Article:

Li, Xiang, Li, Yu-Ning orcid.org/0000-0003-1473-0146, Zhang, Li Xin et al. (1 more author) (2024) Inference for high-dimensional linear expectile regression with de-biasing method. Computational Statistics & Data Analysis. 107997. ISSN 0167-9473

https://doi.org/10.1016/j.csda.2024.107997

Inference for high-dimensional linear expectile regression with de-biasing method

Xiang Li[a,*], Yu-Ning Li[b], Li-Xin Zhang[d,a], Jun Zhao[c,**]

[a]*Zhejiang University, 866 Yuhangtang Rd, Hangzhou, 310058, China*
[b]*School for Business and Society, University of York, Heslington, York, YO10 5DD, United Kingdom*
[c]*School of Computer and Computing Science, Hangzhou City University, No. 48 Huzhou Street, Hangzhou, 310015, China*
[d]*School of Statistics and Mathematics, Zhejiang Gongshang University, No. 18 Xuezheng Street, Hangzhou, 310018, China*

Abstract

The methodology for the inference problem in high-dimensional linear expectile regression is developed. By transforming the expectile loss into a weighted-least-squares form and applying a de-biasing strategy, Wald-type tests for multiple constraints within a regularized framework are established. An estimator for the pseudo-inverse of the generalized Hessian matrix in high dimension is constructed using general amenable regularizers, including Lasso and SCAD, with its consistency demonstrated through a novel proof technique. Simulation studies and real data applications demonstrate the efficacy of the proposed test statistic in both homoscedastic and heteroscedastic scenarios.

Keywords: Amenable regularizer, De-biased Lasso, High-dimensional inference, Precision matrix estimation, Weighted least squares
2020 MSC: 62F05, 62F12, 62J12

1. Introduction

High-dimensional datasets have become increasingly prevalent in many fields, such as finance and genetics, presenting significant challenges for traditional regression methods such as least squares. In particular, these methods can lead to over-fitting and unreliable results in the high-dimensional setting, that is when the number of variables exceeds the sample size. By imposing sparsity constraints on the model coefficients, regularization techniques, such as the Lasso, have been developed to address these limitations. Nonetheless, sparsity constraints bring in complexities for high-dimensional inference, calling for further investigation and tailored methodologies.

Despite notable advancements made in the area of high-dimensional estimation and inference, a considerable portion of the existing literature primarily focuses on regression in mean, providing insights solely into the conditional mean of the response variable. To offer a more comprehensive understanding of the underlying distributional information, alternative regression methods such as quantile regression and expectile

*The Matlab codes are available at: https://github.com/XiangListat2024/expectileHDI

**Corresponding author at: School of Computer and Computing Science, Hangzhou City University, 310015, China
Email address: zhaojun@hzcu.edu.cn (Jun Zhao)

regression (e.g., Koenker, 2005; Newey and Powell, 1987) have been developed to estimate the conditional quantiles and expectiles, respectively, of the response variable. In recent years, there has been a growing interest in expectile regression applied to high-dimensional data, leading to the development of several regularized expectile regression methods. For instance, Wirsik et al. (2019) introduce an l_0-regularized expectile regression model with a Whittaker smoother, specifically tailored for modelling accelerometer data. Gu and Zou (2016) study the expectile regression with Lasso (l_1 regularizer) and nonconvex penalties. Further contributions to this domain include the works of Zhao et al. (2018), Zhao and Zhang (2018), Liao et al. (2019), Ciuperca (2021) and Xu et al. (2021), which study the expectile regression with SCAD, adaptive Lasso, and elastic-net regularizers and show the oracle properties of the estimators. Recently, Zhao et al. (2022) and Man et al. (2024) study the robust expectile regression in high dimensions using the generalized Huber loss with the generally folded concave regularizer and reweighted l_1 regularizer, respectively.

While the estimation of high-dimensional expectile regression has been well-studied, the statistical inference of such models remains an active area of investigation. The existing literature on the inference of expectile regression primarily focuses on the low-dimensional setting since Newey and Powell (1987). Zhao and Zhang (2018) develop de-biased estimation for inference of expectile regression when the number of variables is fixed. Additionally, Jiang et al. (2021) extend it to the single-index expectile model. Recently, Song et al. (2021) consider the inference for large-scale data using the divide and conquer algorithm, while Li et al. (2022a) introduce a weighted cumulative sum type statistic for a longitudinal multikink expectile regression model.

When it comes to the high-dimensional setting, the penalized estimators suffer from non-negligible bias due to the effect of the massive dimensionality, which makes them unfeasible to be used directly for inference. To overcome such a problem, van de Geer et al. (2014), Zhang and Zhang (2014) and Javanmard and Montanari (2014) introduce the de-biasing (or de-sparsifying) procedure with penalized least squares to correct the bias of the initial estimator so that the de-biased estimator can be used for inference purpose. Inspired by such an idea, many recent papers have focused on relevant generalizations for the de-biasing approach. Cai and Guo (2017) study the optimal expected lengths of confidence intervals for general linear functions of a high-dimensional regression vector from both the minimax and the adaptive perspectives. Dezeure et al. (2017) propose bootstrap methodology for individual and simultaneous inference in high-dimensional linear models with possibly non-Gaussian and heteroscedastic errors. Cai et al. (2023) study the generalized linear models (GLMs) with binary outcomes via optimization in quadratic form. Along this line of research, Chronopoulos et al. (2022) further make an extension to the time series framework while Li et al. (2023) combine the de-biasing approach with transfer learning. However, to our knowledge, no literature has discussed the statistical inference for the high dimensional GLMs in the context of the expectile regression.

In this paper, we aim to extend the study of the inference of linear expectile regression to the high-dimensional setting, which allows the number of explanatory variables to be much larger than the sample size. The main contributions of this article, the fundamental novelty of the proposed methodology and its connection to some recent literature are summarized as follows.

- Unlike the equal-weight methods for bias correction in the high-dimensional mean regression models (e.g., Zhang and Zhang, 2014; Javanmard and Montanari, 2014), we develop the bias correction procedure for high-dimensional expectile regression models with the expectile-specified random weights. Such a weighting strategy is essential to the de-biasing procedure of the inference problems in high-dimensional generalized linear models, e.g., van de Geer et al. (2014); Cai et al. (2023), and can be of independent interest to study. Furthermore, we extend the study by van de Geer et al. (2014) who consider a twice differentiable and local Lipstchiz loss function which excludes the expectile loss function. We develop an alternative proving strategy to show that the estimation errors in the preliminary estimator have little impact on the estimation of the expectile-specified random weights.

- We extend the choice of the regularizer to a broader category, that is the amenable regularizers discussed by Loh and Wainwright (2015) and Loh (2017). This extension provides an avenue for the application of a wider range of penalties, such as the SCAD regularizer, which exhibits favorable properties for high-dimensional model selection.

- Our paper also delves into precision matrix estimation in a high-dimensional setting, further advancing previous research by incorporating random weights and considering estimation errors in those weights. This extends the results by Fan et al. (2009) and Loh and Wainwright (2015) which also utilize the non-convex regularizer in the precision matrix estimation.

- we establish a test statistic for multivariate testing within the high-dimensional expectile regression framework. This test is applied in two empirical study cases, one in finance, revealing that momentary supply may possess predictive power in stock returns, and the other in genomics, screening important genes related to the *TLR8* gene.

The rest of the article is organized as follows. In Section 2, we introduce the high-dimensional expectile regression and the inference problem. In Section 3, we introduce the expectile-specified weighting matrix and propose the de-biased estimator of penalized expectile regression. Then we approximate the pseudo-inverse of generalized Hessian matrix by the node-wise regression. In Section 4, we establish a Wald-type test based on the asymptotic normality of the de-biased estimator. Section 5 provides numerical results. Section 6 applies the proposed method to a financial dataset and a genetic dataset. Section 7 contains more discussion along with possible future work.

Notations. For a $p-$dimensional vector $\boldsymbol{v} = (v_1, ..., v_p)^\top \in \mathbb{R}^p$, and $1 \leq q < \infty$, we define $\|\boldsymbol{v}\|_q = (\sum_{i=1}^p |v_i|^q)^{1/q}$, and particularly, we denote by $\|\boldsymbol{v}\|_\infty = \max_{1\leq i\leq p} |v_i|$ the infinity norm of a vector. Denote $\|\boldsymbol{v}\|_0 = |\mathrm{supp}(\boldsymbol{v})|$, where $\mathrm{supp}(\boldsymbol{v}) = \{i; v_i \neq 0\}$ is the active set. Moreover, we denote by $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{u}^\top \boldsymbol{v}$ the inner product of $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$. Furthermore, for $\mathcal{A} \subseteq \{1, ..., p\}$, denote $|\mathcal{A}|$ by the cardinality of $\mathcal{A}$. We let

$\boldsymbol{v}_{\mathcal{A}} = (v_i)_{i\in\mathcal{A}}$ and $\mathcal{A}^C$ be the complement of $\mathcal{A}$. For a differentiable function $g : \mathbb{R}^p \to \mathbb{R}$, we denote by $\nabla g(\boldsymbol{v}) = \partial g(\boldsymbol{v})/\partial \boldsymbol{v}$ and $\nabla_{\mathcal{A}} g(\boldsymbol{v}) = \partial g(\boldsymbol{v})/\partial \boldsymbol{v}_{\mathcal{A}}$. For a matrix $\mathbf{Q} = (Q_{ij})$, we denote by $\mathbf{Q}^\top$ the transpose of the matrix $\mathbf{Q}$, $\|\mathbf{Q}\|_\infty = \max_{ij} |Q_{ij}|$ the element-wise sup-norm, $\|\mathbf{Q}\|_{l_1} = \max_j \sum_i |Q_{ij}|$ the the l_1 norm, and $\|\mathbf{Q}\|_{l_\infty} = \max_i \sum_j |Q_{ij}|$ the the l_∞ norm. For symmetric matrix $\mathbf{Q}$, we denote by $\lambda_{\min}(\mathbf{Q})$ and $\lambda_{\max}(\mathbf{Q})$ its minimal and maximal eigenvalues. Particularly, we denote by $\mathbf{I}$ the unit matrix and by $\boldsymbol{e}_j$ its $j-$th column. Let $a \vee b$ and $a \wedge b$ denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively; and let $a_n \asymp b_n$ denote that $a_n = O(b_n)$ and $b_n = O(a_n)$ hold jointly.

2. High-dimensional expectile regression and the inference problem

In this section, we introduce the high-dimensional expectile regression framework with amenable regularizers, focusing on addressing the inference problem within this context.

We start by introducing the asymmetric squared loss function, or the so-called expectile loss function (e.g., Newey and Powell, 1987),

$$\rho_\tau(u) = |\tau - \mathbb{I}(u < 0)|u^2 = \begin{cases} \tau u^2, & u \geq 0, \\ (1-\tau)u^2, & u < 0, \end{cases}$$

where $\tau \in (0, 1)$ is a positive constant. Then the $\tau-$th expectile of a random variable Y can be defined as

$$m_\tau(Y) = \arg\min_m \mathsf{E}[\rho_\tau(Y - m)].$$

The expectile loss function assigns different weights on the squared loss associated with the sign of the residual $Y - m_\tau(Y)$. It is easy to check that $m_\tau(Y - m_\tau(Y)) = 0$ and $m_{1/2}(Y) = \mathsf{E}[Y]$. It is also worth mentioning that the expectile loss function is not second-order differentiable due to the discontinuity of the first-order derivative at 0.

Next, we introduce the high-dimensional expectile framework. Consider a set of n independent and identically distributed multivariate random variables $\{y_i, \boldsymbol{X}_i\}$, for $i \in \{1, ..., n\}$, where y_i is a scalar variable and $\boldsymbol{X}_i = (x_{i1}, ..., x_{ip})^\top$ is a p-dimensional covariate. The dimension p is allowed to grow with the sample size n and can even be much larger than n. We assume that they are generated by the following high-dimensional expectile linear model:

$$y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^* + \epsilon_i, \tag{1}$$

where $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathsf{E}[\rho_\tau(y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})]$ is a p-dimensional vector of parameters, ϵ_i is the error term. We omit the subscript τ of $\boldsymbol{\beta}^*$ hereafter when there is no confusion. For identification purpose, we assume that $\{\epsilon_i\}, i \in \{1, ..., n\}$ are i.i.d distributed and satisfy $m_\tau(\epsilon_i|\boldsymbol{X}_i) = 0$, or equivalently $\mathsf{E}\,[\rho'_\tau(\epsilon_i)|\boldsymbol{X}_i] = 0$. This framework allows a general conditional heteroscedastic setting or the so-called location-scale framework, that is $\epsilon_i = \sigma(\boldsymbol{X}_i)z_i$, where z_i is the innovation that is independent of $\boldsymbol{X}_i$ and the scale function $\sigma(\boldsymbol{X}_i)$ can be either of a linear form (e.g., Gu and Zou, 2016) or a non-parametric form (e.g., Fan and Yao, 1998).

Under this high-dimensional framework, a popular approach to estimate the coefficient $\boldsymbol{\beta}$ is given by the regularized asymmetric-least-squares (ALS),

$$\hat{\boldsymbol{\beta}} = \underset{\|\boldsymbol{\beta}\|_1 \leq R}{\arg\min} \left(L_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta})\right) = \underset{\|\boldsymbol{\beta}\|_1 \leq R}{\arg\min} \left(\frac{1}{2n} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) \right), \tag{2}$$

where $L_n(\boldsymbol{\beta})$ is the expectile loss function and the constant $1/2$ is introduced for the sake of simplicity in the statistical results. Regarding the regularizer $P_\lambda(\boldsymbol{\beta})$ with a tuning parameter $\lambda > 0$, we consider the amenable category which is developed by Loh and Wainwright (2015) and formally stated in Loh and Wainwright (2017). This category encompasses both convex and non-convex regularizers, including the Lasso, MCP, SCAD, among others. Due to the potential non-convexity of the regularizer, following Loh and Wainwright (2015, 2017), we add the constraint $\|\boldsymbol{\beta}\|_1 \leq R$ in order to ensure that a global minimum $\hat{\boldsymbol{\beta}}$ exists in (2), where R is a tuning parameter whose value may depend on n and p. Additionally, we specify the definition of the amenable regularizer as follows:

Definition 1. (Amenable regularizer). Suppose that the regularizer $P_\lambda(\boldsymbol{\beta})$ is separable across each coordinate, that is,

$$P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_\lambda(\beta_j),$$

where $p_\lambda : \mathbb{R} \to \mathbb{R}$ is some scalar function. If p_λ satisfies the following conditions (i)–(vi), we say that P_λ (or p_λ) is μ-amenable. Moreover, if p_λ further satisfies the condition (vii), we say that P_λ (or p_λ) is $(\mu, \gamma)-$amenable.

(i) The scalar function $p_\lambda(\cdot)$ is symmetric around zero and $p_\lambda(0) = 0$.

(ii) The scalar function $p_\lambda(\cdot)$ is non-decreasing on $\mathbb{R}^+$.

(iii) The scalar function $\dfrac{p_\lambda(t)}{t}$ is non-increasing on $\mathbb{R}^+$.

(iv) The scalar function $p_\lambda(t)$ is differentiable for all $t \neq 0$.

(v) $\lim_{t \to 0^+} p'_\lambda(t) = \lambda$.

(vi) There exists a constant $\mu > 0$ such that the scalar function $p_\lambda(t) + \dfrac{\mu}{2} t^2$ is convex.

(vii) There exists a constant $\gamma > 0$ such that $p'_\lambda(t) = 0$ for all $t \geq \gamma\lambda$.

The conditions (i)–(iii) are relatively mild and feasible for a large variety of regularizers, as also discussed in Zhang and Zhang (2012). The conditions (iv) and (v) exclude some unfeasible regularizers, such as the capped l_1 regularizer, which is not differentiable at many points on the positive real line, and the ridge regularizer due to its behavior at the origin. The condition (vi) represents for the weak convexity which is proposed by Vial (1982), and it can be viewed as a curvature constraint that controls the non-convexity of the regularizer. Moreover, the condition (vii) ensures the oracle property of the corresponding estimator and a similar condition can be found in Wang et al. (2014). More discussion and properties of the amenable regularizers can be found in Loh (2017) and Loh and Wainwright (2015, 2017).

Remark 1. The Lasso regularizer, $p_\lambda(u) = |u|$, is convex and $0-$amenable. However, it is not $(0, \gamma)-$amenable for any $0 < \gamma < \infty$.

Remark 2. The Smoothly clipped absolute deviation (SCAD) regularizer is non-convex, following Fan and Li (2001), takes the form

$$p_\lambda(u) = \lambda|u| \cdot \mathbb{I}(|u| \le \lambda) + \frac{2a\lambda|u| - u^2 - \lambda^2}{2(a-1)} \cdot \mathbb{I}(\lambda < |u| \le a\lambda) + \frac{(a+1)\lambda^2}{2} \cdot \mathbb{I}(a\lambda < |u|),$$

where $a > 2$ is a fixed constant. Moreover, the SCAD regularizer is $(\mu, \gamma)-$amenable, with $\mu = \dfrac{1}{a-1}$ and $\gamma = a$.

Our main interest falls on the statistical inference for the parameter vector $\boldsymbol{\beta}^*$ in the high-dimensional expectile linear regression model (1). To be specific, given a certain expectile level τ, we aim to test, for a given p_0-by-p full row-rank hypothesis matrix $\mathbf{H}$ and a p_0-dimensional vector $\boldsymbol{c}$,

$$\mathsf{H}_0 : \mathbf{H}\boldsymbol{\beta}^* = \boldsymbol{c} \quad \text{versus} \quad \mathsf{H}_1 : \mathbf{H}\boldsymbol{\beta}^* \neq \boldsymbol{c}, \tag{3}$$

where p_0 is the number of constraints and we further assume it to be a fixed integer given as a prior. For example, when $p_0 = 1$, $\boldsymbol{c} = 0$, and $\mathbf{H} = (1, 0, 0, \ldots, 0)$, we have $\mathsf{H}_0 : \beta_1^* = 0$. When $p_0 = 2$, $\boldsymbol{c} = (0, 0)^\top$, and $\mathbf{H} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \end{pmatrix}$, we have $\mathsf{H}_0 : \beta_1^* = \beta_2^* = 0$.

