Graduate School of Economics, Hitotsubashi University
Discussion Paper Series No. 2025-01

Expected Shortfall Regression for High-Dimensional Additive Models

Toshio HONDA and Po-Hsiang PENG

February 2025

Expected Shortfall Regression for High-Dimensional Additive Models

Toshio HONDA and Po-Hsiang PENG

February 10, 2025

Abstract

The expected shortfall (ES) regression can be a powerful and useful tool to analyze the relation between the response variable and the covariates through the conditional mean. As is well-known, there is no single loss function for expected shortfall estimation and there is a suitable loss function for joint estimation of quantile and expected shortfall. In addition to them, recently a very useful two-step procedure for ES regression was proposed : carry out quantile regression and then estimate the ES regression model by applying the least squares method. This procedure is successful due to the Neyman orthogonality. Then high dimensional linear regression models was considered based on the the findings. By exploiting those results, we assume additive models for both quantile and expected shortfall in the high-dimensional setting and consider the group Lasso and SCAD estimators. We establish the oracle inequality and the oracle property for them. Our theoretical results also imply that quantile estimation does not affect ES estimation asymptotically. We also present numerical results that demonstrate satisfactory performance in model selection, estimation accuracy, and prediction error for a moderate sample size together with an empirical study.

Keywords: expected shortfall; quantile regression; group Lasso; group SCAD; B-spline basis; additive models.

1 Introduction

Suppose that we have n i.i.d observations, $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$ is a high-dimensional covariate vector. We allow p/n to go to infinity fast. When we analyze how the covariate vector affects the response variable, the expected shortfall (ES) regression can be a powerful and useful tool to analyze the relation between the

Toshio Honda is Professor, Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan (Email: t.honda@r.hit-u.ac.jp). Po-Hsiang Peng is Ph. D. student, Institute of Statistics and Data Science, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013, R.O.C. (Email: seanpph@gmail.com).

response variable and the covariate vector through both the conditional mean and quantile. The expected shortfall has been also a very common risk measure. Although there is no single loss function for ES estimation as in [14], there is a suitable loss function for joint estimation of quantile and expected shortfall. See [13] for the unconditional expected shortfall and [10] and [23] for the expected shortfall regression.

Recently a very useful two-step procedure for ES regression was proposed in [1] : carry out quantile regression and then estimate the ES regression model by applying the least squares method. This is successful due to the Neyman orthogonality. Based on the findings in [1], robust estimation in ES linear regression was proposed in [16], nonparametric regression of fixed dimension was dealt with in [30], and high-dimensional linear regression models was considered in [31]. See [8] and [19] for nonparametric ES estimation in the setups of fixed dimension. We assume additive models for both quantile and expected shortfall in the high-dimensional setting and consider the group Lasso and SCAD estimators. We establish the oracle inequality and the oracle property in Theorems 2-3 in Section 3. Our theoretical results imply that quantile estimation does not affect ES estimation asymptotically. Section 4 presents simulation studies for both the group Lasso and the group SCAD estimators. In terms of model selection consistency, estimation accuracy, and prediction error, the group SCAD estimator demonstrates superior performance. Furthermore, for a moderate sample size, it is comparable to (or only slightly underperforms) the benchmark that assumes all active features are known.

Specifically, we assume sparse high-dimensional additive models for both the conditional τ-th quantile and expected shortfall on $\boldsymbol{X}_i$:

$$\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i) = \mu_1 + \sum_{j=1}^{p} g_j(X_{ij}) \tag{1}$$

and

$$\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i) = \mathrm{E}[Y_i I\{Y_i \le \mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)\} \,|\, \boldsymbol{X}_i] \tag{2}$$

$$= \mu_2 + \sum_{j=1}^{p} h_j(X_{ij}), \tag{3}$$

where $X_{ij} \in [0,1]$ for all $j \in [p]$ and $[p]$ means $\{1, \ldots, p\}$.

We denote the active index set by S and this S is assumed to be common to $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ and $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ for simplicity of presentation. This S always includes the constant term, which has index 0, and we write $S_- \coloneqq S \setminus \{0\}$, $s \coloneqq |S|$, and $s_- \coloneqq |S_-|$,

where $|A|$ is the number of elements of a set A. In this paper, s is sufficiently small compared to n. Besides, we assume that $g_j(x_j)$ and $h_j(x_j)$ are sufficiently smooth and actually they are assumed to be twice continuously differentiable as we describe in Section 3.

Recently a lot of high-dimensional datasets are available and accordingly suitable statistical procedures have been proposed, the Lasso in [25], the SCAD in [11], and so on. There has been a huge amount of theoretical and applied research on high-dimensional models since these papers and a seminal paper [4] on the theoretical results on the Lasso. For example, see [7], [15], [26], [28], [12], and the references therein for commonly used procedures, their properties, and recent developments. Quantile regression has been a useful and insightful tool for statistical analysis. See [20] for details on quantile regression. As for high-dimensional quantile regression models, for example, see [3], [18], [17] and the references therein.

This paper is highly motivated by [31]. The paper considered the Lasso and the debiased Lasso for ES linear regression models in the high-dimensional setting. The authors derived the oracle inequalities and the asymptotic normality, respectively. The Lasso for time series data are considered in [2]. The oracle inequalities are not given in [2]. By improving the proofs in [31], we derive the oracle inequality for the group Lasso. However, our structured nonparametric regression model is more flexible than linear models and the treatment of approximation error and group structure is not trivial. There are also some other significant differences in the proofs between [31] and the present paper. See comments after Assumption A8 in Section 3. Instead of debiasing, we propose the group SCAD estimator in this paper since the debiasing for additive models look a little too complicated. Debiasing for additive models is a topic of future research. Due to the commonly used assumption in the least squares Lasso literature, Assumption A6, we don't allow heavy-tailed errors. See Remark 1 at the end of Section 3 about the conditional expectation between two conditional quantiles.

This paper is organized as follows. In Section 2, we describe our estimation procedures in detail. We present our theoretical results with the assumptions in Section 3. The results of our numerical studies are given in Section 4. We prove the theorems in Section 5 and conclude this paper with our concluding remarks in Section 6. The proofs of technical lemmas are given in the Appendix.

We introduce some notation here. For a vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|$, $\|\boldsymbol{a}\|_1$, and $\boldsymbol{a}^T$ mean the Euclidean norm, the ℓ_1 norm, and the transpose, respectively. For a symmetric matix

A, we denote the minimum and maximum eigenvalues of a symmetric matix A by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. We write $\|f\|_{L_2}$ and $\|f\|_{L_\infty}$ for the L_2 and L_∞ norms of a function $f(x)$ on $[0,1]$, respectively. $C, C_1, \ldots$ are generic positive constants and their values may vary from line to line. We use $D_0, D_1, \ldots$ as generic positive constants in theorems and lemmas. We use For real numbers a and b, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. The conditional expectation on $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ is denoted by $\mathrm{E}[\cdot|\{\boldsymbol{X}_\ell\}]$.

2 Estimation of expected shortfall

Hereafter we assume that $p > n$. If not, replace p with $p_n = p \vee n$ in the theorems, the lemmas, and the proofs.

We denote our basis for approximating $g_j(x_j)$ and $h_j(x_j)$ by

$$\Psi_j(x_j) = (\psi_{j1}(x_j), \ldots, \psi_{jm}(x_j))^T.$$

For example, we apply the Gram-Schmidt orthonormalization to the B-spline basis on $[0,1]$ w.r.t. some measure as in [17] and then remove the constant element to obtain the basis to be compatible with the identification condition on $g_j(x_j)$ and $h_j(x_j)$. Such bases constructed from the B-spline basis on $[0,1]$ work well. Specifically, they satisfy Assumptions A3 and A4(1) and the conditions on ξ_i and η_i under Assumption A2 below.

We should put some identification condition on $g_j(x_j)$ and $h_j(x_j)$. For example,

$$\int_0^1 g_j(x_j)dx_j = 0,\ j \in [p],\ \text{and}\ \int_0^1 h_j(x_j)dx_j = 0,\ j \in [p]$$

or

$$\int_0^1 g_j(x_j)f_j(x_j)dx_j = 0,\ j \in [p],\ \text{and}\ \int_0^1 h_j(x_j)f_j(x_j)dx_j = 0,\ j \in [p],$$

where $f_j(x_j)$ is the marginal density of X_{1j}. Then $\Psi_j(x_j)$ should satisfy

$$\int_0^1 \psi_{jk}(x_j)dx_j = 0,\ j \in [p] \text{ and } k \in [m],\ \text{or}$$
$$\int_0^1 \psi_{jk}(x_j)f_j(x_j)dx_j = 0,\ j \in [p] \text{ and } k \in [m], \tag{4}$$

accordingly. Practically we should use the empirical measure for the latter. If we adopt the former identification, we have $\Psi_j(x_j) = \Psi(x_j), j \in [p]$ for a suitable common basis $\Psi(x)$.

Since this paper is motivated by [31], we borrow some notation from [31] like $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in (5), $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in (6), and $Z_i(\boldsymbol{\beta})$ in (8).

Let ν be the index of smoothness of $g_j(x_j)$ and $h_j(x_j)$ and we assume $\nu = 2$ and take $m = c_m n^{1/(2\nu+1)}$ in this paper. Define the covariate vector $\boldsymbol{W}_i$ for the additive models by

$$\begin{aligned}\boldsymbol{W}_i &\coloneqq (1, \Psi_1^T(X_{i1}), \ldots, \Psi_p^T(X_{ip}))^T \\ &= (W_{i0}, \boldsymbol{W}_{i1}^T, \ldots, \boldsymbol{W}_{ip}^T)^T \in \mathbb{R}^{pm+1}.\end{aligned}$$

Under Assumption A2 below, there exists $\overline{\boldsymbol{\beta}} = (\overline{\beta}_0, \overline{\boldsymbol{\beta}}_1^T, \ldots, \overline{\boldsymbol{\beta}}_p^T)^T \in \mathbb{R}^{pm+1}$ satisfying

$$\begin{aligned}\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i) &= \mu_1 + \sum_{j=1}^p g_j(X_{ij}) \\ &= \overline{\beta}_0 + \sum_{j=1}^p \Psi_j^T(X_{ij})\overline{\boldsymbol{\beta}}_j + \eta_i \\ &= \boldsymbol{W}_i^T \overline{\boldsymbol{\beta}} + \eta_i,\end{aligned} \tag{5}$$

where $\eta_i^2 = O(n^{-2\nu/(2\nu+1)})$ uniformly in i and in addition $\sum_{i=1}^n \eta_i = 0$ without loss of generality. If not, we should replace η_i with $\eta_i - n^{-1}\sum_{\ell=1}^n \eta_\ell$. Then $\overline{\beta}_0$ depends on $\{\boldsymbol{X}_\ell\}$. However, the dependence does not affect the theorems or the proofs at all.

Similarly, there exists $\overline{\boldsymbol{\theta}} = (\overline{\theta}_0, \overline{\boldsymbol{\theta}}_1^T, \ldots, \overline{\boldsymbol{\theta}}_p^T)^T \in \mathbb{R}^{pm+1}$ satisfying

$$\begin{aligned}\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i) &= \mu_2 + \sum_{j=1}^p h_j(X_{ij}) \\ &= \overline{\theta}_0 + \sum_{j=1}^p \Psi_j^T(X_{ij})\overline{\boldsymbol{\theta}}_j + \xi_i \\ &= \boldsymbol{W}_i^T \overline{\boldsymbol{\theta}} + \xi_i,\end{aligned} \tag{6}$$

where $\xi_i^2 = O(n^{-2\nu/(2\nu+1)})$ uniformly in i with $\nu = 2$ and in addition $\sum_{i=1}^n \xi_i = 0$ without loss of generality. See Assumption A2 below.

Hereafter we use the same partition of $(pm+1)$-dimensional vectors as $\overline{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\theta}}$.

We estimate $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ by exploiting the group Lasso as in [18], whose theoretical results are reproduced as Theorem 1 here, and apply the group Lasso to estimation of $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ by employing the estimate of $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in a similar way to [31]. Then we establish useful theoretical results for estimation of $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in Theorems 2-3. Instead

of debiasing, we consider the group SCAD in estimating $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$.

First we begin with the group Lasso for $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in [18].

The group Lasso for $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{pm+1}}{\operatorname{argmin}} \left[n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{W}_i^T \boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} w_j \|\boldsymbol{\beta}_j\| \right], \tag{7}$$

where $\rho_\tau(\cdot)$ is the check function for the τ-th quantiel, β_0 is not included in the Lasso penalty, and

$$w_j = \begin{cases} 1, & j = 0 \\ \sqrt{m}, & j \in [p] \end{cases}.$$

As stated before, we can find the theory in [18] for this group Lasso estimator and we employ this group Lasso estimator for our ES estimation again. Note that we use a little different penalty from that in [18]. However, there is no theoretical difference due to Lemma 1 in this paper.

