Electronic Journal of Statistics
ISSN: 1935-7524

# Supplementary material for “Efficient Estimation in Expectile Regression Using Envelope Model”

**Tuo Chen**

*University of Florida*
*e-mail:* chentuo@ufl.edu

**Zhihua Su**

*University of Florida*
*e-mail:* zhihuasu@stat.ufl.edu

**Yi Yang**

*McGill University*
*e-mail:* yi.yang6@mcgill.ca
**and**

**Shanshan Ding**

*University of Delaware*
*e-mail:* sding@udel.edu

## Contents

# 1. Technical Proof

## 1.1. *Proof of Lemma 1*

If $\boldsymbol{\Gamma}_\pi$ takes the form of $(\mathbf{I}_{u_\pi}, \mathbf{A}^T)^T$, then any basis matrix of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$ has its first $u_\pi$ rows being a non-singular matrix. We denote $\boldsymbol{\Gamma}^*_\pi$ as an orthonormal basis matrix of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$ and $\boldsymbol{\Gamma}^*_{0\pi}$ as an orthonormal basis matrix of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)^\perp$. Thus, $(\boldsymbol{\Gamma}^*_\pi,\ \boldsymbol{\Gamma}^*_{0\pi})$ is an orthogonal matrix and we rewrite it as a 2 by 2 block matrix:

$$(\boldsymbol{\Gamma}^*_\pi, \boldsymbol{\Gamma}^*_{0\pi}) = \begin{pmatrix} \boldsymbol{\Gamma}^*_{\pi1} & \boldsymbol{\Gamma}^*_{0\pi1} \\ \boldsymbol{\Gamma}^*_{\pi2} & \boldsymbol{\Gamma}^*_{0\pi2} \end{pmatrix},$$

where $\boldsymbol{\Gamma}^*_{\pi1}$ is the matrix containing the first $u_\pi$ rows of $\boldsymbol{\Gamma}^*_\pi$. Since both $(\boldsymbol{\Gamma}^*_\pi,\ \boldsymbol{\Gamma}^*_{0\pi})$ and $\boldsymbol{\Gamma}^*_{\pi1}$ are non-singular, the schur complement of $\boldsymbol{\Gamma}^*_{\pi1}$, donated by $\boldsymbol{Q}$, is nonsingular. In this case, the inverse of $(\boldsymbol{\Gamma}^*_\pi,\ \boldsymbol{\Gamma}^*_{0\pi})$ is

$$\begin{aligned}(\boldsymbol{\Gamma}^*_\pi, \boldsymbol{\Gamma}^*_{0\pi})^{-1} &= \begin{pmatrix} \boldsymbol{\Gamma}^{*-1}_{\pi1} + \boldsymbol{\Gamma}^{*-1}_{\pi1}\boldsymbol{\Gamma}^*_{0\pi1}\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}^*_{\pi2}\boldsymbol{\Gamma}^{*-1}_{\pi1} & -\boldsymbol{\Gamma}^{*-1}_{\pi1}\boldsymbol{\Gamma}^*_{0\pi1}\boldsymbol{Q}^{-1} \\ -\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}^*_{\pi2}\boldsymbol{\Gamma}^{*-1}_{\pi1} & \boldsymbol{Q}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Gamma}^{*T}_{\pi1} & \boldsymbol{\Gamma}^{*T}_{\pi2} \\ \boldsymbol{\Gamma}^{*T}_{0\pi1} & \boldsymbol{\Gamma}^{*T}_{0\pi2} \end{pmatrix}.\end{aligned}$$

The second equality sign in the equation above comes from the fact that the inverse of an orthogonal matrix is the transpose of the orthogonal matrix. It turns out that $\boldsymbol{\Gamma}^{*T}_{0\pi2} = \boldsymbol{Q}^{-1}$, which are nonsingular. Therefore, $\boldsymbol{\Gamma}_{0\pi2}$ is nonsingular and invertible. It indicates that $\boldsymbol{\Gamma}^*_{0\pi}$ has its last $(p - u_\pi)$ rows being a nonsingular matrix. Then, we can decompose $\boldsymbol{\Gamma}^*_{0\pi}$ as

$$\boldsymbol{\Gamma}^*_{0\pi} = \begin{pmatrix} \boldsymbol{\Gamma}^*_{0\pi1} \\ \boldsymbol{\Gamma}^*_{0\pi2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}^*_{0\pi1}\boldsymbol{\Gamma}^{*-1}_{0\pi2} \\ \mathbf{I}_{p-u_\pi} \end{pmatrix} \boldsymbol{\Gamma}^*_{0\pi2} \equiv \begin{pmatrix} \mathbf{B} \\ \mathbf{I}_{p-u_\pi} \end{pmatrix} \boldsymbol{\Gamma}^*_{0\pi2} \equiv \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Gamma}^*_{0\pi2}. \quad (1.1)$$

Apparently, $\boldsymbol{\Gamma}_{0\pi}$ is a basis matrix of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)^\perp$ and we have $\boldsymbol{\Gamma}^T_\pi\boldsymbol{\Gamma}_{0\pi} = 0$, which means $\mathbf{B} + \mathbf{A}^T = 0$ and $\mathbf{B} =$ -$\mathbf{A}^T$. Therefore, $\boldsymbol{\Gamma}_{0\pi}$ takes the form of $(-\mathbf{A}, \mathbf{I}_{p-u_\pi})^T$.

## 1.2. *Proof of Theorem 1*

We apply Theorem 3.3 of [5] to derive the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$. There are five conditions $(i)$–$(v)$ in their Theorem and we need to check them. We denote $e(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})]$.

Based on the conditions (C1)–(C3) and Theorem 3 of [4], we have $\tilde{\boldsymbol{\theta}}_1 \xrightarrow{p} \boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{10}$ is the unique point satisfying $\mathrm{E}_{\boldsymbol{\theta}_0}[s_1(\mathbf{Z};\boldsymbol{\theta}_{10})] = 0$. Therefore, it is obvious that $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0$ is the unique point in $\boldsymbol{\Theta}$ satisfying $e(\boldsymbol{\theta}) = 0$.

Because $\tilde{\boldsymbol{\theta}}$ is the minimizer of $||e_n(\boldsymbol{\theta})||$, the condition $(i)$ holds. The conditions $(ii)$ and $(v)$ automatically hold given (C4) and (C5). By Central Limit Theorem, $\sqrt{n}(e_n(\boldsymbol{\theta}_0) - \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z}_i;\boldsymbol{\theta}_\mathbf{0})]) \xrightarrow{d} \mathcal{N}(0, \mathbf{G})$ , where $\mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z}_i;\boldsymbol{\theta}_0)] = e(\boldsymbol{\theta}_0) = 0$ and $\mathbf{G} = \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta}_0)s(\mathbf{Z};\boldsymbol{\theta}_0)^T]$. The condition $(iv)$ holds.

To prove the condition (*iii*) holds, we need to prove the following Lemma as first.

**Lemma 1.** $\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta_n} \|e_n(\boldsymbol{\theta}) - e(\boldsymbol{\theta}) - e_n(\boldsymbol{\theta}_0)\| = o_p(n^{-1/2})$, *where* $\delta_n$ *is any sequence of positive numbers with limitation 0.*

*Proof.* Let $w_j$, $\mu_j$, $\sigma_j$, $s_{1,j}$, $s_{2,j}$ and $s_{3,j}$ represent the $j$th component of $\mathbf{W}$, $\boldsymbol{\mu}_{\mathbf{X}}$, $\text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}})$, $s_1(\mathbf{Z};\boldsymbol{\theta}_1)$, $s_2(\mathbf{Z};\boldsymbol{\theta}_2)$ and $s_3(\mathbf{Z};\boldsymbol{\theta}_2)$ respectively. For any $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ and $j = 1, \ldots, p+1$,

$$\begin{aligned}|s_{1,j}(\mathbf{Z};\boldsymbol{\theta}_1) - s_{1,j}(\mathbf{Z};\boldsymbol{\theta}_1^*)|^2 &= w_j^2\big((Y-\mathbf{W}^T\boldsymbol{\theta}_1)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi\right| \\ &\quad -(Y-\mathbf{W}^T\boldsymbol{\theta}_1^*)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)-\pi\right|\big)^2.\end{aligned}$$

If $I(Y<\mathbf{W}^T\boldsymbol{\theta}_1) = I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)$, then

$$\begin{aligned}&\big((Y-\mathbf{W}^T\boldsymbol{\theta}_1)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi\right| - (Y-\mathbf{W}^T\boldsymbol{\theta}_1^*)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)-\pi\right|\big)^2 \\ &= (I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi)(Y-\mathbf{W}^T\boldsymbol{\theta}_1 - Y + \mathbf{W}^T\boldsymbol{\theta}_1^*))^2 \\ &= (I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi)\big(\mathbf{W}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)\big)^2 \\ &\leq \big(\mathbf{W}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)\big)^2 \leq \|\mathbf{W}\|^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1\|^2.\end{aligned}$$

If $I(Y<\mathbf{W}^T\boldsymbol{\theta}_1) \neq I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)$, then

$$\begin{aligned}&\big((Y-\mathbf{W}^T\boldsymbol{\theta}_1)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi\right| - (Y-\mathbf{W}^T\boldsymbol{\theta}_1^*)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)-\pi\right|\big)^2 \\ &\leq (Y-\mathbf{W}^T\boldsymbol{\theta}_1 - Y + \mathbf{W}^T\boldsymbol{\theta}_1^*))^2 \\ &= \big(\mathbf{W}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)\big)^2 \leq \|\mathbf{W}\|^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1\|^2.\end{aligned}$$

Therefore, by condition (C2), there exists a positive constant $c_1$ such that

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{\theta}_0}\big[\sup_{\boldsymbol{\theta}^*:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\leq\delta_n}|s_{1,j}(\mathbf{Z};\boldsymbol{\theta}_1) - s_{1,j}(\mathbf{Z};\boldsymbol{\theta}_1^*)|^2\big] \\ \leq&\mathrm{E}_{\boldsymbol{\theta}_0}\big[w_j^2\|\mathbf{W}\|^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1\|^2\big] \leq \delta_n^2\mathrm{E}_{\boldsymbol{\theta}_0}\big[\|\mathbf{W}\|^4\big] \leq c_1\delta_n^2.\end{aligned} \tag{1.2}$$

Let $\mu_j^*$, $\sigma_j^*$ and $\text{vech}[\cdot]_\text{j}$ represent the $j$th component of $\boldsymbol{\mu}_{\mathbf{X}}^*$, $\text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}}^*)$ and $\text{vech}[\cdot]$. Then for $j = 1, \ldots, (p+1)p/2$,
$|s_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2) - s_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2^*)|^2 = \big(\sigma_j - \text{vech}[(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})^T]_j - \sigma_j^* + \text{vech}[(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}^*)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}^*)^T]_j\big)^2$. By (C2), it is easy to verify there exists a positive constant $c_2$ such that

$$\mathrm{E}_{\boldsymbol{\theta}_0}\big[\sup_{\boldsymbol{\theta}^*:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\leq\delta_n}|s_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2) - s_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2^*)|^2\big] \leq c_2\delta_n^2. \tag{1.3}$$

Similarly, for $j = 1, \ldots, p$, there exists a positive constant $c_3$ such that

$$\mathrm{E}_{\boldsymbol{\theta}_0}\big[\sup_{\boldsymbol{\theta}^*:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\leq\delta_n}|s_{3,j}(\mathbf{Z};\boldsymbol{\theta}_2) - s_{3,j}(\mathbf{Z};\boldsymbol{\theta}_2^*)|^2\big] = \mathrm{E}_{\boldsymbol{\theta}_0}\big[\sup_{\boldsymbol{\theta}^*:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\leq\delta_n}(\mu_j-\mu_j^*)^2\big] \leq c_3\delta_n^2. \tag{1.4}$$

Combining the results in (1.2), (1.3) and (1.4), we know $s(\mathbf{Z};\boldsymbol{\theta})$ is $L^2(P)$ continuous at $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By applying Lemma 2.17 in [5], we have

$$\begin{aligned} & n^{-1/2} \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} \left\| \sum_{i=1}^{n} \{ s(\mathbf{Z};\boldsymbol{\theta}) - \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})] - s(\mathbf{Z};\boldsymbol{\theta}_0)\} \right\| \\ =& n^{-1/2} \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} \| n e_n(\boldsymbol{\theta}) - n e(\boldsymbol{\theta}) - n e_n(\boldsymbol{\theta}_0)\| = o_p(1). \end{aligned}$$

Thus, $\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} \|e_n(\boldsymbol{\theta}) - e(\boldsymbol{\theta}) - e_n(\boldsymbol{\theta}_0)\| = o_p(n^{-1/2})$. □

With the result of Lemma 2,

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} \frac{\|e_n(\boldsymbol{\theta}) - e(\boldsymbol{\theta}) - e_n(\boldsymbol{\theta}_0)\|}{n^{-1/2} + \|e_n(\boldsymbol{\theta})\| + \|e(\boldsymbol{\theta})\|} \le o_p(1).$$

The condition $(iii)$ holds. We have already verified all the conditions of Theorem 3.3 in [5]. With the result of Theorem 3.3, we have

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{G}\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}),$$

where $\mathbf{C} = \left.\frac{\partial \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ and $\mathbf{G} = \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta}_0)s(\mathbf{Z};\boldsymbol{\theta}_0)^T]$.