3. The de-biasing method under the expectile framework

In this section, we establish a test statistic for hypothesis (3) by developing the de-biasing method under the expectile framework.

3.1. The de-biased estimator

To begin with, we introduce a $\boldsymbol{\beta}-$related weighting matrix $\mathbf{W}_{\boldsymbol{\beta}} = \mathsf{diag}(w_{\boldsymbol{\beta},1}, ..., w_{\boldsymbol{\beta},n})$, where

$$w_{\boldsymbol{\beta},i} = |\tau - \mathbb{I}(y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} < 0)|^{1/2}. \tag{4}$$

Consequently, the expectile loss function in (2) can be rewritten as follows,

$$L_n(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^n w_{\boldsymbol{\beta},i}^2 (y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 = \frac{1}{2n} \|(\mathbf{W}_{\boldsymbol{\beta}} \boldsymbol{Y} - \mathbf{W}_{\boldsymbol{\beta}} \mathbf{X} \boldsymbol{\beta})\|_2^2,$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$.

Featured by the weighting matrix $\mathbf{W}_{\boldsymbol{\beta}}$, the expectile loss function shares a similarity with the Generalized Least Squares (GLS) loss function. However, it is noteworthy to point out that the weighting matrix depends on $\boldsymbol{\epsilon}$ and also depends on $\mathbf{X}$ when $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$. Then inspired by the de-biased Lasso proposed by van de Geer et al. (2014), we introduce the de-biased estimator of regularized expectile regression (e.g., Gu and Zou, 2016) as follows,

$$\hat{\boldsymbol{\beta}}_{de} = \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}} (\mathbf{W}_{\hat{\boldsymbol{\beta}}} \mathbf{X})^\top \mathbf{W}_{\hat{\boldsymbol{\beta}}} \hat{\boldsymbol{\epsilon}} / n, \tag{5}$$

where $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}$ is an estimate of the inverse of the covariance matrix (i.e., the precision matrix) of $w_{\boldsymbol{\beta}^*,i}\boldsymbol{X}_i$. Obtaining $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}$ poses a twofold challenge. Firstly, $w_{\boldsymbol{\beta}^*,i}$ is not directly observable, and we must estimate it using $w_{\hat{\boldsymbol{\beta}},i}$. Secondly, given that $p > n$, the (weighted) sample covariance matrix is non-invertible. While the second challenge is well-documented and solved in the literature, the first one requires further attention, which we address in Section 3.2. To be more specific, the de-biased estimator allows the following decomposition,

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{de} - \boldsymbol{\beta}^* &= \hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top \mathbf{W}^2_{\boldsymbol{\beta}^*}\boldsymbol{\epsilon}/n - (\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top \mathbf{W}^2_{\boldsymbol{\beta}^*}\boldsymbol{\epsilon}/n - \hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top \mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\boldsymbol{\epsilon}/n) - (\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} - \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&:= \hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top \mathbf{W}^2_{\boldsymbol{\beta}^*}\boldsymbol{\epsilon}/n - \boldsymbol{\Delta}^{(1)} - \boldsymbol{\Delta}^{(2)},
\end{aligned} \tag{6}
$$

where

$$
\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \mathbf{X}_{\boldsymbol{\beta}}^\top \mathbf{X}_{\boldsymbol{\beta}}/n, \quad \text{and} \quad \mathbf{X}_{\boldsymbol{\beta}} = \mathbf{W}_{\boldsymbol{\beta}}\mathbf{X}.
$$

The term $\boldsymbol{\Delta}^{(1)}$ is related to the approximation error of the weighting matrix $\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}$ with respect to $\mathbf{W}^2_{\boldsymbol{\beta}^*}$ and the term $\boldsymbol{\Delta}^{(2)}$ is related to the approximation of the inverse of the covariance matrix and the bias of the initial estimation. To utilize the proposed estimator for inference, it is essential to establish the asymptotic normality of $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top \mathbf{W}^2_{\boldsymbol{\beta}^*}\boldsymbol{\epsilon}/\sqrt{n}$ and demonstrate the negligibility of $\sqrt{n}\|\boldsymbol{\Delta}^{(1)}\|_\infty$ and $\sqrt{n}\|\boldsymbol{\Delta}^{(2)}\|_\infty$.

3.2. The pseudo-inverse from node-wise regression

In this section, we employ a widely-used node-wise regression method (e.g., Meinshausen and Bühlmann, 2006) to construct $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}$, which also serves as a pseudo-inverse of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$.

Let $\boldsymbol{X}_{\boldsymbol{\beta},(j)}$ denote the j-th column of the weighted design matrix $\mathbf{X}_{\boldsymbol{\beta}}$, and $\mathbf{X}_{\boldsymbol{\beta},(-j)}$ represent the weighted design matrix without the j-th column. To handle the high dimensionality, we apply node-wise regression within the classic regularized framework. Specifically, for each $j \in \{1, \ldots, p\}$, we seek the solution $\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}$ that minimizes the following objective function:

$$
\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} := \underset{\|\boldsymbol{\varphi}\|_1 \le R_j}{\arg\min} \left(\|\boldsymbol{X}_{\hat{\boldsymbol{\beta}},(j)} - \mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}\boldsymbol{\varphi}\|_2^2/(2n) + Q_{\lambda_j}(\boldsymbol{\varphi}) \right), \tag{7}
$$

where $\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}$ is a $(p-1)$-dimensional vector with elements $\hat{\varphi}_{\hat{\boldsymbol{\beta}},jl}, l \in \{1, \ldots, p\}, l \neq j$, and $Q_{\lambda_j}(\cdot)$ is the amenable regularizer with tuning parameter λ_j. Similar to the tuning parameter R in (2), R_j is a tuning parameter whose value may depend on n and p. Following Loh and Wainwright (2015), the constraint $\|\boldsymbol{\varphi}\|_1 \le R_j$ ensures a global minimum. In real applications, we can choose R and R_j to be sufficiently large numbers (e.g., Luo and Gao, 2022) or based on preliminary estimation (e.g., Jiang et al., 2023). By letting $\hat{\varphi}_{\hat{\boldsymbol{\beta}},jj} = -1$, we can define

$$
\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\beta}}} = (-\hat{\varphi}_{\hat{\boldsymbol{\beta}},ij}) = \begin{pmatrix} 1 & -\hat{\varphi}_{\hat{\boldsymbol{\beta}},12} & \cdots & -\hat{\varphi}_{\hat{\boldsymbol{\beta}},1p} \\ -\hat{\varphi}_{\hat{\boldsymbol{\beta}},21} & 1 & \cdots & -\hat{\varphi}_{\hat{\boldsymbol{\beta}},2p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\varphi}_{\hat{\boldsymbol{\beta}},p1} & -\hat{\varphi}_{\hat{\boldsymbol{\beta}},p2} & \cdots & 1 \end{pmatrix}.
$$

Moreover, we define a diagonal matrix $\hat{\mathbf{D}}^2_{\hat{\boldsymbol{\beta}}} = \mathsf{diag}(\hat{\phi}^2_{\hat{\boldsymbol{\beta}},1}, ..., \hat{\phi}^2_{\hat{\boldsymbol{\beta}},p})$, with

$$\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j} = \boldsymbol{X}^\top_{\hat{\boldsymbol{\beta}},(j)}(\boldsymbol{X}_{\hat{\boldsymbol{\beta}},(j)} - \mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n = \boldsymbol{X}^\top_{\hat{\boldsymbol{\beta}},(j)}\mathbf{X}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\beta}},j}/n.$$

Then we define the pseudo-inverse of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ as $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}$ with

$$\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}} = \hat{\mathbf{D}}^{-2}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\beta}}}.$$

With the definitions, we can write $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}},j} = \hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\beta}},j}/\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j}$ for $j \in \{1, \ldots, p\}$, where $\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\beta}},j}$ and $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}},j}$ are the j-th column of $\hat{\boldsymbol{\Phi}}^\top_{\hat{\boldsymbol{\beta}}}$ and $\hat{\boldsymbol{\Theta}}^\top_{\hat{\boldsymbol{\beta}}}$, respectively.

Remark 3. Although $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ is symmetric, its approximated inverse $\hat{\Theta}_{\hat{\boldsymbol{\beta}}}$ is not guaranteed to be symmetric. Moreover, the value of $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}$ relies on $\hat{\boldsymbol{\beta}}$, which is different from the classic de-biasing procedure under the mean regression and the classic high-dimensional precision matrix estimation problem, e.g., Friedman et al. (2008), Cai et al. (2011), and Liu and Wang (2017).

3.3. Testing procedure

We summary the testing procedure as follows,

- Step 1: Obtain the ALS estimator $\hat{\boldsymbol{\beta}}$ by solving the optimization problem (2) under the expectile framework with an amenable regularizer.
- Step 2: Estimate $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}$, the pseudo-inverse of the generalized Hessian matrix $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$, by applying the node-wise regression method (7) on the expectile-weighted design matrix $\mathbf{X}_{\hat{\boldsymbol{\beta}}}$ column by column.
- Step 3: Obtain the de-biased estimator $\hat{\boldsymbol{\beta}}_{de}$ by (5).
- Step 4: Construct a Wald-type test statistic,

$$T_{\mathbf{H}} = (\mathbf{H}(\hat{\boldsymbol{\beta}}_{de} - \boldsymbol{\beta}^*))^\top\hat{\boldsymbol{\Omega}}^{-1}_{\mathbf{H}}\mathbf{H}(\hat{\boldsymbol{\beta}}_{de} - \boldsymbol{\beta}^*), \tag{8}$$

for the hypothesis testing (3), where $\hat{\boldsymbol{\Omega}}_{\mathbf{H}} = \mathbf{H}\hat{\boldsymbol{\Omega}}\mathbf{H}^\top$ is the estimator of $\boldsymbol{\Omega}_{\mathbf{H}} = \mathbf{H}\boldsymbol{\Omega}\mathbf{H}^\top$, the asymptotic variance of $\mathbf{H}(\hat{\boldsymbol{\beta}}_{de} - \boldsymbol{\beta}^*)$ with $\boldsymbol{\Omega}$ being the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}_{de}$ defined by

$$\boldsymbol{\Omega} = (\omega_{ij}) = \boldsymbol{\Theta}_{\boldsymbol{\beta}^*}\mathsf{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top w^4_{\boldsymbol{\beta}^*,i}\epsilon_i^2\right]\boldsymbol{\Theta}^\top_{\boldsymbol{\beta}^*}, \tag{9}$$

and $\hat{\boldsymbol{\Omega}}$ its estimator defined by

$$\hat{\boldsymbol{\Omega}} = (\hat{\omega}_{ij}) = \hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top w^4_{\hat{\boldsymbol{\beta}},i}\hat{\epsilon}_i^2\right)\hat{\boldsymbol{\Theta}}^\top_{\hat{\boldsymbol{\beta}}}.$$

The test statistic asymptotically follows a χ^2 distribution with the degrees of freedom equal to the rank of $\mathbf{H}$, which we will show in Theorem 4. Given a statistical significance level α, the rejection region corresponding to the proposed test is

$$\left\{\mathbf{X} : T_{\mathbf{H}} > \chi^2_\alpha(p_0)\right\},$$

where p_0 is the rank of $\mathbf{H}$ and $\chi^2_\alpha(p_0)$ the upper α-quantile of a χ^2 distribution with p_0 degrees of freedom.

4. Assumptions and asymptotic results

In this section, we impose some technical conditions and establish statistical properties of the de-biased estimators.

4.1. Assumptions

Assumption 1. $(\boldsymbol{X}_i^\top, \epsilon_i)$ are i.i.d. random vectors with $\mathsf{E}[w^2_{\boldsymbol{\beta}^*,i}\epsilon_i|\boldsymbol{X}_i] = 0$, where $w_{\boldsymbol{\beta}^*,i}$ is defined by (4). Furthermore, for some positive constants $c_1, \ldots, c_6$,

(i) $\max_j \mathsf{E}[x_{ij}^4] < c_1$, $\|\mathbf{X}\|_\infty = O_p(K)$, $\Pr\{|x_{ij}| > c_2K\}) = o((\ln p)/n)$, where K is allowed to depend on both n and p;

(ii) $\mathsf{E}[\epsilon_i^4] < c_3$, $\max_{i=1}^n |\epsilon_i| = O_p(K)$, $\Pr\{|\epsilon_i| > c_4K\}) = o((\ln p)/n)$, and $\sup_{x\in(-\infty,+\infty)} f_\epsilon(x) < \infty$, where $f_\epsilon(\cdot)$ is the probability density function of ϵ;

(iii) $K^2\sqrt{\ln p/n} \to 0$ as $n, p \to \infty$;

(iv) $0 < c_5 < \lambda_{\min}(\boldsymbol{\Sigma}) < \lambda_{\max}(\boldsymbol{\Sigma}) < c_6 < \infty$.

The moment condition $\mathsf{E}[w^2_{\boldsymbol{\beta}^*,i}\epsilon_i|\boldsymbol{X}_i] = 0$ is the identification condition for the expectile regression. Since we do not impose independence between $\boldsymbol{X}_i$ and ϵ_i, this assumption allows for conditional heteroscedastic cases. The upper bound for $\|\mathbf{X}\|_\infty$ in Assumption 1(i) is also considered in (D1) of van de Geer et al. (2014), which is very general, as it includes the case for bounded variables with $K = 1$ and sub-Gaussian variables with $K = \sqrt{\ln(n \vee p)}$. More generally, heavy tail cases may also be included. For example, if X_{ij} and ϵ_i have bounded $(4 + \delta)$th moment, we can verify using Markov's inequality that the tail probability assumption holds with $K = (np)^{1/(4+\delta)}$. Assumption 1(ii) imposes a similar condition on ϵ_i.

Assumption 1(iii) imposes a restriction between n and p, and a similar assumption is considered in (D2) of van de Geer et al. (2014). When $\boldsymbol{X}_i$ and ϵ_i are bounded variables, then $K = 1$ and the assumption simplifies to $\sqrt{\ln p/n} \to 0$. When $\boldsymbol{X}_i$ and ϵ_i are sub-Gaussian variables, then $K = \sqrt{\ln(n \vee p)}$ and the assumption simplifies to $\ln(n \vee p)\sqrt{\ln p/n} \to 0$. In these two situations, we allow for the ultra-high-dimensional setting with $\ln(p) = O(n^\delta)$ and $0 < \delta < 1/3$. If X_{ij} and ϵ_i have bounded $(4 + \delta)$th moment, we can check that Assumption 1(iii) holds when $p \asymp n^\eta$ with $\delta > 4\eta$.

4.2. Preliminary theoretical results

We start by stating the l_2 error bound, the l_1 error bound and the prediction error bound of the ALS estimators $\hat{\boldsymbol{\beta}}$. Similar results have been proved by Gu and Zou (2016) and Li et al. (2022b) under the sub-Gaussian settings. Denoting by $\mathcal{A} = \{j; \beta_j^* \neq 0\}$ the active set of the covariates and its cardinality by $s := |\mathcal{A}|$, we have the following proposition.

Proposition 1. ***(The bounds for the initial estimators).*** *Suppose that Assumption 1 is satisfied. If $P_\lambda(\boldsymbol{\beta})$ is a μ-amenable regularizer,* $\min\{\tau, 1-\tau\}\lambda_{\min}(\boldsymbol{\Sigma}) > 3\mu/4$ *and the tuning parameters R and λ defined* *in* (2) *are properly chosen such that* $\|\boldsymbol{\beta}^*\|_1 \leq R$ *and* $\lambda \geq c_7 R\sqrt{\ln p/n}$ *for some large positive constant* c_7*, the estimator given by the optimization (2) follows,*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(\sqrt{s}\lambda), \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p(s\lambda), \quad \textit{and} \quad \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/n = O_p(s\lambda^2).$$

We note that when $\boldsymbol{X}_i$ follows a sub-Gaussian distribution, in conjunction with the conclusion of Corollary 1 in Loh and Wainwright (2015), when the sample size satisfies $n \geq c_7(R^2 \vee s)\ln p$, the choice of λ can be changed to $\lambda \asymp \sqrt{\ln p/n}$. This allows us to omit R and streamline the subsequent results.

Next, we discuss the asymptotic property of the de-biased estimator. From (6), we can see that the weighting matrix $\mathbf{W}_{\hat{\boldsymbol{\beta}}}$ is involved in both $\boldsymbol{\Delta}^{(1)}$ and $\boldsymbol{\Delta}^{(2)}$ as well as the construction of $\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}}}$. Thus a key step to derive the asymptotic theory is to show that the weighting matrix $\mathbf{W}_{\hat{\boldsymbol{\beta}}}$ can be replaced by $\mathbf{W}_{\boldsymbol{\beta}^*}$ with little cost. Indeed, we have the following conclusion.

