Conformal Prediction Under Nonignorable Missingness

Menghan Yi[1], Yingying Zhang[1], Yanlin Tang[1*] and Huixia Judy Wang[2*]

[1]East China Normal University and [2]Rice University

Supplementary Material

The supplementary materials include the following: first, the technical proofs for Theorems 1–3, presented in Sections S1.1–S1.3, respectively; second, the high-dimensional extensions in Section S2, including theoretical discussion in Section S2.1 and simulation results in Section S2.2; and finally, additional numerical studies in Section S3, including conditional coverage evaluation in Section S3.1 and supplementary simulation results in Section S3.2.

S1 Technical Proofs

S1.1 Proof of Theorem 1

Without loss of generality, assume the non-missing calibration set $\mathcal{D}_c^{\text{obs}}$ is indexed by $\{1,\ldots,n\}$, while the test data is indexed by $\{n+1\}$. Define the random variable $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$ and its observed value $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ for $i = 1,\ldots,n+1$. According to the definition by Tibshirani et al. (2019), the random variables $\{\mathbf{Z}_i, i = 1,\ldots,n+1\}$ are referred to as weighted exchangeable if their joint density function can be factorized as

$$f(\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}) = \prod_{i=1}^{n+1} \phi_i(\mathbf{z}_i) \cdot g(\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}), \tag{S1.1}$$

where $\phi_i(\cdot)$ is a certain adjustment function, and g is a permutation-invariant function; that is, for any permutation σ of the set $\{1, \ldots, n+1\}$,

$$g(\mathbf{z}_{\sigma(1)}, \ldots, \mathbf{z}_{\sigma(n+1)}) = g(\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}).$$

In our data setup, $\mathbf{Z}_1$ to $\mathbf{Z}_n$ are observed data points drawn from the conditional distribution $f(\mathbf{z} \mid \delta = 1)$, while $\mathbf{Z}_{n+1}$ is drawn from the marginal distribution $f(\mathbf{z})$. Since all variables are independent, the joint density can be factorized as:

$$\begin{aligned} f(\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}) &= \prod_{i=1}^{n} f(\mathbf{z}_i|\delta = 1) f(\mathbf{z}_{n+1}) \\ &= \frac{f(\mathbf{z}_{n+1})}{f(\mathbf{z}_{n+1}|\delta = 1)} \prod_{i=1}^{n+1} f(\mathbf{z}_i|\delta = 1) = w(\mathbf{z}_{n+1}) \prod_{i=1}^{n+1} f(\mathbf{z}_i|\delta = 1). \end{aligned} \tag{S1.2}$$

Therefore, the distribution of our data satisfies the weighted exchangeability definition (S1.1), where $\phi_i(\mathbf{z}) = 1$ for $i = 1, \ldots, n$, and $\phi_i(\mathbf{z}) = w(\mathbf{z})$ for $i = n+1$, and $g(\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}) = \prod_{i=1}^{n+1} f(\mathbf{z}_i \mid \delta = 1)$ is permutation invariant. We define the event $\mathcal{E}_{\mathbf{Z}}$ to indicate that two unordered collections are equal:

$$\mathcal{E}_{\mathbf{Z}} = \big\{[\mathbf{Z}_1, \ldots, \mathbf{Z}_{n+1}] = [\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}]\big\},$$

which means that each $\mathbf{Z}_i$ takes a value from the set $\{\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}\}$, and the collection $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_{n+1}\}$ matches $\{\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}\}$ as a multiset, regardless of order. Let the nonconformity score of $\mathbf{z}_i$ be defined as $r_i = R(\mathbf{z}_i)$ for $i = 1, \ldots, n+1$. By weighted exchangeability, we can calculate

$$\begin{aligned}
\mathbb{P}\{R_{n+1} = r_i \mid \mathcal{E}_{\mathbf{Z}}, \mathcal{D}_t\} &= \mathbb{P}\{\mathbf{Z}_{n+1} = \mathbf{z}_i \mid \mathcal{E}_{\mathbf{Z}}, \mathcal{D}_t\} \\
&= \frac{\sum_{\sigma(n+1)=i} f\big(\mathbf{z}_{\sigma(1)}, \ldots, \mathbf{z}_{\sigma(n+1)}\big)}{\sum_{\sigma} f\big(\mathbf{z}_{\sigma(1)}, \ldots, \mathbf{z}_{\sigma(n+1)}\big)} \\
&\overset{(S1.2)}{=} \frac{\sum_{\sigma(n+1)=i} w\big(\mathbf{z}_{\sigma(n+1)}\big) \prod_{j=1}^{n+1} f(\mathbf{z}_j|\delta = 1)}{\sum_{\sigma} w\big(\mathbf{z}_{\sigma(n+1)}\big) \prod_{j=1}^{n+1} f(\mathbf{z}_j|\delta = 1)} \\
&= \frac{w(\mathbf{z}_i)}{\sum_{j=1}^{n+1} w(\mathbf{z}_j)},
\end{aligned}$$

where the summations are taken over permutations σ of the numbers $1, \ldots, n+1$. Note that this conditional probability is well-defined, since Assumption 1 means that under both $F(\mathbf{z}|\delta = 1)$ and $F(\mathbf{z})$ it is almost surely that $w(\mathbf{z}) < \infty$. Therefore, the distribution of R_{n+1} is

$$(R_{n+1} \mid \mathcal{E}_{\mathbf{Z}}, \mathcal{D}_t) \sim F_R(r) = \sum_{i=1}^{n+1} \frac{w(\mathbf{x}_i, y_i)}{\sum_{j=1}^{n+1} w(\mathbf{x}_j, y_j)} \mathbb{I}\left(r_i \leq r\right), \tag{S1.3}$$

which depends on the true weight function $w(\mathbf{x}, y)$.

To establish the properties under the estimated weights, we generate a new sample using the estimated weight function $\widehat{w}(\mathbf{x}, y)$,

$$(\widetilde{\mathbf{X}}_{n+1}, \widetilde{Y}_{n+1}) \sim \widetilde{f}(\mathbf{x}, y) \quad \text{with} \quad \widetilde{f}(\mathbf{x}, y) = \widehat{w}(\mathbf{x}, y) f(\mathbf{x}, y|\delta = 1). \tag{S1.4}$$

Since the expectation of $\widehat{g}(\mathbf{U})\, q(Y, \widehat{\boldsymbol{\gamma}})$ is finite under the theorem's assumption, the estimated weight $\widehat{w}(\mathbf{X}, Y)$ is finite almost surely. This implies that

$\widehat{w}(\mathbf{X}, Y)$ can be normalized to have expectation one. Consequently, $\widetilde{f}(\mathbf{x}, y)$ defines a valid probability density function. Let $\widetilde{\mathbf{Z}}_{n+1} = (\widetilde{\mathbf{X}}_{n+1}, \widetilde{Y}_{n+1})$, and denote its nonconformity score by $\widetilde{R}_{n+1} = R(\widetilde{\mathbf{Z}}_{n+1})$. Define the event $\mathcal{E}_{\widetilde{\mathbf{Z}}}$ to indicate that two unordered collections are equal: $\mathcal{E}_{\widetilde{\mathbf{Z}}} = \big\{[\mathbf{Z}_1, \ldots, \mathbf{Z}_n, \widetilde{\mathbf{Z}}_{n+1}] = [\mathbf{z}_1, \ldots, \mathbf{z}_n, \widetilde{\mathbf{z}}_{n+1}]\big\}$. Using the same argument as for (S1.3), we have

$$(\widetilde{R}_{n+1} \mid \mathcal{E}_{\widetilde{\mathbf{Z}}}, \mathcal{D}_t) \sim F_{\widetilde{R}}(r) = \sum_{i=1}^{n} \frac{\widehat{w}(\mathbf{x}_i, y_i)}{\sum_{j=1}^{n} \widehat{w}(\mathbf{x}_j, y_j) + \widehat{w}(\widetilde{\mathbf{x}}_{n+1}, \widetilde{y}_{n+1})} \mathbb{I}(r_i \le r)$$
$$+ \frac{\widehat{w}(\widetilde{\mathbf{x}}_{n+1}, \widetilde{y}_{n+1})}{\sum_{j=1}^{n} \widehat{w}(\mathbf{x}_j, y_j) + \widehat{w}(\widetilde{\mathbf{x}}_{n+1}, \widetilde{y}_{n+1})} \mathbb{I}(\widetilde{r}_{n+1} \le r),$$

which is well-defined, since $\widehat{w}(\mathbf{z}) < \infty$ almost surely under both $F(\mathbf{z} \mid \delta = 1)$ and $F(\mathbf{z})$. Let $Q_{\widetilde{R}}(\tau)$ denote the τ-th quantile of the distribution $F_{\widetilde{R}}$, defined as $Q_{\widetilde{R}}(\tau) = \inf\{r : F_{\widetilde{R}}(r) \ge \tau\}$. According to the proposed definition of prediction set, we have $\{\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{\mathbf{X}}_{n+1})\} \Leftrightarrow \{\widetilde{R}_{n+1} \le Q_{\widetilde{R}}(1-\alpha)\}$. Therefore, by the definition of the quantile, it follows that:

$$\begin{aligned} \mathbb{P}\big\{\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{\mathbf{X}}_{n+1}) \mid \mathcal{D}_t\big\} &= \mathbb{E}\big[\mathbb{P}\big\{\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{\mathbf{X}}_{n+1}) \mid \mathcal{E}_{\widetilde{Z}}, \mathcal{D}_t\big\}\big] \\ &= \mathbb{E}\big[\mathbb{P}\{\widetilde{R}_{n+1} \le Q_{\widetilde{R}}(1-\alpha) \mid \mathcal{E}_{\widetilde{Z}}, \mathcal{D}_t\}\big] \\ &\ge 1 - \alpha. \end{aligned} \tag{S1.5}$$

