QUANTILE INDEX REGRESSION

Yingying Zhang[1], Qianqian Zhu[2], Yuefeng Si[3] and Guodong Li[3]

East China Normal University[1]

Shanghai University of Finance and Economics[2]

The University of Hong Kong[3]

Supplementary Material

This supplementary material gives proofs of Proposition 1 in Section S1, theorems for low-dimensional CQR estimation in Sections S2 and those for high-dimensional regularized estimation in Section S3. The technical proofs for four auxiliary lemmas are proved in Section S4. It also includes additional results for the simulation and empirical analysis in Sections S5 and S6, respectively.

S1 Proof of Proposition 1

Proof of Proposition 1. We first consider the Tukey lambda distribution which has the form of

$$Q(\tau, \theta) = \theta_1 + \theta_2 \left\{ \frac{\tau^{\theta_3} - (1-\tau)^{\theta_3}}{\theta_3} \right\},$$

where $\theta_1 \in \mathbb{R}$, $\theta_2 > 0$, $\theta_3 < 0$. Consider four arbitrary quantile levels $0 < \tau_1 < \tau_2 < \tau_3 < \tau_4 < 1$, and two arbitrary index vectors $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\theta}_3)'$ and $\theta = (\theta_1, \theta_2, \theta_3)'$ such that

$$Q(\tau_j, \widetilde{\theta}) = Q(\tau_j, \theta) \text{ for all } 1 \leq j \leq 4. \tag{S1.1}$$

We show that $\widetilde{\theta} = \theta$ in the following.

The first step is to prove $\widetilde{\theta}_3 = \theta_3$ using the proof by contradiction. Suppose that $\widetilde{\theta}_3 \neq \theta_3$ and, without loss of generality, we assume that $\theta_3 < \widetilde{\theta}_3 < 0$. Denote $f_j(\theta_3) = \tau_j^{\theta_3} - (1-\tau_j)^{\theta_3}$ for $1 \leq j \leq 4$. From (S1.1), we have

$$\frac{f_1(\widetilde{\theta}_3) - f_2(\widetilde{\theta}_3)}{f_3(\widetilde{\theta}_3) - f_2(\widetilde{\theta}_3)} - \frac{f_1(\theta_3) - f_2(\theta_3)}{f_3(\theta_3) - f_2(\theta_3)} = 0, \tag{S1.2}$$

and

$$\frac{f_4(\widetilde{\theta}_3) - f_2(\widetilde{\theta}_3)}{f_3(\widetilde{\theta}_3) - f_2(\widetilde{\theta}_3)} - \frac{f_4(\theta_3) - f_2(\theta_3)}{f_3(\theta_3) - f_2(\theta_3)} = 0. \tag{S1.3}$$

Let us fix θ_3, $\widetilde{\theta}_3$, τ_2 and τ_3. As a result, $\kappa_1 = f_2(\widetilde{\theta}_3)$, $\kappa_2 = f_2(\theta_3)$, $\kappa_3 = f_3(\theta_3) - f_2(\theta_3)$ and $\kappa_4 = f_3(\widetilde{\theta}_3) - f_2(\widetilde{\theta}_3)$ are all fixed values. Denote

$$F(\tau) = \kappa_3 \left\{\tau^{\widetilde{\theta}_3} - (1-\tau)^{\widetilde{\theta}_3} - \kappa_1\right\} - \kappa_4 \left\{\tau^{\theta_3} - (1-\tau)^{\theta_3} - \kappa_2\right\}, \tag{S1.4}$$

$$\dot{F}(\tau) = \kappa_3 \widetilde{\theta}_3 \{\tau^{\widetilde{\theta}_3 - 1} + (1-\tau)^{\widetilde{\theta}_3 - 1}\} + \kappa_4 \theta_3 \{\tau^{\theta_3 - 1} + (1-\tau)^{\theta_3 - 1}\},$$

and

$$G(\tau) = \frac{\tau^{\widetilde{\theta}_3 - 1} + (1-\tau)^{\widetilde{\theta}_3 - 1}}{\tau^{\theta_3 - 1} + (1-\tau)^{\theta_3 - 1}},$$

where $\dot{F}(\cdot)$ is the derivative function of $F(\cdot)$, and $\dot{F}(\tau) = 0$ if and only if $G(\tau) = \kappa_4 \theta_3 / (\kappa_3 \widetilde{\theta}_3)$. Note that equations (S1.2) and (S1.3) correspond to $F(\tau_1) = 0$ and $F(\tau_4) =$

0, respectively. Moreover, it can be verified that $F(\tau_2) = 0$ and $F(\tau_3) = 0$, i.e. the equation $F(\tau) = 0$ has at least four different solutions. As a result, the equation $\dot{F}(\tau) = 0$ or $G(\tau) = \kappa_4\theta_3/(\kappa_3\widetilde{\theta}_3)$ has at least three different solutions. While it is implied by Lemma S1 that the equation $G(\tau) = \kappa_4\theta_3/(\kappa_3\widetilde{\theta}_3)$ has at most two different solutions. Due to the contradiction, we prove that $\widetilde{\theta}_3 = \theta_3$, and it is readily to further verify that $(\widetilde{\theta}_1, \widetilde{\theta}_2) = (\theta_1, \theta_2)$.

We next consider the generalized extreme value distribution (GEVD) with quantile function as follows

$$Q(\tau, \theta) = \theta_1 + \theta_2 \frac{1 - (-\log \tau)^{\theta_3}}{\theta_3},$$

where $\theta = (\theta_1, \theta_2, \theta_3)'$, and $\theta_1 \in \mathbb{R}$, $\theta_2 > 0$ and $\theta_3 < 0$. Consider three arbitrary quantile levels $0 < \tau_1 < \tau_2 < \tau_3 < 1$, and two arbitrary index vectors $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\theta}_3)'$ and $\theta = (\theta_1, \theta_2, \theta_3)'$ such that

$$Q(\tau_j, \widetilde{\theta}) = Q(\tau_j, \theta) \text{ for all } 1 \le j \le 3. \tag{S1.5}$$

We show that $\widetilde{\theta} = \theta$ in the following. We first prove $\widetilde{\theta}_3 = \theta_3$ using the proof by contradiction. Denote $f_j(\theta_3) = \{1 - (-\log \tau_j)^{\theta_3}\}/\theta_3$ for $1 \le j \le 3$. It holds that $Q(\tau_j, \widetilde{\theta}) = \theta_1 + \theta_2 f_j(\theta_3)$. From (S1.5), we have

$$\frac{f_2(\widetilde{\theta}_3) - f_1(\widetilde{\theta}_3)}{f_3(\widetilde{\theta}_3) - f_2(\widetilde{\theta}_3)} = \frac{f_2(\theta_3) - f_1(\theta_3)}{f_3(\theta_3) - f_2(\theta_3)}, \tag{S1.6}$$

Let us fix θ_3, $\widetilde{\theta}_3$, τ_1 and τ_2. Then $\kappa_1 = f_2(\widetilde{\theta}_3)$, $\kappa_2 = f_2(\theta_3)$, $\kappa_3 = f_2(\theta_3) - f_1(\theta_3)$ and $\kappa_4 = f_2(\widetilde{\theta}_3) - f_1(\widetilde{\theta}_3)$ are all fixed values. Denote

$$F(\tau) = \kappa_3 \left\{ \frac{1 - (-\log \tau)^{\widetilde{\theta}_3}}{\widetilde{\theta}_3} - \kappa_1 \right\} - \kappa_4 \left\{ \frac{1 - (-\log \tau)^{\theta_3}}{\theta_3} - \kappa_2 \right\}, \tag{S1.7}$$

$$\dot{F}(\tau)=\frac{\kappa_3(-\log\tau)^{\widetilde{\theta}_3-1}}{\tau}-\frac{\kappa_4(-\log\tau)^{\theta_3-1}}{\tau},$$

and $G(\tau)=(-\log\tau)^{\widetilde{\theta}_3-\theta_3}$, where $\dot{F}(\cdot)$ is the derivative function of $F(\cdot)$, and $\dot{F}(\tau)=0$ if and only if $G(\tau)=\kappa_4/\kappa_3$. Since the derivative function of $G(\cdot)$ is $\dot{G}(\tau)=-(\widetilde{\theta}_3-\theta_3)(-\log\tau)^{\widetilde{\theta}_3-\theta_3-1}/\tau$, it holds that $\dot{G}(\tau)>0$ if $\theta_3>\widetilde{\theta}_3$ and $\dot{G}(\tau)<0$ if $\theta_3<\widetilde{\theta}_3$. This implies that $\dot{F}(\tau)=0$ or $G(\tau)=\kappa_4/\kappa_3$ has at most one root. Thus $F(\tau)=0$ has at most two different roots. However, we have $F(\tau_1)=F(\tau_2)=0$ by (S1.5) and $F(\tau_3)=0$ by (S1.6). This indicates that $F(\tau)=0$ has at least three different roots. Due to the contradiction, we prove that $\widetilde{\theta}_3=\theta_3$, and it is easy to verify that $(\widetilde{\theta}_1,\widetilde{\theta}_2)=(\theta_1,\theta_2)$.

Finally, we consider the generalized Pareto distribution (GPD) with quantile function as follows

$$Q(\tau,\theta)=\theta_1+\theta_2\frac{1-(1-\tau)^{\theta_3}}{\theta_3},$$

where $\theta=(\theta_1,\theta_2,\theta_3)'$, and $\theta_1\in\mathbb{R}$, $\theta_2>0$ and $\theta_3<0$. The proof follows the same line as for the GEVD, with $f_j(\theta_3)=[1-(1-\tau)^{\theta_3}]/\theta_3$ for $1\le j\le 3$,

$$F(\tau)=\kappa_3\left\{\frac{1-(1-\tau)^{\widetilde{\theta}_3}}{\widetilde{\theta}_3}-\kappa_1\right\}-\kappa_4\left\{\frac{1-(1-\tau)^{\theta_3}}{\theta_3}-\kappa_2\right\}, \tag{S1.8}$$

$$\dot{F}(\tau)=\kappa_3(1-\tau)^{\widetilde{\theta}_3-1}-\kappa_4(1-\tau)^{\theta_3-1},$$

$$G(\tau)=(1-\tau)^{\widetilde{\theta}_3-\theta_3}\quad\text{and}\quad\dot{G}(\tau)=-(\widetilde{\theta}_3-\theta_3)(1-\tau)^{\widetilde{\theta}_3-\theta_3-1}.$$

As a result, the proof of this proposition is complete. $\square$

S2 Proofs for low-dimensional CQR Estimation

This section gives technical proofs of Theorems 1-3 in Section 2.2. Two auxiliary lemmas are also presented at the end of this subsection: Lemma S1 is used for the proof of Proposition 1, and Lemma S2 is for that of Theorem 3. The proofs of these two auxiliary lemmas are given in Section S4.

Proof of Theorem 1. We first prove the uniqueness of β_0. Denote $\bar{L}_k(\boldsymbol{\beta}) = E[\rho_{\tau_k}\{Y - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}))\}]$. From model (2.2), $\boldsymbol{\beta}_0$ is the minimizer not only of $\bar{L}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \bar{L}_k(\boldsymbol{\beta})$, but also of $\bar{L}_k(\boldsymbol{\beta})$ for all $1 \leq k \leq K$. Suppose that $\boldsymbol{\beta}_0^*$ is another minimizer of $\bar{L}(\boldsymbol{\beta})$, and then it is also the minimizer of $\bar{L}_k(\boldsymbol{\beta})$ for all $1 \leq k \leq K$.

Note that, for $u \neq 0$,

$$\begin{aligned} \rho_\tau(u-v) - \rho_\tau(u) &= -v\psi_\tau(u) + (u-v)[I(0 > u > v) - I(0 < u < v)] \\ &= -v\psi_\tau(u) + \int_0^v [I(u \leq s) - I(u \leq 0)]ds, \end{aligned} \tag{S2.1}$$

where $\psi_\tau(u) = \tau - I(u < 0)$; see Knight (1998). Let $U^{(k)} = Y - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}_0))$ and $V^{(k)} = Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}_0^*)) - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}_0))$. It holds that, for each $1 \leq k \leq K$,

$$0 = \bar{L}_k(\boldsymbol{\beta}_0^*) - \bar{L}_k(\boldsymbol{\beta}_0) = E\{(U^{(k)} - V^{(k)})[I(0 > U^{(k)} > V^{(k)}) - I(0 < U^{(k)} < V^{(k)})]\},$$

which implies that $V^{(k)} = Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}_0^*)) - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}_0)) = 0$ with probability one. This, together with the conditions in Proposition 1 and the monotonic link functions, leads to the fact that $\mathbf{X}'\boldsymbol{\beta}_{j0}^* = \mathbf{X}'\boldsymbol{\beta}_{j0}$ for all $1 \leq j \leq d$. We then have $\boldsymbol{\beta}_0^* = \boldsymbol{\beta}_0$ since $E(\boldsymbol{X}\boldsymbol{X}')$ is positive definite. This accomplishes the proof of the uniqueness of $\boldsymbol{\beta}_0$. $\square$

Proof of Theorem 2. Note that $L_n(\boldsymbol{\beta}) = n^{-1}\sum_{k=1}^K \sum_{i=1}^n \rho_{\tau_k}\{Y_i - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}_i, \boldsymbol{\beta}))\}$ and $\bar{L}(\boldsymbol{\beta}) = E[\sum_{k=1}^K \rho_{\tau_k}\{Y - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}))\}]$. Denote $\Phi(X_i, \beta) = \sum_{k=1}^K \phi_k(\mathbf{X}_i, \boldsymbol{\beta})$, where $\phi_k(\mathbf{X}_i, \boldsymbol{\beta}) = \rho_{\tau_k}\{Y_i - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}_i, \boldsymbol{\beta}))\} - \rho_{\tau_k}\{Y_i - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}_i, \boldsymbol{\beta}_0))\}$. It follows that

$$\bar{L}(\boldsymbol{\beta}) - \bar{L}(\boldsymbol{\beta}_0) = E\{\Phi(\mathbf{X}_i, \boldsymbol{\beta})\} \quad \text{and} \quad L_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^n \Phi(\mathbf{X}_i, \boldsymbol{\beta}).$$

