Adaptive estimation for high-dimensional quantile regression with misspecification and nonresponse

Wei Xiong, Dianliang Deng, Wanying Zhang and Dehui Wang

Liaoning University, University of Regina and Jilin University

Supplementary Material

In the supplementary materials, we provide the proofs of theoretical results in S1, the explanation of regularity conditions in S2. We also explore the relationship between different types of asymptotic risk optimality in S3, and supplement additional numerical results and real-data analysis in S4.

S1 Theoretical proofs

This part contains the proofs for Sections 2–3, along with auxiliary lemmas that facilitate the derivations.

S1.1 Proofs in Section 2.1

Lemma 1. *Let* $\beta_{\tau,j,0}^{(k)}$ *be the unique minimizer of* $E\{\rho_\tau(y-\bar{q}_\tau^{(k)}(\boldsymbol{u})-\beta_{\tau,j}z_j)-\rho_\tau(y)\}$ *among* $\delta_k \in \Delta$. *Under conditions* (C.1)–(C.3) *and* (C.6), *for any* $A_1>0$ *and* $0<\alpha\le\min(\alpha_1,\alpha_2)$, *there exist positive constants* A_2 *and* A_3 *such that for sufficiently large* n,

$$\Pr\left\{\sup_{\delta_k\in\Delta}\max_{1\le j\le p}\left|\hat{\beta}_{\tau,j}^{(k)}-\beta_{\tau,j,0}^{(k)}\right|\ge A_1n^{-\alpha}\right\}\le 2pK\exp(-A_2n^{1-4\alpha})+pK\exp(-A_3n^{1-2\alpha}).$$

Proof of Lemma 1. Denote that

$$B_n^{(k)}(\beta_{\tau,j}) = n^{-1}\sum_{i=1}^{n}\hat{W}_i\rho_\tau(y_i - \hat{q}_\tau^{(k)}(\boldsymbol{u}_i) - \beta_{\tau,j}z_{ij}),$$

$$B^{(k)}(\beta_{\tau,j}) = E\left\{\hat{W}\rho_\tau(y - \hat{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j}z_j)\right\},$$

$$\bar{B}^{(k)}(\beta_{\tau,j}) = E\rho_\tau(y - \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j}z_j).$$

Obviously $\hat{\beta}_{\tau,j,0}^{(k)} = \arg\min B_n^{(k)}(\beta_{\tau,j})$ and $\beta_{\tau,j,0}^{(k)}$ is the unique minimizer of $\bar{B}^{(k)}(\beta_{\tau,j})$. Since $B_n^{(k)}(\beta_{\tau,j})$ is directly approximated by $B^{(k)}(\beta_{\tau,j})$ for sufficiently large n, we firstly focus on the relationship between $\bar{B}^{(k)}(\beta_{\tau,j})$ and $B^{(k)}(\beta_{\tau,j})$, which is impacted by the convergence rate of the response mechanism.

We divide the proof into 3 steps:

Step 1. Let $W = r/\pi(\boldsymbol{x}, y)$, the correct specification of response mechanism leads $E(W - 1|\boldsymbol{x}, y) = 0$. Note that $1/\pi(\boldsymbol{x}, y) \le A_\pi^{-1} < \infty$ uniformly over $(\boldsymbol{x}, y)$. From (C.2) we have $\|\pi(\boldsymbol{x}, y) - \hat{\pi}(\boldsymbol{x}, y)\|_{L_2} \le C_\pi n^{-\alpha_2}$ for some $C_\pi > 0$, by Markov's inequality and Jensen's inequality, it follows that for any $\varepsilon > 0$, there exists $M_1 = C_\pi A_\pi^{-2}\varepsilon^{-1}$ such that

$$\begin{aligned}\Pr\left(|\hat{W} - W| \ge M_1 n^{-\alpha_2}\right) &\le \Pr\left(|\pi(\boldsymbol{x}, y) - \hat{\pi}(\boldsymbol{x}, y)| \ge \frac{C_\pi}{\varepsilon}n^{-\alpha_2}\right)\\ &\le \varepsilon C_\pi^{-1}n^{\alpha_2}\|\pi(\boldsymbol{x}, y) - \hat{\pi}(\boldsymbol{x}, y)\|_{L_1} \le \varepsilon.\end{aligned}$$

This indicates $|\hat{W} - W| = O_p(n^{-\alpha_2})$ (Section 1.2.5 of Serfling (1980)). On the other hand,

$$\begin{aligned}&\left|B^{(k)}(\beta_{\tau,j}^{(k)}) - \bar{B}^{(k)}(\beta_{\tau,j}^{(k)})\right|\\ \le&\left|E\left\{\rho_\tau\left(y - \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j}^{(k)}z_j - (\hat{q}_\tau^{(k)}(\boldsymbol{u}) - \bar{q}_\tau^{(k)}(\boldsymbol{u}))\right) - \rho_\tau\left(y - \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j}^{(k)}z_j\right)\right\}\right|\\ &+\left|E\left(\hat{W} - W\right)\left\{\rho_\tau\left(y - \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j}^{(k)}z_j - (\hat{q}_\tau^{(k)}(\boldsymbol{u}) - \bar{q}_\tau^{(k)}(\boldsymbol{u}))\right) - \rho_\tau\left(y - \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j}^{(k)}z_j\right)\right\}\right|\\ &+\left|E\left(\hat{W} - W\right)\rho_\tau\left(y - \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j}^{(k)}z_j\right)\right|\\ :=&\,|I_1(\beta_{\tau,j}^{(k)})| + |I_2(\beta_{\tau,j}^{(k)})| + |I_3(\beta_{\tau,j}^{(k)})|.\end{aligned}$$

By the definition of $\beta_{\tau,j,0}^{(k)}$ we have for any $\beta_{\tau,j}^{(k)}$ in a small neighborhood of $\beta_{\tau,j,0}^{(k)}$,

$$|I_3(\beta_{\tau,j}^{(k)})| \leq \left|E\left(\hat{W}-W\right)^2\right|^{1/2} O(1) = O(n^{-\alpha_2}).$$

To deal with $|I_2(\beta_{\tau,j}^{(k)})|$, the identity of Knight (1998) implies that

$$\begin{aligned}
&\left|I_2(\beta_{\tau,j}^{(k)})\right| \\
&\leq \left|E(\hat{W}-W)\left(\hat{q}_\tau^{(k)}(\boldsymbol{u})-\bar{q}_\tau^{(k)}(\boldsymbol{u})\right)\left\{I\left(y-\bar{q}_\tau^{(k)}(\boldsymbol{u})-\beta_{\tau,j}^{(k)}z_j\leq 0\right)-\tau\right\}\right| \\
&\quad + \left|E(\hat{W}-W)\int_0^{\hat{q}_\tau^{(k)}(\boldsymbol{u})-\bar{q}_\tau^{(k)}(\boldsymbol{u})}\left\{I\left(y-\bar{q}_\tau^{(k)}(\boldsymbol{u})-\beta_{\tau,j}^{(k)}z_j\leq s\right)-I\left(y-\bar{q}_\tau^{(k)}(\boldsymbol{u})-\beta_{\tau,j}^{(k)}z_j\leq 0\right)\right\}\mathrm{d}s\right| \\
&:= \left|I_{21}(\beta_{\tau,j}^{(k)})\right| + \left|I_{22}(\beta_{\tau,j}^{(k)})\right|.
\end{aligned}$$

Note that $|I(\cdot)-\tau| \leq \max(\tau, 1-\tau)$, we have

$$|I_{21}(\beta_{\tau,j}^{(k)})| \leq \max(\tau,1-\tau)\cdot\left|E(\hat{W}-W)^2\right|^{1/2}\left|E\left(\hat{q}_\tau^{(k)}(\boldsymbol{u})-\bar{q}_\tau^{(k)}(\boldsymbol{u})\right)^2\right|^{1/2} = O(n^{-(\alpha_1+\alpha_2)}).$$

The boundedness of $|I_{22}(\beta_{\tau,j}^{(k)})|$ is determined by Cauchy-Schwarz inequality, which implies

$$\begin{aligned}
|I_{22}(\beta_{\tau,j}^{(k)})| &\leq \left|E(\hat{W}-W)^2\right|^{1/2}\left|E\left(\hat{q}_\tau^{(k)}(\boldsymbol{u})-\bar{q}_\tau^{(k)}(\boldsymbol{u})\right)\int_0^{\hat{q}_\tau^{(k)}(\boldsymbol{u})-\bar{q}_\tau^{(k)}(\boldsymbol{u})} 1\cdot\mathrm{d}s\right|^{1/2} \\
&= O(n^{-(\alpha_1+\alpha_2)}).
\end{aligned}$$

On the other hand, above argument can be translated into $|I_1(\beta_{\tau,j}^{(k)})|$ such that

$$\begin{aligned}
|I_1(\beta_{\tau,j}^{(k)})| &\leq \max(\tau,1-\tau)\cdot\left|E\left(\hat{q}_\tau^{(k)}(\boldsymbol{u})-\bar{q}_\tau^{(k)}(\boldsymbol{u})\right)\right| + \left|E\int_0^{\hat{q}_\tau^{(k)}(\boldsymbol{u})-\bar{q}_\tau^{(k)}(\boldsymbol{u})} 1\cdot\mathrm{d}s\right| \\
&\leq O(n^{-\alpha_1}).
\end{aligned}$$

Therefore, let $\alpha_3 = \min(\alpha_1,\alpha_2)$, for some positive B_1 it follows that

$$0 \leq \left|B^{(k)}(\beta_{\tau,j}^{(k)}) - \bar{B}^{(k)}(\beta_{\tau,j}^{(k)})\right| \leq B_1 n^{-\alpha_3}.$$

Step 2. For $\delta \neq 0$, let $\beta_{\tau,j}^{(k)} = \beta_{\tau,j,0}^{(k)} + \delta$. According to Lemma 2 of Hjort and Pollard (1993),

$$\begin{aligned}\Pr\left\{\left|\hat{\beta}_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}\right| \geq |\delta|\right\} \leq \Pr\Bigg\{ & \sup_{|\beta_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}| \leq |\delta|} \left|B_n^{(k)}(\beta_{\tau,j}^{(k)}) - B^{(k)}(\beta_{\tau,j,0}^{(k)})\right| + B_1 n^{-\alpha_3} \\ & \geq \frac{1}{2} \inf_{|\beta_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}| = |\delta|} \left(\bar{B}^{(k)}(\beta_{\tau,j}^{(k)}) - \bar{B}^{(k)}(\beta_{\tau,j,0}^{(k)})\right)\Bigg\}.\end{aligned} \tag{S1.1}$$

Let $\bar{y}^{(k)} = y - \bar{q}_\tau^{(k)}(\boldsymbol{u})$ and $\delta = ln^{-\alpha}$ with $|l| > 0$, the similar argument of Lemma A3 in He, Wang and Hong (2013) follows that

$$\begin{aligned}\bar{B}^{(k)}(\beta_{\tau,j}^{(k)}) - \bar{B}^{(k)}(\beta_{\tau,j,0}^{(k)}) &= E\left\{z_j \delta\left(I(\bar{y}^{(k)} \leq z_j \beta_{\tau,j,0}^{(k)}) - \tau\right)\right\} \\ &\quad + E\left\{\int_0^{z_j\delta}\left[I(y \leq \bar{q}^{(k)}(\boldsymbol{u}) + z_j\beta_{\tau,j,0}^{(k)} + s) - I(y \leq \bar{q}^{(k)}(\boldsymbol{u}) + z_j\beta_{\tau,j,0}^{(k)})\right]\mathrm{d}s\right\} \\ &= B_2 n^{-2\alpha},\end{aligned}$$

where $B_2 > 0$. The last equation is formed by the fact

$$E\{z_j(F_{\bar{y}^{(k)}}(z_j\beta_{\tau,j,0}^{(k)}|z_j) - \tau)\} = 0,$$

thus we have

$$\frac{1}{2} \inf_{|\beta_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}| = |l|n^{-\alpha}} \left(\bar{B}^{(k)}(\beta_{\tau,j}^{(k)}) - \bar{B}^{(k)}(\beta_{\tau,j,0}^{(k)})\right) \geq B_2 n^{-2\alpha}.$$

Step 3. Note that when $2\alpha < \alpha_3$, there exists $b > 0$ such that

$$\begin{aligned}&\Pr\left\{\sup_{|\beta_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}| \leq |l|n^{-\alpha}} \left|B_n^{(k)}(\beta_{\tau,j}) - B^{(k)}(\beta_{\tau,j,0}^{(k)})\right| + B_1 n^{-\alpha_3} \geq B_2 n^{-2\alpha}\right\} \\ &\leq \Pr\left\{\left|B_n^{(k)}(\beta_{\tau,j,0}^{(k)}) - B^{(k)}(\beta_{\tau,j,0}^{(k)})\right| \geq \frac{1}{2} b n^{-2\alpha}\right\} \\ &\quad + \Pr\left\{\sup_{|\beta_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}| \leq |l|n^{-\alpha}} \left|B_n^{(k)}(\beta_{\tau,j}^{(k)}) - B_n^{(k)}(\beta_{\tau,j,0}^{(k)}) + B^{(k)}(\beta_{\tau,j,0}^{(k)}) - B^{(k)}(\beta_{\tau,j}^{(k)})\right| \geq \frac{1}{2} b n^{-2\alpha}\right\} \\ &:= J_1 + J_2.\end{aligned}$$

In J_1,

$$\left|B_n^{(k)}(\beta_{\tau,j,0}^{(k)}) - B^{(k)}(\beta_{\tau,j,0}^{(k)})\right| = n^{-1}\left|\sum_{i=1}^{n}\{V_i - E(V_i)\}\right|,$$

where $V_i = \hat{W}_i\rho_\tau(y_i - \hat{q}_\tau^{(k)}(\boldsymbol{u}_i) - \beta_{\tau,j,0}^{(k)}z_{ij})$. Obviously $|V_i|$ is uniformly bounded by (C.6), combining Lipschitz continuity of ρ_τ we have $|V_i - E(V_i)| < \infty$ and $D(V_i - E(V_i)) \le \sigma^2 < \infty$. Employing Bernstein's inequality (Lemma 2.2.9 of van der Vaart and Wellner (1996)) it follows that

$$J_1 \le 2\exp\left(-\frac{b^2n^{1-4\alpha}/4}{2\sigma^2 + bn^{-2\alpha}/3}\right) \le 2\exp(-A_2n^{1-4\alpha}). \tag{S1.2}$$

To demonstrate the upper-bound of J_2, Massart's concentration inequality provides a convenient tool which is introduced in Lemma 8.5 of He, Wang and Hong (2013). Let $U_i = \hat{W}_i\rho_\tau(y_i - \hat{q}_\tau^{(k)}(\boldsymbol{u}_i) - \beta_{\tau,j}^{(k)}z_{ij}) - \hat{W}_i\rho_\tau(y_i - \hat{q}_\tau^{(k)}(\boldsymbol{u}_i) - \beta_{\tau,j,0}^{(k)}z_{ij})$, we have

$$|U_i| \le 2\hat{W}_i\left|z_{i,j}(\beta_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)})\right| \le A^*n^{-\alpha}$$

for some $A^* > 0$ and

$$E\left\{\sup_{|\beta_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}| \le A_1n^{-\alpha}} n^{-1}\left|\sum_{i=1}^{n}(U_i - EU_i)\right|\right\} \le A^{**}n^{-\alpha-1/2}$$

for some $A^{**} > 0$. Therefore,

$$J_2 \le \exp\left(-A_3n^{1-2\alpha}\right)$$

for positive α and $A_1 = |l|$. Combining (S1.1) and (S1.2), we have

$$\Pr\left\{\left|\hat{\beta}_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}\right| \ge A_1n^{-\alpha}\right\} \le 2\exp(-A_2n^{1-4\alpha}) + \exp(-A_3n^{1-2\alpha})$$

for all $1 \le j \le p$ and δ_k. By Boole's inequality,

$$\Pr\left\{\sup_{\delta_k\in\Delta}\max_{1\le j\le p}\left|\hat{\beta}_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}\right| \ge A_1n^{-\alpha}\right\} \le \sum_{k=1}^{K}\sum_{j=1}^{p}\Pr\left\{\left|\hat{\beta}_{\tau,j}^{(k)} - \beta_{\tau,j,0}^{(k)}\right| \ge A_1n^{-\alpha}\right\},$$

The lemma is established. □

Lemma 2 (Strong consistency of $\hat{\xi}_\tau^{(k)}$)**.** *For each $\delta_k \in \Delta$, let $\xi_{\tau,0}^{(k)}$ be the true value of $\xi_\tau^{(k)}$. If conditions* (C.1)–(C.3), (C.5)–(C.6) *hold, $\hat{\xi}_\tau^{(k)}$ converges to $\xi_{\tau,0}^{(k)}$ with probability one as $n \to \infty$.*

