Supplement: Proofs, Technical Details and Additional Simulations for “Envelope Quantile Regression”

Shanshan Ding, Zhihua Su, Guangyu Zhu, and Lan Wang

A Proof of Theorem 1

Since the estimating equations (3.1) contain non-smoothing functions, we apply the result in Theorem 3.3 of Pakes and Pollard (1989) to derive the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$. In order to use their result, we firstly need to check the conditions (i)-(v) in their theorem. The conditions (ii) and (v) automatically hold given (C2) and (C3). Since $\mathrm{E}_{\boldsymbol{\theta}_0}[h_n(\boldsymbol{\theta}_0)] = 0$, the condition (iv) can be easily verified by the central limit theorem and (C3). Hence we only need to check the conditions (i) and (iii) in their paper.

By the subgradient condition of quantile regression (Koenker, 2005), there exists a vector $\mathbf{v}$ with components $|v_i| < 1$ such that

$$||h_{1,n}(\tilde{\boldsymbol{\theta}}_1)|| = n^{-1}||(\mathbf{W}_i v_i : i \in \Upsilon)|| = o_p(n^{-1/2}), \text{ by condition (C3),}$$

where Υ denotes a p-element subset of $\{1, 2, \cdots, n\}$ and $\mathbf{W}_i = (1, \mathbf{X}_i^T)^T$. Since $h_{2,n}(\tilde{\boldsymbol{\theta}}_2)$ and $h_{3,n}(\tilde{\boldsymbol{\theta}}_2)$ are both equal to zero, we have $||h_n(\tilde{\boldsymbol{\theta}})|| = ||h_{1,n}(\tilde{\boldsymbol{\theta}}_1)|| = o_p(n^{-1/2})$. Hence the condition (i) in Pakes and Pollard (1989) is satisfied.

To prove their condition (iii), it suffices to show the following Lemma 1. Let $h(\boldsymbol{\theta}) =$

$\mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z}_i, \boldsymbol{\theta})] = \mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z}, \boldsymbol{\theta})]$.

Lemma 1. *Under (C1) and (C3), for every sequence of positive numbers $\delta_n = o(1)$,*

$$\sup_{\boldsymbol{\theta}:||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\leq\delta_n} ||h_n(\boldsymbol{\theta}) - h(\boldsymbol{\theta}) - h_n(\boldsymbol{\theta}_0)|| = o_p(n^{-1/2}). \tag{A.1}$$

Proof. For notational simplicity, in the following we omit the subscript '$\mathbf{X}$' in $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$. Let w_j, μ_j, σ_j, $g_{i,j}$, $i = 1, 2, 3$, denote the jth components of $\mathbf{W}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $g_1(\mathbf{Z}; \boldsymbol{\theta}_1)$, $g_2(\mathbf{Z}; \boldsymbol{\theta}_2)$ and $g_3(\mathbf{Z}; \boldsymbol{\theta}_2)$, respectively. Let c be some positive constant. Then for any $\boldsymbol{\theta}^\star \in \boldsymbol{\Theta}$ and $j = 1, \ldots, p+1$,

$$\Big| \, g_{1,j}(\mathbf{Z}; \boldsymbol{\theta}_1) - g_{1,j}(\mathbf{Z}; \boldsymbol{\theta}_1^\star) \, \Big|^2 \leq w_j^2 \, \Big| \, I(Y < \mathbf{W}^T\boldsymbol{\theta}_1) - I(Y < \mathbf{W}^T\boldsymbol{\theta}_1^\star) \, \Big|$$

It can be shown that

$$\sup_{\boldsymbol{\theta}^\star:||\boldsymbol{\theta}-\boldsymbol{\theta}^\star||\leq\delta_n} \Big| \, I(Y < \mathbf{W}^T\boldsymbol{\theta}_1) - I(Y < \mathbf{W}^T\boldsymbol{\theta}_1^\star) \, \Big| \leq ||\mathbf{W}|| \big[I(Y < \mathbf{W}^T\boldsymbol{\theta}_1 + \delta_n) - I(Y < \mathbf{W}^T\boldsymbol{\theta}_1 - \delta_n) \big].$$

Hence by (C1) and (C3), there exists a positive constant c, such that

$$\begin{aligned} & \mathrm{E}\Big(\sup_{\boldsymbol{\theta}^\star:||\boldsymbol{\theta}-\boldsymbol{\theta}^\star||\leq\delta_n} \Big| \, g_{1,j}(\mathbf{Z}; \boldsymbol{\theta}_1) - g_{1,j}(\mathbf{Z}; \boldsymbol{\theta}_1^\star) \, \Big|^2 \, \Big) \\ \leq \; & \mathrm{E}\Big(w_j^2 ||\mathbf{W}|| \big[I(Y < \mathbf{W}^T\boldsymbol{\theta}_1 + \delta_n) - I(Y < \mathbf{W}^T\boldsymbol{\theta}_1 - \delta_n) \big] \Big) \\ \leq \; & \mathrm{E}\Big(||\mathbf{W}||^3 \big[F_Y(\mathbf{W}^T\boldsymbol{\theta}_1 + \delta_n|\mathbf{X}) - F_Y(\mathbf{W}^T\boldsymbol{\theta}_1 - \delta_n|\mathbf{X}) \big] \Big) \; \leq \; c\delta_n. \end{aligned} \tag{A.2}$$

Let $\mu_j^\star$ and $\sigma_j^\star$ be the jth components of $\boldsymbol{\mu}^\star$ and $\boldsymbol{\Sigma}^\star$, respectively, and let vech$[\,\cdot\,]_j$ be the jth component of the corresponding vech$[\,\cdot\,]$. Then for any $j = 1, \ldots, s$,

$$\Big|\ g_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2) - g_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2^\star)\ \Big|^2 = \Big|\ \sigma_j - \text{vech}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T)]_j - \sigma_j^\star + \text{vech}[(\mathbf{X}-\boldsymbol{\mu}^\star)(\mathbf{X}-\boldsymbol{\mu}^\star)^T)]_j\ \Big|^2 .$$

By Condition (C3), it is easy to verify that

$$\mathrm{E}\Big(\sup_{\boldsymbol{\theta}^\star:||\boldsymbol{\theta}-\boldsymbol{\theta}^\star||\le\delta_n}\Big|\ g_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2) - g_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2^\star)\ \Big|^2\Big) \le k\delta_n^2, \text{ for any } j = 1, \ldots, s, \tag{A.3}$$

where k is some positive constant. Next for any $j = 1, \ldots, p$, we have

$$E\Big(\sup_{\boldsymbol{\theta}^\star:||\boldsymbol{\theta}-\boldsymbol{\theta}^\star||\le\delta_n}\Big|\ g_{3,j}(\mathbf{Z};\boldsymbol{\theta}_2) - g_{3,j}(\mathbf{Z};\boldsymbol{\theta}_2^\star)\ \Big|^2\Big) = \mathrm{E}\Big(\sup_{\boldsymbol{\theta}^\star:||\boldsymbol{\theta}-\boldsymbol{\theta}^\star||\le\delta_n}\Big|\ \mu_j - \mu_j^\star\ \Big|^2) \le \delta_n^2. \tag{A.4}$$

The results in (A.2), (A.3), and (A.4) together imply that $g(\mathbf{Z};\boldsymbol{\theta})$ belongs to the "type IV class" of Andrews (1994) and is $\mathcal{L}^2(P)$–continuous at $\boldsymbol{\theta}$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. For details regarding this statement, see (5.3) in Andrews (1994). In addition, by (C3), $\mathrm{E}_{\boldsymbol{\theta}_0}[g_{i,j}^2] < \infty$ for all $i = 1, 2, 3$ and the corresponding indices j. Thus, by applying Lemma 2.17 in Pakes and Pollard (1989), we have

$$n^{-1/2}\sup_{\boldsymbol{\theta}:||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\le\delta_n}\Big|\Big|\sum_{i=1}^{n}\Big\{g(\mathbf{Z};\boldsymbol{\theta}) - \mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z};\boldsymbol{\theta})] - g(\mathbf{Z};\boldsymbol{\theta}_0)\Big\}\Big|\Big| = o_p(1).$$

This result can also be obtained by using Theorem 3 in Chen, Linton, and Van Keilegom (2003) while omitting the infinite dimensional parameter h. □

So far, we have verified the conditions (i)-(v) in Theorem 3.3 of Pakes and Pollard (1989). In addition, it is easy to show that $\tilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$. Therefore, using their theorem 3.3, we can directly obtain

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\Big(0, (\mathbf{U}_1\mathbf{U}_1^T)^{-1}\mathbf{U}_1\mathbf{V}_1\mathbf{U}_1^T(\mathbf{U}_1\mathbf{U}_1^T)^{-1}\Big),$$

where $\mathbf{U}_1 = \frac{\partial \mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z};\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \nabla \mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z};\boldsymbol{\theta}_0)]$ and $\mathbf{V}_1 = \mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z};\boldsymbol{\theta}_0)g^T(\mathbf{Z};\boldsymbol{\theta}_0)]$.