By Knight's identity at (S2.1) and Taylor expansion, together with the condition of $E\max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\Theta}\|\partial Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}))/\partial\boldsymbol{\beta}\| < \infty$, we can verify that

$$\begin{aligned} E\sup_{\beta\in\Theta}|\Phi(\mathbf{X}_i, \boldsymbol{\beta})| &\le \sum_{k=1}^K E\sup_{\boldsymbol{\beta}\in\Theta}|Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta})) - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}_0))| \\ &\le \sum_{k=1}^K E\sup_{\boldsymbol{\beta}\in\Theta}\left\|\frac{\partial Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}))}{\partial\boldsymbol{\beta}}\right\| \sup_{\boldsymbol{\beta}\in\Theta}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \infty. \end{aligned}$$

Here $\{(Y_i, \mathbf{X}_i')', i = 1, \ldots, n\}$ are independent and identically distributed samples, and Θ is a compact metric space. Moreover, $\Phi(\mathbf{X}_i, \boldsymbol{\beta})$ is a measurable function of $\mathbf{X}_i$ in Euclidean space for each $\boldsymbol{\beta} \in \Theta$, and a continuous function of $\boldsymbol{\beta} \in \Theta$ for each $\mathbf{X}_i$. Then by the uniform law of large numbers in Lemma 2.4 of Newey and McFadden (1994), we have $\sup_{\boldsymbol{\beta}\in\Theta}|n^{-1}\sum_{i=1}^n \Phi(\mathbf{X}_i, \boldsymbol{\beta}) - E\{\Phi(\mathbf{X}_i, \boldsymbol{\beta})\}| = o_p(1)$, that is

$$\sup_{\boldsymbol{\beta}\in\Theta}|L_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta}_0) - [\bar{L}(\boldsymbol{\beta}) - \bar{L}(\boldsymbol{\beta}_0)]| = o_p(1). \tag{S2.2}$$

Note that $\bar{L}(\boldsymbol{\beta})$ is a continuous function with respect to $\boldsymbol{\beta}$ and, from Theorem 1, $\boldsymbol{\beta}_0$ is the unique minimizer of $\bar{L}(\boldsymbol{\beta})$. As a result, for each $\delta > 0$,

$$\epsilon = \inf_{\boldsymbol{\beta}\in B_\delta^c(\boldsymbol{\beta}_0)} \bar{L}(\boldsymbol{\beta}) - \bar{L}(\boldsymbol{\beta}_0) > 0,$$

where $B_\delta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \delta\}$ and $B^c_\delta(\boldsymbol{\beta}_0)$ is its complement set, and hence

$$\left\{ \inf_{\boldsymbol{\beta} \in B^c_\delta(\boldsymbol{\beta}_0)} L_n(\boldsymbol{\beta}) \leq L_n(\boldsymbol{\beta}_0) \right\} \subseteq \left\{ \sup_{\boldsymbol{\beta} \in B^c_\delta(\boldsymbol{\beta}_0)} |L_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta}_0) - [\bar{L}(\boldsymbol{\beta}) - \bar{L}(\boldsymbol{\beta}_0)]\}| \geq \epsilon \right\}. \tag{S2.3}$$

Note that

$$1 = P\left\{ L_n(\widehat{\boldsymbol{\beta}}_n) \leq L_n(\boldsymbol{\beta}_0) \right\} \leq P\left\{ \widehat{\boldsymbol{\beta}}_n \in B_\delta(\boldsymbol{\beta}_0) \right\} + P\left\{ \inf_{\boldsymbol{\beta} \in B^c_\delta(\boldsymbol{\beta}_0)} L_n(\boldsymbol{\beta}) \leq L_n(\boldsymbol{\beta}_0) \right\}$$

which together with (S2.2) and (S2.3), implies that

$$P\{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \delta\} \geq 1 - P\left\{ \inf_{\boldsymbol{\beta} \in B^c_\delta(\boldsymbol{\beta}_0)} L_n(\boldsymbol{\beta}) \leq L_n(\boldsymbol{\beta}_0) \right\} \to 1,$$

as $n \to \infty$. This accomplishes the proof of consistency. □

Proof of Theorem 3. For simplicity, we denote $Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}_i, \boldsymbol{\beta}))$ by $Q_{ik}(\boldsymbol{\beta})$ in the whole proof of this theorem. Let $S_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta}_0)$ and, from Knight's identity at (S2.1), we have

$$\begin{aligned} S_n(\boldsymbol{\beta}) =& \sum_{k=1}^K \sum_{i=1}^n [\rho_{\tau_k}\{Y_i - Q_{ik}(\boldsymbol{\beta})\} - \rho_{\tau_k}\{Y_i - Q_{ik}(\boldsymbol{\beta}_0)\}] \\ =& \sum_{k=1}^K \sum_{i=1}^n \Big[\{Q_{ik}(\boldsymbol{\beta}) - Q_{ik}(\boldsymbol{\beta}_0)\}\{I(e_{ik} < 0) - \tau_k\} \\ & + \int_0^{Q_{ik}(\boldsymbol{\beta}) - Q_{ik}(\boldsymbol{\beta}_0)} \{I(e_{ik} \leq s) - I(e_{ik} \leq 0)\} ds \Big]. \end{aligned}$$

Note that, by Taylor expansion, $Q_{ik}(\boldsymbol{\beta}) - Q_{ik}(\boldsymbol{\beta}_0) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \partial Q_{ik}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta} + 0.5(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'(\partial^2 Q_{ik}(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}')(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, where $\boldsymbol{\beta}^*$ is a vector between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$, defined by $\boldsymbol{\beta}^* = (1-t)\boldsymbol{\beta}_0 + t\boldsymbol{\beta}$ with some $0 < t < 1$. Let $u = \boldsymbol{\beta} - \boldsymbol{\beta}_0$,

$$q_{1ik}(u) = u' \frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \quad \text{and} \quad q_{2ik}(u) = 0.5 u' \frac{\partial^2 Q_{ik}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} u.$$

We then decompose $S_n(\boldsymbol{\beta})$ into

$$\begin{aligned} S_n(\boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{i=1}^{n}\Big[&\{q_{1ik}(u)+q_{2ik}(u)\}\{I(e_{ik}<0)-\tau_k\} \\ &+ \int_0^{q_{1ik}(u)+q_{2ik}(u)}\{I(e_{ik}\le s)-I(e_{ik}\le 0)\}ds\Big] \\ = -u'T_n + \Pi_{1n}(u) + \Pi_{2n}(u) + \Pi_{3n}(u), \end{aligned}$$

where

$$\begin{aligned} T_n &= \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\{\tau_k - I(e_{ik}\le 0)\}, \\ \xi_{ik}(u) &= \int_0^{q_{1ik}(u)}\{I(e_{ik}\le s)-I(e_{ik}\le 0)\}ds, \\ \Pi_{1n}(u) &= \sum_{k=1}^{K}\sum_{i=1}^{n}\left[\xi_{ik}(u) - E\left\{\xi_{ik}(u)|\mathbf{X}_i\right\}\right], \\ \Pi_{2n}(u) &= \sum_{k=1}^{K}\sum_{i=1}^{n}E\left\{\xi_{ik}(u)|\mathbf{X}_i\right\}, \end{aligned}$$

and

$$\Pi_{3n}(u) = \sum_{k=1}^{K}\sum_{i=1}^{n}\Big[q_{2ik}(u)\left\{I(e_{ik}<0)-\tau_k\right\} + \int_{q_{1ik}(u)}^{q_{1ik}(u)+q_{2ik}(u)}\left\{I(e_{ik}\le s)-I(e_{ik}\le 0)\right\}ds\Big].$$

First, by the central limit theorem, we can show that

$$\frac{1}{\sqrt{n}}T_n = \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\{\tau_k - I(e_{ik}\le 0)\} \to N(0,\Omega_0) \qquad \text{(S2.4)}$$

in distribution as $n\to\infty$. Note that, from Theorem 2, $\widehat{u}_n = \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = o_p(1)$ and then, by applying Lemma S2,

$$\begin{aligned}S_n(\widehat{\boldsymbol{\beta}}_n) &= -\widehat{u}_n' T_n + \Pi_{1n}(\widehat{u}_n) + \Pi_{2n}(\widehat{u}_n) + \Pi_{3n}(\widehat{u}_n) \\ &= -\widehat{u}_n' T_n + \frac{1}{2}(\sqrt{n}\widehat{u}_n)'\Omega_1(\sqrt{n}\widehat{u}_n) + o_p(\sqrt{n}\|\widehat{u}_n\| + n\|\widehat{u}_n\|^2) \qquad \text{(S2.5)} \\ &\geq -\|\sqrt{n}\widehat{u}_n\|\left\{\|\frac{1}{\sqrt{n}}T_n\| + o_p(1)\right\} + n\|\widehat{u}_n\|^2\left\{\frac{\lambda_{\min}(\Omega_1)}{2} + o_p(1)\right\},\end{aligned}$$

where $\lambda_{\min}(\Omega_1)$ is the minimum eigenvalue of Ω_1. This, together with the fact that $S_n(\widehat{\boldsymbol{\beta}}_n) = L_n(\widehat{\boldsymbol{\beta}}_n) - L_n(\boldsymbol{\beta}_0) \leq 0$, implies that

$$\sqrt{n}\|\widehat{u}_n\| \leq \left\{\frac{\lambda_{\min}(\Omega_1)}{2} + o_p(1)\right\}^{-1}\left\{\left\|\frac{1}{\sqrt{n}}T_n\right\| + o_p(1)\right\} = O_p(1). \qquad \text{(S2.6)}$$

Denote $u_n^\star = n^{-1}\Omega_1^{-1}T_n$ and, from (S2.5) and (S2.6),

$$S_n(\widehat{\boldsymbol{\beta}}_n) = \frac{1}{2}(\sqrt{n}\widehat{u}_n)'\Omega_1(\sqrt{n}\widehat{u}_n) - (\sqrt{n}\widehat{u}_n)'\Omega_1(\sqrt{n}u_n^\star) + o_p(1).$$

Moreover, since $\sqrt{n}u_n^\star = O_p(1)$, equation (S2.5) still holds when $\widehat{u}_n$ is replaced by $u_n^\star$, and then

$$S(\boldsymbol{\beta}_0 + u_n^\star) = -\frac{1}{2}(\sqrt{n}u_n^\star)'\Omega_1(\sqrt{n}u_n^\star) + o_p(1),$$

which leads to

$$\begin{aligned}0 \geq S(\widehat{\boldsymbol{\beta}}_n) - S(\boldsymbol{\beta}_0 + u_n^\star) &= \frac{1}{2}(\sqrt{n}\widehat{u}_n - \sqrt{n}u_n^\star)'\Omega_1(\sqrt{n}\widehat{u}_n - \sqrt{n}u_n^\star) + o_p(1) \\ &\geq \frac{\lambda_{\min}(\Omega_1)}{2}\|\sqrt{n}\widehat{u}_n - \sqrt{n}u_n^\star\|^2 + o_p(1).\end{aligned}$$

As a result, from (S2.4),

$$\sqrt{n}\widehat{u}_n = \sqrt{n}u_n^\star + o_p(1) = \Omega_1^{-1}n^{-1/2}T_n + o_p(1) \to N(\mathbf{0}, \Omega_1^{-1}\Omega_0\Omega_1^{-1})$$

in distribution as $n \to \infty$. The proof of this theorem is complete. $\square$

Lemma S1. *Consider the function of $G(\tau)$ defined in the proof of Proposition 1 with $\theta_3 < \widetilde{\theta}_3 < 1$. It holds that, (1) for $\tau > 0.5$, $G(\tau)$ is strictly decreasing, (2) for $\tau < 0.5$, $G(\tau)$ is strictly increasing.*

Lemma S2. *Suppose that the conditions of Theorem 3 hold. For any sequence of random variables $\{u_n\}$ with $u_n = o_p(1)$, it holds that*

(a) $\Pi_{1n}(u_n) = o_p(\sqrt{n}\|u_n\| + n\|u_n\|^2)$,

(b) $\Pi_{2n}(u_n) = 0.5(\sqrt{n}u_n)'\Omega_1(\sqrt{n}u_n) + o_p(n\|u_n\|^2)$, *and*

(c) $\Pi_{3n}(u_n) = o_p(n\|u_n\|^2)$,

where $\Pi_{1n}(u)$, $\Pi_{2n}(u)$ and $\Pi_{3n}(u)$ are defined in the proof of Theorem 3.

S3 Proofs for high-dimensional regularized estimation

This subsection first conducts deterministic analysis at Lemma S3, and then stochastic analysis at Lemmas S4 and S5. The proof of Theorem 4 follows from the deterministic analysis in Lemma S3 and stochastic analysis in Lemmas S4 and S5. The detailed proofs for Lemmas S4 and S5 are given in Section S4.