According to [4], we know that

$$\left.\frac{\partial \mathrm{E}_{\boldsymbol{\theta}_0}[s_1(\mathbf{Z};\boldsymbol{\theta}_1)]}{\partial \boldsymbol{\theta}_1^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\mathrm{E}_{\boldsymbol{\theta}_0}\left[\mathbf{W}\mathbf{W}^T \left| I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi \right| \right].$$

As a result, it is easy to give the expression of $\mathbf{C}$ as

$$\mathbf{C} = \begin{pmatrix} -\mathrm{E}_{\boldsymbol{\theta}_0}\left[\mathbf{W}\mathbf{W}^T \left| I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi \right| \right] & 0 & 0 \\ 0 & \mathbf{I}_{p(p+1)/2} & 0 \\ 0 & 0 & \mathbf{I}_p \end{pmatrix}.$$

Next, we give the expression of $\mathbf{G}$ in the form of $(\mathbf{G}_{ij})_{i,j=1,2,3}$. It is easy to check

$\mathbf{G}_{11} = \mathrm{E}_{\boldsymbol{\theta}_0}[s_1(\mathbf{Z};\boldsymbol{\theta}_{10})s_1(\mathbf{Z};\boldsymbol{\theta}_{10})^T] = \mathrm{E}_{\boldsymbol{\theta}_0}\left[\mathbf{W}\mathbf{W}^T(Y - \mathbf{W}^T\boldsymbol{\theta}_{10})^2 \left| I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi \right|^2\right]$;
$\mathbf{G}_{22} = \mathrm{E}_{\boldsymbol{\theta}_0}[s_2(\mathbf{Z};\boldsymbol{\theta}_{20})s_2(\mathbf{Z};\boldsymbol{\theta}_{20})^T] = \mathrm{Var}_{\boldsymbol{\theta}_0}\{\mathrm{vech}[(\mathbf{X}-\boldsymbol{\mu}_0)(\mathbf{X}-\boldsymbol{\mu}_0)^T]\}$;
$\mathbf{G}_{33} = \mathrm{E}_{\boldsymbol{\theta}_0}[s_3(\mathbf{Z};\boldsymbol{\theta}_{20})s_3(\mathbf{Z};\boldsymbol{\theta}_{20})^T] = \mathrm{Var}_{\boldsymbol{\theta}_0}[\mathbf{X}]$;
$\mathbf{G}_{23} = \mathrm{E}_{\boldsymbol{\theta}_0}[s_2(\mathbf{Z};\boldsymbol{\theta}_{20})s_3(\mathbf{Z};\boldsymbol{\theta}_{20})^T] = \mathrm{E}_{\boldsymbol{\theta}_0}\{\mathrm{vech}[(\mathbf{X}-\boldsymbol{\mu}_0)(\mathbf{X}-\boldsymbol{\mu}_0)^T](\boldsymbol{\mu}_0 - \mathbf{X})^T\}$
and

$$\begin{aligned} \mathbf{G}_{12} &= \mathrm{E}_{\boldsymbol{\theta}_0}[s_1(\mathbf{Z};\boldsymbol{\theta}_{10})s_2(\mathbf{Z};\boldsymbol{\theta}_{20})^T] \\ &= \mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathbf{W}s_2(\mathbf{Z};\boldsymbol{\theta}_{20})^T \mathrm{E}_{\boldsymbol{\theta}_0}\left[(Y - \mathbf{W}^T\boldsymbol{\theta}_{10})\left| I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi\right| \,\middle|\, \mathbf{W}\right]\right\} = 0. \end{aligned}$$

Similarly, $\mathbf{G}_{13} = 0$. Since $\mathbf{C}$ is full rank and symmetric, we have

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{C}^{-1}\mathbf{G}\ \mathbf{C}^{-1}).$$

We complete the proof of Theorem 1.

### 1.3. Proof of Theorem 2

For notation simplicity, let $Q_n(\boldsymbol{\theta}) = e_n^T(\boldsymbol{\theta})\hat{\boldsymbol{\Delta}}e_n(\boldsymbol{\theta})$ and $Q(\boldsymbol{\theta}) = e^T(\boldsymbol{\theta})\boldsymbol{\Delta}e(\boldsymbol{\theta})$, where $\boldsymbol{\Delta} = \mathbf{G}^{-1} = \{\mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta}_0)s(\mathbf{Z};\boldsymbol{\theta}_0)^T]\}^{-1}$ and $\hat{\boldsymbol{\Delta}} = \left[\frac{1}{n}\sum_{i=1}^n s(\mathbf{Z}_i;\psi(\hat{\boldsymbol{\zeta}}^*))s(\mathbf{Z}_i;\psi(\hat{\boldsymbol{\zeta}}^*))^T\right]^{-1}$. Let $l_n(\boldsymbol{\gamma}) = e_n(\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0)$ and $l(\boldsymbol{\gamma}) = e(\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0)$. Let $T_n(\boldsymbol{\gamma}) = l_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l_n(\boldsymbol{\gamma})$ and $T(\boldsymbol{\gamma}) = l^T(\boldsymbol{\gamma})\boldsymbol{\Delta}l(\boldsymbol{\gamma})$. In addition, let $\epsilon_n(\boldsymbol{\gamma}) = [l_n(\boldsymbol{\gamma}) - l_n(0) - l(\boldsymbol{\gamma})]/[1+\|\boldsymbol{\gamma}\|]$, $\kappa_n(\boldsymbol{\gamma}) = \epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma}) + 2l_n(0)^T\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma})$ and $\rho_n(\boldsymbol{\gamma}) = n[T_n(\boldsymbol{\gamma}) - \kappa_n(\boldsymbol{\gamma}) - T_n(0) - \hat{\mathbf{D}}^T\boldsymbol{\gamma}/\sqrt{n} - T(\boldsymbol{\gamma})]$, where $\hat{\mathbf{D}} = 2\mathbf{C}\hat{\boldsymbol{\Delta}}l_n(0)$. We firstly prove three Lemmas.

**Lemma 2.** *Under the same conditions in Theorem 2,* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

*Proof.* Let $\mathcal{F} = \{s(\mathbf{Z};\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Based on the fact that $s(\mathbf{Z};\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ and conditions (C2) and (C5), it is easy to verify $\mathcal{F}$ satisfies all the conditions of Uniform Law of Large Numbers. Therefore, we have $\sup_{\boldsymbol{\theta}:\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|e_n(\boldsymbol{\theta}) - e(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$. As a result, $Q_n(\boldsymbol{\theta})$ uniformly converges to $Q(\boldsymbol{\theta})$ in probability in the domain $\boldsymbol{\Theta}_e = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and } \boldsymbol{\theta} = \psi(\boldsymbol{\zeta})\}$. Since $\boldsymbol{\Theta}_e$ is compact, $Q(\boldsymbol{\theta})$ is continuous and $\boldsymbol{\theta}_0$ is the unique minimizer of $Q(\boldsymbol{\theta})$, all the conditions of Theorem 2.1 in [3] are satisfied. By the result of Theorem 2.1, we have $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. □

**Lemma 3.** *Under the same conditions in Theorem 2,* $\sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}\|/\sqrt{n}\leq\delta_n} \frac{|\rho_n(\boldsymbol{\gamma})|}{\|\boldsymbol{\gamma}\|(1+\|\boldsymbol{\gamma}\|)} = o_p(1)$*, where* $\delta_n$ *is any sequence of positive numbers with limitation 0.*

*Proof.* From the definition of $\epsilon_n(\boldsymbol{\gamma})$, we can decompose $T_n(\boldsymbol{\gamma})$ as

$$\begin{aligned} T_n(\boldsymbol{\gamma}) &= (1+\|\boldsymbol{\gamma}\|)^2\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma}) + l_n^T(0)\hat{\boldsymbol{\Delta}}l_n(0) + l^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma}) \\ &\quad + 2(1+\|\boldsymbol{\gamma}\|)\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l_n(0) + 2(1+\|\boldsymbol{\gamma}\|)\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma}) + 2l_n^T(0)\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma}). \end{aligned}$$

It can be shown that $|\rho_n(\boldsymbol{\gamma})|/(\|\boldsymbol{\gamma}\|(1+\|\boldsymbol{\gamma}\|)) \leq \sum_{i=1}^n B_j(\boldsymbol{\gamma})$, where

$B_1(\boldsymbol{\gamma}) = n(2+\|\boldsymbol{\gamma}\|)\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma})/(1+\|\boldsymbol{\gamma}\|)$, $B_2(\boldsymbol{\gamma}) = 2n|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l_n(0)|/(1+\|\boldsymbol{\gamma}\|)$,
$B_3(\boldsymbol{\gamma}) = 2n|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma})|/\|\boldsymbol{\gamma}\|$, $B_4(\boldsymbol{\gamma}) = n|2l_n^T(0)\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma}) - \hat{\mathbf{D}}^T\boldsymbol{\gamma}/\sqrt{n}|/(\|\boldsymbol{\gamma}\|(1+\|\boldsymbol{\gamma}\|))$
$B_5(\boldsymbol{\gamma}) = n|l^T(\boldsymbol{\gamma})(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})l(\boldsymbol{\gamma})|/(\|\boldsymbol{\gamma}\|(1+\|\boldsymbol{\gamma}\|))$.

From Lemma 2, we know $\sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}\|/\sqrt{n}\leq\delta_n} \|\epsilon_n(\boldsymbol{\gamma})\|)^2 = o_p(n^{-1/2})$. We define $\nu = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\|/\sqrt{n} \leq \delta_n\}$ and consider $B_1$–$B_5$ separately. We have

$$\sup_\nu B_1(\boldsymbol{\gamma}) = n\sup_\nu \frac{2+\|\boldsymbol{\gamma}\|}{1+\|\boldsymbol{\gamma}\|}\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma}) \leq n\left\|\hat{\boldsymbol{\Delta}}\right\|\sup_\nu \frac{2+\|\boldsymbol{\gamma}\|}{1+\|\boldsymbol{\gamma}\|}(\sup_\nu \|\epsilon_n(\boldsymbol{\gamma})\|)^2 = o_p(1) \text{ and,}$$

$$\begin{aligned} \sup_\nu B_2(\boldsymbol{\gamma}) &\leq \sup_\nu 2n|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l_n(0)| \leq 2n\sup_\nu \|\epsilon_n(\boldsymbol{\gamma})\|\left\|\hat{\boldsymbol{\Delta}}\right\|\|l_n(0)\| \\ &= 2\left\|\hat{\boldsymbol{\Delta}}\right\|\left\|\sqrt{n}l_n(0)\right\|\sqrt{n}\sup_\nu \|\epsilon_n(\boldsymbol{\gamma})\| = o_p(1). \end{aligned}$$

By Taylor expansion, $l(\boldsymbol{\gamma}) = e(\boldsymbol{\gamma}/\sqrt{n} + \boldsymbol{\theta}_0) = \mathbf{C}\boldsymbol{\gamma}/\sqrt{n} + o(\boldsymbol{\gamma}/\sqrt{n})$. Thus,

$$\begin{aligned}
\sup_{\nu} B_3(\boldsymbol{\gamma}) &= \sup_{\nu} 2n|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}(e(\boldsymbol{\gamma}/\sqrt{n} + \boldsymbol{\theta}_0))|/\left\|\boldsymbol{\gamma}\right\| \\
&\leq \sup_{\nu} 2n \left\|\epsilon_n(\boldsymbol{\gamma})\right\| \left\|\hat{\boldsymbol{\Delta}}\right\| (\left\|\mathbf{C}\right\| \left\|\boldsymbol{\gamma}\right\| /\sqrt{n} + o(\left\|\boldsymbol{\gamma}\right\| /\sqrt{n}))/\left\|\boldsymbol{\gamma}\right\| \\
&= 2\left\|\hat{\boldsymbol{\Delta}}\right\| (\left\|\mathbf{C}\right\| + o(1))\sqrt{n} \sup_{\nu} \left\|\epsilon_n(\boldsymbol{\gamma})\right\| \\
&= o_p(1)
\end{aligned}$$

$$\begin{aligned}
\sup_{\nu} B_4(\boldsymbol{\gamma}) &= \sup_{\nu} n|2l_n^T(0)\hat{\boldsymbol{\Delta}}(l(\boldsymbol{\gamma}) - \mathbf{C}\boldsymbol{\gamma}/\sqrt{n})|/(\left\|\boldsymbol{\gamma}\right\| (1 + \left\|\boldsymbol{\gamma}\right\|)) \\
&\leq 2n \sup_{\nu} \left\|l_n(0)\right\| \left\|\hat{\boldsymbol{\Delta}}\right\| o(1/\sqrt{n}) \\
&= 2o(1) \left\|\hat{\boldsymbol{\Delta}}\right\| \sqrt{n} \left\|l_n(0)\right\| \\
&= o_p(1).
\end{aligned}$$

Finally,

$$\begin{aligned}
\sup_{\nu} B_5(\boldsymbol{\gamma}) &\leq \sup_{\nu} n \left\|l(\boldsymbol{\gamma})\right\|^2 \left\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right\| /(\left\|\boldsymbol{\gamma}\right\| (1 + \left\|\boldsymbol{\gamma}\right\|)) \\
&\leq \sup_{\nu} \left\|\boldsymbol{\gamma}\right\|^2 (\left\|\mathbf{C}\right\| + o(1))^2 \left\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right\| / \left\|\boldsymbol{\gamma}\right\|^2 \\
&= \left\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right\| (\left\|\mathbf{C}\right\| + o(1))^2 = o_p(1).
\end{aligned}$$

Therefore, $\sup_{\nu} \frac{|\rho_n(\boldsymbol{\gamma})|}{\left\|\boldsymbol{\gamma}\right\|(1+\left\|\boldsymbol{\gamma}\right\|)} = o_p(1)$. □

Before stating the next Lemma, we define $\hat{\boldsymbol{\gamma}} = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Note that $T_n(\boldsymbol{\gamma})$ is minimized at $\hat{\boldsymbol{\gamma}}$.