The group Lasso for $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$:

Writing $Q_i = \mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ and $S_i = \mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$, define

$$Z_i(Q_i) \coloneqq (Y_i - Q_i) I\{Y_i \le Q_i\} + \tau Q_i = Y_i I\{Y_i \le Q_i\} + Q_i[\tau - I\{Y_i \le Q_i\}].$$

Then we have

$$\mathrm{E}\{Z_i(Q_i) \,|\, \boldsymbol{X}_i\} = \tau S_i - \tau Q_i + \tau Q_i = \tau S_i.$$

Besides, the addition of $Q_i[\tau - I\{Y_i \le Q_i\}]$ deletes the effect of estimation of Q_i by making the score function insensitive to estimation of conditional quantile(the Neyman orthogonality). See [9] for the Neyman orthogonality. This interesting and useful observation is given in [1] and the results on high-dimensional linear regression models for $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in [31] are based on this observation. The authors of [31] considered the lasso and debiased lasso estimators and proved the oracle inequality and the asymptotic normality, respectively.

We consider additive models for both $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ and $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$. Additive models are more flexible than linear models and allow interpretation. Using the results for $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in [18] and the Lasso theory, we establish the oracle inequality for the group Lasso estimator in Theorem 2 and the "oracle" property for the group SCAD estimator in Theorem 3. The "oracle" property here means not only the active index set detection but also that the estimation of $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ does not affect the estimation of $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$

asymptotically. We verify our Theorem 2 by following the standard argument for the Lasso and improving the proof of Theorem 4.1 in [31]. There are some significant differences between the proofs. See comments after Assumption A8 in Section 3.

With $Z_i(\boldsymbol{\beta}) = Z_i(\boldsymbol{W}_i^T\boldsymbol{\beta})$ for simplicity of notation, this $Z_i(\boldsymbol{\beta})$ is represented as

$$Z_i(\boldsymbol{\beta}) = (Y_i - \boldsymbol{W}_i^T\boldsymbol{\beta})I\{Y_i \le \boldsymbol{W}_i^T\boldsymbol{\beta}\} + \tau\boldsymbol{W}_i^T\boldsymbol{\beta}. \tag{8}$$

The group Lasso estimator $\boldsymbol{W}_i^T\widehat{\boldsymbol{\theta}}$ is given by

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{pm+1}}{\operatorname{argmin}} \Big[(2n)^{-1}\sum_{i=1}^{n}(Z_i(\widehat{\boldsymbol{\beta}}) - \tau\boldsymbol{W}_i^T\boldsymbol{\theta})^2 + \tau\lambda_2\sum_{j=1}^{p} w_j\|\boldsymbol{\theta}_j\|\Big],$$

where $\widehat{\boldsymbol{\beta}}$ is from the group Lasso for $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ and θ_0 is not included in the Lasso penalty as before.

The group SCAD for $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$:

Instead of debiasing in [31], we present the SCAD estimator together with the oracle property since debiasing for additive models is too complicated. Debiasing is a topic of future research.

The group SCAD estimator $\boldsymbol{W}_i^T\widetilde{\boldsymbol{\theta}}$ is given by

$$\widetilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{pm+1}}{\operatorname{argmin}} \Big[(2n)^{-1}\sum_{i=1}^{n}(Z_i(\widehat{\boldsymbol{\beta}}) - \tau\boldsymbol{W}_i^T\boldsymbol{\theta})^2 + \sum_{j=1}^{p}\mathrm{SCAD}_{\lambda_3}(\|\boldsymbol{\theta}_j\|)\Big],$$

where we don't include θ_0 in the SCAD penalty and $\mathrm{SCAD}_\lambda(\cdot)$ is the SCAD penalty satisfying

$$\frac{d}{du}\mathrm{SCAD}_\lambda(u) = \mathrm{sign}(u)\Big(\lambda I(|u| \le \lambda) + \frac{(a\lambda - |u|)\vee 0}{a-1}I(|u| > \lambda)\Big),$$

where $\mathrm{SCAD}_\lambda(0) = 0$, $a > 2$, and $a = 3.7$ is recommended in the literature.

3 Theoretical results

In this section, we present our main results, Theorems 2-3. Theorem 1 is taken from [18]. We make some comments on the proof of Theorem 1 and prove Theorems 2-3 in Section 5. We begin with some notation and then state the assumptions for these theorems.

We define several matrices from $\boldsymbol{W}_i$.

$$\Sigma \coloneqq \mathrm{E}\{\boldsymbol{W}_1\boldsymbol{W}_1^T\} \in \mathbb{R}^{(pm+1)\times(pm+1)} \quad \text{and} \quad \Sigma_j \coloneqq \mathrm{E}\{\boldsymbol{W}_{1j}\boldsymbol{W}_{1j}^T\} \in \mathbb{R}^{m\times m},\ j \in [p].$$

Their sample versions are

$$\widehat{\Sigma} \coloneqq \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_i\boldsymbol{W}_i^T \in \mathbb{R}^{(pm+1)\times(pm+1)} \quad \text{and} \quad \widehat{\Sigma}_j \coloneqq \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_{1j}\boldsymbol{W}_{1j}^T \in \mathbb{R}^{m\times m},\ j \in [p].$$

We need some other matrices for the group SCAD.

$$\begin{aligned}\Sigma_S &\coloneqq \mathrm{E}\{\boldsymbol{W}_{1S}\boldsymbol{W}_{1S}^T\} \in \mathbb{R}^{(s_-m+1)\times(s_-m+1)},\\ \Sigma_{S\cup\{\ell\}} &\coloneqq \mathrm{E}\{\boldsymbol{W}_{1S\cup\{\ell\}}\boldsymbol{W}_{1S\cup\{\ell\}}^T\} \in \mathbb{R}^{(sm+1)\times(sm+1)},\ \ell \in S^c,\end{aligned}$$

where $\boldsymbol{W}_{iS} = (\boldsymbol{W}_{ij}^T)_{j\in S}^T \in \mathbb{R}^{s_-m+1}$, $\boldsymbol{W}_{iS\cup\{\ell\}} = (\boldsymbol{W}_{ij}^T)_{j\in S\cup\{\ell\}}^T \in \mathbb{R}^{sm+1}$, and S^c is the complement of S. We denote their sample versions by $\widehat{\Sigma}_S$ and $\widehat{\Sigma}_{S\cup\{\ell\}}$, respectively.

In addition to (1) and (3), we need the following assumptions. Assumption A1 below is Assumption (D3) in [18] and mainly used in Theorem 1. We define u_i by $u_i = Y_i - Q_i$ and u_i satisfies $\mathrm{P}(u_i \le 0|\boldsymbol{X}_i) = \tau$.

Assumption A1 Let $F_u(u|\boldsymbol{X}_1)$ be the conditional distribution function of u_1 w.r.t. $\boldsymbol{X}_1$. It has a continuously differentiable density $f_u(u|\boldsymbol{X}_1)$ and the density satisfies

$$C_{fL} \le f_u(u|\boldsymbol{X}_1) \le C_{fU} \text{ on } [-c_f, c_f] \text{ and } |f_u'(u|\boldsymbol{X}_1)| \le C_{fU}'$$

for some positive C_{fL}, C_{fU}, c_f, and C_{fU}'.

The next one is about $g_j(x_j)$ and $h_j(x_j)$ in (1) and (3). It is actually Assumption (D4) in [18].

Assumption A2 $g_j(x_j)$ and $h_j(x_j)$, $j \in S_-$, are twice continuously differentiable, namely $\nu = 2$, and they also satisfy

$$\sum_{j\in S_-}(\|g\|_{L_\infty} + \|g'\|_{L_\infty} + \|g''\|_{L_\infty}) < C_g \text{ and } \sum_{j\in S_-}(\|h\|_{L_\infty} + \|h'\|_{L_\infty} + \|h''\|_{L_\infty}) < C_h$$

for some positive C_g and C_h. Then we take a suitable $\Psi(x_j)$ with $m = c_m n^{1/(2\nu+1)} = c_m n^{1/5}$ for some positive c_m.

Assumption A2 allows s to increase. Besides, the assumptions on η_i and ξ_i in (5) and (6) hold when we use the basis constructed from the B-spline basis on $[0, 1]$ as in

[17]. See Corollary 6.26 in [24] about the spline function approximation. The basis also meets the next assumption, which is Assumption (D5) in [18]. Assumption A4(1) also holds if $C_1 < f_j(x_j) < C_2$ uniformly in $j \in [p]$ for some positive C_1 and C_2, where $f_j(x_j)$ is the marginal density of X_{1j} as in (4).

Assumption A3 $\max_{i,j} \|\boldsymbol{W}_{ij}\| = O(m^{1/2})$ and $\max_{i,j,k} |W_{ijk}| \leq W_{\max} = O(m^{1/2})$. This $W_{\max}$ is not a random variable.

Assumptions A4(1) and A5 are Assumptions (D7) and (D8) in [18], respectively. They are standard ones in the high-dimensional literature. Assumption A4(2) is used in Theorem 3 and more restrictive than Assumption A4(1). However, we need to have $C_1 < \lambda_{\min}(\Sigma_S) \leq \lambda_{\max}(\Sigma_S) < C_2$ for statistical inference of the true model.

Assumption A4

(1) For some positive C_L and C_U,

$$C_L \leq \lambda_{\min}(\Sigma_j) \leq \lambda_{\max}(\Sigma_j) \leq C_U \text{ for all } j \in [p].$$

(2) For some positive C'_L and C'_U,

$$C'_L \leq \lambda_{\min}(\Sigma_{S\cup\{\ell\}}) \leq \lambda_{\max}(\Sigma_{S\cup\{\ell\}}) \leq C'_U \text{ for all } \ell \in S^c.$$

Before Assumption A5, we define $\mathbb{C}$, $\phi_{\min}$, and $\phi_{\max}$, which are also common in the Lasso literature. Assumption A5 is Assumption (D8) in [18].

$$\mathbb{C} = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^{pm+1} \middle| \sum_{j\in S^c} w_j \|\boldsymbol{\alpha}_j\| \leq c_0 \sum_{j\in S} w_j \|\boldsymbol{\alpha}_j\| \right\}$$

for some c_0 larger than 3 chosen in [18] and $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T)^T$. Any c_0 larger than 3 works well.

$$\phi_{\min} = \min_{\boldsymbol{\alpha}\in\mathbb{C}} \frac{\boldsymbol{\alpha}^T \Sigma \boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|^2} \leq \max_{\boldsymbol{\alpha}\in\mathbb{C}} \frac{\boldsymbol{\alpha}^T \Sigma \boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|^2} = \phi_{\max}$$

Assumption A5 For some positive $C_{\phi L}$ and $C_{\phi U}$, we have

$$C_{\phi L} \leq \phi_{\min} \leq \phi_{\max} \leq C_{\phi U}.$$

The next assumption is the latter half of Condition 4.1 in [31] and a standard one in the least squares regression Lasso literature. However, it excludes heavy tailed cases. This is because the least squares method is applied in estimating the conditional ES in this paper.

Recalling that $u_i = Y_i - Q_i$, we define ϵ_i by

$$\epsilon_i \coloneqq u_i \wedge 0 - \mathrm{E}\{u_i \wedge 0 | \boldsymbol{X}_i\} = u_i \wedge 0 - \tau S_i. \tag{9}$$

Note that this ϵ_i is the error term in the least squares regression of the second step.

Assumption A6 For some positive σ_u and B_u, we have

$$\mathrm{E}\{\epsilon_1^2 | \boldsymbol{X}_1\} \le \sigma_u^2 \quad \text{and} \quad \mathrm{E}\{|\epsilon_1|^k | \boldsymbol{X}_1\} \le \frac{k! \sigma_u^2 B_u^{k-2}}{2}, \ k \ge 3.$$

The next assumption is about the order of s and p. The first one assures $|\boldsymbol{W}_i^T \boldsymbol{\delta}| \to 0$ since

$$|\boldsymbol{W}_i^T \boldsymbol{\delta}| \le C(1 + c_0)\sqrt{ms}\|\boldsymbol{\delta}\|$$

for $\boldsymbol{\delta} \in \mathbb{C}$ such that $\|\boldsymbol{\delta}\| \le t_n \sqrt{sm(1 + \log p/m)/n}$, where t_n appears in Theorem 1 and note that any t_n going to ∞ sufficiently slowly works in Theorem 1. Therefore we don't need any assumptions related to the third moment of W_{ijk}. Such assumptions are seen in [31]. The second one is used in the group Lasso and the last one is employed in the group SCAD. Recall $m = c_m n^{1/5}$ in Assumption A2. The latter half of the last one is necessary to check the local optimality condition of the oracle estimator. The second and third ones are more restrictive than the first one.

Assumption A7

(1)
$$\frac{ms}{\sqrt{n}}\Big(1 + \frac{\log p}{m}\Big)^{1/2} \to 0$$

(2)
$$\sqrt{\frac{m^3 s^2}{n \log p}} \frac{\log p + m}{m} \to 0$$

(3)
$$\sqrt{\frac{m^3 s^3}{n}} \frac{\log p + m}{m} \to 0 \text{ and } \frac{\min_{j \in S_-} \|h_j\|_{L_2}}{(m \log p/n)^{1/2}} \to \infty$$

We state the theoretical results for the group Lasso for $\mathcal{Q}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in [18]. Actually Theorem 1 is Corollary 3.1(i) in [18]. We give some comments on the assumptions of this theorem in Section 5.