We first treat the observed data, $\{(Y_i, \boldsymbol{X}_i')', i = 1, ..., n\}$, to be deterministic. Consider the loss function $L_n(\boldsymbol{\beta}) = \sum_{i=1}^n \sum_{k=1}^K \rho_{\tau_k}\{Y_i - Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}_i, \boldsymbol{\beta}))\}$, and its first-order Taylor-series error

$$\mathcal{E}_n(\Delta) = n^{-1}L_n(\boldsymbol{\beta}_0 + \Delta) - n^{-1}L_n(\boldsymbol{\beta}_0) - \langle n^{-1}\nabla L_n(\boldsymbol{\beta}_0), \Delta\rangle,$$

where $\Delta \in \mathbb{R}^{dp}$, $\langle \cdot, \cdot \rangle$ is the inner product, $e_{ik} = Y_i - Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}_i, \boldsymbol{\beta}_0))$, $\psi_\tau(u) = \tau - I(u < 0)$, and $\nabla L_n(\boldsymbol{\beta}) = \sum_{i=1}^n \sum_{k=1}^K \psi_\tau(e_{ik}) \partial Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}))/\partial \boldsymbol{\beta}$ is a subgradient of $L_n(\boldsymbol{\beta})$.

Definition 1. Loss function $L_n(\cdot)$ satisfies the local restricted strong convexity (LRSC) condition if

$$\mathcal{E}_n(\Delta) \geq \alpha \|\Delta\|_2^2 - \eta \sqrt{\frac{\log p}{n}} \|\Delta\|_1 \;\text{ for all } \Delta \text{ such that }\; 0 < r \leq \|\Delta\|_2 \leq R,$$

where $\alpha, \eta > 0$, and $\|\cdot\|_1$ and $\|\cdot\|_2$ are ℓ_1 and ℓ_2 norms, respectively.

The above LRSC condition has a larger tolerance term compared with that in Loh and Wainwright (2015), which has a form of $(\log p/n)\|\Delta\|_1^2$. Similar tolerance term can also be found in Fan et al. (2019) for high-dimensional generalized trace regression. It is ready to establish an upper bound for estimation errors when the penalty λ is appropriately selected.

Lemma S3. *Suppose that the regularizer $p_\lambda(\cdot)$ satisfies Assumption 4, and loss function $L_n(\cdot)$ satisfies the LRSC condition with $\alpha > \mu/4$ and $r = \frac{12\eta\sqrt{s}}{(4\alpha-\mu)L}\sqrt{\frac{\log p}{n}}$. If the tuning parameter λ satisfies that*

$$\lambda \geq \frac{4}{L} \max \left\{ \|n^{-1} \nabla L(\boldsymbol{\beta}_0)\|_\infty, \eta \sqrt{\frac{\log p}{n}} \right\},$$

then the minimizer $\widetilde{\boldsymbol{\beta}}_n$ over the set of $\Theta = \mathcal{B}_R(\boldsymbol{\beta}_0)$ satisfies the error bounds

$$\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq \frac{6\sqrt{s}\lambda L}{4\alpha - \mu} \quad \textit{and} \quad \|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_1 \leq \frac{24 s \lambda L}{4\alpha - \mu}.$$

Proof of Lemma S3. Denote $\widetilde{\Delta}_n = \widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$, and it holds that $\|\widetilde{\Delta}_n\|_2 \leq R$. Note that this lemma naturally holds if $\|\widetilde{\Delta}_n\|_2 \leq (3\sqrt{s}\lambda)/(4\alpha - \mu)$. As a result, we only need to consider the case with $(3\sqrt{s}\lambda)/(4\alpha-\mu) \leq \|\widetilde{\Delta}_n\|_2 \leq R$, and it can be verified that

$$r = \frac{12\eta\sqrt{s}}{(4\alpha-\mu)L}\sqrt{\frac{\log p}{n}} \leq \|\widetilde{\Delta}_n\|_2 \leq R.$$

Note that $n^{-1}L_n(\widetilde{\boldsymbol{\beta}}_n)+p_\lambda(\widetilde{\boldsymbol{\beta}}_n) \leq n^{-1}L_n(\boldsymbol{\beta}_0)+p_\lambda(\boldsymbol{\beta}_0)$, and then $\mathcal{E}_n(\widetilde{\Delta}_n) \leq p_\lambda(\boldsymbol{\beta}_0) - p_\lambda(\widetilde{\boldsymbol{\beta}}_n) - \langle n^{-1}\nabla L_n(\boldsymbol{\beta}_0), \Delta\rangle$. This, together with the LRSC condition and Holder's inequality, implies that

$$\begin{aligned}
\alpha\|\widetilde{\Delta}_n\|_2^2 \leq & p_\lambda(\boldsymbol{\beta}_0) - p_\lambda(\widetilde{\boldsymbol{\beta}}_n) + \left(\eta\sqrt{\frac{\log p}{n}} + \|n^{-1}\nabla L(\boldsymbol{\beta}_0)\|_\infty\right)\|\widetilde{\Delta}_n\|_1 \\
\leq & p_\lambda(\boldsymbol{\beta}_0) - p_\lambda(\widetilde{\boldsymbol{\beta}}_n) + \frac{\lambda L}{2}\|\widetilde{\Delta}_n\|_1 \\
\leq & p_\lambda(\boldsymbol{\beta}_0) - p_\lambda(\widetilde{\boldsymbol{\beta}}_n) + \frac{1}{2}\left\{p_\lambda(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \frac{\mu}{2}\|\widetilde{\Delta}_n\|_2^2\right\} \\
\leq & p_\lambda(\boldsymbol{\beta}_0) - p_\lambda(\widetilde{\boldsymbol{\beta}}_n) + \frac{1}{2}\left\{p_\lambda(\widetilde{\boldsymbol{\beta}}_n) + p_\lambda(\boldsymbol{\beta}_0) + \frac{\mu}{2}\|\widetilde{\Delta}_n\|_2^2\right\},
\end{aligned}$$

where the last inequality follows from the subadditivity of $p_\lambda(\cdot)$, while the penultimate inequality is by Assumption 4; see also Lemma 4 in Loh and Wainwright (2015). As a result,

$$0 < \left(\alpha - \frac{\mu}{4}\right)\|\widetilde{\Delta}_n\|_2^2 \leq \frac{3}{2}p_\lambda(\boldsymbol{\beta}_0) - \frac{1}{2}p_\lambda(\widetilde{\boldsymbol{\beta}}_n). \tag{S3.1}$$

Moreover, from Lemma 5 in Loh and Wainwright (2015), it holds that

$$0 \leq 3p_\lambda(\boldsymbol{\beta}_0) - p(\widetilde{\boldsymbol{\beta}}_n) \leq \lambda L(3\|(\widetilde{\Delta}_n)_A\|_1 - \|(\widetilde{\Delta}_n)_{A^c}\|_1), \tag{S3.2}$$

where A is the index set of the s largest elements of $\widetilde{\boldsymbol{\beta}}_n$ in magnitude. Combining (S3.1) and (S3.2), we have

$$\left(\alpha - \frac{\mu}{4}\right)\|\widetilde{\Delta}_n\|_2^2 \leq \frac{3\lambda L}{2}\|(\widetilde{\Delta}_n)_A\|_1 \leq \frac{3\lambda L\sqrt{s}}{2}\|\widetilde{\Delta}_n\|_2.$$

As a result,

$$\|\widetilde{\Delta}_n\|_2 \le \frac{6\sqrt{s}\lambda L}{4\alpha - \mu}.$$

It is also implied by (S3.2) that $\|(\widetilde{\Delta}_n)_{A^c}\|_1 \le 3\|(\widetilde{\Delta}_n)_A\|_1$, which leads to

$$\|\widetilde{\Delta}_n\|_1 \le \|(\widetilde{\Delta}_n)_A\|_1 + \|(\widetilde{\Delta}_n)_{A^c}\|_1 \le 4\|(\widetilde{\Delta}_n)_A\|_1 \le 4\sqrt{s}\|\widetilde{\Delta}_n\|_2.$$

This accomplishes the proof of this lemma. □

We next conduct the stochastic analysis to verify that the "good" event and LRSC condition hold with high probability in Lemmas S4 and S5, respectively. Their technical proofs can be found in Section S4.

Lemma S4. *If Assumption 5 holds, then*

$$\|n^{-1}\nabla L(\boldsymbol{\beta}_0)\|_\infty \le C_S\sqrt{\frac{\log p}{n}} \tag{S3.3}$$

with probability at least $1 - c_1 p^{-c_2}$ *for some positive constants* c_1, c_2 *and* C_S.

Lemma S5. *Suppose that Assumptions 4-6 hold. Given the sample size* $n \ge c' \log p$ *for some* $c' > 0$, *it holds that*

$$\mathcal{E}_n(\Delta) \ge \alpha\|\Delta\|_2^2 - \eta\sqrt{\frac{\log p}{n}}\|\Delta\|_1 \text{ for all } r \le \|\Delta\|_2 \le R$$

with probability at least $1 - c_1 p^{-c_2} - K\log(\sqrt{d}pr/r_l)p^{-c^2}$ *for any* $c > 1$, *where* $\alpha = 0.5 f_{\min}\lambda^0_{\min}$, $\eta = KC_E d 2^{d+1} + 2KL_Q C_X c + C_S$.

Proof of Theorem 4. The proof of this theorem follows from the deterministic analysis in Lemma S3 and stochastic analysis in Lemmas S4, S5. □

S4 Proofs of four auxiliary lemmas

S4.1 Proofs of Lemmas S1 and S2

Proof of Lemma S1. Since $G(\tau)$ is symmetric about $\tau = 0.5$, we only need to show (1). Let $\widetilde{a} = \widetilde{\theta}_3 - 1$, $a = \theta_3 - 1$. Function $G(\tau)$ can be rewritten into

$$G(\tau) = \frac{\tau^{\widetilde{a}} + (1-\tau)^{\widetilde{a}}}{\tau^{a} + (1-\tau)^{a}},$$

and its derivative function has the form of $\dot{G}(\tau) = H(\tau)/\{\tau^a + (1-\tau)^a\}^2$, where

$$H(\tau) = \{\widetilde{a}\tau^{\widetilde{a}-1} - \widetilde{a}(1-\tau)^{\widetilde{a}-1}\}\{\tau^a + (1-\tau)^a\} - \{a\tau^{a-1} - a(1-\tau)^{a-1}\}\{\tau^{\widetilde{a}} + (1-\tau)^{\widetilde{a}}\}.$$

Note that

$$H(\tau) = \underbrace{(\widetilde{a}-a)\tau^{\widetilde{a}+a-1} - (\widetilde{a}-a)(1-\tau)^{\widetilde{a}+a-1}}_{A} +$$
$$\underbrace{\tau^{\widetilde{a}-1}(1-\tau)^{a-1}(\widetilde{a}(1-\tau) + a\tau) - (1-\tau)^{\widetilde{a}-1}\tau^{a-1}(a(1-\tau) + \widetilde{a}\tau)}_{B}.$$

Since $1-\tau < \tau$, $\widetilde{a} + a - 1 < -1$, $\widetilde{a} - a > 0$, we have $A < 0$. If we can show $B < 0$, the proof is completed. Because $\widetilde{a}, a < 0$, $B < 0$ is equivalent to

$$\frac{\widetilde{a}(1-\tau) + a\tau}{\widetilde{a}\tau + a(1-\tau)} > \left(\frac{1-\tau}{\tau}\right)^{\widetilde{a}-a}. \tag{S4.1}$$

Since $\tau > 0.5$, $\widetilde{a} > a$, we have $0 < \{(1-\tau)/\tau\}^{\widetilde{a}-a} < 1$ and $\{\widetilde{a}(1-\tau) + a\tau\}/\{\widetilde{a}\tau + a(1-\tau)\} > 1$. Then the equation (S4.1) holds. $\square$

Proof of Lemma S2. **(a).** Denote

$$f_{ik}(u) = \frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \int_0^1 I(e_{ik} \leq u' \frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} s) - I(e_{ik} \leq 0) ds,$$

and $D_n(u) = n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n}\left[f_{ik}(u) - E\left\{f_{ik}(u)|\mathbf{X}_i\right\}\right]$. It then holds that $\xi_{ik}(u) = u' f_{ik}(u)$ and

$$\Pi_{1n}(u_n) = \sum_{k=1}^{K}\sum_{i=1}^{n}\left[\xi_{ik}(u_n) - E\left\{\xi_{ik}(u_n)|\mathbf{X}_i\right\}\right] = \sqrt{n}u_n' D_n(u_n).$$

To prove (a), it is sufficient to show that, for any $\eta > 0$, $\sup_{\|u\|\leq\eta}\|D_n(u)\|/(1+\sqrt{n}\|u\|) = o_p(1)$. Let $D_n^{(jk)}(u) = n^{-1/2}\sum_{i=1}^{n}\left[f_{ik}^{(j)}(u) - E\left\{f_{ik}^{(j)}(u)|\mathbf{X}_i\right\}\right]$, where $f_{ik}^{(j)}(u)$ is the jth element of $f_{ik}(u)$. We next use the bracketing method in Pollard (1985) to prove that

$$\sup_{\|u\|\leq\eta}\frac{|D_n^{(jk)}(u)|}{1+\sqrt{n}\|u\|} = o_p(1)$$

for each $1 \leq k \leq K$ and $1 \leq j \leq dp$. Without confusion, we abbreviate $f_{ijk}(u_n)$ and $D_n^{(jk)}(u)$ to $f_i(u_n)$ and $D_n(u)$, respectively, in the following proof for simplicity.