Lemma 1. ***(The bounds for the weighting factor).*** *Under the assumptions of Proposition 1, we have*

$$|w^2_{\boldsymbol{\beta}^*,i} - w^2_{\hat{\boldsymbol{\beta}},i}| \leq \mathbb{I}\left(|\epsilon_i| \leq |\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|\right),$$

holds point-wisely for $i \in \{1, \ldots, n\}$ *and*

$$\frac{1}{n}\sum_{i=1}^{n} |w^2_{\hat{\boldsymbol{\beta}},i} - w^2_{\boldsymbol{\beta}^*,i}| = O_p\left(Ks\lambda\right). \tag{10}$$

Since $\hat{\boldsymbol{\beta}}$ is a consistent estimator, Lemma 1 indicates that the weighting factor $w^2_{\hat{\boldsymbol{\beta}},i}$ is equal to $w^2_{\boldsymbol{\beta}^*,i}$ with high probability. Therefore the average approximation error of the weighting factor converges in probability. This convergence rate differs from that in the GLM scenario, where the second-order derivative of the loss function holds the Lipschitz property, e.g., assumption (C1) in van de Geer et al. (2014) and (L1) in Cai et al. (2023). Additionally, van de Geer et al. (2014) assume in (D4) that $|w_{\hat{\boldsymbol{\beta}},i} - w_{\boldsymbol{\beta}^*,i}| \leq |\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|$ and this directly implies that $\frac{1}{n}\sum_{i=1}^n |w^2_{\hat{\boldsymbol{\beta}},i} - w^2_{\boldsymbol{\beta}^*,i}| \leq \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/n = O_p(s\lambda^2)$, regardless of the condition on the maximum absolute value of the design matrix. Lemma 1 shows that under Assumption 1 in the expectile framework, the convergence rate is slower, which results in slower convergence rate in the subsequent node-wise regressions.

4.2.1. Node-wise regressions for the generalized Hessian matrix

Let's denote the inverse of the population Hessian matrix of the weighted design matrix $\mathbf{X}_{\boldsymbol{\beta}^*}$ by $\boldsymbol{\Theta}_{\boldsymbol{\beta}^*} = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}^*} = (\mathsf{E}[\mathbf{X}^\top_{\boldsymbol{\beta}^*}\mathbf{X}_{\boldsymbol{\beta}^*}]/n)^{-1}$. For each $j \in \{1, \ldots, p\}$, the residual of the node-wise regression is denoted by $\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j} = \boldsymbol{X}_{\boldsymbol{\beta}^*,(j)} - \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}$, where $\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} := \arg\min_{\boldsymbol{\varphi}} \mathsf{E}\|X_{\boldsymbol{\beta}^*,(j)} - \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}\|_2^2$ is the corresponding true coefficient. Moreover, we use $s_j = \|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_0$ to denote sparsity, and $\phi^2_{\boldsymbol{\beta}\cdot j} = \mathsf{E}\|\boldsymbol{\varrho}_{\boldsymbol{\beta}\cdot j}\|_2^2/n$ for population variance. For the sake of simplicity, we take $s^{**} = \max_j s_j$ and $\lambda^{**} = \max_j \lambda_j$ in the subsequent of this paper.

Theorem 1. ***(The uniform bounds for the node-wise estimators).*** *Suppose all the assumptions in Proposition 1 are satisfied. For each $j \in \{1, \ldots, p\}$, if $Q_{\lambda_j}(\boldsymbol{\varphi})$ is $\mu-$amenable and the tuning parameters R_j and λ_j defined in (7) are properly chosen such that $\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 \leq R_j$ and $\lambda_j \geq c_8((\max_j R_j \sqrt{\ln p/n}) \vee (s\sqrt{s^{**}}K^3\lambda))$ for some large positive constant c_8, then the estimator given by (7) satisfies the error bounds:*

$$\max_{1\leq j\leq p} \|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 = O_p(s^{**}\lambda^{**}), \quad \max_{1\leq j\leq p} \|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2 = O_p(\sqrt{s^{**}}\lambda^{**}),$$

and

$$\max_{1\leq j\leq p} \|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})\|_2^2/n = O_p(s^{**}(\lambda^{**})^2).$$

Theorem 2. ***(The uniform bounds for the precision matrix).*** *Suppose all the conditions in Theorem 1 hold. Additionally, if we further assume $\sqrt{s^{**}}\lambda^{**} = o(1)$, then*

$$\max_{1\leq j\leq p} \left|\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j} - \phi^2_{\boldsymbol{\beta}^*,j}\right| = O_p(\sqrt{s^{**}}\lambda^{**}),$$

$$\max_{1\leq j\leq p} \|\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\Theta}_{\boldsymbol{\beta}^*,j}\|_1 = O_p(s^{**}\lambda^{**}), \quad \textit{and} \quad \max_{1\leq j\leq p} \|\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\Theta}_{\boldsymbol{\beta}^*,j}\|_2 = O_p(\sqrt{s^{**}}\lambda^{**}).$$

Remark 4. In the framework of high-dimensional asymptotics, the sample size n and the number of dimension p approach infinity simultaneously. The sparsity s and s_j, for $j \in \{1, \cdots, p\}$, the bound K, and the constraints R and R_j, for $j \in \{1, \cdots, p\}$, may be of $O(1)$ or diverge, as the sample size n and the dimension p grow to infinity.

Theorem 1 and 2 provide uniformly convergence of the node-wise regression estimators and the precision matrix, respectively. When the conditions $K \asymp 1$, $R \asymp 1$ and $s \asymp s^{**} \asymp 1$ are simultaneously satisfied, we can choose $\lambda \asymp \lambda_j \asymp \sqrt{\ln p/n}$, for $j \in \{1, \ldots, p\}$. In such a case, the l_1 and l_2 bounds for both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}$ have reached their optimal bounds under the regularized framework, see Cai and Guo (2017), Sun and Zhang (2012) and Verzelen (2012). Compared to the results in Theorem 3.2 of van de Geer et al. (2014) which choose $\lambda_j \asymp K\sqrt{\ln p/n}$, we may choose $\lambda_j \asymp RR_js\sqrt{s^{**}}K^3\sqrt{\ln p/n}$. The primary cause of this difference is the second-order non-Lipschitz property of the expectile loss function. Moreover, we do not impose the assumption that the design matrix $\mathbf{X}$ has i.i.d sub-Gaussian rows as Assumption 2.1 in Zhang and Cheng (2017), nor do we assume any extra condition like $\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\varphi_{\boldsymbol{\beta}^*,j}\|_\infty = O(K)$ in (D1) of van de Geer et al. (2014). Additionally, the use of the non-convex regularizer and the uniform rate under consideration also contribute to this disparity. For the same reason, the convergence rate is slower than that in Thoerem 3.2 of van de Geer et al. (2014).

4.3. Main results

We establish the probabilistic upper bound of $\boldsymbol{\Delta}^{(1)}$ and $\boldsymbol{\Delta}^{(2)}$ as defined in (6) and the asymptotic normality of the de-bias estimator under the expectile framework in the following two theorems, respectively.

Theorem 3. ***($\sqrt{\mathbf{n}}$ negligiblity of $\boldsymbol{\Delta}^{(1)}$ and $\boldsymbol{\Delta}^{(2)}$).*** *Suppose that all the conditions in Theorem 1 hold and additionally we assume that $s\sqrt{s^{**}}\lambda\lambda^{**} = o(n^{-1/2})$, then*

$$\|\boldsymbol{\Delta}^{(1)}\|_\infty = o_p(n^{-1/2}), \quad \textit{and} \quad \|\boldsymbol{\Delta}^{(2)}\|_\infty = o_p(n^{-1/2}).$$

Theorem 4. ***(Asymptotic normality for the de-biased estimator).*** *Suppose that all the conditions in Theorem 3 hold, then for the p_0-by-p hypothesis matrix $\mathbf{H}$ defined in (3) with $\|\mathbf{H}\|_{l_\infty} = O(1)$ and $\mathbf{H}\boldsymbol{\Omega}\mathbf{H}^\top \to \boldsymbol{\Upsilon}$, as $p \to \infty$, where $\boldsymbol{\Omega}$ is defined in (9), p_0 is a fixed integer and $\boldsymbol{\Upsilon}$ is a full rank p_0-by-p_0 matrix, we have*

$$\sqrt{n}\mathbf{H}\left(\hat{\boldsymbol{\beta}}_{de} - \boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}_{p_0}(\mathbf{0}, \boldsymbol{\Upsilon}), \quad \textit{as } n, p \to \infty.$$

Furthermore, if $\mathsf{E}[\epsilon_i^8] < c_3$, $\max_j \mathsf{E}[x_{ij}^8] < c_1$, $K^4\sqrt{\ln p/n} = o(1)$, and $(s^{**})^{3/2}\lambda^{**} = o(1)$, then

$$\|\hat{\boldsymbol{\Omega}}_{\mathbf{H}} - \boldsymbol{\Upsilon}\|_\infty = o_p(1),$$

where $\hat{\boldsymbol{\Omega}}_{\mathbf{H}}$ is defined in (8). Consequently, we have

$$(\mathbf{H}(\hat{\boldsymbol{\beta}}_{de} - \boldsymbol{\beta}^*))^\top \hat{\boldsymbol{\Omega}}_{\mathbf{H}}^{-1} \mathbf{H}(\hat{\boldsymbol{\beta}}_{de} - \boldsymbol{\beta}^*) \xrightarrow{d} \chi^2(p_0), \quad \textit{as } n, p \to \infty.$$

Compared to the Theorem 2.4 in van de Geer et al. (2014), the additional 8th moment conditions on ϵ_i and x_{ij} are essential to guarantee the consistency of $\hat{\boldsymbol{\Omega}}_{\mathbf{H}}$. The assumption $K^4\sqrt{\ln p/n} \to 0$ is valid even when X_{ij} and ϵ_i do not necessarily follow sub-Gaussian or strongly bounded distributions. For instance, when s and s^{**} are fixed positive integers, $p \asymp n^\eta$, and X_{ij} and ϵ_i have bounded $(8+\delta)$th moment for some $\delta > 8\eta$, it is easy to check that $K^4\sqrt{\ln p/n} \to 0$ holds.

Although various regularizers within the amenable category yield the same asymptotic results in terms of the l_1 and l_2 bounds when estimating corresponding coefficients, non-convex regularizers offer distinct advantages in variable selection. This is further elaborated in Loh and Wainwright (2015), Zhao et al. (2018), among others. By reducing falsely selected coordinates in both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varphi}}^2_{\boldsymbol{\beta},j}$ in the estimation of the pseudo-inverse of the sample covariance matrix, our proposed test with non-convex regularizers may achieve extra efficacy. We show in the simulation that the test constructed using the SCAD regularizer performs better than that constructed using the Lasso regularizer in most situations, especially for the heteroscedasticity case.

5. Simulation study

In this section, we conduct simulation studies to evaluate the finite sample performance of the proposed Wald-type test under the expectile framework. Our primary focus is on assessing its ability to control Type I error and analyzing its local power in various scenarios. Meanwhile, recall that there are totally two regularizers, $P_\lambda(\cdot)$ and $Q_{\lambda_j}(\cdot)$, separately used in our testing procedure. We also exhibit a strong interest in the performance of test statistics derived from different regularizers within the amenable category. To illustrate this, we exemplify the convex regularizer with the Lasso and the non-convex regularizer with the SCAD and consider the following two methods:

1) Convex regularizer method, where both $P_\lambda(\cdot)$ and $Q_{\lambda_j}(\cdot)$ are chosen to be the Lasso, denoted as 'Lasso-Lasso'.

2) Non-convex regularizer method, where both $P_\lambda(\cdot)$ and $Q_{\lambda_j}(\cdot)$ are chosen to be the SCAD, denoted as 'SCAD-SCAD'.

Given that expectile regression (2) can be applied to detect heteroscedasticity in high-dimensional data due to its asymmetric weighting on squared loss, we consider both the homoscedastic models studied in van de Geer et al. (2014) and heteroscedastic models discussed in Gu and Zou (2016), Wang et al. (2012), and Zhao et al. (2018). The entire simulation study is conducted using MATLAB 2022b, with each simulation procedure repeated 1,000 times. The tuning parameters λ and λ_j are selected via the $10-$fold cross-validation. In our simulation, the influence of different choices of sufficiently large values of R and R_j on the results is negligible. To balance computational costs, we chose to set both R and R_j to infinity without further tuning. Additionally, we set $\gamma = 3.7$ for the SCAD regularizer, as suggested in Fan and Li (2001).

5.1. Simulation results under the homoscedastic case

The response variable is generated from the following linear expectile model:

$$y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^* + (\epsilon_i - \mathsf{E}\rho_\tau(\epsilon_i)),$$

for $i \in \{1, \ldots, n\}$, where the error term $\epsilon_i - \mathsf{E}\rho_\tau(\epsilon_i)$ ensures that the identification condition is satisfied.

To generate the covariates $\boldsymbol{X}_i$, we draw them independently from the multivariate normal distribution $\mathcal{N}_p(0, \boldsymbol{\Sigma})$. We specify two distinct covariance matrices as the design choices, which are studied by van de Geer et al. (2014) and Liu and Wang (2017), respectively:

$$\begin{aligned} &\text{Toeplitz:} \quad \boldsymbol{\Sigma} = (\Sigma_{jk}) \text{ with } \Sigma_{jk} = \xi^{|j-k|}, \\ &\text{Scale-free graph:} \quad \boldsymbol{\Sigma} = \left[\mathbf{D}\left[\mathbf{A} + (|\lambda_{min}(\mathbf{A})| + 0.2)\mathbf{I}\right]\mathbf{D}\right]^{-1}, \end{aligned}$$

where $\mathbf{A} \in \mathbb{R}^{p\times p}$ is an adjacency matrix associated with a certain graph. In this matrix, the nonzero off-diagonal elements A_{jk}, $|j-k| \le \varsigma$ are set to 0.3, and the diagonal elements are set to 0. Moreover, $\mathbf{D} \in \mathbb{R}^{p\times p}$ is a diagonal matrix where $D_{jj} = 1$ for $j \in \{1, \ldots, p/2\}$ and $D_{jj} = 3$ for $j \in \{p/2+1, \ldots, p\}$. Notably, the Toeplitz covariance matrix leads to tridiagonal precision inverse matrix, while the sparsity of the inverse matrix of the Scale-free graph corr depends on the adjacency relationships recorded by $\mathbf{A}$. For the parameters in different choices of the covariance matrix, we consider $\xi = 0.25, 0.5$ and 0.75 for the Toeplitz while $\varsigma = 10, 20$, and p for the Scale-free graph case.

To generate $\boldsymbol{\beta}^*$, we let $\beta_1^* = k/\sqrt{n}$, $k \in \{0, \ldots, 6\}$ in different settings. For the rest of the coefficients, we consider two scenarios:

1) The Dirac measure case: $\beta_i^* = 1$ if $i \in \mathcal{K}$ and $\beta_i^* = 0$ if $i \notin \mathcal{K} \cup \{1\}$.

2) The Uniform random measure case: $\beta_i^* \sim \mathcal{U}(0, 2)$ if $i \in \mathcal{K}$ and $\beta_i^* = 0$ if $i \notin \mathcal{K} \cup \{1\}$.

where $\mathcal{K} = \mathcal{K}^4$ or $\mathcal{K}^{10}$, with

$$\mathcal{K}^4 \in \{6, 12, 15, 20\}, \quad \text{and} \quad \mathcal{K}^{10} \in \{5, 6, 7, 8, 9, 10, 11, 12, 15, 20\}.$$

We denote the two scenarios as Dirac 4 (Dirac 10) and Unif 4 (Unif 10), respectively.

To generate ϵ_i, two distributions are taken into consideration:

1) The standard normal distribution $\mathcal{N}(0, 1)$,

2) The student-t distribution with 4 degrees of freedom, denoted as t_4.

The t_4 distribution is a heavy-tailed distribution, which helps to evaluate the performance of our proposed test under heavy-tailed scenario.

The sample size is set as $n = 300$, the dimension of the covariates is set as $p = 200, 400$ and 600 and the expectile level is set as $\tau = 0.1, 0.5$ and 0.9. Note that when the expectile level $\tau = 0.5$, the asymmetric squared loss becomes exactly equivalent to the squared loss. The corresponding de-biased estimator then aligns with that under the regularized least square framework, as studied in van de Geer et al. (2014); Javanmard and Montanari (2014) and Zhang and Zhang (2014).

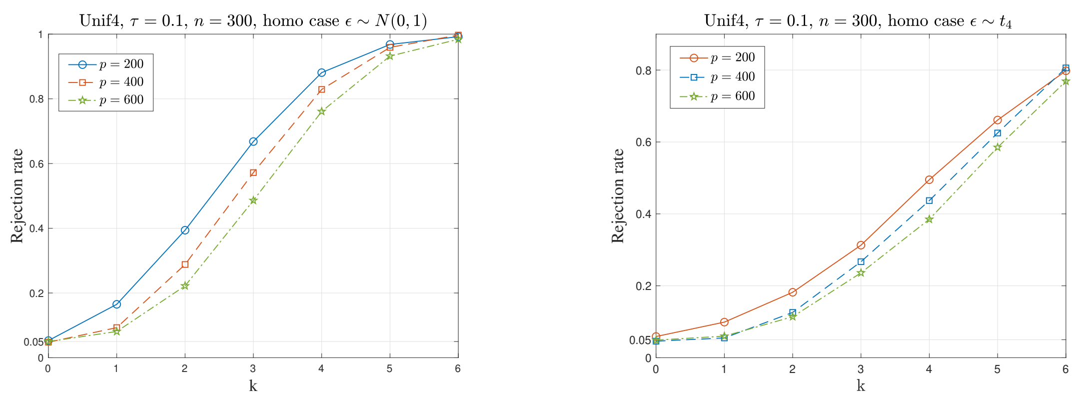

Figure 1: The empirical power and Type I error of the proposed de-biased test for the expectile regression with $\tau = 0.1$ and $\beta_1^* = k/\sqrt{n}$ under the Unif4, homoscedastic case with $n = 300$ and $p \in \{200, 400, 600\}$, calculated from 1000 replicates. The null hypothesis is $\mathsf{H}_0 : \beta_1^* = 0$ and the significant level is set as 5%.