Let $d_{\mathrm{TV}}(F_1, F_2)$ denote the total-variation (TV) distance between distributions F_1 and F_2. Then according to the definition of TV distance,

$$\begin{aligned} &\big|\mathbb{P}\big\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}) \mid \mathcal{D}_t, \mathcal{D}_c\big\} - \mathbb{P}\big\{\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{\mathbf{X}}_{n+1}) \mid \mathcal{D}_t, \mathcal{D}_c\big\}\big| \\ &\le d_{\mathrm{TV}}\big(F(\mathbf{x}, y), \widetilde{F}(\mathbf{x}, y)\big), \end{aligned}$$

where $\widetilde{F}(\mathbf{x}, y)$ is the distribution function of $\widetilde{f}(\mathbf{x}, y)$. Taking expectation over $\mathcal{D}_c$, we have

$$\begin{aligned}\mathbb{P}\big\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}) \mid \mathcal{D}_t\big\} &\geq \mathbb{P}\big\{\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{\mathbf{X}}_{n+1}) \mid \mathcal{D}_t\big\} - d_{\mathrm{TV}}\big(F(\mathbf{x}, y), \widetilde{F}(\mathbf{x}, y)\big) \\ &\overset{(S1.5)}{\geq} 1 - \alpha - d_{\mathrm{TV}}\big(F(\mathbf{x}, y), \widetilde{F}(\mathbf{x}, y)\big). \qquad \text{(S1.6)}\end{aligned}$$

By the integral definition of TV distance,

$$\begin{aligned}d_{\mathrm{TV}}\big(F(\mathbf{x}, y), \widetilde{F}(\mathbf{x}, y)\big) =& \frac{1}{2}\int |dF(\mathbf{x}, y) - d\widetilde{F}(\mathbf{x}, y)| \\ =& \frac{1}{2}\int |\widehat{w}(\mathbf{x}, y) - w(\mathbf{x}, y)| dF(\mathbf{x}, y|\delta = 1) \\ =& \frac{1}{2}\mathbb{E}|\widehat{w}(\mathbf{X}, Y) - w(\mathbf{X}, Y)| \\ =& \frac{1}{2}\mathbb{E}\big|\widehat{g}(\mathbf{U})q(Y, \widehat{\boldsymbol{\gamma}}) - g(\mathbf{U})q(Y, \boldsymbol{\gamma})\big|, \qquad \text{(S1.7)}\end{aligned}$$

where expectation is taken under $(\mathbf{X}, Y) \sim f(\mathbf{x}, y|\delta = 1)$. Finally, taking expectation with respect to $\mathcal{D}_t$ in inequality (S1.6), we have

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1})\} \geq 1 - \alpha - \frac{1}{2}\mathbb{E}|\widehat{g}(\mathbf{U})q(Y, \widehat{\boldsymbol{\gamma}}) - g(\mathbf{U})q(Y, \boldsymbol{\gamma})|.$$

S1.2 Proof of Theorem 2

Let $\mathcal{A} = \{A_j\}$ be a partition of $\mathcal{X}$, and let $A_n(\mathbf{X}_{n+1})$ be the subset of the calibration set whose covariates fall in the same partition element as $\mathbf{X}_{n+1}$:

$$A_n(\mathbf{X}_{n+1}) = \big\{(\mathbf{X}_i, Y_i) \in \mathcal{D}_c^{\mathrm{obs}} : \mathbf{X}_i \in A_j \text{ and } \mathbf{X}_{n+1} \in A_j\big\}. \qquad \text{(S1.8)}$$

It follows that, conditional on $\mathbf{X}_{n+1} \in A_j$, the pair $(\mathbf{X}_{n+1}, Y_{n+1})$ and the subset $A_n(\mathbf{X}_{n+1})$ remain weighted exchangeable, with their joint distribution sharing the same form as in (S1.2). By applying the same argument and derivation as in the proof of Theorem 1, we obtain a coverage guarantee analogous to (S1.5):

$$\mathbb{P}\big\{\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{\mathbf{X}}_{n+1}) \mid \widetilde{\mathbf{X}}_{n+1} \in A_j\big\} \geq 1-\alpha, \tag{S1.9}$$

where $\widetilde{\mathbf{X}}_{n+1}$ and $\widetilde{Y}_{n+1}$ follow the same joint distribution as in (S1.4). By the definition of the total variation (TV) distance, we obtain

$$\begin{aligned} & \mathbb{P}\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}),\ \mathbf{X}_{n+1} \in A_j\} \\ \geq\ & \mathbb{P}\{\widetilde{Y}_{n+1} \in \widehat{C}(\widetilde{\mathbf{X}}_{n+1}),\ \widetilde{\mathbf{X}}_{n+1} \in A_j\} - d_{\mathrm{TV}}(F(\mathbf{x},y), \widetilde{F}(\mathbf{x},y)). \\ \overset{(S1.9)}{\geq}\ & (1-\alpha)\mathbb{P}\{\widetilde{\mathbf{X}}_{n+1} \in A_j\} - d_{\mathrm{TV}}(F(\mathbf{x},y), \widetilde{F}(\mathbf{x},y)). \end{aligned}$$

Similarly, using the definition of the total variation distance, we have

$$\mathbb{P}\{\widetilde{\mathbf{X}}_{n+1} \in A_j\}\ \geq\ \mathbb{P}\{\mathbf{X}_{n+1} \in A_j\} - d_{\mathrm{TV}}\big(F(\mathbf{x}), \widetilde{F}(\mathbf{x})\big).$$

By combining the results above, we conclude that

$$\begin{aligned} & \mathbb{P}\big\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}),\ \mathbf{X}_{n+1} \in A_j\big\} \\ \geq\ & (1-\alpha)\big\{\mathbb{P}\{\mathbf{X}_{n+1} \in A_j\} - d_{\mathrm{TV}}\big(F(\mathbf{x}), \widetilde{F}(\mathbf{x})\big)\big\} - d_{\mathrm{TV}}\big(F(\mathbf{x},y), \widetilde{F}(\mathbf{x},y)\big). \end{aligned}$$

This implies that

$$\begin{aligned} & \mathbb{P}\big\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}) \mid \mathbf{X}_{n+1} \in A_j\big\} \geq 1-\alpha-\widehat{\Delta}_{\mathrm{loc}}, \\ \widehat{\Delta}_{\mathrm{loc}} &= \frac{(1-\alpha)d_{\mathrm{TV}}\big(F(\mathbf{x}), \widetilde{F}(\mathbf{x})\big) + d_{\mathrm{TV}}\big(F(\mathbf{x},y), \widetilde{F}(\mathbf{x},y)\big)}{\mathbb{P}\{\mathbf{X}_{n+1} \in A_j\}}. \end{aligned} \tag{S1.10}$$

Let $\bar{q}(\mathbf{x}, \gamma) = \int q(y, \boldsymbol{\gamma}) dF(y|\mathbf{x}, \delta = 1)$ and $\bar{q}(\mathbf{x}, \widehat{\gamma}) = \int q(y, \widehat{\boldsymbol{\gamma}}) dF(y|\mathbf{x}, \delta = 1)$. By the integral definition of TV distance, we have

$$\begin{aligned} d_{\mathrm{TV}}\big(F(\mathbf{x}), \widetilde{F}(\mathbf{x})\big) &= \frac{1}{2} \int |dF(\mathbf{x}) - d\widetilde{F}(\mathbf{x})| \\ = \ & \frac{1}{2} \int \Big| \int \{w(\mathbf{x}, y) - \widehat{w}(\mathbf{x}, y)\} dF(y \mid \mathbf{x}, \delta = 1) \Big| dF(\mathbf{x} \mid \delta = 1) \\ = \ & \frac{1}{2} \int \Big| g(\mathbf{x}) \bar{q}(\mathbf{x}, \gamma) - \widehat{g}(\mathbf{x}) \bar{q}(\mathbf{x}, \widehat{\gamma}) \Big| dF(\mathbf{x} \mid \delta = 1) \\ = \ & \frac{1}{2} \mathbb{E}_{\mathbf{X} \sim f(\mathbf{x}|\delta=1)} |g(\mathbf{X}) \bar{q}(\mathbf{X}, \gamma) - \widehat{g}(\mathbf{X}) \bar{q}(\mathbf{X}, \widehat{\gamma})|, \end{aligned} \tag{S1.11}$$

where the first equation holds because

$$dF(\mathbf{x}) = \left\{ \int w(\mathbf{x}, y) dF(y \mid \mathbf{x}, \delta = 1) \right\} \cdot dF(\mathbf{x} \mid \delta = 1).$$

By substituting equations (S1.11) and (S1.7) into (S1.10), we obtain an exact lower bound for the local coverage. Alternatively, we can directly apply the inequality $d_{\mathrm{TV}}(F(\mathbf{x}), \widetilde{F}(\mathbf{x})) \leq d_{\mathrm{TV}}(F(\mathbf{x}, y), \widetilde{F}(\mathbf{x}, y))$ to (S1.10) and then substitute (S1.7), which yields the conclusion of Theorem 2.