**Lemma 4.** *Under the same conditions in Theorem 2, $\left\|\hat{\boldsymbol{\gamma}}\right\| = O_p(1)$.*

*Proof.* Let $\nu$ be the same defined in Lemma 4. Firstly,

$$\begin{aligned}
\sup_{\nu} |\kappa_n(\boldsymbol{\gamma})| &= \sup_{\mu} |\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma}) + 2l_n(0)^T\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma})| \\
&\leq \left\|\hat{\boldsymbol{\Delta}}\right\| (\sup_{\mu} \left\|\epsilon_n(\boldsymbol{\gamma})\right\|)^2 + \left\|\hat{\boldsymbol{\Delta}}\right\| \sup_{\mu} \left\|l_n(0)\right\| \left\|\epsilon_n(\boldsymbol{\gamma})\right\| \\
&= o_p(n^{-1}).
\end{aligned}$$

Since $T_n(\hat{\boldsymbol{\gamma}}) \leq T_n(0)$ and $\hat{\boldsymbol{\gamma}} \in \nu$, $T_n(\hat{\boldsymbol{\gamma}}) - \kappa_n(\hat{\boldsymbol{\gamma}}) = T_n(\hat{\boldsymbol{\gamma}}) + o_p(n^{-1}) \leq T_n(0) + o_p(n^{-1})$. We define

$$M = -n[T_n(\hat{\boldsymbol{\gamma}}) - \kappa_n(\hat{\boldsymbol{\gamma}}) - T_n(0) - o_p(n^{-1})] = -\rho_n(\hat{\boldsymbol{\gamma}}) - \sqrt{n}\hat{\mathbf{D}}^T\hat{\boldsymbol{\gamma}} - nT(\hat{\boldsymbol{\gamma}}) + o_p(1) \geq 0.$$

By Taylor expansion, we have $T(\hat{\boldsymbol{\gamma}}) = \hat{\boldsymbol{\gamma}}^T\mathbf{H}\hat{\boldsymbol{\gamma}}/2n + o(\left\|\hat{\boldsymbol{\gamma}}\right\|^2/n)$, where $\mathbf{H} = n\frac{\partial^2 T(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\boldsymbol{\gamma}^T}\Big|_{\boldsymbol{\gamma}=0} = \frac{\partial^2 Q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 2\mathbf{C}\mathbf{G}^{-1}\mathbf{C}$. Because $\mathbf{H}$ is a positive definite matrix by (C4), there exists a positive constant $c$ such that with probability one $T(\hat{\gamma}) \geq c \left\|\boldsymbol{\gamma}\right\|^2/n$. Therefore, by applying Lemma 4, we have

$$\begin{aligned}
M &\leq \|\hat{\boldsymbol{\gamma}}\| (1 + \|\hat{\boldsymbol{\gamma}}\|) o_p(1) + \sqrt{n} \left\|\hat{\mathbf{D}}\right\|^T \|\hat{\boldsymbol{\gamma}}\| - c\|\hat{\boldsymbol{\gamma}}\|^2 + o_p(1) \\
&\leq \|\hat{\boldsymbol{\gamma}}\| (1 + \|\hat{\boldsymbol{\gamma}}\|) o_p(1) + 2\sqrt{n}\,\|\mathbf{C}\| \left\|\hat{\boldsymbol{\Delta}}\right\| \|l_n(0)\| \, \|\hat{\boldsymbol{\gamma}}\| - c\|\hat{\boldsymbol{\gamma}}\|^2 + o_p(1) \\
&= \|\hat{\boldsymbol{\gamma}}\| (1 + \|\hat{\boldsymbol{\gamma}}\|) o_p(1) + O_p(1)\,\|\hat{\boldsymbol{\gamma}}\| - c\|\hat{\boldsymbol{\gamma}}\|^2 + o_p(1) \\
&= [-c + o_p(1)]\,\|\hat{\boldsymbol{\gamma}}\|^2 + \|\hat{\boldsymbol{\gamma}}\|\, O_p(1) + o_p(1).
\end{aligned}$$

Since $M \geq 0$ ,

$$(c - o_p(1))\,\|\hat{\boldsymbol{\gamma}}\|^2 - O_p(1)\,\|\hat{\boldsymbol{\gamma}}\| \leq o_p(1) \Longrightarrow \|\hat{\boldsymbol{\gamma}}\|^2 - O_p(1)\,\|\hat{\boldsymbol{\gamma}}\| \leq o_p(1) \Longrightarrow \hat{\boldsymbol{\gamma}} = O_p(1).$$

□

To prove Theorem 2, we define $Z_n(\boldsymbol{\gamma}) = n[T_n(\boldsymbol{\gamma}) - T_n(0)]$. Obviously, $Z_n(\boldsymbol{\gamma})$ is minimized at $\hat{\boldsymbol{\gamma}}$. Based on Lemma 4, Lemma 5 and Taylor expansion, we have

$$Z_n(\boldsymbol{\gamma}) = \sqrt{n}\hat{\mathbf{D}}^T\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{H}\boldsymbol{\gamma} + o(\|\boldsymbol{\gamma}\|^2) + \rho_n(\boldsymbol{\gamma}) + n\kappa_n(\boldsymbol{\gamma}) \xrightarrow{d} \mathbf{N}^T\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{H}\boldsymbol{\gamma},$$

where $\mathbf{N}$ is a random vector distributed as $\mathcal{N}(0, 4\mathbf{C}\mathbf{G}^{-1}\mathbf{C})$. We define $Z(\boldsymbol{\gamma}) = \mathbf{N}^T\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{H}\boldsymbol{\gamma}$. By Corollary 5.58 in [8], we have $\hat{\boldsymbol{\gamma}} \xrightarrow{d} \tilde{\boldsymbol{\gamma}}$, where

$$\tilde{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_e}{\operatorname{argmin}} Z(\boldsymbol{\gamma}) = \underset{\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_e}{\operatorname{argmin}} \frac{1}{2}(\boldsymbol{\gamma} + \mathbf{H}^{-1}\mathbf{N})^T\mathbf{H}(\boldsymbol{\gamma} + \mathbf{H}^{-1}\mathbf{N}).$$

The parameter vector $\boldsymbol{\gamma}$ is overparameterized. We apply Proposition 4.1 in [6] to solve this problem. The discrepancy function can be formed as

$$F(x, \xi) = \frac{1}{2}(\frac{\boldsymbol{\gamma}}{\sqrt{n}} + \frac{\mathbf{H}^{-1}\mathbf{N}}{\sqrt{n}})^T\mathbf{H}(\frac{\boldsymbol{\gamma}}{\sqrt{n}} + \frac{\mathbf{H}^{-1}\mathbf{N}}{\sqrt{n}}).$$

It is easy to check this discrepancy function satisfies Shapiro's assumptions and $\frac{\partial^2 F}{\partial\xi\xi^T} = \mathbf{H}$. In addition, $-\mathbf{H}^{-1}\mathbf{N} \xrightarrow{d} \mathcal{N}(0, \mathbf{C}^{-1}\mathbf{G}\ \mathbf{C}^{-1})$. Therefore, by Proposition 4.1 in [6], we have $\tilde{\boldsymbol{\gamma}} \xrightarrow{d} \mathcal{N}(0, \Lambda_g)$, where $\Lambda_g = \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\mathbf{C}\mathbf{G}^{-1}\mathbf{C}\boldsymbol{\Psi})^{\dagger}\boldsymbol{\Psi}^{\mathrm{T}}$. Hence,

$$\hat{\boldsymbol{\gamma}} = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\mathbf{C}\mathbf{G}^{-1}\mathbf{C}\boldsymbol{\Psi})^{\dagger}\boldsymbol{\Psi}^{\mathrm{T}}).$$

We complete the proof of Theorem 2.

### *1.4. Corollary*

*Proof.* Let $\boldsymbol{\Upsilon} = \mathbf{C}^{-1}\mathbf{G}\mathbf{C}^{-1}$. According to the results in Theorem 1 and Theorem 2,

$$\begin{aligned} avar(\sqrt{n}\tilde{\boldsymbol{\theta}}) - avar(\sqrt{n}\hat{\boldsymbol{\theta}}) &= \mathbf{C}^{-1}\mathbf{G}\ \mathbf{C}^{-1} - \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\mathbf{C}\mathbf{G}^{-1}\mathbf{C}\boldsymbol{\Psi})^{\dagger}\boldsymbol{\Psi}^{\mathrm{T}} \\ &= \boldsymbol{\Upsilon} - \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Psi})^{\dagger}\boldsymbol{\Psi}^T \\ &= \boldsymbol{\Upsilon}^{1/2}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}})\boldsymbol{\Upsilon}^{1/2} \\ &= \boldsymbol{\Upsilon}^{1/2}\mathbf{Q}_{\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}}\boldsymbol{\Upsilon}^{1/2} \\ &\geq 0. \end{aligned}$$

□

## 2. Simulations Under Variable Selection Settings (Without Envelope Structure)

In this section, we investigate the performance of the ER model, the EER model, the boosting model and the sparse ER model under the settings in which sparsity structure exists but no (nontrivial) envelope structure exists. In this case, $u_\pi = p$, and the EER model degenerates to the ER model. We consider the following settings:

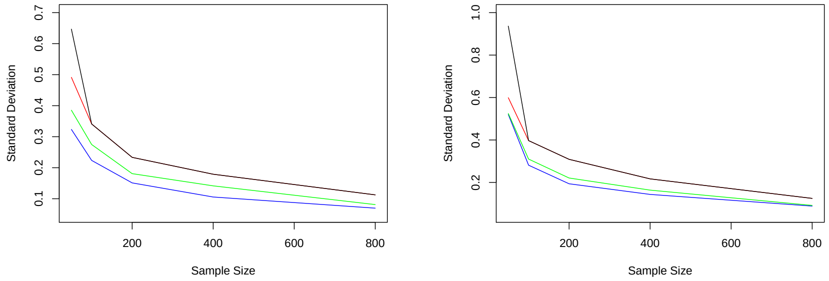
$$Y_i = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i + (2 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i)\epsilon_i, \qquad \text{for} \quad i = 1, \ldots, n.$$

We set $p = 6$ and $p_A = 3$, where $p_A$ denotes the number of active predictors. Both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ were $p$-dimensional vectors. The first $p_A$ elements in $\boldsymbol{\alpha}_1$ were 4 and the rest $p - p_A$ elements were 0. The first $p_A$ elements in $\boldsymbol{\alpha}_2$ were 0.1 and the rest $p - p_A$ elements were 0. The error term $\epsilon$ was generated from standard normal distribution $\epsilon \sim \mathcal{N}(0, 1)$.

Based upon the settings, the $\pi$th conditional expectile of $Y$ had the following form

$$f_\pi(Y|\mathbf{X}) = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X} + (2 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X})f_\pi(\epsilon) = 3 + 2f_\pi(\epsilon) + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon))^T\mathbf{X},$$

where $f_\pi(\epsilon)$ represented the $\pi$th expectile of the error distribution. Thus the coefficients were contained in $\boldsymbol{\beta}_\pi = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon)$ and the last $p - p_A$ elements of $\boldsymbol{\beta}_\pi$ were 0. This means that the first $p_A$ predictors were active predictors, and the rest were inactive. The predictor vector $\mathbf{X}$ followed a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$. The upper left $p_A \times p_A$ block of $\boldsymbol{\Sigma}_{\mathbf{X}}$ was a diagonal matrix with diagonal elements being 1, 2 and 4. The bottom right block was a $(p - p_A) \times (p - p_A)$ diagonal matrix with diagonal elements being 8, 16 and 32. The off-diagonal blocks of $\boldsymbol{\Sigma}_{\mathbf{X}}$ were $3\mathbf{M}$ and $3\mathbf{M}^T$, where $\mathbf{M}$ was a randomly generated $p_A \times (p - p_A)$ orthogonal matrix (generated using `randortho` function in `R` package `pracma`). In this case, the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_\pi) = \mathbb{R}^p$ and the EER model reduces to the ER model since no immaterial



(a) $\pi = 0.5$

(b) $\pi = 0.9$

Fig 1: Comparison of the sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator.

information is present. Therefore, for this scenario, the sparsity structure exists but no (nontrivial) envelope structure exists.

We varied the sample size $n$ from 50 to 800. For each sample size, 100 replications were generated. For each replication, we computed the EER estimator ($u_\pi$ chosen by RCV), the ER estimator, the boosting estimator as well as the sparse ER estimator of $\boldsymbol{\beta}_\pi$. For each element in $\boldsymbol{\beta}_\pi$, we computed the sample standard deviation from the 100 EER estimators, 100 ER estimators, 100 boosting estimators and 100 sparse ER estimators. We took expectile levels 0.50 and 0.90 as examples. The results of a randomly chosen nonzero element in $\boldsymbol{\beta}_\pi$ with $\pi = 0.50$ and $\pi = 0.90$ are summarized in Figure 1.

In each panel of Figure 1, the line for the EER estimator almost overlaps with the line for the ER estimator when sample size exceeds 100. This is expected as when RCV selected $u_\pi = p$ and the EER estimator degenerates to the ER estimator. With small sample size, there was a little variation in the model selection for the EER model, so the EER estimator was more variable than the ER estimator. The efficiency gains from the sparse ER model and the boosting model is obvious under this setting. Take $n = 200$ as an example, the standard deviation is 0.23 for the ER or EER estimator, 0.18 for the boosting estimator and 0.15 for the sparse ER estimator for $\pi = 0.5$. The efficient gains is because that the boosting estimator and the sparse ER estimator correctly identified the underlying sparsity structure. Therefore, in the case where there is sparsity structure but no (nontrivial) envelope structure, the boosting estimator and the sparse ER estimator achieves more efficiency gains than the EER estimator. However, if the sparsity structure and the envelope structure both exist, the EER estimator may be more efficient than the boosting and sparse ER estimator as shown in Section 3.