Theorem 1. *Suppose Assumptions A1, A2, A3, A4, A5, and A7(1) hold. If we take*

$t_n \to \infty$ *satisfying*

$$t_n\sqrt{\frac{(ms)^2}{n}\Big(1+\frac{\log p}{m}\Big)} \to 0 \text{ and set } \lambda_1 = \frac{t_n}{\sqrt{n}}\Big(1+\sqrt{\frac{\log p}{m}}\Big),$$

then we have

$$\|\widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}\| \le C_\beta t_n\sqrt{\frac{ms}{n}\Big(1+\frac{\log p}{m}\Big)} \text{ and } \widehat{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}} \in \mathbb{C}$$

with probability tending to 1 for some sufficiently large C_β.

Hereafter we assume a suitable $t_n \to \infty$ sufficiently slowly is chosen and that it satisfies the assumption of Theorem 1. We also assume that this t_n satisfies

$$t_n\sqrt{\frac{m^3s^2}{n\log p}}\frac{\log p + m}{m} \to 0 \quad \text{and} \quad t_n\sqrt{\frac{m^3s^3}{n}}\frac{\log p + m}{m} \to 0 \tag{10}$$

in addition to Assumptions A7(2) and A7(3), respectively.

Before we state our main results Theorems 2-3, we define a few subsets related to Theorem 1 and matrices like $\widehat{\Sigma}$ and $\widehat{\Sigma}_j$.

$$\begin{aligned}
\Omega_0 &\coloneqq \Big\{\frac{C_L}{2} \le \lambda_{\min}(\widehat{\Sigma}_j) \le \lambda_{\max}(\widehat{\Sigma}_j) \le 2C_U \text{ for all } j \in [p]\Big\} \\
&\qquad \cap \Big\{\frac{C_{\phi L}}{2} \le \min_{\boldsymbol{\alpha}\in\mathbb{C}} \frac{\boldsymbol{\alpha}^T\widehat{\Sigma}\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|^2} \le \max_{\boldsymbol{\alpha}\in\mathbb{C}} \frac{\boldsymbol{\alpha}^T\widehat{\Sigma}\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|^2} \le 2C_{\phi U}\Big\}, \\
\Omega_1 &\coloneqq \Omega_0 \cap \{\widehat{\beta} - \overline{\beta} \in \mathbb{C} \text{ and } \|\widehat{\beta} - \overline{\beta}\| \le r_Q\}, \qquad (11) \\
\Omega_2 &\coloneqq \Omega_1 \cap \Omega_0', \qquad (12)
\end{aligned}$$

where

$$\begin{aligned}
\Omega_0' &\coloneqq \Big\{\frac{C_L'}{2} \le \lambda_{\min}(\widehat{\Sigma}_{S\cup\{\ell\}}) \le \lambda_{\max}(\widehat{\Sigma}_{S\cup\{\ell\}}) \le 2C_U' \text{ for all } \ell \in S^c.\Big\}, \\
r_Q &\coloneqq C_\beta t_n\sqrt{\frac{ms}{n}\Big(1+\frac{\log p}{m}\Big)} \text{ from Theorem 1.} \qquad (13)
\end{aligned}$$

As in Lemma 1 in Section 5, $\mathrm{P}(\Omega_k) \to 1$, $k = 1, 2$, under Assumption A8 if the results in Theorem 1 hold.

Assumption A8

$$\frac{m^3s^2\log p}{n} \to 0.$$

We state the theoretical results for the group Lasso for $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in Theorem 2 and

prove Theorem 2 in Section 5. Theorem 2 may be a group Lasso version of Theorem 4.1 in [31]. However, there are some significant differences between them.

1. We don't use any assumptions on the third or fourth moment on W_{ijk}. Besides, we don't assume that Σ is positive definite.

2. The treatment of the group Lasso is not trivial and needs suitable technical tools for the group structure. Besides, we have to cope with approximation errors η_i in (5) and ξ_i in (6). We also have to check the relations between n, s, and m very carefully. We consider Lemma 2 in a general form while addressing the group structure. Therefore our lemma 2 is exploited in the proof of Theorem 3 for the group SCAD estimator.

3. We don't introduce a convex set as in (S.1.3) in [31].

Theorem 2. *Suppose Assumptions A1, A2, A3, A4, A5, A6, A7(2) and A8 hold. If we set* $\lambda_2 = D_0\sqrt{\log p/n}$*, where* D_0 *depends on the assumptions and* D_1 *and* D_2 *below, then we have*

$$\|\widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}}\| \leq \frac{3\lambda_2(ms)^{1/2}}{\tau\phi_{\min}} \text{ and } \widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}} \in \mathbb{C}$$

with probability larger than $\mathrm{P}(\Omega_1) - n^{-D_1} - \exp\{-D_2(\log p + m)\}$*. We can take any positive* D_1 *and* D_2*.*

Theorem 2 implies that

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{W}_i^T\widehat{\boldsymbol{\theta}} - S_i)^2 = O_p\Big(n^{-4/5} + \frac{ms\log p}{n}\Big) = O_p(n^{-4/5}s\log p).$$

Our group Lasso estimator satisfies the oracle inequality and works as an initial value of other methods like the group SCAD, the adaptive group Lasso, and so on.

We state the theoretical results for the group SCAD for $\mathcal{S}_\tau(Y_i \,|\, \boldsymbol{X}_i)$ in Theorem 3 and prove the theorem in Section 5. Before the theorem, we define the oracle estimator :

$$\widetilde{\boldsymbol{\theta}}_S \coloneqq \Big(\frac{\tau^2}{n}\sum_{i=1}^{n}\boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T\Big)^{-1}\frac{\tau}{n}\sum_{i=1}^{n}\boldsymbol{W}_{iS}Z_i(\widehat{\boldsymbol{\beta}})$$

and the oracle estimator with Q_i given :

$$\breve{\boldsymbol{\theta}}_S \coloneqq \Big(\frac{\tau^2}{n}\sum_{i=1}^{n}\boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T\Big)^{-1}\frac{\tau}{n}\sum_{i=1}^{n}\boldsymbol{W}_{iS}Z_i(Q_i).$$

The difference between λ_2 and λ_3 in Theorems 2-3 is just due to the weight w_j in the group Lasso penalty.

Theorem 3. *Suppose Assumptions A1, A2, A3, A4, A5, A6, A7(3), and A8 hold. If we set* $\lambda_3 = D_0\sqrt{m \log p/n}$, *where* D_0 *and* D_1 *depend on the assumptions and* D_2 *and* D_3 *below, then* $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\theta}}_S^T, \mathbf{0}_{S^c}^T)^T$ *is a local solution to our group SCAD objective function and*

$$\|\widetilde{\boldsymbol{\theta}}_j - \breve{\boldsymbol{\theta}}_j\| = o(\sqrt{m/n}) \tag{14}$$

uniformly in $j \in S_-$ *with probability larger than* $\mathrm{P}(\Omega_2) - n^{-D_2} - \exp\{-D_3(\log p + m)\}$. *We can take any positive* D_2 *and* D_3.

The latter half of Theorem 3 suggests that estimation of the conditional quantile does not affect estimation of the conditional ES regression model asymptotically. The order $o(1)$ of the latter in (14) comes from the latter in (10).

Remark 1. *Suppose that we are interested in* $\tau = a, b (a < b)$ *and our assumptions hold for* $\tau = a, b$. *Write* $S_{ia}, Q_{ia}, Z_{ia}(Q_{ia})$ *for* $S_i, Q_i, Z_i(Q_i)$ *for* $\tau = a$ *and* $S_{ib}, Q_{ib}, Z_{ib}(Q_{ib})$ *for* $S_i, Q_i, Z_i(Q_i)$ *for* $\tau = b$, *respectively. Then*

$$\mathrm{E}\{Z_{1b}(Q_{1b}) - Z_{1a}(Q_{1a})|\boldsymbol{X}\} = \mathrm{E}\{Y_1 I(Q_{1a} < Y_1 \le Q_{1b})|\boldsymbol{X}_1\}.$$

When we estimate $\mathrm{E}\{Y_1 I(Q_{1a} < Y \le Q_{1b})|\boldsymbol{X}_1\}/(b-a)$, *we should appeal to the least squares method from* $Z_{ib}(\widehat{\boldsymbol{\beta}}_b) - Z_{ia}(\widehat{\boldsymbol{\beta}}_a)$ *to* $\boldsymbol{W}_i$. *When we prove the theoretical results, we should deal with* $Z_{ib}(\widehat{\boldsymbol{\beta}}_b)$ *and* $Z_{ia}(\widehat{\boldsymbol{\beta}}_a)$ *separately to prove the so-called deviation conditions and obtain the results like Theorems 2-3. We may be able to relax Assumption A6 and it is a topic of future research.*

4 Numerical studies

4.1 Simulation Studies

In this subsection, we investigate the finite-sample performance of the proposed method. We begin by detailing how the observed features X'_{ij} are transformed into the covariate vectors $\boldsymbol{W}_{ij}$. Next, we describe the process by which new features $X'_{\mathrm{new},j}$ are mapped to $\boldsymbol{W}_{\mathrm{new},j}$. We then present the algorithm and its implementation in R, including the procedures for tuning-parameter selection. These procedures are used to solve the underlying optimization problem, yielding $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\theta}}$, and $\widetilde{\boldsymbol{\theta}}$. To further refine these estimates,

we introduce a re-fit mechanism. Finally, we examine two simulation scenarios in which the features follow different distributions. We evaluate the performance of our approach in terms of estimation accuracy, prediction error, and model-selection consistency in each of these scenarios.

We demonstrate the proposed method using simulated features drawn from a Beta distribution. In cases where the observed data X'_{ij} lie outside the unit interval $[0, 1]$, we first apply min-max normalization to rescale them to the unit interval. Specifically, we define

$$X_{ij} = \frac{X'_{ij} - m_j}{M_j - m_j},$$

where $M_j = \max_{1 \le i \le n} X'_{ij}$ and $m_j = \min_{1 \le i \le n} X'_{ij}$. To construct the covariate vector $\boldsymbol{W}_{ij}$, we first generate the B-spline basis matrix

$$\boldsymbol{B}_j(\boldsymbol{\chi}_j) = \Big(\boldsymbol{B}_{j0}(\boldsymbol{\chi}_j),\, \boldsymbol{B}_{j1}(\boldsymbol{\chi}_j),\, \ldots,\, \boldsymbol{B}_{jm}(\boldsymbol{\chi}_j)\Big) \in \mathbb{R}^{n \times (m+1)},$$

where $\boldsymbol{\chi}_j = (X_{1j}, \ldots, X_{nj})^T$. The B-spline basis is of degree 2, includes an intercept term, and uses knots determined by equally spaced empirical quantiles of $\boldsymbol{\chi}_j$. The function `bs` from the `splines` package in R can be employed to generate $\boldsymbol{B}_j(\boldsymbol{\chi}_j)$. Next, we construct the matrix $\boldsymbol{\Omega}_j(\boldsymbol{\chi}_j) \in \mathbb{R}^{n \times (m+1)}$, whose columns are given by

$$\boldsymbol{\Omega}_j(\boldsymbol{\chi}_j) = \Big[\mathbf{1}_n,\, \boldsymbol{\chi}_j,\, \boldsymbol{B}_{j1}(\boldsymbol{\chi}_j),\, \ldots,\, \boldsymbol{B}_{j,m-1}(\boldsymbol{\chi}_j)\Big],$$

where $\mathbf{1}_n$ is a vector of ones. We then apply a "scaled" QR decomposition to $\boldsymbol{\Omega}_j(\boldsymbol{\chi}_j)$, such that $\boldsymbol{Q}_j \boldsymbol{R}_j = \boldsymbol{\Omega}_j(\boldsymbol{\chi}_j)$ and $\boldsymbol{Q}_j^T \boldsymbol{Q}_j = (n/m)\, \boldsymbol{I}_{m+1}$, where $\boldsymbol{I}_{m+1}$ is the identity matrix. Finally, we define the covariate vector $\boldsymbol{W}_{ij}$ as the i-th row of the matrix $\boldsymbol{Q}_j$, excluding the first element that corresponds to the intercept. For our simulation experiments, we set $m = \lceil n^{1/5} \rceil$.

For new observations $X'_{\text{new},j}$, we construct $\boldsymbol{W}_{\text{new},j}$ using the previously determined M_j, m_j, $\boldsymbol{B}_j(\cdot)$, and $\boldsymbol{R}_j$ to maintain consistency with the training procedure. Specifically, we first apply the transformation

$$X_{\text{new},j} = \max\Big(0, \min\Big(1, \frac{X'_{\text{new},j} - m_j}{M_j - m_j}\Big)\Big),$$

which ensures $X_{\text{new},j}$ is constrained to $[0, 1]$. We then form the new covariate vector by computing $\boldsymbol{\Omega}_j(X_{\text{new},j})\, \boldsymbol{R}_j^{-1}$ and excluding the first element to obtain $\boldsymbol{W}_{\text{new},j}$.