S1.3 Proof of Theorem 3

Lemma 1. *Let $G_f(t \mid \mathbf{x})$ be the conditional CDF of $f(Y \mid \mathbf{X})$, and let $\widehat{G}_{\widehat{f}}(t \mid \bar{A}(\mathbf{x}))$ be the local empirical CDF of $\widehat{f}(Y \mid \mathbf{X})$ constructed from $\bar{A}(\mathbf{x}) := A_n(\mathbf{x}) \cup \{(\mathbf{x}, y)\}$, where $A_n(\mathbf{x})$ is the calibration subset in* (S1.8) *from the same cluster as* $\mathbf{x}$*. They are defined as*

$$\begin{aligned} G_f(t \mid \mathbf{x}) &= \mathbb{P}\{f(Y \mid \mathbf{X}) \leq t \mid \mathbf{X} = \mathbf{x}\}, \\ \widehat{G}_{\widehat{f}}(t \mid \bar{A}(\mathbf{x})) &= \sum_{i \in \bar{A}(\mathbf{x})} \frac{w_i}{\sum_{j \in \bar{A}(\mathbf{x})} w_j} \mathbb{I}\{\widehat{f}(Y_i \mid \mathbf{X}_i) \leq t\}, \end{aligned} \tag{S1.12}$$

where the weights w_i *are assumed to be known. These two CDFs are used to construct the corresponding prediction intervals for* Y *given* $\mathbf{X}$ *as follows:*

$$C(\mathbf{X}; \alpha) = \big\{y \in \mathcal{Y} : G_f(f(y \mid \mathbf{X}) \mid \mathbf{X}) > \alpha\big\},$$

$$\widehat{C}(\mathbf{X}; \alpha) = \big\{y \in \mathcal{Y} : \widehat{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}) \mid \bar{A}(\mathbf{X})) > \alpha\big\}.$$

Under Assumptions 2–5, it follows that:

$$\mathbb{P}\{Y \in C(\mathbf{X}; \alpha) \Delta \widehat{C}(\mathbf{X}; \alpha)\} = o(1).$$

Proof of Lemma 1. To evaluate the discrepancy between $C(\mathbf{X}; \alpha)$ and $\widehat{C}(\mathbf{X}; \alpha)$, it is essential to quantify the deviation between the oracle CDF and its empirical counterpart in (S1.12). Let $G_{\widehat{f}}(t \mid \mathbf{X}) = \mathbb{P}\{\widehat{f}(Y \mid \mathbf{X}) \leq t \mid \mathbf{X}\}$ denote the conditional CDF of $\widehat{f}(Y \mid \mathbf{X})$. Then, by the triangle inequality,

$$\begin{aligned} &\sup_{y \in \mathcal{Y}} \big| G_f(f(y \mid \mathbf{X}) \mid \mathbf{X}) - \widehat{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}) \mid \bar{A}(\mathbf{X})) \big| \\ \leq\ & \sup_{y \in \mathcal{Y}} \big| G_f(f(y \mid \mathbf{X}) \mid \mathbf{X}) - G_{\widehat{f}}(f(y \mid \mathbf{X}) \mid \mathbf{X}) \big| \\ &+ \sup_{y \in \mathcal{Y}} \big| G_{\widehat{f}}(f(y \mid \mathbf{X}) \mid \mathbf{X}) - G_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}) \mid \mathbf{X}) \big| \\ &+ \sup_{y \in \mathcal{Y}} \big| G_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}) \mid \mathbf{X}) - \widehat{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}) \mid \bar{A}(\mathbf{X})) \big| \\ :=\ & \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3. \end{aligned}$$

Let $B_n = \{\sup_{y \in \mathcal{Y}} |\widehat{f}(y \mid \mathbf{X}) - f(y \mid \mathbf{X})| \geq \eta_n\}$, where η_n is from Assumption 3. On the complement event B_n^c, it can be shown that:

$$\begin{aligned}
\mathrm{I}_1 &\leq \sup_z \left| G_f(z \mid \mathbf{X}) - G_{\widehat{f}}(z \mid \mathbf{X}) \right| \\
&= \sup_z \left| \int_{\{y: f(y|\mathbf{X}) \leq z\}} f(y \mid \mathbf{X}) dy - \int_{\{y: \widehat{f}(y|\mathbf{X}) \leq z\}} f(y|\mathbf{X}) dy \right| \\
&\leq \sup_z \int_{\{y: f(y|\mathbf{X}) \leq z, \widehat{f}(y|\mathbf{X}) > z\}} f(y|\mathbf{X}) dy + \sup_z \int_{\{y: f(y|\mathbf{X}) > z, \widehat{f}(y|\mathbf{X}) \leq z\}} f(y|\mathbf{X}) dy \\
&\leq \sup_z \int_{\{y: z - \eta_n \leq f(y|\mathbf{X}) \leq z\}} f(y|\mathbf{X}) dy + \sup_z \int_{\{y: z < f(y|\mathbf{X}) \leq z + \eta_n\}} f(y|\mathbf{X}) dy \\
&= \sup_z \left| G_f(z|\mathbf{X}) - G_f(z - \eta_n|\mathbf{X}) \right| + \sup_z \left| G_f(z + \eta_n|\mathbf{X}) - G_f(z|\mathbf{X}) \right| \\
&= o_P(1),
\end{aligned}$$

where the last equality follows from the Lipschitz continuity of G_f in Assumption 4. Similarly, on the event B_n^c, we have:

$$\begin{aligned}
\mathrm{I}_2 \ &\leq \sup_{|z_1 - z_2| \leq \eta_n} \left| G_{\widehat{f}}(z_1 \mid \mathbf{X}) - G_{\widehat{f}}(z_2 \mid \mathbf{X}) \right| \\
&\leq \sup_{|z_1 - z_2| \leq \eta_n} \left| G_f(z_1 + \eta_n \mid \mathbf{X}) - G_f(z_2 - \eta_n \mid \mathbf{X}) \right| \\
&\leq \quad \sup_z \left| G_f(z + 2\eta_n \mid \mathbf{X}) - G_f(z - 2\eta_n \mid \mathbf{X}) \right| = o_P(1).
\end{aligned}$$

Since $B_n \subset \{\sup_{\mathbf{x} \in \bar{A}(\mathbf{X})} \sup_y |\widehat{f}(y \mid \mathbf{x}) - f(y \mid \mathbf{x})| \geq \eta_n\}$, by Assumption 3,

$$\mathbb{P}(B_n) \leq \mathbb{P}\big(\sup_{\mathbf{x} \in \bar{A}(\mathbf{X})} \sup_y \left| \widehat{f}(y \mid \mathbf{x}) - f(y \mid \mathbf{x}) \right| \geq \eta_n \big) \leq \rho_n = o(1).$$

Consequently, $\mathrm{I}_1 = o_P(1)$ and $\mathrm{I}_2 = o_P(1)$.

For I_3, let $G_{\widehat{f}}(t \mid \bar{A}(\mathbf{X})) = \mathbb{P}\{\widehat{f}(Y \mid \mathbf{X}) \leq t \mid \mathbf{X} \in \bar{A}(\mathbf{X})\}$ denote the conditional CDF of $\widehat{f}(Y \mid \mathbf{X})$ given $\bar{A}(\mathbf{X})$. By the triangle inequality,

$$\mathrm{I}_3 \leq \sup_z \left|\widehat{G}_{\widehat{f}}(z \mid \bar{A}(\mathbf{X})) - G_{\widehat{f}}(z \mid \mathbf{X})\right|$$

$$\leq \sup_z \left|\widehat{G}_{\widehat{f}}(z \mid \bar{A}(\mathbf{X})) - G_{\widehat{f}}(z \mid \bar{A}(\mathbf{X}))\right| + \sup_z \left|G_{\widehat{f}}(z \mid \bar{A}(\mathbf{X})) - G_{\widehat{f}}(z \mid \mathbf{X})\right|.$$