## 3. Simulations Under Variable Selection Settings (With Envelope Structure)

In this section, we investigate the performance of the EER model when the underlying model has both the sparsity structure and the envelope structure. We consider the following simulation settings:

$$Y_i = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i)\epsilon_i, \qquad \text{for} \quad i = 1, \ldots, n.$$

We set $p = 12$, $u_\pi = 2$ and $p_A = 6$, where $p_A$ denotes the number of active predictors. Both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ were $p$-dimensional vectors. The first $p_A$ elements in $\boldsymbol{\alpha}_1$ were 4 and the rest $p - p_A$ elements were 0. The first $p_A$ elements in $\boldsymbol{\alpha}_2$ were 0.1 and the rest $p - p_A$ elements were 0. Four types of error distribution were used to generate $\epsilon$: standard normal distribution $\epsilon \sim \mathcal{N}(0,1)$, student's $t$–distribution with 4 degrees of freedom $\epsilon \sim t_4$, mixed normal distribution $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$, and exponential distribution $\epsilon \sim \mathrm{Exp}(1)$.

Based upon the settings, the $\pi$th conditional expectile of $Y$ had the following form

$$f_\pi(Y|\mathbf{X}) = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X} + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X})f_\pi(\epsilon) = 3 + 8f_\pi(\epsilon) + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon))^T\mathbf{X},$$

where $f_\pi(\epsilon)$ represented the $\pi$th expectile of the error distribution. Thus $\boldsymbol{\beta}_\pi = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon)$ and the last $p - p_A$ elements of $\boldsymbol{\beta}_\pi$ were 0, which means only the first $p_A$ components in $\mathbf{X}$ were active predictors. The predictor vector $\mathbf{X}$ followed a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T + \boldsymbol{\Phi}_0\boldsymbol{\Lambda}_0\boldsymbol{\Phi}_0^T$, where $\boldsymbol{\Lambda}$ was a $u_\pi \times u_\pi$ diagonal matrix with diagonal elements 100 and 9, and $\boldsymbol{\Lambda}_0$ was a $2 \times 2$ block matrix. The upper left block of $\boldsymbol{\Lambda}_0$ was a $(p_A - u_\pi) \times (p_A - u_\pi)$ identity matrix and the bottom right block was a $(p - p_A) \times (p - p_A)$ identity matrix. The off-diagonal blocks of $\boldsymbol{\Lambda}_0$ were $0.8\boldsymbol{\Lambda}_{0*}$ and $0.8\boldsymbol{\Lambda}_{0*}^T$ where $\boldsymbol{\Lambda}_{0*}$ was a randomly generated $(p - p_A) \times (p_A - u_\pi)$ semi-orthogonal matrix. The matrix $\boldsymbol{\Phi} \in \mathbb{R}^{p \times u_\pi}$ was a semi-orthogonal matrix with the first $p_A/2$ rows being $(\sqrt{3}/3,\ 0)$, the following $p_A/2$ rows being $(0,\ \sqrt{3}/3)$ and the remaining $p - p_A$ rows being $(0, 0)$. The matrix $\boldsymbol{\Phi}_0 \in \mathbb{R}^{p \times (p - u_\pi)}$ was a semi-orthogonal matrix that satisfied $\boldsymbol{\Phi}^T\boldsymbol{\Phi}_0 = 0$. Since $\boldsymbol{\alpha}_1 = \boldsymbol{\Phi} \cdot (4\sqrt{3}, 4\sqrt{3})^T$ and $\boldsymbol{\alpha}_2 = \boldsymbol{\Phi} \cdot (\sqrt{3}/10, \sqrt{3}/10)^T$, $f_\pi(Y|\mathbf{X})$ and $\mathbf{X}$ satisfied the EER model with $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_\pi) = \mathrm{span}(\boldsymbol{\Phi})$.

Under this setting, we repeated the sample standard deviations comparison, the prediction performance comparison and the RCV performance examination as described in Section 5 of the paper. To be noted, here for the sample standard deviations comparison, we randomly choose an active component of $\boldsymbol{\beta}_\pi$ to display the outcomes. All results are given in Figures 2 – 7 and Tables 1 – 2.

Figure 2 shows substantial efficiency gains from the EER model in the estimation of $\boldsymbol{\beta}_\pi$. In all the plots with different error distributions and expectile levels $\pi$, the sample standard deviations of the EER estimators are much smaller than the sample standard deviations of the ER estimators, the boosting estimators and the sparse ER estimators under all sample sizes. As variable selection methods, the boosting model and the sparse ER model are more efficient than the

ER model since they correctly identify the underlying sparse structure. However, they do not account for the immaterial information in $\mathbf{X}$ in the estimation. The EER model can still be more efficient than the boosting model and the sparse ER model if the variation of the immaterial part has a large effect on estimation, such as in this example.

Figure 3 indicates that the bootstrap standard deviation is a good approximation to the actual sample standard deviation. Table 1 and Figures 4 – 7 summarize the RMSEs under the EER model, the ER model, the boosting model and the sparse ER model with different error distributions. We can see a notable improvement of the prediction performance for the EER model. Take Table 1 (a) as an example, the EER model reduces the average RMSE by about 40% comparing with the ER model, by about 30% comparing with the boosting model and by about 60% comparing with the sparse ER model. Both the sparse ER model and the boosting model identifies the active predictors. But the sparse ER model tends to put more shrinkage on the nonzero coefficients, while the boosting estimator does not over shrink the nonzero coefficients. Therefore, we notice that the boosting estimator has a better prediction performance than the ER estimator, but the sparse ER estimator has the largest prediction error.

Table 2 summaries the fraction that RCV selects the true dimension $u_\pi = 2$. RCV selects the true dimension more than 90% of the time when sample size reaches 100. And it still gives an accuracy over 75% with a small sample size 25.

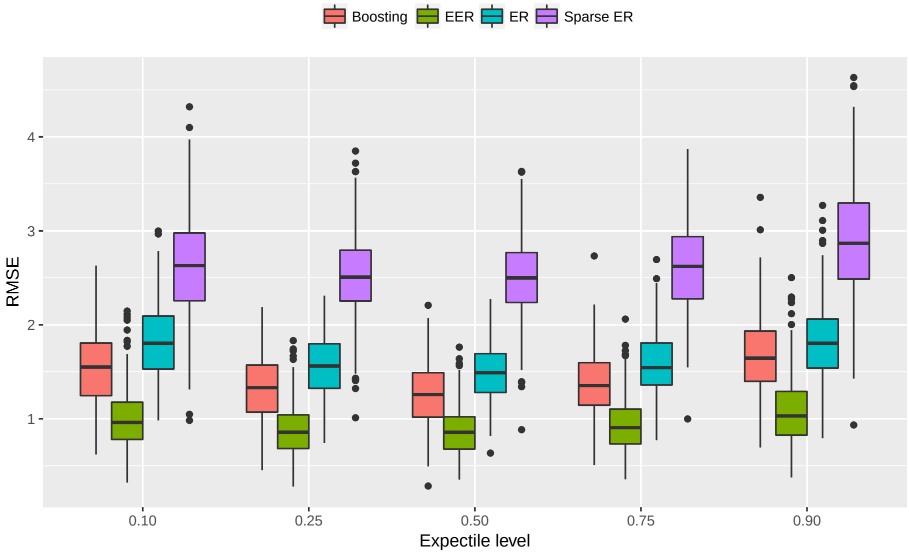
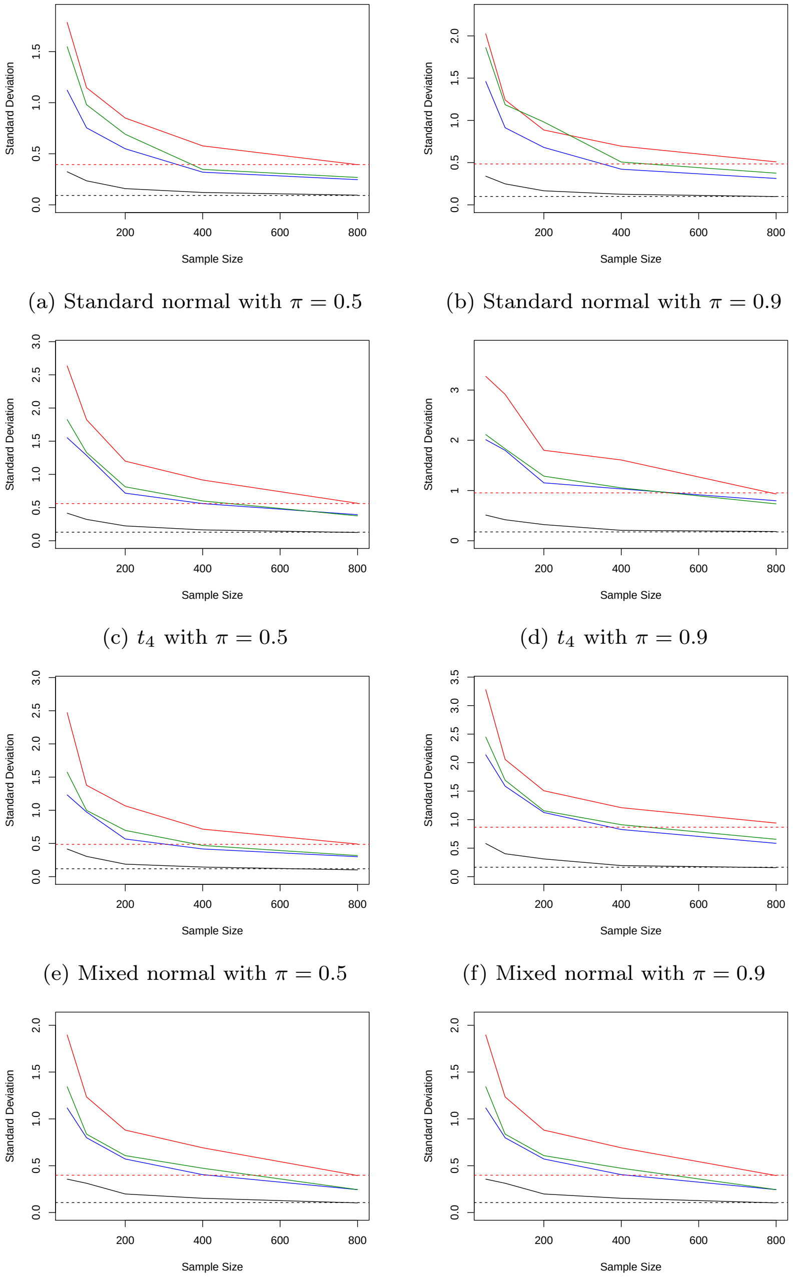


Fig 4: Boxplots of RMSEs under the four models with $\epsilon \sim \mathcal{N}(0,1)$.



(a) Standard normal with $\pi = 0.5$

(b) Standard normal with $\pi = 0.9$

(c) $t_4$ with $\pi = 0.5$

(d) $t_4$ with $\pi = 0.9$

(e) Mixed normal with $\pi = 0.5$

(f) Mixed normal with $\pi = 0.9$

(g) Exp(1) with $\pi = 0.5$

(h) Exp(1) with $\pi = 0.9$

Fig 2: Sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator. The horizontal lines mark the asymptotic standard deviations of the ER estimator (the upper line in each panel) and the EER estimator (the lower line in each panel).

TABLE 1
*The average RMSEs of the 300 replications under the four models with different error distributions.*

(a) $\epsilon \sim \mathcal{N}(0,1)$

| | EER | ER | Boosting | Sparse ER |
|---|---|---|---|---|
| $\pi = 0.10$ | 1.05 | 1.82 | 1.54 | 2.65 |
| $\pi = 0.25$ | 0.88 | 1.57 | 1.32 | 2.51 |
| $\pi = 0.50$ | 0.87 | 1.49 | 1.26 | 2.50 |
| $\pi = 0.75$ | 0.94 | 1.58 | 1.38 | 2.62 |
| $\pi = 0.90$ | 1.09 | 1.83 | 1.67 | 2.89 |

(b) $\epsilon \sim t_4$

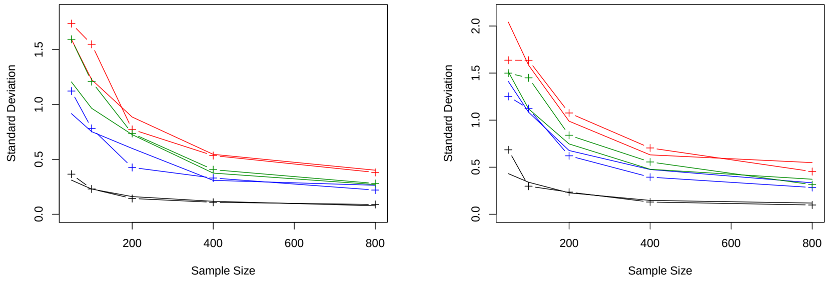
| | EER | ER | Boosting | Sparse ER |
|---|---|---|---|---|
| $\pi = 0.10$ | 1.77 | 3.42 | 2.89 | 6.00 |
| $\pi = 0.25$ | 1.20 | 2.40 | 2.02 | 4.84 |
| $\pi = 0.50$ | 1.08 | 2.09 | 1.77 | 4.52 |
| $\pi = 0.75$ | 1.24 | 2.39 | 2.06 | 5.04 |
| $\pi = 0.90$ | 1.79 | 3.41 | 3.06 | 6.54 |

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

| | EER | ER | Boosting | Sparse ER |
|---|---|---|---|---|
| $\pi = 0.10$ | 1.15 | 2.16 | 1.86 | 3.53 |
| $\pi = 0.25$ | 1.01 | 1.83 | 1.55 | 3.35 |
| $\pi = 0.50$ | 1.01 | 1.81 | 1.54 | 3.55 |
| $\pi = 0.75$ | 1.20 | 2.15 | 1.85 | 4.24 |
| $\pi = 0.90$ | 1.72 | 3.06 | 2.72 | 5.59 |

(d) $\epsilon \sim$ Exp(1)

| | EER | ER | Boosting | Sparse ER |
|---|---|---|---|---|
| $\pi = 0.10$ | 0.64 | 0.74 | 0.67 | 1.85 |
| $\pi = 0.25$ | 0.73 | 1.04 | 0.90 | 2.42 |
| $\pi = 0.50$ | 0.93 | 1.51 | 1.29 | 3.17 |
| $\pi = 0.75$ | 1.33 | 2.22 | 1.95 | 4.19 |
| $\pi = 0.90$ | 1.96 | 3.31 | 2.99 | 5.53 |



(a) Standard normal with $\pi = 0.5$

(b) Standard normal with $\pi = 0.9$

Fig 3: Sample standard deviations and bootstrap standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator. Lines with "+" mark the bootstrap standard deviations for the corresponding estimators.