To assess the ability of the test statistic in controlling the Type I error and local power in homoscedastic scenarios, we firstly conduct a hypothesis test on a single coefficient at a given expectile level τ,

$$\mathsf{H}_{\tau,0} : \beta_1^* = 0 \quad \text{versus} \quad \mathsf{H}_{\tau,1} : \beta_1^* \neq 0. \tag{11}$$

Table 1 shows that test statistics derived from different regularizers exhibit similar efficacy in controlling Type I error and local power under the Dirac 4 scenario with diverse Toeplitz designs. However, the use of the non-convex SCAD regularizer incurs a noticeably higher computational burden, even when implementing the LLA strategy by Zou and Li (2008), than the Lasso method. Figure 1 shows that as the dimension grows, the empirical power of the proposed test gradually diminishes. Furthermore, in scenarios where errors follow

Table 1: The empirical Type I error and power obtained by different regularizers under the homoscedastic case with $n = 300, p = 400$, Dirac 4, $\tau = 0.1$, standard normal error and Toeplitz design from 1000 replicates, and the average computing time for each repetition. The CPU time records the average computation time for each repetition under different methods. The null hypothesis is $\mathsf{H}_0 : \beta_1^* = 0$ and the significant level is set as 5%.

Toeplitz	$\epsilon \sim \mathcal{N}(0,1)$	$\beta_1^* = k/\sqrt{n}$							
Method		$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	CPU time (s)
Lasso-Lasso	$\xi = 0.25$	5.90%	13.50%	35.80%	65.40%	87.10%	96.80%	99.50%	1.056
	$\xi = 0.50$	5.20%	11.80%	35.00%	64.20%	85.20%	95.60%	99.10%	
	$\xi = 0.75$	4.50%	11.90%	28.10%	52.30%	74.20%	87.50%	95.70%	
SCAD-SCAD	$\xi = 0.25$	4.80%	12.30%	35.30%	68.80%	89.00%	95.30%	99.60%	4.252
	$\xi = 0.50$	4.60%	12.50%	33.10%	67.20%	86.70%	94.10%	99.20%	
	$\xi = 0.75$	4.90%	10.80%	26.00%	59.30%	88.50%	89.10%	95.20%	

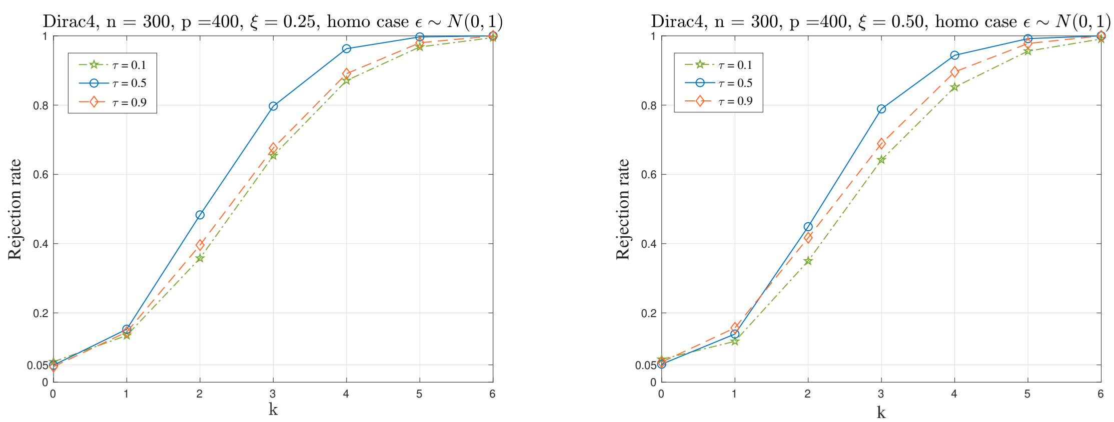

Figure 2: The empirical power and Type I error of the proposed de-biased test for the expectile regression under the Dirac 4, homoscedastic case with $n = 300$, $\beta_1^* = k/\sqrt{n}$, $\tau \in \{0.1, 0.5, 0.9\}$ and Toeplitz design calculated from 1000 replicates. The null hypothesis is $\mathsf{H}_0 : \beta_1^* = 0$ and the significant level is set as 5%.

a t-distribution, the test's empirical power is relatively low compared to scenarios with normally distributed errors. Figure 2 demonstrates that the rejection rate of our proposed test is comparable at $\tau = 0.1$ and $\tau = 0.9$. Notably, the test exhibits marginally higher empirical power at $\tau = 0.5$, which corresponds to the test in van de Geer et al. (2014), in comparison to the other two expectile levels.

Following the results presented in Table 2, we observe that the two methods exhibit similar test power at $\varsigma = 10$. Notably, as ς increases from 20 to p, the 'SCAD-SCAD' method outperforms the 'Lasso-Lasso' method in controlling the empirical Type I error along with local power. This suggests that the non-convex penalizer might be more effective in scenarios where the inverse of the covariance matrix is not particularly sparse.

Next we consider a group test on $\boldsymbol{\beta}^*_{\mathcal{G}}$, specifically,

$$\mathsf{H}_0 : \boldsymbol{\beta}^*_{\mathcal{G}} = \mathbf{0} \quad \text{versus} \quad \mathsf{H}_1 : \boldsymbol{\beta}^*_{\mathcal{G}} \neq \mathbf{0}, \tag{12}$$

where $\mathcal{G} = \{1, 3, 4\}$. Analogously, we set the value of β^*_1 varies from $0/\sqrt{n}$ to $6/\sqrt{n}$ while we should note that β^*_3 and β^*_4 are consistently zero. Table 3 reports the empirical Type I error and power for multiple tests on group $\mathcal{G}$ across various settings on the true coefficients, which is consistent with the result on the test of single component β^*_1 under various coefficient setups in our simulation study.

Table 2: The empirical Type I error and power obtained by different regularizers under the Dirac 4, homoscedastic case with $n = 300, p = 400$, $\tau = 0.1$, standard normal error and Scale-free design from 1000 replicates. The null hypothesis is $\mathsf{H}_0 : \beta^*_1 = 0$ and the significant level is set as 5%.

Scale-free	$\epsilon \sim \mathcal{N}(0,1)$	$\beta^*_1 = k/\sqrt{n}$						
Method		$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Lasso-Lasso	$\varsigma = 10$	5.10%	7.10%	12.10%	30.10%	49.60%	72.10%	86.70%
	$\varsigma = 20$	5.30%	6.90%	7.70%	16.50%	31.70%	61.30%	76.70%
	$\varsigma = 400$	7.50%	47.50%	93.80%	99.90%	100.00%	100.00%	100.00%
SCAD-SCAD	$\varsigma = 10$	4.70%	7.20%	19.70%	41.30%	64.00%	88.30%	92.30%
	$\varsigma = 20$	5.00%	7.30%	16.10%	29.40%	46.60%	66.80%	81.20%
	$\varsigma = 400$	4.80%	45.10%	86.20%	99.30%	100.00%	100.00%	100.00%

5.2. Simulation results under the heteroscedastic case

The response variable is generated from the following sparse linear model:

$$y = x_6 + x_{12} + x_{15} + x_{20} + 0.7\Phi(x_1)\epsilon, \tag{13}$$

where the covariate $(x_1, \cdots, x_p)^\top$ and the error term ϵ are generated as in the homoscedastic case. Additionally, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Note further that the covariate x_1 plays an essential role in revealing the potential heteroscedasticity, since the magnitude of

Table 3: The empirical Type I error and power for group $\mathcal{G}$ under the homoscedastic case with $n = 300, p = 400$, $\xi = 0.50$, $\tau = 0.1$, standard normal error and Toeplitz design from 1000 replicates. The null hypothesis is $\mathsf{H}_0 : \beta_1^* = \beta_3^* = \beta_4^* = 0$ and the significant level is set as 5%.

Toeplitz	$\beta_1^* = k/\sqrt{n},\ \beta_3^* = \beta_4^* = 0,\quad \mathsf{H}_0 : \beta_1^* = \beta_3^* = \beta_4^* = 0.$						
	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Dirac4	5.30%	14.70%	30.90%	57.40%	80.80%	93.50%	98.90%
Dirac10	6.30%	14.00%	30.30%	55.10%	77.00%	91.50%	97.60%
Unif4	4.70%	13.20%	32.60%	59.40%	81.70%	95.10%	99.00%
Unif10	6.50%	15.70%	31.70%	56.20%	78.50%	92.00%	99.30%

the pseudo-true coefficient for x_1 by the means of expectation under expectile framework is strictly distinct from 0 for any $\tau \neq 0.5$. Thus, many studies associate the presence of heteroscedasticity with a non-zero estimate of the coefficient corresponding to x_1, see Gu and Zou (2016) and Zhao et al. (2018).

To evaluate the capability of our proposed test in detecting heteroscedasticity, we continue to employ the test (11) for single coordinate to obtain the rejection rates under a given statistical significance level α at different expectile levels. Furthermore, we are intrigued by the significance of covariates that exhibit high correlation with x_1 and, as a result, may be influenced by such heteroscedasticity. Consequently, we also conduct an additional test to further investigate the performance of our proposed test in controlling the Type I error and local power,

$$\mathsf{H}_{\tau,0} : \beta_2^* = 0 \quad \text{and} \quad \mathsf{H}_{\tau,1} : \beta_2^* \neq 0,$$

where β_2^* is the coefficient of x_2 in the heteroscedastic model. To examine the local power of the additional test, we set $\beta_2^* = k/\sqrt{n}$, for $k \in \{0, ..., 6\}$ in data generating process (13).

Table 4: The frequency that x_1 is selected by the expectile-based estimator deduced by Lasso and SCAD under the heteroscedastic case with $n = 300$, $\xi = 0.5$, Toeplitz design, different expectile level τ and error distribution from 1000 replicates.

Toeplitz	Lasso				SCAD			
Dirac4	$\epsilon \sim \mathcal{N}(0,1)$		$\epsilon \sim t_4$		$\epsilon \sim \mathcal{N}(0,1)$		$\epsilon \sim t_4$	
	$p=400$	$p=600$	$p=400$	$p=600$	$p=400$	$p=600$	$p=400$	$p=600$
$\tau = 0.1$	85.60%	83.70%	83.10%	68.80%	86.10%	79.50%	74.60%	69.10%
$\tau = 0.5$	5.70%	1.60%	3.30%	1.80%	0.00%	0.00%	0.00%	0.00%
$\tau = 0.9$	91.00%	78.60%	88.90%	69.20%	93.80%	87.30%	78.60%	72.20%

Table 4 presents the performance of the expectile-based estimators deduced by Lasso and the SCAD in detecting the heterogeneity, as studied in Zhao et al. (2018), denoted as 'Lasso' and 'SCAD', respectively. In the case of $\tau = 0.5$, the 'Lasso' tends to select x_1 more frequently, while the 'SCAD' is less inclined to select x_1. Table 5 records the empirical rejection rate of (13) deduced by our proposed test statistic under

Table 5: The empirical rejection rate of the test (11) deduced by our proposed test statistic with the 'Lasso-Lasso' method or the 'SCAD-SCAD' method under the heteroscedastic case (13) with $n = 300$, $\xi = 0.5$, Toeplitz design, different expectile level τ and error distribution from 1000 replicates. The null hypothesis is $\mathsf{H}_0 : \beta_1^* = 0$ and the significant level is set as 5%.

Toeplitz	Lasso-Lasso				SCAD-SCAD			
Dirac4	$\epsilon \sim \mathcal{N}(0,1)$		$\epsilon \sim t_4$		$\epsilon \sim \mathcal{N}(0,1)$		$\epsilon \sim t_4$	
	$p = 400$	$p = 600$	$p = 400$	$p = 600$	$p = 400$	$p = 600$	$p = 400$	$p = 600$
$\tau = 0.1$	90.50%	89.70%	82.40%	74.90%	98.20%	96.60%	98.30%	96.10%
$\tau = 0.5$	5.40%	4.80%	5.00%	4.70%	5.10%	5.00%	4.70%	4.70%
$\tau = 0.9$	89.70%	87.70%	79.70%	80.70%	95.10%	94.40%	93.20%	96.50%

the null hypothesis at 5% significant level. For the case where $\tau = 0.5$, the heterogeneity of the model (13) has little impact on the coefficient of x_1 and our proposed test effectively controls the empirical Type I error across various combinations of the dimension p and error distribution. Meanwhile, for the case where $\tau = 0.1$ and $\tau = 0.9$, the empirical rejection rate of our proposed test is relatively larger than the significant level. This observation demonstrates that our test can effectively identify the heteroscedasticity of the model. Furthermore, under the Dirac 4 Toeplitz design, the test statistic derived using the 'SCAD-SCAD' method exhibits a relatively superior capability in detecting heteroscedasticity compared to that derived from the 'Lasso-Lasso' method.

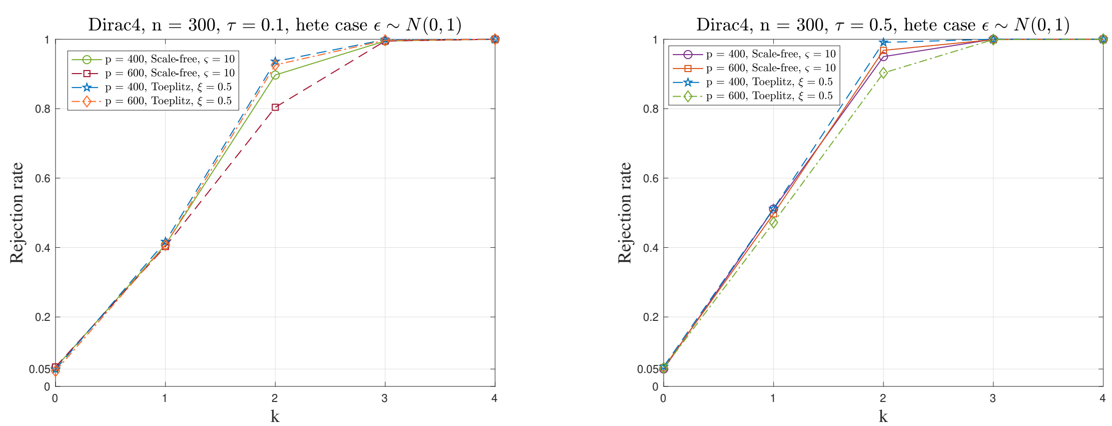

Figure 3: The rejection rate of the proposed de-biased test for the expectile regression with different τ under the heteroscedasticity case (13) with $n = 300$ and $p = 400, 600$, $\beta_2^* = k/\sqrt{n}$, and standard normal error, calculated from 1000 replicates. The null hypothesis is $\mathsf{H}_0 : \beta_2^* = 0$ and the significant level is set as 5%.

Figure 3 depicts the empirical Type I error associated with the coefficient of x_2 across various expectile levels τ, as p increases from 400 to 600. Our proposed test effectively controls the empirical Type I error at both $\tau = 0.1$ and $\tau = 0.5$, across different dimensions p and design setups, demonstrating the test's viability in heteroscedastic scenarios. Furthermore, the empirical power reaches 1 when $\beta_2^* = 4/\sqrt{n}$ for all scenarios depicted in Figure 3, which indicates that our proposed test statistic possesses considerable ability to detect potential active covariates in cases of heteroscedasticity. Notably, in comparison with the homoscedastic case, the trend in local power becomes more pronounced, suggesting that heteroscedasticity significantly impacts the inference for some coordinate of the covariate closely associated with x_1.

Lastly, we consider group tests (12) with $\mathcal{G}_1 = \{1,2,3\}$ and $\mathcal{G}_2 = \{2,3,4\}$, where group $\mathcal{G}_1$ includes the critical variable x_1, while group $\mathcal{G}_2$ encompasses three variables predominantly affected by heteroscedasticity. Moreover, we employ the Dirac 4 pattern to generate the true coefficients. In contrast to experimental designs with homoscedastic group tests, due to the nonlinear heteroscedastic structure of x_1, we only consider altering different dimensions p, expectile levels τ and types of errors to observe their rejection rates in group tests during simulations. In Table 6, the rejection rates for both group tests, although lower than those of the previously presented univariate empirical Type I errors, are comparatively close. This demonstrates that our proposed multivariate testing approach maintains commendable performance under the heteroscedastic settings.

Table 6: The rejection rate calculated by different method, dimension p, expectile level τ and error type for group $\mathcal{G}_1$ and $\mathcal{G}_2$ under the heteroscedastic case (13) with $n = 300$, $\xi = 0.50$ and Toeplitz design from 1000 replicates. The null hypothesis is $\mathsf{H}_0 : \boldsymbol{\beta}^*_{\mathcal{G}} = 0$ and the significant level is set as 5%.

		Lasso-Lasso				SCAD-SCAD			
		$\epsilon \sim \mathcal{N}(0,1)$		$\epsilon \sim t_4$		$\epsilon \sim \mathcal{N}(0,1)$		$\epsilon \sim t_4$	
group	τ value	$p=400$	$p=600$	$p=400$	$p=600$	$p=400$	$p=600$	$p=400$	$p=600$
$\mathcal{G}_1 = \{1,2,3\}$	$\tau = 0.1$	75.80%	73.40%	65.60%	56.40%	95.50%	94.10%	87.70%	83.50%
	$\tau = 0.5$	4.80%	5.50%	4.60%	4.20%	4.60%	4.50%	4.40%	5.00%
	$\tau = 0.9$	78.90%	72.90%	63.70%	61.30%	92.60%	87.70%	86.40%	84.50%
$\mathcal{G}_2 = \{2,3,4\}$	$\tau = 0.1$	4.60%	4.20%	4.30%	4.50%	4.50%	4.40%	3.90%	5.70%
	$\tau = 0.5$	5.40%	5.40%	4.80%	4.70%	4.10%	4.90%	4.70%	4.30%
	$\tau = 0.9$	5.20%	6.20%	4.70%	6.10%	4.50%	4.70%	4.90%	4.50%

6. Real data analysis

6.1. Financial and macro data

In this empirical study, we examine the relationship between monetary policy measures and stock market returns using the FRED-MD dataset, available on the Fred-MD website [1]. The dataset comprises 127 U.S. macroeconomic variables observed monthly from January 1959 to September 2023. These macroeconoic variables are classified into eight groups: consumption, orders and inventories; housing; interest and exchange rates; labour market; money and credit; output and income; prices; and the stock market. More detailed

[1] https://research.stlouisfed.org/econ/mccracken/fred-databases/

description can be found in McCracken and Ng (2016). We designate the monthly returns of S&P 500 as the dependent variable and employ the other 126 variables, lagged by 1, 2, and 3 months, as explanatory variables. Notably, we focus on the Monetary Base (BOGMBASE) and M1 Money Stock (M1SL) as key variables of interest.