For the first term, since the weights are a.s. bounded under Assumption 2, Lemma S.2 of Yi et al. (2025) (the weighted DKW inequality under distribution shift) implies that for any $\varepsilon > 0$,

$$\begin{aligned}&\mathbb{P}\big\{ \sup_z \left|\widehat{G}_{\widehat{f}}(z \mid \bar{A}(\mathbf{X})) - G_{\widehat{f}}(z \mid \bar{A}(\mathbf{X}))\right| \geq \varepsilon\big\} \\ \leq& 2\left\{\left\lceil \frac{2}{(\varepsilon - r_n)_+}\right\rceil + 1\right\} \exp\{-C\bar{n}(\varepsilon - r_n)_+^2\},\end{aligned}$$

where $\bar{n}$ denotes the number of samples in the subset $\bar{A}(\mathbf{X})$, $r_n = O(\bar{n}^{-1})$, and the constant C depends on the upper bound of the weights. Under Assumption 5, $\bar{n} \to \infty$, it follows that $\sup_z \left|\widehat{G}_{\widehat{f}}(z \mid \bar{A}(\mathbf{X})) - G_{\widehat{f}}(z \mid \bar{A}(\mathbf{X}))\right| = o_P(1)$. For the second term, by the law of iterated expectations, we have

$$\begin{aligned}&\sup_z \left|G_{\widehat{f}}(z \mid \bar{A}(\mathbf{X})) - G_{\widehat{f}}(z \mid \mathbf{X})\right| \\ =\ & \sup_z \left|\mathbb{E}\big[G_{\widehat{f}}(z \mid \mathbf{X}') - G_{\widehat{f}}(z \mid \mathbf{X}) \mid \mathbf{X}' \in \bar{A}(\mathbf{X})\big]\right| \\ \leq\ & \sup_z \mathbb{E}\Big[\left|G_{\widehat{f}}(z \mid \mathbf{X}') - G_{\widehat{f}}(z \mid \mathbf{X})\right| \mid \mathbf{X}' \in \bar{A}(\mathbf{X})\Big] \\ \leq\ & \sup_z \mathbb{E}\Big[\left|G_{\widehat{f}}(z \mid \mathbf{X}') - G_f(z \mid \mathbf{X}')\right| \mid \mathbf{X}' \in \bar{A}(\mathbf{X})\Big] \qquad \text{(S1.13)} \\ & + \sup_z \mathbb{E}\Big[\left|G_f(z \mid \mathbf{X}') - G_f(z \mid \mathbf{X})\right| \mid \mathbf{X}' \in \bar{A}(\mathbf{X})\Big] \qquad \text{(S1.14)}\end{aligned}$$

$$+\sup_z \mathbb{E}\Big[\big|G_f(z \mid \mathbf{X}) - G_{\widehat{f}}(z \mid \mathbf{X})\big| \mid \mathbf{X}' \in \bar{A}(\mathbf{X})\Big], \qquad \text{(S1.15)}$$

where (S1.13) and (S1.15) converge to zero in probability by the local uniform consistency in Assumption 3, while (S1.14) also converges to zero due to the Lipschitz continuity and diameter shrinking in Assumptions 4–5. It then follows that $\mathrm{I}_3 = o_P(1)$. Putting all the pieces together, we have

$$\sup_{y\in\mathcal{Y}} \big|G_f(f(y \mid \mathbf{X}) \mid \mathbf{X}) - \widehat{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}) \mid \bar{A}(\mathbf{X}))\big| = o_P(1).$$

Let $\lambda_n = o(1)$ be a sequence such that $D_n = \{\sup_{y\in\mathcal{Y}} |G_f(f(y|\mathbf{X}) \mid \mathbf{X}) - \widehat{G}_{\widehat{f}}(\widehat{f}(y|\mathbf{X}) \mid \bar{A}(\mathbf{X}))| > \lambda_n\}$ and $\mathbb{P}(D_n) = o(1)$. Therefore, on the event D_n^c, it follows from the definition of the prediction set that

$$\begin{aligned} C(\mathbf{X};\alpha)\Delta\widehat{C}(\mathbf{X};\alpha) = &\big\{y : G_f(f(y \mid \mathbf{X}) \mid \mathbf{X}) > \alpha,\ \widehat{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}) \mid \bar{A}(\mathbf{X})) \le \alpha\big\} \\ &\cup \big\{y : G_f(f(y \mid \mathbf{X}) \mid \mathbf{X}) \le \alpha,\ \widehat{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}) \mid \bar{A}(\mathbf{X})) > \alpha\big\} \\ \subseteq &\ \big\{y : \big|G_f\{f(y \mid \mathbf{X}) \mid \mathbf{X}\} - \alpha\big| \le \lambda_n\big\}. \end{aligned}$$

Under Assumption 4, where G_f is Lipschitz continuous, we have

$$\begin{aligned} \mathbb{P}\big\{Y \in C(\mathbf{X};\alpha)\Delta\widehat{C}(\mathbf{X};\alpha)\big\} \le &\ \mathbb{P}\big\{Y \in C(\mathbf{X};\alpha)\Delta\widehat{C}(\mathbf{X};\alpha), D_n^c\big\} + \mathbb{P}(D_n) \\ \le &\ \mathbb{P}\left(|G_f\{f(Y \mid \mathbf{X}) \mid \mathbf{X}\} - \alpha| \le \lambda_n\right) + o(1) \\ \le &\ 2\lambda_n + o(1) = o(1). \end{aligned}$$

□

Lemma 2. *Define the weighted empirical CDFs based on the true weights and the estimated weights of $\widehat{f}(Y_{n+1}|\mathbf{X}_{n+1})$ as follows:*

$$\begin{aligned}\widehat{G}_{\widehat{f}}(t) = \sum_{i=1}^{n} & \frac{w(\mathbf{X}_i, Y_i)}{\sum_{j=1}^{n} w(\mathbf{X}_j, Y_j) + w(\mathbf{X}_{n+1}, y)} \mathbb{I}\{\widehat{f}(Y_i \mid \mathbf{X}_i) \leq t\} \\ & + \frac{w(\mathbf{X}_{n+1}, y)}{\sum_{j=1}^{n} w(\mathbf{X}_j, Y_j) + w(\mathbf{X}_{n+1}, y)} \mathbb{I}\{\widehat{f}(y \mid \mathbf{X}_{n+1}) \leq t\}, \\ \widetilde{G}_{\widehat{f}}(t) = \sum_{i=1}^{n} & \frac{\widehat{w}(\mathbf{X}_i, Y_i)}{\sum_{j=1}^{n} \widehat{w}(\mathbf{X}_j, Y_j) + \widehat{w}(\mathbf{X}_{n+1}, y)} \mathbb{I}\{\widehat{f}(Y_i \mid \mathbf{X}_i) \leq t\} \\ & + \frac{\widehat{w}(\mathbf{X}_{n+1}, y)}{\sum_{j=1}^{n} \widehat{w}(\mathbf{X}_j, Y_j) + \widehat{w}(\mathbf{X}_{n+1}, y)} \mathbb{I}\{\widehat{f}(y \mid \mathbf{X}_{n+1}) \leq t\},\end{aligned}$$

where y is a candidate test point for Y_{n+1}. Then, under Assumption 2,

$$\sup_{y \in \mathcal{Y}} \left| \widehat{G}_{\widehat{f}}\big(\widehat{f}(y \mid \mathbf{X}_{n+1})\big) - \widetilde{G}_{\widehat{f}}\big(\widehat{f}(y \mid \mathbf{X}_{n+1})\big) \right| = o_P(1).$$

Proof of Lemma 2. To simplify notation, let

$$w_i = w(\mathbf{X}_i, Y_i), \quad \widehat{w}_i = \widehat{w}(\mathbf{X}_i, Y_i), \qquad i = 1, \ldots, n,$$

$$w_{n+1}(y) = w(\mathbf{X}_{n+1}, y), \quad \widehat{w}_{n+1}(y) = \widehat{w}(\mathbf{X}_{n+1}, y).$$

Define $S_{n+1}(y) = \sum_{i=1}^{n} w_i + w_{n+1}(y)$ and $\widehat{S}_{n+1}(y) = \sum_{i=1}^{n} \widehat{w}_i + \widehat{w}_{n+1}(y)$. Based on the definitions of $\widehat{G}_{\widehat{f}}(t)$ and $\widetilde{G}_{\widehat{f}}(t)$, we have

$$\begin{aligned}& \sup_{y \in \mathcal{Y}} |\widehat{G}_{\widehat{f}}\big(\widehat{f}(y \mid \mathbf{X}_{n+1})\big) - \widetilde{G}_{\widehat{f}}\big(\widehat{f}(y \mid \mathbf{X}_{n+1})\big)| \\ \leq & \sup_{y \in \mathcal{Y}} \underbrace{\sum_{i=1}^{n} \left| \frac{\widehat{w}_i}{\widehat{S}_{n+1}(y)} - \frac{w_i}{S_{n+1}(y)} \right|}_{\mathrm{II}_1(y)} + \sup_{y \in \mathcal{Y}} \underbrace{\left| \frac{\widehat{w}_{n+1}(y)}{\widehat{S}_{n+1}(y)} - \frac{w_{n+1}(y)}{S_{n+1}(y)} \right|}_{\mathrm{II}_2(y)}.\end{aligned}$$

Furthermore, we can compute $\mathrm{II}_1(y)$ and $\mathrm{II}_2(y)$ as follows.

