.25	$1.91_{(0.07)}$	$1.90_{(0.06)}$	$2.29_{(0.09)}$	$1.57_{(0.06)}$	$1.52_{(0.04)}$	$1.84_{(0.11)}$
		0.50	$1.80_{(0.06)}$	$1.80_{(0.06)}$	$2.05_{(0.07)}$	$1.47_{(0.04)}$	$1.54_{(0.04)}$	$1.35_{(0.04)}$
		0.75	$1.85_{(0.07)}$	$1.84_{(0.06)}$	$2.04_{(0.08)}$	$1.44_{(0.04)}$	$1.55_{(0.05)}$	$1.38_{(0.06)}$
		0.90	$2.01_{(0.07)}$	$2.00_{(0.07)}$	$2.38_{(0.10)}$	$1.49_{(0.05)}$	$1.67_{(0.06)}$	$1.79_{(0.09)}$
200	20	0.10	$1.44_{(0.04)}$	$1.46_{(0.04)}$	$1.66_{(0.06)}$	$1.52_{(0.05)}$	$1.30_{(0.05)}$	$1.60_{(0.06)}$
		0.25	$1.24_{(0.03)}$	$1.24_{(0.03)}$	$1.39_{(0.04)}$	$1.26_{(0.04)}$	$1.15_{(0.03)}$	$1.39_{(0.06)}$
		0.50	$1.16_{(0.03)}$	$1.16_{(0.03)}$	$1.28_{(0.03)}$	$1.10_{(0.03)}$	$1.11_{(0.02)}$	$1.15_{(0.03)}$
		0.75	$1.14_{(0.03)}$	$1.14_{(0.03)}$	$1.25_{(0.04)}$	$1.01_{(0.03)}$	$1.15_{(0.02)}$	$1.04_{(0.04)}$
		0.90	$1.18_{(0.04)}$	$1.20_{(0.04)}$	$1.30_{(0.04)}$	$0.99_{(0.04)}$	$1.27_{(0.03)}$	$1.07_{(0.04)}$
	50	0.10	$1.42_{(0.05)}$	$1.43_{(0.05)}$	$1.53_{(0.05)}$	$1.40_{(0.05)}$	$1.20_{(0.05)}$	$1.48_{(0.05)}$
		0.25	$1.27_{(0.03)}$	$1.26_{(0.03)}$	$1.41_{(0.04)}$	$1.18_{(0.03)}$	$1.11_{(0.02)}$	$1.23_{(0.04)}$
		0.50	$1.22_{(0.03)}$	$1.21_{(0.02)}$	$1.37_{(0.03)}$	$1.11_{(0.02)}$	$1.11_{(0.02)}$	$1.16_{(0.03)}$
		0.75	$1.25_{(0.03)}$	$1.25_{(0.03)}$	$1.43_{(0.05)}$	$1.08_{(0.03)}$	$1.16_{(0.03)}$	$1.12_{(0.05)}$
		0.90	$1.31_{(0.03)}$	$1.31_{(0.03)}$	$1.43_{(0.04)}$	$1.06_{(0.03)}$	$1.21_{(0.03)}$	$1.08_{(0.03)}$

We apply its corrected version, which is available at http://lib.stat.cmu.edu/datasets/boston_corrected.txt as in Wu and Liu (2009). In total, there are 506 observations and 16 variables, among which *CMEDV* (corrected median value of owner-occupied homes) is the response and the other 14 non-constant predictors include *LON* (longitude), *LAT* (latitude), *CRIM* (crime rate by town), *ZN* (proportion of residential land zoned for large lots by town), *INDUS* (proportion non-retail business acres per town), *CHAS* (Charles River dummy: 1 if tract bounds the river; 0 if not), *NOX* (nitrogen oxide concentration), *RM* (average number of rooms), *AGE* (proportion of owner-occupied homes built prior to 1940), *DIS* (weighted distances to five employment centers in Boston), *TAX* (property tax rate), *PTRATIO* (pupil-teacher ratio by town), *B* (black population proportion), and *LSTAT* (proportion of lower socioeconomic status population). As in Wu and Liu (2009), we standardize the response *CMEDV* and the 14 continuous predictors. In the application of the penalized expectile regression, we consider *CMEDV* as the response, *CHAS* and all the other standardized continuous predictors and their corresponding squares as predictors, for a total of 27 predictors.

We follow a similar procedure as in Sect. 4. First, we randomly split the data set into training, validation, and testing data sets with size 200, 200, and 106, respectively. Second, we select the tuning parameter λ with the validation error introduced in

Table 6 Variable selection results with $n = 100$ for Setting 2

n	p	τ		ALASSO			SCAD		
				ET	ES	QT	ET	ES	QT
100	20	0.10	P_1	$0.25_{(0.04)}$	$0.26_{(0.04)}$	$0.25_{(0.04)}$	$0.38_{(0.05)}$	$0.00_{(0.00)}$	$0.40_{(0.05)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{12}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
		0.25	P_1	$0.13_{(0.03)}$	$0.14_{(0.03)}$	$0.20_{(0.04)}$	$0.17_{(0.04)}$	$0.01_{(0.01)}$	$0.30_{(0.05)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{12}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
		0.50	P_1	$0.17_{(0.04)}$	$0.18_{(0.04)}$	$0.18_{(0.04)}$	$0.15_{(0.04)}$	$0.01_{(0.01)}$	$0.26_{(0.04)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{12}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
		0.75	P_1	$0.22_{(0.04)}$	$0.20_{(0.04)}$	$0.25_{(0.04)}$	$0.21_{(0.04)}$	$0.03_{(0.02)}$	$0.32_{(0.05)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{12}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
		0.90	P_1	$0.25_{(0.04)}$	$0.21_{(0.04)}$	$0.33_{(0.05)}$	$0.31_{(0.05)}$	$0.02_{(0.01)}$	$0.40_{(0.05)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{12}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
	50	0.10	P_1	$0.13_{(0.03)}$	$0.12_{(0.03)}$	$0.22_{(0.04)}$	$0.18_{(0.04)}$	$0.00_{(0.00)}$	$0.27_{(0.04)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.97_{(0.02)}$
			P_{12}	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.98_{(0.01)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.97_{(0.02)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$0.98_{(0.01)}$
			P_{20}	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.95_{(0.02)}$
		0.25	P_1	$0.10_{(0.03)}$	$0.06_{(0.02)}$	$0.14_{(0.03)}$	$0.10_{(0.03)}$	$0.00_{(0.00)}$	$0.16_{(0.04)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{12}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.98_{(0.01)}$
		0.50	P_1	$0.05_{(0.02)}$	$0.05_{(0.02)}$	$0.10_{(0.03)}$	$0.06_{(0.02)}$	$0.00_{(0.00)}$	$0.16_{(0.04)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
			P_{12}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$