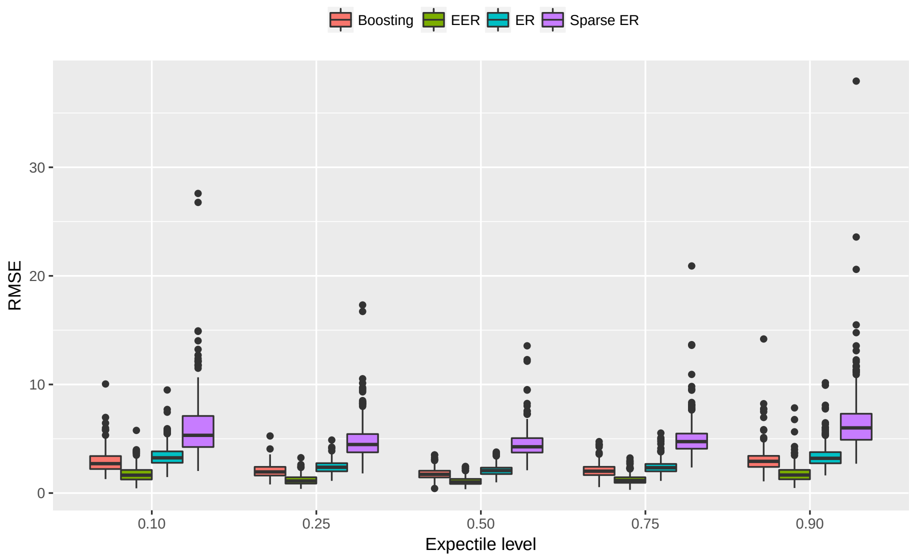


Fig 5: Boxplots of RMSEs under the four models with $\epsilon \sim t_4$.

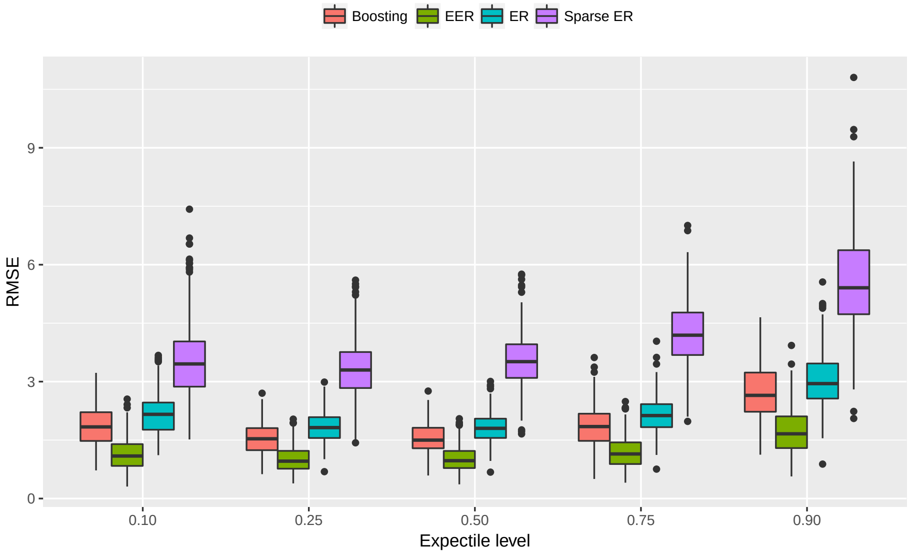


Fig 6: Boxplots of RMSEs under the four models with $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$.

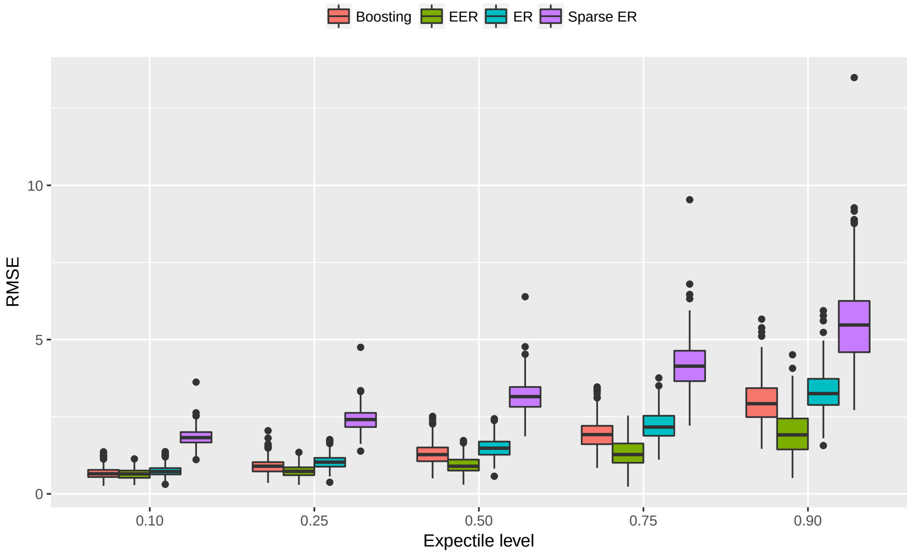


Fig 7: Boxplots of RMSEs under the four models with $\epsilon \sim \text{Exp}(1)$.

Table 2
*The fraction that RCV selects the true $u_\pi$ with different error distributions.*

(a) $\epsilon \sim \mathcal{N}(0,1)$

| | $\pi = 0.10$ | $\pi = 0.25$ | $\pi = 0.50$ | $\pi = 0.75$ | $\pi = 0.90$ |
|---|---|---|---|---|---|
| $n = 25$ | 79% | 83% | 80% | 79% | 82% |
| $n = 50$ | 95% | 94% | 94% | 92% | 88% |
| $n = 100$ | 97% | 97% | 99% | 98% | 96% |
| $n = 200$ | 99% | 100% | 99% | 100% | 99% |
| $n = 400$ | 99% | 100% | 100% | 100% | 100% |
| $n = 800$ | 100% | 100% | 100% | 100% | 100% |

(b) $\epsilon \sim t_4$

| | $\pi = 0.10$ | $\pi = 0.25$ | $\pi = 0.50$ | $\pi = 0.75$ | $\pi = 0.90$ |
|---|---|---|---|---|---|
| $n = 25$ | 84% | 88% | 89% | 88% | 80% |
| $n = 50$ | 90% | 94% | 95% | 91% | 93% |
| $n = 100$ | 98% | 98% | 99% | 99% | 96% |
| $n = 200$ | 100% | 100% | 100% | 100% | 100% |
| $n = 400$ | 100% | 100% | 100% | 100% | 100% |
| $n = 800$ | 100% | 100% | 100% | 100% | 100% |

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

| | $\pi = 0.10$ | $\pi = 0.25$ | $\pi = 0.50$ | $\pi = 0.75$ | $\pi = 0.90$ |
|---|---|---|---|---|---|
| $n = 25$ | 80% | 85% | 88% | 84% | 84% |
| $n = 50$ | 90% | 96% | 96% | 91% | 91% |
| $n = 100$ | 93% | 99% | 99% | 98% | 98% |
| $n = 200$ | 98% | 100% | 100% | 100% | 100% |
| $n = 400$ | 100% | 100% | 100% | 100% | 100% |
| $n = 800$ | 100% | 100% | 100% | 100% | 100% |

(d) $\epsilon \sim \text{Exp}(1)$

| | $\pi = 0.10$ | $\pi = 0.25$ | $\pi = 0.50$ | $\pi = 0.75$ | $\pi = 0.90$ |
|---|---|---|---|---|---|
| $n = 25$ | 78% | 78% | 76% | 79% | 78% |
| $n = 50$ | 95% | 95% | 98% | 94% | 84% |
| $n = 100$ | 99% | 99% | 100% | 99% | 96% |
| $n = 200$ | 100% | 100% | 100% | 99% | 99% |
| $n = 400$ | 100% | 100% | 100% | 100% | 100% |
| $n = 800$ | 100% | 100% | 100% | 100% | 100% |

## 4. Simulations Under No Immaterial Part Settings

In this section, we conduct a simulation study to investigate the performance of the EER model if no immaterial part exists. The data was generated from the following model

$$Y_i = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}} \mathbf{X}_i + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}} \mathbf{X}_i)\epsilon_i, \qquad \text{for} \quad i = 1, \ldots, 800.$$

We set $p = 6$ and each element in $\boldsymbol{\alpha}_1$ was drawn from independent standard normal distribution. Each elements in $\boldsymbol{\alpha}_2$ was 0.1. The predictor vector $\mathbf{X}$ followed a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{P}^T\mathbf{D}\mathbf{P}$, where $\mathbf{P}$ was a randomly generated orthogonal matrix (generated using `randortho` function in `R` package `pracma`), and $\mathbf{D}$ was a diagonal matrix with diagonal elements being 1, 2, 4, 8, 16 and 32. The error $\epsilon$ was generated from the normal distribution $\epsilon \sim \mathcal{N}(0, 5)$.

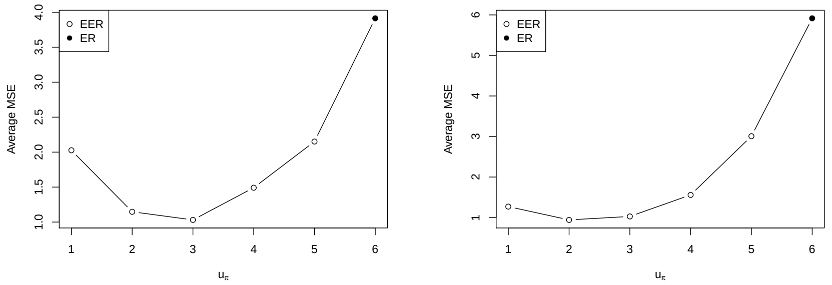
Based upon the settings, the $\pi$th conditional expectile of $Y$ has the following form

$$f_\pi(Y|\mathbf{X}) = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X} + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}) f_\pi(\epsilon) = 3 + 8 f_\pi(\epsilon) + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon))^T \mathbf{X},$$

where $f_\pi(\epsilon)$ represents the $\pi$th expectile of the error distribution, the intercept is $3 + 8f_\pi(\epsilon)$ and the coefficients are $\boldsymbol{\beta}_\pi = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon)$. In this case, $\boldsymbol{\Sigma}_{\mathbf{X}}$ does not have the decomposition as a sum of the variation of the material part (related to $\boldsymbol{\beta}_\pi$) and the variation of the immaterial part. So the envelope subspace is the full space $\mathbb{R}^p$, and there is no immaterial part.

Although no envelope structure is present, we will compute an approximate "EER" estimator and compare it with the ER estimator. We know that under an EER model, the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_\pi) = \mathrm{span}(\boldsymbol{\Gamma}_\pi)$ is spanned by the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$. Therefore, for each $1 \leq u_\pi \leq p$, we approximate $\boldsymbol{\Gamma}_\pi$ by $\hat{\boldsymbol{\Gamma}}_\pi$, which is a $p \times u_\pi$ matrix whose columns were the $u_\pi$ eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ corresponding to the $u_\pi$ largest eigenvalues. We note that under the exact EER model, $\boldsymbol{\Gamma}_\pi$ is chosen to be the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that contains $\boldsymbol{\beta}_\pi$. They may not necessarily be the eigenvectors corresponding to the largest eigenvalues. Since the exact (nontrivial) EER model does not exist here, we are proposing a way to approximate the $\boldsymbol{\Gamma}_\pi$ such that its estimator is least variable. Then we defined the "EER" estimator as $\hat{\boldsymbol{\beta}}_\pi{=}\hat{\boldsymbol{\Gamma}}_\pi\hat{\boldsymbol{\eta}}_\pi$, where $\hat{\boldsymbol{\eta}}_\pi$ was the ER estimator with $Y$ being the response and $\hat{\boldsymbol{\Gamma}}_\pi^T\mathbf{X}$ being the predictors.