To find the expression of $\mathbf{U}_1$, consider

$$\mathrm{E}_{\boldsymbol{\theta}_0}[g_1(\mathbf{Z},\boldsymbol{\theta}_1)] = \mathrm{E}_{\boldsymbol{\theta}_0}\big[\mathbf{W}\big(I(Y < \mathbf{W}^T\boldsymbol{\theta}_1) - \tau\big)\big] = \mathrm{E}_{\boldsymbol{\theta}_0}\big[\mathbf{W}\big(F_Y(\mathbf{W}^T\boldsymbol{\theta}_1|\mathbf{X}) - \tau\big)\big]$$

$$\mathrm{E}_{\boldsymbol{\theta}_0}[g_2(\mathbf{Z},\boldsymbol{\theta}_2)] = \mathrm{vech}(\boldsymbol{\Sigma}) - \mathrm{vech}(\boldsymbol{\Sigma}_0), \;\; \mathrm{E}_{\boldsymbol{\theta}_0}[g_3(\mathbf{Z},\boldsymbol{\theta}_2)] = \mathrm{E}_{\boldsymbol{\theta}_0}(\boldsymbol{\mu} - \mathbf{X}) = \boldsymbol{\mu} - \boldsymbol{\mu}_0,$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the true values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. Recall that $g(\mathbf{Z};\boldsymbol{\theta}) = (g_1^T(\mathbf{Z};\boldsymbol{\theta}_1), g_2^T(\mathbf{Z};\boldsymbol{\theta}_2), g_3^T(\mathbf{Z};\boldsymbol{\theta}_2))^T$. Therefore,

$$\mathbf{U}_1 = \nabla \mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z};\boldsymbol{\theta}_0)] = \begin{pmatrix} \mathrm{E}_{\boldsymbol{\theta}_0}[f_{Y|\mathbf{X}}(\xi_0(\tau|\mathbf{X}))\mathbf{W}\mathbf{W}^T] & 0 & 0 \\ 0 & \mathbf{I}_s & 0 \\ 0 & 0 & \mathbf{I}_p \end{pmatrix}, \tag{A.5}$$

where $\xi_0(\tau|\mathbf{X}) = \mathbf{W}^T\boldsymbol{\theta}_{1,0}$ is the conditional quantile of $Y|\mathbf{X}$ under the true value $\boldsymbol{\theta}_{1,0}$ of $\boldsymbol{\theta}_1$, and $\mathbf{U}_1$ is symmetric. By (C2), $\mathbf{U}_1$ also has a full rank. Hence the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$ can be simplified to

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \ \xrightarrow{d} \ N\Big(0, \mathbf{U}_1^{-1}\mathbf{V}_1\mathbf{U}_1^{-1}\Big).$$

Now we give the expression of $\mathbf{V}_1 = \big(\mathbf{V}_{1,ij}\big)_{i,j=1,2,3}$. Similarly, let $\boldsymbol{\theta}_{2,0}$ denote the true value of $\boldsymbol{\theta}_2$. Then

$$\mathbf{V}_{1,11} = \mathrm{E}_{\boldsymbol{\theta}_0}[g_1(\mathbf{Z};\boldsymbol{\theta}_{1,0})g_1^T(\mathbf{Z};\boldsymbol{\theta}_{1,0})] = \mathrm{E}_{\boldsymbol{\theta}_0}\big[\big(I(Y < \mathbf{W}^T\boldsymbol{\theta}_{1,0}) - \tau\big)^2\mathbf{W}\mathbf{W}^T\big] = \tau(1-\tau)\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{W}\mathbf{W}^T)$$

$$\mathbf{V}_{1,22} = \mathrm{E}_{\boldsymbol{\theta}_0}[g_2(\mathbf{Z};\boldsymbol{\theta}_{2,0})g_2^T(\mathbf{Z};\boldsymbol{\theta}_{2,0})] = \mathrm{var}_{\boldsymbol{\theta}_0}\big\{\mathrm{vech}[(\mathbf{X}-\boldsymbol{\mu}_0)(\mathbf{X}-\boldsymbol{\mu}_0)^T]\big\}$$

$$\mathbf{V}_{1,33} = \mathrm{E}_{\boldsymbol{\theta}_0}[g_3(\mathbf{Z};\boldsymbol{\theta}_{2,0})g_3^T(\mathbf{Z};\boldsymbol{\theta}_{2,0})] = \mathrm{E}_{\boldsymbol{\theta}_0}[(\boldsymbol{\mu}_0-\mathbf{X})(\boldsymbol{\mu}_0-\mathbf{X})^T] = \mathrm{var}_{\boldsymbol{\theta}_0}(\mathbf{X})$$

$$\mathbf{V}_{1,23} = \mathrm{E}_{\boldsymbol{\theta}_0}[g_2(\mathbf{Z};\boldsymbol{\theta}_{2,0})g_3^T(\mathbf{Z};\boldsymbol{\theta}_{2,0})] = \mathrm{E}_{\boldsymbol{\theta}_0}\Big\{\mathrm{vech}[(\mathbf{X}-\boldsymbol{\mu}_0)(\mathbf{X}-\boldsymbol{\mu}_0)^T](\boldsymbol{\mu}_0-\mathbf{X})^T\Big\}.$$

In addition, it is easy to check that $\mathbf{V}_{1,1j} = \mathrm{E}_{\boldsymbol{\theta}_0}[g_1(\mathbf{Z};\boldsymbol{\theta}_{1,0})g_j^T(\mathbf{Z};\boldsymbol{\theta}_{2,0})] = 0$ for j=2,3. Correspondingly, we have

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*) \ \xrightarrow{d} \ N\Big(0, \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}\Big),$$

where

$$\mathbf{U} = \begin{pmatrix} \mathrm{E}_{\boldsymbol{\theta}_0}[f_{Y|\mathbf{X}}(\xi_0(\tau|\mathbf{X}))\mathbf{W}\mathbf{W}^T] & 0 \\ 0 & \mathbf{I}_s \end{pmatrix} \text{ and } \mathbf{V} = \begin{pmatrix} \mathbf{V}_{1,11} & 0 \\ 0 & \mathbf{V}_{1,22} \end{pmatrix}.$$

We thus complete the proof of Theorem 1.

B Proof of Theorem 2

For notational simplicity, let $Q_n(\boldsymbol{\theta}) = h_n^T(\boldsymbol{\theta})\widehat{\boldsymbol{\Delta}}h_n(\boldsymbol{\theta})$ and $Q(\boldsymbol{\theta}) = h^T(\boldsymbol{\theta})\boldsymbol{\Delta}h(\boldsymbol{\theta})$, where $h_n(\boldsymbol{\theta})$ is given by (3.3), $\widehat{\boldsymbol{\Delta}} = \left\{n^{-1}\sum_{i=1}^n g(\mathbf{Z}_i;\tilde{\boldsymbol{\theta}})g^T(\mathbf{Z}_i;\tilde{\boldsymbol{\theta}})\right\}^{-1}$, $h(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z},\boldsymbol{\theta})]$, and $\boldsymbol{\Delta} = V_1^{-1} = \left\{\mathrm{E}_{\boldsymbol{\theta}_0}[g(\mathbf{Z};\boldsymbol{\theta}_0)g^T(\mathbf{Z};\boldsymbol{\theta}_0)]\right\}^{-1}$. Let $l_n(\boldsymbol{\gamma}) = h_n(\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0)$ and $l(\boldsymbol{\gamma}) = h(\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0)$. Thus, $l_n(0) = h_n(\boldsymbol{\theta}_0)$ and $l(0) = h(\boldsymbol{\theta}_0) = 0$. Let $T_n(\boldsymbol{\gamma}) = l_n^T(\boldsymbol{\gamma})\widehat{\boldsymbol{\Delta}}l_n(\boldsymbol{\gamma})$ and $T(\boldsymbol{\gamma}) = l^T(\boldsymbol{\gamma})\boldsymbol{\Delta}l(\boldsymbol{\gamma})$. In addition, let $\varepsilon_n(\boldsymbol{\gamma}) = [l_n(\boldsymbol{\gamma}) - l_n(0) - l(\boldsymbol{\gamma})]/(1+||\boldsymbol{\gamma}||)$, $\kappa_n(\boldsymbol{\gamma}) = \varepsilon_n^T(\boldsymbol{\gamma})\widehat{\boldsymbol{\Delta}}\varepsilon_n(\boldsymbol{\gamma}) + 2l_n^T(0)\widehat{\boldsymbol{\Delta}}\varepsilon_n(\boldsymbol{\gamma})$, and $\rho_n(\boldsymbol{\gamma}) = n[T_n(\boldsymbol{\gamma}) - \kappa_n(\boldsymbol{\gamma}) - T_n(0) - \widehat{\mathbf{D}}^T\boldsymbol{\gamma}/\sqrt{n} - T(\boldsymbol{\gamma})]$, where $\widehat{\mathbf{D}} = 2\mathbf{U}_1\widehat{\boldsymbol{\Delta}}l_n(0)$ with $\mathbf{U}_1$ given by (A.5).

The proof of Theorem 2 relies on the following Lemmas 2-4.

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

Proof. Let $\mathcal{F} = \{g(\mathbf{Z},\boldsymbol{\theta}), \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta}$ is compact under (C3). It is easy to verify that $\mathcal{F}$ is a VC-class. Thus, under the proposed conditions, $\mathcal{F}$ is Glivenko-Cantelli by Theorem 19.4 and Lemma 19.15 in Van der Vaart (1998). That is,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||h_n(\boldsymbol{\theta}) - h(\boldsymbol{\theta})|| \to 0, \text{ a.s.}$$

As a result, $Q_n(\boldsymbol{\theta})$ uniformly converges to $Q(\boldsymbol{\theta})$ in probability. Correspondingly, $Q_n(\boldsymbol{\theta})$ uniformly converges to $Q(\boldsymbol{\theta})$ in probability for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}_e$, where $\boldsymbol{\Theta}_e = \{\boldsymbol{\theta} : \boldsymbol{\theta}\in\boldsymbol{\Theta}$ and $\boldsymbol{\theta} = \psi_0(\boldsymbol{\zeta}_\tau^{'})\}$. As the support of $\boldsymbol{\zeta}_\tau^{'}$ is compact, $\boldsymbol{\Theta}_e$ is compact. Moreover, since $h(\boldsymbol{\theta}_0) = 0$ and $\boldsymbol{\theta}_0$ is the unique root, $\boldsymbol{\theta}_0$ is the unique minimizer of $Q(\boldsymbol{\theta})$. In addition, by (C2), $Q(\boldsymbol{\theta})$ is continuous. Therefore, by applying Theorem 2.1 in Newey and McFadden (1994), we have

$\widehat{\boldsymbol{\theta}}_g \xrightarrow{p} \boldsymbol{\theta}_0$. □

Lemma 3. *Under the same conditions in Theorem 2, for every sequence of positive numbers* $\delta_n = o(1)$,

$$\sup_{\vartheta} \frac{|\rho_n(\boldsymbol{\gamma})|}{||\boldsymbol{\gamma}||(1+||\boldsymbol{\gamma}||)} = o_p(1), \tag{B.1}$$

where $\vartheta = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma}||/\sqrt{n} \leq \delta_n\}$.