$$\begin{aligned}
\mathrm{II}_1(y) = &\ \sum_{i=1}^{n}\left|\frac{\widehat{w}_i\ S_{n+1}(y) - w_i\ \widehat{S}_{n+1}(y)}{\widehat{S}_{n+1}(y)S_{n+1}(y)}\right| \\
= &\ \sum_{i=1}^{n}\left|\frac{\widehat{w}_i\ \{S_{n+1}(y) - \widehat{S}_{n+1}(y)\} + \widehat{S}_{n+1}(y)(\widehat{w}_i - w_i)}{\widehat{S}_{n+1}(y)S_{n+1}(y)}\right| \\
\leq &\ \frac{\widehat{S}_n\ |S_{n+1}(y) - \widehat{S}_{n+1}(y)| + \widehat{S}_{n+1}(y)\sum_{i=1}^{n}|\widehat{w}_i - w_i|}{\widehat{S}_{n+1}(y)S_{n+1}(y)} \\
\leq &\ \frac{|S_{n+1}(y) - \widehat{S}_{n+1}(y)|}{S_{n+1}(y)} + \frac{\sum_{i=1}^{n}|\widehat{w}_i - w_i|}{S_{n+1}(y)} \\
\leq &\ \frac{2\sum_{i=1}^{n}|\widehat{w}_i - w_i|}{S_{n+1}(y)} + \frac{|w_{n+1}(y) - \widehat{w}_{n+1}(y)|}{S_{n+1}(y)}.
\end{aligned}$$

Similarly, we can derive that:

$$\mathrm{II}_2(y) \leq \frac{\sum_{i=1}^{n}|\widehat{w}_i - w_i|}{S_{n+1}(y)} + \frac{2|w_{n+1}(y) - \widehat{w}_{n+1}(y)|}{S_{n+1}(y)}.$$

Under Assumption 2, there exists a constant $C_1 > 0$ such that the weight function is a.s. bounded below by C_1. Combining $\mathrm{II}_1(y)$ and $\mathrm{II}_2(y)$ gives

$$\begin{aligned}
&\sup_{y\in\mathcal{Y}}\left|\widehat{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}_{n+1})) - \widetilde{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}_{n+1}))\right| \\
\leq &\ 3\Big(\frac{\sum_{i=1}^{n}|\widehat{w}_i - w_i|}{S_n} + \frac{\sup_{y\in\mathcal{Y}}|\widehat{w}_{n+1}(y) - w_{n+1}(y)|}{S_n}\Big) \\
\leq &\ \frac{3}{C_1}\Big(\underbrace{\frac{1}{n}\sum_{i=1}^{n}|\widehat{w}_i - w_i|}_{\mathrm{II}_3} + \underbrace{\frac{1}{n}\sup_{y\in\mathcal{Y}}|\widehat{w}_{n+1}(y) - w_{n+1}(y)|}_{\mathrm{II}_4}\Big).
\end{aligned}$$

By the Cauchy–Schwarz inequality, we obtain

$$\mathbb{E}(\mathrm{II}_3) = \mathbb{E}|\widehat{w}(\mathbf{X}, Y) - w(\mathbf{X}, Y)| \leq (\mathbb{E}|\widehat{w}(\mathbf{X}, Y) - w(\mathbf{X}, Y)|^2)^{1/2}.$$

Let $E_n = \{\sup_{y\in\mathcal{Y}} |\widehat{w}(\mathbf{X}, y) - w(\mathbf{X}, y)| \geq \eta_n^{1/3}\}$ and $H_n = \{\mathbb{E}[\sup_{y\in\mathcal{Y}} |\widehat{w}(\mathbf{X}, y) - w(\mathbf{X}, y)|^2 \mid \widehat{w}] \geq \eta_n\}$. Under Assumption 2, there exists a constant $C_2 > 0$ such that $|\widehat{w}(\mathbf{X}, Y) - w(\mathbf{X}, Y)|^2 \leq C_2$ almost surely. Hence,

$$\begin{aligned}\mathbb{E}\,|\widehat{w}(\mathbf{X}, Y) - w(\mathbf{X}, Y)|^2 &= \mathbb{E}\left\{|\widehat{w}(\mathbf{X}, Y) - w(\mathbf{X}, Y)|^2 \cdot \mathbb{I}(H_n)\right\} \\ &\qquad + \mathbb{E}\left\{|\widehat{w}(\mathbf{X}, Y) - w(\mathbf{X}, Y)|^2 \cdot \mathbb{I}(H_n^c)\right\} \\ &\leq C_2\rho_n + \eta_n \mathbb{P}(H_n^c) \ \to \ 0.\end{aligned}$$

Therefore, by Markov's inequality, we have for any $\varepsilon > 0$,

$$\mathbb{P}(\mathrm{II}_3 \geq \varepsilon) \leq \frac{\mathbb{E}(\mathrm{II}_3)}{\varepsilon} \leq \frac{1}{\varepsilon}\left(\mathbb{E}|\widehat{w}(\mathbf{X}, Y) - w(\mathbf{X}, Y)|^2\right)^{1/2} \to 0.$$

Moreover, by the consistency assumption in Assumption 2, we obtain

$$\begin{aligned}&\mathbb{P}(E_n) = \mathbb{E}\{\mathbb{P}(E_n \mid \widehat{w})\mathbb{I}\left(H_n\right)\} + \mathbb{E}\{\mathbb{P}(E_n \mid \widehat{w})\mathbb{I}\left(H_n^c\right)\} \\ &\leq \mathbb{P}\left(H_n\right) + \mathbb{E}\left[\frac{\mathbb{E}\left[\sup_{y\in\mathcal{Y}} |\widehat{w}(\mathbf{X}, y) - w(\mathbf{X}, y)|^2 \mid \widehat{w}\right]}{\eta_n^{2/3}}\mathbb{I}\left(H_n^c\right)\right] \\ &\leq \rho_n + \eta_n^{1/3} = o(1).\end{aligned}$$

Therefore, for any $\varepsilon > 0$ and sufficiently large n, we have $\mathbb{P}(\mathrm{II}_4 \geq \varepsilon) \leq \mathbb{P}(E_n) = o(1)$. Combining all the above results, we obtain

$$\sup_{y\in\mathcal{Y}} \left|\widehat{G}_{\widehat{f}}(\widehat{f}(y|\mathbf{X}_{n+1})) - \widetilde{G}_{\widehat{f}}(\widehat{f}(y|\mathbf{X}_{n+1}))\right| = o_P(1).$$

□

Proof of Theorem 3. Define the local empirical CDF of $\widehat{f}(Y_{n+1} \mid \mathbf{X}_{n+1})$ over the local dataset $\bar{A}(\mathbf{X}_{n+1})$ in (S1.8) as

$$\widehat{G}_{\widehat{f}}(t \mid \bar{A}(\mathbf{X}_{n+1})) = \sum_{i \in A_n(\mathbf{X}_{n+1})} \varpi_i(y)\, \mathbb{I}\{\widehat{f}(Y_i \mid \mathbf{X}_i) \leq t\}$$

$$+ \varpi_{n+1}(y)\, \mathbb{I}\{\widehat{f}(y \mid \mathbf{X}_{n+1}) \leq t\},$$

where y is a candidate of Y_{n+1}. The normalized weights are defined as

$$\varpi_i(y) = \frac{w(\mathbf{X}_i, Y_i)}{\sum_{j \in A_n(\mathbf{X}_{n+1})} w(\mathbf{X}_j, Y_j) + w(\mathbf{X}_{n+1}, y)},$$

$$\varpi_{n+1}(y) = \frac{w(\mathbf{X}_{n+1}, y)}{\sum_{j \in A_n(\mathbf{X}_{n+1})} w(\mathbf{X}_j, Y_j) + w(\mathbf{X}_{n+1}, y)}.$$

By a basic property of distribution functions, the prediction set proposed by Algorithm 2 can be expressed as:

$$\widehat{C}(\mathbf{X}_{n+1}; \alpha) = \big\{y \in \mathcal{Y} : \widehat{G}_{\widehat{f}}(\widehat{f}(y \mid \mathbf{X}_{n+1}) \mid \bar{A}(\mathbf{X}_{n+1})) > \alpha\big\}.$$

It follows from Lemma 1 that $\mathbb{P}\{Y_{n+1} \in C(\mathbf{X}_{n+1}; \alpha) \Delta \widehat{C}(\mathbf{X}_{n+1}; \alpha)\} = o(1)$. Furthermore, according to Lemma 28 in Izbicki et al. (2022), we obtain that $\widehat{C}(\mathbf{X}_{n+1}; \alpha)$ satisfies asymptotic conditional validity. That is, there exists a sequence of sets $\Lambda_n \in \mathcal{X}$ such that $\mathbb{P}\left(\mathbf{X}_{n+1} \in \Lambda_n\right) = 1 - o(1)$, and

$$\sup_{\mathbf{x}_{n+1} \in \Lambda_n} \left| \mathbb{P}\big\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}; \alpha) \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}\big\} - (1 - \alpha) \right| = o(1).$$

Therefore, the conclusion of Theorem 3 is established under the true weight function. Lemma 2 further shows that replacing the true weight function with its estimator preserves asymptotic validity.

□

S2 High-Dimensional Extensions

S2.1 Theoretical Discussion

Our method relies on estimating two components: (1) the propensity score $\mathbb{P}(\delta = 1 \mid \mathbf{X}, Y)$, and (2) the conditional density $f(y \mid \mathbf{x})$. In the main paper, we focus on finite-dimensional covariate settings, and here we discuss how to estimate these components in high-dimensional scenarios and the related theoretical properties.