We generated 100 replications and computed the mean squared error (MSE) $\|\hat{\boldsymbol{\beta}}_\pi - \boldsymbol{\beta}_\pi\|^2$ at expectile levels $\pi = 0.5$ and 0.9 for each replication. The average MSE are summarized in Figure 8. Because the true $u_\pi$ equals $p$, a smaller $u_\pi$ leads to larger bias, but its estimator is less variable. As $u_\pi$ increases, the bias of $\hat{\boldsymbol{\beta}}_\pi$ becomes smaller but its variance becomes larger. When $\pi = 0.5$, the bias-variance tradeoff makes the average MSE reach its minimum 1.03 at $u_\pi = 3$. When $\pi = 0.9$, the average MSE reaches its minimum 0.94 at $u_\pi = 2$. Note that when $u_\pi = p$, the EER estimator $\hat{\boldsymbol{\beta}}_\pi$ reduces to the ER estimator. The MSE of the ER estimator is 3.91 for $\pi = 0.5$ and 5.92 for $\pi = 0.9$. The results shows that when there is no immaterial part, we can still expect to have a smaller MSE from an approximate EER estimator in some cases due to the bias-variance tradeoff.



(a) $\pi = 0.5$

(b) $\pi = 0.9$

Fig 8: Average MSE with respect to $u_\pi$. Note that the MSE corresponding to $u_\pi = p = 6$ is the MSE of the ER estimator.

## 5. Analysis of "state.x77" with Predictors in Original Scale

We perform the same data analysis on the "state.x77" data again, but with the predictors in the original scale instead of the standardized predictors. We first select the dimension of the envelope subspace with RCV. For all quantile levels, RCV selects $u_\pi = 4(= p)$. This indicated that there is no immaterial part in the data and the EER estimator reduces to the ER estimator. In this case, the EER estimator and the ER estimator have the same efficiency. Detailed information about the estimated regression coefficients are provided in the following Table 3. Because $u_\pi = p$, the estimated regression coefficients given by the EER model and the ER model are exactly the same.

TABLE 3
*The estimated regression coefficients for the predictors in the original scale given by the EER model, the ER model, the boosting model and the sparse ER model.*

| | EER | | | | ER | | | |
|---|---|---|---|---|---|---|---|---|
| | Population | Income | Illiteracy | Frost | Population | Income | Illiteracy | Frost |
| $\pi = 0.10$ | $0.29\times10^{-3}$ | $-1.14\times10^{-3}$ | 3.30 | $-9.24\times10^{-3}$ | $0.29\times10^{-3}$ | $-1.14\times10^{-3}$ | 3.30 | $-9.24\times10^{-3}$ |
| $\pi = 0.25$ | $0.26\times10^{-3}$ | $-0.66\times10^{-3}$ | 3.65 | $-5.39\times10^{-3}$ | $0.26\times10^{-3}$ | $-0.66\times10^{-3}$ | 3.65 | $-5.39\times10^{-3}$ |
| $\pi = 0.50$ | $0.22\times10^{-3}$ | $0.06\times10^{-3}$ | 4.14 | $0.58\times10^{-3}$ | $0.22\times10^{-3}$ | $0.06\times10^{-3}$ | 4.14 | $0.58\times10^{-3}$ |
| $\pi = 0.75$ | $0.21\times10^{-3}$ | $0.55\times10^{-3}$ | 4.48 | $5.22\times10^{-3}$ | $0.21\times10^{-3}$ | $0.55\times10^{-3}$ | 4.48 | $5.22\times10^{-3}$ |
| $\pi = 0.90$ | $0.18\times10^{-3}$ | $0.72\times10^{-3}$ | 4.58 | $7.89\times10^{-3}$ | $0.18\times10^{-3}$ | $0.72\times10^{-3}$ | 4.58 | $7.89\times10^{-3}$ |
| | Boosting | | | | Sparse ER | | | |
| | Population | Income | Illiteracy | Frost | Population | Income | Illiteracy | Frost |
| $\pi = 0.10$ | $0.29\times10^{-3}$ | $-1.12\times10^{-3}$ | 3.22 | $-10.00\times10^{-3}$ | $0.10\times10^{-3}$ | $0.00\times10^{-3}$ | 2.28 | $-9.34\times10^{-3}$ |
| $\pi = 0.25$ | $0.23\times10^{-3}$ | $-0.41\times10^{-3}$ | 3.59 | $-5.05\times10^{-3}$ | $0.08\times10^{-3}$ | $0.00\times10^{-3}$ | 2.89 | $-2.95\times10^{-3}$ |
| $\pi = 0.50$ | $0.19\times10^{-3}$ | $0.00\times10^{-3}$ | 3.79 | $0.00\times10^{-3}$ | $0.04\times10^{-3}$ | $0.00\times10^{-3}$ | 2.70 | $0.00\times10^{-3}$ |
| $\pi = 0.75$ | $0.12\times10^{-3}$ | $0.00\times10^{-3}$ | 3.36 | $0.00\times10^{-3}$ | $0.00\times10^{-3}$ | $0.00\times10^{-3}$ | 2.12 | $0.00\times10^{-3}$ |
| $\pi = 0.90$ | $0.06\times10^{-3}$ | $0.00\times10^{-3}$ | 2.92 | $0.00\times10^{-3}$ | $0.00\times10^{-3}$ | $0.00\times10^{-3}$ | 0.73 | $0.00\times10^{-3}$ |

We took a close look at the data and found that the scales of the four predictors are quite different. For example, population varies from 365 to 21198 (thousand) while illiteracy level varies from 0.5 to 2.8 (percent). This makes the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{X}}$ range from 0.17 to $1.99 \times 10^7$ and the eigenvectors are very close to the standard basis vectors, i.e, $(1,0,0,0)^T$, $(0,1,0,0)^T$, $(0,0,1,0)^T$ and $(0,0,0,1)^T$. In this case, if $\boldsymbol{\beta}_\pi$ belongs to an envelope subspace that is a proper subset of $\mathbb{R}^p$, then we are essentially performing variable selection. For example, if the dimension of envelope subspace is $u = 3$, then the envelope subspace is spanned by 3 out of the 4 eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$. Since $\boldsymbol{\beta}_\pi$ lies in the envelope subspace, one component of $\boldsymbol{\beta}_\pi$ has to be 0, which means the corresponding predictor is immaterial to the conditional expectile of the response. The dimension selection results from RCV indicates the EER model finds that all four predictors are material to the conditional expectile of the response at all investigated expectile levels.

This situation is also shared by other dimension reduction based methods such as principal component analysis (PCA). If one component is selected, which corresponds to direction $(0,1,0,0)^T$, then only one variable (income) is included in subsequent analysis. Thus when variables have drastically different scales, PCA normally standardize the variables. We followed this practice and presented the results with standardized predictor variables in Section 6 of the paper.

## 6. Prediction Performance Comparison on “state.x77”

We compared the prediction performance between the ER model and the EER model on “state.x77” using five fold cross-validation repeated with 50 random splits to compute the mean predicted expectile losses. The results are summarized in the following table. The predicted expectile losses from the EER model are slightly larger than those from the ER model. This may due to the criterion we use to select the dimension of the envelope subspace $u_\pi$. We selected $u_\pi$ by RCV with one standard deviation rule. In other words, instead of choosing the dimension that has the minimum RCV, we choose the smallest dimension having RCV less than one standard deviation above the minimum value of RCV. Therefore this criterion tends to select a more parsimonious model by sacrificing some predictive accuracy comparing to the best model. In this case, it is possible that the full model, i.e., the ER model, has an RCV that is closer to the minimum value of RCV compared to the selected model.

Table 4

*Mean of the predicted expectile losses under the ER and the EER model with different expectile levels.*

| | $\pi = 0.10$ | $\pi = 0.25$ | $\pi = 0.50$ | $\pi = 0.75$ | $\pi = 0.90$ |
|---|---|---|---|---|---|
| ER | 1.48 | 2.79 | 3.75 | 3.29 | 2.45 |
| EER | 1.58 | 2.82 | 3.86 | 3.85 | 2.72 |

## 7. Simulation Results at More Expectile Levels

In Section 5 of the paper, we give the results at expectile levels 0.50 and 0.90 in Figure 1. Here we provide the results at expectile levels 0.10, 0.25 and 0.75 in Figure 9.

Figure 9 shows a similar pattern as Figure 1 of the paper. For every error distribution and expectile level, the sample standard deviations of the EER estimators are much smaller than the sample standard deviations of the ER estimators, the boosting estimators and the sparse ER estimators under all sample sizes.

## 8. Analysis of Computational Complexity of the GMM Algorithm

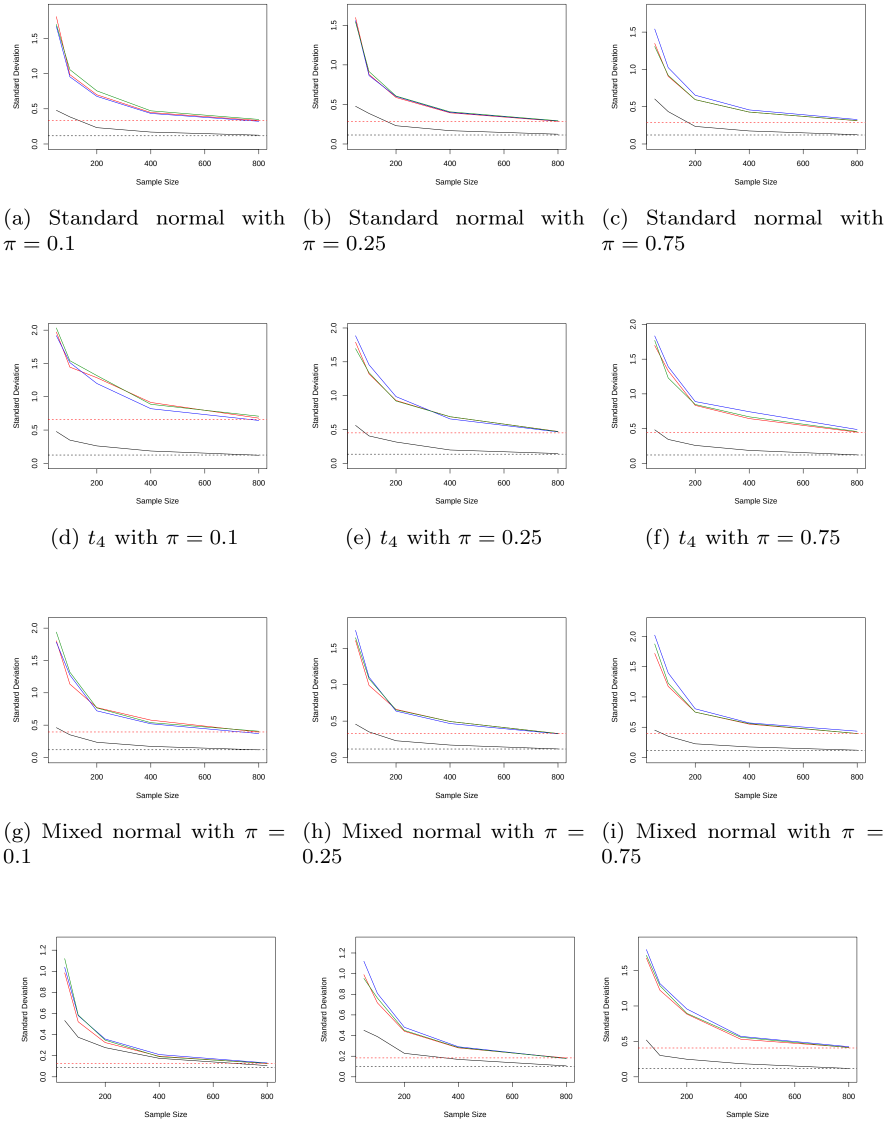
We give an analysis about the computational burden on the parameter estimation approach – generalized method of moments (GMM). There are three steps in GMM and we will count the number of flops for each step.

**Step 1** : Get the intermediate estimator $\hat{\boldsymbol{\zeta}}^*$ by minimizing $e_n^*(\boldsymbol{\zeta})^T e_n^*(\boldsymbol{\zeta})$, where

$$e_n^*(\boldsymbol{\zeta}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \mathbf{W}_i(Y_i - \mu_\pi - \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi)\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi) - \pi\right| \\ \mathrm{vech}(\boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi{\boldsymbol{\Gamma}_\pi}^T + \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}{\boldsymbol{\Gamma}_{0\pi}}^T) - \mathrm{vech}(\mathbf{S}_\mathbf{X}) \\ \boldsymbol{\mu}_\mathbf{X} - \bar{\mathbf{X}} \end{pmatrix}.$$

In this step, we apply Nelder-Mead method to find the minimum of the objective function. It is an iterative method and the number of flops in each iteration is $\mathcal{O}(T_f)$, where $T_f$ represents the number of flops to compute the value of the objective function $e_n^*(\boldsymbol{\zeta})^T e_n^*(\boldsymbol{\zeta})$ for a given $\boldsymbol{\zeta}$ ([7]).