To solve the quantile estimator $\widehat{\boldsymbol{\beta}}$, we formulate the optimization problem (7) as

follows:

$$\begin{aligned}
\min_{\boldsymbol{\beta},\boldsymbol{s},\boldsymbol{\eta}^+,\boldsymbol{\eta}^-} \quad & \frac{\tau}{n}\sum_{i=1}^{n}\eta_i^+ + \frac{1-\tau}{n}\sum_{i=1}^{n}\eta_i^- + \lambda_1\sum_{j=1}^{p} w_j s_j \\
\text{s.t.} \quad & \eta_i^+ - \eta_i^- = Y_i - \boldsymbol{W}_i^T\boldsymbol{\beta}, \quad i = 1,\ldots,n, \\
& \|\boldsymbol{\beta}_j\| \le s_j, \quad j = 1,\ldots,p, \\
& \eta_i^+ \ge 0, \quad i = 1,\ldots,n, \\
& \eta_i^- \ge 0, \quad i = 1,\ldots,n,
\end{aligned}$$

where $\boldsymbol{\eta}^+ = (\eta_1^+,\ldots,\eta_n^+)^T$, $\boldsymbol{\eta}^- = (\eta_1^-,\ldots,\eta_n^-)^T$, and $\boldsymbol{s} = (s_1,\ldots,s_p)^T$. This constitutes a second-order cone programming problem, which we solve using the `Rmosek` package. To determine the appropriate value of λ_1, we follow the procedure outlined in Section 3.3 of [18]. Specifically, define

$$\Lambda = \max_{1\le j\le p}\Big\|\frac{1}{n}\sum_{i=1}^{n}\big(\tau - u_i\big)\boldsymbol{W}_{ij}\Big\|,$$

where u_i are i.i.d. Bernoulli random variables with probability τ. Let $\Lambda(1-\theta|\{\boldsymbol{X}_\ell\})$ denote the conditional $(1-\theta)$-quantile of Λ given $\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\}$. The value of λ_1 is then chosen as $c\Lambda(1-\theta|\{\boldsymbol{X}_\ell\})$ for $\theta \in (0,1)$ and $c > 0$. In the subsequent numerical studies, (c,θ) is set to $(1.15, 0.05)$, and $\Lambda(1-\theta|\{\boldsymbol{X}_\ell\})$ is estimated using 1,000 simulated observations. Once the coefficients are estimated, the model is then re-fitted using only active features, defined as those for which $\|\widehat{\boldsymbol{\beta}}_j\| \ge 10^{-6}$, to further refine the estimation.

For the expected shortfall estimators $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$, we use the `grpreg` package to carry out the estimation. Additionally, a 2-fold cross-validation procedure is applied to determine the appropriate values for the regularization parameters λ_2 and λ_3. Similarly, the model is re-fitted to further refine the estimation using only active features, defined as those with $\|\widehat{\boldsymbol{\theta}}_j\| > 0$ and $\|\widetilde{\boldsymbol{\theta}}_j\| > 0$.

To evaluate performance, we consider four criteria: estimation error (`est`), prediction error (`pre`), the number of true positives (`tp`), and the number of false positives (`fp`). We report these measures for both quantile regression and expected shortfall regression, denoted by subscripts `Q` and `ES`, respectively. Each criterion is computed as the average over 100 repeated experiments. We quantify both estimation and prediction errors using

the relative ℓ_2 error, defined by

$$\mathrm{RelErr}_{\ell_2}(\widehat{\boldsymbol{v}}, \boldsymbol{v}) = \frac{\|\widehat{\boldsymbol{v}} - \boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2},$$

where $\widehat{\boldsymbol{v}}$ denotes the fitted or predicted values, and $\boldsymbol{v}$ denotes the true values. For example, the estimation error for the group SCAD estimator is

$$\mathrm{RelErr}_{\ell_2}\Big(\Big(\boldsymbol{W}_1^T\widetilde{\boldsymbol{\theta}}, \ldots, \boldsymbol{W}_n^T\widetilde{\boldsymbol{\theta}}\Big)^T, \Big(S_\tau(Y_1 \mid \boldsymbol{X}_1), \ldots, S_\tau(Y_n \mid \boldsymbol{X}_n)\Big)^T\Big).$$

Finally, we refer to the method employing group lasso in the second stage as `es-lasso`, and the method employing group SCAD as `es-scad`. An oracle estimator that uses only the active features and is computed via the `esreg` package is referred to as `benchmark`.

Case 1

We adopt the following data generation process for Case 1. For $i = 1, \ldots, n$, the response variable Y_i is generated as

$$Y_i = 3\sin(2\pi X_{i1}) + 3\cos(2\pi X_{i2}) + 3\exp(X_{i3}) + \Big(\exp(-X_{i1}) + X_{i2}^2 + X_{i3}^3\Big)\epsilon_i,$$

where X_{ij}, for $j = 1, \ldots, p$, are independent and uniformly distributed on $[0, 1]$, and ϵ_i are independent standard normal random variables. Under this setting, the signal-to-noise ratio is 6.658. We focus on three values of τ, specifically 0.05, 0.1, and 0.2, while fixing the number of features at $p = 1000$. We then determine two sample sizes, $n = \lceil 90/\tau \rceil$ and $n = \lceil 150/\tau \rceil$, in accordance with the design described in [31] when $s = 3$ in their notation.

Table 1 presents the performance of the `es-lasso`, `es-scad`, and `benchmark` methods, evaluated using `est`, `pre`, `tp`, and `fp`. We first observe that the quantile estimator tends to omit active features when $n = \lceil 90/\tau \rceil$ and $\tau = 0.2$. However, this phenomenon disappears once the sample size increases. In all other cases, the quantile estimator performs comparably to the benchmark across all metrics.

Next, we focus on the performance of the expected shortfall regression. When $n = \lceil 90/\tau \rceil$ and $\tau = 0.2$, both the group SCAD and group Lasso estimators tend to select several irrelevant features. In terms of false positives, the group SCAD estimator even performs worse than the Lasso estimator; however, `es-scad` still provides better estimation and prediction accuracy than `es-lasso`. In all other scenarios, the group

Table 1: Performance of `es-lasso`, `es-scad`, and `benchmark` for Case 1.

$n = \lceil 90/\tau \rceil$									
τ	Method	$\texttt{est}_{\texttt{Q}}$	$\texttt{pre}_{\texttt{Q}}$	$\texttt{tp}_{\texttt{Q}}$	$\texttt{fp}_{\texttt{Q}}$	$\texttt{est}_{\texttt{ES}}$	$\texttt{pred}_{\texttt{ES}}$	$\texttt{tp}_{\texttt{ES}}$	$\texttt{fp}_{\texttt{ES}}$
0.05	`es-lasso`	0.0580	0.0583	3.000	0.020	0.1170	0.1180	3.000	0.93
	`es-scad`	0.0580	0.0583	3.000	0.020	0.0950	0.0951	3.000	0.23
	`benchmark`	0.0576	0.0579	3.000	0.000	0.0913	0.0913	3.000	0.00
0.10	`es-lasso`	0.0733	0.0740	3.000	0.000	0.1530	0.1540	3.000	2.54
	`es-scad`	0.0733	0.0740	3.000	0.000	0.1120	0.1140	3.000	0.65
	`benchmark`	0.0733	0.0741	3.000	0.000	0.1020	0.1040	3.000	0.00
0.20	`es-lasso`	0.1150	0.1160	2.800	0.000	0.2100	0.2120	3.000	6.62
	`es-scad`	0.1150	0.1160	2.800	0.000	0.1870	0.1900	3.000	6.85
	`benchmark`	0.0744	0.0757	3.000	0.000	0.1040	0.1060	3.000	0.00
$n = \lceil 150/\tau \rceil$									
τ	Method	$\texttt{est}_{\texttt{Q}}$	$\texttt{pre}_{\texttt{Q}}$	$\texttt{tp}_{\texttt{Q}}$	$\texttt{fp}_{\texttt{Q}}$	$\texttt{est}_{\texttt{ES}}$	$\texttt{pred}_{\texttt{ES}}$	$\texttt{tp}_{\texttt{ES}}$	$\texttt{fp}_{\texttt{ES}}$
0.05	`es-lasso`	0.0501	0.0501	3.000	0.680	0.0830	0.0831	3.000	0.58
	`es-scad`	0.0501	0.0501	3.000	0.680	0.0693	0.0694	3.000	0.09
	`benchmark`	0.0470	0.0470	3.000	0.000	0.0715	0.0715	3.000	0.00
0.10	`es-lasso`	0.0480	0.0485	3.000	0.000	0.0958	0.0967	3.000	1.18
	`es-scad`	0.0480	0.0485	3.000	0.000	0.0741	0.0749	3.000	0.23
	`benchmark`	0.0477	0.0482	3.000	0.000	0.0722	0.0727	3.000	0.00
0.20	`es-lasso`	0.0634	0.0639	3.000	0.000	0.1170	0.1180	3.000	2.12
	`es-scad`	0.0634	0.0639	3.000	0.000	0.0925	0.0924	3.000	0.81
	`benchmark`	0.0632	0.0636	3.000	0.000	0.0868	0.0867	3.000	0.00

SCAD estimator achieves better performance across all metrics. In fact, for these cases, the performance of `es-scad` is close to the `benchmark`, demonstrating the effectiveness of the proposed method.

Case 2

We use the same setup as in Case 1, except for the distribution of X_{ij}. Specifically, we draw each X_{ij} from an independent Beta distribution with shape parameters $\alpha = \beta = 3$. Compared to Case 1, the mean of the features remains unchanged, while the variance decreases from 1/12 to 1/28.

Table 2 exhibits a pattern similar to the results observed in Case 1. When $n = \lceil 90/\tau \rceil$ and $\tau = 0.2$, the quantile estimator tends to omit relevant features, leading to

Table 2: Performance of `es-lasso`, `es-scad`, and `benchmark` for Case 2.

$n = \lceil 90/\tau \rceil$									
τ	Method	$\texttt{est}_{\texttt{Q}}$	$\texttt{pre}_{\texttt{Q}}$	$\texttt{tp}_{\texttt{Q}}$	$\texttt{fp}_{\texttt{Q}}$	$\texttt{est}_{\texttt{ES}}$	$\texttt{pred}_{\texttt{ES}}$	$\texttt{tp}_{\texttt{ES}}$	$\texttt{fp}_{\texttt{ES}}$
0.05	`es-lasso`	0.0819	0.0826	2.95	0.03	0.1860	0.1870	3.00	1.86
	`es-scad`	0.0819	0.0826	2.95	0.03	0.1360	0.1380	2.99	0.54
	`benchmark`	0.0710	0.0715	3.00	0.00	0.1070	0.1070	3.00	0.00
0.10	`es-lasso`	0.0741	0.0743	2.98	0.00	0.1990	0.1960	3.00	3.35
	`es-scad`	0.0741	0.0743	2.98	0.00	0.1300	0.1290	2.99	0.84
	`benchmark`	0.0722	0.0725	3.00	0.00	0.1060	0.1060	3.00	0.00
0.20	`es-lasso`	0.2120	0.2150	2.22	0.00	0.2980	0.2960	3.00	8.38
	`es-scad`	0.2120	0.2150	2.22	0.00	0.2530	0.2530	3.00	6.36
	`benchmark`	0.0787	0.0800	3.00	0.00	0.1190	0.1210	3.00	0.00
$n = \lceil 150/\tau \rceil$									
τ	Method	$\texttt{est}_{\texttt{Q}}$	$\texttt{pre}_{\texttt{Q}}$	$\texttt{tp}_{\texttt{Q}}$	$\texttt{fp}_{\texttt{Q}}$	$\texttt{est}_{\texttt{ES}}$	$\texttt{pred}_{\texttt{ES}}$	$\texttt{tp}_{\texttt{ES}}$	$\texttt{fp}_{\texttt{ES}}$
0.05	`es-lasso`	0.0572	0.0572	3.00	0.01	0.1270	0.1270	3.00	1.43
	`es-scad`	0.0572	0.0572	3.00	0.01	0.0950	0.0952	3.00	0.39
	`benchmark`	0.0571	0.0571	3.00	0.00	0.0854	0.0851	3.00	0.00
0.10	`es-lasso`	0.0585	0.0587	3.00	0.00	0.1310	0.1310	3.00	1.80
	`es-scad`	0.0585	0.0587	3.00	0.00	0.0988	0.0993	3.00	0.54
	`benchmark`	0.0585	0.0587	3.00	0.00	0.0885	0.0893	3.00	0.00
0.20	`es-lasso`	0.0609	0.0620	3.00	0.00	0.1610	0.1610	3.00	3.87
	`es-scad`	0.0609	0.0620	3.00	0.00	0.1170	0.1190	3.00	1.33
	`benchmark`	0.0624	0.0636	3.00	0.00	0.0920	0.0930	3.00	0.00

poor estimation and prediction. In all other cases, however, its performance remains satisfactory. Focusing on the expected shortfall regression, the group SCAD estimator outperforms the group Lasso estimator on nearly all metrics. In terms of true positives when $n = \lceil 90/\tau \rceil$, `es-scad` selects 0.01 less active features, which represents only a minor discrepancy. Hence, `es-scad` can be regarded as superior to `es-lasso`. Compared to Case 1, the group SCAD estimator's performance is not as close to the benchmark. Nonetheless, except for the scenario where $n = \lceil 90/\tau \rceil$ and $\tau = 0.2$, its performance remains satisfactory. This example demonstrates that, regardless of whether the features follow a uniform or a bell-shaped distribution, the proposed method sustains a high-level performance.