Table 6 continued

n	p	τ		ALASSO			SCAD		
				ET	ES	QT	ET	ES	QT
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
		0.75	P_1	$0.09_{(0.03)}$	$0.10_{(0.03)}$	$0.15_{(0.04)}$	$0.09_{(0.03)}$	$0.00_{(0.00)}$	$0.16_{(0.04)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$
			P_{12}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$
		0.90	P_1	$0.13_{(0.03)}$	$0.14_{(0.03)}$	$0.20_{(0.04)}$	$0.22_{(0.04)}$	$0.00_{(0.00)}$	$0.26_{(0.04)}$
			P_6	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.96_{(0.02)}$
			P_{12}	$0.99_{(0.01)}$	$0.99_{(0.01)}$	$0.97_{(0.02)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$0.96_{(0.02)}$
			P_{15}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.97_{(0.02)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.94_{(0.02)}$
			P_{20}	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.99_{(0.01)}$	$1.00_{(0.00)}$	$1.00_{(0.00)}$	$0.97_{(0.02)}$

Sect. 3.2. Then, we evaluate the test error by the average expectile check loss on the testing data. We replicate 100 times and report its average values. We still consider the five different expectiles $\tau = 0.1, 0.25, 0.5, 0.75$ and 0.9. Although the underlying distribution is unknown, by assuming the normal distribution, we apply the SALES and penalized quantile regression and compare their performance. For the quantile regression, we use the quantiles hat correspond to the five expectiles above, i.e., $\alpha = 0.19, 0.33, 0.5, 0.66$ and 0.8.

Table 7 displays the test error and variable selection results for cases of $p = 14$ and $p = 27$. In the table, TE is the average expectile check loss on the testing data. Size refers to the average number of predictors selected in the final model. For the penalized expectile regression `ET` and `ES`, both SCAD and adaptive LASSO produce similar test errors. Although they are not directly comparable, both penalized expectile regression methods yield smaller test errors than penalized quantile regression across different segments of the distribution. In terms of variable selection, for both SCAD and adaptive LASSO penalties, `ET` tends to produce sparser models than `ES` and `QT` for $p = 14$. When $p = 27$, `ALASSO QT` produces sparser models than the others except at $\tau = 0.1$, and `SCAD ET` tends to produce larger models.

Figure 1 shows the estimated coefficients by the proposed penalized expectile regression with both ALASSO and SCAD when $p = 14$. It can be seen that some regression coefficients vary as τ changes, for example *CHAS*, *DIS*, and *TAX*. In particular, the estimate for *TAX* shows an interesting pattern because the estimate is increasing as τ increases and is close to zero when $\tau = 0.5$. Hence, we can interpret that the property tax rate is not an important variable in the mean function, but is important in other expectiles.

Table 7 Test errors and variable selection results for real data

p	τ		ALASSO			SCAD		
			ET	ES	QT	ET	ES	QT
14	0.10	TE	$0.058_{(0.001)}$	$0.056_{(0.001)}$	$0.099_{(0.002)}$	$0.056_{(0.001)}$	$0.056_{(0.001)}$	$0.096_{(0.002)}$
		Size	$8.83_{(0.17)}$	$11.81_{(0.25)}$	$9.96_{(0.13)}$	$12.26_{(0.24)}$	$11.26_{(0.23)}$	$12.48_{(0.17)}$
	0.25	TE	$0.105_{(0.002)}$	$0.104_{(0.002)}$	$0.135_{(0.004)}$	$0.104_{(0.002)}$	$0.105_{(0.002)}$	$0.132_{(0.004)}$
		Size	$9.90_{(0.14)}$	$11.92_{(0.16)}$	$10.31_{(0.14)}$	$12.15_{(0.22)}$	$11.70_{(0.23)}$	$12.15_{(0.17)}$
	0.50	TE	$0.154_{(0.004)}$	$0.153_{(0.004)}$	$0.163_{(0.005)}$	$0.154_{(0.004)}$	$0.154_{(0.004)}$	$0.160_{(0.005)}$
		Size	$9.64_{(0.15)}$	$10.95_{(0.19)}$	$9.91_{(0.15)}$	$10.83_{(0.23)}$	$11.15_{(0.24)}$	$11.69_{(0.20)}$
	0.75	TE	$0.164_{(0.004)}$	$0.164_{(0.004)}$	$0.175_{(0.006)}$	$0.164_{(0.004)}$	$0.164_{(0.004)}$	$0.174_{(0.006)}$
		Size	$9.17_{(0.19)}$	$10.34_{(0.23)}$	$9.06_{(0.17)}$	$9.98_{(0.26)}$	$10.55_{(0.29)}$	$10.40_{(0.25)}$
	0.90	TE	$0.133_{(0.004)}$	$0.133_{(0.004)}$	$0.173_{(0.007)}$	$0.134_{(0.004)}$	$0.132_{(0.004)}$	$0.172_{(0.006)}$
		Size	$8.94_{(0.21)}$	$10.24_{(0.27)}$	$8.31_{(0.20)}$	$9.20_{(0.29)}$	$9.50_{(0.36)}$	$9.80_{(0.31)}$
27	0.10	TE	$0.066_{(0.003)}$	$0.065_{(0.003)}$	$0.080_{(0.003)}$	$0.069_{(0.003)}$	$0.063_{(0.003)}$	$0.084_{(0.003)}$
		Size	$10.99_{(0.62)}$	$12.58_{(0.71)}$	$13.90_{(0.31)}$	$14.16_{(0.77)}$	$13.70_{(0.62)}$	$17.54_{(0.57)}$
	0.25	TE	$0.090_{(0.003)}$	$0.089_{(0.003)}$	$0.100_{(0.004)}$	$0.092_{(0.004)}$	$0.089_{(0.003)}$	$0.104_{(0.005)}$
		Size	$16.35_{(0.53)}$	$17.24_{(0.49)}$	$15.25_{(0.26)}$	$20.67_{(0.55)}$	$17.13_{(0.54)}$	$18.99_{(0.55)}$
	0.50	TE	$0.108_{(0.004)}$	$0.108_{(0.004)}$	$0.115_{(0.005)}$	$0.110_{(0.004)}$	$0.111_{(0.004)}$	$0.114_{(0.005)}$
		Size	$17.33_{(0.32)}$	$19.17_{(0.39)}$	$15.74_{(0.32)}$	$21.11_{(0.54)}$	$19.15_{(0.55)}$	$20.35_{(0.59)}$
	0.75	TE	$0.104_{(0.004)}$	$0.102_{(0.004)}$	$0.115_{(0.005)}$	$0.103_{(0.004)}$	$0.107_{(0.004)}$	$0.112_{(0.005)}$
		Size	$18.16_{(0.30)}$	$20.19_{(0.41)}$	$16.32_{(0.34)}$	$21.73_{(0.49)}$	$20.74_{(0.57)}$	$21.57_{(0.54)}$
	0.90	TE	$0.080_{(0.003)}$	$0.078_{(0.003)}$	$0.104_{(0.005)}$	$0.079_{(0.003)}$	$0.080_{(0.003)}$	$0.103_{(0.004)}$
		Size	$17.68_{(0.43)}$	$20.11_{(0.45)}$	$15.88_{(0.33)}$	$21.39_{(0.50)}$	$20.74_{(0.57)}$	$20.14_{(0.48)}$