Proof. Based on the definition of $\varepsilon_n(\boldsymbol{\gamma})$, we have

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

and $T_n(0) = l_n^T(0)\widehat{\boldsymbol{\Delta}}l_n(0)$. Consequently, it can be shown that $|\rho_n(\boldsymbol{\gamma})|/[||\boldsymbol{\gamma}||(1+||\boldsymbol{\gamma}||)] = \sum_{j=1}^5 B_j(\boldsymbol{\gamma})$, where

$$\begin{aligned} & B_1(\boldsymbol{\gamma}) = n(||\boldsymbol{\gamma}||+2)\varepsilon_n^T(\boldsymbol{\gamma})\widehat{\boldsymbol{\Delta}}\varepsilon_n(\boldsymbol{\gamma})/(1+||\boldsymbol{\gamma}||), \;\; B_2(\boldsymbol{\gamma}) = 2n|\varepsilon_n^T(\boldsymbol{\gamma})\widehat{\boldsymbol{\Delta}}l_n(0)|/(1+||\boldsymbol{\gamma}||), \\ & B_3(\boldsymbol{\gamma}) = 2n|\varepsilon_n^T(\boldsymbol{\gamma})\widehat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma})|/||\boldsymbol{\gamma}||, \;\; B_4(\boldsymbol{\gamma}) = n|2l_n^T(0)\widehat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma}) - \widehat{\mathbf{D}}^T\boldsymbol{\gamma}/\sqrt{n}|/[||\boldsymbol{\gamma}||(1+||\boldsymbol{\gamma}||)], \\ & B_5(\boldsymbol{\gamma}) = n|l^T(\boldsymbol{\gamma})(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})l(\boldsymbol{\gamma})|/[||\boldsymbol{\gamma}||(1+||\boldsymbol{\gamma}||)]. \end{aligned}$$

To prove Lemma 3, it suffices to show that $\sup_{\vartheta} B_j(\boldsymbol{\gamma}) = o_p(1)$ for all $j = 1, \ldots, 5$. From Lemma 1, we know that $\sup_{\vartheta} ||\varepsilon_n(\boldsymbol{\gamma})|| = o_p(n^{-1/2})$. Then under Condition (C2), $\sup_{\vartheta} B_1(\boldsymbol{\gamma}) \leq \sup_{\vartheta} n||\varepsilon_n(\boldsymbol{\gamma})||^2||\widehat{\boldsymbol{\Delta}}||(||\boldsymbol{\gamma}||+2)/(1+||\boldsymbol{\gamma}||) = n\sup_{\vartheta}||\varepsilon_n(\boldsymbol{\gamma})||^2 O_p(1) = o_p(1)$, and $\sup_{\vartheta} B_2(\boldsymbol{\gamma}) \leq \sqrt{n}\sup_{\vartheta}||\varepsilon_n(\boldsymbol{\gamma})||O_p(1) = o_p(1)$. By Taylor expansion, $l(\boldsymbol{\gamma}) = \mathbf{U}_1\boldsymbol{\gamma}/\sqrt{n}+$

$o(\boldsymbol{\gamma}/\sqrt{n})$. Thus, $\sup_\vartheta B_3(\boldsymbol{\gamma}) \leq 2\sqrt{n}\sup_\vartheta ||\varepsilon_n(\boldsymbol{\gamma})||||\widehat{\boldsymbol{\Delta}}||(||\mathbf{U}_1||||\boldsymbol{\gamma}||+o(||\boldsymbol{\gamma}||)/||\boldsymbol{\gamma}|| \leq \sqrt{n}\sup_\vartheta ||\varepsilon_n(\boldsymbol{\gamma})||O_p(1) = o_p(1)$, and $\sup_\vartheta B_4(\boldsymbol{\gamma}) = 2n\sup_\vartheta |l_n^T(0)\widehat{\boldsymbol{\Delta}}[l(\boldsymbol{\gamma}) - \mathbf{U}_1\boldsymbol{\gamma}/\sqrt{n}]|/[||\boldsymbol{\gamma}||(1 + ||\boldsymbol{\gamma}||)] \leq \sqrt{n}||l_n(0)||||\widehat{\boldsymbol{\Delta}}||o_p(1) = o_p(1)$. Finally, since $\sqrt{n}||l(\boldsymbol{\gamma})|| \leq ||\boldsymbol{\gamma}||[||\mathbf{U}_1|| + o(1)]$, $\sup_\vartheta B_5(\boldsymbol{\gamma}) \leq \sup_\vartheta n||l(\boldsymbol{\gamma})||^2||\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}||/[||\boldsymbol{\gamma}||(1 + ||\boldsymbol{\gamma}||)] \leq \sup_\vartheta ||\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}||O_p(1) = o_p(1)$.

□

Note that $T_n(\boldsymbol{\gamma})$ is minimized at $\widehat{\boldsymbol{\gamma}}_g = \sqrt{n}(\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0)$ under enveloping.

Lemma 4. *Under the same conditions in Theorem 2,* $||\widehat{\boldsymbol{\gamma}}_g|| = \sqrt{n}||\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0|| = O_p(1)$.

Proof. Let ϑ be the same defined as in Lemma 3. First, consider

$$\sup_\vartheta |\kappa_n(\boldsymbol{\gamma})| \leq O_p(1)\sup_\vartheta(||\varepsilon_n(\boldsymbol{\gamma})||^2 + 2||\varepsilon_n(\boldsymbol{\gamma})||||l_n(0)||) \leq o_p(n^{-1}). \tag{B.2}$$

Note that under the envelope setting, $T_n(\widehat{\boldsymbol{\gamma}}_g) \leq T_n(0)$ and by Lemma 2, $\widehat{\boldsymbol{\gamma}}_g \in \vartheta$. Hence $T_n(\widehat{\boldsymbol{\gamma}}_g) - \kappa_n(\boldsymbol{\gamma}) \leq T_n(\widehat{\boldsymbol{\gamma}}_g) + o_p(n^{-1}) \leq T_n(0) + o_p(n^{-1})$. Correspondingly, we have

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

By Taylor expansion, $T(\widehat{\boldsymbol{\gamma}}_g) = \widehat{\boldsymbol{\gamma}}_g^T\mathbf{H}\widehat{\boldsymbol{\gamma}}_g/(2n) + o(||\widehat{\boldsymbol{\gamma}}_g||^2/n)$, where $\mathbf{H} = n\left.\frac{\partial^2 T(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\boldsymbol{\gamma}^T}\right|_{\boldsymbol{\gamma}=0} = \left.\frac{\partial^2 Q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 2\mathbf{U}_1\boldsymbol{\Delta}\mathbf{U}_1 = 2\mathbf{U}_1\mathbf{V}_1^{-1}\mathbf{U}_1$. Since $\mathbf{H}$ is positive definite by (C2), there exists a constant $c > 0$, such that with probability approaching one, $T(\widehat{\boldsymbol{\gamma}}_g) \geq c||\widehat{\boldsymbol{\gamma}}_g||^2/n$. Therefore, by applying Lemma 3, we have $M \leq ||\widehat{\boldsymbol{\gamma}}_g||(1+||\widehat{\boldsymbol{\gamma}}_g||)o_p(1) + ||\widehat{\boldsymbol{\gamma}}_g||O_p(1) - c||\widehat{\boldsymbol{\gamma}}_g||^2 + o_p(1) = [-c + o_p(1)]||\widehat{\boldsymbol{\gamma}}_g||^2 + ||\widehat{\boldsymbol{\gamma}}_g||O_p(1) + o_p(1)$.