Without loss of generality, we assume that $\partial Q_{ik}(\boldsymbol{\beta}_0)/\partial\beta_j \geq 0$. Let $\mathcal{F} = \{f_i(u) : \|u\| \leq \eta\}$ be a collection of functions indexed by u. For any fixed $\epsilon > 0$ and $0 < \delta \leq \eta$, there exists a sequence of small cubes $\{B_{\epsilon\delta/C_1}(u_l)\}_{l=1}^{L_\epsilon}$ to cover $B_\delta(0)$, where $B_r(\zeta)$ is an open neighborhood of ζ with radius r, C_1 is a constant defined later, L_ϵ is an integer less than $c_0\epsilon^{-dp}$ and c_0 is a constant independent of ϵ and δ. Moreover, we assume that $U_l(\delta) \subseteq B_{\epsilon\delta/C_1}(u_l)$, and $\{U_l(\delta)\}_{l=1}^{L_\epsilon}$ forms a partition of $B_\delta(0)$. For any $u \in U_l(\delta)$, we define the bracketing functions as

$$f_i^{\pm}(u) = \frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial\beta_j}\int_0^1 \digamma_{ik}\left\{u'\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}s \pm \frac{\epsilon\delta}{C_1}\left\|\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\|\right\}ds,$$

where $\digamma_{ik}(s) = I(e_{ik} \leq s) - I(e_{ik} \leq 0)$. It then holds that, for all $u \in U_l(\delta)$,

$$f_i^-(u_l) \leq f_i(u) \leq f_i^+(u_l) \tag{S4.2}$$

and

$$E[f_i^+(u_l) - f_i^-(u_l)|\mathbf{X}_i] \leq \frac{2\epsilon\delta}{C_1} \sup_{y,x} f_Y(y|x) \left\| \frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\|^2 := \frac{\epsilon\delta\Delta_i}{C_1}, \tag{S4.3}$$

where $\Delta_i = 2\sup_{y,x} f_Y(y|x)\|\partial Q_{ik}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|^2$. Define the event

$$E_n = \left\{ \frac{1}{nC_1} \sum_{i=1}^n \Delta_i < 2 \right\}.$$

By taking $C_1 = E\Delta_i$ and applying the law of large numbers, we have $P(E_n) \to 1$ as $n \to \infty$.

Put $\delta_m = 2^{-m}\eta$. Denote $B(m) = B_{\delta_m}(0)$ for simplicity, and let $A(m) = B(m)/B(m+1)$ be the annulus. Fix $\epsilon > 0$, and assume that $\{U_l(\delta_m)\}_{l=1}^{L_\epsilon}$ is a partition of $B(m)$. We first consider the upper tail. For any $u \in U_l(\delta_m)$, if event E_n holds, then

$$\begin{aligned} D_n(u) &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[f_i^+(u_l) - E\left\{ f_i^-(u_l)|\mathbf{X}_i \right\} \right] = D_n^+(u_l) + \frac{1}{\sqrt{n}} \sum_{i=1}^n E\left\{ f_i^+(u_l) - f_i^-(u_l)|\mathbf{X}_i \right\} \\ &\leq D_n^+(u_l) + \sqrt{n}\epsilon\delta_m \left\{ \frac{1}{nC_1} \sum_{i=1}^n \Delta_i \right\} \leq D_n^+(u_l) + 2\sqrt{n}\epsilon\delta_m, \end{aligned}$$

where

$$D_n^+(u_l) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[f_i^+(u_l) - E\left\{ f_i^+(u_l)|\mathbf{X}_i \right\} \right].$$

Note that

$$
\begin{aligned}
&E\left\{D_n^+(u_l)\right\}^2 \leq \frac{1}{n}\sum_{i=1}^n E\left\{f_i^+(u_l)\right\}^2 = \frac{1}{n}\sum_{i=1}^n E[E\{(f_i^+(u_l))^2|\mathbf{X}_i\}] \\
\leq & \frac{1}{n}\sum_{i=1}^n E\left[\left\{\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \beta_j}\right\}^2 \int_0^1 E\left[\left|F_{ik}\left\{u_l'\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}s + \frac{\epsilon\delta_m}{C_1}\left\|\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right\|\right\}\right| |\mathbf{X}_i\right] ds\right] \\
\leq & \frac{1}{n}\sum_{i=1}^n E\left[\left\{\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \beta_j}\right\}^2 \sup_{|y|\leq C\delta_m\|\partial Q_{ik}(\boldsymbol{\beta}_0)/\partial\beta\|} |F_Y(Q_{ik}(\boldsymbol{\beta}_0)+y|\mathbf{X}_i) - F_Y(Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i)|\right] \\
\leq & C\delta_m \sup_{y,x} f_Y(y|x) E\left\|\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right\|^3 := \pi(\delta_m),
\end{aligned}
$$

where $\pi(\delta_m)$ goes to 0 as δ_m goes to 0. Moreover, for any $u \in A(m)$, $1 + \sqrt{n}\|u\| > \sqrt{n}\delta_{m+1} = \sqrt{n}\delta_m/2$. As a result,

$$
\begin{aligned}
&P\left\{\sup_{u\in A(m)} \frac{D_n(u)}{1+\sqrt{n}\|u\|} > 6\epsilon, E_n\right\} \\
&\quad \leq P\left\{\sup_{u\in A(m)} D_n(u) > 3\sqrt{n}\epsilon\delta_m, E_n\right\} \leq P\left\{\max_{1\leq l\leq L_\epsilon}\sup_{u\in U_l(\delta_m)\cap A(m)} D_n(u) > 3\sqrt{n}\epsilon\delta_m, E_n\right\} \\
&\quad \leq P\left\{\max_{1\leq l\leq L_\epsilon} D_n^+(u_l) > \sqrt{n}\epsilon\delta_m, E_n\right\} \leq L_\epsilon \max_{1\leq l\leq L_\epsilon} \frac{E\left\{D_n^+(u_l)\right\}^2}{n\epsilon^2\delta_m^2} \leq L_\epsilon \frac{\pi(\delta_m)}{n\epsilon^2\delta_m^2}.
\end{aligned}
$$

Similarly, we can obtain the same bound for the lower tail, and hence

$$
P\left\{\sup_{u\in A(m)} \frac{|D_n(u)|}{1+\sqrt{n}\|u\|} > 6\epsilon, E_n\right\} \leq 2L_\epsilon \frac{\pi(\delta_m)}{n\epsilon^2\delta_m^2}.
$$

We split the set of $\{u : \|u\| \leq \eta\}$ into $B(m_n+1)$ and $B(m_n+1)^c = \cup_{m=0}^{m_n} A(m)$, where m_n satisfies $n^{-1/2} \leq 2^{-m_n} < 2n^{-1/2}$. It can be verified that

$$
\begin{aligned}
&P\left\{\sup_{u\in B(m_n+1)^c}\frac{|D_n(u)|}{1+\sqrt{n}\|u\|}>6\epsilon\right\}\\
&\leq \sum_{m=0}^{m_n}P\left\{\sup_{u\in A(m)}\frac{|D_n(u)|}{1+\sqrt{n}\|u\|}>6\epsilon, E_n\right\}+P(E_n^c)\leq \sum_{m=0}^{m_n}\frac{2L_\epsilon\pi(\delta_m)}{n\epsilon^2\eta^2}2^{2m}+P(E_n^c)\\
&\leq \frac{1}{n}\sum_{m=0}^{m_\epsilon-1}\frac{CL_\epsilon}{\epsilon^2\eta^2}2^{2m}+\frac{\epsilon}{n}\sum_{m=m_\epsilon}^{m_n}2^{2m}+P(E_n^c)\leq O(\frac{1}{n})+4\epsilon+P(E_n^c),
\end{aligned}
\tag{S4.4}
$$

where m_ϵ at the last line is chosen such that $2L_\epsilon\pi_n(\delta_m)/(\epsilon^2\eta^2) < \epsilon$ for all $m > m_\epsilon$, since $\pi(\delta_m) \to 0$ as $k \to \infty$. Consider the set of $B(m_n + 1)$. For any $u \in U_l(\delta_{m_n+1})$, by using a similar argument, we can show that $D_n(u) \leq D_n^+(u_l) + 2\sqrt{n}\epsilon\delta_{m_n+1} \leq D_n^+(u_l) + 2\epsilon$. As a result, due to the fact that $1 + \sqrt{n}\|u\| > 1$,

$$
P\left\{\sup_{u\in B(m_n+1)}\frac{D_n(u)}{1+\sqrt{n}\|u\|}>3\epsilon, E_n\right\}\leq P\left\{\max_{1\leq l\leq L_\epsilon}D_n^+(u_l)>\epsilon, E_n\right\}\leq \frac{L_\epsilon\pi(\delta_{m_n+1})}{\epsilon^2},
$$

and the bound for the lower tail can be obtained similarly. Thus,

$$
P\left\{\sup_{u\in B(m_n+1)}\frac{|D_n(u)|}{1+\sqrt{n}\|u\|}>3\epsilon\right\}\leq \frac{2L_\epsilon\pi(\delta_{m_n+1})}{\epsilon^2}+P(E_n^c),
\tag{S4.5}
$$

which, together with (S4.4) and the fact that $\pi(\delta_{m_n+1}) \to 0$ and $P(E_n^c) \to 0$ as $n \to \infty$, accomplished the proof of (a).

(b). Note that

$$\begin{aligned}
\Pi_{2n}(u) &= \sum_{k=1}^{K}\sum_{i=1}^{n} E\left\{\xi_{ik}(u)|\mathbf{X}_i\right\} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} \int_0^{q_{1ik}(u)} \left[F_Y\left(s+Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i\right) - F_Y\left(Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i\right)\right] ds \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} \left\{ \int_0^{q_{1ik}(u)} f_Y\left(Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i\right) sds \right. \\
&\qquad\qquad \left. + \int_0^{q_{1ik}(u)} \left[f_Y(\zeta_k^*|\mathbf{X}_i) - f_Y\left(Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i\right)\right] sds \right\} \\
&= (\sqrt{n}u)' K_{1n}(\sqrt{n}u) + (\sqrt{n}u)' K_{2n}(u)(\sqrt{n}u),
\end{aligned} \tag{S4.6}$$

where ζ_k^* is between $Q_{ik}(\boldsymbol{\beta}_0)$ and $s + Q_{ik}(\boldsymbol{\beta}_0)$,

$$K_{1n} = \frac{1}{2n}\sum_{k=1}^{K}\sum_{i=1}^{n} f_Y\left(Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i\right) \frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}'},$$

and

$$K_{2n}(u) = \frac{1}{n\|u\|^2}\sum_{k=1}^{K}\sum_{i=1}^{n} \int_0^{q_{1ik}(u)} \{f_Y(\zeta_k^*|\mathbf{X}_i) - f_Y\left(Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i\right)\} sds.$$

Note that

$$\begin{aligned}
&\sup_{\|u\|\le\eta} |K_{2n}(u)| \le \sup_{\|u\|\le\eta} \frac{1}{n\|u\|^2}\sum_{k=1}^{K}\sum_{i=1}^{n} \int_{-|q_{1ik}(u)|}^{|q_{1ik}(u)|} \{|f_Y(\zeta_k^*|\mathbf{X}_i) - f_Y\left(Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i\right)|\} sds \\
&\le \frac{1}{2n}\sum_{k=1}^{K}\sum_{i=1}^{n} \left\{ \sup_{|y|\le C\eta\|\partial Q_{ik}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|} |f_Y(y+Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i) - f_Y\left(Q_{ik}(\boldsymbol{\beta}_0)|\mathbf{X}_i\right)| \right\} \left\| \frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\|^2,
\end{aligned}$$

which, together with Assumption 3, implies that $\sup_{\|u\|\le\eta} |K_{2n}(u)| = o_p(1)$, and hence $K_{2n}(u_n) = o_p(1)$. Moreover, by the law of large numbers, $K_{1n} = 0.5\Omega_1 + o_p(1)$. Thus, the proof of (b) is accomplished.

(c). Note that

$$\Pi_{3n}(u) = \sum_{k=1}^{K}\sum_{i=1}^{n}\left[q_{2ik}(u)\{I(e_{ik}<0)-\tau_k\} + \int_{q_{1ik}(u)}^{q_{1ik}(u)+q_{2ik}(u)}\{I(e_{ik}\leq s)-I(e_{ik}\leq 0)\}ds\right]$$
$$= (\sqrt{n}u)'\left\{\sum_{k=1}^{K}K_{3n}(\boldsymbol{\beta}^*)\right\}(\sqrt{n}u) + K_{4n}(u),$$

where $q_{1ik}(u) = u'\partial Q_{ik}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}$, $q_{2ik}(u) = 0.5u'\partial^2 Q_{ik}(\boldsymbol{\beta}_k^*)/(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}')u$,

$$K_{3n}(\boldsymbol{\beta}^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 Q_{ik}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\{I(e_{ik}<0)-\tau_k\}$$

and

$$K_{4n}(u) = \sum_{k=1}^{K}\sum_{i=1}^{n}\int_{q_{1ik}(u)}^{q_{1ik}(u)+q_{2ik}(u)} F_{ik}(s)ds.$$

For $K_{3n}(\boldsymbol{\beta}^*)$, it holds that $E[\sup_{\boldsymbol{\beta}^*\in\Theta}\|\partial^2 Q_{ik}(\boldsymbol{\beta}^*)/(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}')\||I(e_{ik}<0)-\tau_k|] < \infty$, and $E[\partial^2 Q_{ik}(\boldsymbol{\beta}^*)/(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}')\{I(e_{ik}<0)-\tau_k\}] = 0$. Then, by applying Theorem 4.2.1 in Amemiya (1985), we have

$$\sup_{\boldsymbol{\beta}^*\in\Theta}|K_{3n}(\boldsymbol{\beta}^*)| = o_p(1).$$

On the other hand,

$$\frac{K_{4n}(u)}{n\|u\|^2} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_{0}^{q_{2ik}(u)/\|u\|^2} F_{ik}(\|u\|^2 s + q_{1ik}(u))ds := \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}J_{ik}(u).$$