- Because $\mathbf{X}_i$ is a $p$-dimensional vector, $\boldsymbol{\Gamma}_\pi$ is a $p$ by $u_\pi$ matrix and $\boldsymbol{\eta}_\pi$ is a $u_\pi$-dimensional vector, it takes $\mathcal{O}(pu_\pi)$ flops to compute $\mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi$. Afterwards, because $Y_i$, $\mu_\pi$ and $\mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi$ are scalars, it takes $\mathcal{O}(1)$ flops to compute $(Y_i-\mu_\pi-\mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi)$ and $\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi) - \pi\right|$. Next, because $\mathbf{W}_i$ is a $(p+1)$-dimensional vector, it takes $\mathcal{O}(p)$ flops to compute the product $\mathbf{W}_i(Y_i-\mu_\pi-\mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi)\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi) - \pi\right|$. Finally, we need to perform the above multiplication for each sample and then take the average. Hence, the total number of flops to compute the first line in $e_n^*(\boldsymbol{\zeta})$ is $\mathcal{O}(npu_\pi)$.
- Because $\boldsymbol{\Gamma}_\pi$ is a $p$ by $u_\pi$ matrix and $\boldsymbol{\Omega}_\pi$ is a $u_\pi$ by $u_\pi$ matrix, it takes $\mathcal{O}(p^2u_\pi)$ flops to compute $\boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi{\boldsymbol{\Gamma}_\pi}^T$. In addition, because $\boldsymbol{\Gamma}_{0\pi}$ is a $p$ by $(p-u_\pi)$ matrix and $\boldsymbol{\Omega}_{0\pi}$ is a $(p-u_\pi)$ by $(p-u_\pi)$ matrix, it takes $\mathcal{O}(p(p-u_\pi)^2) = \mathcal{O}(p^3)$ flops to compute $\boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}{\boldsymbol{\Gamma}_{0\pi}}^T$. Afterwards, it takes $\mathcal{O}(np^2)$ flops to compute $\mathbf{S}_\mathbf{X} = \sum_{i=1}^n(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})^T/n$. Finally, because $\mathrm{vech}(\boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi{\boldsymbol{\Gamma}_\pi}^T + \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}{\boldsymbol{\Gamma}_{0\pi}}^T)$ and $\mathrm{vech}(\mathbf{S}_\mathbf{X})$ are $\mathcal{O}(p^2)$-dimensional vectors, it takes $\mathcal{O}(p^2)$ flops to compute the difference between them. Hence, the total number of flops to compute the second line in $e_n^*(\boldsymbol{\zeta})$ is $\mathcal{O}(np^2 + p^3)$.



(a) Standard normal with $\pi = 0.1$

(b) Standard normal with $\pi = 0.25$

(c) Standard normal with $\pi = 0.75$

(d) $t_4$ with $\pi = 0.1$

(e) $t_4$ with $\pi = 0.25$

(f) $t_4$ with $\pi = 0.75$

(g) Mixed normal with $\pi = 0.1$

(h) Mixed normal with $\pi = 0.25$

(i) Mixed normal with $\pi = 0.75$

(j) Exp(1) with $\pi = 0.1$

(k) Exp(1) with $\pi = 0.25$

(l) Exp(1) with $\pi = 0.75$

Fig 9: Comparison of the sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator. The horizontal lines mark the asymptotic standard deviations of the ER estimator (the upper line in each panel) and the EER estimator (the lower line in each panel).

- Because $\boldsymbol{\mu}_{\mathbf{X}}$ is a $p$-dimensional vector and $\bar{\mathbf{X}}$ is also a $p$-dimensional vector, it takes $\mathcal{O}(p)$ flops to compute the difference between them. Hence, the total number of flops to compute the third line in $e_n^*(\boldsymbol{\zeta})$ is $\mathcal{O}(p)$.

To sum up, it takes $\mathcal{O}(np^2 + p^3)$ flops to compute $e_n^*(\boldsymbol{\zeta})$. Once we have $e_n^*(\boldsymbol{\zeta})$, it takes another $\mathcal{O}(p^2)$ flops to compute the objective function $e_n^*(\boldsymbol{\zeta})^T e_n^*(\boldsymbol{\zeta})$. So the total number of flops to compute the value of the objective function is $T_f$=$\mathcal{O}(np^2 + p^3)$. Therefore, the number of flops in each iteration of the Nelder-Mead algorithm is $\mathcal{O}(np^2 + p^3)$.

**Step 2** : Compute the scale matrix

$$\hat{\boldsymbol{\Delta}} = \left[\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*)) s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T\right]^{-1},$$

where

$$s(\mathbf{Z}; \psi(\boldsymbol{\zeta})) = \begin{pmatrix} \mathbf{W}(Y - \mu_\tau - \mathbf{X}^T\boldsymbol{\Gamma}_\tau\boldsymbol{\eta}_\tau)\left|I(Y < \mu_\tau + \mathbf{X}^T\boldsymbol{\Gamma}_\tau\boldsymbol{\eta}_\tau) - \tau\right| \\ \text{vech}(\boldsymbol{\Gamma}_\tau\boldsymbol{\Omega}_\tau{\boldsymbol{\Gamma}_\tau}^T + \boldsymbol{\Gamma}_{0\tau}\boldsymbol{\Omega}_{0\tau}{\boldsymbol{\Gamma}_{0\tau}}^T) - \text{vech}\{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\} \\ \boldsymbol{\mu}_{\mathbf{X}} - \mathbf{X}. \end{pmatrix}$$

In this step, we firstly compute the matrix $\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*)) s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T$. Following similar calculations in Step 1, it takes $\mathcal{O}(p^3)$ flops to compute $s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))$. Upon we get $s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))$, it takes another $\mathcal{O}(p^4)$ flops to compute the multiplication $s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*)) s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T$. We need to do this multiplication for each sample and then take the average, then the number of flops to get the matrix $\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*)) s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T$ is $\mathcal{O}(np^4)$. Afterwards, we need to solve for the inversion of the matrix. Matrix inversion takes $\mathcal{O}(m^3)$ flops for an $m$ by $m$ matrix. In our case, it takes $\mathcal{O}(p^6)$ flops for the matrix inversion. So the number of flops in this step is $\mathcal{O}(np^4+p^6)$.

**Step 3** : Obtain the GMM estimator $\hat{\boldsymbol{\zeta}}$ by minimizing $e_n^*(\boldsymbol{\zeta})^T\hat{\boldsymbol{\Delta}}e_n^*(\boldsymbol{\zeta})$. Similar as Step 1, we apply Nelder-Mead method to find the minimum of the objective function $e_n^*(\boldsymbol{\zeta})^T\hat{\boldsymbol{\Delta}}e_n^*(\boldsymbol{\zeta})$. It takes $\mathcal{O}(np^2 + p^3)$ to compute $e_n^*(\boldsymbol{\zeta})$. Once we get $e_n^*(\boldsymbol{\zeta})$, it takes another $\mathcal{O}(p^4)$ flops to compute the objective function $e_n^*(\boldsymbol{\zeta})^T\hat{\boldsymbol{\Delta}}e_n^*(\boldsymbol{\zeta})$. So the total number of flops to compute the value of the objective function is $T_f = \mathcal{O}(np^2 + p^4)$. Therefore, the number of flops in each iteration of the Nelder-Mead algorithm is $\mathcal{O}(np^2 + p^4)$.

## 9. Comparison Between EQR and EER

In this section, we compare the performance between the envelope quantile regression (EQR; [1]) model and the EER model with simulated data and S&P 500 data.

For the simulated data, we use same settings in Section 5 of the paper. Since the true underlying distributions are known, we are able to map expectiles to quantiles under each distribution. For example, 0.19 quantile is identical to 0.10 expectile under the standard normal distribution $\epsilon \sim \mathcal{N}(0, 1)$. The mappings for the four error distributions considered in the simulation are shown in the following Table 5.

Table 5
*The mappings between expectiles and quantiles under the four types of distributions.*

(a) $\epsilon \sim \mathcal{N}(0, 1)$

| Expectile levels $\pi$ | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 |
|---|---|---|---|---|---|
| Quantile levels $\alpha$ | 0.19 | 0.33 | 0.50 | 0.67 | 0.81 |

(b) $\epsilon \sim \mathrm{Exp}(1)$

| Expectile levels $\pi$ | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 |
|---|---|---|---|---|---|
| Quantile levels $\alpha$ | 0.34 | 0.48 | 0.63 | 0.77 | 0.87 |

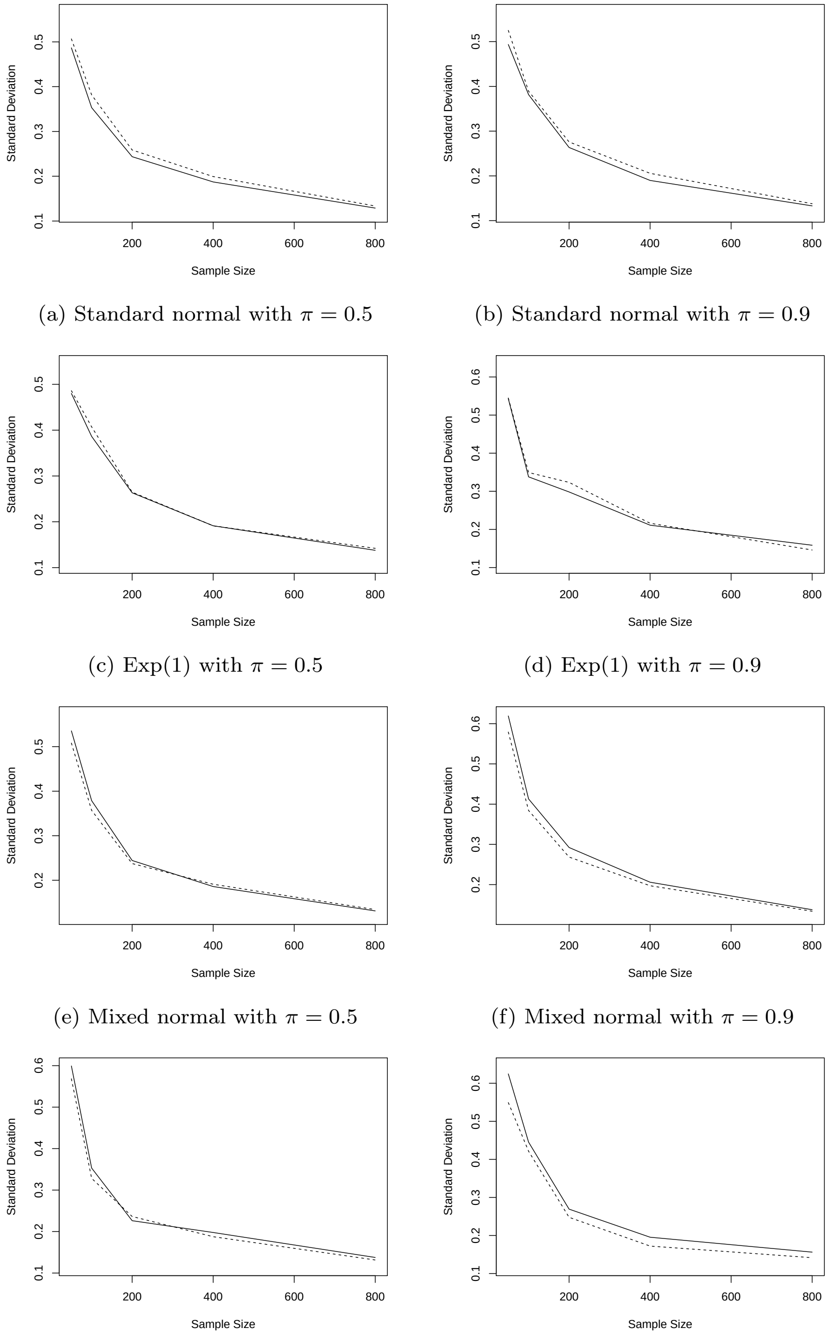
(c) $\epsilon \sim 0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(1, 5)$

| Expectile levels $\pi$ | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 |
|---|---|---|---|---|---|
| Quantile levels $\alpha$ | 0.19 | 0.34 | 0.52 | 0.70 | 0.84 |

(d) $\epsilon \sim t_4$

| Expectile levels $\pi$ | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 |
|---|---|---|---|---|---|
| Quantile levels $\alpha$ | 0.16 | 0.30 | 0.50 | 0.70 | 0.84 |

We repeated the simulation in Section 5 of the manuscript for the EQR model. Then we compared the sample standard deviations of the EER estimator and the EQR estimator, as well as the prediction performance. Note that the expectile levels investigated for the EER model were still 0.10, 0.25, 0.50, 0.75 and 0.90, while their corresponding quantile level mappings given in Table 5 were used for the EQR model. The results are summarized in Figure 10 and Tables 6.



(a) Standard normal with $\pi = 0.5$

(b) Standard normal with $\pi = 0.9$

(c) Exp(1) with $\pi = 0.5$

(d) Exp(1) with $\pi = 0.9$

(e) Mixed normal with $\pi = 0.5$

(f) Mixed normal with $\pi = 0.9$

(g) $t_4$ with $\pi = 0.5$

(h) $t_4$ with $\pi = 0.9$

Fig 10: Sample standard deviations. Solid lines mark the EER estimator. Dashed lines mark the EQR estimator.