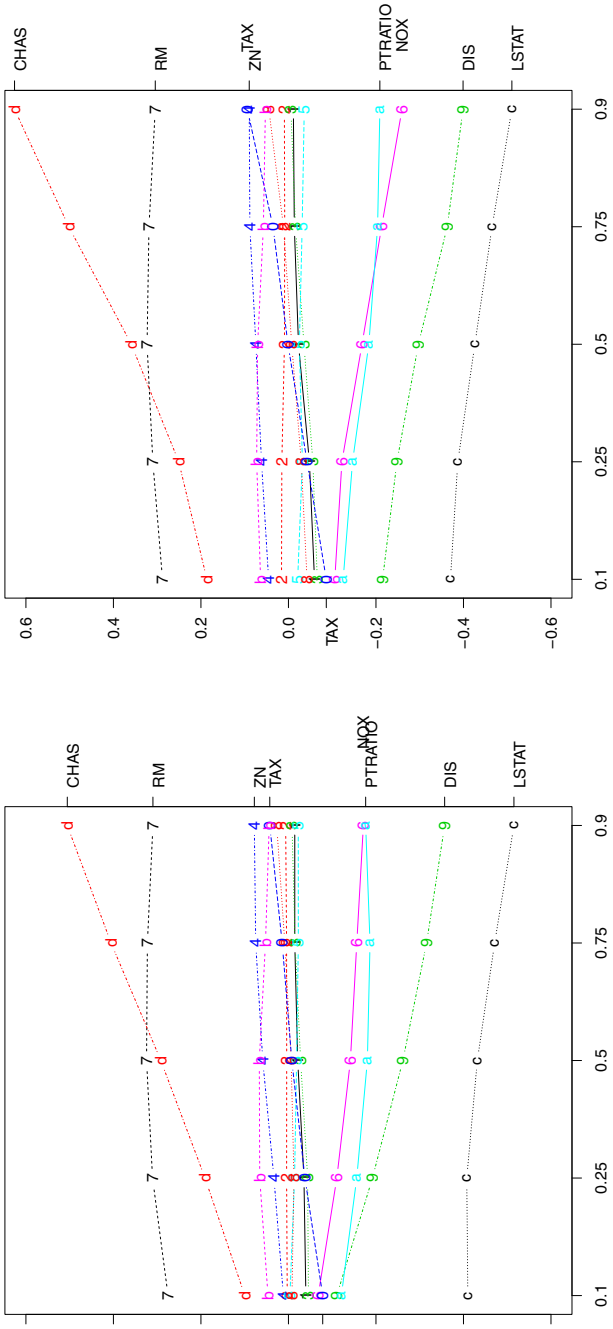

Fig. 1 Regression coefficient estimates at different τ values with $p = 14$ for the proposed penalized expectile regression

6 Discussion

In this work, we propose the penalized expectile regression using SCAD and adaptive LASSO penalties and show their oracle theoretical properties introduced by Fan and Li (2001), Zou (2006), and Wu and Liu (2009). Also, we compare the proposed method with the penalized quantile regression in Wu and Liu (2009) through a simulation study and real data analysis, and demonstrate its superior performance in terms of estimation and variable selection when heteroscedasticity is present. However, these advantages come at a cost: (i) although there is one-to-one match between the expectiles and quantiles, the expectiles do not have strong interpretability, and (ii) when data are homogeneous with some outliers rather than heterogeneous, penalized expectile regression is more affected by these outliers, while penalized quantile regression is more robust. The proposed work shows some attractive aspects of penalized expectile regression, and we recommend combining its use with penalized quantile regression.

An important problem in the field of penalized regression is construction of confidence intervals for the model coefficients. As Javanmard and Montanari (2014) point out, it is challenging to derive an exact sampling distribution of the parameter estimators in penalized regression due to the use of the optimization procedure, which is a main obstacle for conducting statistical inference using confidence intervals or hypothesis testing. Javanmard and Montanari (2014) propose an efficient algorithm for obtaining confidence intervals and p values by constructing a de-biased version of regularized M-estimators in a high-dimensional setting. Zhang and Zhang (2014) propose a hypothesis testing procedure for high-dimensional data using a low-dimensional projection approach. Lockhart et al. (2014) develop the covariance test statistic to determine the significance of regression coefficients in the sequence of models visited along the LASSO solution path. Also, resampling methods for hypothesis testing have been studied; Chatterjee and Lahiri (2010) apply the residual bootstrap approach to the LASSO estimator, and Minnier et al. (2011) propose a perturbation-based procedure to approximate the distribution of a general class of penalized parameter estimates, which leads to the estimation of the covariance matrix and confidence regions. We suggest statistical inference for penalized quantile and expectile regression as our future work.

Acknowledgements This work is part of the first author's dissertation. The third author's research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2007611).

Appendix: Proofs of theorems

Proof of Theorem 1

Following Wu and Liu (2009), it is sufficient to show that for any given $\delta > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|\mathbf{u}\|=C} R_n(\boldsymbol{\beta}_0 + \mathbf{u}/\sqrt{n}) > R_n(\boldsymbol{\beta}_0)\right\} \geq 1 - \delta. \tag{10}$$

It implies that there exists a local minimizer satisfying $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-\frac{1}{2}})$. Now consider

$$\begin{aligned} & R_n(\boldsymbol{\beta}_0 + \mathbf{u}/\sqrt{n}) - R_n(\boldsymbol{\beta}_0) \\ &= \sum_{i=1}^{n} \Big(\rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{x}_i^{\mathrm{T}}\mathbf{u}/\sqrt{n}) - \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) \Big) \\ &\quad + n \sum_{j=1}^{p} \Big(p_{\lambda_n}(|\beta_{j0} + u_j/\sqrt{n}|) - p_{\lambda_n}(|\beta_{j0}|) \Big). \end{aligned}$$

