

Supplementary Materials

EC.1. Proof of Theorem 1 (Convergence in Distribution)

EC.1.1. Proof of Lemma 3

For $i = 1, 5$,

$$\mathbb{E} \left[\left| h \left(\hat{Q}(U|\mathbf{x}^*) \right) - h \left(F_Y^{-1}(U|\mathbf{x}^*) \right) \right| \cdot \mathbb{I}\{U \in I_i\} \right] \leq 2M_h \cdot \Pr(U \in I_i) = \frac{2M_h \cdot \epsilon}{10 \max(M_h, 1)} \leq \frac{\epsilon}{5}.$$

For $i = 2, 4$,

$$\mathbb{E} \left[\left| h \left(\hat{Q}(U|\mathbf{x}^*) \right) - h \left(F_Y^{-1}(U|\mathbf{x}^*) \right) \right| \cdot \mathbb{I}\{U \in I_i\} \right] \leq 2M_h \cdot \Pr(U \in I_i) = \frac{2M_h}{m} \leq \frac{\epsilon}{5},$$

provided that $m \geq 10 \max(M_h, 1)/\epsilon$. $\square$

EC.1.2. Proof of Lemma 4

Note that

$$\left| h \left(\hat{Q}(U|\mathbf{x}^*) \right) - h \left(F_Y^{-1}(U|\mathbf{x}^*) \right) \right| \cdot \mathbb{I}\{U \in I_3\} \leq L_h \cdot \sup_{\tau \in I_3} \left| \hat{Q}(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*) \right|. \quad (\text{EC.1.0})$$

By Proposition 2, as $n, m \rightarrow \infty$,

$$\sup_{\tau \in I_3} \left| \hat{Q}(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*) \right| \xrightarrow{\mathbb{P}} 0.$$

Therefore,

$$\left| h \left(\hat{Q}(U|\mathbf{x}^*) \right) - h \left(F_Y^{-1}(U|\mathbf{x}^*) \right) \right| \cdot \mathbb{I}\{U \in I_3\} \xrightarrow{\mathbb{P}} 0. \quad (\text{EC.1.1})$$

Moreover, because h is bounded,

$$\left| h \left(\hat{Q}(U|\mathbf{x}^*) \right) - h \left(F_Y^{-1}(U|\mathbf{x}^*) \right) \right| \cdot \mathbb{I}\{U \in I_3\} \leq 2M_h. \quad (\text{EC.1.2})$$

It then follows from (EC.1.1), (EC.1.2), and the dominated convergence theorem that

$$\mathbb{E} \left[\left| h \left(\hat{Q}(U|\mathbf{x}^*) \right) - h \left(F_Y^{-1}(U|\mathbf{x}^*) \right) \right| \cdot \mathbb{I}\{U \in I_3\} \right] \rightarrow 0$$

as $m, n \rightarrow \infty$. Therefore, for a given $\epsilon > 0$, there exist sufficiently large M and N , such that

$$\mathbb{E} \left[\left| h \left(\hat{Q}(U|\mathbf{x}^*) \right) - h \left(F_Y^{-1}(U|\mathbf{x}^*) \right) \right| \cdot \mathbb{I}\{U \in I_3\} \right] \leq \frac{\epsilon}{5},$$

for all $m \geq M, n \geq N$. $\square$

EC.1.3. Proof of Proposition 2

The notation $\tilde{\beta}_j$ is hereinafter employed to represent $\tilde{\beta}(\tau_j)$ for the sake of simplicity, while $\hat{\beta}(\tau)$ denotes QRGMM's estimation of $\beta(\tau)$ through interpolation for any given value of τ . Notice that for any given subinterval $[\tau_L, \tau_U] \subset (0, 1)$ with $0 < \tau_L < \tau_U < 1$ and any given m , since τ_L and τ_U are not required to be symmetric, there are three cases for the relation between $[\tau_L, \tau_U]$ and $[\frac{1}{m}, 1 - \frac{1}{m}]$, that is, $[\frac{1}{m}, 1 - \frac{1}{m}] \subset [\tau_L, \tau_U]$, $[\tau_L, \tau_U] \subseteq [\frac{1}{m}, 1 - \frac{1}{m}]$, and $[\tau_L, \tau_U]$ intersects with $[\frac{1}{m}, 1 - \frac{1}{m}]$. We prove Proposition 2 for each of these three cases, respectively.

Case 1: $[\frac{1}{m}, 1 - \frac{1}{m}] \subset [\tau_L, \tau_U]$. For $\tau \in [\tau_L, \tau_1]$, we have $\hat{\beta}(\tau) = \tilde{\beta}_1$. Write

$$\hat{\beta}(\tau) - \beta(\tau) = \tilde{\beta}_1 - \beta(\tau) = \tilde{\beta}_1 - \beta(\tau_1) + \beta(\tau_1) - \beta(\tau).$$

For $\tau \in [\tau_j, \tau_{j+1})$, denote $\frac{\tau_{j+1} - \tau}{\tau_{j+1} - \tau_j}$ as α , then we have $\hat{\beta}(\tau) = \alpha \tilde{\beta}(\tau_j) + (1 - \alpha) \tilde{\beta}_{j+1}$. Write

$$\begin{aligned} \hat{\beta}(\tau) - \beta(\tau) &= \alpha (\tilde{\beta}_j - \beta(\tau_j)) + (1 - \alpha) (\tilde{\beta}_{j+1} - \beta(\tau_{j+1})) + \alpha (\beta(\tau_j) - \beta(\tau)) + (1 - \alpha) (\beta(\tau_{j+1}) - \beta(\tau)). \end{aligned} \quad (\text{EC.1.3})$$