Table 7 presents the results of the de-biased estimation and inference across expectile levels $\tau = 0.1, 0.3, 0.5, 0.7,$ and 0.9. The estimates of regression coefficients associated with BOGMBASE and M1SL are reported alongside their standard deviations. Additionally, the table displays the p-values for separate group tests on BOGMBASE and M1SL with lags 1, 2, and 3 using our proposed test statistics. Our results indicate that the de-biased estimates of $\mathrm{M1SL}_{t-1}$ and $\mathrm{M1SL}_{t-3}$ are statistically significant at 10% significance level under the $\tau = 0.9$ case while the de-biased estimates of the regression coefficients related to M1SL are not significant under other expectile levels. Moreover, the group test on the M1SL is statistically significant at 5% significance level under the $\tau = 0.9$ case, suggesting that M1SL can be used to predict the upper expectiles of the distribution of S&P 500 returns. Furthermore, the differences in the results of the de-biased estimates at given expectile levels, along with the disparity in the significance of group tests for variables BOGMBASE and M1SL, suggest heteroscedasticity in the regression model. This finding is consistent with the results reported by Taamouti (2015), who demonstrated that while money supply appears to have no impact on stock returns using a nonparametric Granger causality test in mean, employing a quantile regression-based test reveals a statistically significant influence.

6.2. Gene data

In this section, we utilize a microarray dataset to demonstrate the effectiveness of our proposed de-biased expectile test statistic. The dataset comprises microarray gene expression profiling of peripheral blood from 119 healthy women in a multi-ethnic study of atherosclerosis cohort aged 50 or above conducted by Huang et al. (2011). The dataset is publicly available at the NCBI Gene Expression Omnibus data repository [2] under accession number GSE20129.

Huang et al. (2011) investigate the impact of the innate immune system on the development of atherosclerosis by analyzing gene profiles from 119 patients' peripheral blood. Their case study identifies the toll-like receptors (TLRs) signaling pathway as crucial in activating the innate immune response in atherosclerosis cases. Particularly, the *TLR8* gene is highlighted as a key gene associated with atherosclerosis. To further study the relation between this key gene and the other genes, Fan et al. (2017) regress the expression of the *TLR8* gene on the expressions of another 464 genes from 12 different related-pathways (TLR, CCC, CIR, IFNG, MAPK, RAPO, EXAPO, INAPO, DRS, NOD, EPO, CTR). They find that their methods are able to select more genes than Lasso, potentially valuable for subsequent confirmatory studies. Furthermore, their results reveal that the regression residuals have heavy right tail and skewed distribution. Zhao et al. (2018) further demonstrate that the skewness and kurtosis of the expression of the *TLR8* gene deviates from those

[2] http://www.ncbi.nlm.nih.gov/geo/

Table 7: High-dimensional expectile regression results of S&P 500 on the other 126 variables with lags 1–3 as explanatory variables. The estimates of regression coefficients associated with BOGMBASE and M1SL are reported alongside their standard errors. "***", "**", "*" indicate the statistical significance at 1%, 5%, and 10% levels, respectively.

	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
M1SL_{t-1}	-0.021	-0.023	-0.032	-0.049	-0.072*
	(0.1065)	(0.0742)	(0.0599)	(0.0511)	(0.0423)
M1SL_{t-2}	0.105	0.066	0.056	0.030	0.002
	(0.1085)	(0.0709)	(0.0557)	(0.0468)	(0.0399)
M1SL_{t-3}	-0.135	-0.070	-0.065	-0.074	-0.084*
	(0.0870)	(0.0538)	(0.0504)	(0.0484)	(0.0459)
BOGMBASE_{t-1}	-0.022	0.018	0.036	0.043	0.048
	(0.1288)	(0.0919)	(0.0768)	(0.0638)	(0.0568)
BOGMBASE_{t-2}	0.058	0.028	0.013	0.008	0.019
	(0.0742)	(0.0600)	(0.0540)	(0.0482)	(0.0417)
BOGMBASE_{t-3}	-0.090	-0.040	-0.023	-0.010	-0.021
	(0.0733)	(0.0510)	(0.0436)	(0.0430)	(0.0389)
H_0: $\beta_{M1SL_{t-1}} = \beta_{M1SL_{t-2}} = \beta_{M1SL_{t-3}} = 0$					
p-value	0.444	0.479	0.350	0.188	0.047
H_0: $\beta_{BOGMBASE_{t-1}} = \beta_{BOGMBASE_{t-2}} = \beta_{BOGMBASE_{t-3}} = 0$					
p-value	0.625	0.837	0.889	0.897	0.752

of a normal distribution. In addition, they show the heteroscedastic feature of the data by implementing penalized expectile regressions with different expectile levels. Motivated by this, we apply our proposed de-biased expectile test to the expectile regression model where the response variable is the expression of the *TLR8* gene and the explanatory variables are the expressions of another 464 genes with expectile levels $\tau = 0.1, 0.3, 0.5, 0.7$, and 0.9.

Table 8: Significant genes in the expectile regression at the significance level $\alpha = 0.05$, and their corresponding coefficients estimated using the de-biasing method, obtained at expectile levels $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The genes found in Fan et al. (2017) are in boldface.

$\tau = 0.1$		$\tau = 0.3$		$\tau = 0.5$		$\tau = 0.7$		$\tau = 0.9$	
IFI6	0.142	***IFI6***	0.148	***MAPK1***	0.299	*IFNA2*	-0.162	*IL1A*	-0.314
RASGRP1	0.212	***MAPK1***	0.265	*MAPK14*	0.241	*MAPK10*	-0.101	***CSF3***	-0.227
MAPK1	0.171	*MAPK14*	0.211	***PSMB8***	0.141	*MAPK7*	-0.180	*IFNGR2*	-0.206
PRKCH	-0.164	***PSMB8***	0.146	***KPNB1***	0.198	***MAPK1***	0.209	*IFNA2*	-0.180
DUSP1	0.136	***KPNB1***	0.178	***BCL2L11***	-0.184	*PSMA5*	-0.198	***AKT3***	-0.172
AKT1	0.231	*CASP6*	-0.135	*CFLAR*	0.365	*SPTAN1*	-0.232	*MAP3K14*	-0.393
TLR3	0.121	***BCL2L11***	-0.208	***CRK***	0.250	***KPNB1***	0.182	*MAPK7*	-0.202
DAPK2	0.129	*CFLAR*	0.277			*CFLAR*	0.421	*NRAS*	-0.345
PSMD4	-0.176	***CRK***	0.203			*CBL*	-0.174	*MAP3K11*	-0.133
PSMB8	0.152					*PTK2B*	-0.097	*MAP3K4*	-0.264
STK24	0.231					***CRK***	0.370	*AKT2*	-0.199
PRKCQ	-0.130							*PSMD6*	-0.244
KPNB1	0.216							*UBA52*	-0.181
HIST1H1D	-0.165							*PSMC1*	-0.149
CASP6	-0.186							*VIM*	-0.178
BCL2L11	-0.238							*CFLAR*	0.576
AIM2	0.202							*NOD2*	-0.230
ZAP70	-0.113							*CBL*	-0.220
								IL17A	-0.188
								CRK	0.535
								CDC42	-0.188
								CD247	-0.239

Table 8 reports the significant genes along with their de-biased estimates. For clarity, we bold the overlapping genes found in Fan et al. (2017). In the $\tau = 0.5$ case, where the expectile regression is equivalent to the mean regression, there are a total of seven genes significant at $\alpha = 5\%$. Our approach identifies six additional genes compared with the sole gene *CRK* obtained using the Lasso estimation in Fan et al. (2017). Furthermore, the significant genes selected by our proposed de-biasing method at different expectile levels include a subset of genes identified by the adaptively weighted Lasso (R-Lasso) method in Fan et al.

(2014) and the regularized approximate quadratic estimator with an L_1 regularizer (RA-Lasso) in Fan et al. (2017). Notably, at $\tau = 0.1$ and $\tau = 0.9$, the significant genes identified overlap with those selected at $\tau = 0.5$, aligning with the findings in Zhao et al. (2018). Moreover, both the significant genes and their corresponding de-biased estimates vary with changing expectile levels, indicating the presence of heteroscedasticity in this data.

In summary, examining the heteroscedastic pattern through expectile linear regression at various expectile levels, along with the identification of significant covariates by our proposed test, may provide new insights into depicting the relationship between the target gene *TLR8* and other related genes.

7. Conclusion

In this article, we develop a bias correction procedure to conduct statistical inference for high-dimensional sparse linear models under the expectile settings. Due to the second-order non-Lipschitz property of the expectile loss function, our approach deviates from that of van de Geer et al. (2014). An alternative proving strategy is developed to demonstrate that the estimation errors in the preliminary estimator have little impact on the estimation of the expectile-specified random weights, which is pivotal in revealing the asymptotic normality of the de-biased estimator. A Wald-type test statistic is then established for multivariate testing. Simulation results indicate that the test we proposed performs well in controlling the Type I error rate and demonstrates good local power under both the homoscedastic and heteroscedastic cases, even with heavy-tailed errors. Furthermore, compared with the method studied in Zhao et al. (2018), The simulation demonstrates that our proposed test statistic possesses ability to detect heteroscedasticity based on different expectile levels.

The empirical study on the FRED-MD dataset demonstrates that under the expectile settings, monetary supply may possess a certain predictive ability for stock returns, aligning with the conclusions in Taamouti (2015). Meanwhile, the empirical results on the selected gene data reveal a list of genes that exhibit significant explanatory ability for the expression of the *TLR8* gene, which, along with the corresponding de-biased estimates, vary across different expectile levels. Such findings not only highlight the presence of heteroscedasticity within the data, as discussed in Zhao et al. (2018), but also provide a novel perspective on the relationships between the *TLR8* gene and other genes, which consists a subset of genes identified in Fan et al. (2017).

It is important to note that our results are applicable to a broader regularization framework, where the regularizers can be either convex or non-convex, provided they belong to the amenable category. While the use of non-convex regularizers introduces complexity to the optimization problem, their superior performance in variable selection may allow for more effective control over Type I errors and yields higher local power compared to convex regularizers. This assertion is further corroborated by our simulation results, particularly in scenarios with less sparse Hessian matrices. However, when addressing hypothesis testing in general sparse scenarios, even though the results using convex and non-convex regularizers are similar, it is crucial to note the considerably heavy computational burden brought about by the use of non-convex regularizers.

Future research directions involve extending our theoretical results to facilitate simultaneous inference for high-dimensional components of a large parameter vector in sparse expectile linear models, in which the dimension of the parameter vector of interest is allowed to grow with the sample size. Moreover, we acknowledge that our assumption regarding the existence of the 8th moment of the observations can be overly restrictive for certain datasets. Therefore, there is a need for the development of more robust expectile regression methods and Hessian matrix estimation methods to address datasets with less stringent moment assumptions.

Acknowledgments

We thank the Editor, Associate Editor and referees. Zhang's research was supported by grants from the NSF of China (Grant Nos.U23A2064 and 12031005). Zhao's research was supported by the MOE Project of Humanities and Social Sciences (No. 21YJCZH235), the Hangzhou Joint Fund of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHZY24A010002 and Scientific Research Foundation of Hangzhou City University (No.J-202315).

Appendix A: Proof of the main results

We start with two auxiliary lemmas, where the proof can be found in Loh and Wainwright (2015).

Lemma A.1. *If $P_\lambda : \mathbb{R}^p \to \mathbb{R}$ is an amenable regularizer defined in Definition 1, then the function $\lambda\|\boldsymbol{\beta}\|_1 - P_\lambda(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$ is everywhere differentiable. Moreover, the function $\mu\|\boldsymbol{\beta}\|_2^2/2 + P_\lambda(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1$ is convex and for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have*

$$\lambda\|\boldsymbol{\beta}\|_1 \leq P_\lambda(\boldsymbol{\beta}) + \frac{\mu}{2}\|\boldsymbol{\beta}\|_2^2.$$

Lemma A.2. *Suppose that the regularizer P_λ satisfies the conditions (i)-(vi) in Definition 1. If $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is $k-$sparse, then for any $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\xi P_\lambda(\boldsymbol{\beta}^*) - P_\lambda(\boldsymbol{\beta}) > 0$ and $\xi \geq 1$, we have*

$$\xi P_\lambda(\boldsymbol{\beta}^*) - P_\lambda(\boldsymbol{\beta}) \leq \xi\|\Delta\boldsymbol{\beta}_{\mathcal{B}}\|_1 - \|\Delta\boldsymbol{\beta}_{\mathcal{B}^C}\|_1,$$

where $\Delta\boldsymbol{\beta} := \boldsymbol{\beta}^* - \boldsymbol{\beta}$ and $\mathcal{B}$ is the index set of the k largest elements of $\boldsymbol{\beta}$ in magnitude.

A.1. Proof of Proposition 1

Lemma A.3. *Under Assumption 1, we have*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty = O_p(\sqrt{\ln p/n}). \tag{A.1}$$

Proof. Define $\sigma_{i,jk} = x_{ij}x_{ik}$, $\bar{\sigma}_{i,jk} = \sigma_{i,jk}I\{|\sigma_{i,jk}| \le CK^2\}$ and $\tilde{\sigma}_{i,jk} = \sigma_{i,jk}I\{|\sigma_{i,jk}| > CK^2\}$. Then by Bernstein inequality and union bound inequality, we have

$$\Pr\left(\max_{1\le j,k\le p}\left|\frac{1}{n}\sum_{i}^{n}(\bar{\sigma}_{i,jk} - \mathsf{E}[\bar{\sigma}_{i,jk}])\right| \ge t\right) \le p^2\exp\left[-\frac{(nt)^2/2}{\kappa^2 + CK^2nt/3}\right],$$

where $\kappa^2 = \sum_{i=1}^{n}\mathsf{Var}[\bar{\sigma}_{i,jk}] = O(n)$. And this further implies that

$$\max_{1\le j,k\le p}\left|\frac{1}{n}\sum_{i}^{n}(\bar{\sigma}_{i,jk} - \mathsf{E}[\bar{\sigma}_{i,jk}])\right| = O_p(\sqrt{\ln p/n}) + O_p(K^2\ln p/n). \tag{A.2}$$

On the other hand,

$$|\mathsf{E}[\tilde{\sigma}_{i,jk}]| \le (\mathsf{E}[\sigma_{i,jk}^2]\Pr\{|\sigma_{i,jk}| > CK^2\})^{1/2} = o(\sqrt{\ln p/n}),$$

then we have

$$\begin{aligned}
&\Pr\left(\max_{1\le j,k\le p}\left|\frac{1}{n}\sum_{i}^{n}(\tilde{\sigma}_{i,jk} - \mathsf{E}[\tilde{\sigma}_{i,jk}])\right| > C\max\{\sqrt{\ln p/n}, K^2\ln p/n\}\right)\\
\le\ &\Pr\left(\max_{1\le j,k\le p}\left|\frac{1}{n}\sum_{i}^{n}\tilde{\sigma}_{i,jk}\right| > \frac{C}{2}\max\{\sqrt{\ln p/n}, K^2\ln p/n\}\right)\\
\le\ &\Pr\left(\max_{1\le j,k\le p}\max_{1\le i\le n}|\sigma_{i,jk}| > CK^2\right) = o(1),
\end{aligned} \tag{A.3}$$

where the last step follows from the fact that

$$\Pr\left(\max_{1\le j,k\le p}\max_{1\le i\le n}|\sigma_{i,jk}| > CK^2\right) = \Pr\left(\max_{1\le j\le p}\max_{1\le i\le n}|x_{i,j}| > \sqrt{C}K\right)$$

and the assumption $\|\mathbf{X}\|_\infty = O_p(K)$.