Because $p'_{\lambda_n}(\theta) = \lambda_n\{I(\theta \le \lambda_n) + \frac{(a\lambda_n - \theta)_+}{(a-1)\lambda_n} I(\theta > \lambda_n)\} \ge 0$ for some $a > 2$ and $\theta > 0$, and $p_{\lambda_n}(0) = 0$,

$$n(p_{\lambda_n}(|\beta_{j0} + u_j/\sqrt{n}|) - p_{\lambda_n}(|\beta_{j0}|)) = n(p_{\lambda_n}(|u_j/\sqrt{n}|) - p_{\lambda_n}(0)) \ge 0$$

for $j = q+1, \ldots, p$. Hence,

$$\begin{aligned} & R_n(\boldsymbol{\beta}_0 + \mathbf{u}/\sqrt{n}) - R_n(\boldsymbol{\beta}_0) \\ &\ge \sum_{i=1}^{n} \Big(\rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{x}_i^{\mathrm{T}}\mathbf{u}/\sqrt{n}) - \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) \Big) \\ &\quad + n \sum_{j=1}^{q} \Big(p_{\lambda_n}(|\beta_{j0} + u_j/\sqrt{n}|) - p_{\lambda_n}(|\beta_{j0}|) \Big). \end{aligned} \tag{11}$$

We first consider the second term on the right-hand side of (11). For $j = 1, \ldots, q$,

$$\begin{aligned} & n(p_{\lambda_n}(|\beta_{j0} + u_j/\sqrt{n}|) - p_{\lambda_n}(|\beta_{j0}|)) \\ &= n\Big(p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0}) \frac{u_j}{\sqrt{n}} + \frac{p''_{\lambda_n}(|\beta_{j0}|)}{2} \Big(\frac{u_j}{\sqrt{n}}\Big)^2 + o\Big(\frac{p''_{\lambda_n}(|\beta_{j0}|)}{n}\Big) \Big) \\ &= O\Big(\sqrt{n} \max_{1 \le j \le q} p'_{\lambda_n}(|\beta_{j0}|) + \max_{1 \le j \le q} p''_{\lambda_n}(|\beta_{j0}|) \Big). \end{aligned}$$

For large n,

$$\begin{aligned} p'_{\lambda_n}(|\beta_{j0}|) &= \lambda_n \Big(I(|\beta_{j0}| \le \lambda_n) + \frac{(a\lambda_n - |\beta_{j0}|)_+}{(a-1)\lambda_n} I(|\beta_{j0}| > \lambda_n) \Big) \\ &= \frac{(a\lambda_n - |\beta_{j0}|)_+}{a-1} \to 0 \text{ as } \lambda_n \to 0, \\ p''_{\lambda_n}(|\beta_{j0}|) &= -\frac{1}{a-1} I(\lambda_n < |\beta_{j0}| < a\lambda_n) \to 0 \text{ as } \lambda_n \to 0. \end{aligned}$$

Denote the first and second derivatives of $\rho_\tau(\epsilon_i - t)$ at $t = 0$ as follows:

$$g_\tau(\epsilon_i) = \rho'_\tau(\epsilon_i - t)\,|_{t=0} = -2\tau\epsilon_i I(\epsilon_i \ge 0) - 2(1-\tau)\epsilon_i I(\epsilon_i < 0),$$

$$h_\tau(\epsilon_i) = \rho_\tau^{''}(\epsilon_i - t)\mid_{t=0} = 2\tau I(\epsilon_i \geq 0) + 2(1-\tau)I(\epsilon_i < 0).$$

Then $\mathrm{E}(g_\tau(\epsilon_i)) = 0$. Denote $\mathrm{Var}(g_\tau(\epsilon_i)) = \sigma_{g_\tau}^2$, $\mathrm{E}(h_\tau(\epsilon_i)) = \mu_{h_\tau} > 0$ and $\mathrm{Var}(h_\tau(\epsilon_i)) = \sigma_{h_\tau}^2, i = 1, \ldots, n$. According to model (3.1), $\epsilon_i = y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0, i = 1, \ldots, n$. Now we consider the first term on the right-hand side of (11):

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(\rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{x}_i^{\mathrm{T}}\mathbf{u}/\sqrt{n}) - \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\Big) \\
&\quad = \sum_{i=1}^{n}\Big(g_\tau(\epsilon_i)\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{u}}{\sqrt{n}} + \frac{h_\tau(\epsilon_i)}{2}\Big(\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{u}}{\sqrt{n}}\Big)^2 + o\Big(\frac{1}{n}\Big)\Big).
\end{aligned}$$

We note that

$$\begin{aligned}
\sum_{i=1}^{n} g_\tau(\epsilon_i)\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{u}}{\sqrt{n}} &= \mathrm{E}\left(\sum_{i=1}^{n} g_\tau(\epsilon_i)\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{u}}{\sqrt{n}}\right) + O_p\left(\sqrt{\mathrm{Var}\left(\sum_{i=1}^{n} g_\tau(\epsilon_i)\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{u}}{\sqrt{n}}\right)}\right) \\
&= O_p\left(\sqrt{\mathbf{u}^{\mathrm{T}}\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\mathbf{u}\sigma_{g_\tau}^2}\right),
\end{aligned}$$

and

$$\begin{aligned}
\sum_{i=1}^{n}\frac{h_\tau(\epsilon_i)}{2}\left(\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{u}}{\sqrt{n}}\right)^2 &= \frac{\mu_{h_\tau}}{2}\mathbf{u}^{\mathrm{T}}\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\mathbf{u} + O_p\left(\sqrt{\frac{1}{4}\sum_{i=1}^{n}\left(\mathbf{u}^{\mathrm{T}}\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\mathbf{u}\right)^2\sigma_{h_\tau}^2}\right) \\
&= \frac{\mu_{h_\tau}}{2}\mathbf{u}^{\mathrm{T}}\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\mathbf{u} + o_p(1).
\end{aligned}$$

Therefore, $R_n(\boldsymbol{\beta}_0 + \mathbf{u}/\sqrt{n}) - R_n(\boldsymbol{\beta}_0)$ is dominated by $\frac{\mu_{h_\tau}}{2}\mathbf{u}^{\mathrm{T}}\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\mathbf{u}$, for $\|\mathbf{u}\| = C$, where C is sufficiently large. In conclusion, there exists a local minimizer of $R_n(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}^{(\mathrm{SCAD})}$, such that $\|\hat{\boldsymbol{\beta}}^{(\mathrm{SCAD})} - \boldsymbol{\beta}_0\| = O_p(n^{-\frac{1}{2}})$, if $\lambda_n \to 0$ as $n \to \infty$. □