For any $\eta > 0$,

$$\sup_{\|u\|\leq\eta}|J_{ik}(u)| \leq \int_{-\Lambda}^{\Lambda}\left[F_{ik}\left\{\eta^2\Lambda + \eta\left\|\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\|\right\} - F_{ik}\left\{-\eta^2\Lambda - \eta\left\|\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\|\right\}\right]ds,$$

and

$$
\begin{aligned}
& E\left\{\sup_{\|u\|\leq\eta}|J_{ik}(u)|\right\} \\
& \leq 2E\left[\Lambda\left\{F_Y\left(\eta^2\Lambda+\eta\left\|\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\||\mathbf{X}_i\right)-F_Y\left(-\eta^2\Lambda-\eta\left\|\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\||\mathbf{X}_i\right)\right\}\right] \\
& \leq 4\sup_{y,x} f_Y(y|x)\eta^2 E\left(\Lambda^2\right)+4\sup_{y,x} f_Y(y|x)\eta E\left(\Lambda\left\|\frac{\partial Q_{ik}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\|\right)\to 0,
\end{aligned}
$$

as $\eta \to 0$, where $\Lambda = 0.5\sup_{\boldsymbol{\beta}\in\Theta}\|\partial^2 Q_{ik}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'\|$. Thus, $K_{4n}(u_n) = o_p(n\|u_n\|^2)$. This completes the proof of (c). □

S4.2 Proofs for Lemmas S4 and S5

Proof of Lemma S4. Note that

$$
n^{-1}\nabla L(\boldsymbol{\beta}_0)=\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\psi_{\tau_k}(e_{ik})\frac{\partial Q(\tau_k,\boldsymbol{\theta}(\boldsymbol{X}_i,\boldsymbol{\beta}_0))}{\partial\boldsymbol{\gamma}}\otimes\boldsymbol{X}_i.
$$

where $e_{ik}=Y_i-Q(\tau_k,\boldsymbol{\theta}(\boldsymbol{X}_i,\boldsymbol{\beta}_0))$, $\psi_\tau(u)=\tau-I(u<0)$, and $\boldsymbol{X}_i=(X_{1i},...,X_{pi})'$. For $1\leq j\leq d$, denote $\xi_j(\boldsymbol{X}_i)=\partial Q(\tau_k,\boldsymbol{\theta}(\boldsymbol{X}_i,\boldsymbol{\beta}_0))/\partial\gamma_j$ and, from Assumption 5, it holds that $|\xi_j(\boldsymbol{X}_i)|\leq L_Q$.

It can be verified that, conditional on $\boldsymbol{X}_i$, $\psi_{\tau_k}(e_{ik})$ is sub-Gaussian with the parameter of 0.5, and hence, for any $\delta>0$,

$$
E\exp[\delta n^{-1}\psi_{\tau_k}(e_{ik})\xi_j(\boldsymbol{X}_i)X_{li}]\leq E\exp\{[\delta n^{-1}\xi_j(\boldsymbol{X}_i)X_{li}]^2/8\}\leq\exp\{[\delta n^{-1}L_QC_X]^2/8\}.
$$

As a result, for each $t > 0$, $1 \leq j \leq d$ and $1 \leq l \leq p$,

$$\begin{aligned}
&P\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\psi_{\tau_k}(e_{ik})\xi_j(\boldsymbol{X}_i)X_{li} > t\right) \\
&\leq \inf_{\delta>0}\exp(-\delta t)\prod_{i=1}^{n}\prod_{k=1}^{K}E\exp[\delta n^{-1}\psi_{\tau_k}(e_{ik})\xi_j(\boldsymbol{X}_i)X_{li}] \\
&\leq \inf_{\delta>0}\exp\left(\frac{n^{-1}K(L_QC_X)^2}{8}\delta^2 - \delta t\right) \leq \exp\left(\frac{-2nt^2}{K(L_QC_X)^2}\right),
\end{aligned}$$

which implies that

$$P\left(\|n^{-1}\nabla L(\boldsymbol{\beta}_0)\|_\infty \geq t\right) \leq \exp\left(\frac{-2nt^2}{K(L_QC_X)^2} + \log(2dp)\right).$$

By letting $t = C_S\sqrt{\log p/n}$ with $C_S > \sqrt{0.5K}L_QC_X$, we accomplish the proof with $c_1 = 2d$ and $c_2 = 2C_S^2/[K(L_QC_X)^2] - 1 > 0$. $\square$

Proof of Lemma S5. We first show the strong convexity of $\bar{L}(\boldsymbol{\beta}) = E[n^{-1}L_n(\boldsymbol{\beta})] = E[\sum_{k=1}^{K}\rho_{\tau_k}\{Y - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}))\}]$. Let $Q^*(\boldsymbol{\beta}) = (Q(\tau_1, \boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\beta})), \ldots, Q(\tau_K, \boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\beta})))'$ and, by Taylor expansion,

$$E\|Q^*(\boldsymbol{\beta}) - Q^*(\boldsymbol{\beta}_0)\|_2^2 = E\left\|\frac{\partial Q^*(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}}\Delta\right\|_2^2 = \Delta'\Omega_2(\boldsymbol{\beta}^*)\Delta,$$

where $\Delta = \boldsymbol{\beta} - \boldsymbol{\beta}_0$, and $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$. Note that $\partial\bar{L}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta} = 0$ and hence, by Knight's identity at (S2.1) and Assumption 6, it can be verified that

$$\begin{aligned}
&\bar{\mathcal{E}}(\Delta) := \bar{L}(\boldsymbol{\beta}) - \bar{L}(\boldsymbol{\beta}_0) - \langle \Delta, \partial \bar{L}(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta} \rangle = \bar{L}(\boldsymbol{\beta}) - \bar{L}(\boldsymbol{\beta}_0) \\
=&E\left(\sum_{k=1}^{K} \int_0^{Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\beta})) - Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\beta}_0))} \{F_{Y|\boldsymbol{X}}(Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\beta}_0)) + s) - F_{Y|\boldsymbol{X}}(Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\beta}_0)))\} ds \right) \\
\geq& 0.5 f_{\min} E\|Q^*(\boldsymbol{\beta}) - Q^*(\boldsymbol{\beta}_0)\|_2^2 \geq 0.5 f_{\min} \lambda_{\min}^0 \|\Delta\|_2^2
\end{aligned} \tag{S4.7}$$

uniformly for $\{\Delta \in \mathbb{R}^{dp} : \|\Delta\|_2 \leq R\}$.

For $1 \leq k \leq K$, denote $L_n^{(k)}(\boldsymbol{\beta}) = \sum_{i=1}^n \rho_{\tau_k}\{Y_i - Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}_i, \boldsymbol{\beta}))\}$ and $\bar{L}_k(\boldsymbol{\beta}) = E[n^{-1} L_n^{(k)}(\boldsymbol{\beta})] = E[\rho_{\tau_k}\{Y - Q(\tau_k, \boldsymbol{\theta}(\mathbf{X}, \boldsymbol{\beta}))\}]$. Note that $L_n(\boldsymbol{\beta}) = \sum_{k=1}^K L_n^{(k)}(\boldsymbol{\beta})$ and $\bar{L}(\boldsymbol{\beta}) = \sum_{k=1}^K \bar{L}_k(\boldsymbol{\beta})$. Let $\mathcal{E}_k^*(\Delta) = |n^{-1} L_n^{(k)}(\boldsymbol{\beta}) - n^{-1} L_n^{(k)}(\boldsymbol{\beta}_0) - \{\bar{L}_k(\boldsymbol{\beta}) - \bar{L}_k(\boldsymbol{\beta}_0)\}|$, and we next prove that, uniformly for $r \leq \|\Delta\|_2 \leq R$,

$$\mathcal{E}_k^*(\Delta) \leq C_{\mathcal{E}} \sqrt{\frac{\log p}{n}} \|\Delta\|_1, \tag{S4.8}$$

with probability at least $1 - \log(\sqrt{dp} r / r_l) p^{-c^2}$ for any $c > 1$, where $C_{\mathcal{E}} = C_E d 2^{d+1} + 2 L_Q C_X c$. As in Theorem 9.34 in Wainwright (2019), we use the peeling argument, which is a common strategy in empirical process theory.

Tail bound for fixed radii: Define a set $\mathcal{C}(r_1) := \left\{\Delta \in \mathbb{R}^{dp} : \|\Delta\|_1 \leq r_1\right\}$ for a fixed radii $r_1 > 0$, and a random variable $A_n(r_1) = r_1^{-1} \sup_{\boldsymbol{\beta} \in \mathcal{C}(r_1)} \mathcal{E}_k(\Delta)$. We next show that, for any $t > 0$,

$$A_n(r_1) \leq C_E d 2^d \sqrt{\frac{\log p}{n}} + L_Q C_X \sqrt{\frac{t}{n}}. \tag{S4.9}$$

with probability at least $1 - e^{-t}$.

For $1 \leq i \leq n$, denote $\boldsymbol{W}_i = (Y_i, \boldsymbol{X}_i')'$. Note that random variable $A(r_1)$ has a form of $f(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)$, and it is guaranteed by Assumption 5 that

$$|f(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_i, \ldots, \boldsymbol{W}_n) - f(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{i\prime}, \ldots, \boldsymbol{W}_n)| \leq n^{-1} L_Q C_X,$$

i.e., if we replace $\boldsymbol{W}_i$ by $\boldsymbol{W}_{i'}$, while keep other $\boldsymbol{W}_j$ fixed, then $A(r_1)$ changes by at most $n^{-1}L_Q C_X$. As a result, by the bounded differences inequality and for any $t > 0$,

$$A_n(r_1) \leq E[A_n(r_1)] + L_Q C_X \sqrt{\frac{t}{n}} \tag{S4.10}$$

with probability at least $1 - e^{-t}$.

In addition, it is implied by Assumption 5 that, for all $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{dp}$,

$$|\rho_{\tau_k} \{Y_i - Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}_i, \boldsymbol{\beta}))\} - \rho_{\tau_k}\{Y_i - Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}))\}| \leq L_Q \sum_{l=1}^{d} |\boldsymbol{X}_i'(\boldsymbol{\beta}_l - \widetilde{\boldsymbol{\beta}}_l)|,$$

which leads to

$$\begin{aligned}
E[A_n(r_1)] \leq & \frac{2}{nr_1} E\left(\sup_{\boldsymbol{\beta} \in \mathcal{C}(r_1)} \left| \sum_{i=1}^{n} \epsilon_i \left[\rho_{\tau_k}\{Y_i - Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}_i, \boldsymbol{\beta}))\} - \rho_{\tau_k}\{Y_i - Q(\tau_k, \boldsymbol{\theta}(\boldsymbol{X}_i, \boldsymbol{\beta}_0))\}\right] \right| \right) \\
\leq & \frac{C2^d}{nr_1} E\left(\sup_{\boldsymbol{\beta} \in \mathcal{C}(r_1)} \left| \sum_{l=1}^{d} \sum_{i=1}^{n} V_{il} (\boldsymbol{X}_i'(\boldsymbol{\beta}_l - \boldsymbol{\beta}_{0l})) \right| \right) \leq \frac{C2^d}{n} E\left(\sum_{l=1}^{d} \left\| \sum_{i=1}^{n} V_{il} \boldsymbol{X}_i \right\|_\infty \right) \\
\leq & \frac{Cd2^d}{n} E\left(\left\| \sum_{i=1}^{n} V_{i1} \boldsymbol{X}_i \right\|_\infty \right) \leq C_E d 2^d \sqrt{\frac{\log p}{n}},
\end{aligned} \tag{S4.11}$$

where $\{\epsilon_i\}$ and $\{V_{il}\}$ are $i.i.d.$ Rademacher and standard Gaussian random variables, respectively, the first inequality is due to the symmetrization theorem (Loh and Wainwright, 2015, Lemma 12), the second one is by the multivariate contraction theorem (van de Geer, 2016, Theorem 16.3), the third one is due to the fact $|\boldsymbol{X}'\boldsymbol{\beta}| \leq \|\boldsymbol{X}\|_\infty \|\boldsymbol{\beta}\|_1$, and the last one is by Lemma 16 of Loh and Wainwright (2015) given the sample size of $n \geq c' \log p$ for some $c' > 0$. The upper bound at (S4.9) holds by combining (S4.10) and (S4.11).

Extension to uniform radii via peeling: Define a sequence of sets $\Theta_l := \{\Delta \in \mathbb{R}^{dp} : 2^{l-1} r \leq \|\Delta\|_1 \leq 2^l r\}$ with $1 \leq l \leq N = \log(\sqrt{dp}R/r)$. It can be verified that $\{\Delta \in \mathbb{R}^{dp} : r \leq \|\Delta\|_2 \leq R\} \subseteq \{\Delta \in \mathbb{R}^{dp} : r \leq \|\Delta\|_1 \leq \sqrt{dp}R\} \subseteq \cup_{l=1}^N \Theta_l$. As a result,

$$
\begin{aligned}
P & \left(\mathcal{E}_k^*(\Delta) \geq C_{\mathcal{E}} \sqrt{\frac{\log p}{n}} \|\Delta\|_1, \Delta \in \cup_{l=1}^N \Theta_l\right) \\
& \leq \sum_{l=1}^N P\left(\mathcal{E}_k^*(\Delta) \geq 2^{l-1} r C_{\mathcal{E}} \sqrt{\frac{\log p}{n}}, \Delta \in \Theta_l\right) \\
& = \sum_{l=1}^N P\left(\mathcal{E}_k^*(\Delta) \geq (C_E d 2^d)(2^k r) \sqrt{\frac{\log p}{n}} + L_Q C_X (2^k r) \sqrt{\frac{c^2 \log p}{n}}, \Delta \in \Theta_l\right) \\
& \leq \sum_{l=1}^N P\left(A(2^l r) \geq C_E d 2^d \sqrt{\frac{\log p}{n}} + L_Q C_X \sqrt{\frac{c^2 \log p}{n}}\right),
\end{aligned}
$$

where $C_{\mathcal{E}} = C_E d 2^{d+1} + 2 L_Q C_X c$. By applying (S4.9), it holds that

$$
P\left(\mathcal{E}_k^*(\Delta) \geq C_{\mathcal{E}} \sqrt{\frac{\log p}{n}} \|\Delta\|_1, r \leq \|\Delta\|_2 \leq R\right) \leq \sum_{k=1}^N e^{-c^2 \log p} = \log(\sqrt{dp}R/r) p^{-c^2},
$$

i.e. (S4.8) holds.