Combing (A.2), (A.3), and the assumption $K^2\sqrt{\ln p/n} = o(1)$, we prove (A.1). □

Lemma A.4. *Under Assumption 1, we have*

$$\|\mathbf{X}^\top\mathbf{W}_{\boldsymbol{\beta}^*}^2\boldsymbol{\epsilon}/n\|_\infty = O_p(\sqrt{\ln p/n}). \tag{A.4}$$

Proof. Following a similar argument as in the proof of Lemma A.3, we can prove the result. □

Proof of Proposition 1. Recall that $\nabla L_n(\boldsymbol{\beta}^*) = -\mathbf{X}^\top\mathbf{W}_{\boldsymbol{\beta}^*}^2\boldsymbol{\epsilon}/n$ and let $z_\infty^* = \|\nabla L_n(\boldsymbol{\beta}^*)\|_\infty$. By Lemma 4 of Gu and Zou (2016) and letting $\Delta\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, we can prove that

$$\min\{\tau, 1-\tau\}\|\mathbf{X}\Delta\boldsymbol{\beta}\|_2^2/n \le \langle\nabla L_n(\hat{\boldsymbol{\beta}}) - \nabla L_n(\boldsymbol{\beta}^*), \Delta\boldsymbol{\beta}\rangle. \tag{A.5}$$

On the left hand side, we have

$$\|\mathbf{X}\Delta\boldsymbol{\beta}\|_2^2/n = \Delta\boldsymbol{\beta}^\top\Sigma\Delta\boldsymbol{\beta} + \Delta\boldsymbol{\beta}^\top(\hat{\Sigma} - \Sigma)\Delta\boldsymbol{\beta} \ge \lambda_{\min}(\boldsymbol{\Sigma})\|\Delta\boldsymbol{\beta}\|_2^2 - \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_\infty\|\Delta\boldsymbol{\beta}\|_1^2, \tag{A.6}$$

and on the right hand side, we have

$$\langle\nabla L_n(\hat{\boldsymbol{\beta}}) - \nabla L_n(\boldsymbol{\beta}^*), \Delta\boldsymbol{\beta}\rangle \le \langle-\nabla P_\lambda(\hat{\boldsymbol{\beta}}) - \nabla L_n(\boldsymbol{\beta}^*), \Delta\boldsymbol{\beta}\rangle, \tag{A.7}$$

because any solution $\hat{\boldsymbol{\beta}}$ of the optimization problem (2) satisfies the first-order necessary condition, i.e.,

$$\langle \nabla L_n(\hat{\boldsymbol{\beta}}) + \nabla P_\lambda(\hat{\boldsymbol{\beta}}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \rangle \geq 0, \quad \text{for any } \|\boldsymbol{\beta}\|_1 \leq R.$$

Furthermore, noting that the $\mu-$amenable regularizer holds the following property,

$$\langle \nabla P_\lambda(\hat{\boldsymbol{\beta}}), -\Delta\boldsymbol{\beta} \rangle \leq P_\lambda(\boldsymbol{\beta}^*) - P_\lambda(\hat{\boldsymbol{\beta}}) + \frac{\mu}{2}\|\Delta\boldsymbol{\beta}\|_2^2,$$

we can get

$$\langle -\nabla P_\lambda(\hat{\boldsymbol{\beta}}) - \nabla L_n(\boldsymbol{\beta}^*), \Delta\boldsymbol{\beta} \rangle \leq z_\infty^* \|\Delta\boldsymbol{\beta}\|_1 + P_\lambda(\boldsymbol{\beta}^*) - P_\lambda(\hat{\boldsymbol{\beta}}) + \frac{\mu}{2}\|\Delta\boldsymbol{\beta}\|_2^2. \tag{A.8}$$

Combining (A.5)–(A.8), we obtain

$$\begin{aligned}
\min\{\tau, 1-\tau\}\lambda_{\min}(\boldsymbol{\Sigma})\|\Delta\boldsymbol{\beta}\|_2^2 &\leq (z_\infty^* + \min\{\tau, 1-\tau\}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_\infty \|\Delta\boldsymbol{\beta}\|_1)\|\Delta\boldsymbol{\beta}\|_1 + P_\lambda(\boldsymbol{\beta}^*) - P_\lambda(\hat{\boldsymbol{\beta}}) + \frac{\mu}{2}\|\Delta\boldsymbol{\beta}\|_2^2 \\
&\leq (z_\infty^* + 2\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_\infty R) \cdot \frac{1}{\lambda}(P_\lambda(\Delta\boldsymbol{\beta}) + \frac{\mu}{2}\|\Delta\boldsymbol{\beta}\|_2^2) + P_\lambda(\boldsymbol{\beta}^*) - P_\lambda(\hat{\boldsymbol{\beta}}) + \frac{\mu}{2}\|\Delta\boldsymbol{\beta}\|_2^2,
\end{aligned}$$

where the last step follows from $\|\Delta\boldsymbol{\beta}\|_1 \leq \|\hat{\boldsymbol{\beta}}\|_1 + \|\boldsymbol{\beta}^*\|_1 \leq 2R$ and Lemma A.1. Note further that the regularizer is sub-addictivie, i.e., $P_\lambda(\Delta\boldsymbol{\beta}) \leq P_\lambda(\boldsymbol{\beta}^*) + P_\lambda(\hat{\boldsymbol{\beta}})$ and $\min\{\tau, 1-\tau\}\lambda_{\min}(\boldsymbol{\Sigma}) > 3\mu/4$. Under the event

$$\mathcal{E} = \{\lambda \geq 4z_\infty^*\} \cap \{\lambda \geq 8\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_\infty R\},$$

we deduce that

$$0 \leq (\min\{\tau, 1-\tau\}\lambda_{\min}(\boldsymbol{\Sigma}) - \frac{3}{4}\mu)\|\Delta\boldsymbol{\beta}\|_2^2 \leq \frac{3}{2}P_\lambda(\boldsymbol{\beta}^*) - \frac{1}{2}P_\lambda(\hat{\boldsymbol{\beta}}). \tag{A.9}$$

Denote $\mathcal{A}_1$ the index set of the s largest components of $\Delta\boldsymbol{\beta}$ in magnitude, which may be slightly different from the active set $\mathcal{A}$ in most cases. By the property of the $\mu-$amenable regularizer (see Lemma 5 in Loh and Wainwright (2015)), we have

$$0 < \frac{3}{2}P_\lambda(\boldsymbol{\beta}^*) - \frac{1}{2}P_\lambda(\hat{\boldsymbol{\beta}}) \leq \frac{3}{2}\lambda\|\Delta\boldsymbol{\beta}_{\mathcal{A}_1}\|_1 - \frac{1}{2}\lambda\|\Delta\boldsymbol{\beta}_{\mathcal{A}_1^c}\|_1 \leq \frac{3}{2}\sqrt{s}\lambda\|\Delta\boldsymbol{\beta}_{\mathcal{A}_1}\|_2 \leq \frac{3}{2}\sqrt{s}\lambda\|\Delta\boldsymbol{\beta}\|_2. \tag{A.10}$$

Then by combining (A.9) and (A.10), it yields the l_2-bound

$$\|\Delta\boldsymbol{\beta}\|_2 \leq \frac{6\sqrt{s}\lambda}{4\min\{\tau, 1-\tau\}\lambda_{\min}(\boldsymbol{\Sigma}) - 3\mu},$$

and the l_1-bound can be derived using (A.10), i.e.,

$$\|\Delta\boldsymbol{\beta}\|_1 = \|\Delta\boldsymbol{\beta}_{\mathcal{A}_1}\|_1 + \|\Delta\boldsymbol{\beta}_{\mathcal{A}_1^c}\|_1 \leq 4\|\Delta\boldsymbol{\beta}_{\mathcal{A}_1}\|_1 \leq 4\sqrt{s}\|\Delta\boldsymbol{\beta}\|_2.$$

Next, from (A.5) and (A.7) we can obtain the following result for the prediction error bound,

$$\|\mathbf{X}\Delta\boldsymbol{\beta}\|_2^2/n \leq \frac{s\lambda^2}{\min\{\tau, 1-\tau\}}\left(\frac{12}{4\min\{\tau, 1-\tau\}\lambda_{\min}(\boldsymbol{\Sigma}) - 3\mu} + \frac{36\mu}{(4\min\{\tau, 1-\tau\}\lambda_{\min}(\boldsymbol{\Sigma}) - 3\mu)^2}\right).$$

Finally, by Lemmas A.3 and A.4 and choosing λ such that $\lambda \geq c_7 R\sqrt{\ln p/n}$ for some large positive constant c_7, we can prove that $\Pr(\mathcal{E}) \to 1$, and we complete the proof of Proposition 1.

$\square$

Proof of Lemma 1. Since

$$|w^2_{\boldsymbol{\beta}^*,i} - w^2_{\hat{\boldsymbol{\beta}},i}| = |1-2\tau| \cdot \mathbb{I}(|w^2_{\boldsymbol{\beta}^*,i} - w^2_{\hat{\boldsymbol{\beta}},i}| > 0),$$

and

$$\{|w^2_{\boldsymbol{\beta}^*,i} - w^2_{\hat{\boldsymbol{\beta}},i}| > 0\} \subseteq \{|\epsilon_i| \leq |\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|\},$$

then we have

$$|w^2_{\boldsymbol{\beta}^*,i} - w^2_{\hat{\boldsymbol{\beta}},i}| \leq \mathbb{I}(|\epsilon_i| \leq |\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|),$$

holds point-wisely for any $i \in \{1, \ldots, n\}$. By Markov's inequality,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n |w^2_{\hat{\boldsymbol{\beta}},i} - w^2_{\boldsymbol{\beta}^*,i}| \geq \delta_1\right) \leq \frac{1}{n\delta_1}\sum_{i=1}^n \Pr\left(\left|\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right| > \delta_2\right) + \frac{1}{n\delta_1}\sum_{i=1}^n \Pr\left(|\epsilon_i| \leq \delta_2\right), \tag{A.11}$$

holds for any $\delta_1, \delta_2 > 0$. Note that by Hölder's inequality,

$$\frac{1}{n}\sum_{i=1}^n \Pr\left(\left|\boldsymbol{X}_i^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right| > \delta_2\right) \leq \frac{1}{n}\sum_{i=1}^n \Pr\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \|\boldsymbol{X}_i\|_\infty > \delta_2\right) = \Pr\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \|\boldsymbol{X}_i\|_\infty > \delta_2\right) \tag{A.12}$$

and

$$\frac{1}{n}\sum_{i=1}^n \Pr\left(|\epsilon_i| \leq \delta_2\right) \leq 2\delta_2 \cdot \sup_{x \in (-\delta_2, +\delta_2)} f_\epsilon(x) \leq 2\delta_2 \cdot \sup_{x \in (-\infty, +\infty)} f_\epsilon(x). \tag{A.13}$$

Thus, by combining (A.11)–(A.13), and taking $\delta_1 = c_9\delta_2$ and $\delta_2 = c_{10}Ks\lambda$, where c_9 and c_{10} are large enough positive constants, we can prove (10). $\square$

A.2. Proof of Theorem 1 and Theorem 2

Proof of Theorem 1. We start with the $\mathbf{W}_{\boldsymbol{\beta}^*}$-weighted node-wise regression, that is

$$\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)} = \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} + \boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}, \tag{A.14}$$

where $\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)}$ is the $j-$th column of $\mathbf{X}_{\boldsymbol{\beta}^*} = \mathbf{W}_{\boldsymbol{\beta}^*}\mathbf{X}$, and $\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}$ is defined by

$$\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} = \arg\min_{\boldsymbol{\varphi}} \mathsf{E}\|\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)} - \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}\|_2^2 = \arg\min_{\boldsymbol{\varphi}} \mathsf{E}[(x_{\boldsymbol{\beta}^*,(j),1} - \boldsymbol{X}^\top_{\boldsymbol{\beta}^*,(-j),1}\boldsymbol{\varphi})^2],$$

where $x_{\boldsymbol{\beta}^*,(j),1}$ is the first element of $\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)}$ and $\boldsymbol{X}^\top_{\boldsymbol{\beta}^*,(-j),1}$ is the first row of $\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}$. Furthermore, we denote the variance of the residual by

$$\phi^2_{\boldsymbol{\beta}^*,j} = \mathsf{E}\|\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}\|_2^2/n = \mathsf{E}[\varrho^2_{\boldsymbol{\beta}^*,j,1}], \quad j \in \{1, \ldots, p\}.$$

Thus we have $\phi^2_{\boldsymbol{\beta}^*,j} = 1/\Theta_{\boldsymbol{\beta}^*,jj}$. Recall that $\boldsymbol{\Theta}_{\boldsymbol{\beta}^*,j} = (-\varphi_{\boldsymbol{\beta}^*,j,1}/\phi^2_{\boldsymbol{\beta}^*,j}, ..., -\varphi_{\boldsymbol{\beta}^*,j,p}/\phi^2_{\boldsymbol{\beta}^*,j})^\top$. Using the eigenvalue condition,

$$1/\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}^*}) \leq \Theta_{\boldsymbol{\beta}^*,jj} \leq 1/\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}^*}),$$

and

$$\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\beta}^*})\|\mathbf{\Theta}_{\boldsymbol{\beta}^*,j}\|_2^2 \le \mathbf{\Theta}_{\boldsymbol{\beta}^*,j}^\top \mathbf{\Sigma}_{\boldsymbol{\beta}^*}\mathbf{\Theta}_{\boldsymbol{\beta}^*,j} = \Theta_{\boldsymbol{\beta}^*,jj} = 1/\phi^2_{\boldsymbol{\beta}^*,j},$$

then we obtain

$$\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\beta}^*}) \le \phi^2_{\boldsymbol{\beta}^*,j} \le \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\beta}^*}), \tag{A.15}$$

and

$$\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2 = \|\phi^2_{\boldsymbol{\beta}^*,j}\mathbf{\Theta}_{\boldsymbol{\beta}^*,j}\|_2^2 \le \phi^2_{\boldsymbol{\beta}^*,j}/\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\beta}^*}).$$

This further implies

$$\max_{1\le j\le p}\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 \le \max_{1\le j\le p}\sqrt{s_j}\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2 \le \max_{1\le j\le p}\sqrt{s_j\Sigma_{\boldsymbol{\beta}^*,jj}/\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\beta}^*})} = O(\sqrt{s^{**}}), \tag{A.16}$$

and

$$\max_{1\le j\le p}\|\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}\|_\infty = \max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol{\beta}^*}\mathbf{\Theta}_{\boldsymbol{\beta}^*,j}\|_\infty \cdot \max_{1\le j\le p}\phi^2_{\beta^*,j} = O_p(\sqrt{s^{**}}K).$$

Then by left-producting the diagonal matrix $\mathbf{W}_{\hat{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta}^*}^{-1}$ on both sides of (A.14), we have the $\mathbf{W}_{\hat{\boldsymbol{\beta}}}$-weighted node-wise regression,

$$\boldsymbol{X}_{\hat{\boldsymbol{\beta}},(j)} = \mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} + \mathbf{W}_{\hat{\boldsymbol{\beta}}}\mathbf{W}_{\boldsymbol{\beta}^*}^{-1}\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}.$$

By the definition of $\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}$ in (7), we have

$$\|\boldsymbol{X}_{\hat{\boldsymbol{\beta}},(j)} - \mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}\|_2^2/n + Q_{\lambda_j}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}) \le \|\boldsymbol{X}_{\hat{\boldsymbol{\beta}},(j)} - \mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2/n + Q_{\lambda_j}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}),$$

which further indicates that

$$\|\mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})\|_2^2/n + Q_{\lambda_j}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}) \le 2\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}^\top\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})/n + Q_{\lambda_j}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}). \tag{A.17}$$

Then by Hölder's inequality, we have

$$\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}^\top\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})/n \le (\iota_{1j}+\iota_{2j})\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1, \tag{A.18}$$

where $\iota_{1j} = \|\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}^\top\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty$ and $\iota_{2j} = \|\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}^\top(\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*} - \mathbf{I})\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty$. For ι_{1j} in (A.18), by Lemma A.3, (A.15), and (A.16), we obtain

$$\begin{aligned}\max_{1\le j\le p}\iota_{1j} = \max_{1\le j\le p}\|\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}^\top\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty &= \max_{1\le j\le p}\|\phi^2_{\boldsymbol{\beta}^*,j}\mathbf{\Theta}^\top_{\boldsymbol{\beta}^*,j}(\mathbf{X}^\top_{\boldsymbol{\beta}^*}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n - \mathbf{\Sigma}_{\boldsymbol{\beta}^*,-j})\|_\infty \\ &= O_p(\sqrt{s^{**}\ln p/n}),\end{aligned} \tag{A.19}$$

where $\mathbf{\Sigma}_{\boldsymbol{\beta}^*,-j} = [\mathbf{X}^\top_{\boldsymbol{\beta}^*}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n]$. For ι_{2j} in (A.18), by Hölder's inequality we have

$$\begin{aligned}\max_{1\le j\le p}\iota_{2j} &= \max_{1\le j\le p}\|\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}^\top(\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*} - \mathbf{I})\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty \\ &= \max_{1\le j\le p}\max_{1\le q\le p,q\ne j}\frac{1}{n}\left|\sum_{i=1}^n\frac{w^2_{\hat{\boldsymbol{\beta}},i} - w^2_{\boldsymbol{\beta}^*,i}}{w^2_{\boldsymbol{\beta}^*,i}}\varrho_{\boldsymbol{\beta}^*,ji}x_{\boldsymbol{\beta}^*,iq}\right| \\ &\le \max_{1\le j\le p}\max_{1\le q\le p,q\ne j}\max_{1\le i\le n}|\varrho_{\boldsymbol{\beta}^*,ji}x_{\boldsymbol{\beta}^*,il}|\cdot\frac{1}{n}\sum_{i=1}^n|w^2_{\hat{\boldsymbol{\beta}},i} - w^2_{\boldsymbol{\beta}^*,i}|, \\ &= O_p(\sqrt{s^{**}}K^2\cdot sK\lambda) = O_p(s\sqrt{s^{**}}K^3\lambda),\end{aligned} \tag{A.20}$$

where the last step follows from (A.16), the results in Proposition 1 and Lemma 1.