As $M \geq 0$ and $-c + o_p(1) < 0$ with probability approaching one, it follows that $||\widehat{\boldsymbol{\gamma}}_g||^2 - 2||\widehat{\boldsymbol{\gamma}}_g||O_p(1) \leq o_p(1)$. Hence $[||\widehat{\boldsymbol{\gamma}}_g|| - O_p(1)]^2 \leq O_p(1)$ and $|||\widehat{\boldsymbol{\gamma}}_g|| - O_p(1)| \leq O_p(1)$, indicating that $||\widehat{\boldsymbol{\gamma}}_g|| = O_p(1)$. □

To show Theorem 2, let $Z_n(\boldsymbol{\gamma}) = n[T_n(\boldsymbol{\gamma}) - T_n(0)]$. Under the envelope setting, $Z_n(\boldsymbol{\gamma})$ is minimized at $\widehat{\boldsymbol{\gamma}}_g$. Based on the results in Lemmas 3 and 4, and Taylor expansion, we see that

$$\begin{aligned} Z_n(\boldsymbol{\gamma}) &= \sqrt{n}\widehat{\mathbf{D}}^T\boldsymbol{\gamma} + nT(\boldsymbol{\gamma}) + o_p(1) = \sqrt{n}\widehat{\mathbf{D}}^T\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{H}\boldsymbol{\gamma} + o_p(1) \\ &\xrightarrow{d} \mathbf{N}^T\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{H}\boldsymbol{\gamma} =: Z(\boldsymbol{\gamma}), \end{aligned} \tag{B.3}$$

where $\mathbf{N} = N(0, 4\mathbf{U}_1\mathbf{V}_1^{-1}\mathbf{U}_1)$ because $\sqrt{n}\widehat{\mathbf{D}} = 2\sqrt{n}\mathbf{U}_1\widehat{\boldsymbol{\Delta}}l_n(0)$. Therefore, by Lemma 4 and the argmax theorem (Corollary 5.58 in Van der Vaart, 1998), we have $\widehat{\boldsymbol{\gamma}}_g \xrightarrow{d} \tilde{\boldsymbol{\gamma}}$, where

$$\tilde{\boldsymbol{\gamma}} = \underset{\frac{\boldsymbol{\gamma}}{\sqrt{n}}+\boldsymbol{\theta}_0=\psi_0(\boldsymbol{\zeta}'_\tau)}{\text{argmin}} Z(\boldsymbol{\gamma}) = \underset{\frac{\boldsymbol{\gamma}}{\sqrt{n}}+\boldsymbol{\theta}_0=\psi_0(\boldsymbol{\zeta}'_\tau)}{\text{argmin}} \frac{1}{2n}(\boldsymbol{\gamma} + \mathbf{H}^{-1}\mathbf{N})^T\mathbf{H}(\boldsymbol{\gamma} + \mathbf{H}^{-1}\mathbf{N}). \tag{B.4}$$

Since the parameter vector $\boldsymbol{\gamma}$ is overparameterized, we next apply Shapiro (1986) to establish the asymptotic distribution of $\tilde{\boldsymbol{\gamma}}$. We form the discrepancy function $F(\mathbf{x}, \boldsymbol{\xi})$ in Shapiro (1986) as

$$F(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{2}(\frac{\boldsymbol{\gamma}}{\sqrt{n}} + \frac{\mathbf{H}^{-1}\mathbf{N}}{\sqrt{n}})^T\mathbf{H}(\frac{\boldsymbol{\gamma}}{\sqrt{n}} + \frac{\mathbf{H}^{-1}\mathbf{N}}{\sqrt{n}}) \tag{B.5}$$

where $\mathbf{x}$ and $\boldsymbol{\xi}$ in our context represent $-\mathbf{H}^{-1}\mathbf{N}/\sqrt{n}$ and $\boldsymbol{\gamma}/\sqrt{n}$, respectively. It is easy to check that (B.5) satisfies Shapiro's assumptions 1-5 and $\frac{\partial^2 F}{\partial \boldsymbol{\xi}\boldsymbol{\xi}^T} = \mathbf{H} = 2\mathbf{U}_1\mathbf{V}_1^{-1}\mathbf{U}_1$. In addition, $-\mathbf{H}^{-1}\mathbf{N} \xrightarrow{d} N(0, \mathbf{U}_1^{-1}\mathbf{V}_1\mathbf{U}_1^{-1})$. Let $\boldsymbol{\Psi}_1 = \partial\psi_1(\boldsymbol{\zeta}_\tau^{'})/\partial\boldsymbol{\zeta}_\tau^{'T}$. Therefore, by applying Proposition 4.1 of Shapiro (1986), we have $\tilde{\boldsymbol{\gamma}} \xrightarrow{d} N(0, \boldsymbol{\Lambda}_g)$, where $\boldsymbol{\Lambda}_g =$

$$\boldsymbol{\Psi}_1(\boldsymbol{\Psi}_1^T\mathbf{H}\boldsymbol{\Psi}_1)^\dagger\boldsymbol{\Psi}_1^T\mathbf{H} \cdot \mathrm{avar}(-\mathbf{H}^{-1}\mathbf{N}) \cdot \mathbf{H}\boldsymbol{\Psi}_1(\boldsymbol{\Psi}_1^T\mathbf{H}\boldsymbol{\Psi}_1)^\dagger\boldsymbol{\Psi}_1^T = \boldsymbol{\Psi}_1(\boldsymbol{\Psi}_1^T\mathbf{U}_1\mathbf{V}_1^{-1}\mathbf{U}_1\boldsymbol{\Psi}_1)^\dagger\boldsymbol{\Psi}_1^T.$$

Hence $\widehat{\boldsymbol{\gamma}}_g = \sqrt{n}(\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{\Lambda}_g)$.

Finally, recall that $\widehat{\boldsymbol{\theta}}_g^*$ is the envelope GMM estimator of $\boldsymbol{\theta}^* = (\mu_\tau, \boldsymbol{\beta}_\tau^T, \mathrm{vech}(\boldsymbol{\Sigma}_\mathbf{X})^T)^T$ and is a sub vector of $\widehat{\boldsymbol{\theta}}_g$. Since

$$\boldsymbol{\Psi}_1 = \begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \mathbf{I}_p \end{pmatrix}, \quad \mathbf{U}_1 = \begin{pmatrix} \mathbf{U} & 0 \\ 0 & \mathbf{I}_p \end{pmatrix}, \quad \text{and} \quad \mathbf{V}_1 = \begin{pmatrix} \mathbf{V} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{V}_{1,33} \end{pmatrix},$$

where $\mathbf{A} = (0, \mathbf{V}_{1,23}^T)^T$, it can be verified that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_g^* - \boldsymbol{\theta}_0^*) \xrightarrow{d} N\Big(0,\ \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\mathbf{U}\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\Psi})^\dagger\boldsymbol{\Psi}^T\Big).$$

Note that $\boldsymbol{\Psi} = \partial\psi(\boldsymbol{\zeta}_\tau)/\partial\boldsymbol{\zeta}_\tau^T$. To give the expression of $\boldsymbol{\Psi}$, we introduce the contraction and expansion matrices that connect the 'vec' and 'vech' operators. For instance, for any symmetric matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$, $\mathrm{vech}(\mathbf{A}) = \mathbf{C}_m\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\mathbf{A}) = \mathbf{E}_m\mathrm{vech}(\mathbf{A})$, where $\mathbf{C}_m \in \mathbb{R}^{m(m+1)/2\times m^2}$ and $\mathbf{E}_m \in \mathbb{R}^{m^2\times m(m+1)/2}$ are the unique contraction and expansion matrices (Henderson and Searle, 1979). After some algebra, it can be shown that

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Phi}_\tau & \boldsymbol{\eta}_\tau^T \otimes \mathbf{I}_p & 0 & 0 \\ 0 & 0 & \boldsymbol{\Psi}_{33} & \mathbf{C}_p(\boldsymbol{\Phi}_\tau \otimes \boldsymbol{\Phi}_\tau)\mathbf{E}_{u_\tau} & \mathbf{C}_p(\boldsymbol{\Phi}_{0\tau} \otimes \boldsymbol{\Phi}_{0\tau})\mathbf{E}_{p-u_\tau} \end{pmatrix}, \tag{B.6}$$

where $\boldsymbol{\Psi}_{33} = 2\mathbf{C}_p(\boldsymbol{\Phi}_\tau\boldsymbol{\Omega}_\tau \otimes \mathbf{I}_p - \boldsymbol{\Phi}_\tau \otimes \boldsymbol{\Phi}_{0\tau}\boldsymbol{\Omega}_{0\tau}\boldsymbol{\Phi}_{0\tau}^T)$.

To show the asymptotic efficiency of $\widehat{\boldsymbol{\theta}}_g^*$ relative to $\tilde{\boldsymbol{\theta}}^*$, let $\boldsymbol{\Upsilon} = \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}$. Consider

$$\begin{aligned}\mathrm{avar}(\sqrt{n}\tilde{\boldsymbol{\theta}}^*) - \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\theta}}_g^*) &= \boldsymbol{\Upsilon} - \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Psi})^\dagger\boldsymbol{\Psi}^T = \boldsymbol{\Upsilon}^{1/2}\Big(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}}\Big)\boldsymbol{\Upsilon}^{1/2} \\ &= \boldsymbol{\Upsilon}^{1/2}\mathbf{Q}_{\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}}\boldsymbol{\Upsilon}^{1/2} \geq 0,\end{aligned}$$

where $\mathbf{P}_{\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}} = \boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}(\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Psi})^\dagger\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}^{-1/2}$ is the projection matrix onto the column span of $\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}$, and $\mathbf{Q}_{\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}}$ is the orthogonal projection matrix. Both of them are positive definite (or semidefinite). This completes the proof of Theorem 2.