Proof of Theorem 2

(a) For any $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-\frac{1}{2}}), 0 < \|\boldsymbol{\beta}_2\| \leq Cn^{-\frac{1}{2}}$,

$$\begin{aligned}
&R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}) - R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \boldsymbol{\beta}_2{}^{\mathrm{T}})^{\mathrm{T}}) \\
&\quad = \sum_{i=1}^{n}\Big(\rho_\tau(y_i - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1) - \rho_\tau(y_i - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1 - \mathbf{x}_{i2}^{\mathrm{T}}\boldsymbol{\beta}_2)\Big) - n\sum_{j=q+1}^{p} p_{\lambda_n}(|\beta_j|) \\
&\quad = \sum_{i=1}^{n}\Big(g_\tau(\epsilon_i)\mathbf{x}_{i1}^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) + \frac{h_\tau(\epsilon_i)}{2}\Big(\mathbf{x}_{i1}^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})\Big)^2 + o\Big((\mathbf{x}_{i1}^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}))^2\Big)\Big)
\end{aligned}$$

$$-\sum_{i=1}^{n}\Big(g_\tau(\epsilon_i)\mathbf{x}_i^{\mathrm{T}}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}+\frac{h_\tau(\epsilon_i)}{2}\Big(\mathbf{x}_{i1}^{\mathrm{T}}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}\Big)^2$$
$$+o\Big((\mathbf{x}_{i1}^{\mathrm{T}}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}})^2\Big)\Big)-n\sum_{j=q+1}^{p}p_{\lambda_n}(|\beta_j|). \tag{12}$$

By Condition 2 and following the proof of Theorem 1,

$$\sum_{i=1}^{n}g_\tau(\epsilon_i)\mathbf{x}_{i1}^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})=O_p\left(\sqrt{\sqrt{n}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}}\sum_{i=1}^{n}\frac{\mathbf{x}_{i1}\mathbf{x}_{i1}^{\mathrm{T}}}{n}\sqrt{n}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})\sigma_{g_\tau}^2}\right)$$
$$=O_p(1),$$
$$\sum_{i=1}^{n}g_\tau(\epsilon_i)\mathbf{x}_i^{\mathrm{T}}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$$
$$=O_p\left(\sqrt{\sqrt{n}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\sqrt{n}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}\sigma_{g_\tau}^2}\right)$$
$$=O_p(1),$$
$$\frac{h_\tau(\epsilon_i)}{2}\Big(\mathbf{x}_{i1}^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})\Big)^2=\frac{\mu_{h_\tau}}{2}\sqrt{n}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}}\sum_{i=1}^{n}\frac{\mathbf{x}_{i1}\mathbf{x}_{i1}^{\mathrm{T}}}{n}\sqrt{n}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})+o_p(1)$$
$$=O_p(1),$$
$$\frac{h_\tau(\epsilon_i)}{2}\Big(\mathbf{x}_i^{\mathrm{T}}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}\Big)^2$$
$$=\frac{\mu_{h_\tau}}{2}\sqrt{n}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\sqrt{n}((\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}+o_p(1)=O_p(1).$$

Now we consider the last term on the right-hand side of (12). For $j=q+1,\ldots,p$,

$$\begin{aligned}p_{\lambda_n}(|\beta_j|)&=\lim_{\theta\to0^+}p_{\lambda_n}(\theta)+\lim_{\theta\to0^+}p'_{\lambda_n}(\theta)|\beta_j|+o(|\beta_j|)\\&=\lambda_n|\beta_j|+o(|\beta_j|).\end{aligned}$$

Therefore, $n\sum_{j=q+1}^{p}p_{\lambda_n}(|\beta_j|)=n\lambda_n\Big(\sum_{j=q+1}^{p}\Big(|\beta_j|+o(|\beta_j|/\lambda_n)\Big)\Big)$. Because $\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}=O_p(n^{-\frac{1}{2}}),o(|\beta_j|/\lambda_n)=o\Big(\frac{1}{\sqrt{n}\lambda_n}\Big)$. We note that $\sqrt{n}\lambda_n\to\infty,n\lambda_n\to\infty$ as $n\to\infty$ and $R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}},\mathbf{0}^{\mathrm{T}})^{\mathrm{T}})-R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}},\boldsymbol{\beta}_2{}^{\mathrm{T}})^{\mathrm{T}})$ is dominated by

$$-n\sum_{j=q+1}^{p}p_{\lambda_n}(|\beta_j|).$$

Consequently,

$$R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}) - R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \boldsymbol{\beta}_2{}^{\mathrm{T}})^{\mathrm{T}}) \to -\infty \text{ as } n \to \infty.$$

This completes the proof of part(a) of the theorem. □

(b) From Theorem 1 and part(a), we know $\hat{\boldsymbol{\beta}}_1$ is a root-n consistent local minimizer of $R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}})$. Let $\boldsymbol{\theta}_1 = \sqrt{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})$, i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10} + \boldsymbol{\theta}_1/\sqrt{n}$. Then

$$\begin{aligned} R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}) &= \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1) + n\sum_{j=1}^{q} p_{\lambda_n}(|\beta_j|) \\ &= \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_{10} - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\theta}_1/\sqrt{n}) + n\sum_{j=1}^{q} p_{\lambda_n}(|\beta_{j0} + \theta_j/\sqrt{n}|) \\ &\triangleq Q_n(\boldsymbol{\theta}_1). \end{aligned}$$

Because $\hat{\boldsymbol{\theta}}_1 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{(\mathrm{SCAD})} - \boldsymbol{\beta}_{10})$ is a local minimizer of $Q_n(\boldsymbol{\theta}_1)$,

$$\frac{\partial Q_n(\boldsymbol{\theta}_1)}{\partial \theta_j} \mid_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1} = 0,$$

for $j = 1, \ldots, q$. Now we decompose the derivative of $Q_n(\boldsymbol{\theta}_1)$ by parts:

$$\begin{aligned} \rho_\tau(y_i - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_{10} - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\theta}_1/\sqrt{n}) &= \rho_\tau(\epsilon_i) + g_\tau(\epsilon_i)\Big(-\frac{\mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\theta}_1}{\sqrt{n}}\Big) \\ &\quad + \frac{h_\tau(\epsilon_i)}{2}\Big(-\frac{\mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\theta}_1}{\sqrt{n}}\Big)^2 + o(1), \\ \frac{\partial}{\partial \theta_j}\rho_\tau(y_i - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_{10} - \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\theta}_1/\sqrt{n}) &= -g_\tau(\epsilon_i)\frac{x_{ij}}{\sqrt{n}} + h_\tau(\epsilon_i)\frac{\mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\theta}_1}{n}x_{ij}, \\ p_{\lambda_n}(|\beta_{j0} + \theta_j/\sqrt{n}|) &= p_{\lambda_n}(|\beta_{j0}|) + p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})\frac{\theta_j}{\sqrt{n}} \\ &\quad + \frac{p''_{\lambda_n}(|\beta_{j0}|)}{2}\Big(\frac{\theta_j}{\sqrt{n}}\Big)^2 + o\Big(\frac{1}{n}\Big). \end{aligned}$$