Finally, from (S4.7), (S4.8) and Lemma S4,

$$
\begin{aligned}
\mathcal{E}_n(\Delta) & \geq \bar{\mathcal{E}}(\Delta) - \sum_{k=1}^K \mathcal{E}_k^*(\Delta) - \|n^{-1} \nabla L(\boldsymbol{\beta}_0)\|_\infty \|\Delta\|_1 \\
& \geq 0.5 f_{\min} \lambda_{\min}^0 \|\Delta\|_2^2 - (K C_{\mathcal{E}} + C_S) \sqrt{\frac{\log p}{n}} \|\Delta\|_1
\end{aligned}
$$

uniformly for $r \leq \|\Delta\|_2 \leq R$ with probability at least $1 - c_1 p^{-c_2} - K \log(\sqrt{dp} r / r_l) p^{-c^2}$ for any $c > 1$. This accomplishes the proof. $\square$

S5 Additional simulation studies

S5.1 QIR using GEVD and GPD as quantile functions

This subsection provides additional results for the DGPs of generalized extreme value distribution (GEVD) and generalized Pareto distribution (GPD) as follows

$$
\begin{aligned}
\text{GEVD} : Y_i = Q(U_i, \theta(X_i, \beta)) &= \theta_1(X_i, \beta) + \theta_2(X_i, \beta)\frac{1 - \{-\log(\tau)\}^{\theta_3(X_i,\beta)}}{\theta_3(X_i, \beta)}, \\
\text{GPD} : Y_i = Q(U_i, \theta(X_i, \beta)) &= \theta_1(X_i, \beta) + \theta_2(X_i, \beta)\frac{1 - (1-\tau)^{\theta_3(X_i,\beta)}}{\theta_3(X_i, \beta)}, \qquad \text{(S5.12)}
\end{aligned}
$$

where $\{U_i\}$ are independent and follow Uniform$(0, 1)$, $\mathbf{X}_i = (1, X_{i1}, X_{i2})'$, $\{(X_{i1}, X_{i2})'\}$ is an $i.i.d.$ sequence with 2-dimensional standard normality. The true parameter vector is $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}'_{01}, \boldsymbol{\beta}'_{02}, \boldsymbol{\beta}'_{03})'$, and we set the location parameters $\boldsymbol{\beta}_{01} = (1, 0.5, -1)'$, the scale parameters $\boldsymbol{\beta}_{02} = (1, 0.5, -1)'$ and the tail parameters $\boldsymbol{\beta}_{03} = (1, -1, 1)'$. For the tail index $\theta_3(\mathbf{X}_i, \boldsymbol{\beta})$, before generating the data, we first scale each covariate into the range of $[-0.5, 0.5]$ such that a relatively stable sample can be generated. In addition, g_1, g_2 and g_3 are the inverse of link functions. As in Section 4.1 of the main file, we choose $g_1(x) = x, g_2(x) = \text{softplus}(x)$ and $g_3(x) = 1 - \text{softplus}(x)$, where $\text{softplus}(x) = \log(1 + \exp(x))$ is a smoothed version of $x_+ = \max\{0, x\}$ and hence the name. We consider three sample sizes of $n = 500$, 1000 and 2000, and there are 500 replications for each sample size.

For the data generated by the DGP of GEVD or GPD, we fit them using the quantile index regression (QIR) with the same quantile function and the same link functions. The algorithm for CQR estimation in Section 3 is applied with $K = 10$ and τ_k's being equally spaced over $[\tau_L, \tau_U]$. We consider three quantile ranges of $(\tau_L, \tau_U) = (0.5, 0.99)$, $(0.7, 0.99)$ and $(0.9, 0.99)$ to evaluate the estimation efficiency. Figure S.1 gives the boxplots of three fitted location parameters $\widehat{\boldsymbol{\beta}}_{1n} = (\widehat{\beta}_{1,1}, \widehat{\beta}_{1,2}, \widehat{\beta}_{1,3})'$ for the DGP of GEVD and GPD. It can be seen that both bias and standard deviation decrease as the sample size increase. Moreover, when τ_L decreases, the quantile levels with richer observations will be used for the estimation and, as expected, both bias and standard deviation will decrease. Boxplots for fitted scale and tail parameters show a similar pattern and hence are omitted to save the space. These findings are the same as for the Tukey lambda distribution.

We next evaluate the prediction performance of $Q(\tau^*, \boldsymbol{\theta}(\mathbf{X}, \widehat{\boldsymbol{\beta}}_n))$ at two interesting quantile levels of $\tau^* = 0.991$ and 0.995. As in Section 4.1 of the main file, we consider two values of covariates, $\boldsymbol{X} = (1, 0.1, -0.2)'$ and $(1, 0, 0)'$. Table S.1 presents both the sample mean and standard deviation of prediction errors in terms of squared loss (PESs) across 500 replications for the DGP of GEVD and GPD. As for the DGP of Tukey Lambda distribution in (4.5), the prediction is improved as the sample size becomes larger, and the prediction is more accurate at the 99.1-th quantile level for almost all cases.

We also conduct simulation experiments to evaluate the finite-sample performance of the high-dimensional regularized estimation at (2.4). For the DGPs in (S5.12), we consider $p = 50$, and the true parameter vectors are preserved as in Section 4.2. The sample size is chosen such that $n = \lfloor cs \log p \rfloor$ with $c = 10$, 30, and 50, where $\lfloor x \rfloor$ refers to the largest integer smaller than or equal to x. All other settings are the same as in the low-dimensional case. The algorithm for regularized estimation in Section 3 is used to search for the estimators, and we generate an independent validation set of size $5n$ to select tuning parameter λ by minimizing the composite quantile check loss; see also Wang et al. (2012).

To evaluate the prediction performance of the regularized estimation, Table S.2 lists mean square errors of the predicted conditional quantiles $Q(\tau^*, \boldsymbol{\theta}(\mathbf{X}, \widetilde{\boldsymbol{\beta}}_n))$, as well as the sample standard deviations of prediction errors in squared loss, with $p = 50$ for the DGP of GEVD and GPD, respectively. As for the DGP of Tukey Lambda distribution in Section 4.2, larger sample size leads to much smaller mean square errors. Moreover, when τ_L is larger, the prediction also becomes worse, and it may be due to the lower estimation efficiency. Finally, the prediction at $\tau^* = 0.991$ is more accurate for almost all cases.

To evaluate the performance of variable selection for regularized estimation, Table S.3 reports the selecting results with $p = 50$ and $n = \lfloor cs \log p \rfloor$ for $c = 10$, 30 and 50. When τ_L is larger, P_{AI} decreases, and it indicates the increasing of selection accuracy. In addition, performance improves when sample size gets larger. These findings are the same as for the Tukey lambda distribution.

Overall, it can be seen that the simulation findings are the same for three DGPs of Tukey lambda distribution, GEVD and GPD in estimation, quantile prediction and variable selection. Therefore, in the following we focus on the DGP of Tukey lambda distribution.

S5.2 Sensitivity analysis to the selection of K

We conduct sensitivity analysis of the CQR estimation to the selection of K, using the DGP of Tukey lambda distribution in (4.5) with the same settings. The algorithm for CQR estimation in Section 3 is applied with $K = 5, 10$ and 15, and τ_k's being equally spaced over the quantile range $[\tau_L, \tau_U] = [0.5, 0.99]$.

Figure S.2 gives the boxplots of three fitted location parameters $\widehat{\boldsymbol{\beta}}_{1n} = (\widehat{\beta}_{1,1}, \widehat{\beta}_{1,2}, \widehat{\beta}_{1,3})'$ for $K = 5, 10$ and 15. It can be seen that the results of CQR estimator $\widehat{\boldsymbol{\beta}}_n$ for $K = 5, 10$ and 15 are similar, indicating that the performance of $\widehat{\boldsymbol{\beta}}_n$ is insensitive to the choice of K given a fixed interval $[\tau_L, \tau_U]$. Boxplots for fitted scale and tail parameters show a similar pattern and hence are omitted. Hence, for the selection of τ_ks, it is sensitive for different quantile ranges $[\tau_L, \tau_U]$, while there is not much difference among varying values of K.

We next evaluate the prediction performance of $Q(\tau^*, \boldsymbol{\theta}(\mathbf{X}, \widehat{\boldsymbol{\beta}}_n))$ at two interesting quantile levels of $\tau^* = 0.991$ and 0.995. As in Section 4.1 of the main file, we consider two values of covariates, $\boldsymbol{X} = (1, 0.1, -0.2)'$ and $(1, 0, 0)'$. Table S.4 presents both the sample mean and standard deviation of PESs across 500 replications for $K = 5, 10$ and 15. It is clear that, the prediction performance of CQR estimation is also insensitive to the choice of K given a fixed interval $[\tau_L, \tau_U]$.

S5.3 Sensitivity analysis to the choice of link functions

To assess the robustness of our model to misspecifications due to the link functions, we generate data using alternative link functions that differ from those used in estimation. Actually, different choices of link functions are related to model mis-specification. Note that it is meaningless to evaluate the influence of model mis-specification on the estimator directly, since the true values of parameters under mis-specification are probably undefined. Alternatively, we evaluate the influence of mis-specification due to link functions on conditional quantile prediction.

Specifically, the Tukey lambda distribution is chosen as the quantile function for both the DGP and QIR estimation, and the settings for covariates and true parameter vector are preserved as for DGP (4.5) in the main file. To assess the sensitivity of QIR under different link functions, we examine three mis-specified scenarios for the location, scale and tail indices θ_j's in Cases (i)–(iii), respectively. In each case, data are generated from a specified DGP, and the model is fitted using QIR. Each index θ_j is linked to covariates through a link function $g_j^{-1}(x)$. The default functions used in QIR estimation are as follows

$$g_1(x) = x, \quad g_2(x) = \text{softplus}(x), \quad g_3(x) = 1 - \text{softplus}(x).$$

The three cases for DGP are detailed as follows:

- **Case (i): Misspecified link for θ_1.** We use the DGP with

$$g_1(x) = x + d\text{LeakyReLU}(x, 0.8),$$

while the g_2 and g_3 remain the same as in estimation. Here, $d \geq 0$ is the departure parameter with a larger value indicating greater deviation from the true link function, and $\text{LeakyReLU}(x, 0.8) = x\mathbb{I}(x > 0) + 0.8x\mathbb{I}(x < 0)$. Hence, only g_1 is misspecified for $d > 0$ in this case.

- **Case (ii): Misspecified link for θ_2.** We use the DGP with

$$g_2(x) = \text{softplus}(x) + d\text{ReLU}(x),$$

while g_1 and g_3 match the estimation model. Here, $\text{ReLU}(x) = x\mathbb{I}(x > 0)$. Hence, only g_2 is misspecified for $d > 0$ in this case.

- **Case (iii): Misspecified link for θ_3.** We use the DGP with

$$g_3(x) = 1 - \text{softplus}(x) + d[1 - \text{ReLU}(x)],$$

while the g_1 and g_2 align with the estimation model. Hence, only g_3 is mis-specified for $d > 0$ in this case.

The algorithm for CQR estimation in Section 3 is applied with $K = 10$ and τ_k's being equally spaced over three quantile ranges of $(\tau_L, \tau_U) = (0.5, 0.99)$, $(0.7, 0.99)$ and $(0.9, 0.99)$. We consider three departure levels of $d = 0, 0.5$ or 1, and $d = 0$ corresponds to the correctly specified case. Table S.5 reports both the mean absolute errors and mean square errors across 500 replications for the predicted conditional quantiles at $\tau^* = 0.991$ and $\mathbf{X} = (1, 0.1, -0.2)'$. It can be seen that the misspecification due to link functions makes the prediction less accurate, especially when the departure level increases. Moreover, the errors due to the misspecification in location index θ_1 are the smallest while are the largest in the tail index θ_3. This indicates that the prediction is not sensitive to the departure of location index θ_1, while it is most sensitive to the departure of tail index θ_3. In addition, the errors due to the departure of tail index θ_3 decrease as the quantile range decreases. As a result, when the model is misspecified in tail index, we may choose the interval $[\tau_L, \tau_U]$ narrow and closer to the target quantile level τ^* such that a better result can be achieved.

S5.4 Prediction comparison with other methods

In our simulation studies, we have compared the performance of linear quantile regression (LQR), extremal quantile regression (EQR), degenerated QIR (dQIR), and QIR across three DGPs. In each case, the covariates are kept consistent with those used in DGP (4.5) of the main paper, and the response variable follows a Tukey lambda distribution with $n = 2000$. The three DGPs are described as follows:

- **DGP1:** Location index $\theta_1 = \mathbf{X}'\boldsymbol{\beta}_1$, scale index $\theta_2 = \mathbf{X}'\boldsymbol{\beta}_2$ and tail index $\theta_3 = -0.3$, where $\boldsymbol{\beta}_1 = (1, 0.5, -1)'$ and $\boldsymbol{\beta}_2 = (3, 0.3, 0.3)'$. This setup corresponds to the EQR and dQIR models.
- **DGP2:** Location index $\theta_1 = \mathbf{X}'\boldsymbol{\beta}_1$, scale index $\theta_2 = \text{softplus}(\mathbf{X}'\boldsymbol{\beta}_2)$ and tail index $\theta_3 = -0.3$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (1, 0.5, -1)'$. This represents a simplified version of

the QIR model with a constant tail index and a nonlinear link function for the scale index.