Derivation of the asymptotic variance under i.i.d. errors

Under the i.i.d. error assumption, we have $\mathrm{E}_{\boldsymbol{\theta}_0}[f_{Y|\mathbf{X}}(\xi_0(\tau|\mathbf{X}))\mathbf{W}\mathbf{W}^T] = \mathrm{E}_{\boldsymbol{\theta}_0}[f(\xi(\tau))\mathbf{W}\mathbf{W}^T] = f(\xi(\tau))\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{W}\mathbf{W}^T)$. When $\mathrm{E}(\mathbf{X}) = 0$, it follows that

$$\mathrm{avar}(\sqrt{n}\tilde{\boldsymbol{\beta}}_\tau) = \tau(1-\tau)[f(\xi(\tau))\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{X}\mathbf{X}^T)]^{-1}\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{X}\mathbf{X}^T)][f(\xi(\tau))\mathrm{E}_{\boldsymbol{\theta}_0}(\mathbf{X}\mathbf{X}^T)]^{-1} = \frac{\tau(1-\tau)}{f^2(\xi(\tau))}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}.$$

Let $\mathbf{J} = \mathbf{U}\mathbf{V}^{-1}\mathbf{U}$. Next we note that the asymptotic variance of $\widehat{\boldsymbol{\theta}}_g^*$ depends only on the column space of $\boldsymbol{\Psi}$. We thus can replace $\boldsymbol{\Psi}$ with any matrix $\boldsymbol{\Psi}^*$ that has the same column space as $\boldsymbol{\Psi}$. Following Cook, Li, and Chiaromonte (2010), we choose the most convenient $\boldsymbol{\Psi}^* = \text{blockdiag}\big(1, \boldsymbol{\Psi}_1^*\big)$ that can make $\boldsymbol{\Psi}^{*^T}\mathbf{J}\boldsymbol{\Psi}^*$ block diagonal, where

$$\boldsymbol{\Psi}_1^* = \begin{pmatrix} \boldsymbol{\Phi}_\tau & \boldsymbol{\eta}_\tau^T \otimes \boldsymbol{\Phi}_{0\tau} & 0 & 0 \\ 0 & \boldsymbol{\Psi}_{33}^* & \mathbf{C}_p(\boldsymbol{\Phi}_\tau \otimes \boldsymbol{\Phi}_\tau)\mathbf{E}_{u_\tau} & \mathbf{C}_p(\boldsymbol{\Phi}_{0\tau} \otimes \boldsymbol{\Phi}_{0\tau})\mathbf{E}_{p-u_\tau} \end{pmatrix} \equiv \Big(\boldsymbol{\Psi}_{1,1}^* \, \boldsymbol{\Psi}_{1,2}^* \, \boldsymbol{\Psi}_{1,3}^* \, \boldsymbol{\Psi}_{1,4}^*\Big)$$

and $\boldsymbol{\Psi}_{33}^* = 2\mathbf{C}_p(\boldsymbol{\Phi}_\tau\boldsymbol{\Omega}_\tau \otimes \boldsymbol{\Phi}_{0\tau} - \boldsymbol{\Phi}_\tau \otimes \boldsymbol{\Phi}_{0\tau}\boldsymbol{\Omega}_{0\tau})$. With this $\boldsymbol{\Psi}^*$, we can also construct a block diagonal matrix $\boldsymbol{\Psi}^{**} = \text{blockdiag}\big(1, \boldsymbol{\Psi}_1^{**}\big)$, where

$$\boldsymbol{\Psi}_1^{**} = \begin{pmatrix} \mathbf{I}_{u_\tau} & \boldsymbol{\eta}_\tau^T \otimes \boldsymbol{\Phi}_\tau^T & 0 & 0 \\ 0 & \mathbf{I}_{u_\tau} \otimes \boldsymbol{\Phi}_{0\tau}^T & 0 & 0 \\ 0 & 2\mathbf{C}_{u_\tau}(\boldsymbol{\Omega}_\tau \otimes \boldsymbol{\Phi}_\tau^T) & \mathbf{I}_{u_\tau(u_\tau+1)/2} & 0 \\ 0 & 0 & 0 & \mathbf{I}_{(p-u_\tau)(p-u_\tau+1)/2,} \end{pmatrix}$$

such that $\boldsymbol{\Psi} = \boldsymbol{\Psi}^*\boldsymbol{\Psi}^{**}$ and $\boldsymbol{\Psi}^{**}$ has the full row rank. Accordingly, we can write $\mathbf{J} = \text{blockdiag}\big(\frac{f^2(\xi(\tau))}{\tau(1-\tau)}, \ \frac{f^2(\xi(\tau))}{\tau(1-\tau)}\boldsymbol{\Sigma}_\mathbf{X}, \ [\text{var}_{\boldsymbol{\theta}_0}\big\{\text{vech}[(\mathbf{X}-\boldsymbol{\mu}_0)(\mathbf{X}-\boldsymbol{\mu}_0)^T]\big)\}]^{-1} = \text{blockdiag}\big(\frac{f^2(\xi(\tau))}{\tau(1-\tau)}, \ \mathbf{J}_1\big)$. When $\mathbf{X}$ is normal, the moment estimator of $\boldsymbol{\Sigma}_\mathbf{X}$ obtained from the estimating equation (3.1) is asymptotically equivalent to the MLE of $\boldsymbol{\Sigma}_\mathbf{X}$ with the Fisher information $\mathbf{F}_{\boldsymbol{\Sigma}_\mathbf{X}} = \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_\mathbf{X}^{-1} \otimes \boldsymbol{\Sigma}_\mathbf{X}^{-1})\mathbf{E}_p$ (Cook *et al.*, 2013). Thus $\mathbf{J}_1 = \text{blockdiag}\Big(\frac{f^2(\xi(\tau))}{\tau(1-\tau)}\boldsymbol{\Sigma}_\mathbf{X}, \ \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_\mathbf{X}^{-1} \otimes \boldsymbol{\Sigma}_\mathbf{X}^{-1})\mathbf{E}_p\Big)$, where $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Phi}_\tau\boldsymbol{\Omega}_\tau\boldsymbol{\Phi}_\tau^T + \boldsymbol{\Phi}_{0\tau}\boldsymbol{\Omega}_{0\tau}\boldsymbol{\Phi}_{0\tau}^T$ under enveloping. Then after matrix multiplication, we see that $\boldsymbol{\Psi}_1^{*^T}\mathbf{J}_1\boldsymbol{\Psi}_1^*$ and $\boldsymbol{\Psi}^{*^T}\mathbf{J}\boldsymbol{\Psi}^*$ are block diagonal matrices, and $\boldsymbol{\Psi}_1^*(\boldsymbol{\Psi}_1^{*^T}\mathbf{J}_1\boldsymbol{\Psi}_1^*)^\dagger \boldsymbol{\Psi}_1^{*^T} = \sum_{j=1}^4 \boldsymbol{\Psi}_{1,j}^*(\boldsymbol{\Psi}_{1,j}^{*^T}\mathbf{J}_1\boldsymbol{\Psi}_{1,j}^*)^\dagger\boldsymbol{\Psi}_{1,j}^{*^T}$. Therefore, $\text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{g,\tau}) = \boldsymbol{\Phi}_\tau(\boldsymbol{\Psi}_{1,1}^{*^T}\mathbf{J}_1\boldsymbol{\Psi}_{1,1}^*)^\dagger\boldsymbol{\Phi}_\tau^T +$

$(\boldsymbol{\eta}_\tau^T \otimes \boldsymbol{\Phi}_{0\tau})(\boldsymbol{\Psi}_{1,2}^{*^T}\mathbf{J}_1\boldsymbol{\Psi}_{1,2}^*)^\dagger(\boldsymbol{\eta}_\tau \otimes \boldsymbol{\Phi}_{0\tau}^T)$. After matrix multiplication and utilizing Corollary D.1 and (S4.8) in the supplement of Cook, Li, and Chiaromonte (2010) , we have $\boldsymbol{\Psi}_{1,1}^{*^T}\mathbf{J}_1\boldsymbol{\Psi}_{1,1}^* = \frac{f^2(\xi(\tau))}{\tau(1-\tau)}\boldsymbol{\Omega}_\tau$ and $\boldsymbol{\Psi}_{1,2}^{*^T}\mathbf{J}_1\boldsymbol{\Psi}_{1,2}^* = \frac{f^2(\xi(\tau))}{\tau(1-\tau)}\boldsymbol{\eta}_\tau\boldsymbol{\eta}_\tau^T \otimes \boldsymbol{\Omega}_{0\tau} + \frac{1}{2}\boldsymbol{\Psi}_{33}^{*^T}\mathbf{E}_p^T(\boldsymbol{\Sigma}_\mathbf{X}^{-1} \otimes \boldsymbol{\Sigma}_\mathbf{X}^{-1})\mathbf{E}_p\boldsymbol{\Psi}_{33}^* = \frac{f^2(\xi(\tau))}{\tau(1-\tau)}\boldsymbol{\eta}_\tau\boldsymbol{\eta}_\tau^T \otimes \boldsymbol{\Omega}_{0\tau} + \boldsymbol{\Omega}_\tau \otimes \boldsymbol{\Omega}_{0\tau}^{-1} + \boldsymbol{\Omega}_\tau^{-1} \otimes \boldsymbol{\Omega}_{0\tau} - 2\mathbf{I}_{u_\tau} \otimes \mathbf{I}_{p-u_\tau}$. This completes the proof.

C Proof of Theorem 3

Following the proof of Theorem 1, it can be similarly shown that

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*) \xrightarrow{d} N\Big(0, \mathbf{U}_{pe}^{-1}\mathbf{V}_{pe}\mathbf{U}_{pe}^{-1}\Big),$$

where

$$\mathbf{U}_{pe} = \begin{pmatrix} \mathrm{E}_{\boldsymbol{\theta}_0}[f_{Y|\mathbf{X}}(\xi_0(\tau|\mathbf{X}))\mathbf{W}\mathbf{W}^T] & 0 \\ 0 & \mathbf{I}_{s_1} \end{pmatrix} \text{ and } \mathbf{V}_{pe} = \begin{pmatrix} \mathbf{V}_{1,11} & 0 \\ 0 & \mathbf{V}_{1,22} \end{pmatrix},$$

where $\mathbf{V}_{1,11} = \tau(1-\tau)\mathrm{E}_{\boldsymbol{\theta}_0}[\mathbf{W}\mathbf{W}^T]$ and $\mathbf{V}_{1,22} = \mathrm{var}_{\boldsymbol{\theta}_0}\{\mathrm{vech}[(\mathbf{X}_1 - \boldsymbol{\mu}_{\mathbf{X}_1,0})(\mathbf{X}_1 - \boldsymbol{\mu}_{\mathbf{X}_1,0})^T]\}$.