4.2 Empirical Study

High-performance concrete (HPC) is widely used in modern civil engineering because of its superior strength, durability, and cost-effectiveness compared to conventional concrete [22]. Accurate prediction of compressive strength is crucial for ensuring structural safety and optimizing material design. However, most existing studies focus on conditional mean estimation [5, 29], potentially overlooking tail risks. In this subsection, we present an illustrative example of our proposed method by constructing an expected shortfall regression model. Rather than developing a formal model or drawing definitive conclusions, our aim is to elucidate the underlying approach. By highlighting the potential applications of our approach, this demonstration paves the way for more comprehensive investigations in future research.

The empirical analysis employs the HPC dataset originally compiled by [29]. After excluding observations with zero values, 225 samples remain, each containing eight predictor variables: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age. The target variable is concrete compressive strength, measured in megapascals (MPa). Prior research indicates that the relationship between these mixture components and compressive strength is inherently nonlinear. This nonlinearity limits the applicability of traditional expected shortfall regression, thereby underscoring the need for an additive expected shortfall regression model to effectively capture these complex effects.

To assess the performance of our proposed high-dimensional additive expected shortfall regression method in more challenging settings, we artificially inflate the dimensionality by adding extra features drawn independently from a standard normal distribution. This approach tests the method's ability to distinguish between relevant variables and noise as the number of predictors increases. Specifically, we evaluate the method at quantile levels 0.2, 0.3, and 0.4. At each quantile level, the method is tested on three datasets augmented to contain 200, 300, and 400 predictors. We anticipate that our method will consistently identify the truly relevant predictors while effectively excluding the noise variables. Implementation details match those used in the simulation study. In particular, we set the number of basis functions to $m = \lceil n^{1/5} \rceil = 3$. The penalty parameter λ_1 is chosen as $1.15 \cdot \Lambda(0.95 \mid \{\boldsymbol{X}_\ell\})$ based on 1,000 simulated observations, and λ_2 and λ_3 are selected via two-fold cross-validation.

Table 3 summarizes the selection results for `es-lasso` and `es-scad`. The "Indices" column indicates which of the original eight predictors were selected, and the "Noise"

column shows the number of artificially added variables that were falsely selected. For quantile regression, no spurious variables entered the model. Regarding expected shortfall, `es-lasso` generally selects more irrelevant features than `es-scad`, mirroring the pattern observed in our simulation studies. One exception occurs at $(\tau, p) = (0.2, 200)$, where the group SCAD estimator selects as many irrelevant features as the group LASSO estimator. This result is expected, as lower quantile levels require more data for accurate estimation. Furthermore, the selection of the true predictors remains stable across the three datasets, reinforcing the reliability of the proposed method. Although this example is merely illustrative, it demonstrates that by incorporating additional features, our approach can effectively exclude irrelevant variables and pave the way for more comprehensive analyses in future research.

Table 3: Selection results on HPC data.

τ	p	Method	Quantile		Expected Shortfall	
			Indices	Noise	Indices	Noise
0.2	200	`es-lasso`	{8}	0	{1,2,3,4,5,8}	10
		`es-scad`	{8}	0	{1,2,3,4,5,8}	10
	300	`es-lasso`	{8}	0	{4,5,8}	1
		`es-scad`	{8}	0	{4,5,8}	1
	400	`es-lasso`	{8}	0	{1,2,3,4,5,8}	17
		`es-scad`	{8}	0	{1,3,4,5,8}	7
0.3	200	`es-lasso`	{1,8}	0	{1,3,4,5,8}	2
		`es-scad`	{1,8}	0	{1,3,4,5,8}	1
	300	`es-lasso`	{1,8}	0	{1,3,4,5,8}	8
		`es-scad`	{1,8}	0	{1,3,4,5,8}	1
	400	`es-lasso`	{1,8}	0	{1,3,4,5,8}	8
		`es-scad`	{1,8}	0	{1,3,4,5,8}	1
0.4	200	`es-lasso`	{1,8}	0	{1,3,4,5,6,8}	1
		`es-scad`	{1,8}	0	{1,3,4,5,6,8}	1
	300	`es-lasso`	{1,4,8}	0	{1,3,4,5,6,8}	10
		`es-scad`	{1,4,8}	0	{1,2,3,4,5,6,8}	1
	400	`es-lasso`	{1,8}	0	{1,3,4,5,6,8}	10
		`es-scad`	{1,8}	0	{1,3,4,5,6,8}	5

5 Proofs

We give the proofs of the theorems and some necessary lemmas in this section. We prove those lemmas in the Appendix. We begin with the comments on the proof of Theorem 1.

Comments on the proof Theorem 1. In [18], the author considered additive quantile regression models in Sction 4 and proved the theoretical results by applying Theorem 3.1 and Corollary 3.1. Assumptions (D1) and (D2) hold in our setup and Assumptions (D3), (D5), (D7), and (D8) are also assumed in our Theorem 1. Besides, Assumption A2 and the condition on m, s, and $\log p$ in Theorem 1 cover Assumption (D9). Our Assumption A2 assures Assumptions (D4) and (D6) there. In Section 4 in [18], the author assumed s was bounded just to assure the order of the approximation error in (1), η_i. Since we use Assumption A2, the boundedness of s is not necessary. Thus all the assumptions for quantile estimation are satisfied. □

In Lemma 1 below, we prove that $\mathrm{P}(\Omega_k) \to 1$, $k = 1, 2$, if the results in Theorem 1 hold. We verify this lemma in the Appendix.

Lemma 1. *If Assumptions A3, A4(1)-(2), A5, and A8 hold, we have that*

$$\mathrm{P}(\Omega_0) \to 1 \quad \textit{and} \quad \mathrm{P}(\Omega_0') \to 1.$$

In Lemmas 2-4, we deal with the so-called deviation condition in the Lasso literature. We need some notation for these lemmas.

Suppose that we have $\boldsymbol{V}_{ij} = (V_{ij1}, \ldots, V_{ijm})^T \in \mathbb{R}^m$, $i = 1, \ldots, n$ and $j \in \mathcal{A} \subset [p]$, where $\boldsymbol{V}_{ij}$ depends on $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. $\boldsymbol{W}_{ij}$ is an example of $\boldsymbol{V}_{ij}$ and we actually apply these lemmas to $\boldsymbol{W}_{ij}$ in the proof of Theorem 2. In the proofs of Theorems 2-3, we evaluate

$$\sup_{j \in \mathcal{A}} \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_{ij} \{ Z_i(\widehat{\boldsymbol{\beta}}) - \tau \boldsymbol{W}_i^T \overline{\boldsymbol{\theta}} \} \right\| \tag{15}$$

by using Lemmas 2-4. Note that $0 \notin \mathcal{A}$. For $j = 0$, we use these lemmas only with $V_{i0} \equiv 1$. Then Lemma 2 holds without ξ_i since $\sum_{i=1}^n \xi_i = 0$ and Lemmas 3 and 4 clearly hold with $\sqrt{m \log p/n}$ replaced by $\sqrt{\log p/n}$.

We define

$$\Omega_\omega := \Omega_0 \cap \{ \lambda_{\max}(\widehat{\Omega}_j) \leq \ C_{\omega U} \text{ for all } j \in \mathcal{A} \},$$

where $\widehat{\Omega}_j = n^{-1}\sum_{i=1}^n \boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T \in \mathbb{R}^{m\times m}$ and $C_{\omega U}$ is some positive constant. Besides, let $\max_{i,j,k}|V_{ijk}| \le V_{\max}$ and $V_{\max}$ is not a random variable and may go to infinity.

To evaluate (15), we should deal with

$$B_1 \coloneqq \sup_{j\in\mathcal{A}} \sup_{\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\in B_{\mathbb{C}}(r_Q)} \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{V}_{ij}\{Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}})\} \right\|, \tag{16}$$

where $B_{\mathbb{C}}(r_Q) = \mathbb{C} \cap \{\|\boldsymbol{\beta} - \overline{\boldsymbol{\beta}}\| \le r_Q\}$,

$$B_2 \coloneqq \sup_{j\in\mathcal{A}} \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{V}_{ij}(Z_i(\overline{\boldsymbol{\beta}}) - \mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}]) \right\|, \tag{17}$$

$$B_3 \coloneqq \sup_{j\in\mathcal{A}} \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{V}_{ij}(\mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}] - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}}) \right\|. \tag{18}$$

In Lemma 2, we evaluate B_1 by modifying the proof of Lemma S.1.1 in [31]. The treatment of approximation errors, the special structure of $\boldsymbol{W}_{ij}$, and the group structure is not trivial. Besides, we consider general $\boldsymbol{V}_{ij}$ and apply this more general Lemma 2 to Theorem 3 for the group SCAD estimator.

In Lemmas 2-4 here, $\mathrm{P}_{\Omega_\omega}(\cdot)$ means the probability limited to Ω_ω, not a conditional probability on Ω_ω. It implies that $\mathrm{P}_{\Omega_\omega}(A) = \mathrm{P}(A\cap\Omega_\omega) \le \mathrm{P}(\Omega_\omega)$ for any event A. Note that $V_{\max}\{(\log p + m)/m\}t_n\{m^2s^2/(n\log p)\}^{1/2} \to 0$ for a suitably chosen t_n under Assumption A7(2) or A7(3) as in (10).

Lemma 2. *Suppose that Assumptions A1, A2, A3, A4(1), A5, and A7(1) hold. Then there exists a positive constant D_1 depending on D_2 such that*

$$\mathrm{P}_{\Omega_\omega}\left(B_1 \le D_1\left(V_{\max}\frac{\log p + m}{m}t_n\sqrt{\frac{m^2s^2}{n\log p}}\right)\sqrt{\frac{m\log p}{n}}\right) \ge \mathrm{P}(\Omega_\omega) - \exp\{-D_2(\log p + m)\}.$$

We can choose any positive D_2. Besides note that t_n is from Theorem 1 and r_Q in (13) and that any t_n going to ∞ slowly works well.

In Lemma 3, we consider B_2 by using Assumption A6. This kind of lemma is a standard one in the Lasso literature. The latter halves of Lemmas 3-4 are used in the proof of Theorem 3.

Lemma 3. *Suppose that Assumptions A1, A2, and A6 hold and $V_{\max}(\log p/n)^{1/2} \to 0$.*

Then there exist positive constants D_1 and D_2 depending on D_3 such that

$$B_2 \leq D_1\sqrt{\frac{m\log p}{n}} \quad \textit{and}$$

$$\sup_{j\in\mathcal{A}}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_{ij}\Big(Z_i(\overline{\boldsymbol{\beta}})-\mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}]-\epsilon_i\Big)\right\| \leq D_2 n^{-2/5}\sqrt{\frac{m\log p}{n}}$$

with probability larger than $\mathrm{P}(\Omega_\omega)-n^{-D_3}$. *We can choose any positive* D_3.

In Lemma 4, we deal with the approximation error in B_3.

Lemma 4. *Suppose that Assumptions A1 and A2. Then there exist positive constants D_1 and D_2 such that we have on Ω_ω,*

$$B_3 \leq D_1\Big\{\frac{1}{n}\sum_{i=1}^{n}(\xi_i^2+\eta_i^4)\Big\}^{1/2} \quad \textit{and}$$

$$\sup_{j\in\mathcal{A}}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_{ij}\Big(\mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}]-\tau\boldsymbol{W}^T\overline{\boldsymbol{\theta}}-\tau\xi_i\Big)\right\| \leq D_2\Big(\frac{1}{n}\sum_{i=1}^{n}\eta_i^4\Big)^{1/2}.$$

We repeat that the latter half of Lemma 4 holds without ξ_i on the RHS for $V_{i0}\equiv 1$ since $\sum_{i=1}^{n}\xi_i=0$.