Denote by the population covariance matrix $\mathbf{\Sigma}_{-j,-j} = \mathsf{E}[\mathbf{X}_{(-j)}^\top \mathbf{X}_{(-j)}/n]$ and the sample covariance $\hat{\mathbf{\Sigma}}_{-j,-j} = \mathbf{X}_{(-j)}^\top \mathbf{X}_{(-j)}/n$, respectively. Note further that

$$\lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\min}(\mathbf{\Sigma}_{-j,-j}) \quad \text{and} \quad \|\mathbf{\Sigma}_{-j,-j} - \hat{\mathbf{\Sigma}}_{-j,-j}\|_\infty \leq \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_\infty,$$

then we have

$$\begin{aligned}
\|\mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})\|_2^2/n &\geq \min\{\tau, 1-\tau\} \cdot \|\mathbf{X}_{(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})\|_2^2/n \\
&\geq \min\{\tau, 1-\tau\} \left(\lambda_{\min}(\mathbf{\Sigma}_{-j,-j})\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2 - \|\mathbf{\Sigma}_{-j,-j} - \hat{\mathbf{\Sigma}}_{-j,-j}\|_\infty \|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1^2\right) \\
&\geq \min\{\tau, 1-\tau\} \left(\lambda_{\min}(\mathbf{\Sigma})\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2 - \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_\infty \|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1^2\right).
\end{aligned} \tag{A.21}$$

Since $\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 \leq R_j$, we have $\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 \leq \|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}\|_1 + \|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 \leq 2R_j$. Combining (A.17) with (A.18) and (A.21), we can derive that

$$\begin{aligned}
&0 \leq \min\{\tau, 1-\tau\}\lambda_{\min}(\mathbf{\Sigma})\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2 \\
&\quad \leq \left\{2(\iota_{1j} + \iota_{2j}) + 2\min\{\tau, 1-\tau\}R_j\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_\infty\right\}\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 + Q_{\lambda_j}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}) - Q_{\lambda_j}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}),
\end{aligned}$$

where ι_{1j} and ι_{2j} are defined in (A.18). Under the event

$$\mathcal{E}_j = \{\lambda_j \geq 8(\iota_{1j} + \iota_{2j})\} \cap \{\lambda_j \geq 8\min\{\tau, 1-\tau\}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_\infty R_j\},$$

by invoking Lemma A.1 and applying the sub-addictive property of the amenable regularizer in Proposition 1, it yields that

$$\min\{\tau, 1-\tau\}\lambda_{\min}(\mathbf{\Sigma})\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2 \leq \frac{3}{2}Q_{\lambda_j}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}) - \frac{1}{2}Q_{\lambda_j}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}) + \frac{3\mu}{4}\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2.$$

Since

$$\min\{\tau, 1-\tau\}\lambda_{\min}(\mathbf{\Sigma}) \geq 3\mu/4,$$

then

$$0 \leq (\min\{\tau, 1-\tau\}\lambda_{\min}(\mathbf{\Sigma}) - 3\mu/4)\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2 \leq \frac{3}{2}Q_{\lambda_j}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}) - \frac{1}{2}Q_{\lambda_j}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}).$$

Letting $s_j = \|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_0$, applying the conclusion in Lemma A.2, and following the similar argument in the proof of Proposition 1, we obtain the l_2 bound

$$\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2 \leq \frac{6\sqrt{s_j}\lambda_j}{4\min\{\tau, 1-\tau\}\lambda_{\min}(\mathbf{\Sigma}) - \mu}, \tag{A.22}$$

also the l_1 bound

$$\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 \leq \frac{24 s_j\lambda_j}{4\min\{\tau, 1-\tau\}\lambda_{\min}(\mathbf{\Sigma}) - \mu}, \tag{A.23}$$

and finally the prediction error bound

$$\|\mathbf{X}_{(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})\|_2^2/n \leq \frac{s_j\lambda_j^2}{\min\{\tau, 1-\tau\}}\left(\frac{12}{4\min\{\tau, 1-\tau\}\lambda_{\min}(\mathbf{\Sigma}) - \mu} + \frac{36\mu}{(4\min\{\tau, 1-\tau\}\lambda_{\min}(\mathbf{\Sigma}) - \mu)^2}\right). \tag{A.24}$$

Lastly, we can prove $\Pr(\bigcap_{j=1}^{p}\mathcal{E}_j) \to 1$ as $n \to \infty$ by (A.19) and (A.20) and choosing λ_j such that $\lambda_j \geq c_8((\max_j R_j\sqrt{\ln p/n}) \vee (s\sqrt{s^{**}}K^3\lambda))$ for $j \in \{1,\ldots,p\}$. By replacing s_j and λ_j in (A.22)–(A.24) with s^{**} and λ^{**}, respectively, we completes the proof of Theorem (1). $\square$

Proof of Theorem 2. Recall that

$$\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j} = \boldsymbol{X}^\top_{\hat{\boldsymbol{\beta}},(j)}(\boldsymbol{X}_{\hat{\boldsymbol{\beta}},(j)} - \mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n.$$

Since

$$\boldsymbol{X}_{\hat{\boldsymbol{\beta}},(j)} = \mathbf{W}_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-1}_{\boldsymbol{\beta}^*}\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)} = \mathbf{W}_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-1}_{\boldsymbol{\beta}^*}(\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} + \boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}),$$

and

$$\boldsymbol{X}_{\hat{\boldsymbol{\beta}},(j)} - \mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} = \mathbf{W}_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-1}_{\boldsymbol{\beta}^*}(\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j} + \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} - \hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})), \tag{A.25}$$

then we can write

$$\begin{aligned}
\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j} - \phi^2_{\boldsymbol{\beta}^*,j} &= \left[\boldsymbol{X}^\top_{\boldsymbol{\beta}^*,(j)}(\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)} - \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n - \phi^2_{\boldsymbol{\beta}^*,j}\right] \\
&\quad + \left[\boldsymbol{X}^\top_{\boldsymbol{\beta}^*,(j)}(\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*} - \mathbf{I})(\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)} - \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n\right] \\
&=: \iota_{3j} + \iota_{4j}.
\end{aligned}$$

For ι_{3j}, we have

$$\begin{aligned}
&\left|\boldsymbol{X}^\top_{\boldsymbol{\beta}^*,(j)}(\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)} - \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n - \phi^2_{\boldsymbol{\beta}^*,j}\right| \\
\leq\ &\left|\boldsymbol{\varrho}^\top_{\boldsymbol{\beta}^*,j}\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}/n - \phi^2_{\boldsymbol{\beta}^*,j}\right| + \left|\boldsymbol{\varrho}^\top_{\boldsymbol{\beta}^*,j}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} - \hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n\right| \\
&+ \left|\boldsymbol{\varrho}^\top_{\boldsymbol{\beta}^*,j}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}/n\right| + \left|\boldsymbol{\varphi}^\top_{\boldsymbol{\beta}^*,j}\mathbf{X}^\top_{\boldsymbol{\beta}^*,(-j)}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} - \hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n\right| \\
=:\ &\iota_{3j,1} + \iota_{3j,2} + \iota_{3j,3} + \iota_{3j,4}.
\end{aligned} \tag{A.26}$$

By (A.16), we can prove

$$\max_{1\leq j\leq p}\iota_{3j,1} = \max_{1\leq j\leq p}\left|\phi^4_{\boldsymbol{\beta}^*,j}\boldsymbol{\Theta}^\top_{\boldsymbol{\beta}^*,j}(\mathbf{X}^\top_{\boldsymbol{\beta}^*}\mathbf{X}_{\boldsymbol{\beta}^*}/n - \boldsymbol{\Sigma}_{\boldsymbol{\beta}^*})\boldsymbol{\Theta}_{\boldsymbol{\beta}^*,j}\right| = O_p(s^{**}\sqrt{\ln p/n}). \tag{A.27}$$

Moreover, by Hölder's inequality and (A.19), we have

$$\begin{aligned}
\max_{1\leq j\leq p}\iota_{3j,2} &\leq \max_{1\leq j\leq p}\|\boldsymbol{\varrho}^\top_{\boldsymbol{\beta}^*,j}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty \cdot \max_{1\leq j\leq p}\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} - \hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}\|_1 \\
&= O_p(\sqrt{s^{**}\ln p/n})O_p(s^{**}\lambda^{**}) = O_p((s^{**})^{3/2}\lambda^{**}\sqrt{\ln p/n}),
\end{aligned} \tag{A.28}$$

and

$$\max_{1\leq j\leq p}\iota_{3j,3} \leq \max_{1\leq j\leq p}\|\boldsymbol{\varrho}^\top_{\boldsymbol{\beta}^*,j}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty \cdot \max_{1\leq j\leq p}\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1 = O_p(s^{**}\sqrt{\ln p/n}). \tag{A.29}$$

Note further that if we have $\sqrt{s^{**}}\lambda^{**} = o(1)$, then

$$
\begin{aligned}
& \|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2/n \\
\le\ & \left|\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2/n - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}^\top \mathsf{E}\left[\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}^\top \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\right]\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\right| + \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}^\top \mathsf{E}\left[\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}^\top \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\right]\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} \\
\le\ & \|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}^\top \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n - \mathsf{E}\left[\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}^\top \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\right]\|_\infty \|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1^2 + \lambda_{\max}(\Sigma_{\boldsymbol{\beta}^*})\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_2^2 \\
\le\ & O_p(\sqrt{\ln p/n})O(s^{**}) + O(1) = O(1).
\end{aligned}
$$

Meanwhile, by the Cauchy–Schwarz inequality,

$$
\begin{aligned}
\max_{1\le j\le p} \iota_{3j,4} &\le \max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}/\sqrt{n}\|_2 \cdot \max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} - \hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/\sqrt{n}\|_2 \\
&\le \max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}/\sqrt{n}\|_2 \cdot \sqrt{\max\{\tau, 1-\tau\}}\max_{1\le j\le p}\|\mathbf{X}_{(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})/\sqrt{n}\|_2 \\
&= O_p(\sqrt{s^{**}}\lambda^{**}).
\end{aligned} \tag{A.30}
$$

Thus, combining (A.27)–(A.30) we prove that

$$
\max_{1\le j\le p}|\iota_{3j}| = O_p(s^{**}\sqrt{\ln p/n}) + O_p((s^{**})^{3/2}\lambda^{**}\sqrt{\ln p/n}) + O_p(s^{**}\sqrt{\ln p/n}) + O_p(\sqrt{s^{**}}\lambda^{**}) = O_p(\sqrt{s^{**}}\lambda^{**}).
$$

Now we turn to ι_{4j}. Similar to ι_{2j}, we have

$$
\begin{aligned}
|\iota_{4j}| \le & \left|\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)}^\top(\mathbf{W}_{\hat{\boldsymbol{\beta}}}^2\mathbf{W}_{\boldsymbol{\beta}^*}^{-2} - \mathbf{I})(\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)} - \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})/n\right| + \left|\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)}^\top(\mathbf{W}_{\hat{\boldsymbol{\beta}}}^2\mathbf{W}_{\boldsymbol{\beta}^*}^{-2} - \mathbf{I})\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})/n\right| \\
\le & \max_{1\le q\le p, q\ne j}\frac{1}{n}\left|\sum_{i=1}^n \frac{w_{\hat{\boldsymbol{\beta}},i}^2 - w_{\boldsymbol{\beta}^*,i}^2}{w_{\boldsymbol{\beta}^*,i}^2}\varrho_{\boldsymbol{\beta}^*,j,i}x_{\boldsymbol{\beta}^*,ij}\right| + \max_{1\le q\le p, q\ne j}\frac{1}{n}\left|\sum_{i=1}^n \frac{w_{\hat{\boldsymbol{\beta}},i}^2 - w_{\boldsymbol{\beta}^*,i}^2}{w_{\boldsymbol{\beta}^*,i}^2}|x_{\boldsymbol{\beta}^*,ij}||\boldsymbol{X}_{\boldsymbol{\beta}^*,(-j),i}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})|\right|.
\end{aligned}
$$

Since $\|\boldsymbol{X}_{\boldsymbol{\beta}^*,(j)}\|_\infty \le \|\boldsymbol{X}_{(j)}\|_\infty \le \|\mathbf{X}\|_\infty$, and

$$
\max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j})\|_\infty \le \max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}\|_\infty \cdot \max_{1\le j\le p}\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j} - \hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}\|_1 = O_p(Ks^{**}\lambda^{**}),
$$

then by Hölder's inequality and (A.20), we get

$$
\max_{1\le j\le p}|\iota_{4j}| \le O_p(s\sqrt{s^{**}}K^3\lambda) + O_p(Ks\lambda)O_p(K)O_p(Ks^{**}\lambda^{**}) = O_p(s\sqrt{s^{**}}K^3\lambda). \tag{A.31}
$$

Thus, by (A.27) and (A.31) we prove that

$$
\max_{1\le j\le p}|\hat{\phi}_{\hat{\boldsymbol{\beta}},j}^2 - \phi_{\boldsymbol{\beta}^*,j}^2| = O_p(\sqrt{s^{**}}\lambda^{**}) + O_p(s\sqrt{s^{**}}K^3\lambda) = O_p(\sqrt{s^{**}}\lambda^{**}). \tag{A.32}
$$

Furthermore, by (A.15) we have

$$
\max_{1\le j\le p}|1/\hat{\phi}_{\hat{\boldsymbol{\beta}},j}^2 - 1/\phi_{\boldsymbol{\beta}^*,j}^2| = O_p(\sqrt{s^{**}}\lambda^{**}). \tag{A.33}
$$

Finally, combining (A.32) and (A.33), we have

$$
\begin{aligned}
\max_{1\le j\le p}\left\|\hat{\boldsymbol{\Theta}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\Theta}_{\boldsymbol{\beta}^*,j}\right\|_1 &= \max_{1\le j\le p}\left\|\hat{\boldsymbol{\Phi}}_{\hat{\boldsymbol{\beta}},j}/\hat{\phi}_{\hat{\boldsymbol{\beta}},j}^2 - \boldsymbol{\Phi}_{\boldsymbol{\beta}^*,j}/\phi_{\boldsymbol{\beta}^*,j}^2\right\|_1 \\
&\le \max_{1\le j\le p}\left(\|\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j} - \boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1/\hat{\phi}_{\hat{\boldsymbol{\beta}},j}^2\right) + \max_{1\le j\le p}\left(\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}\|_1(1/\hat{\phi}_{\hat{\boldsymbol{\beta}},j}^2 - 1/\phi_{\boldsymbol{\beta}^*,j}^2)\right) \\
&= O_p(s^{**}\lambda^{**}) + O(\sqrt{s^{**}})O_p(\sqrt{s^{**}}\lambda^{**}) = O_p(s^{**}\lambda^{**}).
\end{aligned}
$$

Analogously,

$$\begin{aligned}\max_{1\le j\le p}\|\hat{\mathbf{\Theta}}_{\hat{\boldsymbol\beta},j}-\mathbf{\Theta}_{\boldsymbol\beta^*,j}\|_2 &\le \max_{1\le j\le p}\left(\|\hat{\boldsymbol\varphi}_{\hat{\boldsymbol\beta},j}-\boldsymbol\varphi_{\boldsymbol\beta^*,j}\|_2/\hat\phi^2_{\hat{\boldsymbol\beta},j}\right)+\max_{1\le j\le p}\left(\|\boldsymbol\varphi_{\boldsymbol\beta^*,j}\|_2(1/\hat\phi^2_{\hat{\boldsymbol\beta},j}-1/\phi^2_{\boldsymbol\beta^*,j})\right)\\ &= O_p(\sqrt{s^{**}}\lambda^{**})+O(1)O_p(\sqrt{s^{**}}\lambda^{**})=O_p(\sqrt{s^{**}}\lambda^{**}).\end{aligned}$$

Now we complete our proof for Theorem 2. □

A.3. Proof of Theorem 3

Proof of Theorem 3. (i) Regarding $\mathbf{\Delta}^{(1)}$, by Hölder's inequality, we have

$$\|\mathbf{\Delta}^{(1)}\|_\infty\le\|\hat{\mathbf{\Theta}}_{\hat{\boldsymbol\beta}}\mathbf{X}^\top\|_\infty\|\mathbf{W}^2_{\hat{\boldsymbol\beta}}\boldsymbol\epsilon/n-\mathbf{W}^2_{\boldsymbol\beta^*}\boldsymbol\epsilon/n\|_1. \tag{A.34}$$

Note that

$$\begin{aligned}\|\hat{\mathbf{\Theta}}_{\hat{\boldsymbol\beta}}\mathbf{X}^\top\|_\infty=O(\max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol\beta^*}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol\beta},j}\|_\infty)&\le O\left(\max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol\beta^*}\mathbf{\Theta}_{\boldsymbol\beta^*,j}\|_\infty+\max_{1\le j\le p}\|\mathbf{X}_{\boldsymbol\beta^*}(\hat{\mathbf{\Theta}}_{\hat{\boldsymbol\beta},j}-\mathbf{\Theta}_{\boldsymbol\beta^*,j})\|_\infty\right)\\ &\le O_p(\sqrt{s^{**}}K)+O_p(K)O_p(s^{**}\lambda^{**})=O_p(\sqrt{s^{**}}K).\end{aligned} \tag{A.35}$$

Moreover, by the arguments in Lemma 1, we have

$$\begin{aligned}\|\mathbf{W}^2_{\hat{\boldsymbol\beta}}\boldsymbol\epsilon/n-\mathbf{W}^2_{\boldsymbol\beta^*}\boldsymbol\epsilon/n\|_1&\le\frac{1}{n}\sum_{i=1}^n\left|w^2_{\hat{\boldsymbol\beta},i}-w^2_{\boldsymbol\beta^*,i}\right||\boldsymbol X_i^\top(\hat{\boldsymbol\beta}-\boldsymbol\beta^*)|\\ &\le O_p(Ks\lambda)\cdot\|\mathbf{X}\|_\infty\|\boldsymbol\beta^*-\hat{\boldsymbol\beta}\|_1=O_p(K^2s^2\lambda^2).\end{aligned} \tag{A.36}$$

Combining (A.34)–(A.36), we have

$$\|\mathbf{\Delta}^{(1)}\|_\infty\le O_p(\sqrt{s^{**}}K)O_p(K^2s^2\lambda^2)=O_p(s\lambda\lambda^{**})=o_p(n^{-1/2}),$$

as $s\sqrt{s^{**}}\lambda^{**}\lambda=o(n^{-1/2})$ by assumption.