For $\tau \in [\tau_{m-1}, \tau_U]$, we have $\hat{\beta}(\tau) = \tilde{\beta}_{m-1}$. Write

$$\hat{\beta}(\tau) - \beta(\tau) = \tilde{\beta}_{m-1} - \beta(\tau) = \tilde{\beta}_{m-1} - \beta(\tau_{m-1}) + \beta(\tau_{m-1}) - \beta(\tau).$$

Thus, recall that $F_Y^{-1}(\tau|\mathbf{x}^*) = \beta(\tau)^\top \mathbf{x}^*$, we then have

$$\begin{aligned} & \sup_{\tau \in [\tau_L, \tau_U]} |\hat{Q}(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*)| \\ & \leq \max_{\substack{\tau_j \in [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} \|\tilde{\beta}_j - \beta(\tau_j)\| \cdot \|\mathbf{x}^*\| + \sup \left\{ \sup_{\tau \in [\tau_L, \tau_1]} |F_Y^{-1}(\tau_1|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*)|, \right. \\ & \quad \left. \sup_{\substack{\tau \in [\tau_{j-1}, \tau_{j+1}) \\ 2 \leq j \leq m-2}} |F_Y^{-1}(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau_j|\mathbf{x}^*)|, \sup_{\tau \in [\tau_{m-1}, \tau_U]} |F_Y^{-1}(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau_{m-1}|\mathbf{x}^*)| \right\} \\ & \leq \max_{\substack{\tau_j \in [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} \|\tilde{\beta}_j - \beta(\tau_j)\| \cdot \|\mathbf{x}^*\| + \sup \left\{ |F_Y^{-1}(\tau_1|\mathbf{x}^*) - F_Y^{-1}(\tau_L|\mathbf{x}^*)|, \right. \\ & \quad \left. \sup_{1 \leq j \leq m-2} |F_Y^{-1}(\tau_j|\mathbf{x}^*) - F_Y^{-1}(\tau_{j+1}|\mathbf{x}^*)|, |F_Y^{-1}(\tau_U|\mathbf{x}^*) - F_Y^{-1}(\tau_{m-1}|\mathbf{x}^*)| \right\} \\ & \leq \max_{\substack{\tau_j \in [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} \|\tilde{\beta}_j - \beta(\tau_j)\| \cdot \|\mathbf{x}^*\| + \frac{1}{m} \left(\inf_{\tau \in [\tau_L, \tau_U]} f_Y(F_Y^{-1}(\tau|\mathbf{x}^*)|\mathbf{x}^*) \right)^{-1}, \end{aligned}$$

where the last inequality is due to Assumption (2.b), the mean value theorem, and for any $\tau \in [\tau_L, \tau_U]$,

$$\left(F_Y^{-1}(\tau|\mathbf{x}^*) \right)' = \left(f_Y \left(F_Y^{-1}(\tau|\mathbf{x}^*) \right) \right)^{-1} \leq \left(\inf_{\tau \in [\tau_L, \tau_U]} f_Y \left(F_Y^{-1}(\tau|\mathbf{x}^*) \right) \right)^{-1}.$$

Case 2: $[\tau_L, \tau_U] \subseteq [\frac{1}{m}, 1 - \frac{1}{m}]$. Compared with Case 1, this case is more tricky due to the problem caused by the two-side small intervals near the two endpoints τ_L and τ_U , (e.g. $[\tau_j, \tau_U]$, if $\tau_U \in [\tau_j, \tau_{j+1}]$ for some j), where $\hat{\beta}(\tau)$ are interpolated by $\tilde{\beta}_j$ and $\tilde{\beta}_{j+1}$, however, $\tilde{\beta}_{j+1}$ is outside $[\tau_L, \tau_U]$, we can't ensure its uniform convergence by Proposition 1. To solve this problem, we compress each end of the interval $[\tau_L, \tau_U]$ inward by $\frac{1}{m}$, that is, we only consider the interval $[\tau_L + \frac{1}{m}, \tau_U - \frac{1}{m}]$, and it will converge to (τ_L, τ_U) as $m \rightarrow \infty$. Then, notice that for any $\tau \in [\tau_L + \frac{1}{m}, \tau_U - \frac{1}{m}]$, we can now always find $[\tau_j, \tau_{j+1})$ subject to $\tau \in [\tau_j, \tau_{j+1}) \subseteq [\tau_L, \tau_U]$. By equation (EC.1.4), we have