Proof of Theorem 2. We focus on the events included in Ω_1 in the proof. Write

$$LS(\boldsymbol{\theta}) \coloneqq \frac{1}{2n}\sum_{i=1}^{n}(Z_i(\widehat{\boldsymbol{\beta}})-\tau\boldsymbol{W}_i^T\boldsymbol{\theta})^2 \text{ and then}$$
$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}}\Big\{LS(\boldsymbol{\theta})+\tau\lambda_2\sum_{j=1}^{p}w_j\|\boldsymbol{\theta}_j\|\}.$$

Notice the following inequality and expression :

$$LS(\widehat{\boldsymbol{\theta}})+\tau\lambda_2\sum_{j=1}^{p}w_j\|\widehat{\boldsymbol{\theta}}_j\| \leq LS(\overline{\boldsymbol{\theta}})+\tau\lambda_2\sum_{j=1}^{p}w_j\|\overline{\boldsymbol{\theta}}_j\|$$

and

$$LS(\widehat{\boldsymbol{\theta}}) = \frac{1}{2n}\sum_{i=1}^{n}\{(Z_i(\widehat{\boldsymbol{\beta}})-\tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}})+\tau\boldsymbol{W}_i^T(\overline{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}})\}^2.$$

They yield the basic inequality of the Lasso as in [15] :

$$\tau^2(\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}})^T\widehat{\Sigma}(\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}) \le \tau\lambda_2\Big(\sum_{j=1}^p w_j\|\overline{\boldsymbol{\theta}}_j\| - \sum_{j=1}^p w_j\|\widehat{\boldsymbol{\theta}}_j\|\Big) + (\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}})^T\frac{\tau}{n}\sum_{i=1}^n \boldsymbol{W}_i^T(Z_i(\widehat{\boldsymbol{\beta}}) - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}}). \tag{19}$$

We closely examine the two terms of the RHS of (19).

With $\widehat{\boldsymbol{\delta}}_j \coloneqq \widehat{\boldsymbol{\theta}}_j - \overline{\boldsymbol{\theta}}_j$,

$$\begin{aligned}\sum_{j=1}^p w_j\|\overline{\boldsymbol{\theta}}_j\| - \sum_{j=1}^p w_j\|\widehat{\boldsymbol{\theta}}_j\| &= \sum_{j\in S_-} w_j\|\overline{\boldsymbol{\theta}}_j\| - \sum_{j\in S_-} w_j\|\widehat{\boldsymbol{\delta}}_j + \overline{\boldsymbol{\theta}}_j\| - \sum_{j\in S^c} w_j\|\widehat{\boldsymbol{\delta}}_j\| \\ &\le \sum_{j\in S_-} w_j\|\widehat{\boldsymbol{\delta}}_j\| - \sum_{j\in S^c} w_j\|\widehat{\boldsymbol{\delta}}_j\|\end{aligned} \tag{20}$$

As for the second term, note that

$$\begin{aligned}\frac{\tau}{n}\sum_{i=1}^n \boldsymbol{W}_i(Z_i(\widehat{\boldsymbol{\beta}}) - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}}) = \frac{\tau}{n}\sum_{i=1}^n \boldsymbol{W}_i\{Z_i(\widehat{\boldsymbol{\beta}}) - Z_i(\overline{\boldsymbol{\beta}})\} \\ + \frac{\tau}{n}\sum_{i=1}^n \boldsymbol{W}_i(Z_i(\overline{\boldsymbol{\beta}}) - \mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}]) \\ + \frac{\tau}{n}\sum_{i=1}^n \boldsymbol{W}_i(\mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}] - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}}).\end{aligned} \tag{21}$$

Define Ω_{tmp} by

$$\Omega_{\mathrm{tmp}} \coloneqq \{\lambda_2 \ge 2(E_1+E_2+E_3)\} \cap \Omega_1, \tag{22}$$

where $\boldsymbol{V}_{ij} = \boldsymbol{W}_{ij}$ in B_1, B_2, and B_3,

$$\begin{aligned}E_1 &\coloneqq \frac{B_1}{\sqrt{m}} \vee \sup_{\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\in B_{\mathbb{C}}(r_Q)} \Big|\frac{1}{n}\sum_{i=1}^n\{Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}})\}\Big|, \\ E_2 &\coloneqq \frac{B_2}{\sqrt{m}} \vee \Big|\frac{1}{n}\sum_{i=1}^n(Z_i(\overline{\boldsymbol{\beta}}) - \mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}])\Big|, \\ E_3 &\coloneqq \frac{B_3}{\sqrt{m}} \vee \Big|\frac{1}{n}\sum_{i=1}^n(\mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}] - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}})\Big|.\end{aligned}$$

Recall that $w_0 = 1$ and $w_j = \sqrt{m}$ for $j \in [p]$. See also the comment after (15) for $V_{i0} \equiv 1$.

By (19)-(22) , we have on Ω_{tmp},

$$\begin{aligned}\tau^2(\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}})^T\widehat{\Sigma}(\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}) &\le \lambda_2\tau\Big(\sum_{j\in S_-} w_j\|\widehat{\boldsymbol{\delta}}_j\| - \sum_{j\in S^c} w_j\|\widehat{\boldsymbol{\delta}}_j\|\Big) \\ &\quad + \frac{\lambda_2\tau}{2}\Big(\sum_{j\in S} w_j\|\widehat{\boldsymbol{\delta}}_j\| + \sum_{j\in S^c} w_j\|\widehat{\boldsymbol{\delta}}_j\|\Big) \\ &\le \frac{\lambda_2\tau}{2}\Big(3\sum_{j\in S} w_j\|\widehat{\boldsymbol{\delta}}_j\| - \sum_{j\in S^c} w_j\|\widehat{\boldsymbol{\delta}}_j\|\Big).\end{aligned} \tag{23}$$

This implies $\widehat{\boldsymbol{\delta}} \in \mathbb{C}$ since c_0 from [18] is larger than 3. (23) also implies that

$$\begin{aligned}\frac{\tau^2}{2}\phi_{\min}\|\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}\|^2 &\le \frac{3\lambda_2\tau}{2}\sum_{j\in S} w_j\|\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}\| \\ &\le \frac{3\lambda_2\tau}{2}\sqrt{s}\Big(\sum_{j\in S} w_j^2\|\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}\|^2\Big)^{1/2} \\ &\le \frac{3\lambda_2\tau}{2}\sqrt{ms}\|\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}\|.\end{aligned}$$

Thus we have on Ω_{tmp},

$$\|\widehat{\boldsymbol{\theta}}-\overline{\boldsymbol{\theta}}\| \le \frac{3\lambda_2(ms)^{1/2}}{\tau\phi_{\min}}. \tag{24}$$

By (22), Assumption A7(2), and Lemmas 2-4 with $\boldsymbol{V}_{ij} = \boldsymbol{W}_{ij}$, we have (24) with probability larger than $\mathrm{P}(\Omega_1) - n^{-D_1} - \exp\{-D_2(\log p + m)\}$ if we take a sufficiently large D_0. Hence the proof is complete. □

Proof of Theorem 3. Define Δ_i by

$$\begin{aligned}\Delta_i &:= Z_i(\widehat{\boldsymbol{\beta}}) - \tau\boldsymbol{W}_{iS}^T\overline{\boldsymbol{\theta}}_S \\ &= \{Z_i(\widehat{\boldsymbol{\beta}}) - Z_i(\overline{\boldsymbol{\beta}})\} + [Z_i(\overline{\boldsymbol{\beta}}) - \mathrm{E}\{Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}\}] + [\mathrm{E}\{Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}\} - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}}] \\ &= \epsilon_i + \tau\xi_i + \{Z_i(\widehat{\boldsymbol{\beta}}) - Z_i(\overline{\boldsymbol{\beta}})\} \\ &\quad + [Z_i(\overline{\boldsymbol{\beta}}) - \mathrm{E}\{Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}\} - \epsilon_i] + [\mathrm{E}\{Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}\} - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}} - \tau\xi_i]\end{aligned} \tag{25}$$

Then the oracle estimators have the following expressions :

$$\widetilde{\boldsymbol{\theta}}_S = \overline{\boldsymbol{\theta}}_S + \Big(\frac{\tau^2}{n}\sum_{i=1}^n \boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T\Big)^{-1}\frac{\tau}{n}\sum_{i=1}^n \boldsymbol{W}_{iS}\Delta_i, \tag{26}$$

$$\breve{\boldsymbol{\theta}}_S = \overline{\boldsymbol{\theta}}_S + \Big(\frac{\tau^2}{n}\sum_{i=1}^n \boldsymbol{W}_{iS}\boldsymbol{W}_{iS}^T\Big)^{-1}\frac{\tau}{n}\sum_{i=1}^n \boldsymbol{W}_{iS}(\epsilon_i + \tau\xi_i). \tag{27}$$

We evaluate the difference between them by using Lemma 2 and the latter halves of Lemmas 3-4. See the third line of (25).

A sufficient condition for $\widetilde{\boldsymbol{\theta}}$ in Theorem 3 to be a local solution of the SCAD objective function is

$$\|\widetilde{\boldsymbol{\theta}}_j\| > a\lambda_3, \quad j \in S_-, \tag{28}$$

where a is from the SCAD penalty, and

$$\left\| \frac{\tau}{n} \sum_{i=1}^{n} \boldsymbol{W}_{ij} \{ Z_i(\widehat{\boldsymbol{\beta}}) - \tau \boldsymbol{W}_{iS}^T \widetilde{\boldsymbol{\theta}}_S \} \right\| < \lambda_3, \quad j \in S^c. \tag{29}$$

We begin with (29). By (26), we have

$$\frac{\tau}{n} \sum_{i=1}^{n} \boldsymbol{W}_{ij} \{ Z_i(\widehat{\boldsymbol{\beta}}) - \tau \boldsymbol{W}_{iS}^T \widetilde{\boldsymbol{\theta}}_S \} = \frac{\tau}{n} \sum_{i=1}^{n} \boldsymbol{W}_{ij} \Delta_i - \widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1} \frac{\tau}{n} \sum_{i=1}^{n} \boldsymbol{W}_{iS} \Delta_i, \tag{30}$$

where $\widehat{\Sigma}_{jS} = n^{-1} \sum_{i=1}^{n} \boldsymbol{W}_{ij} \boldsymbol{W}_{iS}^T$.

We have dealt with the first term on the RHS in (30) and we evaluate the second term by using Lemmas 2-4 with $\boldsymbol{V}_{ij} = \widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{W}_{iS}$. Then $\mathcal{A} = S^c$ and

$$\widehat{\Omega}_j = \widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1} \widehat{\Sigma}_{jS}^T.$$

On Ω_2, $\widehat{\Sigma}_j - \widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1} \widehat{\Sigma}_{jS}^T$ is positive definite, which implies

$$\lambda_{\max}(\widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1} \widehat{\Sigma}_{jS}^T) \leq 2C_U'. \tag{31}$$

Since

$$\boldsymbol{V}_{ij} = \left(\widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1/2} \right) \left(\widehat{\Sigma}_S^{-1/2} \boldsymbol{W}_{iS} \right) \quad \text{and} \quad \max_i \|\boldsymbol{W}_{iS}\| \leq C_1 (sm)^{1/2}$$

for some positive C_1, we have

$$\max_{i,j,k} |V_{ijk}| = O((ms)^{1/2}). \tag{32}$$

By (31), (32), and Assumption A7(3), we can apply Lemmas 2-4. Thus if we take a sufficiently large D_0 in λ_3, we have

$$\max_{j \in S^c} \left\| \frac{\tau}{n} \sum_{i=1}^{n} \boldsymbol{W}_{ij} \Delta_i \right\| < \frac{\lambda_3}{2} \quad \text{and} \quad \max_{j \in S^c} \left\| \widehat{\Sigma}_{jS} \widehat{\Sigma}_S^{-1} \frac{\tau}{n} \sum_{i=1}^{n} \boldsymbol{W}_{iS} \Delta_i \right\| < \frac{\lambda_3}{2} \tag{33}$$

with the probability specified in Theorem 3.

Finally we consider (28) and prove it by showing the convergence rate of $\|\widetilde{\boldsymbol{\theta}}_j - \overline{\boldsymbol{\theta}}_j\|$. We need some more notation to exploit Lemmas 2-4 again with $\mathcal{A} = S_-$. Recall we have not imposed any penalty on $\theta_0 (j = 0)$. We concentrate on S_- here. For $j \in S_-$, we define

$$\overline{\boldsymbol{W}}_j \coloneqq (\boldsymbol{W}_{1j}, \ldots, \boldsymbol{W}_{nj})^T \in \mathbb{R}^{n \times m}, \quad \overline{\boldsymbol{W}}_{S\setminus\{j\}} \coloneqq (\overline{\boldsymbol{W}}_\ell)_{\ell \in S\setminus\{j\}} \mathbb{R}^{n \times (m(s-2)+1)},$$

$$P_{S\setminus\{j\}} \coloneqq \overline{\boldsymbol{W}}_{S\setminus\{j\}} (\overline{\boldsymbol{W}}_{S\setminus\{j\}}^T \overline{\boldsymbol{W}}_{S\setminus\{j\}})^{-1} \overline{\boldsymbol{W}}_{S\setminus\{j\}}^T, \quad \text{and} \quad \boldsymbol{\Delta} \coloneqq (\Delta_1, \ldots, \Delta_n)^T.$$

By some standard manipulation, we obtain the following representation :

$$\widetilde{\boldsymbol{\theta}}_j - \overline{\boldsymbol{\theta}}_j = \{n^{-1} \tau \overline{\boldsymbol{W}}_j^T (I - P_{S\setminus\{j\}}) \overline{\boldsymbol{W}}_j\}^{-1} n^{-1} \overline{\boldsymbol{W}}_j^T (I - P_{S\setminus\{j\}}) \boldsymbol{\Delta}. \tag{34}$$

We have on Ω_2,

$$C_2^{-1} \le \lambda_{\min}(n^{-1} \overline{\boldsymbol{W}}_j^T (I - P_{S\setminus\{j\}}) \overline{\boldsymbol{W}}_j) \le \lambda_{\max}(n^{-1} \overline{\boldsymbol{W}}_j^T (I - P_{S\setminus\{j\}}) \overline{\boldsymbol{W}}_j) \le C_2, \ j \in S_-,$$

for some positive constant C_2. Therefore we have only to consider

$$n^{-1} \overline{\boldsymbol{W}}_j^T (I - P_{S\setminus\{j\}}) \boldsymbol{\Delta} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{W}_{ij} \Delta_i - \widehat{\Sigma}_{jS\setminus\{j\}} \widehat{\Sigma}_{S\setminus\{j\}}^{-1} \frac{1}{n} \sum_{i=1}^n \boldsymbol{W}_{iS\setminus\{j\}} \Delta_i \tag{35}$$

where $\widehat{\Sigma}_{jS\setminus\{j\}} \coloneqq n^{-1} \overline{\boldsymbol{W}}_j^T \overline{\boldsymbol{W}}_{S\setminus\{j\}}$, $\widehat{\Sigma}_{S\setminus\{j\}} \coloneqq n^{-1} \overline{\boldsymbol{W}}_{S\setminus\{j\}}^T \overline{\boldsymbol{W}}_{S\setminus\{j\}}$, and $\boldsymbol{W}_{iS\setminus\{j\}} \coloneqq (\boldsymbol{W}_{i\ell})_{\ell \in S\setminus\{j\}}$. For $\breve{\boldsymbol{\theta}}_j$, Δ_i is just replaced with $\epsilon_i + \tau \xi_i$ in (35).