Therefore, as $n \to \infty$,

$$n\frac{\partial}{\partial \theta_j}p_{\lambda_n}(|\beta_{j0} + \theta_j/\sqrt{n}|) = n\Big(p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})\frac{1}{\sqrt{n}} + p''_{\lambda_n}(|\beta_{j0}|)\frac{\theta_j}{n}\Big) \to 0. \quad (13)$$

From the proof of Theorem 1, (13) holds. Plugging them in $\frac{\partial Q_n(\boldsymbol{\theta}_1)}{\partial \theta_j} \mid_{\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_1} = 0$, for $j = 1, \ldots, q$, we have

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \Big(- g_\tau(\epsilon_i) \frac{x_{ij}}{\sqrt{n}} + h_\tau(\epsilon_i) \frac{\mathbf{x}_{i1}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_1}{n} x_{ij} \Big) \\
&\quad + \sum_{j=1}^{q} n \Big(p'_{\lambda_n}(|\beta_{j0}|) \mathrm{sgn}(\beta_{j0}) \frac{1}{\sqrt{n}} + p''_{\lambda_n}(|\beta_{j0}|) \frac{\hat{\theta}_j}{n} \Big), \\
\mu_{h_\tau} \sum_{i=1}^{n} \frac{\mathbf{x}_{i1}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_1}{n} x_{ij} &= \sum_{i=1}^{n} \Big(g_\tau(\epsilon_i) \frac{x_{ij}}{\sqrt{n}} - \frac{(h_\tau(\epsilon_i) - \mu_{h_\tau}) \mathbf{x}_{i1}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_1}{n} x_{ij} \Big) \\
&\quad - \sum_{j=1}^{q} n \Big(p'_{\lambda_n}(|\beta_{j0}|) \mathrm{sgn}(\beta_{j0}) \frac{1}{\sqrt{n}} + p''_{\lambda_n}(|\beta_{j0}|) \frac{\hat{\theta}_j}{n} \Big), \\
\mu_{h_\tau} \sum_{i=1}^{n} \frac{\mathbf{x}_{i1} \mathbf{x}_{i1}^{\mathrm{T}}}{n} \hat{\boldsymbol{\theta}}_1 &= \sum_{i=1}^{n} g_\tau(\epsilon_i) \frac{\mathbf{x}_{i1}}{\sqrt{n}} - \sum_{i=1}^{n} \frac{(h_\tau(\epsilon_i) - \mu_{h_\tau}) \mathbf{x}_{i1} \mathbf{x}_{i1}^{\mathrm{T}}}{n} \hat{\boldsymbol{\theta}}_1 - \sum_{j=1}^{q} \mathbf{m}_{\lambda_n}(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\beta}_{10}), \\
\hat{\boldsymbol{\theta}}_1 &= \left(\mu_{h_\tau} \sum_{i=1}^{n} \frac{\mathbf{x}_{i1} \mathbf{x}_{i1}^{\mathrm{T}}}{n} \right)^{-1} \left(\sum_{i=1}^{n} g_\tau(\epsilon_i) \frac{\mathbf{x}_{i1}}{\sqrt{n}} - \sum_{i=1}^{n} \frac{(h_\tau(\epsilon_i) - \mu_{h_\tau}) \mathbf{x}_{i1} \mathbf{x}_{i1}^{\mathrm{T}}}{n} \hat{\boldsymbol{\theta}}_1 \right. \\
&\quad \left. - \sum_{j=1}^{q} \mathbf{m}_{\lambda_n}(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\beta}_{10}) \right),
\end{aligned}
$$

where $\mathbf{m}_{\lambda_n}(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\beta}_{10})$ is defined as a q-dimensional vector with the jth element $n\big(p'_{\lambda_n}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})\frac{1}{\sqrt{n}} + p''_{\lambda_n}(|\beta_{j0}|)\frac{\hat{\theta}_j}{n}\big)$. According to (13) and Condition 2, as $n \to \infty$, $\mu_{h_\tau} \sum_{i=1}^{n} \frac{\mathbf{x}_{i1}\mathbf{x}_{i1}^{\mathrm{T}}}{n} \to \mu_{h_\tau} \Sigma_{11}$, $\sum_{i=1}^{n} \frac{(h_\tau(\epsilon_i) - \mu_{h_\tau})\mathbf{x}_{i1}\mathbf{x}_{i1}^{\mathrm{T}}}{n} \to 0$, and $\mathbf{m}_{\lambda_n}(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\beta}_{10}) \to 0$. In addition, $\mathrm{E}\Big(g_\tau(\epsilon_i)\frac{\mathbf{x}_{i1}}{\sqrt{n}}\Big) = \mathbf{0}$, $i = 1, \ldots, n$,

$$
\begin{aligned}
&\mathrm{Var}\Big(\sum_{i=1}^{n} g_\tau(\epsilon_i) \frac{\mathbf{x}_{i1}}{\sqrt{n}} \Big) = \sigma_{g_\tau}^2 \frac{\sum_{i=1}^{n} \mathbf{x}_{i1} \mathbf{x}_{i1}^{\mathrm{T}}}{n} \to \sigma_{g_\tau}^2 \Sigma_{11}, \\
&\sum_{i=1}^{n} \mathrm{E}\left(\| g_\tau(\epsilon_i) \frac{\mathbf{x}_{i1}}{\sqrt{n}} \|^2 I\Big(\| g_\tau(\epsilon_i) \frac{\mathbf{x}_{i1}}{\sqrt{n}} \| > \xi \Big) \right) \le \sum_{i=1}^{n} \frac{\mathrm{E} \| g_\tau(\epsilon_i) \frac{\mathbf{x}_{i1}}{\sqrt{n}} \|^4}{\xi^2} \\
&\quad = \frac{1}{\xi^2} \mathrm{E}\Big(g_\tau^4(\epsilon_i) \Big) \sum_{i=1}^{n} \left(\frac{\mathbf{x}_{i1}^{\mathrm{T}} \mathbf{x}_{i1}}{n} \right)^2 \to 0,
\end{aligned}
$$

for any $\xi > 0$. Applying Lindeberg–Feller CLT, we have

$$
\sum_{i=1}^{n} g_\tau(\epsilon_i) \frac{\mathbf{x}_{i1}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathbf{w_1} \sim N(\mathbf{0}, \sigma_{g_\tau}^2 \Sigma_{11}).
$$