$$\begin{aligned}
& \sup_{\tau \in [\tau_L + \frac{1}{m}, \tau_U - \frac{1}{m}]} |\hat{Q}(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*)| \\
& \leq \max_{\substack{\tau_j \in [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} \|\tilde{\beta}_j - \beta(\tau_j)\| \cdot \|\mathbf{x}^*\| + \sup \left\{ \sup_{\substack{\tau \in [\tau_j, \tau_{j+1}) \subseteq [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} |F_Y^{-1}(\tau_j|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*)|, \right. \\
& \quad \left. \sup_{\substack{\tau \in [\tau_j, \tau_{j+1}) \subseteq [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} |F_Y^{-1}(\tau_{j+1}|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*)| \right\} \\
& \leq \max_{\substack{\tau_j \in [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} \|\tilde{\beta}_j - \beta(\tau_j)\| \cdot \|\mathbf{x}^*\| + \sup_{\substack{\tau_j \in [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} |F_Y^{-1}(\tau_{j+1}|\mathbf{x}^*) - F_Y^{-1}(\tau_j|\mathbf{x}^*)| \\
& \leq \max_{\substack{\tau_j \in [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} \|\tilde{\beta}_j - \beta(\tau_j)\| \cdot \|\mathbf{x}^*\| + \frac{1}{m} \left(\inf_{\tau \in [\tau_L, \tau_U]} f_Y(F_Y^{-1}(\tau|\mathbf{x}^*)|\mathbf{x}^*) \right)^{-1}.
\end{aligned}$$

Case 3: $[\tau_L, \tau_U]$ intersects with $[\frac{1}{m}, 1 - \frac{1}{m}]$. This case is just a combination of Case 1 and Case 2. Following a similar proof, we can still have

$$\begin{aligned}
& \sup_{\tau \in [\tau_L + \frac{1}{m}, \tau_U - \frac{1}{m}]} |\hat{Q}(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*)| \\
& \leq \max_{\substack{\tau_j \in [\tau_L, \tau_U] \\ 1 \leq j \leq m-1}} \|\tilde{\beta}_j - \beta(\tau_j)\| \cdot \|\mathbf{x}^*\| + \frac{1}{m} \left(\inf_{\tau \in [\tau_L, \tau_U]} f_Y(F_Y^{-1}(\tau|\mathbf{x}^*)|\mathbf{x}^*) \right)^{-1}. \tag{EC.1.4}
\end{aligned}$$

Thus, whichever case it is, we all have (EC.1.4). Hence, as $n, m \rightarrow \infty$, by Proposition 1, we have

$$\sup_{\tau \in (\tau_L, \tau_U)} |\hat{Y}(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*)| \xrightarrow{\mathbb{P}} 0. \quad \square$$

EC.1.4. Proof of Theorem 1

We prove the convergence in distribution of $Y(\mathbf{x}^*)$ by Lemma 2. Suppose h is any given bounded, Lipschitz continuous function on $\mathcal{Y}$ such that there exist positive constants M_h and L_h that, $|h(y)| \leq M_h$, for all $y \in \mathcal{Y}$, and $|h(y) - h(y')| \leq L_h|y - y'|$, for all y and y' . Then, it is sufficient to show that

$$\mathbb{E}h(\hat{Q}(U|\mathbf{x}^*)) \rightarrow \mathbb{E}h(F_Y^{-1}(U|\mathbf{x}^*)),$$

as $n, m \rightarrow \infty$, where the distribution of $F_Y^{-1}(U|\mathbf{x}^*)$ given $\mathbf{x}^*$ is the conditional distribution $F_Y(\cdot|\mathbf{x}^*)$.

For any given $\epsilon_1 > 0$, by Lemma 4, given $\tau_L = \frac{\epsilon_1}{10(M_h \vee 1)}$, $\tau_U = 1 - \frac{\epsilon_1}{10(M_h \vee 1)}$, for any $\epsilon_2 > 0$, there exist sufficiently large $N(\epsilon_1, \epsilon_2)$, $M(\epsilon_1, \epsilon_2)$, such that for any $n \geq N(\epsilon_1, \epsilon_2)$, $m \geq M(\epsilon_1, \epsilon_2)$, we have

$$\mathbb{E} \left[\left| \left(h(\hat{Q}(U|\mathbf{x}^*)) - h(F_Y^{-1}(U|\mathbf{x}^*)) \right) \right| \mathbb{I}\{U \in I_3\} \right] \leq \frac{\epsilon_2}{5}. \quad (\text{EC.1.5})$$

Then, take sufficiently large $M'(\epsilon_1, \epsilon_2) = M(\epsilon_1, \epsilon_2) \vee \frac{10((M_h \vee 1))}{\epsilon_1}$ such that for any $n \geq N(\epsilon_1, \epsilon_2)$, $m \geq M'(\epsilon_1, \epsilon_2)$, we have that $\frac{1}{m} \leq \frac{\epsilon_1}{10(M_h \vee 1)}$ and (EC.1.5). Notice that by the analysis of Proposition 2, (EC.1.5) keeps valid no matter what relationship between $[\tau_L, \tau_U]$ and $[\frac{1}{m}, 1 - \frac{1}{m}]$ is.