Proof of Lemma 2. Obviously $\xi_{\tau,0}^{(k)}$ is the minimizer of $E\{\rho_\tau(y - \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \xi)\}$. Based on the same proof of Lemma 1, for any $A_4 > 0$ there exist positive A_5 and q such that

$$\Pr\left\{\sup_{\delta_k \in \Delta}\left|\hat{\xi}_\tau^{(k)} - \xi_{\tau,0}^{(k)}\right| \geq A_4 n^{-\alpha}\right\} \leq 3K\exp(-A_5 n^{1-4\alpha}) \leq \frac{3}{A_5 n^{1+q}}.$$

Therefore, for every $\varepsilon_2 > 0$,

$$\sum_{n=1}^{\infty}\Pr\left\{\sup_{\delta_k \in \Delta}\left|\hat{\xi}_\tau^{(k)} - \xi_{\tau,0}^{(k)}\right| \geq \varepsilon_2\right\} \leq \sum_{n=1}^{\infty}\Pr\left\{\sup_{\delta_k \in \Delta}\left|\hat{\xi}_\tau^{(k)} - \xi_{\tau,0}^{(k)}\right| \geq A_4 n^{-\alpha}\right\} < \infty.$$

By Theorem 1.3.4 of Serfling (1980), the lemma is proved. □

Lemma 3. *Under conditions* (C.1), (C.3) *and* (C.6), *for $\delta_k \in \Delta$ and $1 \leq j \leq p$,*

$$\beta_{\tau,j,0}^{(k)} = \arg\min_{\beta} E\{w_\tau(\boldsymbol{u}, z_j) R_\tau^2(\boldsymbol{u}, z_j; \beta)\},$$

where $R_\tau(\boldsymbol{u}, z_j; \beta) = \beta z_j + \bar{q}_\tau^{(k)}(\boldsymbol{u}) - \beta_{\tau,j,0} z_j - q_\tau(\boldsymbol{u})$, $\beta_{\tau,j,0} = \arg\min_\beta E\rho_\tau(y - q(\boldsymbol{u}) - \beta z_j)$ and

$$w_\tau(\boldsymbol{u}, z_j) = \int_0^1 (1-s) f_{\bar{y}}\left(s(\beta z_j + \bar{q}_\tau^{(k)}(\boldsymbol{u})) + (1-s)(\beta_{\tau,j,0} z_j + q_\tau(\boldsymbol{u}))\middle| z_j\right) \mathrm{d}s.$$

Proof of Lemma 3. See Theorem 1 of Angrist et al. (2006). □

Proof of Theorem 1. For (i), note that

$$\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n} = n^{-1}\sum_{i=1}^{n}\left|\hat{\beta}_{\tau,j}^{(k)} z_{ij} - \hat{\xi}_\tau^{(k)}\right|,$$

according to condition (C.4), one can obtain that

$$
\begin{aligned}
&\Pr\left\{\inf_{\delta_k\in\Delta}\max_{j\in\mathcal{I}^c}\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n}\geq\sup_{\delta_k\in\Delta}\min_{j\in\mathcal{I}}\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n}\right\}\\
\leq&\Pr\left\{\left(\inf_{\delta_k\in\Delta}\max_{j\in\mathcal{I}^c}+\sup_{\delta_k\in\Delta}\min_{j\in\mathcal{I}}\right)n^{-1}\sum_{i=1}^{n}\left|z_{ij}\left(\hat{\beta}_{\tau,j}^{(k)}-\beta_{\tau,j,0}^{(k)}\right)\right|+\left(\inf_{\delta_k\in\Delta}+\sup_{\delta_k\in\Delta}\right)\left|\hat{\xi}_{\tau}^{(k)}-\xi_{\tau}^{(k)}\right|\right.\\
&\qquad\left.\geq\left(\sup_{\delta_k\in\Delta}\min_{j\in\mathcal{I}}-\inf_{\delta_k\in\Delta}\max_{j\in\mathcal{I}^c}\right)n^{-1}\sum_{i=1}^{n}\left|\bar{q}_{\tau}^{(k)}(\boldsymbol{u}_i)+\beta_{\tau,j,0}^{(k)}z_{ij}-\xi_{\tau,0}^{(k)}\right|\right\}\\
\leq&\Pr\left\{\sup_{\delta_k\in\Delta}\max_{1\leq j\leq p}n^{-1}\sum_{i=1}^{n}\left|z_{ij}\left(\hat{\beta}_{\tau,j}^{(k)}-\beta_{\tau,j,0}^{(k)}\right)\right|+\sup_{\delta_k\in\Delta}2\left|\hat{\xi}_{\tau}^{(k)}-\xi_{\tau}^{(k)}\right|\geq c\right\}.
\end{aligned}
$$

Therefore, for any $0<\varepsilon^*<c$,

$$
\begin{aligned}
&\Pr\left\{\inf_{\delta_k\in\Delta}\max_{j\in\mathcal{I}^c}\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n}\geq\sup_{\delta_k\in\Delta}\min_{j\in\mathcal{I}}\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n}\right\}\\
\leq&\Pr\left\{\sup_{\delta_k\in\Delta}\max_{1\leq j\leq p}\left|\hat{\beta}_{\tau,j}^{(k)}-\beta_{\tau,j,0}^{(k)}\right|\cdot\left(n^{-1}\sum_{i=1}^{n}|z_{ij}|\right)\geq\varepsilon^*\right\}+\Pr\left\{\sup_{\delta_k\in\Delta}\left|\hat{\xi}_{\tau}^{(k)}-\xi_{\tau}^{(k)}\right|\geq c-\varepsilon^*\right\}\\
:=&\,S_1+S_2.
\end{aligned}
$$

Firstly, $S_2=o(1)$ by Lemma 2 and Theorem 2 (c) of Ferguson (1996). For the first term of the right side, we have

$$
\begin{aligned}
S_1\leq&\Pr\left\{\sup_{\delta_k\in\Delta}\max_{1\leq j\leq p}\left|\hat{\beta}_{\tau,j}^{(k)}-\beta_{\tau,j,0}^{(k)}\right|\geq\sqrt{\frac{\varepsilon^*}{2}}\right\}+\Pr\left\{\max_{1\leq j\leq p}n^{-1}\sum_{i=1}^{n}\left(|z_{ij}|-E|z_{ij}|\right)\geq\sqrt{\frac{\varepsilon^*}{2}}\right\}\\
&+\Pr\left\{\sup_{\delta_k\in\Delta}\max_{1\leq j\leq p}\left|\hat{\beta}_{\tau,j}^{(k)}-\beta_{\tau,j,0}^{(k)}\right|\cdot\max_{1\leq j\leq p}E|z_{ij}|\geq\frac{\varepsilon^*}{2}\right\}\\
:=&\,S_{11}+S_{12}+S_{13}.
\end{aligned}
$$

By Lemma 1 and (C.3), it is easy to show that for sufficiently large n, S_{11} and S_{13} are both controlled by $o(1)$. On the other hand, together with Bernstein's inequality and (C.5),

$$
S_{12}\leq\sum_{i=1}^{p}\Pr\left\{n^{-1}\sum_{i=1}^{n}\left(|z_{ij}|-E|z_{ij}|\right)\geq\sqrt{\frac{\varepsilon^*}{4}}\right\}=o(1).
$$

Hence

$$\Pr\left\{\inf_{\delta_k\in\Delta}\max_{j\in\mathcal{I}^c}\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n}\geq\sup_{\delta_k\in\Delta}\min_{j\in\mathcal{I}}\left\|\hat{f}_j^{(k)}(\tau)\right\|_{1,n}\right\}\leq o(1)\to 0,\quad \text{as}\ \ n\to\infty.$$

To prove (ii), we set $k=1$ without loss of generality. Define

$$\left\|f_j^{(1)}(\tau)\right\|_{1,n}=n^{-1}\sum_{i=1}^{n}|\beta_{\tau,j}^{(1)}z_{ij}-\xi_\tau^{(1)}|,$$

and by triangle inequality, one can see that

$$\begin{aligned}\left|\left\|\hat{f}_j^{(1)}(\tau)\right\|_{1,n}-\left\|f_j^{(1)}(\tau)\right\|_{1,n}\right|&=\left|n^{-1}\sum_{i=1}^{n}\left|\hat{\beta}_{\tau,j}^{(k)}z_{ij}-\hat{\xi}_\tau^{(k)}\right|-n^{-1}\sum_{i=1}^{n}\left|\beta_{\tau,j,0}^{(1)}z_{ij}-\xi_\tau^{(1)}\right|\right|\\&\leq\left|\hat{\beta}_{\tau,j}^{(k)}-\beta_{\tau,j,0}^{(1)}\right|\cdot\left(n^{-1}\sum_{i=1}^{n}|z_{ij}|\right)+\left|\hat{\xi}_\tau^{(k)}-\xi_\tau^{(1)}\right|.\end{aligned}$$

Similar to the previous discussion, we have

$$\begin{aligned}&\Pr\left\{\max_{1\leq j\leq p}\left|\left\|\hat{f}_j^{(1)}(\tau)\right\|_{1,n}-\left\|f_j^{(1)}(\tau)\right\|_{1,n}\right|>0\right\}\\&\leq\Pr\left\{\max_{1\leq j\leq p}\left|\hat{\beta}_{\tau,j}^{(1)}-\beta_{\tau,j,0}^{(1)}\right|\cdot\left(n^{-1}\sum_{i=1}^{n}|z_{ij}|\right)>0\right\}+\Pr\left\{\left|\hat{\xi}_\tau^{(1)}-\xi_\tau^{(1)}\right|>0\right\}=o(1).\end{aligned}$$

Note that $\xi_\tau^{(1)}$ is the "true" τ-th unconditional quantile of $\bar{y}$ and Lemma 3 reflects $\beta_{\tau,j,0}^{(1)}=\beta_{\tau,j,0}$, the "true coefficient" of $Q_\tau(\bar{y}|z_j)$, as the target function is nonnegative. In this case,

$$\begin{aligned}\Pr\left(\mathcal{I}_\tau\subset\bigcup_{\delta_k\in\Delta}\hat{\mathcal{I}}_\tau^{(k)}\right)&\geq\Pr\left(\mathcal{I}_\tau\subset\hat{\mathcal{I}}_\tau^{(1)}\right)\geq\Pr\left(\min_{j\in\mathcal{I}_\tau}\left\|\hat{f}_j^{(1)}(\tau)\right\|_{1,n}\geq v_n\right)\\&\geq\Pr\left(\min_{j\in\mathcal{I}_\tau}\|f_j(\tau)\|_{1,n}-\max_{j\in\mathcal{I}_\tau}\left|\left\|\hat{f}_j^{(1)}(\tau)\right\|_{1,n}-\|f_j(\tau)\|_{1,n}\right|\geq v_n\right)\\&=1-\Pr\left(\max_{j\in\mathcal{I}_\tau}\left|\left\|\hat{f}_j^{(1)}(\tau)\right\|_{1,n}-\|f_j(\tau)\|_{1,n}\right|\geq\min_{j\in\mathcal{I}_\tau}\|f_j(\tau)\|_{1,n}-v_n\right)\\&\geq 1-o(1).\end{aligned}$$

This completes the proof. □

S1.2 Proofs in Section 2.3–2.4

Lemma 4. *Let $e_\tau^* = e_\tau - Ee_\tau$, under condition* (C.8)*, e_τ^* is a centered sub-exponential random variable and*

(i) $E\exp(t|e_\tau|) \le 2\exp(8e^2C_{2,\tau}^2t^2 + C_{2,\tau}t)$, *for* $\forall|t| \le 1/(4eC_{2,\tau})$.

(ii) $E\{|e_\tau|^2\exp(t|e_\tau|)\} \le 16\sqrt{2}C_{2,\tau}^2\exp(16e^2C_{2,\tau}^2t^2 + C_{2,\tau}t)$, *for* $\forall|t| \le 1/(8eC_{2,\tau})$.

Proof of Lemma 4. It is obviously seen that e_τ^* is a centered sub-exponential variable such that

$$\|e_\tau^*\|_{\mathrm{SEXP}} = \sup_{k\ge1}\frac{1}{k}\left\{E|e_\tau - Ee_\tau|^k\right\}^{1/k} \le \|e_\tau\|_{\mathrm{SEXP}} + \sup_{k\ge1}\frac{1}{k}E|e_\tau| \le 2C_{2,\tau},$$

hence for $|t| \le 1/(4eC_{2,\tau})$,

$$E\exp(t|e_\tau|)\{E\exp(t|Ee_\tau|)\}^{-1} \le E\exp(t|e_\tau^*|) \le 2\exp(8e^2C_{2,\tau}^2t^2),$$

where the second inequality is formed by Lemma 2 (a) of Gu and Zou (2019). This directly verifies (i) with (C.9). To prove (ii), note that

$$E\{|e_\tau|^2\exp(t|e_\tau|)\} \le (E|e_\tau|^4)^{1/2}\{E\exp(2t|e_\tau|)\}^{1/2},$$

where $(E|e_\tau|^4)^{1/2} \le 16C_{2,\tau}^2$ and $E\exp(2t|e_\tau|) \le 2\exp(32e^2C_{2,\tau}^2t^2 + 2C_{2,\tau}t)$ for $2|t| \le 1/(4eC_{2,\tau})$. This completes the proof. □

Proof of Theorem 2. We only focus on the case that $B = 1$ with $\{b_1, ..., b_n\} = \{1, ..., n\}$ and omit the index "b" for simplicity of expression. Homologous processes are implemented in the completely consistent situation of Section 2.2 according with the convexity of check loss. Let $h_W^*(x) = \exp\{-\lambda W\rho_{\tau,a}(x)\}$ and

$$p_N^n = \sum_{(k,m)\in\hat{\mathcal{S}}_\tau}\omega_{(k,m)}\prod_{i=N+1}^{n}h_{\hat{W}_i}^*\left(y_i - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_i)\right),$$

it is straightforward to obtain

$$\begin{aligned}
p_N^n = & \sum_{(k,m)\in\hat{\mathcal{S}}_\tau} \frac{\omega_{(k,m)}}{\sum_{(k,m)\in\hat{\mathcal{S}}_\tau}\omega_{(k,m)}} h^*_{\hat{W}_{N+1}}\left(y_{N+1}-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_{N+1})\right) \\
& \times \sum_{(k,m)\in\hat{\mathcal{S}}_\tau} \frac{\omega_{(k,m)}\prod_{i=N+1}^{N+2} h^*_{\hat{W}_i}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)}{\sum_{(k,m)\in\hat{\mathcal{S}}_\tau}\omega_{(k,m)} h^*_{\hat{W}_{N+1}}\left(y_{N+1}-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_{N+1})\right)} \\
& \times\cdots\times \sum_{(k,m)\in\hat{\mathcal{S}}_\tau} \frac{\omega_{(k,m)}\prod_{i=N+1}^{n} h^*_{\hat{W}_i}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)}{\sum_{(k,m)\in\hat{\mathcal{S}}_\tau}\omega_{(k,m)}\prod_{i=N+1}^{n-1} h^*_{\hat{W}_{N+1}}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)} \\
= & \prod_{i=N+1}^{n}\sum_{(k,m)\in\hat{\mathcal{S}}_\tau} \hat{\Omega}_{(k,m),i} h^*_{\hat{W}_i}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right).
\end{aligned}$$

and

$$0 < h^*_W(x) = \exp\{-\lambda W\rho_{\tau,a}(x)\} \le \exp\{-\lambda W\rho_\tau(x)\} := h_W(x).$$

Therefore for the i-th copy, denote ζ as a discrete measure induced by the event $A_{k,m} = \{\text{the } (k,m)\text{-th candidate in } \hat{\mathcal{S}}_\tau \text{ is selected}\}$ such that $\zeta(A_{k,m}) = \hat{\Omega}_{(k,m),i}$, which implies

$$-\log(p_N^n) = -\sum_{i=N+1}^{n}\log\left\{E_\zeta h^*_{\hat{W}_i}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}.$$

Let $m_{k,\zeta}(\boldsymbol{x})$ be the k-th ($k \ge 2$) central moment of $\boldsymbol{x}$ with respect to ζ, Lemma 3.6.1 of Catoni (2004) indicates that

$$\begin{aligned}
\log\left\{E_\zeta h^*_{\hat{W}_i}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \le & -\lambda E_\zeta\left\{\hat{W}_i\rho_\tau\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \\
& + \frac{\lambda^2}{2} m_{2,\zeta}\left\{\hat{W}_i\rho_{\tau,a}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}\exp\left\{\lambda\max\left(0, M_{\zeta_\gamma,i}\right)\right\},
\end{aligned} \tag{S1.3}$$

where

$$M_{\zeta_\gamma,i} = \sup_{\gamma\in[0,\lambda]} \frac{m_{3,\zeta_\gamma}\left\{\hat{W}_i\rho_{\tau,a}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}}{m_{2,\zeta_\gamma}\left\{\hat{W}_i\rho_{\tau,a}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}},$$

and the induce measure is

$$\zeta_\gamma = \frac{\hat{\Omega}_{(k,m),i} \exp\left\{-\gamma \hat{W}_i \rho_{\tau,a}(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i))\right\}}{\sum_{(k,m)\in\hat{\mathcal{S}}_{\tau,N}} \hat{\Omega}_{(k,m),i} \exp\left\{-\gamma \hat{W}_i \rho_{\tau,a}(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i))\right\}}, \qquad \gamma \in [0,\lambda].$$