The rest of the proof is similar to the proof of Theorem 2 by first following the steps in Lemmas 2-4 to show $\sqrt{n}$-consistency of $\widehat{\boldsymbol{\theta}}_{pe}^*$. Then apply the argmax theorem and Shapiro's Proposition 4.1 to establish asymptotic normality and asymptotic efficiency. We thus omit the details. The gradient matrix $\mathbf{G}$ is given by

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Psi}_\tau & \boldsymbol{\eta}_\tau^T \otimes \mathbf{I}_{p_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{I}_{p_2} & 0 & 0 \\ 0 & 0 & \mathbf{G}_{43} & 0 & \mathbf{C}_{p_1}(\boldsymbol{\Psi}_\tau \otimes \boldsymbol{\Psi}_\tau)\mathbf{E}_{d_\tau} & \mathbf{C}_{p_1}(\boldsymbol{\Psi}_{0\tau} \otimes \boldsymbol{\Psi}_{0\tau})\mathbf{E}_{p_1 - d_\tau} \end{pmatrix},$$

where $\mathbf{G}_{43} = 2\mathbf{C}_{p_1}(\boldsymbol{\Psi}_\tau \boldsymbol{\Omega}_\tau \otimes \mathbf{I}_{p_1} - \boldsymbol{\Psi}_\tau \otimes \boldsymbol{\Psi}_{0\tau} \boldsymbol{\Omega}_{0\tau} \boldsymbol{\Psi}_{0\tau}^T)$.

D Additional simulations

D.1 Effects of estimation uncertainty on efficiency

When the immaterial variation is substantial and the envelope dimension is large, the EQR might still outperform the standard QR. To show this, we performed an additional simulation study, where the data generating model is similar as that in Section 5 of the manuscript but the envelope dimension is chosen to be much larger with $u = 8$ (close to the full dimension $p = 10$). In addition, we set $\boldsymbol{\Phi} = \left(\begin{smallmatrix} \mathbf{I}_6 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma} \end{smallmatrix}\right)$, where $\boldsymbol{\Gamma}$ was a 4×2 matrix with the first 2 rows being $(\sqrt{2}/2, 0)$ and the other 2 rows being $(0, \sqrt{2}/2)$. The first six elements of the vector $\boldsymbol{\eta}_1$ were 5 and the last two elements were $5\sqrt{2}$, and $\boldsymbol{\eta}_2$ was a vector with the first five elements being 0, the sixth element being 0.1 and the last two elements being $\sqrt{2}/10$. The matrix $\boldsymbol{\Omega}$ was a diagonal matrix with diagonal elements being $(10, 20, \ldots, 80)$ and $\boldsymbol{\Omega}_0 = \mathbf{I}_2$. The other terms were kept the same as in Section 5. In this example, the immaterial variation is relatively large as $||\boldsymbol{\Omega}_0^{-1}|| \gg ||\boldsymbol{\Omega}^{-1}||$. The following figures and table demonstrate the comparison results of the estimated standard deviations, MSEs and squared biases between EQR and QR for $\tau = 0.5$.

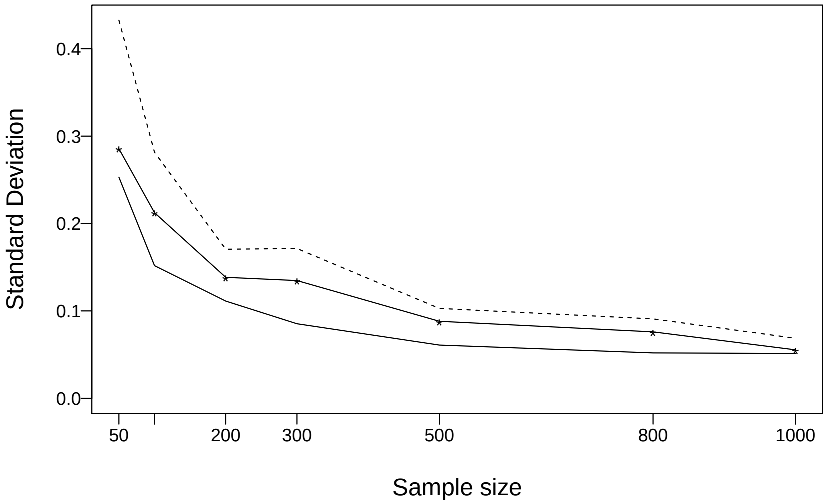

Figure D.1: Comparison of the estimated standard deviations. The line — marks the standard deviations of the EQR estimator with true u and the line – – marks the standard deviations of the standard QR estimator. The line with "*" marks the EQR estimator with u selected by RCV.

Table D.1: Comparison of the estimated standard deviations

n	QR	EQR(true u)	EQR(selected u)	ratio QR/EQR(selected u)
50	0.433	0.253	0.286	1.515
100	0.282	0.152	0.212	1.326
200	0.171	0.111	0.138	1.232
300	0.171	0.085	0.135	1.272
500	0.103	0.061	0.088	1.167
800	0.091	0.052	0.076	1.196
1000	0.069	0.051	0.055	1.239

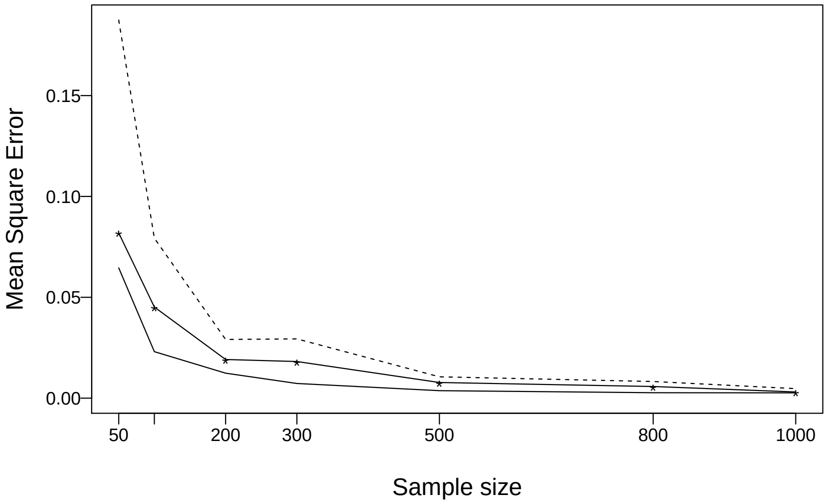

Figure D.2: Comparison of the MSEs. The line — marks the MSEs of the EQR estimator with true u and the line – – marks the MSEs of the standard QR estimator. The line with "*" marks the EQR estimator with u selected by RCV.

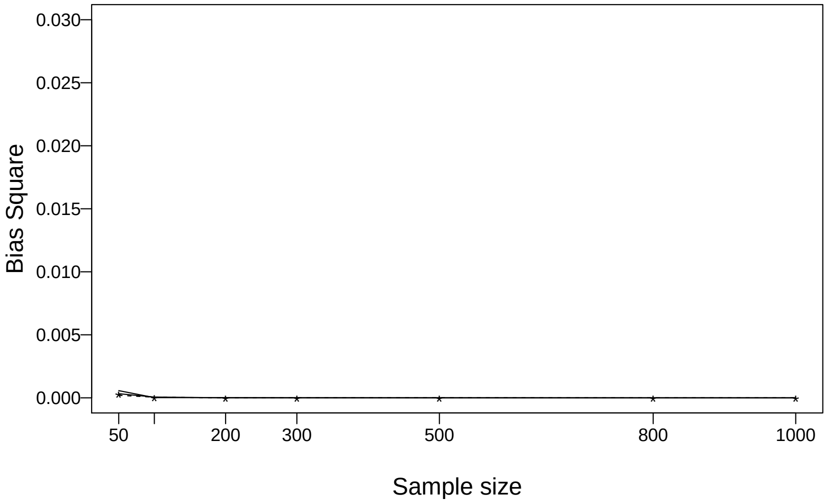

Figure D.3: Comparison of the squared biases. The line — marks the squared biases of the EQR estimator with true u and the line – – marks the squared biases of the standard QR estimator. The line with "*" marks the EQR estimator with u selected by RCV.

We see from Figures D.1-D.3 and Table D.1 that the EQR shows smaller estimated standard deviations and smaller MSEs than the standard QR under both the true and selected envelope dimensions, while it provides very similar estimation biases as QR. The EQR outperforms QR in this case.

When the immaterial variation is relatively small while the envelope dimension is large, the efficiency gains from enveloping might be inadequate to overcome the cost of uncertainty in estimating the envelope subspace and parameters, resulting in relatively close or worse performance of EQR compared to QR. The following simulation illustrates this possibility. Under the setting of the last simulation, we now set η_1 to be a vector with the first six elements being 1 and the last two elements being $\sqrt{2}$. The matrix $\mathbf{\Omega}$ was chosen to be a diagonal matrix with diagonal elements $(1, 2, \ldots, 8)$ and $\mathbf{\Omega}_0$ was $100\mathbf{I}_2$. The other parameters remain unchanged. In this case, the immaterial variation is relatively small as $||\mathbf{\Omega}_0^{-1}|| \ll ||\mathbf{\Omega}^{-1}||$. The comparison results are summarized in Figures D.4–D.6 and Table D.2 for $\tau = 0.5$.