As in evaluating (30), we obtain by Lemmas 2-4 with $\boldsymbol{V}_{ij} = \widehat{\Sigma}_{jS\setminus\{j\}} \widehat{\Sigma}_{S\setminus\{j\}}^{-1} \boldsymbol{W}_{iS\setminus\{j\}}$, (25)-(27), and (34)-(35) that

$$\|\widetilde{\boldsymbol{\theta}}_j - \overline{\boldsymbol{\theta}}_j\| \le D_1 \sqrt{\frac{m \log p}{n}} \text{ and } \|\widetilde{\boldsymbol{\theta}}_j - \breve{\boldsymbol{\theta}}_j\| = O\Big(\frac{\log p + m}{m} t_n \sqrt{\frac{m^3 s^3}{n \log p}} \Big) \sqrt{\frac{m \log p}{n}} \Big) \tag{36}$$

uniformly in $j \in S_-$ for some positive D_1 with the specified probability in the theorem if we take a sufficiently large D_0 in λ_3. For $\widetilde{\boldsymbol{\theta}}_j - \overline{\boldsymbol{\theta}}_j$, we employed Lemmas 2-4 for the last three terms in the second line of (25). For $\widetilde{\boldsymbol{\theta}}_j - \breve{\boldsymbol{\theta}}_j$, we employed Lemmas 2-4 for the last three terms in the third line of (25).

By the latter half of Assumption 7(3), we have

$$\frac{\|\overline{\boldsymbol{\theta}}_j\|}{(m \log p/n)^{1/2}} \to \infty. \tag{37}$$

The former half of (36) and (37) imply (28). Hence the proof is complete.

□

6 Conclusion

We proposed a two-step procedure for additive ES models with high-dimensional covariates. We also assumed additive models for conditional quantiles. We considered the group Lasso and the group SCAD and successfully provided the oracle inequality and proved the oracle property for the group Lasso and the group SCAD, respectively. Our numerical studies showed the desirable performances of the proposed models and two-step procedures.

Acknowledgements.

This research is financially supported by JSPS KAKENHI Grant Number JP 24K14850 (HONDA) and Taiwan NSTC Grant Number 112-2122-M-007-001-MY3 (PENG).

References

[1] S. Barendse. Efficiently weighted estimation of tail and interquantile expectations. 2020.

[2] S. Barendse. Expected shortfall lasso. *arXiv preprint arXiv:2307.01033*, 2023.

[3] A. Belloni and V. Chernozhukov. $l1$-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39:82–130, 2011.

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

[5] C. Bilim, C. D. Atiş, H. Tanyildizi, and O. Karahan. Predicting the compressive strength of ground granulated blast furnace slag concrete using artificial neural network. *Advances in Engineering Software*, 40.

[6] S. Boucheron, G. Lugosi, and P P. Massart. *A Nonasymptotic Theory of Independence. Concentration Inequalities*. Oxford University Press, Oxford, 2013.

[7] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods Theory and Applications*. Springer, New York, Dordrecht, Heidelberg, London, 2011.

[8] S. X. Chen. Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics*, 6:87–107, 2008.

[9] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68, 2018.

[10] T. Dimitriadis and S. Bayer. A joint quantile and expected shortfall regression framework. *Electron. J. Statist.*, 13:1823–1871, 2019.

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

[12] J. Fan, R. Li, C.-H. Zhang, and H. Zou. *Statistical Foundations of Data Science*. CRC press, Boca Raton, 2020.

[13] T. Fissler and J. F. Ziegel. Higher order elicitability and osband's principle. *The Annals of Statistics*, 44:1680–1707, 2016.

[14] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity*. CRC press, Boca Raton, 2015.

[16] X. He, K. M. Tan, and W. X. Zhou. Robust estimation and inference for expected shortfall regression with many regressors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85:1223–1246, 2023.

[17] T. Honda, C.-K. Ing, and W.-Y. Wu. Adaptively weighted group lasso for semiparametric quantile regression models. *Bernoulli*, 25:3311–3338, 2019.

[18] K. Kato. Group lasso for high dimensional sparse quantile regression models. *arXiv preprint arXiv:1103.1458*, 2011.

[19] K. Kato. Weighted nadaraya-watson estimation of conditional expected shortfall. *Journal of Financial Econometrics*, 10:265–291, 2012.

[20] R. Koenker. *Quantile Regression*. Cambridge University Press, New York, 2005.

[21] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[22] P. K. Mehta and P. J. M. Monteiro. *Concrete: Microstructure, Properties, and Materials*. McGraw-Hill, New York, 3rd edition, 2006.

[23] A. J. Patton, J. F. Ziegel, and R. Chen. Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211:388–413, 2019.

[24] L. L. Schumaker. *Spline Functions: Basic Theory 3rd ed*. Cambridge University Press, Cambridge, 2007.

[25] R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58:267–288, 1996.

[26] S. van de Geer. *Estimation and testing under sparsity*. Springer, Switzerland, 2016.

[27] A. D. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

[28] M. Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019.

[29] I.-C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998.

[30] M. Yu, Y. Wang, S. Xie, K. M. Tan, and W.-X. Zhou. Estimation and inference for nonparametric expected shortfall regression over rkhs. *Journal of the American Statistical Association*, just-accepted, 2024.

[31] S. Zhang, X. He, K. M. Tan, and W. X. Zhou. High-dimensional expected shortfall regression. *Journal of the American Statistical Association*, just-accepted, 2025.

7 Appendix(to be the supplementary material)

We prove Lemmas 1-4 here.

Proof of Lemma 1. In the proof, we only verify that

$$\mathrm{P}\Big(\frac{C_{\phi L}}{2} \le \min_{\boldsymbol{\alpha}\in\mathbb{C}} \frac{\boldsymbol{\alpha}^T\widehat{\Sigma}\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|^2} \le \max_{\boldsymbol{\alpha}\in\mathbb{C}} \frac{\boldsymbol{\alpha}^T\widehat{\Sigma}\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|^2} \le 2C_{\phi U}\Big) \to 1$$

since we can evaluate the probabilities of the other events in the same way.

First we define Δ by

$$\Delta \coloneqq \max_{0\le j,k\le pm} |(\widehat{\Sigma}-\Sigma)_{jk}|.$$

As is well known in the Lasso literature, we have that for $\boldsymbol{\alpha}\in\mathbb{C}$,

$$\begin{aligned} |\boldsymbol{\alpha}^T(\widehat{\Sigma}-\Sigma)\boldsymbol{\alpha}| &\le \|\boldsymbol{\alpha}\|_1^2\Delta \le \Big(|\alpha_0| + \sum_{j=1}^{p} w_j\|\boldsymbol{\alpha}_j\|\Big)^2\Delta \\ &\le (c_0+1)^2\Big(|\alpha_0| + \sum_{j\in S_-} w_j\|\boldsymbol{\alpha}_j\|\Big)^2\Delta \le ms(c_0+1)^2\|\boldsymbol{\alpha}\|^2\Delta. \end{aligned} \tag{38}$$

Thus we have only to prove $ms\Delta\to 0$ in probability.

Assumptions A3 and A4(1) imply that

$$\mathrm{E}\{W_{1jk}^4\} \le C_1\mathrm{E}\{mW_{1jk}^2\} \le C_2 m. \tag{39}$$

uniformly in j and k for some positive C_1 and C_2.

By using (39) and applying Bernstein's inequality (see pp.102-103 in [27]) to each element of $\widehat{\Sigma}-\Sigma$, we have only to follow the standard argument to prove

$$\mathrm{P}\Big(\Delta > C_3\sqrt{\frac{m\log p}{n}}\Big) \le 2\exp(-C_4\log p) \tag{40}$$

for any positive C_4 if we take a sufficiently large C_3. The desired result follows from Assumptions A5 and A8 and (40). Hence the proof is complete. □

Proof of Lemma 2. Define $\boldsymbol{\delta} \coloneqq \boldsymbol{\beta}-\overline{\boldsymbol{\beta}} \in B_{\mathbb{C}}(r_Q)$ here and write $Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}})$ as

$$Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}}) = (\{Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}})\} - \mathrm{E}[Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}]) + \mathrm{E}[Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}]. \tag{41}$$

We begin with the second term on the RHS of (41) and define

$$R_i(\boldsymbol{\delta}) \coloneqq \boldsymbol{W}_i^T\boldsymbol{\beta} - Y_i = \boldsymbol{W}_i^T\boldsymbol{\delta} - \eta_i.$$

This $r_i(\boldsymbol{\delta})$ satisfies

$$|R_i(\boldsymbol{\delta})|^2 \leq 2\boldsymbol{\delta}^T\boldsymbol{W}_i\boldsymbol{W}_i^T\boldsymbol{\delta} + 2\eta_i^2 \to 0$$

uniformly in i.

Note that

$$\mathrm{E}[Z_i(\boldsymbol{\beta})|\{\boldsymbol{X}_\ell\}] = \int_{-\infty}^{R_i(\boldsymbol{\delta})}(t - R_i(\boldsymbol{\delta}))f_u(t|\boldsymbol{X}_i)dt + \tau\boldsymbol{W}_i^T\boldsymbol{\beta} \tag{42}$$

and

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{E}[Z_i(\boldsymbol{\beta})|\{\boldsymbol{X}_\ell\}] = \{\tau - F_u(R_i(\boldsymbol{\delta})|\boldsymbol{X}_i)\}\boldsymbol{W}_i. \tag{43}$$

(42) and (43) imply that

$$\mathrm{E}[Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}] = \int_0^1\{\tau - F_u(R_i(t\boldsymbol{\delta})|\boldsymbol{X}_i)\}dt\boldsymbol{W}_i^T\boldsymbol{\delta}.$$

Hence we have

$$\begin{aligned}|\mathrm{E}[Z_i(\boldsymbol{\beta}) - Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}]| &\leq \frac{C_{fU}}{2}\{|\boldsymbol{W}_i^T\boldsymbol{\delta}|(|\boldsymbol{W}_i^T\boldsymbol{\delta}| + |\eta_i|)\} \\ &\leq \frac{C_{fU}}{2}\Big\{\frac{3}{2}|\boldsymbol{W}_i^T\boldsymbol{\delta}|^2 + \frac{1}{2}|\eta_i|^2\Big\}.\end{aligned}$$

We have on Ω_ω,

$$\frac{1}{n}\sum_{i=1}^n|\boldsymbol{W}_i^T\boldsymbol{\delta}|^2|V_{ijk}| \leq V_{\max}\boldsymbol{\delta}^T\widehat{\Sigma}\boldsymbol{\delta} \leq 2V_{\max}C_{\phi U}\|\boldsymbol{\delta}\|^2$$

and

$$\Big\|\frac{1}{n}\sum_{i=1}^n|\eta_i|^2\boldsymbol{V}_{ij}\Big\| \leq C_{\omega U}^{1/2}\Big(\frac{1}{n}\sum_{i=1}^n\eta_i^4\Big)^{1/2} \leq C_1n^{-4/5}$$

for some positive C_1. For the latter, see the argument at the end of the proof of Lemma 4.

Therefore we have on Ω_ω,

$$\sup_{j\in\mathcal{A}}\sup_{\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\in B_{\mathbb{C}}(r_Q)}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_{ij}\mathrm{E}[Z_i(\boldsymbol{\beta})-Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}]\right\|\le C_2(n^{-4/5}+m^{1/2}V_{\max}C_{\phi U}r_Q^2) \quad (44)$$

for some positive C_2.