Note that condition (C.2) implies $\sup_{1\le i\le n}|\hat{W}_i| \le A_\pi^{-1}$. On the other hand, $\rho_{\tau,a}(\cdot)$ is constructed by the sum of original loss and a quadratic part. For any $(k,m) \in \hat{\mathcal{S}}_\tau$, the strong convexity of the squared loss function derives that

$$\begin{aligned}\sup_{(k,m)\in\mathcal{S}_\tau} \left|\left(y - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x})\right)^2 - e_\tau^2\right| \le 2|e_\tau| \sup_{(k,m)\in\mathcal{S}_\tau} \left|\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right| \\ + 2 \sup_{(k,m)\in\mathcal{S}_\tau} \left|\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right|^2,\end{aligned}$$

and

$$\begin{aligned}&E_\zeta\left\{\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)^2 - \left(y_i - E_\zeta \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)^2\right\}^2 \\ \le &\left\{2|e_\tau| + 4 \sup_{(k,m)\in\hat{\mathcal{S}}_\tau} \left|\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right|\right\}^2 E_\zeta\left\{\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i) - E_\zeta \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right\}^2.\end{aligned}$$

Applying the same idea of step 1 in Lemma 1, we have

$$\begin{aligned}M_{\zeta_\gamma,i} &\le \sup_{\gamma\in[0,\lambda]} \sup_{(k_1,m_1)\in\hat{\mathcal{S}}_\tau} \left|\hat{W}_i \rho_{\tau,a}\left(y_i - \hat{Q}^{(k_1,m_1)}_{\tau,N}(y|\boldsymbol{x}_i)\right) - E_{\zeta_\gamma}\left\{\hat{W}_i \rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}\right| \\ &\le \left|\hat{W}_i\right| \sup_{\gamma\in[0,\lambda]} \sup_{\substack{(k_1,m_1)\in\hat{\mathcal{S}}_\tau \\ (k_2,m_2)\in\hat{\mathcal{S}}_\tau}} \left|\rho_{\tau,a}\left(y_i - \hat{Q}^{(k_1,m_1)}_{\tau,N}(y|\boldsymbol{x}_i)\right) - \rho_{\tau,a}\left(y_i - \hat{Q}^{(k_2,m_2)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right| \\ &\le 2\left|\hat{W}_i\right|\left[\{\max(\tau,1-\tau)+1\} C_{1,\tau} + a_n \sup_{(k,m)\in\hat{\mathcal{S}}_\tau} \left|\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)^2 - e_{\tau,i}^2\right|\right] \\ &\le \overline{C} A_\pi^{-1} C_{1,\tau} + 2a_n A_\pi^{-1} C_{1,\tau}^2 + 4a_n A_\pi^{-1} C_{1,\tau}|e_{\tau,i}|,\end{aligned}$$

and

$$
\begin{aligned}
& m_{2,\zeta}\left\{\hat{W}_i\rho_{\tau,a}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \\
&\le \left|\hat{W}_i\right|^2 E_\zeta\left\{\rho_{\tau,a}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)-\rho_{\tau,a}\left(y_i-E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}^2 \\
&\le 2\left|\hat{W}_i\right|^2\{\max(\tau,1-\tau)+1\}^2 E_\zeta\left\{\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)-E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right\}^2 \\
&\quad + 2a_n^2\left|\hat{W}_i\right|^2 E_\zeta\left\{\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)^2-\left(y_i-E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)^2\right\}^2 \\
&\le \left\{1/2A_\pi^{-2}\overline{C}^2+16a_n^2A_\pi^{-2}C_{1,\tau}^2+8a_n^2A_\pi^{-2}|e_{\tau,i}|^2\right\}E_\zeta\left\{\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)-E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right\}^2,
\end{aligned}
$$

where $\overline{C}=2(\max(\tau,1-\tau)+1)$. By the proof of Theorem 1 in Shan and Yang (2009),

$$
\begin{aligned}
& E_\zeta\left\{\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)-E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right\}^2 \\
&\le a_n^{-1}E_\zeta\left\{\rho_{\tau,a}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)-\rho_{\tau,a}\left(y_i-E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\},
\end{aligned}
$$

hence (S1.3) can be written as

$$
\begin{aligned}
\log\left\{E_\zeta h^*_{\hat{W}_i}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \le & -\lambda\left\{\hat{W}_iE_\zeta\rho_\tau\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \\
& + \frac{\lambda^2}{a_n}\left\{1/2A_\pi^{-2}\overline{C}^2+16a_n^2A_\pi^{-2}C_{1,\tau}^2+8a_n^2A_\pi^{-2}|e_{\tau,i}|^2\right\} \\
& \cdot E_\zeta\left\{\rho_{\tau,a}\left(y_i-\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)-\rho_{\tau,a}\left(y_i-E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \\
& \cdot \exp\{\lambda(\overline{C}A_\pi^{-1}C_{1,\tau}+2a_nA_\pi^{-1}C_{1,\tau}^2)+4\lambda a_nA_\pi^{-1}C_{1,\tau}|e_{\tau,i}|\}.
\end{aligned}
$$

Using Lemma 4 and taking the expectation (denoted by E_i) of both sides in (S1.3) with respect to y_i among $\boldsymbol{x}_i\cup(\boldsymbol{x}_l,y_l)_{l=1}^{i-1}$, one can derive that for $\mathcal{M}_{1,\tau}\lambda\le(8e)^{-1}$,

$$
\begin{aligned}
&E_i \log\left\{E_\zeta h^*_{\hat{W}_i}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \\
\le& -\lambda E_i\left\{(\hat{W}_i - W_i)E_\zeta \rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} - \lambda E_i E_\zeta \rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right) \\
&+ \frac{\lambda^2}{a_n}\left(2\mathcal{K}_{1,\tau}\exp\left\{(\mathcal{M}_{1,\tau}+\mathcal{K}_{2,\tau})\lambda + \mathcal{M}_{2,\tau}\lambda^2\right\} + 8\sqrt{2}\mathcal{M}^2_{1,\tau}C^{-2}_{1,\tau}\exp\left\{(\mathcal{M}_{1,\tau}+\mathcal{K}_{2,\tau})\lambda + 2\mathcal{M}_{2,\tau}\lambda^2\right\}\right) \\
&\cdot E_i\left[E_\zeta\left\{\rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right) - \rho_{\tau,a}\left(y_i - E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}\right] \\
:=& -\lambda E_i\left\{(\hat{W}_i - W_i)E_\zeta \rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} - \lambda E_i E_\zeta \rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right) \\
&+ \lambda^2(\overline{\mathcal{K}}_{1,\tau}+\overline{\mathcal{K}}_{2,\tau})E_i\left[E_\zeta\left\{\rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right) - \rho_{\tau,a}\left(y_i - E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}\right].
\end{aligned}
$$

where $\mathcal{M}_{1,\tau} = 4a_nA^{-1}_\pi C_{1,\tau}C_{2,\tau}$, $\mathcal{M}_{2,\tau} = 8e^2\mathcal{M}^2_{1,\tau}$, $\mathcal{K}_{1,\tau} = 0.5A^{-2}_\pi\overline{C}^2 + \mathcal{M}^2_{1,\tau}C^{-2}_{2,\tau}$ and $\mathcal{K}_{2,\tau} = A^{-1}_\pi\overline{C}C_{1,\tau} + 2a_nA^{-1}_\pi C^2_{1,\tau}$. Further, if λ also satisfies $\lambda^2(\overline{\mathcal{K}}_{1,\tau}+\overline{\mathcal{K}}_{2,\tau}) \le \lambda$, we have with probability one,

$$
\begin{aligned}
&E_i \log\left\{E_\zeta h^*_{\hat{W}_i}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \\
\le& -\lambda E_i\left\{\rho_{\tau,a}\left(y_i - E_\zeta\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} - \lambda E_i\left\{(\hat{W}_i - W_i)E_\zeta\rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}.
\end{aligned}
$$

Together with Jensen's inequality and $\rho_{\tau,a}(v) \ge \rho_\tau(v)$, above inequality is reformed as

$$
\begin{aligned}
E\log(1/p^n_N) =& -\sum_{i=N+1}^{n} E\left(E_i\log\left[E_\zeta\exp\left\{-\lambda\rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}\right]\right) \\
&\ge \lambda\sum_{i=N+1}^{n} E\rho_\tau\left(y_i - \sum_{(k,m)\in\hat{\mathcal{S}}_\tau}\hat{\Omega}_{(k,m),i}\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right) \\
&\quad + \lambda\sum_{i=N+1}^{n} E\left\{(\hat{W}_i - W_i)\sum_{(k,m)\in\hat{\mathcal{S}}_\tau}\hat{\Omega}_{(k,m),i}\rho_{\tau,a}\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\} \\
&\ge \lambda\sum_{i=N+1}^{n} E\rho_\tau\left(y_i - \sum_{(k,m)\in\hat{\mathcal{S}}_\tau}\hat{\Omega}_{(k,m),i}\hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right) \\
&\quad + \lambda\sum_{i=N+1}^{n} E\left\{(\hat{W}_i - W_i)\inf_{(k,m)\in\hat{\mathcal{S}}_\tau}\rho_\tau\left(y_i - \hat{Q}^{(k,m)}_{\tau,N}(y|\boldsymbol{x}_i)\right)\right\}.
\end{aligned}
$$

On the other hand, for $(k,m) \in \hat{\mathcal{S}}_\tau$,

$$-\log(p_N^n) \le \log(1/\omega_{(k,m)}) + \lambda \sum_{i=N+1}^{n} \hat{W}_i \rho_{\tau,a}\left(y_i - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_i)\right),$$

it follows that

$$\begin{aligned}\frac{1}{n-N} E\log(1/p_N^n) \le & \lambda E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) + \lambda E\left\{(\hat{W}-W)\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\} \\ & + \frac{\log(1/\omega_{(k,m)})}{(n-N)} + a_n\lambda(C_{1,\tau}^2 + 2C_{2,\tau}^2)\{1 + O(n^{-\alpha_2})\}.\end{aligned}$$

Using the convexity of the check loss, we have for any $(k,m) \in \hat{\mathcal{S}}_\tau$,

$$\begin{aligned}E\rho_\tau\left(y - \hat{Q}_\tau^{\text{mix}}(y|\boldsymbol{x})\right) &\le \frac{1}{n-N}\sum_{i=N+1}^{n} E\rho_\tau\left(y - \sum_{(k,m)\in\hat{\mathcal{S}}_\tau} \hat{\Omega}_{(k,m),i}\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) \\ &\le E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) + E\left\{(\hat{W}-W)\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\} \\ &\quad - E\left\{(\hat{W}-W)\inf_{(k,m)\in\hat{\mathcal{S}}_\tau}\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\} \\ &\quad + \frac{\log(1/\omega_{(k,m)})}{\lambda(n-N)} + a_n(C_{1,\tau}^2 + 2C_{2,\tau}^2)\left\{1 + O(n^{-\alpha_2})\right\}.\end{aligned}$$

The remaining is to determine the convergence rate of

$$I := E\left\{(\hat{W}-W)\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\} - E\left\{(\hat{W}-W)\inf_{(k,m)\in\hat{\mathcal{S}}_\tau}\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\}.$$

Note that

$$\begin{aligned}I &\le \left|E\left[(\hat{W}-W)\left\{\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) - \rho_\tau\left(y - Q_\tau(y|\boldsymbol{x})\right)\right\}\right]\right| \\ &\quad + \left|E\left[(\hat{W}-W)\left\{\inf_{(k,m)\in\hat{\mathcal{S}}_\tau}\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) - \rho_\tau\left(y - Q_\tau(y|\boldsymbol{x})\right)\right\}\right]\right| \\ &:= |I_1| + |I_2|.\end{aligned}$$

By Lagrange's mean theorem and (C.10),

$$
\begin{aligned}
& E \int_0^{\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})-Q_\tau(y|\boldsymbol{x})} \left[I(y \le Q_\tau(y|\boldsymbol{x}) + s) - I(y \le Q_\tau(y|\boldsymbol{x})) \right] \mathrm{d}s \\
&= O(1) E \left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x}) \right)^2,
\end{aligned}
$$

one can derive that

$$
\begin{aligned}
|I_1| &\le \left\{ E(\hat{W} - W)^2 \right\}^{1/2} \left[E \left\{ \rho_\tau \left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) \right) - \rho_\tau \left(y - Q_\tau(y|\boldsymbol{x}) \right) \right\}^2 \right]^{1/2} \\
&\le \left\{ E \left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x}) \right)^4 \right\}^{1/2} O(n^{-\alpha_2}) \\
&= O(n^{-\alpha_2 - 2\alpha_{km}}),
\end{aligned}
$$

and

$$
\begin{aligned}
|I_2| &\le \left\{ E(\hat{W} - W)^2 \right\}^{1/2} \left[E \inf_{(k,m) \in \hat{\mathcal{S}}_\tau} \left\{ \rho_\tau \left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) \right) - \rho_\tau \left(y - Q_\tau(y|\boldsymbol{x}) \right) \right\}^2 \right]^{1/2} \\
&\le \inf_{(k,m) \in \hat{\mathcal{S}}_\tau} O\left(n^{-\alpha_2 - 2\alpha_{km}} \right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
E\rho_\tau \left(y - \hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x}) \right) \le \inf_{(k,m) \in \hat{\mathcal{S}}_\tau} & \left\{ E\rho_\tau \left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) \right) + \frac{\log(1/\omega_{(k,m)})}{\lambda(n-N)} + O(n^{-\alpha_2 - 2\alpha_{km}}) \right\} \\
& + a_n (C_{1,\tau}^2 + 2C_{2,\tau}^2) \left\{ 1 + O(n^{-\alpha_2}) \right\}.
\end{aligned}
$$

The proof is completed. □

Derivation the inequality in Remark 2. For simplicity, We only care about the case that $B = 1$. Observed that

$$
E\rho_\tau \left(y - \hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x}) \right) \le \inf_{(k,m) \in \hat{S}_\tau} \left\{ E\rho_\tau \left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) \right) + \frac{\log(1/\omega_{(k,m)})}{\lambda(n-N)} + O(n^{-\alpha_2 - 2\alpha_{km}}) \right\} + O(a_n).
$$

Combined the identity of the quantile loss function and iterated expectation,

$$
\begin{aligned}
&E\int_0^{\hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})-Q_\tau(y|\boldsymbol{x})}[I(e_\tau\le s)-I(e_\tau\le 0)]\,\mathrm{d}s\\
&\le \inf_{(k,m)\in\hat{S}_\tau}\left\{E\int_0^{\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})-Q_\tau(y|\boldsymbol{x})}[I(e_\tau\le s)-I(e_\tau\le 0)]\,\mathrm{d}s+\frac{\log(1/\omega_{(k,m)})}{\lambda(n-N)}+O(n^{-\alpha_2-2\alpha_{km}})\right\}+O(a_n).
\end{aligned}
$$

Let ξ_x lie between $Q_\tau(y|\boldsymbol{x})$ and $Q_\tau(y|\boldsymbol{x})+s$, it follows that

$$
\begin{aligned}
&\frac{1}{2}E\left\{f_{y|\boldsymbol{x}}(\xi_x)\left(\hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})-Q_\tau(y|\boldsymbol{x})\right)^2\right\}\\
&\le \inf_{(k,m)\in\hat{S}_\tau}\left[\frac{1}{2}E\left\{f_{y|\boldsymbol{x}}(\xi_x)\left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})-Q_\tau(y|\boldsymbol{x})\right)^2\right\}+\frac{\log(1/\omega_{(k,m)})}{\lambda(n-N)}+O(n^{-\alpha_2-2\alpha_{km}})\right]+O(a_n).
\end{aligned}
$$

Therefore, the conclusion is established by taking $C=\max\{\sup_t f_{y|x}(t)/2,1\}\cdot(\inf_t f_{y|x}(t))^{-1}$. □

Lemma 5. *Under same conditions in Theorem 1, if* $\|Q_{\tau,N}^{(k_0,m_0)}(y|\boldsymbol{x})-Q_\tau(y|\boldsymbol{x})\|_{L_2}=0$ *for some* $(k_0,m_0)\in\hat{S}_\tau$*, we have for some constant* $c>0$ *and any* $\kappa>0$*,*

$$
\inf_{\substack{(k,m)\in\hat{S}_\tau\\1\le b\le B}}\left\{E\rho_\tau\left(y-\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x})\right)-cn^{-2\alpha_{km}-\kappa}\right\}\le E\rho_\tau\left(y-\hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})\right)
$$

Proof of Lemma 5. We first give the proof for $B=1$. Given $\kappa>0$, note that

$$
\begin{aligned}
&\inf_{(k,m)\in\hat{S}_\tau}E\rho_\tau\left(y-\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)-E\rho_\tau\left(y-\hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})\right)\\
&=\inf_{(k,m)\in\hat{S}_\tau}E\int_{\hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})-Q_\tau(y|\boldsymbol{x})}^{\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})-Q_\tau(y|\boldsymbol{x})}[I(e_\tau\le s)-I(e_\tau\le 0)]\,\mathrm{d}s\\
&\le\inf_{(k,m)\in\hat{S}_\tau}E\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})-\hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})\right|\\
&=\inf_{(k,m)\in\hat{S}_\tau}E\left\{\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})-\hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})\right|I\left(|e_\tau|\le n^{2\alpha_{km}+\kappa}|\hat{\mathcal{S}}_\tau|\right)\right\}\\
&\quad+\inf_{(k,m)\in\hat{S}_\tau}E\left\{\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})-\hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})\right|I\left(|e_\tau|\ge n^{2\alpha_{km}+\kappa}|\hat{\mathcal{S}}_\tau|\right)\right\}\\
&:=I_1+I_2,
\end{aligned}
$$

where the first inequality is obtained by $I(a) - I(b) \in [-1, 1]$. Note that

$$
\begin{aligned}
I_1 &\le E\left|\hat{Q}_{\tau,N}^{(k_0,m_0)}(y|\boldsymbol{x}) - \hat{Q}_{\tau}^{\text{mix}}(y|\boldsymbol{x})\right| \\
&\le E\left|\sum_{(k,m)\neq(k_0,m_0)} \hat{\Omega}_{(k,m)}\left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - \hat{Q}_{\tau,N}^{(k_0,m_0)}(y|\boldsymbol{x})\right)\right| \\
&\le 2E\left\{\sup_{(k,m)\in\hat{S}_\tau}\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right| \cdot \left(\sum_{(k,m)\neq(k_0,m_0)} \hat{\Omega}_{(k,m)}\right)\right\} \\
&\le 2C_{1,\tau} \cdot E\left(\sum_{(k,m)\neq(k_0,m_0)} \hat{\Omega}_{(k,m)}\right).
\end{aligned}
$$

Since the (k_0, m_0)-th candidate is correctly specified, and $\hat{\Omega}_{(k,m)}$ is uniformly bounded over $[0, 1]$. By adopting the same proof of Theorem 4 (p.19), the last inequality converges to 0.