(ii) Regarding $\mathbf{\Delta}^{(2)}$, by Hölder's inequality, we have

$$\|\mathbf{\Delta}^{(2)}\|_\infty\le\|\hat{\mathbf{\Theta}}_{\hat{\boldsymbol\beta}}\hat{\mathbf{\Sigma}}_{\hat{\boldsymbol\beta}}-\mathbf{I}\|_\infty\|\hat{\boldsymbol\beta}-\boldsymbol\beta^*\|_1. \tag{A.37}$$

Recall that

$$\hat\phi^2_{\hat{\boldsymbol\beta},j}=\boldsymbol X^\top_{\hat{\boldsymbol\beta},(j)}(\boldsymbol X_{\hat{\boldsymbol\beta},(j)}-\mathbf{X}_{\hat{\boldsymbol\beta},(-j)}\hat{\boldsymbol\varphi}_{\hat{\boldsymbol\beta},j})/n=\boldsymbol X^\top_{\hat{\boldsymbol\beta},(j)}\mathbf{X}_{\hat{\boldsymbol\beta}}\hat{\mathbf{\Phi}}_{\hat{\boldsymbol\beta},j}/n$$

and

$$\boldsymbol X^\top_{\hat{\boldsymbol\beta},(j)}\mathbf{X}_{\hat{\boldsymbol\beta}}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol\beta},j}/n=1.$$

which indicates that the diagonal elements of $\hat{\mathbf{\Theta}}_{\hat{\boldsymbol\beta}}\hat{\mathbf{\Sigma}}_{\hat{\boldsymbol\beta}}$ are all equal to 1. Thus, by using the argument in

(A.25), we can show that

$$\begin{aligned}
\|\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\hat{\mathbf{\Sigma}}_{\hat{\boldsymbol{\beta}}}-\mathbf{I}\|_\infty &= \max_{1\le j\le p}\left\|\mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}^\top \mathbf{X}_{\hat{\boldsymbol{\beta}}}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}},j}/n\right\|_\infty \\
&\le \max_{1\le j\le p}\left\|\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j}\mathbf{X}_{\hat{\boldsymbol{\beta}},(-j)}^\top \mathbf{X}_{\hat{\boldsymbol{\beta}}}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}},j}/n\right\|_\infty \cdot \max_{1\le j\le p} 1/\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j} \\
&= \max_{1\le j\le p}\left\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}^\top \mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*}\boldsymbol{\varrho}_{\boldsymbol{\beta}^*,j}/n\right\|_\infty \cdot \max_{1\le j\le p} 1/\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j} \\
&\quad + \max_{1\le j\le p}\left\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}^\top \mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}-\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n\right\|_\infty \cdot \max_{1\le j\le p} 1/\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j} \\
&\quad + \max_{1\le j\le p}\left\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}^\top (\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*}-\mathbf{I})\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}-\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/n\right\|_\infty \cdot \max_{1\le j\le p} 1/\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j} \\
&=: (\iota_{5j}+\iota_{6j}+\iota_{7j})\cdot \max_{1\le j\le p} 1/\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j}. \qquad (A.38)
\end{aligned}$$

Recall that $\iota_{1j}=\|\boldsymbol{\varrho}^\top_{\boldsymbol{\beta}^*,j}\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty$ and $\iota_{2j}=\|\boldsymbol{\varrho}^\top_{\boldsymbol{\beta}^*,j}(\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*}-\mathbf{I})\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty$, which are defined in (A.18). For ι_{5j}, by (A.19) and (A.20) we have

$$\iota_{5j}\le \max_{1\le j\le p}\iota_{1j}+\max_{1\le j\le p}\iota_{2j}=O_p(\sqrt{s^{**}\ln p/n})+O_p(s\sqrt{s^{**}}K^3\lambda), \qquad (A.39)$$

By the Cauchy-Schwarz inequality, we have

$$\iota_{6j}\le \max_{1\le j\le p}\max_{1\le q\le p, q\ne j}\Big(\|\boldsymbol{X}_{\boldsymbol{\beta}^*,(q)}/\sqrt{n}\|_2\cdot\|\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}(\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}-\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j})/\sqrt{n}\|_2\Big)=O_p(\sqrt{s^{**}}\lambda^{**}), \qquad (A.40)$$

and

$$\iota_{7j}\le \max_{1\le j\le p}\|\boldsymbol{X}^\top_{\boldsymbol{\beta}^*,(-j)}(\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\mathbf{W}^{-2}_{\boldsymbol{\beta}^*}-\mathbf{I})\mathbf{X}_{\boldsymbol{\beta}^*,(-j)}/n\|_\infty\cdot\max_{1\le j\le p}\|\boldsymbol{\varphi}_{\boldsymbol{\beta}^*,j}-\hat{\boldsymbol{\varphi}}_{\hat{\boldsymbol{\beta}},j}\|_1=O_p(K^2s\lambda K)\cdot O_p(s^{**}\lambda^{**}). \qquad (A.41)$$

Combining (A.37)-(A.41), we can prove that

$$\begin{aligned}
\|\Delta^{(2)}\|_\infty &\le \Big(O_p(\sqrt{s^{**}\ln p/n})+O_p(s\sqrt{s^{**}}K^3\lambda)+O_p(K^3s\lambda)\cdot O_p(s^{**}\lambda^{**})+O_p(\sqrt{s^{**}}\lambda^{**})\Big)\cdot\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 \\
&= O_p(\sqrt{s^{**}}\lambda^{**})\cdot O_p(s\lambda)=o_p(n^{-1/2}),
\end{aligned}$$

given that $\lambda^{**}\ge c_8 s\sqrt{s^{**}}K^3\lambda$, $\max_{1\le j\le p}1/\hat{\phi}^2_{\hat{\boldsymbol{\beta}},j}=O_p(1)$ by (A.15) and (A.33), $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1=O_p(s\lambda)$ by Proposition 1, and $s\sqrt{s^{**}}\lambda^{**}\lambda=o(n^{-1/2})$ by assumption.

□

A.4. Proof of Theorem 4.

Proof of Theorem 4. Recall the decomposition (6),

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{de}-\boldsymbol{\beta}^* &= \hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top\mathbf{W}^2_{\boldsymbol{\beta}^*}\boldsymbol{\epsilon}/n-(\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top\mathbf{W}^2_{\boldsymbol{\beta}^*}\boldsymbol{\epsilon}/n-\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top\mathbf{W}^2_{\hat{\boldsymbol{\beta}}}\boldsymbol{\epsilon}/n)-(\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\hat{\mathbf{\Sigma}}_{\hat{\boldsymbol{\beta}}}-\mathbf{I})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \\
&:= \hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top\mathbf{W}^2_{\boldsymbol{\beta}^*}\boldsymbol{\epsilon}/n-\boldsymbol{\Delta}^{(1)}-\boldsymbol{\Delta}^{(2)}.
\end{aligned}$$

Since Theorem 3 shows that the terms $\sqrt{n}\boldsymbol{\Delta}^{(1)}$ and $\sqrt{n}\boldsymbol{\Delta}^{(2)}$ are negligible, we only need to verify the following two items to conclude the desired results:

(i) The asymptotic normality of $\mathbf{H}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top\mathbf{W}^2_{\boldsymbol{\beta}^*}\boldsymbol{\epsilon}/\sqrt{n}$.

(ii) The estimator of the asymptotic variance-covariance matrix $\hat{\mathbf{\Omega}}_{\mathbf{H}}$ is consistent with $\mathbf{\Omega}_{\mathbf{H}}$.

To prove (i), note that

$$\begin{aligned}\left\|\mathbf{H}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\sum_{i=1}^{n}\boldsymbol{X}_i w_{\boldsymbol{\beta}^*,i}^2\epsilon_i/n-\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\sum_{i=1}^{n}\boldsymbol{X}_i w_{\boldsymbol{\beta}^*,i}^2\epsilon_i/n\right\|_{\infty} &\leq \|\sum_{i=1}^{n}\boldsymbol{X}_i w_{\boldsymbol{\beta}^*,i}^2\epsilon_i/n\|_{\infty}\|\mathbf{H}\|_{l_\infty}\|\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}-\mathbf{\Theta}_{\boldsymbol{\beta}^*}\|_{l_\infty}\\ &= O_p(s^{**}\lambda^{**}\sqrt{\ln p/n})\end{aligned}$$

and that $\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\boldsymbol{X}_i w_{\boldsymbol{\beta}^*,i}^2\epsilon_i$, $i\in\{1,\ldots,n\}$ are i.i.d. sequences of p_0-dimensional vectors with zero mean and finite variance. Thus, given $s^{**}\lambda^{**}\sqrt{\ln p/n}=o(n^{-1/2})$, by the Lindeberg-Levy central limit theorem along with the Slutsky's lemma, it follows that

$$\mathbf{H}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\mathbf{X}^\top\mathbf{W}_{\boldsymbol{\beta}^*}^2\boldsymbol{\epsilon}/\sqrt{n}\xrightarrow{d}\mathcal{N}_{p_0}\left(\mathbf{0},\mathbf{\Omega}_{\mathbf{H}}\right)\quad \textit{with}\quad \mathbf{\Omega}_{\mathbf{H}}=\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\mathsf{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top w_{\boldsymbol{\beta}^*,i}^4\epsilon_i^2\right]\mathbf{\Theta}_{\boldsymbol{\beta}^*}^\top\mathbf{H}^\top.$$

For (ii), note that we further assume $\mathsf{E}[x_{ij}^8]=O(1)$ and $K^4\sqrt{\ln p/n}=o(1)$, then following a similar argument as in the proof of Lemma A.3, we can show that

$$\left\|\hat{\mathbf{Q}}-\mathbf{Q}\right\|_{\infty}=\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top w_{\boldsymbol{\beta}^*,i}^4\epsilon_i^2-\mathsf{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top w_{\boldsymbol{\beta}^*,i}^4\epsilon_i^2\right]\right\|_{\infty}=O_p(\sqrt{\ln p/n}),$$

where $\mathbf{Q}=\mathsf{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top w_{\boldsymbol{\beta}^*,i}^4\epsilon_i^2\right]$ and $\hat{\mathbf{Q}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top w_{\boldsymbol{\beta}^*,i}^4\epsilon_i^2$. Then we have

$$\begin{aligned}&\|\mathbf{H}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\hat{\mathbf{Q}}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}^\top\mathbf{H}^\top-\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\mathbf{Q}\mathbf{\Theta}_{\boldsymbol{\beta}^*}^\top\mathbf{H}^\top\|_{\infty}\\ \leq\ &\|\mathbf{H}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}\hat{\mathbf{Q}}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}^\top\mathbf{H}^\top-\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\hat{\mathbf{Q}}\mathbf{\Theta}_{\boldsymbol{\beta}^*}^\top\mathbf{H}^\top\|_{\infty}+\|\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}(\hat{\mathbf{Q}}-\mathbf{Q})\mathbf{\Theta}_{\boldsymbol{\beta}^*}^\top\mathbf{H}^\top\|_{\infty}\\ \leq\ &\|\hat{\mathbf{Q}}\|_{\infty}\|\mathbf{H}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}-\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\|_{l_\infty}^2+2\|\hat{\mathbf{Q}}\|_{\infty}\|\mathbf{H}\hat{\mathbf{\Theta}}_{\hat{\boldsymbol{\beta}}}-\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\|_{l_\infty}\|\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\|_{l_\infty}+\left\|\hat{\mathbf{Q}}-\mathbf{Q}\right\|_{\infty}\|\mathbf{H}\mathbf{\Theta}_{\boldsymbol{\beta}^*}\|_{l_\infty}^2\\ =\ &O_p((s^{**})^2(\lambda^{**})^2)+O_p(s^{**}\lambda^{**})O_p(\sqrt{s^{**}})+O_p(s^{**}\sqrt{\ln p/n})\\ =\ &O_p((s^{**})^{3/2}\lambda^{**}),\end{aligned}$$

which completes the proof. □

References

Cai, T., Liu, W., Luo, X., 2011. A constrained $l1$ minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association 106, 594–607.

Cai, T.T., Guo, Z., 2017. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. The Annals of Statistics 45, 615–646.

Cai, T.T., Guo, Z., Ma, R., 2023. Statistical inference for high-dimensional generalized linear models with binary outcomes. Journal of the American Statistical Association 118, 1319–1332.

Chronopoulos, I., Chrysikou, K., Kapetanios, G., 2022. High dimensional generalised penalised least squares. arXiv preprint arXiv:2207.07055 .

Ciuperca, G., 2021. Variable selection in high-dimensional linear model with possibly asymmetric errors. Computational Statistics & Data Analysis 155, 107–112.

Dezeure, R., Bühlmann, P., Zhang, C.H., 2017. High-dimensional simultaneous inference with the bootstrap. Test 26, 685–719.

Fan, J., Fan, Y., Barut, E., 2014. Adaptive robust variable selection. Annals of Statistics 42, 324–351.

Fan, J., Feng, Y., Wu, Y., 2009. Network exploration via the adaptive lasso and scad penalties. The Annals of Applied Statistics 3, 521–541.

Fan, J., Li, Q., Wang, Y., 2017. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. Journal of the Royal Statistical Society Series B: Statistical Methodology 79, 247–265.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Fan, J., Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85, 645–660.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9, 432–441.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics 42, 1166–1202.

Gu, Y., Zou, H., 2016. High-dimensional generalizations of asymmetric least squares regression and their applications. Annals of Statistics 44, 2661–2694.

Huang, C.C., Liu, K., Pope, R.M., Du, P., Lin, S., Rajamannan, N.M., Huang, Q.Q., Jafari, N., Burke, G.L., Post, W., et al., 2011. Activated TLR signaling in atherosclerosis among women with lower framingham risk score: the multi-ethnic study of atherosclerosis. PloS one 6, e21067.

Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. The Journal of Machine Learning Research 15, 2869–2909.

Jiang, F., Zhou, Y., Liu, J., Ma, Y., 2023. On high-dimensional poisson models with measurement error: Hypothesis testing for nonlinear nonconvex optimization. Annals of statistics 51, 233.

Jiang, R., Peng, Y., Deng, Y., 2021. Variable selection and debiased estimation for single-index expectile model. Australian & New Zealand Journal of Statistics 63, 658–673.

Koenker, R., 2005. Quantile regression. volume 38. Cambridge university press.

Li, D., Wang, L., Zhao, W., 2022a. Estimation and inference for multikink expectile regression with longitudinal data. Statistics in Medicine 41, 1296–1313.

Li, S., Zhang, L., Cai, T.T., Li, H., 2023. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. Journal of the American Statistical Association , forthcoming.

Li, X., Zhang, Y., Zhao, J., 2022b. An improved algorithm for high-dimensional continuous threshold expectile model with variance heterogeneity. Journal of Statistical Computation and Simulation 92, 1590–1617.

Liao, L., Park, C., Choi, H., 2019. Penalized expectile regression: an alternative to penalized quantile regression. Annals of the Institute of Statistical Mathematics 71, 409–438.

Liu, H., Wang, L., 2017. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. Electronic Journal of Statistics 11, 241–294.

Loh, P.L., 2017. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. The Annals of Statistics , 866–896.

Loh, P.L., Wainwright, M.J., 2015. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Journal of Machine Learning Research 16, 559–616.

Loh, P.L., Wainwright, M.J., 2017. Support recovery without incoherence: A case for nonconvex regularization. The Annals of Statistics 45, 2455–2482.

Luo, B., Gao, X., 2022. High-dimensional robust approximated m-estimators for mean regression with asymmetric data. Journal of Multivariate Analysis 192, 105080.

Man, R., Tan, K.M., Wang, Z., Zhou, W.X., 2024. Retire: Robust expectile regression in high dimensions. Journal of Econometrics , forthcoming.

McCracken, M.W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34, 574–589.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34, 1436–1462.

Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica: Journal of the Econometric Society , 819–847.

Song, S., Lin, Y., Zhou, Y., 2021. Linear expectile regression under massive data. Fundamental Research 1, 574–585.

Sun, T., Zhang, C.H., 2012. Scaled sparse linear regression. Biometrika 99, 879–898.

Taamouti, A., 2015. Stock market's reaction to money supply: a nonparametric analysis. Studies in Nonlinear Dynamics & Econometrics 19, 669–689.

Verzelen, N., 2012. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. Electronic Journal of Statistics 6, 38–90.

Vial, J.P., 1982. Strong convexity of sets and functions. Journal of Mathematical Economics 9, 187–205.

Wang, L., Wu, Y., Li, R., 2012. Quantile regression for analyzing heterogeneity in ultra-high dimension. Journal of the American Statistical Association 107, 214–222.

Wang, Z., Liu, H., Zhang, T., 2014. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. Annals of statistics 42, 2164–2201.

Wirsik, N., Otto-Sobotka, F., Pigeot, I., 2019. Modeling physical activity data using L_0-penalized expectile regression. Biometrical Journal 61, 1371–1384.

Xu, Q., Ding, X., Jiang, C., Yu, K., Shi, L., 2021. An elastic-net penalized expectile regression with applications. Journal of Applied Statistics 48, 2205–2230.

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: Statistical Methodology , 217–242.

Zhang, C.H., Zhang, T., 2012. A general theory of concave regularization for high-dimensional sparse estimation problems. Statistical Science , 576–593.

Zhang, X., Cheng, G., 2017. Simultaneous inference for high-dimensional linear models. Journal of the American Statistical Association 112, 757–768.

Zhao, J., Chen, Y., Zhang, Y., 2018. Expectile regression for analyzing heteroscedasticity in high dimension. Statistics & Probability Letters 137, 304–311.

Zhao, J., Yan, G., Zhang, Y., 2022. Robust estimation and shrinkage in ultrahigh dimensional expectile regression with heavy tails and variance heterogeneity. Statistical Papers , 1–28.

Zhao, J., Zhang, Y., 2018. Variable selection in expectile regression. Communications in Statistics-Theory and Methods 47, 1731–1746.

Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics 36, 1509–1533.