We consider the first term. We deal with $\boldsymbol{\delta}=\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\in B_{\mathbb{C}}(r_Q)$. Define $r_{ijk}(\boldsymbol{\delta})$ by

$$r_{ijk}(\boldsymbol{\delta}) \coloneqq V_{ijk}(Z_i(\boldsymbol{\beta})-Z_i(\overline{\boldsymbol{\beta}})),$$

and notice that

$$|r_{ijk}(\boldsymbol{\delta}_1)-r_{ijk}(\boldsymbol{\delta}_2)|\le V_{\max}|\boldsymbol{W}_i^T(\boldsymbol{\delta}_1-\boldsymbol{\delta}_2)|. \quad (45)$$

Taking $r_{1Q}=\sqrt{s}(c_0+1)r_Q$, we define A_{jk} by

$$A_{jk}\coloneqq\frac{n}{V_{\max}r_{1Q}}\sup_{\boldsymbol{\delta}\in B_{\mathbb{C}}(r_Q)}\left|\frac{1}{n}\sum_{i=1}^{n}(r_{ijk}(\boldsymbol{\delta})-\mathrm{E}[r_{ijk}(\boldsymbol{\delta})|\{\boldsymbol{X}_\ell\}])\right| \quad (46)$$

as in the proof of Lemma S.1.1 in [31]. Here we apply another symmetrization theorem from that cited in [31] and then we adopt some part of the proof of Lemma A.3 in [18]. Hereafter we carefully cope with the group structure.

Let $\{e_i\}$ be a sequence of independent Rademacher variables. By the symmetrization inequality (cf. Lemma 2.3.7 in [27]), we have

$$P_{\Omega_\omega}(A_{jk}\ge t)\le 4\mathrm{E}_{\Omega_\omega}\mathrm{P}\Big[\frac{1}{V_{\max}r_{1Q}}\sup_{\boldsymbol{\delta}\in B_{\mathbb{C}}(r_Q)}\Big|\sum_{i=1}^{n}e_ir_{ijk}(\boldsymbol{\delta})\Big|>\frac{t}{4}\Big|\{\boldsymbol{X}_\ell\}\Big].$$

By Markov's inequality, $P_{\Omega_\omega}(A_{jk}\ge t)$ is bounded by

$$e^{-t\lambda/4}\mathrm{E}_{\Omega_\omega}\mathrm{E}\Big[\exp\Big\{\frac{\lambda}{V_{\max}r_{1Q}}\sup_{\boldsymbol{\delta}\in B_{\mathbb{C}}(r_Q)}\Big|\sum_{i=1}^{n}e_ir_{ijk}(\boldsymbol{\delta})\Big|\Big\}\Big|\{\boldsymbol{X}_\ell\}\Big], \quad (47)$$

where λ is any positive number.

By recalling (45) and applying the contraction theorem (cf. Theorem 4.12 in [21]) , the conditional expectation in (47) is bounded by

$$\mathrm{E}\Big[\exp\Big\{\frac{\lambda}{r_{1Q}}\sup_{\boldsymbol{\delta}\in B_{\mathbb{C}}(r_Q)}\Big|\sum_{i=1}^{n}e_i\boldsymbol{W}_i^T\boldsymbol{\delta}\Big|\Big\}\Big|\{\boldsymbol{X}_\ell\}\Big]\le\mathrm{E}\Big[\exp\Big\{\lambda\sup_{0\le j\le p}\Big\|\sum_{i=1}^{n}e_i\boldsymbol{W}_{ij}\Big\|\Big\}\Big|\{\boldsymbol{X}_\ell\}\Big]. \quad (48)$$

We evaluate the RHS of (48) by exploiting Corollary A.1 in [18] $(p+1)$ times for

each j. Noticing that

$$2\mathrm{E}[Z^2] = 2\sum_{i=1}^{n} \boldsymbol{W}_{ij}^T \boldsymbol{W}_{ij} \le 2C_U mn \ \text{ and } \sigma^2 = \sup_{\|\boldsymbol{\gamma}\|=1} \boldsymbol{\gamma}^T \sum_{i=1}^{n} \boldsymbol{W}_{ij}\boldsymbol{W}_{ij}^T \boldsymbol{\gamma} \le 2C_U n$$

in that corollary, we have an upper bound of the RHS of (48) :

$$16\exp\{\log(p+1) + \lambda C_1\sqrt{n}(\sqrt{m} + \lambda\sqrt{n})\} \tag{49}$$

for some sufficiently large C_1 depending on C_U.

Therefore

$$e^{-t\lambda/4}\mathrm{E}\Big[\exp\Big\{\frac{\lambda}{V_{\max} r_{1Q}} \sup_{\boldsymbol{\delta}\in B_{\mathbb{C}}(r_Q)} \Big|\sum_{i=1}^{n} e_i r_{ijk}(\boldsymbol{\delta})\Big|\Big\}\Big|\{\boldsymbol{X}_\ell\}\Big]$$

is bounded by

$$16\exp\{-t\lambda/4 + \log(p+1) + \lambda C_1\sqrt{n}(\sqrt{m} + \lambda\sqrt{n})\}. \tag{50}$$

and we obtain that

$$P_{\Omega_\omega}(A_{jk} \ge t) \le 16\mathrm{P}(\Omega_\omega)\exp\{-t\lambda/4 + \log(p+1) + \lambda C_1\sqrt{n}(\sqrt{m} + \lambda\sqrt{n})\} \tag{51}$$

If we take $t = C_2\sqrt{n(\log p + m)}$ and $\lambda = C_3\sqrt{(\log p + m)/n}$, the expression inside the exponential in (51) reduces to

$$-\frac{C_2C_3}{4}(\log p + m) + \log(p+1) + C_1C_3\sqrt{(\log p + m)}\{\sqrt{m} + C_3\sqrt{(\log p + m)}\}. \tag{52}$$

By (51) and (52), we choose a sufficient large C_2 for fixed C_1 and C_3 and obtain that

$$\begin{aligned} P(A_{jk} \ge C_2\sqrt{n(\log p + m)}|\Omega_\omega) &\le 16\exp\Big\{-\frac{C_2C_3}{8}(\log p + m)\Big\}, \\ P\Big(\sup_{j\in\mathcal{A},k\in[m]} A_{jk} \ge C_2\sqrt{n(\log p + m)}\Big|\Omega_\omega\Big) &\le 16\exp\Big\{\log(p+1) + \log m - \frac{C_2C_3}{8}(\log p + m)\Big\}. \end{aligned} \tag{53}$$

If $C_2C_3 > 16$ and we can choose such C_2, (53) yields

$$P\Big(\sup_{j\in\mathcal{A},k\in[m]} A_{jk} \ge C_2\sqrt{n(\log p + m)}\Big|\Omega_\omega\Big) \le 16\exp\Big\{-\frac{C_2C_3}{16}(\log p + m)\Big\}. \tag{54}$$

If

$$\sup_{j\in\mathcal{A},k\in[m]} A_{jk} \le C_2\sqrt{n(\log p + m)},$$

we obtain by the definitions of r_{1Q} above (46) and r_Q in (13) that

$$\sup_{\boldsymbol{\delta}\in B_{\mathbb{C}}(r_Q)} \left|\frac{1}{n}\sum_{i=1}^{n}(r_{ijk}(\boldsymbol{\delta}) - \mathrm{E}[r_{ijk}(\boldsymbol{\delta})|\{\boldsymbol{X}_\ell\}])\right| \le C_4 t_n V_{\max}\sqrt{\frac{m^2 s^2}{n\log p}}\frac{\log p + m}{m}\sqrt{\frac{\log p}{n}} \quad (55)$$

uniformly in j and k for some sufficiently large C_4.

(44), (54), and (55) yields the desired result and the proof of the lemma is complete.

□

This kind of result is a standard one in the Lasso literature. We prove this lemma as Lemma S.1.3 in [31].

Proof of Lemma 3. We decompose $Z_i(\overline{\boldsymbol{\beta}})$ as

$$\begin{aligned} Z_i(\overline{\boldsymbol{\beta}}) = {} & (Y_i - Q_i)I(Y_i \le Q_i) + \eta_i I(Y_i \le \boldsymbol{W}_i^T\overline{\boldsymbol{\beta}}) \\ & + (Y_i - Q_i)\{I(Y_i \le \boldsymbol{W}_i^T\overline{\boldsymbol{\beta}}) - I(Y_i \le Q_i)\} + \tau(Q_i - \eta_i). \end{aligned} \quad (56)$$

The last term depends on only $\{\boldsymbol{X}_\ell\}$.

We give the details only for the first term on the RHS here because we can deal with the second and third terms in the same way. Recall that

$$\epsilon_i = (Y_i - Q_i)I(Y_i \le Q_i) - \mathrm{E}\{(Y_i - Q_i)I(Y_i \le Q_i)|\boldsymbol{X}_i\}.$$

By Assumption A6 and Theorem 2.10 in [6], we obtain that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i V_{ijk}\right| \le \sigma_u C_{\omega U}^{1/2}\sqrt{\frac{2t}{n}} + 2B_u V_{\max}\frac{t}{n}$$

with probabilty $(1-2e^{-t})$ conditional on $\{\boldsymbol{X}_\ell\}$ on Ω_ω. If we take $t = (2+C_1)\log p$, then we have

$$\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i V_{ijk}\right| \le C_2\sqrt{\frac{\log p}{n}} \quad (57)$$

for some sufficiently large C_2 with probability $(1 - n^{-C_1})$ conditional on $\{\boldsymbol{X}_\ell\}$ on Ω_ω. Recall that we assume $\log p \ge n$.

As for the second term, notice that

$$\mathrm{E}[\eta_i^2 I(Y_i \le \boldsymbol{W}_i^T\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}] = O(n^{-4/5}) \quad \text{and} \quad |\eta_i I(Y_i \le \boldsymbol{W}_i^T\overline{\boldsymbol{\beta}})| = O(n^{-2/5})$$

uniformly in i. Therefore we have that for any given C_3,

$$\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n V_{ijk}(\eta_i I(Y_i \le \boldsymbol{W}_i^T\overline{\boldsymbol{\beta}}) - \mathrm{E}[\eta_i I(Y_i \le \boldsymbol{W}_i^T\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}])\right| \le C_4 n^{-2/5}\sqrt{\frac{\log p}{n}}. \quad (58)$$

for some sufficiently large C_4 with probability larger than $(1-n^{-C_3})$ conditional on $\{\boldsymbol{X}_\ell\}$ on Ω_ω. We can deal with the third term in the same way.

Then we should take the expectation on Ω_ω and collect terms of $k = 1, \ldots, m$ for each j. Then the desired result follows from (57), (58), and the result for the third term. □

Proof of Lemma 4. First we evaluate $\mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}] - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}}$. Note that

$$\mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}] = \tau S_i - \int_{\boldsymbol{W}_i^T\overline{\boldsymbol{\beta}}}^{Q_i} y f_Y(y|\boldsymbol{X}_i)dy + \boldsymbol{W}_i^T\overline{\boldsymbol{\beta}}\int_{\boldsymbol{W}_i^T\overline{\boldsymbol{\beta}}}^{Q_i} f_Y(y|\boldsymbol{X}_i)dy, \quad (59)$$

where $f_Y(u+Q_i|\boldsymbol{X}_i) = f_u(u|\boldsymbol{X}_i)$ and $u = y - Q_i$.

Thus we have

$$\begin{aligned}\mathrm{E}[Z_i(\overline{\boldsymbol{\beta}})|\{\boldsymbol{X}_\ell\}] - \tau\boldsymbol{W}_i^T\overline{\boldsymbol{\theta}} &= \tau\xi_i - \int_{\boldsymbol{W}_i^T\overline{\boldsymbol{\beta}}}^{Q_i}(y - \boldsymbol{W}_i^T\overline{\boldsymbol{\beta}})f_Y(y|\boldsymbol{X}_i)dy \quad (60)\\ &= \tau\xi_i + O(\eta_i^2)\end{aligned}$$

uniformly in i.

For any $\{\zeta_i\}$, we have on Ω_ω,

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n \zeta_i \boldsymbol{V}_{ij}\right\| &= \sup_{\|\boldsymbol{\alpha}\|=1}\left|\frac{1}{n}\sum_{i=1}^n \zeta_i \boldsymbol{\alpha}^T\boldsymbol{V}_{ij}\right| \quad (61)\\ &\le \left(\frac{1}{n}\sum_{i=1}^n \zeta_i^2\right)^{1/2}\left(\sup_{\|\boldsymbol{\alpha}\|=1}\boldsymbol{\alpha}^T\widehat{\Omega}_j\boldsymbol{\alpha}\right)^{1/2}\\ &\le \lambda_{\max}(\widehat{\Omega}_j)\left(\frac{1}{n}\sum_{i=1}^n \zeta_i^2\right)^{1/2} \le C_{\omega U}\left(\frac{1}{n}\sum_{i=1}^n \zeta_i^2\right)^{1/2}.\end{aligned}$$

The desired results follow from (60) and (61). Hence the proof is complete. □