Let $\alpha = \sup_{(k,m)\in\hat{S}_\tau} \alpha_{km}$. Together with Markov's inequality and Cauchy-Schwarz's inequality,

$$
\begin{aligned}
I_2 &\le \inf_{(k,m)\in\hat{S}_\tau}\left\{E\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - \hat{Q}_{\tau}^{\text{mix}}(y|\boldsymbol{x})\right|^2\right\}^{1/2}\left\{\Pr\left(|e_\tau| \ge n^{2\alpha+\kappa}|\hat{\mathcal{S}}_\tau|\right)\right\}^{1/2} \\
&\le \inf_{(k,m)\in\hat{S}_\tau}\left\{E\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - \hat{Q}_{\tau}^{\text{mix}}(y|\boldsymbol{x})\right|^2\right\}^{1/2} \cdot n^{-2\alpha-\kappa}|\hat{\mathcal{S}}_\tau|^{-1}\left\{E|e_\tau|\right\}^{1/2} \\
&\le \frac{\sqrt{8C_{2,\tau}}}{n^{2\alpha+\kappa}|\hat{\mathcal{S}}_\tau|}\left\{E\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right|^2 + E\left|\sum_{(k,m)\in\hat{S}_\tau} \hat{\Omega}_{(k,m)}\left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right)\right|^2\right\}^{1/2} \\
&\le \frac{\sqrt{8C_{2,\tau}}}{n^{2\alpha+\kappa}|\hat{\mathcal{S}}_\tau|}\left[E\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right|^2 + E\left(\sum_{(k,m)\in\hat{S}_\tau} \hat{\Omega}_{(k,m)}^2\right)\left\{\sum_{(k,m)\in\hat{S}_\tau}\left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right)^2\right\}\right]^{1/2} \\
&\le \frac{\sqrt{8C_{2,\tau}}}{n^{2\alpha+\kappa}|\hat{\mathcal{S}}_\tau|}\left\{E\sup_{(k,m)\in\hat{\mathcal{S}}_\tau}\left|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right|^2 + |\hat{\mathcal{S}}_\tau|^2 E\sup_{(k,m)\in\hat{\mathcal{S}}_\tau}\left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right)^2\right\}^{1/2} \\
&\le n^{-2\alpha-\kappa}\sqrt{8C_{2,\tau}(1 + |\hat{\mathcal{S}}_\tau|^{-2})}C_{1,\tau} := cn^{-2\alpha-\kappa}.
\end{aligned}
$$

This completes the proof. When $B \ge 1$, note that

$$
\inf_{\substack{(k,m)\in\hat{S}_\tau \\ 1\le b\le B}} E\left|\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x}) - \hat{Q}_{\tau}^{\text{mix}}(y|\boldsymbol{x})\right| \le E\left\{B^{-1}\sum_{b=1}^{B}\sum_{(k,m)\in\hat{S}_\tau} \hat{\Omega}_{(k,m)}^b\left|\hat{Q}_{\tau,N,b}^{(k,m)}(y|\boldsymbol{x}) - \hat{Q}_{\tau}^{\text{mix}}(y|\boldsymbol{x})\right|\right\}.
$$

Then the definition of $\hat{Q}_\tau^{\text{mix}}(y|\boldsymbol{x})$ directly fits above steps. □

Proof of Theorem 3. We care about $B = 1$ for simplicity. By the identity of $\rho_\tau(\cdot)$,

$$\inf_{(k,m)\in\hat{S}_\tau} E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) = \inf_{(k,m)\in\hat{S}_\tau} E\int_0^{\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})} [I(e_\tau \le s) - I(e_\tau \le 0)]\,\mathrm{d}s + E\rho_\tau(e_\tau)$$

$$\ge \inf_{(k,m)\in\hat{S}_\tau} \frac{f_{y|x}(\xi_x)}{2} \left\|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right\|_{L_2}^2.$$

Obviously, the order of $\|\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\|_{L_2}^2$ depends on whether $\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})$ is consistent to the true CQF. In this sense, the result of Theorem 3 requires the discussion on the candidate set, i.e., whether $\hat{S}_\tau$ contains the correct model.

When $\hat{S}_\tau$ does not contains the correct model, combining Theorem 1 of Angrist et al. (2006) with (C.7)–(C.10) we can see

$$\inf_{(k,m)\in\hat{S}_\tau} E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) = O(1).$$

Hence the result holds if $\log|\hat{S}_\tau|/\{\lambda(n-N)\} = o(1)$, which satisfies the constraint in (i) of Theorem 3.

When $\hat{S}_\tau$ contains the correct model, Combining Theorem 2 with Lemma 5, we have

$$\inf_{(k,m)\in\hat{S}_\tau} E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) - O(n^{-2\alpha_{k_0m_0}-\kappa})$$

$$\le E\rho_\tau\left(y - \hat{Q}_\tau^{\text{mix}}(y|\boldsymbol{x})\right) \le \inf_{(k,m)\in\hat{S}_\tau} \left\{E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\} + \frac{\log|\hat{S}_\tau|}{\lambda(n-N)} + O(n^{-\alpha_2-2\alpha_{k_0m_0}}) + O(a_n),$$

and

$$\inf_{(k,m)\in\hat{S}_\tau} E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) = O(n^{-2\alpha_{k_0m_0}}).$$

To ensure the excess risk shrinks to 0 as $n \to \infty$, it requires that

$$an^{2\alpha_{k_0m_0}} = o(1), \quad \text{and} \quad \frac{\log|\hat{S}_\tau|}{\lambda(n-N)} n^{2\alpha_{k_0m_0}} = o(1).$$

Hence (ii) of Theorem 3 is satisfied. Therefore, the proof is completed. □

Proof of Theorem 4. Note that $0 \le \hat{\Omega}_{(k,m)} \le 1$ for $(k,m) \in \mathcal{S}_{\tau,\mathrm{cor}}$. The result of the theorem is equivalent to that for any given $\varepsilon > 0$,

$$\Pr(-\log \hat{\Omega}_{\mathrm{cor}} \ge \varepsilon) \to 0, \quad \text{as } n \to \infty.$$

Since for $N+2 \le i \le n$,

$$\hat{\Omega}_{\mathrm{cor},i} = \sum_{(k,m)\in\mathcal{S}_{\tau,\mathrm{cor}}} \frac{\omega_{(k,m)} \exp\left\{-\lambda \sum_{l=N+1}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_l))\right\}}{\sum_{(k,m)\in\hat{\mathcal{S}}_\tau} \omega_{(k,m)} \exp\left\{-\lambda \sum_{l=N+1}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_l))\right\}},$$

we have

$$\begin{aligned}\hat{\Omega}_{\mathrm{cor}} =& (n-N)^{-1} \sum_{i=N+2}^{n} \frac{\sum_{(k,m)\in\mathcal{S}_{\tau,\mathrm{cor}}} \omega_{(k,m)} \exp\left\{-\lambda \sum_{l=N+1}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_l))\right\}}{\sum_{(k,m)\in\hat{\mathcal{S}}_\tau} \omega_{(k,m)} \exp\left\{-\lambda \sum_{l=N+1}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_l))\right\}} \\ &+ \frac{\omega_{(k,m)}|\mathcal{S}_{\tau,\mathrm{cor}}|}{n-N}.\end{aligned}$$

Moreover, by the convexity of logarithmic function, it follows that

$$\begin{aligned}\Pr\left(-\log \hat{\Omega}_{\mathrm{cor}} \ge \varepsilon\right) \le& \frac{1}{(n-N)\varepsilon} \left(\sum_{i=N+2}^{n} E\left[\log \sum_{(k,m)\in\hat{\mathcal{S}}_\tau} \omega_{(k,m)} \exp\left\{-\lambda \sum_{l=N+1}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_l))\right\}\right]\right) \\ &- \frac{1}{(n-N)\varepsilon} \left(\sum_{i=N+2}^{n} E\left[\log \sum_{(k,m)\in\mathcal{S}_{\tau,\mathrm{cor}}} \omega_{(k,m)} \exp\left\{-\lambda \sum_{l=N+1}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_l))\right\}\right]\right) \\ &- \frac{1}{(n-N)\varepsilon} \log\left(\omega_{(k,m)}|\mathcal{S}_{\tau,\mathrm{cor}}|\right).\end{aligned}$$

According to the proof of Theorem 2, one can verify that for either $\mathcal{S} = \hat{\mathcal{S}}_\tau$ or $\mathcal{S}_{\tau,\mathrm{cor}}$,

$$\begin{aligned}&\sum_{i=N+2}^{n} \log \sum_{(k,m)\in\mathcal{S}} \omega_{(k,m)} \exp\left\{-\lambda \sum_{l=N+1}^{i-1} \hat{W}_l \rho_{\tau,a}(y_l - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_l))\right\} \\ =& \sum_{i=N+1}^{n} \log \sum_{(k,m)\in\mathcal{S}} \hat{\Omega}_{(k,m),i} \exp\left\{-\lambda \hat{W}_i \rho_{\tau,a}(y_i - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_i))\right\},\end{aligned}$$

hence for $(k,m) \in \mathcal{S}_{\tau,\mathrm{cor}}$,

$$\begin{aligned}\Pr\left(-\log\hat{\Omega}_{\mathrm{cor}} \geq \varepsilon\right) &\leq \frac{\lambda}{\varepsilon} E\left\{\hat{W}\rho_{\tau,a}\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\} + \frac{\log|\mathcal{S}_{\tau,\mathrm{cor}}| + \log(\omega_{(k,m)}|\mathcal{S}_{\tau,\mathrm{cor}}|)}{(n-N)\varepsilon} \\ &\quad - \frac{\lambda}{(n-N)\varepsilon}\sum_{i=N+1}^{n} E\rho_\tau\left(y - \sum_{m\in\hat{\mathcal{S}}_\tau}\hat{\Omega}_{(k,m),i}\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}_i)\right) + \varepsilon^{-1}O(n^{-\alpha_2-\alpha_{(k,m)}}) \\ &\leq \frac{\lambda}{\varepsilon} E\left\{\hat{W}\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) - \rho_\tau\left(y - \hat{Q}_\tau^{\mathrm{mix}}(y|\boldsymbol{x})\right)\right\} + o(1).\end{aligned}$$

Therefore, for any $(\widetilde{k},\widetilde{m}) \in \hat{\mathcal{S}}_\tau$,

$$0 \leq \Pr\left(-\log\hat{\Omega}_{\mathrm{cor}} \geq \varepsilon\right) \leq \frac{\lambda}{\varepsilon}\left[E\left\{\hat{W}\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\} - \inf_{(\widetilde{k},\widetilde{m})\in\hat{\mathcal{S}}_\tau} E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(\widetilde{k},\widetilde{m})}(y|\boldsymbol{x})\right)\right] + o(1).$$

The remaining task is to demonstrate that the right side of above inequality is $o(1)$. In fact, it is obviously shown that by Theorem 1 (ii) and Lemma 3, $\hat{\mathcal{S}}_\tau$ contains at least one candidate such that $Q_\tau^{(k,m)}(y|\boldsymbol{x}) = q_\tau(\boldsymbol{u}) + \boldsymbol{\beta}_{\tau,1}^\top \boldsymbol{z}_1$ in probability. Denote $e_{\tau,1} = y - q_\tau(\boldsymbol{u}) - \boldsymbol{\beta}_{\tau,1}^\top \boldsymbol{z}_1$, combining (C.10) with the argument of Lemma 2 and Theorem 2(c) of Ferguson (1996), we have

$$\begin{aligned}&\left|E\left\{\hat{W}\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right)\right\} - \inf_{(\widetilde{k},\widetilde{m})\in\hat{\mathcal{S}}_\tau} E\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(\widetilde{k},\widetilde{m})}(y|\boldsymbol{x})\right)\right| \\ &\leq \left|E(\hat{W} - W)\left\{\rho_\tau\left(y - \hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x})\right) - \rho_\tau\left(e_{\tau,1}\right)\right\}\right| \\ &\quad + \left|E\left\{\rho_\tau\left(e_{\tau,1} - \left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right)\right) - \rho_\tau\left(e_{\tau,1}\right)\right\}\right| \\ &\quad + \left|\inf_{(\widetilde{k},\widetilde{m})\in\hat{\mathcal{S}}_\tau} E\left\{\rho_\tau\left(e_{\tau,1} - \left(\hat{Q}_{\tau,N}^{(\widetilde{k},\widetilde{m})}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right)\right) - \rho_\tau\left(e_{\tau,1}\right)\right\}\right| \\ &\leq O(n^{-\alpha_2-\alpha_{km}}) + \left\{E\left(\hat{Q}_{\tau,N}^{(k,m)}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right)^2\right\}^{1/2} \\ &\quad + \inf_{(\widetilde{k},\widetilde{m})\in\hat{\mathcal{S}}_\tau}\left\{E\left(\hat{Q}_{\tau,N}^{(\widetilde{k},\widetilde{m})}(y|\boldsymbol{x}) - Q_\tau(y|\boldsymbol{x})\right)^2\right\}^{1/2} \\ &= o(1),\end{aligned}$$

thus the proof is completed. □

S1.3 Proof in Section 3

Proof of Corollary 1. Let $k(\boldsymbol{x}) = 1/\pi(\boldsymbol{x})$,

$$\widetilde{f}_{\bar{X}}(\bar{\boldsymbol{x}}) = \frac{f_{\bar{X}}(\bar{\boldsymbol{x}})\omega(\bar{\boldsymbol{x}})f^{-1}_{\varepsilon|\boldsymbol{x}}(0)}{\int_{v\in\bar{\mathbb{X}}} f_{\bar{X}}(v)\omega(v)f^{-1}_{\varepsilon|\boldsymbol{x}}(0)\mathrm{d}v},$$

and

$$\widetilde{\omega}(\bar{\boldsymbol{x}}) = \frac{\omega(\bar{\boldsymbol{x}})f^{-1}_{\varepsilon|\boldsymbol{x}}(0)}{\int_{v\in\bar{\mathbb{X}}} f_{\bar{X}}(v)\omega(v)f^{-1}_{\varepsilon|\boldsymbol{x}}(0)\mathrm{d}v}.$$