Since (EC.1.5) is satisfied for any $\epsilon_1, \epsilon_2 > 0$, let $\epsilon_2 = \epsilon_1$, then we have the following statement: for any given bounded Lipschitz continuous function h , any $\epsilon_1 > 0$, $\tau_L = \frac{\epsilon_1}{10(M_h \vee 1)}$, and $\tau_U = 1 - \frac{\epsilon_1}{10(M_h \vee 1)}$, there exist sufficiently large $N = N(\epsilon_1, \epsilon_1)$, $M' = M'(\epsilon_1, \epsilon_1)$, such that for any $n \geq N$, $m \geq M'$, we have that $\frac{1}{m} \leq \frac{\epsilon_1}{10(M_h \vee 1)}$ and (EC.1.5). Hence, it is straightforward to see

$$\begin{aligned} & |\mathbb{E}h(\hat{Q}(U|\mathbf{x}^*)) - \mathbb{E}h(F_Y^{-1}(U|\mathbf{x}^*))| \\ & \leq \mathbb{E} \left[\left| \left(h(\hat{Q}(U|\mathbf{x}^*)) - h(F_Y^{-1}(U|\mathbf{x}^*)) \right) \right| \mathbb{I}\{U \in I_1\} \right] + \mathbb{E} \left[\left| \left(h(\hat{Q}(U|\mathbf{x}^*)) - h(F_Y^{-1}(U|\mathbf{x}^*)) \right) \right| \mathbb{I}\{U \in I_2\} \right] \\ & + \mathbb{E} \left[\left| \left(h(\hat{Q}(U|\mathbf{x}^*)) - h(F_Y^{-1}(U|\mathbf{x}^*)) \right) \right| \mathbb{I}\{U \in I_3\} \right] + \mathbb{E} \left[\left| \left(h(\hat{Q}(U|\mathbf{x}^*)) - h(F_Y^{-1}(U|\mathbf{x}^*)) \right) \right| \mathbb{I}\{U \in I_4\} \right] \\ & + \mathbb{E} \left[\left| \left(h(\hat{Q}(U|\mathbf{x}^*)) - h(F_Y^{-1}(U|\mathbf{x}^*)) \right) \right| \mathbb{I}\{U \in I_5\} \right] \\ & \leq \frac{8M_h \cdot \epsilon_1}{10(M_h \vee 1)} + \frac{\epsilon_1}{5} \leq \epsilon_1, \end{aligned}$$

which completes the proof of Theorem (1.a). Theorem (1.b) can be readily derived from Theorem (1.a) and Assumption (2.a), as the convergence in distribution and the continuity of the limit imply uniform convergence of the CDF (Van der Vaart 2000, Lemma 2.11). $\square$

EC.1.5. Proof of Proposition 3

By Lemma 1, there exists a p -dimensional standard Brownian bridge $\mathbf{B}$ on $[0, 1]$ such that

$$\sqrt{n}D_0^{-1/2}D_1(\tau) [\hat{\beta}(\tau) - \beta(\tau)] \Rightarrow \mathbf{B}(\tau)$$

uniformly for $\tau \in [\tau_\ell, \tau_u]$, this implies that

$$\sqrt{n}D_0^{-1/2}D_1(\tau_j) [\hat{\beta}_j - \beta(\tau_j)] = O_{\mathbb{P}}(1)$$

uniformly for all $\tau_j \in [\tau_\ell, \tau_u]$, $j = 1, \dots, m$. Thus, by Assumptions (2.b) and (3.a), we have $\sup_{\tau \in [\tau_\ell, \tau_u]} \|D_1^{-1}(\tau)\| < \infty$, which implies

$$\max_{\substack{\tau_j \in [\tau_\ell, \tau_u] \\ 1 \leq j \leq m-1}} \sqrt{n} \|\hat{\beta}_j - \beta(\tau_j)\| = O_{\mathbb{P}}(1). \quad \square$$

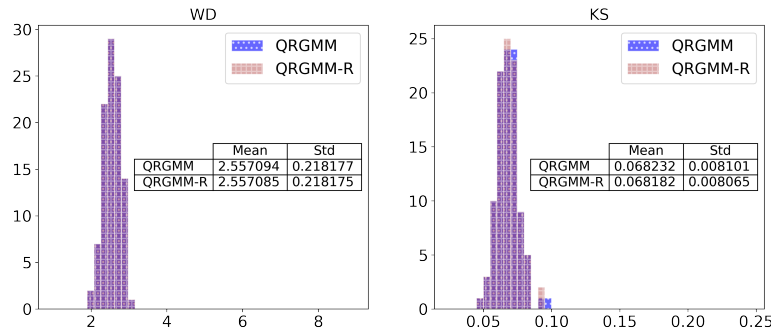
Table EC.1 Generated Samples Approximating Distribution of Y .

	Sample Mean		Sample SD	
	Mean	SD	Mean	SD
Truth	0.796524	0.048984	13.991365	0.050623
QRGMM	0.770990	0.298282	10.316667	0.321973
QRGMM-R	0.770992	0.298280	10.316670	0.321972

EC.2. Numerical Demonstration of QRGMM-R

We demonstrate QRGMM-R's performance using Test Problem 2 from Section 7.1, comparing it with QRGMM. Test Problem 2 exhibits more frequent quantile crossing than Test Problem 1 due to model misspecification, as its true conditional quantile function violates the linear quantile regression assumptions. Over 100 replications, the quantile crossing frequency is 0.0695 (± 0.0202).

We evaluate QRGMM-R using the experimental design from Section 7.1.1, examining sample means, standard deviations, Wasserstein distances, and KS statistics. Results appear in Table EC.1 and Figure EC.1, showing minimal distributional differences between QRGMM and QRGMM-R.

Figure EC.1 Generated Samples Approximating Distribution of Y : WD and KS.

In terms of computational speed, QRGMM-R generates 10^5 samples in 0.0040 seconds compared to 0.0039 seconds for QRGMM, averaged over 100 replications. The rearrangement step thus adds minimal computational overhead while preserving the distributional properties of QRGMM.