By Slutsky's theorem, $\hat{\boldsymbol{\theta}}_1 \xrightarrow{\mathcal{L}} \left(\mu_{h_\tau}\Sigma_{11}\right)^{-1}\mathbf{w_1}$. Then, we can conclude,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{(\mathrm{SCAD})} - \boldsymbol{\beta}_{10}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \sigma_{g_\tau}^2/\mu_{h_\tau}^2\,\Sigma_{11}^{-1}).$$

This completes the proof. □

Proof of Theorem 3

We first prove the asymptotic normality in part (b). Let $\boldsymbol{\theta} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. Then, we have

$$\begin{aligned}
V_n(\boldsymbol{\theta}) &\triangleq R_n(\boldsymbol{\beta}_0 + \boldsymbol{\theta}/\sqrt{n}) - R_n(\boldsymbol{\beta}_0)\\
&= \sum_{i=1}^{n}\Big(g_\tau(\epsilon_i)\Big(-\frac{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}}{\sqrt{n}}\Big) + \frac{h_\tau(\epsilon_i)}{2}\Big(-\frac{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}}{\sqrt{n}}\Big)^2 + o\Big(\frac{1}{n}\Big)\Big)\\
&\quad + n\lambda_n\sum_{j=1}^{p} w_j(|\beta_{j0} + \theta_j/\sqrt{n}| - |\beta_{j0}|)\\
&= \sum_{i=1}^{n} g_\tau(\epsilon_i)\Big(-\frac{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}}{\sqrt{n}}\Big) + \frac{\mu_{h_\tau}}{2}\boldsymbol{\theta}^{\mathrm{T}}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\sum_{i=1}^{n}\Big(\frac{(h_\tau(\epsilon_i) - \mu_{h_\tau})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\Big)\boldsymbol{\theta}\\
&\quad + o_p(1) + n\lambda_n\sum_{j=1}^{p} w_j(|\beta_{j0} + \theta_j/\sqrt{n}| - |\beta_{j0}|). \qquad (14)
\end{aligned}$$

From the proof of Theorem 2,

$$\begin{aligned}
&\sum_{i=1}^{n} g_\tau(\epsilon_i)\frac{\mathbf{x}_i}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathbf{w} \sim N(\mathbf{0}, \sigma_{g_\tau}^2\,\Sigma),\\
&\frac{\mu_{h_\tau}}{2}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n} \to \frac{\mu_{h_\tau}}{2}\Sigma,\\
&\frac{1}{2}\sum_{i=1}^{n}\frac{(h_\tau(\epsilon_i) - \mu_{h_\tau})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n} \to 0.
\end{aligned}$$

Now we consider the last term of (14). For $1 \le j \le q$,

$$w_j \xrightarrow{\mathcal{P}} |\beta_{j0}|^{-\gamma}, \sqrt{n}\big(|\beta_{j0} + \theta_j/\sqrt{n}| - |\beta_{j0}|\big) \to \theta_j\,\mathrm{sgn}(\beta_{j0}).$$

By Slutsky's theorem, $n\lambda_n\sum_{j=1}^{p} w_j\big(|\beta_{j0}+\theta_j/\sqrt{n}|-|\beta_{j0}|\big) \xrightarrow{\mathcal{P}} 0$ because $\sqrt{n}\lambda_n \to 0$ as $n \to \infty$. For $q+1 \le j \le p$, $\beta_{j0} = 0$ and

$$\sqrt{n}(|\beta_{j0} + \theta_j/\sqrt{n}| - |\beta_{j0}|) = |\theta_j|, \sqrt{n}\lambda_n w_j = \lambda_n n^{(\gamma+1)/2}(|\sqrt{n}\tilde{\beta}_j|)^{-\gamma},$$

where $\tilde{\beta}_j$ is the jth element of $\tilde{\boldsymbol{\beta}}$ defined in (3.5) and $\sqrt{n}\tilde{\beta}_j = O_p(1)$. Therefore

$$n\lambda_n \sum_{j=1}^{p} w_j(|\beta_{j0} + \theta_j/\sqrt{n}| - |\beta_{j0}|) \begin{cases} \xrightarrow{\mathcal{P}} \infty & \text{if } \theta_j \neq 0, \\ = 0 & \text{if } \theta_j = 0. \end{cases}$$

Applying Slutsky's theorem again, we have $V_n(\boldsymbol{\theta}) \xrightarrow{\mathcal{L}} V(\boldsymbol{\theta})$ for every $\boldsymbol{\theta}$. Here,

$$V(\boldsymbol{\theta}) = \begin{cases} \dfrac{\mu_{h_\tau}}{2} \boldsymbol{\theta}_1{}^{\mathrm{T}} \Sigma_{11} \boldsymbol{\theta}_1 + \mathbf{w}_1^{\mathrm{T}} \boldsymbol{\theta}_1 & \text{if } \theta_j = 0, q+1 \leq j \leq p, \\ \infty & \text{otherwise.} \end{cases}$$

where $\mathbf{w}_1 = (w_1, w_2, \ldots, w_q)^{\mathrm{T}} \sim N(\mathbf{0}, \sigma_{g_\tau}^2 \Sigma_{11})$ and $\boldsymbol{\theta}_1 = (\theta_1, \theta_2, \ldots, \theta_q)^{\mathrm{T}}$. We note that $V_n(\boldsymbol{\theta})$ is convex and the unique minimum of $V(\boldsymbol{\theta})$ is

$$((-(\mu_{h_\tau} \Sigma_{11})^{-1} \mathbf{w}_1)^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}.$$

With the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{(\mathrm{AL})} - \boldsymbol{\beta}_{10}) = \hat{\boldsymbol{\theta}}_1 \xrightarrow{\mathcal{L}} -(b\Sigma_{11})^{-1} \mathbf{w}_1 \sim N(\mathbf{0}, \sigma_{g_\tau}^2 / \mu_{h_\tau}^2 \, \Sigma_{11}^{-1})$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_2^{(\mathrm{AL})} - \boldsymbol{\beta}_{20}) = \hat{\boldsymbol{\theta}}_2 \xrightarrow{\mathcal{L}} \mathbf{0}$$

where $\hat{\boldsymbol{\theta}}_2 = (\hat{\theta}_{q+1}, \hat{\theta}_{q+2}, \ldots, \hat{\theta}_p)^{\mathrm{T}}$, which proves the asymptotic normality property.