One can define a random variable $\widetilde{\boldsymbol{x}}$ on $\bar{\mathbb{X}}$ with the density $\widetilde{f}_{\bar{X}}$, such that

$$E\left\{\widetilde{\omega}(\bar{\boldsymbol{x}})I(\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}_S)\right\} = \int_{\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}_S} \frac{\omega(\bar{\boldsymbol{x}})f^{-1}_{\varepsilon|\boldsymbol{x}}(0)}{\int_{v\in\bar{\mathbb{X}}} f_{\bar{X}}(v)\omega(v)f^{-1}_{\varepsilon|\boldsymbol{x}}(0)\mathrm{d}v} f_{\bar{X}}(\bar{\boldsymbol{x}})\mathrm{d}\bar{\boldsymbol{x}}$$

$$= \Pr(\widetilde{\boldsymbol{x}}\in\bar{\mathbb{X}}_S) := \widetilde{q}(\bar{\mathbb{X}}_S).$$

Since $\int_{v\in\bar{\mathbb{X}}} f_{\bar{X}}(v)\omega(v)f^{-1}_{\varepsilon|\boldsymbol{x}}(0)\mathrm{d}v$ is nonstochastic and is bounded away from 0 and infinity, it can be regarded as a constant C_ω that will not affect the extremum of the target variance by multiplying a specific probability level $\max\{\tau^2,(1-\tau)^2\}$. Therefore, the question is equivalent to minimizing

$$\begin{aligned}
\boldsymbol{V}^*_\omega(\bar{\mathbb{X}}_S) &= \frac{1}{\widetilde{q}^2(\bar{\mathbb{X}}_S)} E\{\widetilde{\omega}^2(\bar{\boldsymbol{x}})I(\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}_S)k(\boldsymbol{x})\} \\
&= \frac{1}{\widetilde{q}^2(\bar{\mathbb{X}}_S)} \int_{\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}_S} \widetilde{\omega}(\bar{\boldsymbol{x}})k(\boldsymbol{x}) \frac{\omega(\bar{\boldsymbol{x}})f^{-1}_{\varepsilon|\boldsymbol{x}}(0)}{\int_{v\in\bar{\mathbb{X}}} f_{\bar{X}}(v)\omega(v)f^{-1}_{\varepsilon|\boldsymbol{x}}(0)\mathrm{d}v} f_{\bar{X}}(\bar{\boldsymbol{x}})\mathrm{d}\bar{\boldsymbol{x}} \\
&= \frac{1}{\widetilde{q}(\bar{\mathbb{X}}_S)} E\left\{\widetilde{\omega}(\widetilde{\boldsymbol{x}})k(\widetilde{\boldsymbol{x}})|I(\widetilde{\boldsymbol{x}}\in\bar{\mathbb{X}}_S)\right\}.
\end{aligned}$$

Following the argument of Theorem A1 in Crump et al. (2009), the remainder is to calculate $\delta > 0$ to obtain the optimal subspace

$$\bar{\mathbb{X}}_S = \{\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}|\widetilde{\omega}(\bar{\boldsymbol{x}})k(\boldsymbol{x}) \le \delta\}.$$

Define a random variable $u = \widetilde{\omega}(\widetilde{\boldsymbol{x}})k(\widetilde{\boldsymbol{x}})$ with the density f_u, the minimum of

$$\boldsymbol{V}_{\omega}^{*}(\bar{\mathbb{X}}_S) = \frac{E\left\{uI(u \in \bar{\mathbb{X}}_S)\right\}}{\Pr^2(u \in \bar{\mathbb{X}}_S)} = \frac{\int_0^{\delta} u f_u(u)\mathrm{d}u}{\left\{\int_0^{\delta} f_u(u)\mathrm{d}u\right\}^2}$$

is achieved at either $\delta = \sup_{\boldsymbol{x}\in\bar{\mathbb{X}}} k(\boldsymbol{x})$ or $\delta = 2E\{\widetilde{\omega}(\widetilde{\boldsymbol{x}})k(\widetilde{\boldsymbol{x}})|\widetilde{\omega}(\widetilde{\boldsymbol{x}})k(\widetilde{\boldsymbol{x}}) < \delta\}$. Recall the definition of $\widetilde{\boldsymbol{x}}$ and $\widetilde{\omega}(\bar{\boldsymbol{x}})$, let $\gamma = \delta C_{\omega}$,

$$\frac{\gamma}{C_{\omega}} = 2\frac{E\left\{C_{\omega}^{-1}\omega(\widetilde{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0)k(\widetilde{\boldsymbol{x}}) \cdot I\left(C_{\omega}^{-1}\omega(\widetilde{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0)k(\widetilde{\boldsymbol{x}}) < \gamma C_{\omega}^{-1}\right)\right\}}{E\left\{I\left(C_{\omega}^{-1}\omega(\widetilde{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0)k(\widetilde{\boldsymbol{x}}) < \gamma C_{\omega}^{-1}\right)\right\}},$$

it follows that

$$\begin{aligned}
\gamma &= 2\frac{\int_{\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}} \omega^2(\bar{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-2}(0)k(\boldsymbol{x})I\left(\omega(\bar{\boldsymbol{x}})k(\boldsymbol{x}) < \gamma\right) f_{\bar{X}}(\bar{\boldsymbol{x}})\mathrm{d}\bar{\boldsymbol{x}}}{\int_{\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}} \omega(\bar{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0)I\left(\omega(\bar{\boldsymbol{x}})k(\boldsymbol{x}) < \gamma\right) f_{\bar{X}}(\bar{\boldsymbol{x}})\mathrm{d}\bar{\boldsymbol{x}}} \\
&\quad \times \frac{\left\{\int_{\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}} I\left(\omega(\bar{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0)k(\boldsymbol{x}) < \gamma\right) f_{\bar{X}}(\bar{\boldsymbol{x}})\mathrm{d}\bar{\boldsymbol{x}}\right\}^{-1}}{\left\{\int_{\bar{\boldsymbol{x}}\in\bar{\mathbb{X}}} I\left(\omega(\bar{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0)k(\boldsymbol{x}) < \gamma\right) f_{\bar{X}}(\bar{\boldsymbol{x}})\mathrm{d}\bar{\boldsymbol{x}}\right\}^{-1}} \\
&= 2\frac{E\left\{\omega^2(\bar{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-2}(0)k(\boldsymbol{x})\middle|\omega(\bar{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0)k(\boldsymbol{x}) < \gamma\right\}}{E\left\{\omega(\bar{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0)\middle|\omega(\bar{\boldsymbol{x}})f_{\varepsilon|\boldsymbol{x}}^{-1}(0) < \gamma\right\}}.
\end{aligned}$$

The proof is completed. □

S2 Further discussions on regularity conditions

We take account of the order in condition (C.2), and further explain (C.4) and the constraint in Theorem 1 (ii).

S2.1 Discussion on condition (C.2)

To illustrate a specific order in (C.2), we consider the MAR response mechanism with respect to a logistic regressive model:

$$\pi(\boldsymbol{x}, y) = \pi(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{x})}{1 + \exp(\boldsymbol{\theta}^\top \boldsymbol{x})}.$$

Following (3.3) of Wang (2011), it implies $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| = O_p(\sqrt{p/n})$. By the Lipschitz continuity of $\pi(\boldsymbol{x}; \boldsymbol{\theta})$, common regularities (e.g., $0 < \lambda_{\min}(E\boldsymbol{x}\boldsymbol{x}^\top) \le \lambda_{\max}(E\boldsymbol{x}\boldsymbol{x}^\top) < \infty$; the parameter space is compact) follows that for both $\underline{c}$, $\overline{c} > 0$,

$$\underline{c} \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \right\|_{L_2} \le \|\pi(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_n) - \pi(\boldsymbol{x}; \boldsymbol{\theta})\|_{L_2} \le \overline{c} \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \right\|_{L_2}.$$

When p is fixed, the right side of above inequality is obviously controlled by $C_2 n^{-1/2}$. The left side of above inequality is bounded below by $C_1 n^{-1/2}$ (Theorem 8.11 of van der Vaart (2000)) as well. Consequently, $\alpha_2 = 1/2$ as the minimax optimal order of the response probability estimator. Moreover, for nonlinear specifications and the MNAR mechanism, it always has an order that $0 < \alpha_2 < 1/2$ (e.g., Tsybakov (2009); Shan, Li and Ai (2024)).

S2.2 Discussion on condition (C.4)

we first consider the case that Δ contains the consistent estimation, that is, $\|\hat{q}_\tau^{(k)}(\boldsymbol{u}) - q_\tau(\boldsymbol{u})\|_{L_2} = o(1)$ for some $\delta_k \in \Delta$. To ensure the screening consistency, the identifying constraint is as Wang et al. (2023) that

$$\left(\min_{j \in \mathcal{I}_\tau} - \max_{j \in \mathcal{I}_\tau^c} \right) n^{-1} \sum_{i=1}^{n} |\beta_{\tau,j,0} z_{ij} - \xi_\tau| > 0. \tag{S2.1}$$

In fact, above condition directly implies (C.4) because $\beta_{\tau,j,0} = \beta^{(k)}_{\tau,j,0}$, $\xi^{(k)}_\tau = \xi_\tau$ and the left side of (S2.1) is bounded by

$$\left(\sup_{\delta_k\in\Delta}\min_{j\in\mathcal{I}_\tau} - \inf_{\delta_k\in\Delta}\max_{j\in\mathcal{I}^c_\tau}\right) n^{-1}\sum_{i=1}^n \left|\beta^{(k)}_{\tau,j,0}z_{ij} - \xi^{(k)}_\tau\right|.$$

When Δ are all misspecified, note that

$$\min_{j\in\mathcal{I}_\tau} n^{-1}\sum_{i=1}^n |\beta_{\tau,j,0}z_{ij} - \xi_\tau| \le \min_{j\in\mathcal{I}_\tau} n^{-1}\sum_{i=1}^n \left|\beta^{(k)}_{\tau,j,0}z_{ij} - \xi^{(k)}_\tau\right| + \max_{1\le j\le p}\left(\left|\beta^{(k)}_{\tau,j,0} - \beta_{\tau,j,0}\right| n^{-1}\sum_{i=1}^n |z_{ij}| + \left|\xi^{(k)}_\tau - \xi_\tau\right|\right),$$

$$\max_{j\in\mathcal{I}^c_\tau} n^{-1}\sum_{i=1}^n \left|\beta^{(k)}_{\tau,j,0}z_{ij} - \xi^{(k)}_\tau\right| \le \max_{j\in\mathcal{I}^c_\tau} n^{-1}\sum_{i=1}^n |\beta_{\tau,j,0}z_{ij} - \xi_\tau| + \max_{1\le j\le p}\left(\left|\beta^{(k)}_{\tau,j,0} - \beta_{\tau,j,0}\right| n^{-1}\sum_{i=1}^n |z_{ij}| + \left|\xi^{(k)}_\tau - \xi_\tau\right|\right).$$

In this case,

$$\left(\min_{j\in\mathcal{I}_\tau} - \max_{j\in\mathcal{I}^c_\tau}\right) n^{-1}\sum_{i=1}^n |\beta_{\tau,j,0}z_{ij} - \xi_\tau| \le \left(\min_{j\in\mathcal{I}_\tau} - \max_{j\in\mathcal{I}^c_\tau}\right) n^{-1}\sum_{i=1}^n \left|\beta^{(k)}_{\tau,j,0}z_{ij} - \xi^{(k)}_\tau\right| + C_n,$$

which implies (C.4) is rational when the signal strength of significant predictors exceeds the deviation introduced by pseudo parameters. Therefore, (C.4) is regarded as a generalization of (S2.1) for multiple model specifications.

Discussion on the additional constraint in Thoerem 1 (ii)

Thoerem 1 (ii) imposes two additional conditions: (a) $\bar{q}^{(k)}_\tau(\boldsymbol{u}) = q_\tau(\boldsymbol{u})$ for some $\delta_k \in \Delta$; (b) $\min_{j\in\mathcal{I}_\tau} \|f_j(\tau)\|_{1,n} \ge v_n$. The latter is common for identifying the marginal quantile utility, serves as one theoretical basis for consistently screening significant predictors (Fan and Song (2010)). For (a), note that $\bar{q}^{(k)}_\tau(\boldsymbol{u})$ is the pseudo true model of $q_\tau(\boldsymbol{u})$. The constraint is close to standard assumptions in doubly robust inference (Han et al. (2019), Li, Gu and Liu (2020)). To care the existence in our model framework, we assume a linear specification that $q^{(k)}_\tau(\boldsymbol{u}) = \boldsymbol{a}^\top\boldsymbol{u}$. The parameter is estimated as

$$\hat{\boldsymbol{\alpha}}_n = \underset{(\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)})}{\arg\min}\, n^{-1}\sum_{i=1}^n W_i\rho_\tau\left(y_i - \boldsymbol{a}^\top\boldsymbol{u}_i - \boldsymbol{\beta}^\top_{(k)}\boldsymbol{z}_{(k)i}\right),$$

where $\boldsymbol{z}_{(k)i}$ is a prior specification of $\boldsymbol{z}_i$ through δ_k. Following the proof of Lemma 1 and

Angrist et al. (2006), one can show that $\bar{q}_\tau^{(k)}(\boldsymbol{u}) = \bar{\boldsymbol{\alpha}}^\top \boldsymbol{u}$ with

$$\bar{\boldsymbol{\alpha}} = \underset{(\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)})}{\arg\min}\, E\left\{ w_\tau(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) \left(\boldsymbol{\alpha}^\top \boldsymbol{u} + \boldsymbol{\beta}_{(k)}^\top \boldsymbol{z}_{(k)} - q_\tau(\boldsymbol{u}) - \boldsymbol{\beta}_{\text{true}} \boldsymbol{z}_{\text{true}} \right)^2 \right\}, \tag{S2.2}$$

where

$$w_\tau(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) = \int_0^1 (1-t) f_{y|\boldsymbol{x}} \left(t\boldsymbol{\alpha}^\top \boldsymbol{u} + (1-t) q_\tau(\boldsymbol{u}) + t\boldsymbol{\beta}_{(k)}^\top \boldsymbol{z}_{(k)} + (1-t)\boldsymbol{\beta}_{\text{true}} \boldsymbol{z}_{\text{true}} \right) \mathrm{d}t.$$

Note that (S2.2) is equivalent to minimizing

$$E\rho_\tau \left(y - \boldsymbol{a}^\top \boldsymbol{u} - \boldsymbol{\beta}_{(k)}^\top \boldsymbol{z}_{(k)} \right),$$

which has the unique minimum on the parameter space. It can be seen that when $q_\tau(\boldsymbol{u}) = \boldsymbol{\alpha}_{\text{true}}^\top \boldsymbol{u}$, the score function satisfies

$$E\left\{ R_\tau^2(\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) \frac{\partial w_\tau(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)})}{\partial \boldsymbol{\alpha}} + 2R_\tau(\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) w_\tau(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) \boldsymbol{u} \right\} \Big|_{(\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\beta}}_{(k)})} = \boldsymbol{0},$$

$$E\left\{ R_\tau^2(\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) \frac{\partial w_\tau(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)})}{\partial \boldsymbol{\beta}_{(k)}} + 2R(\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) w_\tau(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) \boldsymbol{z}_{(k)} \right\} \Big|_{(\bar{\boldsymbol{\alpha}},\bar{\boldsymbol{\beta}}_{(k)})} = \boldsymbol{0},$$

where $R_\tau(\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) = (\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\text{true}})^\top \boldsymbol{u} + \boldsymbol{\beta}_{(k)}^\top \boldsymbol{z}_{(k)} - \boldsymbol{\beta}_{\text{true}}^\top \boldsymbol{z}_{\text{true}}$. Of this view, a natural solution is $(\bar{\boldsymbol{\alpha}}^\top, \bar{\boldsymbol{\beta}}_{(k)}^\top)^\top = (\boldsymbol{\alpha}_{\text{true}}^\top, \boldsymbol{\beta}_{\text{true}}^\top)^\top$ to ensure $R_\tau(\boldsymbol{\alpha},\boldsymbol{\beta}_{(k)}) = 0$ almost surely. Therefore, condition (b) implies that δ_k ensures both $q_\tau^{(k)}(\boldsymbol{u})$ and $\boldsymbol{\beta}_{(k)}^\top \boldsymbol{z}$ are correctly specified. In practice, this can be implemented via prior information to include as many potentially significant factors as possible.