(1). Propensity Score Estimation

Since nonignorable missingness causes identification issues even in low-dimensional settings, estimating the propensity score in high dimensions becomes even more challenging. Existing methods can address this problem primarily under low-dimensional structural assumptions (such as sparsity), via a two-step strategy: first performing screening or dimension reduction, and then estimating the model in the reduced space. For example, Ding et al. (2020) identifies covariates associated with δ using a Pearson χ^2 statistic, whereas Wang et al. (2021) employs sufficient dimension reduction to simplify both the $Y \mid \mathbf{X}$ and $\delta \mid (\mathbf{X}, Y)$ models. After dimension reduction, both studies construct their propensity score estimators within the instrumental-variable framework of Shao and Wang (2016). In the simulation studies in Section S2.2, we adopt the dimension reduction approach of Wang et al. (2021) following an initial feature screening. Under standard regularity conditions (e.g., conditions for kernel estimation) and a dimension-reduction step with accuracy guarantees, the resulting estimators are consistent and satisfy Assumption 2 in the main paper.

(2). Conditional Density Estimation

In the main paper, we develop a conditional density estimator that is reformulated in terms of the conditional quantile $\widehat{Q}_Y(\tau \mid \mathbf{x}, \delta = 1)$ estimated from the observed data, as given in equations (2.8) and (2.10). This reformulation enables us to leverage existing conditional quantile estimation methods to accommodate high-dimensional settings, and the choice of estimator does not affect the validity of the coverage guarantee. We may use the classical ℓ_1-penalized quantile regression of Belloni and Chernozhukov (2011), which achieves an estimation error of $O_P(\sqrt{(s \log p)/n})$ under standard conditions and a sparsity assumption. Here, p is the number of covariates, s is the number of nonzero coefficients in the true model, and $s \ll p$. Building on this work, Wang et al. (2012) and Tan et al. (2022) achieve notable theoretical and computational improvements by introducing nonconvex penalties such as SCAD and MCP, with the latter further enhanced by convolution smoothing and multi-step weighted ℓ_1. We can also use generalized random forests (Athey et al., 2019), which can handle relatively large p, but their theoretical guarantees are derived under the assumption that p is fixed. In addition, we may incorporate recent methods in high-dimensional quantile regression reviewed in Qiu et al. (2026) into our framework in appropriate settings.

Another efficient strategy in ultrahigh-dimensional settings is to perform feature screening before fitting the regression model. Zhang et al. (2020) show that under nonignorable missingness with a logistic propensity model, the pseudo-response $Y\delta$ can be used for screening because the active set of $Y \mid \mathbf{X}$ is contained in that of $Y\delta \mid \mathbf{X}$. Therefore, under standard regularity conditions and using a screening procedure with the sure-screening property, any downstream regression estimator will remain consistent when the selected model contains the true active set. Following this strategy, in the simulation studies in Section S2.2, we first perform feature screening and then fit linear quantile regression (Koenker, 2005) and quantile regression forests (Athey et al., 2019) using the reduced covariates.

The above methods yield consistent estimates of the conditional quantiles in high-dimensional settings under the regularity conditions specified in their respective papers. Consequently, under these conditions, our conditional density estimator can be shown to satisfy Assumption 3 in the main paper. In addition, by directly using a conditional density estimator, one can implement Algorithm 2 in high-dimensional settings using K-means clustering based on the profile distance Izbicki et al. (2022). If the number of clusters satisfies $K_n \to \infty$ and $K_n = o(n)$, then under mild regularity conditions, such as the marginal density of $\mathbf{X}$ being bounded away from zero and infinity on a compact set, the partition produced by Algorithm 2 can be shown to satisfy Assumption 5 with probability tending to one.

S2.2 Simulation Results

This section presents simulation studies to evaluate our method in high-dimensional settings. We employ a two-step strategy: first performing feature screening to reduce the dimensionality to a finite set of covariates, and then estimating the propensity score and conditional density in the resulting reduced space.

We follow the data-generating mechanism in Section 4 of the main paper but increase the covariate dimension to $p = 300$. The covariates are generated as $U_{ij} \sim N(Z_i/10,\, 1), j = 1, \ldots, 300$, where Z_i is an instrumental variable. As in Section 4, we consider both linear and nonlinear models, with the true outcome Y_i and the missingness indicator depending only on the first 10 components of $\mathbf{U}_i$. In the feature screening step, we fit a penalized linear regression of Y on $\mathbf{X}$ using the observed training data with an Elastic Net penalty, and directly retain the 15 covariates with the largest estimated coefficients. We then proceed in this reduced low-dimensional space as in Section 4 of the main paper: the propensity score is estimated using the method of Wang et al. (2021), and the conditional density is obtained via equations (2.8) and (2.10). For conditional quantile estimation, we use linear quantile regression (LQR) in the linear model and quantile random forests (QRF) in the nonlinear model, respectively.

We compare the four methods considered in the main paper, namely OMNI, Naive, MAR-CP, and MNAR-CP. The results for the linear model with a sample size of $n = 1000$ are presented in Table S1, while those for the nonlinear model with a sample size of $n = 4000$ are presented in Table S2. Consistent with the low-dimensional findings, the Naive and MAR-CP methods usually fail to reach the nominal coverage because they do not correctly adjust for the bias caused by nonignorable missingness, either by ignoring it or by applying a misspecified correction. Both OMNI and the proposed MNAR-CP method achieve the nominal coverage. However, our intervals tend to be wider because OMNI uses fully observed samples and avoids weight estimation, which is unrealistic in practice. In contrast, our method relies on a smaller effective sample size and has additional estimation error that becomes more pronounced in high-dimensional settings. As the sample size increases, our interval lengths become closer to those of OMNI, as shown in Table S2.

Table S1: Average Coverage percentages (AC%) and Average Length (AL) for 90% prediction sets in the high-dimensional setting with $p = 300$ and $n = 1000$, averaged over 500 new subjects and 500 repetitions under the linear model using the LQR estimator.

	Miss.(%)	OMNI	Naive	MAR-CP	MNAR-CP
			(a) Homoskedastic		
AC% (SE×100)	40	90.12(0.08)	88.29(0.11)	87.79(0.24)	**89.68(0.12)**
	50	90.12(0.08)	87.94(0.13)	87.09(0.36)	**89.76(0.14)**
	60	90.12(0.08)	87.40(0.15)	85.75(0.47)	**89.61(0.15)**
AL (SE)	40	3.48(0.01)	3.51(0.01)	3.67(0.06)	5.45(0.12)
	50	3.48(0.01)	3.55(0.01)	3.67(0.02)	6.29(0.15)
	60	3.48(0.01)	3.61(0.01)	3.82(0.05)	6.78(0.16)
			(b) Heteroscedastic		
AC% (SE×100)	40	90.00(0.09)	84.01(0.15)	82.91(0.41)	**88.80(0.18)**
	50	90.00(0.09)	83.23(0.17)	82.02(0.43)	**89.14(0.20)**
	60	90.00(0.09)	82.47(0.20)	81.01(0.48)	**89.34(0.21)**
AL (SE)	40	2.73(0.01)	2.29(0.01)	2.52(0.05)	7.93(0.21)
	50	2.73(0.01)	2.31(0.03)	2.63(0.07)	8.96(0.23)
	60	2.73(0.01)	2.45(0.06)	2.94(0.09)	10.11(0.23)

OMNI: Standard conformal prediction applied to the complete data. Naive: Standard conformal prediction applied to the observed data. MAR-CP: Weighted conformal prediction with MAR weights. MNAR-CP: Proposed weighted conformal prediction with MNAR weights. Values in parentheses are standard errors.

Table S2: Average Coverage percentages (AC%) and Average Length (AL) for 90% prediction sets in the high-dimensional setting with $p = 300$ and $n = 4000$, averaged over 500 new subjects and 500 repetitions under the nonlinear model using the QRF estimator.

	Miss.(%)	OMNI	Naive	MAR-CP	MNAR-CP
			(a) Homoskedastic		
AC% (SE×100)	20	90.01(0.07)	84.16(0.08)	88.01(0.31)	**90.89(0.11)**
	30	89.99(0.06)	82.71(0.09)	87.50(0.41)	**90.90(0.12)**
	40	90.02(0.07)	81.12(0.09)	85.58(0.63)	**90.40(0.12)**
AL (SE)	20	11.07(0.01)	9.33(0.01)	13.51(0.42)	11.72(0.06)
	30	11.06(0.01)	9.01(0.01)	14.78(0.51)	11.86(0.09)
	40	11.05(0.01)	8.68(0.01)	14.27(0.49)	11.65(0.06)
			(b) Heteroscedastic		
AC% (SE×100)	20	89.92(0.07)	82.34(0.09)	86.64(0.52)	**90.84(0.12)**
	30	89.94(0.06)	80.57(0.09)	87.18(0.59)	**90.82(0.13)**
	40	89.93(0.06)	78.55(0.10)	85.16(0.75)	**90.13(0.14)**
AL (SE)	20	10.58(0.01)	8.45(0.01)	13.17(0.44)	11.17(0.06)
	30	10.59(0.01)	8.08(0.01)	15.66(0.56)	11.42(0.10)
	40	10.57(0.01)	7.69(0.01)	14.80(0.52)	11.43(0.12)

OMNI: Standard conformal prediction applied to the complete data. Naive: Standard conformal prediction applied to the observed data. MAR-CP: Weighted conformal prediction with MAR weights. MNAR-CP: Proposed weighted conformal prediction with MNAR weights. Values in parentheses are standard errors.