- **DGP3:** Location index $\theta_1 = \mathbf{X}'\boldsymbol{\beta}_1$, scale index $\theta_2 = \text{softplus}(\mathbf{X}'\boldsymbol{\beta}_2)$ and tail index $\theta_3 = 1 - \text{softplus}(\mathbf{X}'\boldsymbol{\beta}_3)$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (1, 0.5, -1)'$ and $\boldsymbol{\beta}_3 = (1, -1, 1)'$. This setting corresponds to our proposed QIR model in general situations.

We implement the CQR estimation algorithm in Section 3 for both QIR and dQIR using $K = 10$ quantile levels, and τ_k's being equally spaced over three quantile ranges of $(\tau_L, \tau_U) = (0.5, 0.99)$, $(0.7, 0.99)$ and $(0.9, 0.99)$. For the QIR, we set $g_1(x) = x$, $g_2(x) = \text{softplus}(x)$ and $g_3(x) = 1 - \text{softplus}(x)$ in the CQR estimation algorithm. For the dQIR, we set $g_1(x) = g_2(x) = x$ and θ_3 to be an unknown parameter in the CQR estimation algorithm. For the LQR, we use the rq function in R package quantreg developed by Roger Koenker. For the EQR, we use the Twostage function in R package EXQR developed by Wang et al. (2012).

Figure S.3 presents boxplots of the prediction bias for conditional quantiles at $\tau = 0.991$, evaluated at $\mathbf{X} = (1, 0.1, -0.2)'$ and $\mathbf{X} = (1, 0, 0)'$, based on 500 replications. The results show that our proposed QIR method yields unbiased predictions across all settings, as each DGP is a submodel of the general QIR framework. However, for DGP1 and DGP2, the QIR exhibits higher prediction variance due to the inclusion of redundant parameters. More importantly, when the tail index varies with covariates, as in DGP3, alternative methods LQR, EQR and dQIR tend to introduce substantial bias as they omit parameters.

S6 Additional details for the empirical analysis

In Section 5 of the main paper, we briefly introduce the three methods for comparison with QIR: (i.) linear quantile regression (LQR) at the level of τ^* with ℓ_1 penalty in Belloni et al. (2019), (ii.) extremal quantile regression (EQR) in (Wang et al., 2012) adapted to high-dimensional data, and (iii.) degenerated QIR (dQIR) with identity link functions for location and scale indices and a constant tail index. Below we provide further details for these three methods.

- **LQR:** For a given target quantile level such as $\tau^* = 0.991$ or 0.995, we fit a high-dimensional LQR with Lasso penalty to obtain the Lasso-penalized estimator

$$(\check{\alpha}(\tau^*), \check{\boldsymbol{\beta}}(\tau^*)) = \arg\min_{\alpha,\boldsymbol{\beta}} \left[\frac{1}{n} \sum_{i=1}^{n} \rho_{\tau^*}(Y_i - \alpha - \mathbf{X}_i' \boldsymbol{\beta}) + \lambda \sum_{j=1}^{d} |\beta_j| \right], \qquad \text{(S6.13)}$$

  where α is the intercept, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)'$ is the slope vector, and the tuning parameter $\lambda > 0$ is selected by minimizing the composite check loss in the testing set; see further details in the last paragraph of Section 5. Then the τ^*th conditional quantile prediction is $\check{Q}_Y(\tau^* \mid \boldsymbol{X}) = \check{\alpha}(\tau^*) + \mathbf{X}' \check{\boldsymbol{\beta}}(\tau^*)$.

- **EQR:** The EQR first estimates the intermediate conditional quantiles using LQR, and then extrapolates these estimates to the high tails based on the estimated tail index. Specifically, the prediction involves the following two steps:

First, for $\widetilde{K} = \lfloor 4.5 n_{\text{train}}^{1/3} \rfloor = 38$ equally spaced quantile levels $\tau_1 < \tau_2 < \cdots < \tau_{\widetilde{K}}$ in the range $[0.96, 0.99]$, we fit Lasso-penalized LQR to calculate the Lasso-penalized estimator at (S6.13). Then the τ_jth conditional quantile prediction is given by $\check{q}_j = \check{\alpha}(\tau_j) + \boldsymbol{X}'\check{\boldsymbol{\beta}}(\tau_j)$, for $j = 1, \ldots, \widetilde{K}$.

Second, the tail index parameter $\widehat{\gamma}$ is estimated by $\widehat{\gamma} = \frac{1}{\widetilde{K}-1} \sum_{j=2}^{\widetilde{K}} \log \frac{\check{q}_j}{\check{q}_1}$; see also equation (2.12) in Wang et al. (2012). Then the extrapolated prediction at a more extreme quantile level $\tau^* = 0.991$ or 0.995 is given by

$$\check{Q}_Y(\tau^* \mid \boldsymbol{X}) = \left(\frac{1-\tau_1}{1-\tau^*} \right)^{\widehat{\gamma}} \check{q}_1.$$

- **dQIR:** To further bridge the connection between EQR and our QIR, we consider a simplified version of QIR, referred to the degenerated QIR (dQIR), which chooses identity link functions for location and scale indices and a constant tail index. The algorithm for CQR estimation in Section 3 is applied with $K = 10$, τ_k's being equally spaced over $[\tau_L, \tau_U]$, and $g_1(x) = g_2(x) = x$ and θ_3 set to be an unknown parameter.

References

Amemiya, T. (1985). *Advanced econometrics*. Boston: Harvard University Press.

Belloni, A., V. Chernozhukov, and K. Kato (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association 114*, 749–758.

Fan, J., W. Gong, and Z. Zhu (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics 212*, 177–202.

Knight, K. (1998). Limiting distributions for L1 regression estimators under general conditions. *The Annals of Statistics 26*, 755–770.

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

Newey, K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2112–2245.

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory 1*, 295–313.

van de Geer, S. A. (2016). *Estimation and testing under sparsity*. Springer.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wang, H. J., D. Li, and X. He (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association 107*, 1453–1464.

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

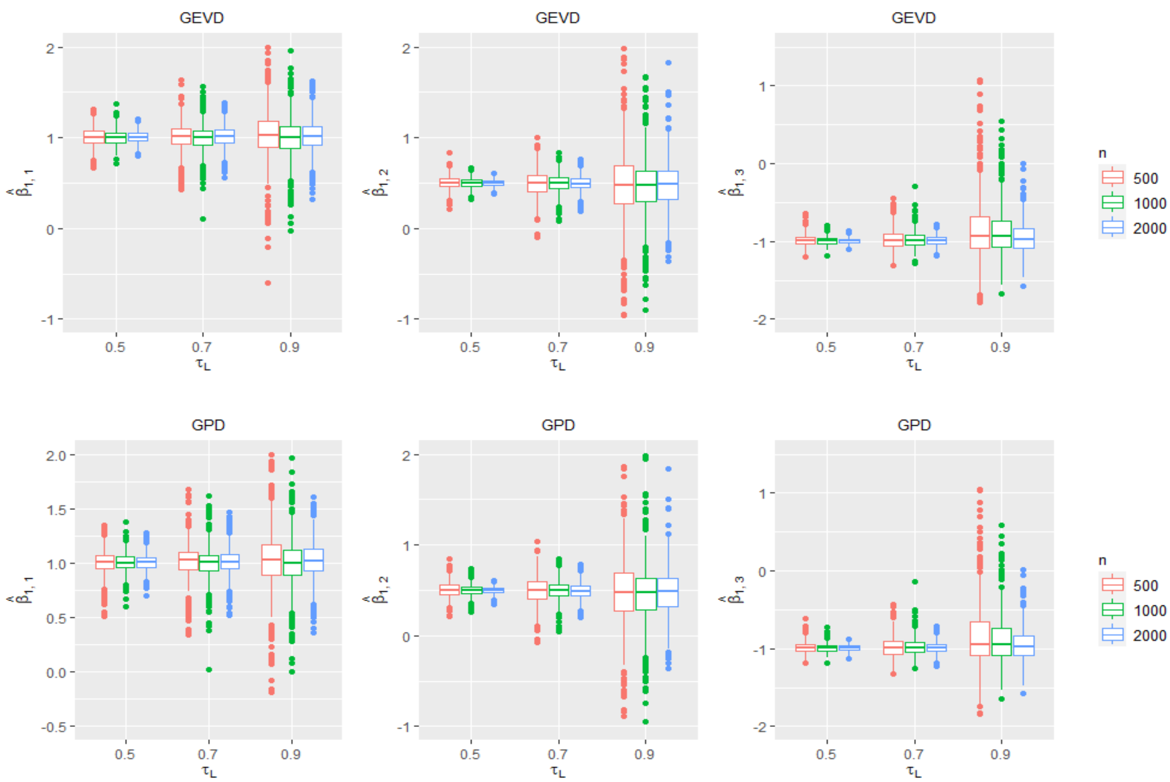

Figure S.1: Boxplots for fitted location parameters of $\widehat{\boldsymbol{\beta}}_{1,1}$ (left panel), $\widehat{\boldsymbol{\beta}}_{1,2}$ (middle panel), and $\widehat{\boldsymbol{\beta}}_{1,3}$ (right panel) under the DGP of GEVD (upper panel) and GPD (lower panel). Sample size is $n = 500$, 1000 or 2000, and the lower bound of quantile range $[\tau_L, \tau_U]$ is $\tau_L = 0.5$, 0.7 or 0.9.

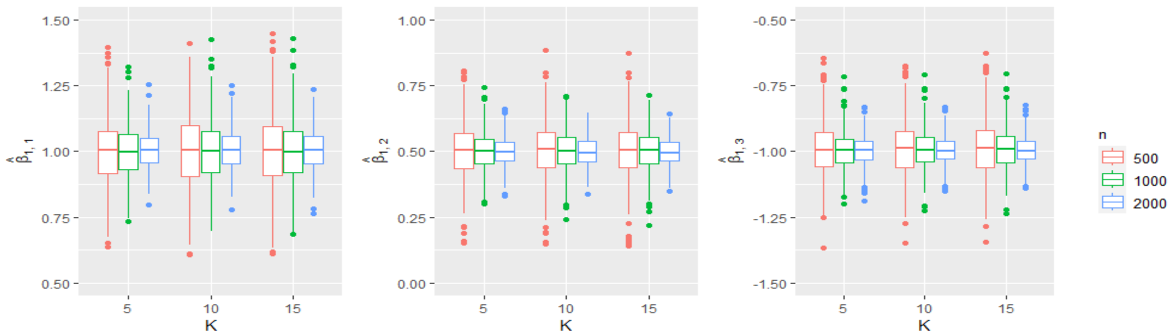

Figure S.2: Boxplots for fitted location parameters of $\widehat{\boldsymbol{\beta}}_{1,1}$ (left panel), $\widehat{\boldsymbol{\beta}}_{1,2}$ (middle panel), and $\widehat{\boldsymbol{\beta}}_{1,3}$ (right panel). Sample size is $n = 500$, 1000 or 2000, and the quantile range $[0.5, 0.99]$ and the number of levels $K = 5$, 10 or 15 are considered for estimation.

Table S.1: Mean square errors of the predicted conditonal quantile $Q(\tau^*, \boldsymbol{\theta}(\mathbf{X}, \widehat{\boldsymbol{\beta}}_n))$ at the level of $\tau^* = 0.991, 0.995$ under the DGP of GEVD or GPD. The values in bracket refer to the corresponding sample standard deviations of prediction errors in squared loss. We denote $\boldsymbol{X} = (1, 0.1, -0.2)'$ as $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X} = (1, 0, 0)'$ as $\boldsymbol{X}^{(2)}$.

			GEVD				GPD			
n	$[\tau_L, \tau_U]$		$\boldsymbol{X}^{(1)}$		$\boldsymbol{X}^{(2)}$		$\boldsymbol{X}^{(1)}$		$\boldsymbol{X}^{(2)}$	
		τ^*	0.991	0.995	0.991	0.995	0.991	0.995	0.991	0.995
500	$[0.5, 0.99]$		1.81(3.40)	3.33(6.90)	5.79(10.69)	13.28(27.15)	1.76(3.38)	3.25(6.80)	5.71(10.90)	13.21(28.37)
	$[0.7, 0.99]$		1.85(3.50)	3.40(6.93)	5.25(9.44)	12.11(25.09)	1.85(3.49)	3.42(6.96)	5.49(11.27)	12.90(31.04)
	$[0.9, 0.99]$		2.88(4.47)	5.32(8.76)	6.98(12.67)	16.15(33.30)	2.85(4.53)	5.31(9.04)	6.68(11.04)	15.47(28.87)
1000	$[0.5, 0.99]$		0.94(1.75)	1.73(3.38)	3.01(4.85)	6.92(12.01)	0.97(1.85)	1.80(3.64)	2.92(4.61)	6.77(11.52)
	$[0.7, 0.99]$		1.18(3.29)	2.15(6.10)	2.91(4.42)	6.69(10.94)	1.23(3.33)	2.26(6.24)	2.90(4.28)	6.68(10.44)
	$[0.9, 0.99]$		1.73(3.19)	3.13(5.66)	3.57(4.78)	8.17(11.66)	1.74(3.27)	3.16(5.81)	3.52(4.71)	8.10(11.54)
2000	$[0.5, 0.99]$		0.80(1.77)	1.49(3.45)	1.62(2.47)	3.74(6.05)	0.83(1.80)	1.55(3.52)	1.58(2.39)	3.69(5.92)
	$[0.7, 0.99]$		1.00(2.18)	1.89(4.41)	1.61(2.85)	3.77(7.68)	1.01(2.25)	1.92(4.55)	1.59(2.58)	3.68(6.64)
	$[0.9, 0.99]$		1.48(3.02)	2.81(6.15)	1.97(2.98)	4.50(7.06)	1.48(3.04)	2.81(6.20)	1.96(2.96)	4.49(7.06)

Table S.2: Mean square errors of the predicted conditonal quantile $Q(\tau^*, \boldsymbol{\theta}(\mathbf{X}, \widehat{\boldsymbol{\beta}}_n))$ at the level of $\tau^* = 0.991, 0.995$ under the DGP of GEVD or GPD with $p = 50$ and $n = \lfloor ck \log p \rfloor$. The values in bracket refer to the corresponding sample standard deviations of prediction errors in squared loss. We denote $\boldsymbol{X} = (1, 0.1, -0.2, 0, \cdots, 0)'$ as $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X} = (1, 0, 0, 0, \cdots, 0)'$ as $\boldsymbol{X}^{(2)}$.