S3 Relationship between asymptotic risk optimalities

In this part, we compare Theorem 3 to classical asymptotic risk optimalities such as Theorem 3.3 in Lu and Su (2015), Corollary 1 in Xu, Zhang and Chiou (2023), and Theorem 2 in Xiong,

Deng and Wang (2025). To distinguish the methods, We uniformly summarize competitors as **"weight-optimization" model average (WOMA)**[1]. For ease of description, we set $\mu = Q_\tau(y|\boldsymbol{x})$, $\mu_{(m)} = Q_\tau^{(m)}(y|\boldsymbol{x})$, $\hat{\mu}_{(m)N} = \hat{Q}_{\tau,N}^{(m)}(y|\boldsymbol{x})$, and replace the candidate set as $\mathcal{M} = \{\mu_{(m)} : m = 1, ..., M_n\}$. Let $\hat{\mu}_{(m)n}$ be the CQF estimator of the m-th candidate corresponding to $\mathcal{D}_n = \{(\boldsymbol{x}_i^\top, y_i, r_i)^\top : i = 1, ..., n\}$, $\hat{\mu}_n(\boldsymbol{w}) = \sum_{m=1}^{M_n} w_m \hat{\mu}_{(m)n}$, and $\xi_n(\boldsymbol{w}) = E\{y - \hat{\mu}_n(\boldsymbol{w})|\mathcal{D}_n\}$ be the out-of-sample prediction risk. The asymptotic risk optimality of WOMA is

$$\frac{\xi_n(\hat{\boldsymbol{w}})}{\inf_{w\in\mathcal{W}} \xi_n(\boldsymbol{w})} = 1 + o_p(1), \tag{S3.1}$$

where the pair $(\boldsymbol{x}^\top, y)^\top$ in (S3.1) is an independent copy of those in $\mathcal{D}_n$.

Without loss of generality, we only focus on the case of $B = 1$. As shown in Lu and Su (2015), Wang et al. (2023), common conditions induce that

$$\inf_{\boldsymbol{w}\in\mathcal{W}} \xi_n(\boldsymbol{w}) \geq E\rho_\tau(e_\tau) - o_p(1).$$

Therefore, two types of asymptotic optimalities are theoretically equivalent if

$$\left| \inf_{m\in\mathcal{M}} \frac{E\rho_\tau(y - \hat{\mu}_{(m)N})}{\inf_{\boldsymbol{w}\in\mathcal{W}} \xi_n(\boldsymbol{w})} - 1 \right| = o_p(1). \tag{S3.2}$$

Since the numerator is the optimal risk of a single candidate, it can be regarded as a weighted estimator such that

$$\inf_{m\in\mathcal{M}} E\rho_\tau(y - \hat{\mu}_{(m)N}) = \xi_n(\boldsymbol{w}_*) + o(1),$$

where $\boldsymbol{w}_* \in \mathcal{W}$ is the weight vector such that its M_0-th element is 1 and all other elements are 0, and the M_0-th candidate has the infimum $E\rho_\tau(y - \hat{\mu}_{(m)N})$. Let $\boldsymbol{w}_0 = \arg\min_{\boldsymbol{w}\in\mathcal{M}} \xi_n(\boldsymbol{w})$, then (S3.2) holds provided that

$$\Pr(\boldsymbol{w}_* = \boldsymbol{w}_0) \to 1.$$

[1] This type of model averaging is named by the form $\hat{\boldsymbol{w}} = \arg\min_{m\in\mathcal{M}} \hat{L}_n(\boldsymbol{w})$, where $\hat{L}_n(\boldsymbol{w})$ incorporates quantile loss with candidate estimators (Yu, Zhang and Liang (2025)). In the area of missing data, it could be rewritten as $\hat{L}_n(\boldsymbol{w}; \hat{\pi})$ via inverse probability weighting.

Unfortunately, above convergence does not hold when $\mathcal{M}$ is rather general. However, for a fully linear CQF (i.e., $\mu = \boldsymbol{\beta}^\top \boldsymbol{x}$), the result can be established if WOMA is parsimonious[2]. This frequently restrict $\mathcal{M}$ to be a nested set that contains all significant predictors.

Now return to a general specification of μ, the natural question is whether we can relax the nested requirement for $\mathcal{M}$. The answer is **Yes**. Inspired by condition (C.7) (i) and Lemma 3, we can see that $\hat{\mu}_{(m)n}$ is L_2 consistent to μ as $\mu_{(m)}$ is correctly specified. Since such specification may not be exclusive, $\boldsymbol{w}_0$ should be respectively expanded to including all correct models.

Of this view, we give the following key conditions:

(S.1) $\sup_{w \in \mathcal{W}} \left| \hat{\xi}_n(\boldsymbol{w}) - \xi_n(\boldsymbol{w}) \right| = 1 + o_p(1)$.

(S.2) As $n \to \infty$, the weight estimator of WOMA is over-consistency, i.e.,

$$\sum_{m \in \mathcal{M}_T} \hat{w}_m \xrightarrow{p} 1,$$

where $\mathcal{M}_T$ is the set of correct models in $\mathcal{M}$.

(S.3) For some constant c, $E\rho_\tau(e_\tau) \geq c$ almost surely.

Condition (S.1) provides a fundamental theoretical medium for validating (S3.1). This condition incorporates essential regularities such as $\|\mu\|_{L_4} = O(1)$, the uniform boundedness of $f_{y|\boldsymbol{x}}(\cdot)$, constraints on the order of M_n and the dimension of candidates. Condition (S.2) is weaker than the parsimony but another important conclusion in the study of WOMA. It demonstrates that the weight estimator asymptotically assigns to all correct candidates (including overfitted models), further ensures the estimator possesses large sample properties that are asymptotically equivalent to the oracle specification. Both (S.1) and (S.2) have emerged as central objects for related inference, which have been extensively discussed in

[2] In the linear specification, "parsimony" demonstrates that the weight corresponding to the minimal-overfitting model is consistent to 1. Such procedure consistently assigns maximal weight to the candidate estimator with minimum out-of-sample prediction risk. See Zhang et al. (2020), Xu and Zhang (2024) for detailed descriptions.

the past decade (Lu and Su (2015), Zhang et al. (2020), Chen et al. (2023), Wang et al. (2023), Xu, Zhang and Chiou (2023), Xiong, Deng and Wang (2025), etc.). Condition (S.3) excludes scenarios in which $n^{1-\alpha_{km}}/\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})$ is not convergence to zero. It further ensures that (S3.1) remains well-defined when the candidate set contains the correct model (Yu, Zhang and Liang (2025)). Combining with Theorem 3, the following result is established:

Theorem S1. *Under conditions* (C.1)–(C.10)*, and* (S.1)–(S.3)*. Suppose $\mathcal{M}$ contains at least one correct model, and (ii) of Theorem 3 holds. Then our adaptive estimator is optimal in the sense that its global risk is asymptotically identical to out-of-sample prediction risk of the best-but-infeasible WOMA estimator, i.e.,*

$$\frac{E\rho_\tau\left(y-\hat{\mu}_{\text{mix}}\right)}{\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}=1+o_p(1).$$

Proof of Theorem S1. Let $\boldsymbol{w}_1$ be the limit weight vector of $\hat{\boldsymbol{w}}$ in (S.1), that is, the sum of elements corresponding to all candidates in $\mathcal{M}_T$ is 1. Note that

$$\frac{E\rho_\tau\left(y-\hat{\mu}_{\text{mix}}\right)}{\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}=\frac{E\rho_\tau\left(y-\hat{\mu}_{\text{mix}}\right)}{\inf_{\boldsymbol{m}\in\mathcal{M}}E\rho_\tau\left(y-\hat{\mu}_{(m)N}\right)}\cdot\frac{\inf_{\boldsymbol{m}\in\mathcal{M}}E\rho_\tau\left(y-\hat{\mu}_{(m)N}\right)}{\xi_n(\boldsymbol{w}_1)}\cdot\frac{\xi_n(\boldsymbol{w}_1)}{\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}. \quad \text{(S3.3)}$$

Since $\mathcal{M}_T\neq\varnothing$ fits (ii) of Theorem 3, the first term on the right side of (S3.3) converges to 1.

On the other side,

$$\begin{aligned}\left|\frac{\xi_n(\boldsymbol{w}_1)-\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}{\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}\right|&\leq\frac{\xi_n(\boldsymbol{w}_1)-\hat{\xi}_n(\hat{\boldsymbol{w}})+\inf_{\boldsymbol{w}\in\mathcal{W}}\hat{\xi}_n(\boldsymbol{w})-\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}{\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}\\&\leq\frac{\xi_n(\boldsymbol{w}_1)-\hat{\xi}_n(\boldsymbol{w}_1)}{\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}+\frac{\hat{\xi}_n(\boldsymbol{w}_1)-\hat{\xi}_n(\hat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})}+\frac{\hat{\xi}_n(\boldsymbol{w}_1)-\xi_n(\boldsymbol{w}_1)+r_n}{\inf_{\boldsymbol{w}\in\mathcal{W}}\xi_n(\boldsymbol{w})},\end{aligned} \quad \text{(S3.4)}$$

where r_n is a non-negative sequence (may depend on τ) such that r_n tends to 0 as $n\to\infty$. The existence of r_n is natural and by Lemma 1.7.1 of Chang and Shi (2003), and r_n has a non-negative decreasing subsequence with a corresponding weight vector subsequence. From

(S.2) and the convexity of the quantile loss, we have

$$\Pr\left(\left|\hat{\xi}_n(\boldsymbol{w}_1) - \hat{\xi}_n(\hat{\boldsymbol{w}})\right| > 0\right) \to 0.$$

Thus (S3.4) is $o_p(1)$ by (S.1). It follows that the last term on the right side of (S3.3) converges to 1 in probability.

Finally, we consider the asymptotic performance of

$$R_n(\boldsymbol{w}_1) := \left|\inf_{\boldsymbol{m}\in\mathcal{M}} E\rho_\tau\left(y - \hat{\mu}_{(m)N}\right) - \xi_n(\boldsymbol{w}_1)\right|.$$

For any given $\varepsilon > 0$,

$$\begin{aligned}\Pr\left(R_n(\boldsymbol{w}_1) \geq 4\varepsilon\bar{C}^{-1}\right) &\leq \Pr\left[\inf_{\boldsymbol{m}\in\mathcal{M}} E\left\{\rho_\tau\left(y - \hat{\mu}_{(m)N}\right) - \rho_\tau(e_\tau)\right\} \geq 2\varepsilon\bar{C}^{-1}\right] \\ &\quad + \Pr\left\{\left|\xi_n(\boldsymbol{w}_1) - E\rho_\tau(e_\tau)\right| \geq 2\varepsilon\bar{C}^{-1}\right\} \\ &:= PR_{1n} + PR_{2n}.\end{aligned}$$

By the identity of Knight (1998), the law of iterated expectations, and mean-value theorem,

$$\begin{aligned}PR_{1n} &= \Pr\left\{\inf_{m\in\mathcal{M}} E\int_0^{\hat{\mu}_{(m)N}-\mu}\left[I\left(y \leq \mu + s\right) - I\left(y \leq \mu\right)\right]\mathrm{d}s \geq 2\varepsilon\bar{C}^{-1}\right\} \\ &\leq \Pr\left\{\inf_{m\in\mathcal{M}}\left\|\hat{\mu}_{(m)N} - \mu\right\|_{L_2}^2 \geq \varepsilon\right\} = 0 \quad \text{as} \quad n \to \infty.\end{aligned}$$

On the other side,

$$\begin{aligned}\left|\xi_n(\boldsymbol{w}_1) - E\rho_\tau(e_\tau)\right| &= E\left\{\rho_\tau\left(e_\tau - \left(\hat{\mu}_n(\boldsymbol{w}_1) - \mu\right)\right) - \rho_\tau\left(e_\tau\right)\middle|\mathcal{D}_n\right\} \\ &\leq \frac{\bar{C}}{2}E\left\{\left(\sum_{m\in\mathcal{M}_T} w_m\hat{\mu}_{(m)n} - \mu\right)^2\middle|\mathcal{D}_n\right\} \\ &\leq \frac{\bar{C}}{2}E\left\{\sup_{m\in\mathcal{M}_T}\left(\hat{\mu}_{(m)n} - \mu\right)^2\middle|\mathcal{D}_n\right\},\end{aligned}$$

where the first equality is followed by the independence between $(\boldsymbol{x}^\top, y)^\top$ and $(\boldsymbol{x}_i^\top, y_i)^\top$'s, and

the last inequality is followed by Jensen's inequality. Combining the consistency of $\hat{\mu}_{(m)n}$ ($m \in \mathcal{M}_T$) with Markov's inequality, we have

$$PR_{2n} \le \Pr\left[E\left\{\sup_{m\in\mathcal{M}_T}\left(\hat{\mu}_{(m)n}-\mu\right)^2\middle|\mathcal{D}_n\right\}\ge\varepsilon\right]\le\varepsilon^{-1}\cdot\sup_{m\in\mathcal{M}_T}\left\|\hat{\mu}_{(m)n}-\mu\right\|_{L_2}^2=o(1).$$

Therefore,

$$\frac{\inf_{\boldsymbol{m}\in\mathcal{M}}E\rho_\tau\left(y-\hat{\mu}_{(m)N}\right)}{\xi_n(\boldsymbol{w}_1)}\xrightarrow{p}1\quad\text{as}\quad n\to\infty.$$

The left side of (S3.3) converges to 1 in probabilty. The proof is completed. □

Table S1: Summary of the equivalence of asymptotic optimalities

Cardinality of $\mathcal{M}_T$	α_m	Equivalence
>0	$<1/4$	✓
	$\ge 1/4$	×
0	– –	×

Together with the discussion of Theorem 3, we can summarize a clear list of conditions for the equivalence (see Table S1). It implies that the asymptotic equivalence must require $|\mathcal{M}_T|>0$, and further restrict the order of $\inf_{m\in\mathcal{M}_T}\left\|\hat{\mu}_{(m)N}-\mu\right\|_{L_2}$ (denoted as α_m). The result can be attributed to either of the following reasons:

(R1) The infimum prediction risk among correct models decays faster than excess risk.

(R2) WOMA assigns weights to the non-optimal models with a positive probability.

Thus, when $|\mathcal{M}_T|=0$, the equivalence is destroyed due to (R2); When $|\mathcal{M}_T|>0$ with the order of $\inf_{m\in\mathcal{M}_T}\|\hat{\mu}_{(m)N}-\mu\|_{L_2}$ faster than $n^{-1/4}$, the equivalence is destroyed due to (R1).

Finally, we roughly explain why $\Pr(\boldsymbol{w}_*=\boldsymbol{w}_0)\nrightarrow 1$ under $|\mathcal{M}_T|=0$. Consider a binary set $\mathcal{M}=\{\mu_{(1)},\mu_{(2)}\}$. Without loss of generality we assume $\mu_{(1)}$ has the infimum prediction risk, i.e., $\boldsymbol{w}_*=(1,0)^\top$. Due to the convexity of $\rho_\tau(\cdot)$, if $\Pr(\boldsymbol{w}_*=\boldsymbol{w}_0)\to 1$, it follows that for any $\bar{\boldsymbol{w}}=(\bar{w}_1,\bar{w}_2)^\top\in\mathcal{M}$ and $\bar{\boldsymbol{w}}\ne\boldsymbol{w}_*$,

$$\Pr\left(\xi_n(\bar{\boldsymbol{w}})\le\xi_n(\boldsymbol{w}_*)\right)\to 0.$$

Note that

$$\xi_n(\bar{\boldsymbol{w}}) - \xi_n(\boldsymbol{w}_*) = E\left\{ \int_{\hat{\mu}_{(1)}-\mu}^{\bar{w}_1\hat{\mu}_{(1)}+(1-\bar{w}_1)\hat{\mu}_{(2)}-\mu} [I(y \le \mu + s) - I(y \le \mu)]\, \mathrm{d}s \middle| \mathcal{D}_n \right\}.$$

It is not hard to prove the uniform integrability under conditions such as (C.7) and (C.8). In this case, an essential condition is $E\left(\xi_n(\bar{\boldsymbol{w}}) - \xi_n(\boldsymbol{w}_*)\right) \to 0$. However,

$$E\left(\xi_n(\bar{\boldsymbol{w}}) - \xi_n(\boldsymbol{w}_*)\right) = E\left[f_{y|\boldsymbol{x}}(\xi_x) \left\{ (\bar{w}_1 - 1)\left(\hat{\mu}_{(1)} - \hat{\mu}_{(2)}\right) \right\}^2 \right] \ge \frac{C}{2}(1-\bar{w}_1)^2 E\left(\hat{\mu}_{(1)} - \hat{\mu}_{(2)}\right)^2. \tag{S3.5}$$

One can see the lower-bound of (S3.5) is positive unless $\bar{w}_1 = 1$ or $\inf_{m_1 \ne m_2} \|\hat{\mu}_{(m_1)} - \mu_{(m_2)}\|_{L_2} = o(1)$. The first is contradictory because $0 < \bar{w}_1 < 1$, and the second always destroys the identification of candidates. Of this aspect, $\boldsymbol{w}_* \ne \boldsymbol{w}_0$ has a positive probability.

To further investigate the problem, we conduct a simple simulation as a counterexample. Set $y_i = x_{i1} + x_{i2} + \varepsilon_i$, where x_{i1} and x_{i2} are i.i.d. sample from $N(0,1)$, and $\varepsilon_i \sim N(0,1)$. The candidate CQF models are $\mu_{(1)i} = \theta x_{1i}$ and $\mu_{(2)i} = \theta x_{2i}$, respectively. We apply JMA procedure to estimate $\hat{\boldsymbol{w}}_0$ of WOMA, which is consistent to $\boldsymbol{w}_0 = (w_1, w_2)^\top$ by Theorem 1 of Yu, Zhang and Liang (2025). The normalized FPR among candidates are defined as $\zeta_1 = E\rho_\tau(y - \hat{\mu}_{(1)})/\xi_n(\hat{\boldsymbol{w}}_0)$, $\zeta_2 = E\rho_\tau(y - \hat{\mu}_{(2)})/\xi_n(\hat{\boldsymbol{w}}_0)$, $\zeta_3 = \inf_{m\in\{1,2\}} E\rho_\tau(y - \hat{\mu}_{(m)})/\xi_n(\hat{\boldsymbol{w}}_0)$.