S3 Additional Numerical Studies

S3.1 Conditional Coverage

This section investigates the conditional coverage performance of the proposed method, including the non-localized prediction from Algorithm 1 and the localized prediction from Algorithm 2.

We generate the covariates $\mathbf{X} = (\mathbf{Z}, \mathbf{U})$ and the missingness indicator $\delta \sim \text{Bern}(\pi)$ following the same setup as in the simulation study in Section 4 of the main paper. For clarity, we focus on linear model with $a = -3.5$, resulting in the following model:

$$Y_i = Z_i + \textstyle\sum_{j=1}^{10} U_{ij} + \epsilon_i,$$

$$\pi_i = \pi(\mathbf{U}_i, Y_i) = 1/\big\{1 + \exp\big(- 3.5 - 0.1 \textstyle\sum_{j=1}^{10} U_{ij} + 0.65 Y_i\}\big)\big\},$$

where we consider only homoscedastic errors with $\epsilon_i \sim N(0, 1)$ and a sample size of $n = 1000$. This setting corresponds to a 40% missing rate. For the proposed method, we use the same procedures for weight and conditional density estimation as in Section 4 of the main paper, with conditional quantiles estimated only via linear quantile regression. Let K be the number of clusters in Algorithm 2. We compare the prediction performance for $K = 1$ (i.e., the non-localized prediction in Algorithm 1), $K = 5$, and $K = 10$.

To evaluate conditional coverage, we fix the covariates for 20 new subjects. These covariates are generated as follows: we first create 20 equally spaced quantile levels ranging from 0.01 to 0.99. For each quantile level, we compute the corresponding marginal quantile for each covariate dimension based on its true distribution. This yields 20 covariate vectors, each consisting of marginal quantiles taken at the same quantile level across all dimensions. The average results over 1000 repeated experiments for each fixed covariate test point are shown in Figure S1, where the x-axis represents the 20 covariate vectors generated at quantile levels from 0.01 to 0.99.

It is observed that the conditional coverage is slightly low for subjects on the left side of Figure S1. This is because these new subjects have extreme covariate values—some near the 0.01 quantile—where training data are sparse and predictions are less reliable without model assumptions. However, as the number of clusters K increases, the conditional coverage improves, which aligns with our theory—larger K provides a better approximation to the conditional distribution, leading to more reliable conditional coverage guarantees for any given covariate. In practice, the selection of K is guided by the predictive objective: if only marginal coverage over the entire covariate space is required, setting $K = 1$ is adequate. However, if conditional coverage is desired for a specific subject—particularly one with atypical covariates—a larger K is generally more appropriate (e.g., 5 to 10 when $n = 1000$) and may be further increased as the sample size grows.

S3.2 Additional Simulation Results

This section presents additional simulation results that complement Section 4 of the main paper. While the main paper reports marginal coverage for the linear model with $n = 1000$ and the nonlinear model with $n = 4000$, we also include results for the nonlinear model with $n = 1000$ in Table S3 and for the linear model with $n = 4000$ in Table S4. In addition, Figure S2 illustrates examples of conditional density estimates under different settings. Since these findings are consistent with those in the main paper, we omit further discussion.

Bibliography

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Belloni, A. and V. Chernozhukov (2011). l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 82–130.

Ding, X., J. Chen, and X. Chen (2020). Regularized quantile regres-

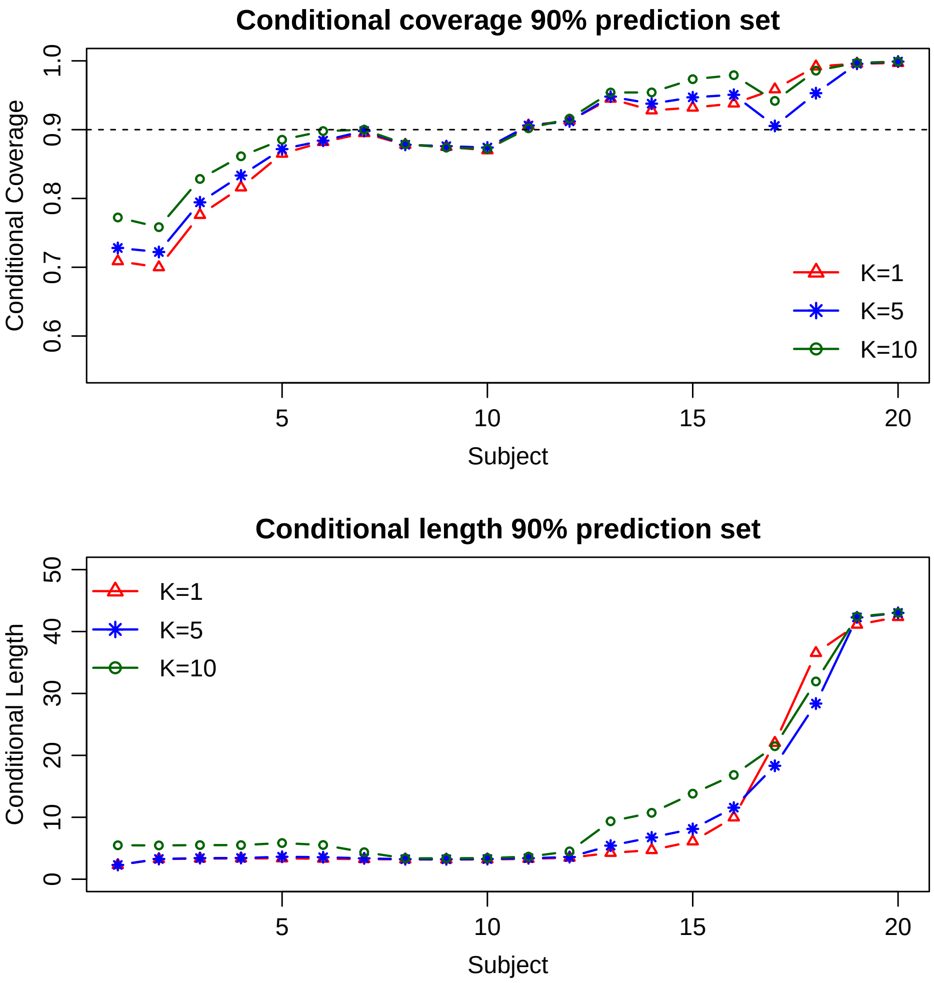

Figure S1: Comparison of conditional coverage and conditional length across different numbers of clusters K. The x-axis represents 20 covariate vectors, each formed by taking the same quantile level (from 0.01 to 0.99) across all variables.

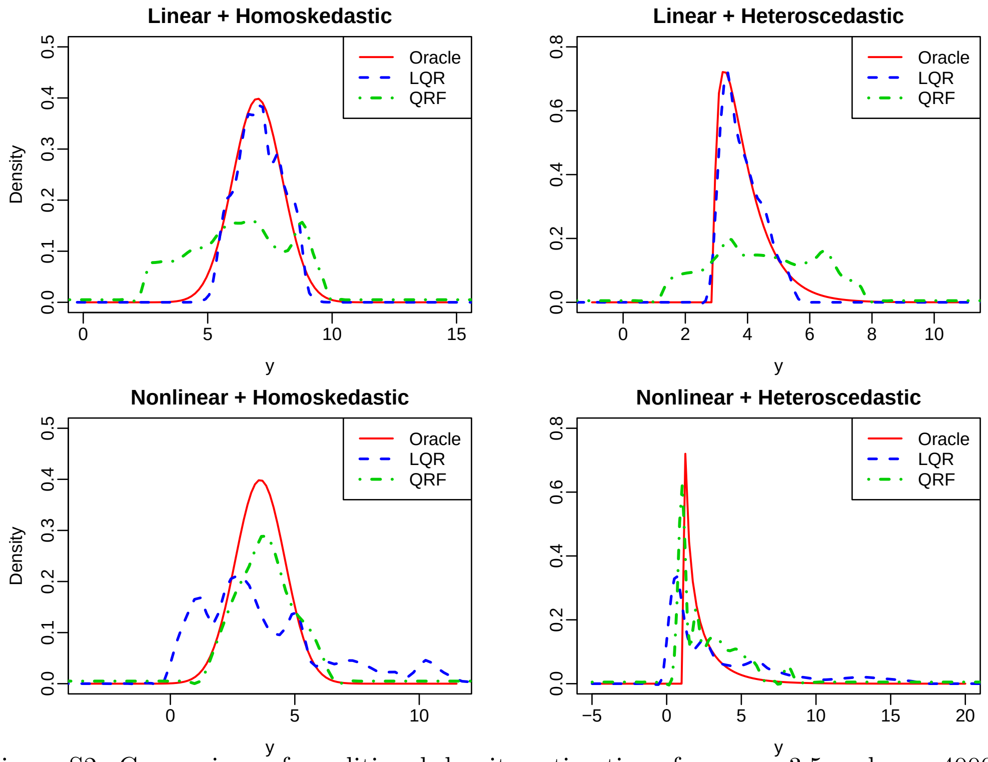

Figure S2: Comparison of conditional density estimations for $a = -3.5$ and $n = 4000$. The oracle density is represented by the red solid line, LQR by the blue dashed line, QRF by the green dash-dotted line.