EC.3. Asymptotic Analysis of QRGMM-R

In this section, we first present Neocleous and Portnoy (2008)'s findings on quantile crossing in linear quantile regression models. We then establish the convergence in distribution of QRGMM-R. Our analysis builds on results from Neocleous and Portnoy (2008) and Chernozhukov et al. (2010).

EC.3.1. Frequency of Quantile Crossing

We first list the assumptions required in Neocleous and Portnoy (2008), which are stronger than Assumptions (2.b) and (3.b).

ASSUMPTION EC.1. (EC.1.a) For any given interval $[\tau_L, \tau_U] \subset (0, 1)$ with $0 < \tau_L < \tau_U < 1$, and any bounded $\mathcal{X}$, there exist constants a, b, c with $a > 0, b < \infty$, and $c < \infty$ such that

$$a \leq f_Y(F_Y^{-1}(\tau|\mathbf{x})|\mathbf{x}) \leq b, \quad |f'_Y(F_Y^{-1}(\tau|\mathbf{x})|\mathbf{x})| \leq c$$

uniformly for $\mathbf{x} \in \mathcal{X}, \tau \in [\tau_L, \tau_U]$.

(EC.1.b) $\|\mathbf{x}_i\| \leq d$ for some constant d uniformly in $i = 1, \dots, n$.

LEMMA EC.1 (**Neocleous and Portnoy (2008)**). Under Assumptions 1, (2.a), (3.a), and EC.1, let m satisfy $\limsup_{n \rightarrow \infty} n^\eta/m > 0$ and $\liminf_{n \rightarrow \infty} n^{1/2}/(m \log n) > 0$ for some $\eta > 0$. Then with probability tending to one, $\hat{Q}(\tau|\mathbf{x})$ is strictly monotonic uniformly on $\{\mathbf{x} \mid \|\mathbf{x}\| \leq d\}$ for $\tau \in [\tau_L + \frac{1}{m}, \tau_U - \frac{1}{m}]$, and

$$\sqrt{n} \left(\hat{\beta}(\tau) - \tilde{\beta}(\tau) \right) = O \left(\sqrt{n}/(m^2) \right) + O_{\mathbb{P}} \left((\log n/m)^{1/2} \right) + O_{\mathbb{P}} \left(n^{-1/4} (\log n)^{1/2} \right)$$

uniformly for $\tau \in [\tau_L + \frac{1}{m}, \tau_U - \frac{1}{m}]$.

The implications of Lemma EC.1 are twofold. First, strict monotonicity of $\hat{Q}(\tau|\mathbf{x})$ over (τ_L, τ_U) is guaranteed when m grows slower than $\sqrt{n}$. Second, when m grows faster than $n^{1/4}$, $\hat{\beta}(\tau)$ maintains the same first-order asymptotic Bahadur representation as $\tilde{\beta}(\tau)$.

EC.3.2. Convergence in Distribution of QRGMM-R

ASSUMPTION EC.2. $Q : (0, 1) \times \mathcal{X} \mapsto \mathbb{R}$ is a continuously differentiable function in both arguments, and $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$.

ASSUMPTION EC.3. $\tau \mapsto Q(\tau|\mathbf{x})$ has $\partial_\tau Q(\tau|\mathbf{x}) > 0$ for each $(\tau, \mathbf{x}) \in (0, 1) \times \mathcal{X}$.

Under Assumption 1, the monotonicity of $Q(\tau|\mathbf{x})$ in τ (Assumption EC.3) follows immediately.

LEMMA EC.2 (**Chernozhukov et al. (2010)**). Let $\mathcal{U}$ be $(0, 1)$ or its subinterval (a, b) , and let $\ell^\infty(\mathcal{U} \times \mathcal{X})$ denote the space of bounded measurable functions $h : \mathcal{U} \times \mathcal{X} \mapsto \mathbb{R}$. Consider an estimated quantile curve $\hat{Q}$ in $\ell^\infty(\mathcal{U} \times \mathcal{X})$. Under Assumptions EC.2 and EC.3, if

$$a_n \left(\hat{Q}(\tau|\mathbf{x}) - Q(\tau|\mathbf{x}) \right) \Rightarrow G(\tau|\mathbf{x})$$

in $\ell^\infty(\mathcal{U} \times \mathcal{X})$ as a stochastic process indexed by $(\tau, \mathbf{x}) \in \mathcal{U} \times \mathcal{X}$, where a_n is a sequence of constants going to infinity as $n \rightarrow \infty$, and $G(\tau|\mathbf{x})$ is a stochastic process with continuous paths, then

$$a_n \left(\hat{Q}^R(\tau|\mathbf{x}) - Q(\tau|\mathbf{x}) \right) \Rightarrow G(\tau|\mathbf{x})$$

uniformly over $\mathcal{U} \times \mathcal{X}$, establishing that the rearranged quantile curve maintains the same first-order asymptotic distribution as the original estimate.