From Figure S1, we can find that although $\hat{\mu}_{(2)}$ always has the infimum prediction risk, the empirical probability of $w_1 \le 1/2$ is around 0.5. It indicates that the weight of $\hat{\mu}_{(2)}$ evaluates far away from 0 and 1 with a positive probability. Moreover, none of ζ_k $(k = 1, 2, 3)$ is close to 1. Therefore, it practically verifies the nonequivalence.

S4 Additional numerical results

In this part, we show some other numerical studies to support the performance of the proposed method, including supplementary results in Section 4 and further simulation designs.

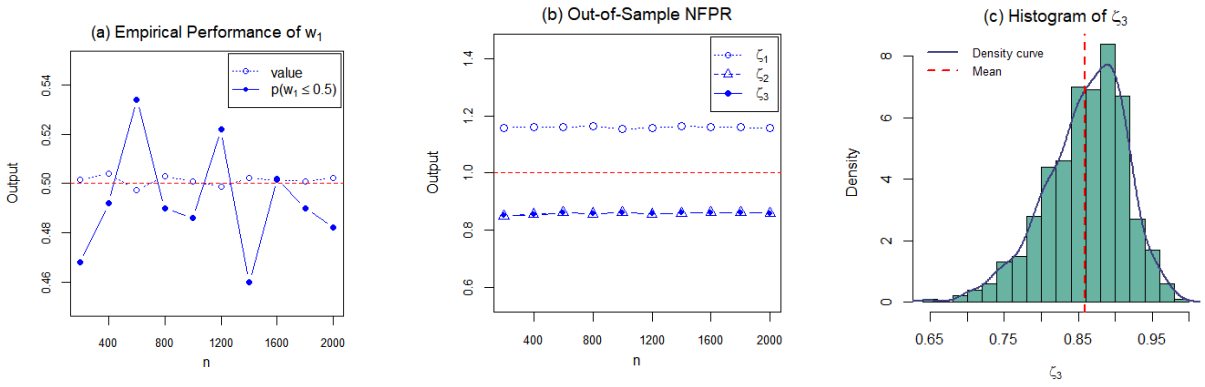

Figure S1: Simulation results under $\tau = 0.05$. The size of out-of-sample observations is 100.

S4.1 Supplementary simulation results in Section 4

Firstly, We present additional results that are omitted in Section 4. We investigate the change of SIS properties when $q_\tau(\boldsymbol{u})$ is misspecified. The subset of $\boldsymbol{z}$ is established by selecting the top $\lfloor 2n/(3\log(n)) \rfloor$ features. We consider the following two criteria: The first is normalized true discovery rate (NTDR):

$$\mathrm{NTDR}(\tau) = \frac{\mathrm{TDR}(\tau)}{\mathrm{TDR}_{full}(\tau)},$$

where the numerator is average true discovery rate obtained from incomplete dataset, and the denominator is the aaverage true discovery rate obtained from full dataset and true $q_\tau(u)$. The second criterion is the minimum model size, which is the smallest number that needs to be included to ensure all active components are selected.

Table S2 presents the result of SIS procedure (2.2). It demonstrates that the misspecified $q_\tau(\boldsymbol{u})$ dose affect the screening process. On the other side, different specification for $q_\tau(\boldsymbol{u})$ derives different active variables, as the NTDR of the union set is almost higher than the single one. Specifically, both q_τ-LM and q_τ-NAM are correct at $\tau = 0.5$, resulting in smaller MMSs. While q_τ-NAM can better capture the nonlinearity, none of specifications are entirely correct for $\tau \neq 0.5$. Therefore, ensembling the screened components from all candidates will be advantageous. An interesting point is that as n increases, the TDR of "Union" is

better than that given by full data and true $q_\tau(\boldsymbol{u})$. This implies that incorporating model uncertainty through multiple candidate specifications and combining screening sets of their marginal quantile utilities can effectively enhance the SIS property.

Table S2: NTDR and MMS (in parentheses) of SIS procedure of $\boldsymbol{z}$ in Section 4.1. †

(τ, ρ)	$n = 100$			$n = 200$			$n = 300$		
	q_τ-LM	q_τ-NAM	Union‡	q_τ-LM	q_τ-NAM	Union	q_τ-LM	q_τ-NAM	Union
(0.05, 0)	0.553(425)	0.587(421)	0.701	0.597(1294)	0.653(1281)	0.715	0.581(4361)	0.651(4342)	0.685
(0.5, 0)	0.697(145)	0.715(148)	0.760	0.844(132)	0.846(129)	0.869	0.902(175)	0.903(176)	0.919
(0.75, 0)	0.737(262)	0.818(197)	0.897	0.875(595)	1.006(295)	1.046	0.997(1750)	1.162(567)	1.204
(0.05, 0.5)	0.588(444)	0.615(446)	0.694	0.668(1360)	0.689(1365)	0.755	0.660(4551)	0.676(4586)	0.731
(0.5, 0.5)	0.762(323)	0.763(324)	0.820	0.860(954)	0.861(947)	0.882	0.936(3053)	0.926(3087)	0.944
(0.75, 0.5)	0.74(363)	0.853(340)	0.927	0.871(994)	0.963(965)	1.006	0.887(3430)	0.992(3190)	1.013

† The missing rate is about 32.42% by setting $\boldsymbol{\theta} = (-1, 1, 1)^\intercal$.
‡ The index set is formed by the union of screened variables from q_τ-LM and q_τ-NAM.

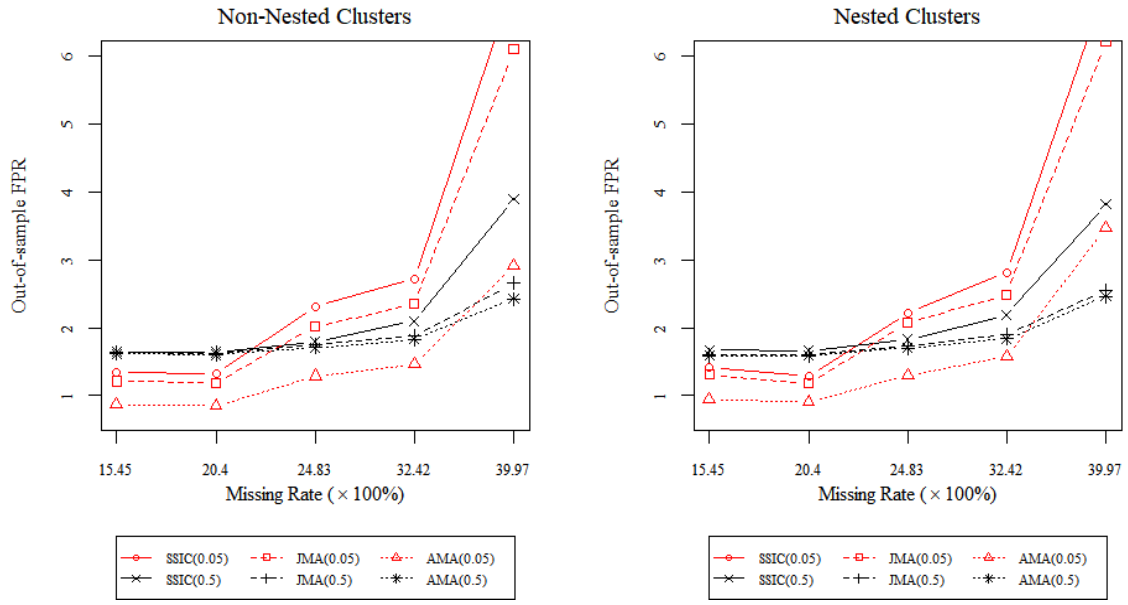

Figure S2: FPR of the MA estimator for CQF across a range of missing rates ($n = 200, \rho = 0.5$).

Figure S2 supplies the performance of three model averaging estimators under different missing rates, with respect to $\rho = 0.5$. We can find that the overall performance is same as that in Figure 1: Our method (AMA) has an outperformance regardless of τ and the missing rate. Besides, FPR of each estimator is almost higher than that with $\rho = 0$ due to

the accuracy of SIS.

Table S3: Out-of-sample FPR and its standard deviation (in parentheses) of AMA estimator via NON-NESTED candidate clusters.†

$(\tau,\ \rho)$	Number of Random Split B				
	1	2	5	10	20
(0.05, 0)	1.97(1.50)	1.80(0.99)	1.69(0.95)	1.63(0.90)	1.64(0.87)
(0.5, 0)	1.34(0.70)	1.26(0.69)	1.24(0.67)	1.20(0.60)	1.18(0.58)
(0.75, 0)	1.75(1.15)	1.58(0.81)	1.45(0.68)	1.43(0.71)	1.42(0.67)
(0.05, 0.5)	1.99(2.26)	2.04(2.34)	1.79(2.19)	1.46(1.66)	1.52(1.17)
(0.5, 0.5)	2.07(0.80)	1.94(0.47)	1.86(0.41)	1.83(0.36)	1.77(0.33)
(0.75, 0.5)	2.21(1.58)	2.09(0.98)	1.98(0.82)	1.84(0.54)	1.82(0.54)

† The missing rate is about 32.42% by setting $\boldsymbol{\theta} = (-1, 1, 1)^\top$.

Table S4: Out-of-sample FPR and its standard deviation (in parentheses) of AMA estimator via NESTED candidate clusters.†

$(\tau,\ \rho)$	Number of Random Split B				
	1	2	5	10	20
(0.05, 0)	2.17(1.27)	1.94(1.14)	1.79(1.13)	1.76(1.13)	1.74(0.86)
(0.5, 0)	1.33(0.69)	1.24(0.65)	1.25(0.65)	1.22(0.62)	1.18(0.59)
(0.75, 0)	1.61(0.82)	1.49(0.82)	1.47(0.70)	1.45(0.69)	1.45(0.69)
(0.05, 0.5)	2.36(2.42)	1.99(1.86)	1.69(1.69)	1.69(1.64)	1.66(1.51)
(0.5, 0.5)	2.01(0.69)	1.94(0.62)	1.87(0.51)	1.84(0.45)	1.79(0.35)
(0.75, 0.5)	2.20(1.02)	1.98(1.11)	1.94(0.70)	1.89(0.65)	1.84(0.66)

† The missing rate is about 32.42% by setting $\boldsymbol{\theta} = (-1, 1, 1)^\top$.

Secondly, we show the performance of AMA among different settings of random split time. Table S3–S4 reveal a consistent decreasing trend in both FPR and SD as B increases. This phenomenon demonstrates the data-adaptation of the proposed algorithm: Through an iterative random partition mechanism, the method effectively reduces potential uncertainty inherent in the fixed dataset. In this sense, it shares the conceptual similarity with the data-robust behavior of cross-validation techniques.

Thirdly, we show the histogram of λ in Figure S3, which is selected by (2.5). One can find that when $\tau = 0.5$ the optimal λ tends to be smaller than other probability levels. Such phenomenon stems from the action of λ: it simultaneously describes the sensitivity of adaptive weights corresponding to the loss of testing sample. In our simulation design,

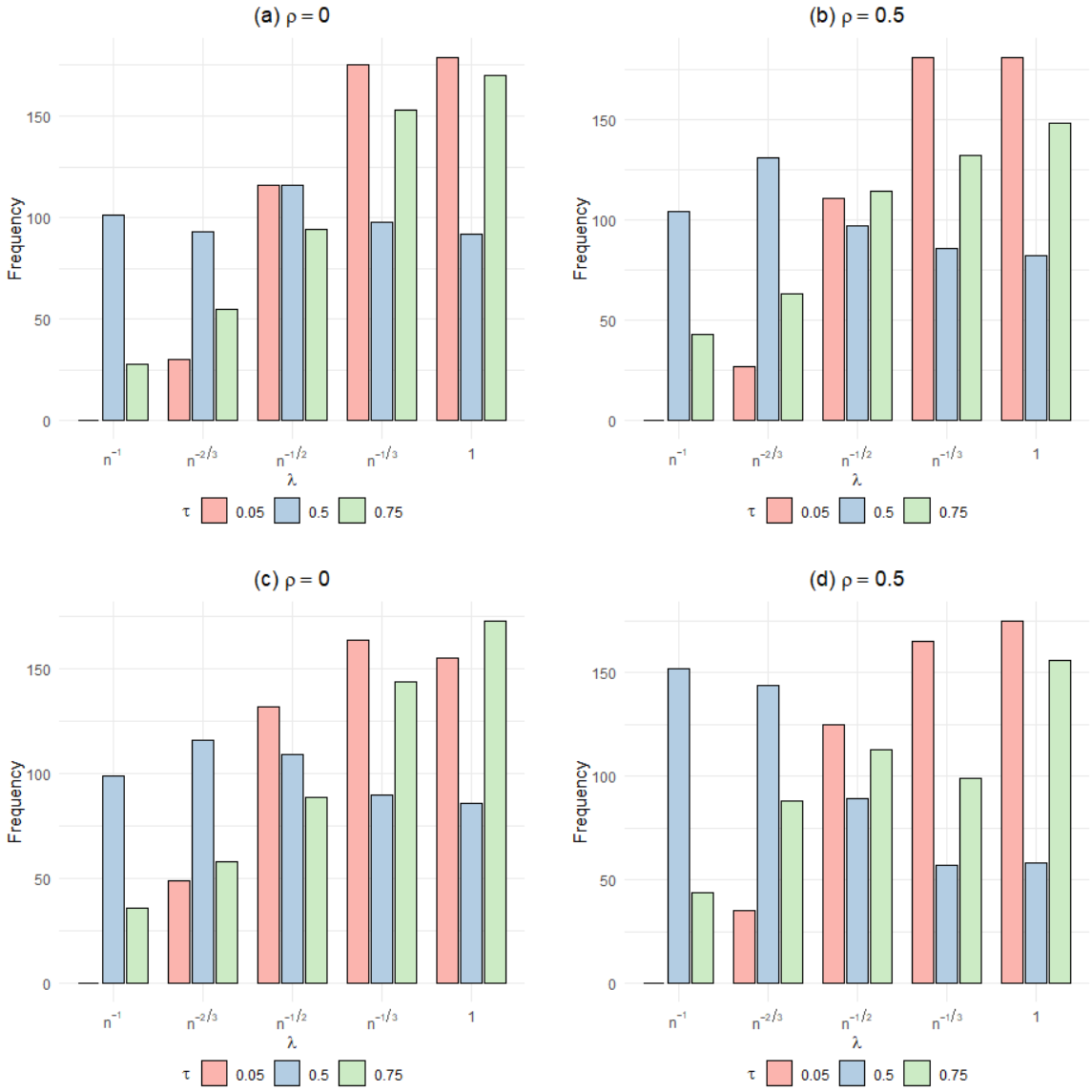

Figure S3: Histogram of the optimal λ in "MA" column of AMA, corresponding to $n = 200$ and 32.42% missing rate. The fist row is plotted by non-nested set and the second row is plotted by nested set.

smaller λ (with $\tau = 0.5$) induce near-uniform weights and a simple averaging procedure. This is mainly because both q_τ-LM and q_τ-NAM are correctly specified, hence weakening the necessity for distinguishing candidates. In contrast, larger λ (with $\tau \neq 0.5$) sharpen the weight differentiation, preferentially matching the model with superior historical loss as the inter-model divergence grows.

In summary, the output of λ also implicitly encodes the relative discriminability among candidate models across quantiles.

S4.2 Additional simulations

Further results for predicting response mechanism

In this part, we compare the predicting accuracy of high-dimensional $\pi(\boldsymbol{x})$ in Section 4.3. Set 100 pairs of out-of-sample observations, we use the following Kullback-Leibler divergence (KLD) as a measure:

$$\mathrm{KLD} = \frac{1}{100}\sum_{i=1}^{100}\left[\dot{r}_i\left\{\mathrm{logit}(\pi(\dot{\boldsymbol{x}}_i)) - \mathrm{logit}(\hat{\pi}(\dot{\boldsymbol{x}}_i))\right\} + \log\left\{\frac{1-\pi(\dot{\boldsymbol{x}}_i)}{1-\hat{\pi}(\dot{\boldsymbol{x}}_i)}\right\}\right].$$

We consider the following competitors: (i) MCP: variable selection by MCP penalty; (ii) GLASSO: group-lasso estimation; (iii) 5-CVMA: 5-folds cross-validation model averaging; (iv) JMA: jackknife model averaging; (v) RJMA: robust jackknife model averaging (3.5); (vi) RCVMA: 5-folds robust cross-validation model averaging. Methods (ii)–(v) are same as those in Section 4.3, and the last is adopted incorporating 5-CVMA with constraints in (3.5). Other settings are same as in Section 4.3.