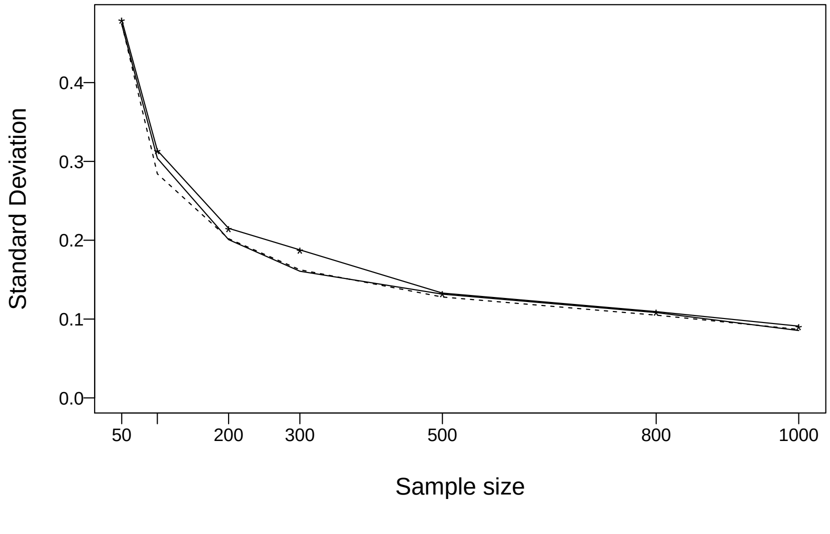

Figure D.4: Comparison of the estimated standard deviations. The line — marks the standard deviations of the EQR estimator with true u and the line – – marks the standard deviations of the standard QR estimator. The line with "*" marks the EQR estimator with u selected by RCV.

Table D.2: Comparison of the estimated standard deviations

n	QR	EQR(true u)	EQR(selected u)	ratio QR/EQR(selected u)
50	0.475	0.474	0.480	0.990
100	0.284	0.304	0.314	0.905
200	0.202	0.201	0.215	0.939
300	0.163	0.161	0.188	0.866
500	0.128	0.132	0.133	0.962
800	0.105	0.108	0.109	0.959
1000	0.087	0.085	0.091	0.954

It can be seen that the EQR shows slightly larger estimated standard deviations and MSEs than QR in this case, especially under the selected envelope dimensions, while the estimation biases from EQR and QR are relatively close.

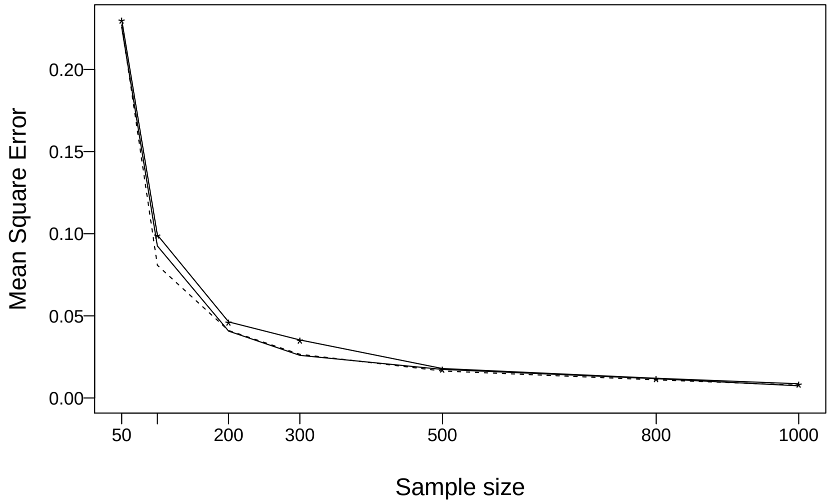

Figure D.5: Comparison of the MSEs. The line — marks the MSEs of the EQR estimator with true u and the line – – marks the MSEs of the standard QR estimator. The line with "*" marks the EQR estimator with u selected by RCV.

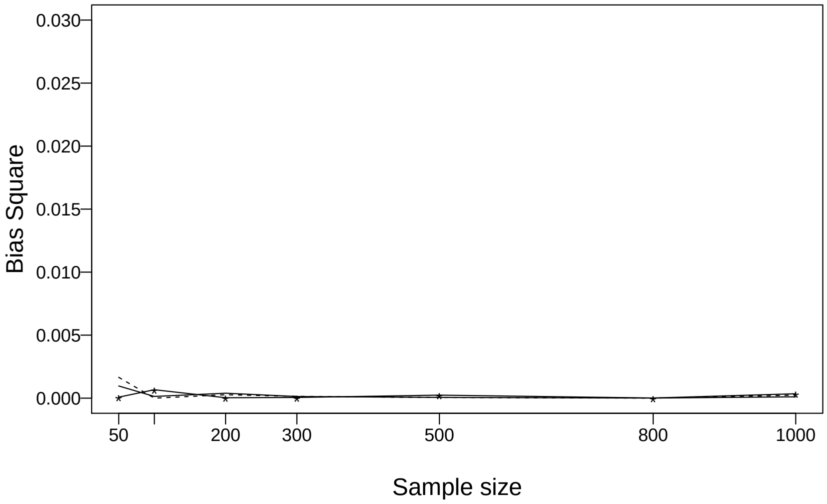

Figure D.6: Comparison of the squared biases. The line — marks the squared biases of the EQR estimator with true u and the line – – marks the squared biases of the standard QR estimator. The line with "*" marks the EQR estimator with u selected by RCV.

D.2 Performance under the violation of the envelope assumption

We report results of a simulation study that illustrates the performance of the EQR estimator when the parameters do not have the envelope structure. We generated the data from the following model

$$Y_i = \mu + \boldsymbol{\alpha}^T \mathbf{X}_i + (5 + \boldsymbol{\gamma}^T \mathbf{X}_i)\epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\mu = 5$, ϵ followed the standard normal distribution with distribution function F_ϵ, the elements in $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ were independent standard normal variates, and $\mathbf{X}$ followed a multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$. The covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$ had the structure $\mathbf{M}\mathbf{D}\mathbf{M}^T$, where $\mathbf{D}$ was a diagonal matrix with diagonal elements $1, \ldots, p$ and elements in $\mathbf{M}$ were independent standard normal variates. Under this setting, $\mu_\tau = \mu + 5F_\epsilon^{-1}(\tau)$, $\boldsymbol{\beta}_\tau = \boldsymbol{\alpha} + \boldsymbol{\gamma} F_\epsilon^{-1}(\tau)$, and $\boldsymbol{\beta}_\tau$ and $\boldsymbol{\Sigma}_\mathbf{X}$ do not follow the envelope structure in (2.5). We set $p = 10$, $n = 50$, and generated 200 replications. For each replication, we computed the EQR estimator of $\boldsymbol{\beta}_\tau$ with $u_\tau = 1, \ldots, 10$. Then the estimation variance, bias and MSE were calculated for each EQR estimator. Note that when $u_\tau = 10$, the EQR estimator reduces to the standard QR estimator. Results for a randomly chosen element in $\boldsymbol{\beta}_\tau$ ($\tau = 0.5$) are displayed in Table D.3.

Table D.3: Estimation variance, bias and MSE for envelope estimators with different u_τ.

u_τ	1	2	3	4	5	6	7	8	9	10
Est. Var.	0.07	0.15	0.19	0.30	0.37	0.37	0.36	0.36	0.41	0.40
Bias	0.60	0.47	0.34	0.24	0.11	0.09	0.09	0.07	0.09	0.05
MSE	0.42	0.38	0.30	0.35	0.38	0.38	0.37	0.37	0.42	0.40

We notice that when u_τ increases, the bias decreases and the estimation variance increases. The MSE reaches its minimum 0.30 when $u_\tau = 3$, which is smaller than the MSE of the QR estimator 0.40. This indicates that when the envelope structure does not hold, it is still worthwhile to compute the EQR estimator because it may have a smaller MSE.

D.3 Results on estimation bias for examples in Section 5

Figures D.7 and D.8 provide the comparison results on the squared estimation biases between EQR and QR estimators under the same settings as Figures 3–6 of the manuscript. The estimation biases of the two methods are generally close except that QR has relatively large bias when $\tau = 0.9$ and the sample size is relatively small.

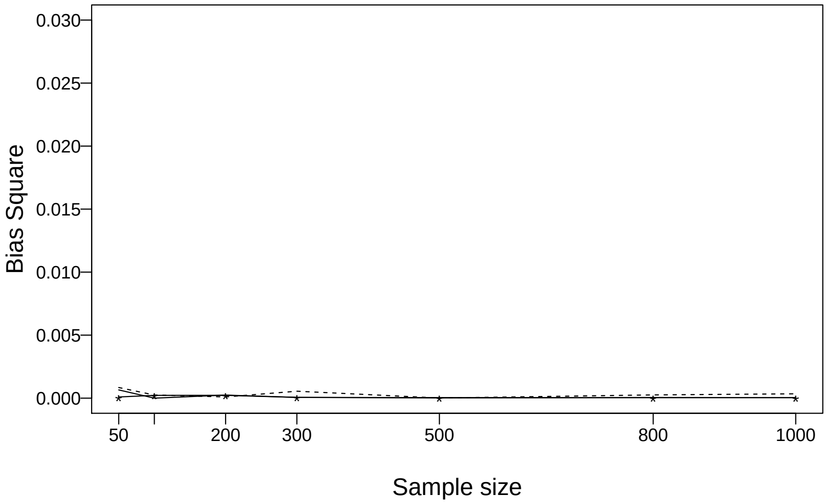

Figure D.7: Comparison of the squared biases ($\tau = 0.5$). The line — marks the squared biases of the EQR estimator with true u and the line – – marks the squared biases of the standard QR estimator. The line with "*" marks the EQR estimator with u selected by RCV.