Table 6
*The average RMSEs of the 300 replications under the EER model and EQR model with different error distributions.*

(a) $\epsilon \sim \mathcal{N}(0,1)$

| $\pi(\alpha)$ | 0.10 (0.19) | 0.25 (0.33) | 0.50 (0.50) | 0.75 (0.67) | 0.90 (0.81) |
|---|---|---|---|---|---|
| EER | 1.05 | 0.94 | 0.93 | 0.98 | 1.12 |
| EQR | 1.06 | 1.00 | 0.98 | 1.01 | 1.09 |

(b) $\epsilon \sim \text{Exp}(1)$

| $\pi(\alpha)$ | 0.10 (0.34) | 0.25 (0.48) | 0.50 (0.63) | 0.75 (0.77) | 0.90 (0.87) |
|---|---|---|---|---|---|
| EER | 0.70 | 0.77 | 0.95 | 1.32 | 2.01 |
| EQR | 0.76 | 0.86 | 1.06 | 1.29 | 1.70 |

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

| $\pi(\alpha)$ | 0.10 (0.19) | 0.25 (0.34) | 0.50 (0.52) | 0.75 (0.70) | 0.90 (0.84) |
|---|---|---|---|---|---|
| EER | 1.19 | 1.02 | 1.02 | 1.21 | 1.71 |
| EQR | 1.11 | 0.99 | 0.99 | 1.07 | 1.29 |

(d) $\epsilon \sim t_4$

| $\pi(\alpha)$ | 0.10 (0.16) | 0.25 (0.30) | 0.50 (0.50) | 0.75 (0.70) | 0.90 (0.84) |
|---|---|---|---|---|---|
| EER | 1.71 | 1.22 | 1.11 | 1.27 | 1.82 |
| EQR | 1.40 | 1.09 | 1.01 | 1.09 | 1.36 |

The estimation efficiency is similar for the EER estimator and the EQR estimator. The sample standard deviations of the EER estimators are very close to those of the EQR estimators, as indicated in Figure 10. Additionally, we observed that under the distributions with relatively smaller variance (standard normal and Exp(1)), the sample standard deviations of the EER estimators are slightly smaller than those of the EQR estimators. While under the distributions with relatively larger variance (mixed normal and $t_4$), the sample standard deviations of the EER estimators become slightly larger than those of the EQR estimators. The observation is consistent with the property that expectiles are more sensitive to extreme values. Under distributions with relatively larger variance, it is more likely to have extreme values in the data, which results in more variant EER estimators than the EQR estimators. Similar observation is shown in Table 6 as well. Under standard normal and Exp(1), the average RMSEs from the EER model are slightly smaller than those from the EQR model for most expectile levels. While under mixed normal and $t_4$, the average RMSEs from the EER model become larger than those from the EQR model.

For S&P 500 data, since we do not have direct knowledge on the underlying distribution, a grid of 23 levels (0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.98 and 0.99) were investigated as quantiles under the EQR model and expectiles under the EER model. To compare prediction performance, we notice that the measure of prediction performance are different for the EQR model and the EER model: EQR model uses the quantile loss and EER model uses the expectile loss. Therefore we firstly computed the mean of the predicted quantile loss for each level under both the EQR model and the EER model. Among the 23 levels, the mean of predicted quantile loss under the EQR model ranges from $3.0{\times}10^{-3}$ to $3.1{\times}10^{-2}$ with an average of $2.0{\times}10^{-2}$. The mean of predicted quantile loss under the EER model ranges from $3.0{\times}10^{-3}$ to $3.1{\times}10^{-2}$ with an average of $2.1{\times}10^{-2}$. Boxplots of the predicted quantile loss under the two models are included in the left panel of Figure 11. Secondly we computed the mean of the predicted expectile loss under both models for each level, and the results are included in the boxplots in the right panel of Figure 11. Among the 23 levels, the mean of the predicted expectile loss under the EQR model ranges from $8.8{\times}10^{-4}$ to $3.3{\times}10^{-3}$ with an average of $2.5{\times}10^{-3}$. The mean of predicted expectile loss under the EER model ranges from $4.7{\times}10^{-4}$ to $3.6{\times}10^{-3}$ with an average of $2.3{\times}10^{-3}$. Based on the ranges and averages, we can not identify a statistically significant difference of the prediction performance between the EQR model and the EER model in this case.

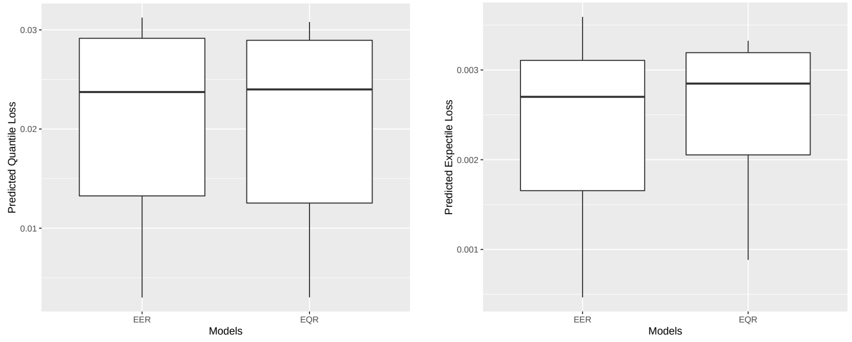


Fig 11: Boxplots of predicted quantile loss and expectile loss for the two models.

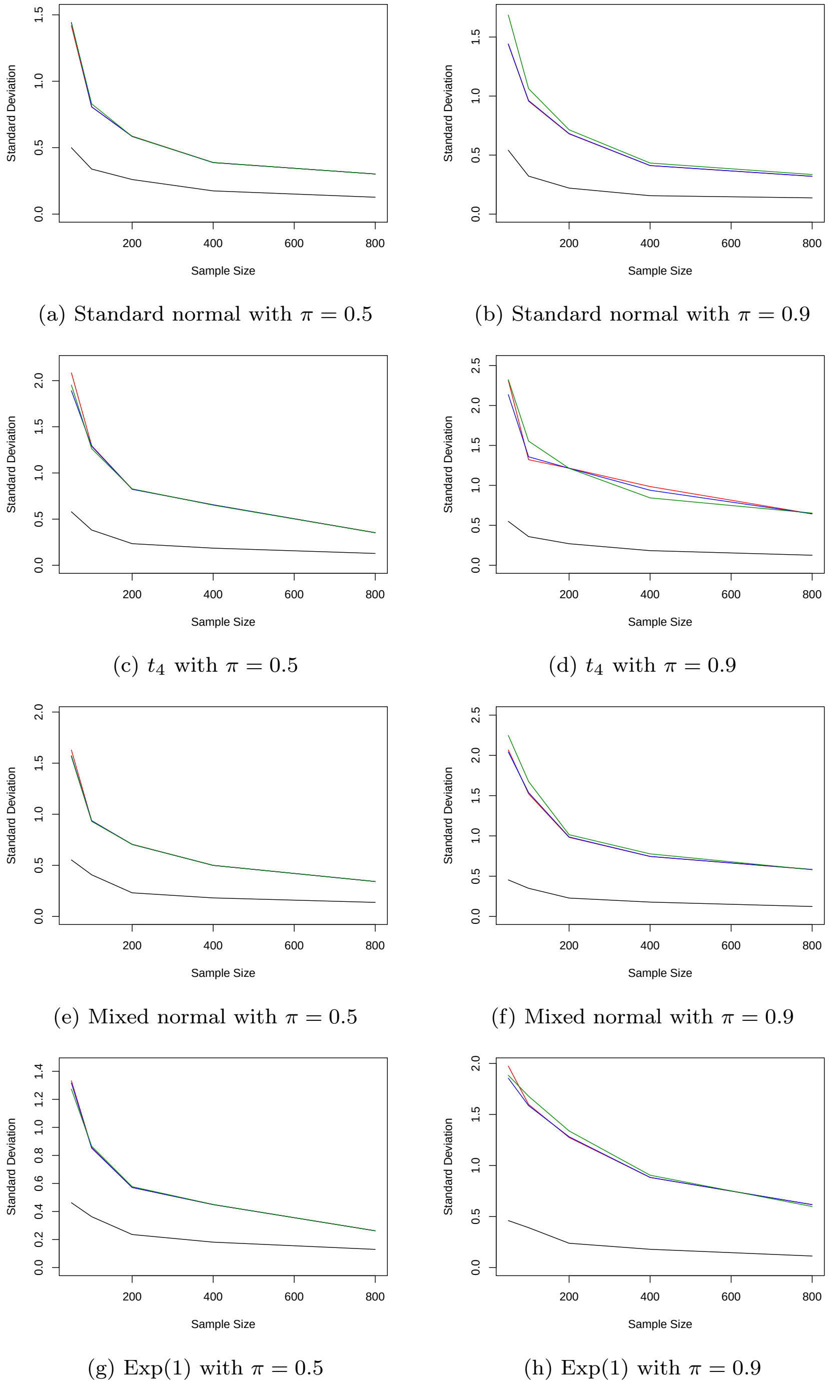
Similar to the relationship between the QR and the ER, the EQR model and the EER model have their unique advantages over each other, and neither approach is uniformly superior. We need to choose the appropriate model based on the goal and context of the problems. For example, if we want to evaluate the potential loss from a portfolio and we are strongly risk averse, then we may use the EER model because it is more sensitive to the extreme losses. If we hope to have a model that is easier to interpret, then we may want to use the EQR model.

## 10. Simulation Results for Sparse Expectile Regression Estimator with an Alternative Tuning Parameter

In the same setting as in Section 5 of the manuscript, we update the results of the sparse ER estimator with a different tuning parameter. The sparse ER estimator was computed by R package `SALES` [2]. It gives two choices of the tuning parameter $\lambda$: $\lambda_{min}$, which is the value of $\lambda$ that minimizes the cross validation error, and $\lambda_{1se}$, which is the largest value of $\lambda$ having its cross validation error within one standard error of the minimum cross validation error. The package takes $\lambda_{1se}$ as the default value for the subsequent variable selection and parameter estimation, and the corresponding results are included in the manuscript. In this section, we update the results using $\lambda_{min}$ as the tuning parameter. We included the ER estimator, EER estimator and the boosting estimator in all figures and tables for completeness. Note that the results for these estimators are unchanged.

The sample standard deviations are included in Figure 12. We do not observe a big difference in sample standard deviation between the sparse ER estimators using $\lambda_{1se}$ (page 14 of the manuscript) and using $\lambda_{min}$. But it seems that the sparse ER estimator has a slightly smaller sample standard deviation with $\lambda_{min}$.

We also calculated the root mean squares errors (RMSE) of the sparse ER estimator using $\lambda_{min}$ as the tuning parameter. The results are in Table 7. Compared to Table 1 of the manuscript (page 16), we notice that the performance of the sparse ER estimator gets better, and its RMSE is very close to the ER estimator and the boosting estimator. It seems that the default value $\lambda_{1se}$ gives a model that is too parsimonious for this settings.



(a) Standard normal with $\pi = 0.5$

(b) Standard normal with $\pi = 0.9$

(c) $t_4$ with $\pi = 0.5$

(d) $t_4$ with $\pi = 0.9$

(e) Mixed normal with $\pi = 0.5$

(f) Mixed normal with $\pi = 0.9$

(g) Exp(1) with $\pi = 0.5$

(h) Exp(1) with $\pi = 0.9$

Fig 12: Comparison of the sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER (with $\lambda_{min}$ as the tuning parameter) estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator.

TABLE 7

*Comparison of the RMSEs, averaged over 300 replications. Using $\lambda_{min}$ as the selected value of $\lambda$ for the sparse ER estimator.*

(a) $\epsilon \sim \mathcal{N}(0, 1)$

| | EER | ER | Boosting | Sparse ER |
|---|---|---|---|---|
| $\pi = 0.10$ | 1.04 | 1.85 | 1.86 | 1.86 |
| $\pi = 0.25$ | 0.93 | 1.60 | 1.60 | 1.60 |
| $\pi = 0.50$ | 0.90 | 1.52 | 1.52 | 1.52 |
| $\pi = 0.75$ | 0.95 | 1.61 | 1.62 | 1.61 |
| $\pi = 0.90$ | 1.10 | 1.87 | 1.91 | 1.88 |

(b) $\epsilon \sim t_4$

| | EER | ER | Boosting | Sparse ER |
|---|---|---|---|---|
| $\pi = 0.10$ | 1.84 | 3.49 | 3.50 | 3.54 |
| $\pi = 0.25$ | 1.28 | 2.46 | 2.47 | 2.47 |
| $\pi = 0.50$ | 1.15 | 2.14 | 2.15 | 2.14 |
| $\pi = 0.75$ | 1.31 | 2.44 | 2.45 | 2.44 |
| $\pi = 0.90$ | 1.85 | 3.46 | 3.51 | 3.47 |

(c) $\epsilon \sim 0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(1, 5)$

| | EER | ER | Boosting | Sparse ER |
|---|---|---|---|---|
| $\pi = 0.10$ | 1.20 | 2.21 | 2.22 | 2.23 |
| $\pi = 0.25$ | 1.05 | 1.87 | 1.87 | 1.88 |
| $\pi = 0.50$ | 1.05 | 1.86 | 1.87 | 1.86 |
| $\pi = 0.75$ | 1.24 | 2.21 | 2.22 | 2.22 |
| $\pi = 0.90$ | 1.75 | 3.14 | 3.18 | 3.16 |

(d) $\epsilon \sim \text{Exp}(1)$

| | EER | ER | Boosting | Sparse ER |
|---|---|---|---|---|
| $\pi = 0.10$ | 0.70 | 0.76 | 0.79 | 0.75 |
| $\pi = 0.25$ | 0.77 | 1.07 | 1.07 | 1.06 |
| $\pi = 0.50$ | 0.96 | 1.54 | 1.54 | 1.54 |
| $\pi = 0.75$ | 1.34 | 2.27 | 2.28 | 2.27 |
| $\pi = 0.90$ | 1.99 | 3.37 | 3.40 | 3.39 |

## References

[1] Ding, S., Su, Z., Zhu, G., and Wang, L. (2019). Envelope quantile regression. *Statistica Sinica*, To appear.

[2] Gu, Y. and Zou, H. (2016). Sales: Elastic net and (adaptive) lasso penalized sparse asymmetric least squares (sales) and coupled sparse asymmetric least squares (cosales) using coordinate descent and proximal gradient algorithms. *R package version 1.0.0*.

[3] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

[4] Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55**, 4, 819–47.

[5] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society* **57**, 5, 1027–1057.

[6] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association* **81**, 393, 142–149.

[7] Singer, S. and Singer, S. (1999). Complexity analysis of nelder-mead search iterations. In *Proceedings of the 1. Conference on Applied Mathematics and Computation, Dubrovnik, Croatia.* PMF–Matematički odjel, Zagreb, 185–196.

[8] Van der Vaart, A. W. (1998). *Asymptotic statistics.* Vol. **3**. Cambridge university press.