Table S5: Out-of-sample KLD and its standard deviation (in parentheses) of response mechanism. ($n = 200$)†

ρ	Methods					
	MCP	GLASSO	JMA	5-CVMA	RJMA	RCVMA
0	0.411(1.316)	0.222(0.138)	0.266(0.137)	**0.137**(*0.062*)	0.158(0.101)	**0.136**(*0.540*)
0.5	0.362(0.540)	0.222(0.157)	0.210(0.101)	**0.112**(*0.052*)	0.143(0.103)	*0.115*(**0.048**)
0.8	0.382(1.648)	0.200(0.163)	0.137(0.067)	**0.088**(**0.050**)	0.131(0.099)	*0.108*(*0.051*)

† The best and the second best outputs are in **bold** and *italic*, respectively.

Results in Table S5 demonstrate that model averaging estimators consistently outperform variable selection methods. Specifically, 5-CVMA achieves better performance than JMA, mainly due to the greater data adaptability and ability for controlling overfitting through multifolds cross-validation. On the other side, RJMA always performs better than JMA, and RCVMA performs nearly as well as 5-CVMA. These suggest that the unbiased constraint does not compromise predicting accuracy. Combining with the result in Section 4.3, we can further conclude that high precision for $\pi(\boldsymbol{x})$ does not contribute for predicting CQF when applying inverse probability weighting. Therefore, implementing the robust enhancing strategy is necessary.

Further simulation studies for nonsparse high-dimensional model, aggregating with model-free feature screening procedures

In this part, we change the true value of $\boldsymbol{\beta}$ in Section 4.3 as

$$\beta_j = (-1)^j \sqrt{8} j^{-1}, \quad j = 1, ..., p,$$

and $n = 200$, $p = 1000$. Other settings of data generating process and candidate models are inherited from Section 4.3. We focus on the following competitors:

(a) SIS-mix: adaptive estimation based on variable screening procedure (2.2), where $q_\tau^{(k)}(\boldsymbol{u})$ is fitted by minimizing $\sum_{i=1}^n W_i \rho_\tau \left(y_i - q_\tau^{(k)}(\boldsymbol{u}_i) - \sum_{j=1}^5 \beta_j z_{ij} \right)$.

(b) DC-mix: adaptive estimation based on the distance-correlation screening procedure (DC-SIS, proposed by Li, Zhong and Zhu (2012)). Since $y|\boldsymbol{x}, r = 1$ has the same distribution with $y|\boldsymbol{x}$ under MAR, the dataset for DC-SIS is $\{(\boldsymbol{x}_i^\top, y_i | r_i = 1) : i = 1, ..., n\}$.

The response mechanism in each competitor is estimated by JMA and RJMA, respectively, hence resulting in total 6 averaged estimators.

The simulation results in Tables S6–S7 reveal several key findings. Firstly, DC-mix performs lower FPR than SIS-mix for $\tau \neq 0.5$, while SIS-mix shows better performance at

$\tau = 0.5$. This demonstrates that the misspecification of $q_\tau(\boldsymbol{u})$ impacts model-based screening effectiveness, further affects the prediction of CQF. Secondly, the robust enhancement from JMA to RJMA consistently improves both FPR and SD across all scenarios, aligning with the performance in Section 4.3. Thirdly, the adaptive estimator combined SIS and DC-SIS candidates keeps the outperformacne, demonstrating that our method can aggregate diverse screening approaches as well. Finally, the nested candidate set consistently outperforms the nonnested counterparts, as the screening process tends to produce underfitted models for the nonsparse case.

Table S6: Out-of-sample FPR and its standard deviation (in parentheses) of CQF estimator via NON-NESTED candidate clusters. †

$(\tau,\ \rho)$	JMA			RJMA		
	SIS-mix	DC-mix	MA‡	SIS-mix	DC-mix	MA
(0.05, 0)	0.788(0.377)	0.727(0.290)	0.733(0.315)	0.793(0.277)	0.700(0.241)	**0.717**(**0.219**)
(0.5, 0)	1.042(0.145)	1.055(0.143)	0.989(0.123)	1.019(0.125)	1.047(0.135)	**0.979**(**0.115**)
(0.75, 0)	1.065(0.216)	1.033(0.247)	0.984(0.212)	0.973(0.175)	0.963(0.179)	**0.906**(**0.151**)
(0.05, 0.5)	0.664(0.537)	0.659(0.336)	0.663(0.462)	0.634(0.232)	0.652(0.523)	**0.650**(**0.277**)
(0.5, 0.5)	0.923(0.119)	0.946(0.125)	0.889(0.114)	0.913(0.112)	0.939(0.130)	**0.883**(**0.104**)
(0.75, 0.5)	0.960(0.218)	0.933(0.245)	0.898(0.224)	0.852(0.138)	0.857(0.143)	**0.809**(**0.130**)
(0.05, 0.8)	0.524(0.265)	0.515(0.246)	0.516(0.258)	0.523(0.217)	0.503(0.224)	**0.509**(**0.222**)
(0.5, 0.8)	0.590(0.091)	0.595(0.090)	0.567(0.080)	0.579(0.083)	0.583(0.081)	**0.556**(**0.072**)
(0.75, 0.8)	0.644(0.182)	0.614(0.177)	0.600(0.177)	0.565(0.104)	0.545(0.092)	**0.529**(**0.089**)

† The best output among "MA" columns is labeled in **bold**.

‡ The candidate set forms as the union of that in SIS-mix and DC-mix.

Table S7: Out-of-sample FPR and its standard deviation (in parentheses) of CQF estimator via NESTED candidate clusters. †

$(\tau,\ \rho)$	JMA			RJMA		
	SIS-mix	DC-mix	MA‡	SIS-mix	DC-mix	MA
(0.05, 0)	0.813(0.391)	0.751(0.301)	0.716(0.297)	0.845(0.299)	0.710(0.222)	**0.688**(**0.209**)
(0.5, 0)	1.042(0.143)	1.053(0.138)	0.990(0.124)	1.019(0.128)	1.046(0.135)	**0.979**(**0.115**)
(0.75, 0)	1.063(0.239)	1.024(0.360)	0.978(0.219)	0.965(0.173)	0.944(0.175)	**0.901**(**0.161**)
(0.05, 0.5)	0.705(0.537)	0.699(0.386)	0.674(0.443)	0.693(0.318)	0.685(0.409)	**0.646**(**0.345**)
(0.5, 0.5)	0.923(0.119)	0.946(0.128)	0.889(0.114)	0.912(0.112)	0.937(0.130)	**0.883**(**0.105**)
(0.75, 0.5)	0.948(0.213)	0.911(0.244)	0.887(0.133)	0.838(0.133)	0.838(0.158)	**0.801**(**0.126**)
(0.05, 0.8)	0.552(0.262)	0.528(0.232)	0.515(0.234)	0.562(0.231)	0.532(0.228)	**0.504**(**0.213**)
(0.5, 0.8)	0.590(0.089)	0.596(0.095)	0.567(0.080)	0.579(0.083)	0.582(0.081)	**0.556**(**0.072**)
(0.75, 0.8)	0.636(0.185)	0.605(0.172)	0.594(0.168)	0.556(0.096)	0.536(0.090)	**0.526**(**0.088**)

† The best output among "MA" columns is labeled in **bold**.

‡ The candidate set forms as the union of that in SIS-mix and DC-mix.

S4.3 Application in gene microarray data

In this section, we apply the proposed method to analyze a microarray dataset from Gene Expression Omnibus (the dataset is available at https://www.ncbi.nlm.nih.gov/geo/, or access via `getGEO("GSE5680")` in `R`), which is collected by the expression quantitative trait locus mapping in the Rattus norvegicus. The aim is to research the gene regulation in the mammalian eye and to identify genetic variation relevant to human eye disease (Scheetz et al. (2006)). The dataset consists of over $31,000$ gene probes individually acquired from total 120 12-week-old eye tissue of hybrid F2 rats and $18,976$ of them are detected at a level sufficient to be considered "expressed", the main purpose of which is to find the genes that are correlated with the gene TRIM32 (presented by the probe "1389163_at") causing Bardet–Biedl syndrome.

we investigate the relationship between the quantile effects of TRIM32 (denoted as y) and the sparsity assumption based on "sufficiently expressed" probes. We first standardize each probe set and select the top 3000 probes with the largest variances. Besides of the linear specification in Wang et al. (2023), we also wonder the nonlinear representation of TRIM32 corresponding to the covariates with the strongest correlation. Thereby u is cho-

sen by the largest absolute value among 3000 probes and $\dim(\boldsymbol{z}) = 2999$, the mission is transformed into evaluating $q_\tau(u) + \sum_{j=1}^{p} \beta_{\tau,j} z_j$ with either linear ($q_\tau^{(1)}(u) = \alpha u$) or nonlinear ($q_\tau^{(2)}(u) = f_\tau(u)$) candidate. Furthermore, to illustrate the methodology in the presence of nonresponse scenario, we artificially induce approximately 36% random missingness among $\{y_i : i = 1, ..., 120\}$ by the mechanism $\text{logit}\{\pi(\boldsymbol{x}; \boldsymbol{\phi})\} = 1.5 + \boldsymbol{\phi}^\top \widetilde{\boldsymbol{x}}$, where $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, ..., \widetilde{x}_{20})^\top$ is a random permutation of $(x_1, ..., x_{20})^\top$, and $\phi_j = (-1)^j$ for $j = 1, ..., 20$.

We randomly divide the dataset by $n_1 = 90$ observations for fitting and others for validation. During the process of CVMA for the response mechanism, we screen the top 50 components in conjunction with p-values less than 0.05 after ranking in decreasing order and set $d = \{5, 10\}$. Besides we use the top 5 components in terms of absolute correlation with y among $\boldsymbol{z}$ to obtain a prior $\hat{q}_\tau^{(k)}(u)$ $(k = 1, 2)$. Throughout the screening process, $v_n = \lfloor n_1 / \log(n_1) \rfloor$, and we use the same clustering procedure in Section 4.1 to establish either non-nested or nested candidate set. Other settings of competitors are employed in Section 4.3.

Table S8: Hybrid F2 rat eye tissue microarray data: out-of-sample FPR and SD (in parentheses) via NON-NESTED clusters. †

τ	G-JMA			5-CVMA			R-AMA		
	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA
0.05	0.410	0.303	0.362	0.677	0.454	0.337	0.243	0.448	**0.246**
	(0.568)	(0.097)	(0.426)	(1.073)	(0.180)	(0.282)	(0.172)	(0.166)	(**0.136**)
0.25	0.775	0.310	0.620	0.790	0.371	0.396	0.386	0.353	**0.316**
	(2.863)	(0.074)	(2.495)	(1.965)	(0.088)	(0.730)	(0.234)	(0.085)	(**0.086**)
0.5	0.457	0.273	0.374	0.584	0.309	**0.307**	0.402	0.302	0.314
	(0.493)	(0.052)	(0.458)	(0.603)	(0.057)	(0.155)	(0.136)	(0.055)	(**0.066**)
0.75	0.434	0.304	0.366	0.575	0.367	0.339	0.356	0.325	**0.291**
	(0.261)	(0.077)	(0.166)	(0.508)	(0.095)	(0.138)	(0.156)	(0.057)	(**0.087**)
0.95	1.665	0.286	1.639	1.190	0.477	2.295	0.215	0.431	**0.210**
	(>10)	(0.106)	(>10)	(6.476)	(0.225)	(>10)	(0.156)	(0.187)	(**0.138**)

† In each row, the best output among "MA" columns is labeled in **bold**.

Table S9: Hybrid F2 rat eye tissue microarray data: out-of-sample FPR and SD (in parentheses) via NESTED clusters. †

τ	G-JMA			5-CVMA			R-AMA		
	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA	q_τ-LM	q_τ-NAM	MA
0.05	0.383	0.323	0.373	0.573	0.474	0.346	0.276	0.444	**0.266**
	(0.312)	(0.109)	(0.258)	(0.591)	(0.191)	(0.202)	(0.261)	(0.156)	(**0.170**)
0.25	0.408	0.317	0.362	0.601	0.369	0.371	0.410	0.352	**0.335**
	(0.220)	(0.077)	(0.142)	(0.553)	(0.090)	(0.268)	(0.256)	(0.076)	(**0.107**)
0.5	0.415	0.286	0.339	0.702	0.319	0.372	0.398	0.308	**0.313**
	(0.357)	(0.056)	(0.118)	(1.357)	(0.064)	(0.638)	(0.127)	(0.060)	(**0.067**)
0.75	0.600	0.289	0.547	0.874	0.349	0.345	0.333	0.341	**0.285**
	(2.060)	(0.058)	(1.852)	(2.728)	(0.076)	(0.229)	(0.137)	(0.062)	(**0.078**)
0.95	0.678	0.313	0.626	0.903	0.454	0.481	0.221	0.411	**0.228**
	(1.529)	(0.108)	(1.376)	(2.125)	(0.210)	(0.784)	(0.153)	(0.157)	(**0.169**)

† In each row, the best output among "MA" columns is labeled in **bold**.

Table S8–S9 present results over 100 random replications. From AMA, we can find that when τ = 0.05 and τ = 0.95, the linear specification of $q_\tau(u)$ outperforms the nonlinear model. However, this feature is obscured in G-JMA and 5-CVMA due to the presence of outliers. Both averaging estimators exhibit poor performance particularly at τ = 0.95. R-AMA uniformly results an adaptive approximation with the optimal estimator, and keeps the satisfactory predictive accuracy and variance under non-nested and nested clusters. This also confirms the conclusions drawn from our numerical simulations.

References

Angrist, J., Chernozhukov, V., Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica* **74**, 539–563.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour*, XXXI-2001 (Vol. 1851). Springer Science & Business Media.

Chang, G. Z., Shi, J. H. (2003). *Mathematical Analysis Course*. University of Science and Technology of China Press (in Chinese). He Fei, China.

Chen, Z., Liao, J., Xu, W., Yang Y. (2023). Multifold cross-validation model averaging for generalized additive partial linear models. *Journal of Computational and Graphical Statistics* **32**, 1649–1659.

Crump, R. K., Imbens, G. W., Mitnik, O. A., Hotz, V. J. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**, 187–199.

Fan, J., Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics* **38**, 3567–3604.

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* Chapman and Hall, London.

Gu, Y., Zou, H. (2019). Aggregated expectile regression by exponential weighting. *Statistica Sinica* **29**, 671–692.

Han, P., Kong, L., Zhao, J., Zhou, X. (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **81**, 305–333.

He, X., Wang, L., Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics* **41**, 342–369.

Hjort, N. L., Pollard, D. (1993). *Asymptotics for minimisers of convex processes.* Statistical Research Report, University of Oslo.

Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *Annals of Statistics* **26**, 755–770.

Li, R., Zhong, W., Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129–1139.

Li, W., Gu, Y., Liu, L. (2020). Demystifying a class of multiply robust estimators. *Biometrika* **107**, 919–933.

Lu, X., Su, L. (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics* **188**, 40–58.

Serfling, R. (1980). *Approximation theorems of mathematical statistics.* John Wiley, New York.

Shan, J., Li, W., Ai, C. (2024). Efficient nonparametric inference of causal mediation effects with nonignorable missing confounders. *arxiv*: 2402.05384.

Shan, K., Yang, Y. (2009). Combining regression quantile estimators. *Statistica Sinica* **19**, 1171—1191.

Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14429–14434.

Tsybakov., A. (2009). *Introduction to nonparametric estimation.* Springer-Verlag, 2009.

van der Vaart, A. W. (2000). *Asymptotic statistics.* Cambridge University Press.

van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes.* New York: Springer.

Wang, L. (2011). Gee analysis of clustered binary data with diverging number of covariates. *Annals of Statistics* **39**, 389–417.

Wang, M., Zhang, X., Wan, A. T. K., You, K., Zou, G. (2023). Jackknife model averaging for high-dimensional quantile regression. *Biometrics* **79**, 178–189.

Xiong, W., Deng, D., Wang, D. (2025). Semiparametric model averaging for high-dimensional quantile regression with nonignorable nonresponse. *arXiv*, 2509.00464.

Xu, W., Zhang, X. (2024). On asymptotic optimality of least squares model averaging when true model is included. *arXiv*: 2411.09258v1.

Xu, W., Zhang, X., Chiou, J. M. (2023). On asymptotic optimality of least squares model averaging when true model is included. *arXiv*: 2310.01970v1.

Yu, D., Zhang, X., Liang, H. (2025). Unified optimal model averaging with a general loss function based on cross-validation. *Journal of the American Statistical Association*, doi.org/10.1080/01621459.2025.2487215.

Zhang, X., Zou, G., Liang, H., Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association* **115**, 972–984.