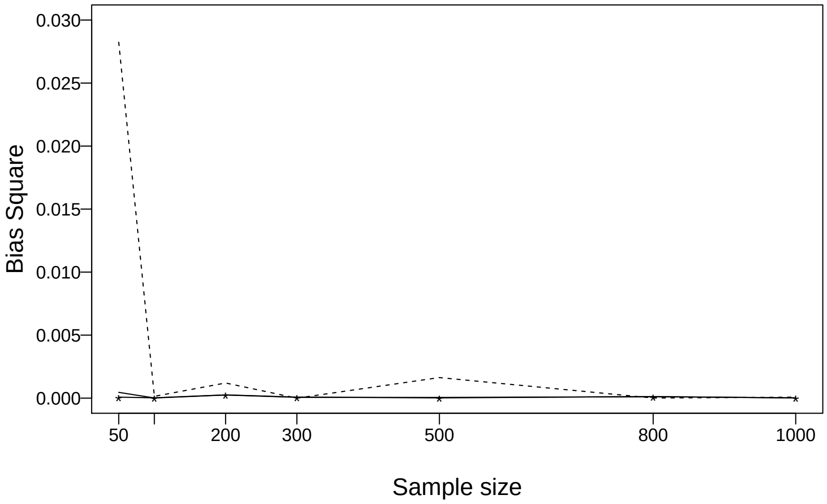

Figure D.8: Comparison of the squared biases ($\tau = 0.9$). The line — marks the squared biases of the EQR estimator with true u and the line – – marks the squared biases of the standard QR estimator. The line with "*" marks the EQR estimator with u selected by RCV.

D.4 Simulation for PEQR

We generated the data from the model

$$Y_i = \mu + \boldsymbol{\alpha}^T \mathbf{X}_{1,i} + \boldsymbol{\beta}_2^T \mathbf{X}_{2,i} + (5 + \boldsymbol{\gamma}^T \mathbf{X}_{1,i})\epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\boldsymbol{\alpha} = \boldsymbol{\Psi}_1 \boldsymbol{\eta}_1$, $\boldsymbol{\gamma} = \boldsymbol{\Psi}_1 \boldsymbol{\eta}_2$, and ϵ follows a standard normal distribution with distribution function F_ϵ. Here $\boldsymbol{\Psi}_1 \in \mathbb{R}^{p_1 \times d}(d < p_1)$ is a semi-orthogonal matrix. Hence $\mu_\tau = \mu + 5F_\epsilon^{-1}(\tau)$, $\boldsymbol{\beta}_{1,\tau} = \boldsymbol{\Psi}_1(\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2 F_\epsilon^{-1}(\tau)) = \boldsymbol{\Psi}_1 \boldsymbol{\eta}_\tau$, $\boldsymbol{\beta}_{2,\tau} = \boldsymbol{\beta}_2$, $\boldsymbol{\Psi}_\tau = \boldsymbol{\Psi}_1$ and $d_\tau = d$, for $0 < \tau < 1$. We set $p_1 = 8$, $p_2 = 2$, $d = 2$ and varied the sample size n from 50 to 1000. We generated $\mathbf{X}_1$ from a multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{\Psi}_1 \boldsymbol{\Omega}_1 \boldsymbol{\Psi}_1^T + \boldsymbol{\Psi}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Psi}_0^T$, where $\boldsymbol{\Psi}_0$ was a completion of $\boldsymbol{\Psi}_1$, and $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_0$ were the corresponding coordinate matrices. We generated $\boldsymbol{\Psi}_1$ with the first $p_1/2$ rows being $(-0.5, 0)$ and the remaining rows being $(0, -0.5)$. The matrix $\boldsymbol{\Omega}_1$ was a diagonal matrix with diagonal elements 25 and 100, $\boldsymbol{\Omega}_0$ was an identity matrix, $\boldsymbol{\eta}_1$ was $(-10, -10)^T$, $\boldsymbol{\eta}_2$ was $(0, -2/\sqrt{80})^T$, and μ was 5. The elements in $\mathbf{X}_2$ were independent Bernoulli random variables with the success probability 0.5, and $\boldsymbol{\beta}_{2,\tau}$ was $(2, 2)^T$. For each sample size, we generated 200 replications, and fit the standard QR model and the PEQR model to each replication. The estimation standard deviation for each element in $\boldsymbol{\beta}_{1,\tau}$ and $\boldsymbol{\beta}_{2,\tau}$ was calculated for both the standard QR estimator and the PEQR estimator. We also generated 200 repetitions from paired bootstrap, and computed the bootstrap standard deviations. The results of two randomly chosen elements in $\boldsymbol{\beta}_{1,\tau}$ and $\boldsymbol{\beta}_{2,\tau}$ for $\tau = 0.5$ are displayed in Figure D.9 and Figure D.10. The efficiency gains for $\boldsymbol{\beta}_{1,\tau}$ are substantial. Across all elements in $\boldsymbol{\beta}_{1,\tau}$, at sample size 1000, the PEQR estimator reduced the estimation standard deviations by 45.1% to 55.9%. The efficiency gains for elements in $\boldsymbol{\beta}_{2,\tau}$ are not very obvious, although we observe that the PEQR estimator is slightly more efficient than the standard QR estimator in Figure D.10.

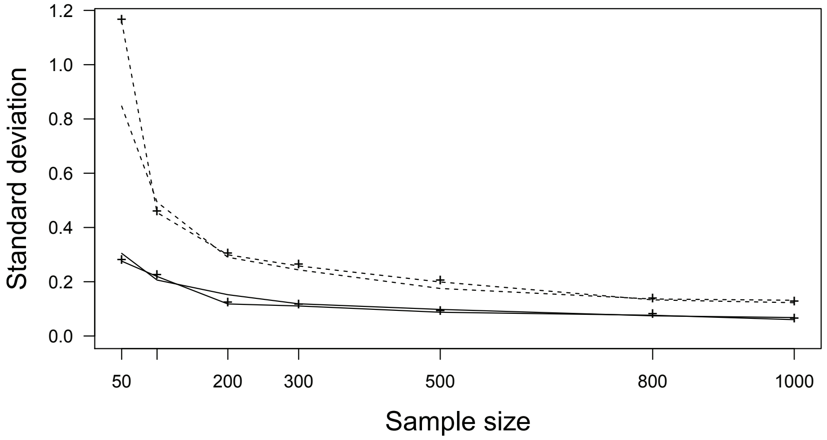

Figure D.9: Comparison of the estimation standard error for an element in $\boldsymbol{\beta}_1$. Lines — mark the PEQR estimator and lines – – mark the standard QR estimator. The lines with "+" mark the bootstrap standard deviations for the corresponding estimators.

We have also examined the performance of RCV on the envelope dimension selection under PEQR for $\tau = 0.5$. Not tabulated here, RCV was stable across all sample sizes, and the fraction that it selected the true d_τ was 94.5% at sample size 50 and gradually increased to 100% at sample size 1000. We repeated this simulation using the selected d_τ to incorporate the model selection variability, as we did in Section 5. Similar results are observed as in Figures 3, 4, 5 and 6: some efficiency is lost compared to the PEQR estimator with the true d_τ, but the PEQR estimator with selected d_τ is still more efficient than the standard QR estimator. In addition, the MSE of the PEQR estimator with

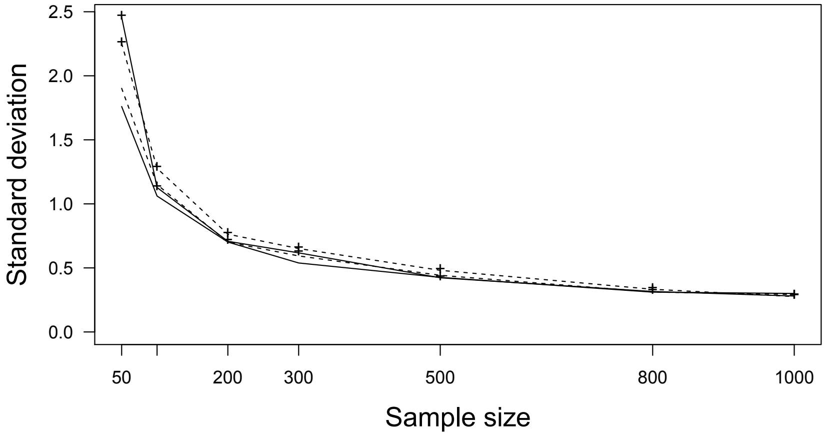

Figure D.10: Comparison of the estimation standard error for an element in $\boldsymbol{\beta}_2$. Lines — mark the PEQR estimator and lines – – mark the standard QR estimator. The lines with "+" mark the bootstrap standard deviations for the corresponding estimators.

selected d_τ is much smaller than the standard QR estimator.

E More information on Nelder-Mead method

Nelder-Mead method or downhill simplex method (Nelder and Mead, 1965) is applied to find the minima of the objective function. Specifically, suppose that the objective function has k variables, the Nelder-Mead method begins with a set of $k+1$ test points arranged as a simplex. It evaluates the objective function on these test points, and orders the values. Based on the values, it finds the worst point as well as the centroid of the other k points. The Nelder-Mead method then performs a series of transformations in order to find a new test point to replace the worst test point, aiming to decrease the value of the objective function at the test points. Most common transformations includes reflection (computes the reflection point of the worst point with respect to the centroid), expansion (the same as reflection, but the point is two times as far to the centroid than the reflection point), and contraction (the middle point between the worst point and the centroid). A candidate point is accepted as a new test point if it is better than the worst point and satisfies some other conditions. If none of the candidate points is accepted, we shrink the simplex towards the best point in the previous step. This procedure is repeated until the volume of the simplex is small enough (Singer and Singer, 1999).

REFERENCES

Andrews, D. W. (1994). Empirical process methods in econometrics. *Handbook of Econometrics*, **4**, 2247–2294.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica: Journal of the Econometric Society*, **71**(5), 1591–1608.

Cook, R., Helland, I., and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B*, **75**(5), 851–877.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, **20**, 927–1010.

Henderson, H. V. and Searle, S. (1979). Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *Canadian Journal of Statistics*, **7**(1), 65–81.

Koenker, R. (2005). *Quantile regression*. Cambridge university press.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, **7**(4), 308–313.

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

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

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

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*, pages 185–196. PMF–Matematički odjel, Zagreb.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.