Note that the output random variable of QRGMM-R is $\hat{Y}^R(\mathbf{x}^*) := \hat{Q}^R(U|\mathbf{x}^*)$, where $U \sim \text{Unif}(0, 1)$. By Lemma 1, under Assumptions 1-3, we know that for any given interval $[\tau_L, \tau_U] \subset (0, 1)$, we have that $\sqrt{n} [\tilde{\beta}(\tau) - \beta(\tau)]$ converge in distribution to a p-dimensional Brownian bridge uniformly over $\tau \in [\tau_L, \tau_U]$. Then under Assumption EC.1, according to Lemma EC.1, as long as m increases faster than $n^{1/4}$, $\sqrt{n} [\hat{Q}(\tau|\mathbf{x}^*) - Q(\tau|\mathbf{x}^*)]$ also converge to a Gaussian process uniformly over $[\tau_L + \frac{1}{m}, \tau_U - \frac{1}{m}] \times \mathcal{X}$, denoted as $G(\tau|\mathbf{x}^*)$. Then by Lemma EC.2, we have the following corollary.

COROLLARY EC.1. *Under Assumptions 1, (2.a), (3.a), EC.1, and EC.2, we have*

$$\sqrt{n} [\hat{Q}^R(\tau|\mathbf{x}^*) - Q(\tau|\mathbf{x}^*)] \Rightarrow G(\tau|\mathbf{x}^*)$$

uniformly over $[\tau_L + \frac{1}{m}, \tau_U - \frac{1}{m}] \times \mathcal{X}$.

This corollary implies that all three estimators $\tilde{Q}(\tau|\mathbf{x}^*)$, $\hat{Q}(\tau|\mathbf{x}^*)$, and $\hat{Q}^R(\tau|\mathbf{x}^*)$ share the same first-order asymptotic distribution. This further indicates that the asymptotic analysis for $\hat{Y}(\mathbf{x}^*)$ remains unaffected by the rearrangement operation of the new algorithm. Then the convergence in distribution of $\hat{Y}^R(\mathbf{x}^*)$, which is restated in a more rigorous form in the following Theorem EC.1, can be easily obtained through Corollary EC.1. As Corollary EC.1 implies that for any given $\mathbf{x}^* \in \mathcal{X}$, as $n, m \rightarrow \infty$,

$$\sup_{\tau \in I_3} |\hat{Q}^R(\tau|\mathbf{x}^*) - F_Y^{-1}(\tau|\mathbf{x}^*)| \xrightarrow{\mathbb{P}} 0,$$

which is aligned with Proposition 2, then according to Lemma 4, we can prove Theorem EC.1 (a rigorous restatement of Theorem 2) following the same way of Theorem 1.

THEOREM EC.1. *Under Assumptions 1, (2.a), (3.a), EC.1, and EC.2, the following results hold:*

(2.a) *For any $\mathbf{x}^* \in \mathcal{X}$, $\hat{Y}^R(\mathbf{x}^*) \Rightarrow Y(\mathbf{x}^*)$ as $n, m \rightarrow \infty$.*

(2.b) *For any $\mathbf{x}^* \in \mathcal{X}$,*

$$\sup_{y \in (-\infty, \infty)} |F_{\hat{Y}^R}(y|\mathbf{x}^*) - F_Y(y|\mathbf{x}^*)| \rightarrow 0,$$

as $n, m \rightarrow \infty$, where $F_{\hat{Y}^R}(y|\mathbf{x}^)$ is the conditional CDF of $\hat{Y}^R(\mathbf{x}^*)$ given $\mathbf{x}^*$.*

EC.4. Supplementary Materials for Section 6

EC.4.1. Multi-output QRGMM Algorithm

To extend the QRGMM framework to multidimensional outputs, we leverage the conditional distribution method, which sequentially generates univariate conditional distributions while preserving the use of single-output quantile regression models, as detailed in Algorithm 2. This approach avoids the complexities of joint quantile definitions and maintains the framework's core structure.

Algorithm 2 Multi-output QRGMM**Offline Stage:**

- Collect a dataset $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \dots, n\}$, and choose a quantile regression model $Q_l(\tau|\mathbf{x}, \mathbf{y}_{[l-1]})$ for the conditional quantile function $F_{Y_l}^{-1}(\tau|\mathbf{x}, \mathbf{y}_{[l-1]})$ for each $l = 1, \dots, d$.
- Choose a positive integer m and create the grid $\{\tau_j := j/m : j = 1, \dots, m-1\}$.
- For each $j = 1, \dots, m-1$ and $l = 1, \dots, d$, fit the model $Q_l(\tau|\mathbf{x}, \mathbf{y}_{[l-1]})$ with quantile level τ_j to the dataset, yielding $\tilde{Q}_l(\tau_j|\mathbf{x}, \mathbf{y}_{[l-1]})$.

Online Stage:

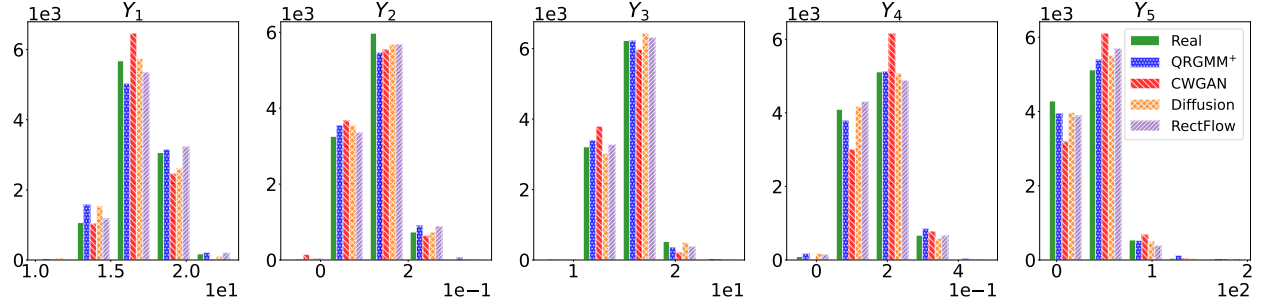
- Observe the covariates $\mathbf{x} = \mathbf{x}^*$.
- Generate $\{u_{l,k} : l = 1, \dots, d, k = 1, \dots, K\}$ independently from $\text{Unif}(0, 1)$.
- For each $l = 1, \dots, d$ and $k = 1, \dots, K$, calculate $\hat{Y}_{l,k} = \hat{Q}_l(u_{l,k}|\mathbf{x}^*, \hat{Y}_{1,k}, \dots, \hat{Y}_{l-1,k})$, where $\tilde{Q}_l(\tau|\mathbf{x}, \mathbf{y}_{[l-1]})$ is the linear interpolator of $\{(\tau_j, \tilde{Q}_l(\tau_j|\mathbf{x}, \mathbf{y}_{[l-1]})) : j = 1, \dots, m-1\}$.
- Output: $\{\hat{\mathbf{Y}}_k : k = 1, \dots, K\}$, where $\hat{\mathbf{Y}}_k = (\hat{Y}_{1,k}, \dots, \hat{Y}_{d,k})$.

EC.4.2. Multi-output Simulation for a Bank Queueing System

We demonstrate the performance of QRGMM⁺, which integrates Cannon (2011)’s neural network quantile regression model into the multi-output QRGMM framework (Algorithm 2), using a bank queueing simulation. Operating from 9:00 AM to 5:00 PM, this simulation focuses on transient rather than steady-state behavior, providing a complex multidimensional testing environment.

The simulator models two customer types with preemptive priority service: priority customers interrupt the service of normal customers and must be completely served before normal customers resume service. Customer arrivals follow Poisson processes with type-specific rates, and customers may renege (prematurely exit) if their waiting time exceeds a Gamma-distributed threshold. The bank operates six service counters, each with an individual queue and exponentially distributed service times at counter-specific rates. The simulator takes ten-dimensional input parameters $\mathbf{x} \in \mathbb{R}^{10}$ comprising two customer arrival rates, six counter service rates, and two reneging distribution parameters. It generates five output metrics $\mathbf{Y}(\mathbf{x}) = (Y_1(\mathbf{x}), Y_2(\mathbf{x}), Y_3(\mathbf{x}), Y_4(\mathbf{x}), Y_5(\mathbf{x}))$: average system time and renege rate for both customer types, plus maximum counter overtime after closing. The simulation is implemented using R’s *simmer* discrete-event-simulation package.

We train QRGMM⁺ and other models using $n = 10^4$ samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, where $\mathbf{x}$ follows a uniform distribution over a hypercube. For evaluation, we generate $K = 10^4$ samples at a fixed $\mathbf{x}^*$ from each trained metamodel and the simulator, comparing their conditional marginal distributions of Y_l given $\mathbf{x}^*$ for each dimension $l = 1, \dots, 5$ in Figure EC.2. The conditional marginal distributions of samples from QRGMM⁺, Diffusion, and RectFlow notably match the true distributions more closely than those from CWGAN across all dimensions.

Figure EC.2 Generated Samples Approximating Distribution of Y_l given $\mathbf{x}^* = \left(\frac{1}{4}, \frac{1}{8}, \frac{1}{15}, \frac{1}{16}, \frac{1}{17}, \frac{1}{19}, \frac{1}{11}, \frac{1}{13}, 18, 12\right)$.

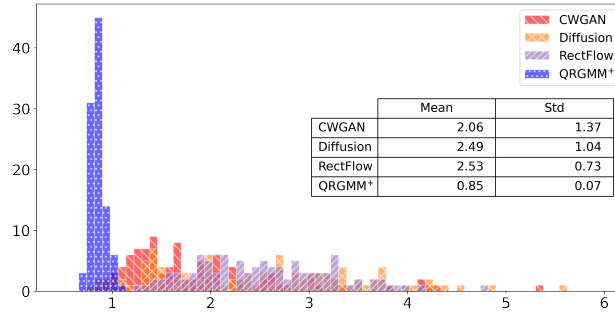
To evaluate unconditional distributions, we use each trained metamodel to produce a dataset $\{(\mathbf{x}'_k, \hat{\mathbf{y}}'_k)\}_{k=1}^K$ and compare it against a test dataset $\{(\mathbf{x}'_k, \mathbf{y}'_k)\}_{k=1}^K$ generated from the simulator. For each dimension $l = 1, \dots, 5$, we calculate the sample mean and sample standard deviation of $\{\hat{y}'_{l,k}\}_{k=1}^K$ or $\{y'_{l,k}\}_{k=1}^K$. We also compute the Wasserstein distance between the multi-dimensional samples from each metamodel and the simulator using the POT Python library for optimal transport computations.

Across 100 replications, we calculate means and standard deviations of all summary statistics (sample mean, sample standard deviation, and Wasserstein distance). As shown in Table EC.2, QRGMM⁺ consistently provides the closest approximation to the true first two moments across all output dimensions, outperforming other models. Moreover, QRGMM⁺ yields substantially smaller Wasserstein distances with lower variability (Figure EC.3), indicating a more accurate and stable recovery of the full conditional distribution for complex stochastic systems.