			GEVD				GPD			
c	$[\tau_L, \tau_U]$		$\boldsymbol{X}^{(1)}$		$\boldsymbol{X}^{(2)}$		$\boldsymbol{X}^{(1)}$		$\boldsymbol{X}^{(2)}$	
		τ^*	0.991	0.995	0.991	0.995	0.991	0.995	0.991	0.995
10	$[0.5, 0.99]$		1.65(2.98)	2.87(5.34)	5.92(15.68)	12.88(38.59)	1.81(3.07)	3.23(5.62)	6.33(11.08)	13.85(25.29)
	$[0.7, 0.99]$		1.63(2.92)	2.81(5.18)	5.42(9.90)	11.52(22.06)	1.76(3.04)	3.17(5.60)	5.98(11.06)	13.20(25.92)
	$[0.9, 0.99]$		2.40(6.51)	4.40(12.96)	6.33(12.00)	13.69(27.78)	2.23(5.10)	4.06(9.91)	5.87(10.71)	12.72(24.76)
30	$[0.5, 0.99]$		0.64(0.83)	1.12(1.45)	2.18(3.06)	4.73(6.77)	0.62(0.80)	1.10(1.42)	2.17(3.15)	4.74(7.20)
	$[0.7, 0.99]$		0.66(0.87)	1.17(1.55)	2.14(2.84)	4.66(6.26)	0.66(0.90)	1.20(1.65)	2.19(3.08)	4.80(6.89)
	$[0.9, 0.99]$		0.85(1.28)	1.55(2.39)	2.52(3.49)	5.62(8.02)	0.83(1.28)	1.53(2.44)	2.47(3.36)	5.52(7.73)
50	$[0.5, 0.99]$		0.37(0.60)	0.64(1.09)	1.22(1.80)	2.63(4.02)	0.35(0.51)	0.62(0.93)	1.19(1.72)	2.56(3.82)
	$[0.7, 0.99]$		0.39(0.61)	0.70(1.12)	1.22(1.84)	2.66(4.13)	0.40(0.62)	0.72(1.16)	1.22(1.81)	2.66(4.10)
	$[0.9, 0.99]$		0.55(1.10)	1.01(2.11)	1.48(2.45)	3.32(5.65)	0.58(1.53)	1.08(2.99)	1.42(2.30)	3.20(5.37)

Table S.3: Selection results for regularized estimation under the DGP of GEVD or GPD with $p = 50$ and $n = \lfloor ck \log p \rfloor$ for $c = 10$, 30 and 50. The values in brackets are the corresponding standard deviations.

		GEVD				GPD			
$[\tau_L, \tau_U]$	c	size	P_{AI}	FP	FN	size	P_{AI}	FP	FN
$[0.5, 0.99]$	10	8.98(0.14)	98.0	0.00(0.00)	0.22(1.56)	9.00(0.16)	97.8	0.01(0.09)	0.13(1.21)
	30	9.00(0.04)	99.8	0.00(0.00)	0.02(0.50)	9.01(0.08)	99.4	0.00(0.05)	0.00(0.00)
	50	9.00(0.00)	100.0	0.00(0.00)	0.00(0.00)	9.00(0.00)	100.0	0.00(0.00)	0.00(0.00)
$[0.7, 0.99]$	10	8.89(0.31)	89.6	0.00(0.00)	1.18(3.50)	9.07(0.49)	85.2	0.09(0.31)	0.60(2.51)
	30	9.00(0.06)	99.6	0.00(0.00)	0.04(0.70)	9.01(0.11)	99.4	0.01(0.08)	0.00(0.00)
	50	9.00(0.00)	100.0	0.00(0.00)	0.00(0.00)	9.00(0.00)	100.0	0.00(0.00)	0.00(0.00)
$[0.9, 0.99]$	10	8.69(0.62)	71.8	0.02(0.14)	3.67(6.76)	8.70(0.59)	72.0	0.02(0.14)	3.56(6.42)
	30	8.92(0.33)	90.4	0.01(0.11)	1.04(3.39)	8.91(0.36)	89.2	0.01(0.11)	1.20(3.73)
	50	8.99(0.18)	96.6	0.01(0.08)	0.24(1.63)	8.99(0.20)	96.0	0.01(0.08)	0.29(1.77)

Table S.4: Mean square errors of the predicted conditonal quantile $Q(\tau^*, \boldsymbol{\theta}(\mathbf{X}, \widehat{\boldsymbol{\beta}}_n))$ at the level of $\tau^* = 0.991$ or 0.995, where the quantile range $[0.5, 0.99]$ and the number of levels $K = 5$, 10 or 15 are considered for estimation. The values in bracket refer to the corresponding sample standard deviations of prediction errors in squared loss.

			$\boldsymbol{X} = (1, 0.1, -0.2)'$		$\boldsymbol{X} = (1, 0, 0)'$	
n	K		0.991	0.995	0.991	0.995
		True	10.36	11.84	15.14	18.85
500	5		1.84(3.49)	3.36(6.82)	7.09(13.79)	16.07(33.84)
	10		1.79(3.29)	3.32(6.65)	5.98(11.50)	14.00(30.47)
	15		2.06(4.10)	3.85(8.27)	5.67(10.75)	13.42(29.11)
1000	5		0.87(1.45)	1.60(2.83)	3.56(7.05)	8.15(17.26)
	10		0.90(1.47)	1.67(2.79)	3.09(5.02)	7.23(12.65)
	15		1.10(1.97)	2.04(3.72)	2.92(4.17)	6.78(10.12)
2000	5		0.68(1.22)	1.26(2.37)	1.85(2.93)	4.29 (7.26)
	10		0.86(2.06)	1.63(4.01)	1.66(2.44)	3.89(6.04)
	15		0.98(2.15)	1.86(4.32)	1.65(2.56)	3.87(6.34)

Table S.5: Mean absolute errors of the predicted conditonal quantile $Q(\tau^*, \boldsymbol{\theta}(\mathbf{X}, \widehat{\boldsymbol{\beta}}_n))$ at the level of $\tau^* = 0.991$ and $\boldsymbol{X} = (1, 0.1, -0.2)'$, under the DGP of Tukey lambda distribution with link functions deviate from the working model with departure levels of $d = 0, 0.5$ or 1. Baseline corresponds to $d = 0$, and the values in bracket refer to the corresponding mean square errors of prediction errors.

n	$[\tau_L, \tau_U]$	Baseline	θ_1(d=0.5)	θ_1(d=1)	θ_2(d=0.5)	θ_2(d=1)	θ_3(d=0.5)	θ_3(d=1)
500	$[0.5, 0.99]$	1.03(1.79)	1.03(1.78)	1.04(1.85)	1.81(5.64)	2.83(14.05)	1.82(4.64)	3.42(16.20)
	$[0.7, 0.99]$	1.08(1.99)	1.15(2.23)	1.22(2.47)	1.89(5.98)	2.82(13.10)	1.58(3.62)	2.88(11.33)
	$[0.9, 0.99]$	1.34(2.94)	1.41(3.22)	1.49(3.67)	2.13(7.66)	3.08(15.86)	1.29(2.61)	2.33(7.42)
1000	$[0.5, 0.99]$	0.74(0.90)	0.75(0.92)	0.78(0.98)	1.36(2.95)	2.11(7.11)	1.58(3.20)	2.84(10.08)
	$[0.7, 0.99]$	0.85(1.22)	0.89(1.37)	0.97(1.55)	1.54(3.81)	2.16(7.35)	1.35(2.43)	2.48(7.54)
	$[0.9, 0.99]$	1.04(1.79)	1.13(2.06)	1.18(2.23)	1.82(5.26)	2.52(9.94)	1.05(1.66)	2.05(5.36)
2000	$[0.5, 0.99]$	0.67(0.86)	0.68(0.83)	0.70(0.86)	1.21(2.51)	1.85(5.90)	1.33(2.38)	2.40(8.61)
	$[0.7, 0.99]$	0.77(1.07)	0.81(1.17)	0.89(1.33)	1.33(2.89)	1.88(5.74)	1.23(2.05)	2.27(6.49)
	$[0.9, 0.99]$	0.95(1.56)	1.01(1.76)	1.08(2.00)	1.61(4.26)	2.20(7.88)	0.95(1.39)	1.76(3.86)

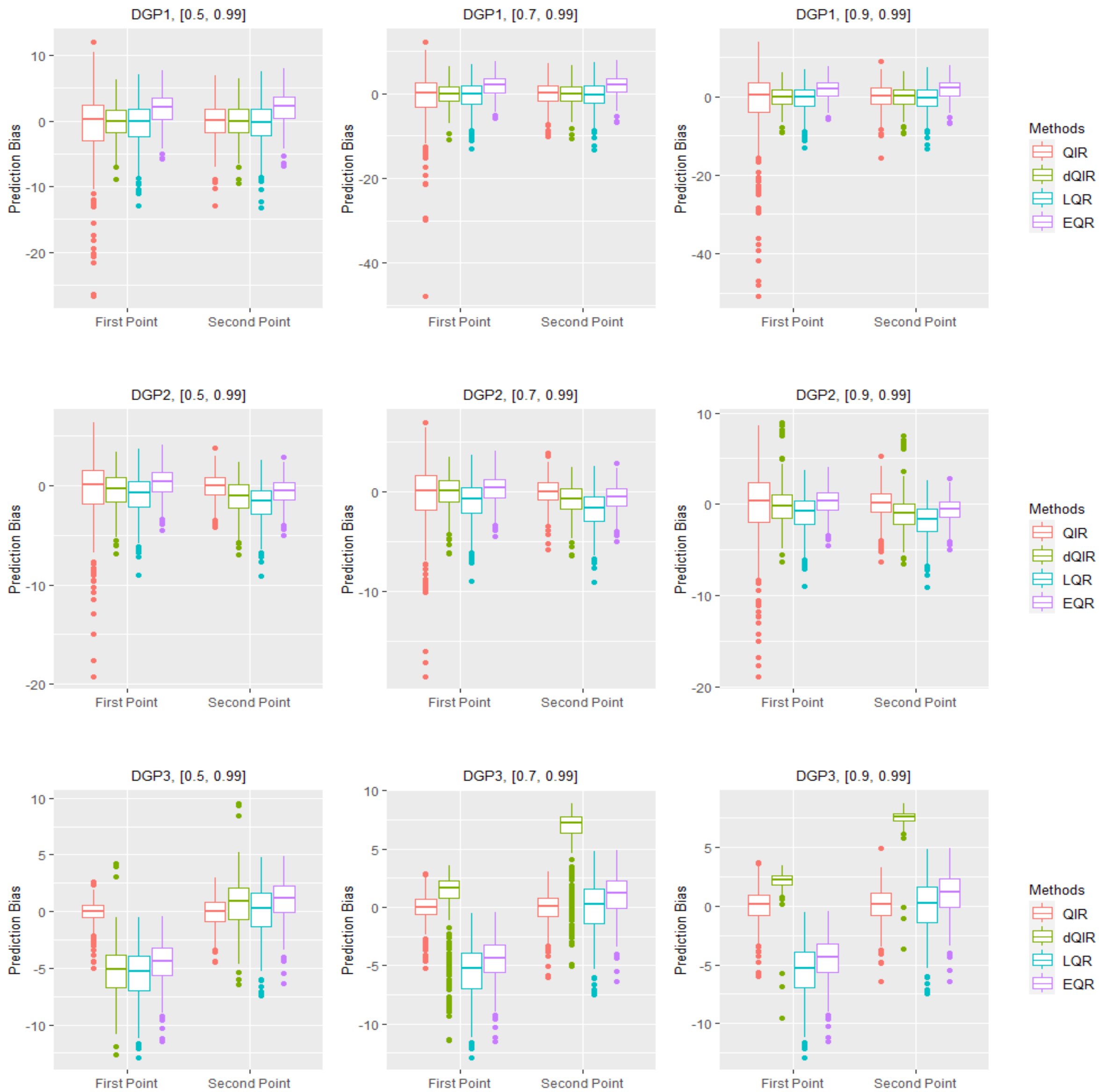

Figure S.3: Boxplots of prediction bias from QIR, dQIR, EQR, and LQR for $\tau^* = 0.991$ at two points, $\boldsymbol{X} = (1, 0.1, -0.2)$ (first point) and $\boldsymbol{X} = (1, 0, 0)$ (second point). The first, second, and third rows correspond to DGP1, DGP2, and DGP3, respectively.