Table S3: Average Coverage percentages (AC%) and Average Length (AL) for 90% prediction sets, across different missing rates and $\widehat{f}$ estimators, averaged over 500 new subjects and 500 repetitions in nonlinear model with $n = 1000$.

	Miss.(%)	$\widehat{f}$	OMNI	Naive	MAR-CP	MNAR-CP
				(a) Homoskedastic		
AC% (SE×100)	20	LQR	90.06(0.08)	84.77(0.11)	86.75(0.28)	**90.50(0.14)**
		QRF	90.12(0.08)	84.44(0.11)	86.19(0.30)	**90.13(0.15)**
	30	LQR	90.06(0.08)	83.51(0.11)	86.14(0.31)	**90.06(0.16)**
		QRF	90.05(0.09)	82.98(0.11)	85.75(0.29)	**89.64(0.16)**
	40	LQR	90.12(0.08)	82.23(0.12)	86.73(0.42)	**89.79(0.18)**
		QRF	90.08(0.08)	81.35(0.11)	85.89(0.42)	**89.27(0.18)**
AL (SE)	20	LQR	11.61(0.04)	9.31(0.03)	12.25(0.38)	20.54(0.61)
		QRF	11.70(0.03)	9.71(0.02)	11.16(0.21)	12.91(0.21)
	30	LQR	11.61(0.04)	8.97(0.03)	12.50(0.41)	21.55(0.62)
		QRF	11.68(0.03)	9.35(0.02)	11.32(0.24)	13.31(0.27)
	40	LQR	11.75(0.04)	8.72(0.03)	15.20(0.54)	23.26(0.66)
		QRF	11.72(0.03)	9.02(0.02)	12.83(0.38)	14.00(0.34)
				(b) Heteroscedastic		
AC% (SE×100)	20	LQR	89.86(0.09)	83.06(0.11)	85.96(0.30)	**90.30(0.17)**
		QRF	89.92(0.09)	82.53(0.11)	85.23(0.35)	**89.85(0.16)**
	30	LQR	89.86(0.09)	81.60(0.13)	86.01(0.37)	**89.66(0.19)**
		QRF	89.91(0.09)	80.77(0.11)	85.11(0.41)	**89.07(0.18)**
	40	LQR	89.97(0.09)	80.16(0.13)	84.70(0.52)	**89.41(0.20)**
		QRF	90.04(0.08)	79.02(0.13)	83.80(0.51)	**88.52(0.21)**
AL (SE)	20	LQR	10.88(0.04)	8.13(0.03)	12.07(0.46)	22.17(0.66)
		QRF	11.20(0.03)	8.80(0.02)	11.00(0.28)	12.91(0.30)
	30	LQR	10.88(0.04)	7.76(0.03)	13.19(0.50)	22.20(0.67)
		QRF	11.20(0.03)	8.41(0.02)	11.51(0.32)	13.56(0.37)
	40	LQR	10.91(0.04)	7.51(0.03)	14.21(0.56)	20.31(0.65)
		QRF	11.23(0.02)	8.04(0.02)	12.00(0.40)	12.73(0.33)

OMNI: Standard conformal prediction applied to the complete data. Naive: Standard conformal prediction applied to the observed data. MAR-CP: Weighted conformal prediction with MAR weights. MNAR-CP: Proposed weighted conformal prediction with MNAR weights. Values in parentheses are standard errors.

Table S4: Average Coverage percentages (AC%) and Average Length (AL) for 90% prediction sets, across different missing rates and $\widehat{f}$ estimators, averaged over 500 new subjects and 500 repetitions in linear model with $n = 4000$.

	Miss.(%)	$\widehat{f}$	OMNI	Naive	MAR-CP	MNAR-CP
				(a) Homoskedastic		
AC% (SE×100)	40	LQR	89.99(0.07)	87.85(0.08)	84.77(0.59)	**89.54(0.08)**
		QRF	90.02(0.07)	79.75(0.09)	81.98(0.26)	**89.23(0.11)**
	50	LQR	89.99(0.07)	87.27(0.09)	85.52(0.41)	**89.40(0.08**)
		QRF	90.01(0.06)	77.31(0.10)	79.77(0.31)	**88.95(0.11)**
	60	LQR	89.99(0.07)	86.72(0.10)	84.15(0.56)	**89.49(0.09)**
		QRF	89.96(0.06)	74.82(0.11)	78.14(0.32)	**88.42(0.13)**
AL (SE)	40	LQR	3.30(0.00)	3.24(0.00)	3.21(0.02)	3.75(0.04)
		QRF	7.61(0.01)	6.98(0.01)	7.77(0.11)	7.78(0.04)
	50	LQR	3.30(0.00)	3.25(0.00)	3.32(0.04)	4.09(0.07)
		QRF	7.61(0.01)	6.96(0.01)	7.81(0.12)	7.95(0.05)
	60	LQR	3.30(0.00)	3.26(0.00)	3.32(0.02)	4.51(0.08)
		QRF	7.60(0.01)	6.95(0.01)	7.94(0.12)	8.09(0.07)
				(b) Heteroscedastic		
AC% (SE×100)	40	LQR	90.08(0.07)	83.47(0.10)	81.16(0.57)	**89.13(0.11)**
		QRF	90.11(0.06)	76.60(0.11)	79.71(0.30)	**88.67(0.12)**
	50	LQR	90.08(0.07)	82.35(0.11)	79.75(0.59)	**89.79(0.12)**
		QRF	90.10(0.07)	73.84(0.12)	77.79(0.35)	**88.06(0.12)**
	60	LQR	90.08(0.07)	81.12(0.12)	79.00(0.55)	**89.48(0.13)**
		QRF	90.04(0.06)	71.07(0.13)	75.84(0.39)	**88.51(0.13)**
AL (SE)	40	LQR	2.57(0.00)	2.04(0.00)	2.32(0.09)	3.45(0.12)
		QRF	7.59(0.01)	6.83(0.01)	7.80(0.12)	7.96(0.08)
	50	LQR	2.57(0.00)	1.97(0.00)	2.22(0.07)	3.98(0.14)
		QRF	7.59(0.01)	6.82(0.01)	7.98(0.13)	8.15(0.10)
	60	LQR	2.57(0.00)	1.91(0.01)	2.38(0.10)	4.71(0.16)
		QRF	7.59(0.01)	6.85(0.01)	8.05(0.12)	8.39(0.10)

OMNI: Standard conformal prediction applied to the complete data. Naive: Standard conformal prediction applied to the observed data. MAR-CP: Weighted conformal prediction with MAR weights. MNAR-CP: Proposed weighted conformal prediction with MNAR weights. Values in parentheses are standard errors.

sion for ultrahigh-dimensional data with nonignorable missing responses. *Metrika 83*(5), 545–568.

Izbicki, R., G. Shimizu, and R. B. Stern (2022). Cd-split and hpd-split: Efficient conformal regions in high dimensions. *The Journal of Machine Learning Research 23*(1), 3772–3803.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Qiu, Z., C. Peng, Y. Tang, and H. J. Wang (2026). Review of recent advances in high-dimensional quantile regression. *Wiley Interdisciplinary Reviews: Computational Statistics*. To appear.

Shao, J. and L. Wang (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika 103*(1), 175–187.

Tan, K. M., L. Wang, and W.-X. Zhou (2022). High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(1), 205–233.

Tibshirani, R. J., R. Foygel Barber, E. Candes, and A. Ramdas (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems 32*.

Wang, L., Y. Wu, and R. Li (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association 107*(497), 214–222.

Wang, L., P. Zhao, and J. Shao (2021). Dimension-reduced semiparametric estimation of distribution functions and quantiles with nonignorable nonresponse. *Computational Statistics & Data Analysis 156*, 107142.

Yi, M., Y. Zhang, Y. Tang, and H. J. Wang (2025). Hierarchical conformal prediction for clustered data with missing responses. Manuscript under review.

Zhang, J., Q. Wang, and J. Kang (2020). Feature screening under missing indicator imputation with non-ignorable missing response. *Computational Statistics & Data Analysis 149*, 106975.

KLATASDS-MOE, School of Statistics, East China Normal University

E-mail: menghany@umich.edu

Academy of Statistics and Interdisciplinary Sciences, East China Normal University

E-mail: yyzhang@fem.ecnu.edu.cn

KLATASDS-MOE, School of Statistics, East China Normal University

E-mail: yltang@fem.ecnu.edu.cn * Corresponding author

Department of Statistics, Rice University

E-mail: jw322@rice.edu * Corresponding author