Next, we show the sparsity property. For any $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-\frac{1}{2}})$, $0 < \|\boldsymbol{\beta}_2\| \leq Cn^{-\frac{1}{2}}$, following the proof of Theorem 2, we have

$$\begin{aligned} & R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}) - R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \boldsymbol{\beta}_2{}^{\mathrm{T}})^{\mathrm{T}}) \\ &= \sum_{i=1}^{n} \left(g_\tau(\epsilon_i) \mathbf{x}_{i1}^{\mathrm{T}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) + \frac{h_\tau(\epsilon_i)}{2} \left(\mathbf{x}_{i1}^{\mathrm{T}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) \right)^2 + o\left((\mathbf{x}_{i1}^{\mathrm{T}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}))^2 \right) \right) \\ &\quad - \sum_{i=1}^{n} \left(g_\tau(\epsilon_i) \mathbf{x}_i^{\mathrm{T}} ((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} + \frac{h_\tau(\epsilon_i)}{2} \left(\mathbf{x}_{i1}^{\mathrm{T}} ((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} \right)^2 \right. \\ &\quad \left. + o_p\left((\mathbf{x}_{i1}^{\mathrm{T}} ((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}})^2 \right) \right) - n\lambda_n \sum_{j=q+1}^{p} w_j(|\beta_j|). \end{aligned} \tag{15}$$

The first two terms are bounded in the same way as the proof of Theorem 2:

$$\sum_{i=1}^{n} g_\tau(\epsilon_i) \mathbf{x}_{i1}^{\mathrm{T}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) = O_p\left(\sqrt{\sqrt{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}} \sum_{i=1}^{n} \frac{\mathbf{x}_{i1} \mathbf{x}_{i1}^{\mathrm{T}}}{n} \sqrt{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) \sigma_{g_\tau}^2} \right)$$

$$= O_p(1),$$

$$\sum_{i=1}^{n} g_\tau(\epsilon_i)\mathbf{x}_i^{\mathrm{T}}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$$

$$= O_p\left(\sqrt{\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}\sigma_{g_\tau}^2}}\right) = O_p(1),$$

$$\frac{h_\tau(\epsilon_i)}{2}\left(\mathbf{x}_{i1}^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})\right)^2$$

$$= \frac{\mu_{h_\tau}}{2}\sqrt{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}\sum_{i=1}^{n}\frac{\mathbf{x}_{i1}\mathbf{x}_{i1}^{\mathrm{T}}}{n}\sqrt{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) + o_p(1) = O_p(1),$$

$$\frac{h_\tau(\epsilon_i)}{2}\left(\mathbf{x}_i^{\mathrm{T}}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}\right)^2$$

$$= \frac{\mu_{h_\tau}}{2}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{n}\sqrt{n}((\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}} + o_p(1) = O_p(1).$$

For the third term on the right-hand side of (15),

$$n\lambda_n\sum_{j=q+1}^{p} w_j|\beta_j| = n^{(\gamma+1)/2}\lambda_n\sqrt{n}\sum_{j=q+1}^{p}(\sqrt{n}\tilde{\beta}_j)^{-\gamma}|\beta_j| \to \infty,$$

because $\sqrt{n}\tilde{\beta}_j = O_p(1)$ and $n^{(\gamma+1)/2}\lambda_n \to \infty$. Therefore,

$$R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}) - R_n((\boldsymbol{\beta}_1{}^{\mathrm{T}}, \boldsymbol{\beta}_2{}^{\mathrm{T}})^{\mathrm{T}}) \to -\infty \text{ as } n \to \infty.$$

This implies $\hat{\boldsymbol{\beta}}_2^{(\mathrm{AL})} = \mathbf{0}$. □

Proof of Corollary 1

From the proof of Theorem 2, it can be shown that

$$\sum_{i=1}^{n} g_\tau(\epsilon_i)\frac{\mathbf{x}_{i1}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathbf{w_1} \sim N(\mathbf{0}, \Sigma_{11}^g),$$

and $\hat{\boldsymbol{\theta}}_1 \xrightarrow{\mathcal{L}} \left(\Sigma_{11}^h\right)^{-1}\mathbf{w_1}$. □

References

Aigner, D., Amemiya, T., Poirier, D. (1976). On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review*, *17*, 377–396.

Belloni, A., Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, *39*, 82–130.
Belloni, A., Chernozhukov, V., Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, *102*, 77–94.
Chatterjee, A., Lahiri, S. N. (2010). Asymptotic properties of the residual bootstrap for Lasso estimators. *Proceedings of the American Mathematical Society*, *138*, 4497–4509.
Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.
Friberg, H. A. (2014) *Rmosek: The r-to-mosek optimization interface*. r package version 1.2.5.1.
Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics*, *22*, 1993–2010.
Gu, Y., Zou, H. (2016). High-dimensional generalizations of asymmetric least squares regression and their applications. *The Annals of Statistics*, *44*, 2661–2694.
Harrison, D., Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Environmental Economics and Management*, *5*, 81–102.
Javanmard, A., Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, *15*, 2869–2909.
Jones, M. C. (1994). Expectiles and m-quantiles are quantiles. *Statistics & Probability Letters*, *20*, 149–153.
Kim, Y., Choi, H., Oh, H. S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, *103*, 1665–1673.
Knight, K., Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*, 1356–1378.
Kocherginsky, M., He, X., Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, *14*(1), 41–55.
Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.
Koenker, R., Mizera, I. (2014). Convex optimization in R. *Journal of Statistical Software*, *60*, 1–23.
Kuan, C. M., Yeh, J. H., Hsu, Y. C. (2009). Assessing value at risk with care, the conditional autoregressive expectile models. *Journal of Econometrics*, *150*, 261–270.
Li, Y., Zhu, J. (2008). l_1-norm quantile regression. *Journal of Computational and Graphical Statistics*, *17*, 1–23.
Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R. (2014). A significance test for the Lasso. *The Annals of Statistics*, *42*, 413–468.
Minnier, J., Tian, T., Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, *106*, 1371–1382.
MOSEK ApS D. (2011). *The mosek optimization tool manual. version 7.0*. https://www.mosek.com.
Newey, W. K., Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, *55*, 819–847.
Schnabel, S. K., Eilers, P. H. C. (2009). Optimal expectile smoothing. *Computational Statistics and Data Analysis*, *53*, 4168–4177.
Sobotka, F., Radice, R., Marra, G., Kneib, T. (2013a). Estimating the relationship between women's education and fertility in Bostwana by using and instumental variable approach to semiparametric expectile regression. *Journal of the Royal Statistical Society, Series B*, *62*, 25–45.
Sobotka, F., Radice, R., Marra, G., Kneib, T. (2013b). On confidence intervals for semiparametric expectile regression. *Statistics and Computing*, *23*, 135–148.
Wang, L., Wu, Y., Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-hign dimension. *Journal of the American Statistical Association*, *107*, 214–222.
Wu, Y., Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, *19*, 801–817.
Yuille, A. L., Rangarajan, A. (2003). The concave–convex procedure. *Neural Computation*, *15*, 915–936.
Zhang, C. H., Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, *76*, 217–242.
Ziegel, J. (2014). Coherence and elicitability. *Mathematical Finance*, *26*, 901–918.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.
Zou, H., Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, *36*, 1509–1533.