Table EC.2 Generated Samples Approximating Distribution of Y : Mean and Standard Deviation.

	Y_1		Y_2		Y_3		Y_4		Y_5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Sample Mean										
Truth	18.2215	0.0193	0.3794	0.0016	15.2626	0.0189	0.3316	0.0011	41.4609	0.2125
QRGMM ⁺	18.2202	0.0239	0.3794	0.0018	15.2605	0.0219	0.3316	0.0014	41.3725	0.2592
CWGAN	18.2490	0.1762	0.3830	0.0186	15.2804	0.1806	0.3309	0.0134	41.5027	2.0023
Diffusion	18.2290	0.0745	0.3794	0.0042	15.2237	0.1161	0.3311	0.0049	39.1964	1.1825
RectFlow	18.2163	0.0436	0.3790	0.0031	15.2213	0.0796	0.3310	0.0034	39.0578	0.7876
Sample SD										
Truth	2.1663	0.0163	0.1769	0.0008	1.9447	0.0130	0.1211	0.0008	19.4521	0.2214
QRGMM ⁺	2.1614	0.0187	0.1767	0.0009	1.9242	0.0200	0.1208	0.0009	18.9545	0.2535
CWGAN	2.1203	0.1498	0.1739	0.0106	1.9015	0.1147	0.1186	0.0090	18.7522	1.4463
Diffusion	2.1553	0.0350	0.1767	0.0016	1.9143	0.0291	0.1210	0.0017	18.4099	0.7548
RectFlow	2.1615	0.0315	0.1769	0.0016	1.9194	0.0336	0.1208	0.0017	17.9479	0.4712

Lastly, we evaluate online sample generation speed across 100 replications. QRGMM⁺ requires on average 0.1426 seconds (std. 0.0237) to generate 10^4 samples, which is slower than CWGAN (0.0434 ± 0.0027

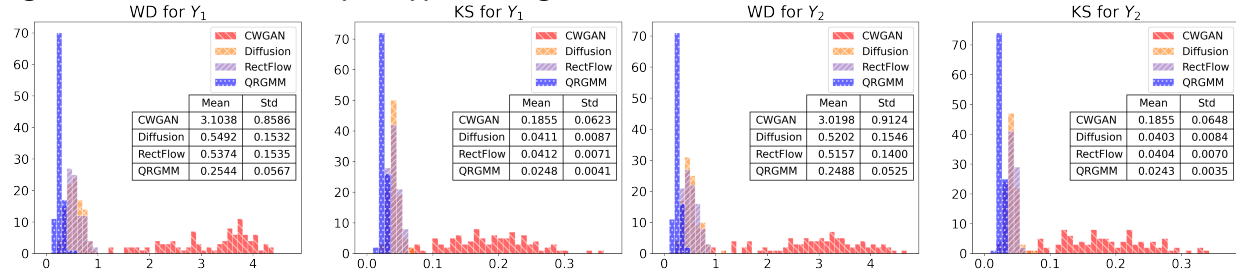
Figure EC.3 Generated Samples Approximating Distribution of Y : Wasserstein Distance.**Table EC.3** Generated Samples Approximating Distribution of Y_1 and Y_2 : Mean and Standard Deviation.

	Sample Mean				Sample SD			
	Y_1		Y_2		Y_1		Y_2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Truth	15.6000	0.0894	15.6290	0.0921	9.5288	0.0553	9.5357	0.0533
QRGMM	15.5018	0.1227	15.5427	0.1178	9.5026	0.0821	9.5127	0.0747
CWGAN	15.4213	2.7809	15.8217	2.6824	8.0462	1.7469	8.0826	1.7796
Diffusion	15.7320	0.2996	15.7170	0.2859	9.9453	0.1748	9.9406	0.1730
RectFlow	15.7210	0.2738	15.7563	0.2511	9.9604	0.1845	9.9406	0.1712

seconds) but comparable to Diffusion (0.1801 ± 0.0876 seconds) and RectFlow (0.0947 ± 0.0010 seconds). The slower speed of QRGMM⁺ stems from its sequential generation process across output dimensions. Nevertheless, all generative metamodels considered here operate well within real-time constraints for this five-dimensional setting. Further improving the computational efficiency of QRGMM⁺ for higher-dimensional output spaces remains an important direction for future research.

EC.5. Supplementary Materials for Section 7

This subsection complements Section 7.2.1 by providing additional results on the unconditional distributions of Y_1 and Y_2 obtained from generated samples. Table EC.3 and Figure EC.4 summarize the performance of all models in approximating these distributions. The results are consistent with the findings in Section 7.1.1, demonstrating that QRGMM achieves better accuracy and stability across multiple summary statistics.

Figure EC.4 Generated Samples Approximating Distribution of Y_1 and Y_2 : WD and KS.

References

- Chernozhukov V, Fernández-Val I, Galichon A (2010) Quantile and probability curves without crossing. *Econometrica* 78(3):1093–1125.
- Neocleous T, Portnoy S (2008) On monotonicity of regression quantile functions. *Statistics & Probability Letters* 78(10):1226–1229.
- Van der Vaart AW (2000) *Asymptotic Statistics* (Cambridge University Press).
- Cannon AJ (2011) Quantile regression neural networks: Implementation in R and application to precipitation down-scaling. *Computers & Geosciences* 37(9):1277–1284.