Biometrika (2023), *pp.* 1–18
Printed in Great Britain

Supplementary material for "Semi-supervised distribution learning"

The supplementary material contains all technical proofs, applications including the semi-supervised conformal p-values and local distributional treatment effects, additional numerical studies with more complex settings, and real data analysis. Codes for the proposed methods and numerical studies can be found on GitHub (`https://github.com/mtwen/BASD`).

S1. Proofs

S1.1. Lemmas

LEMMA S1 (A BERNSTEIN-TYPE INEQUALITY FOR BOUNDED PROCESSES). *Let $\mathcal{T}$ be an index set admitting a countable separant $\mathcal{S}$. Let $X_i = (X_{i,s})_{s\in\mathcal{T}}$ $(i = 1, \dots, n)$ be independent (not necessarily identically distributed) real-valued random variables. Assume that $E(X_{i,s}) = 0$, and $|X_{i,s}| \le 1$ for all $s \in \mathcal{T}$ and $i = 1, \dots, n$. Let $Z = \sup_{s\in\mathcal{T}} \sum_{i=1}^n X_{i,s}$ and let the weak variance Σ^2 and the wimpy variance σ^2 be defined as $\Sigma^2 = E(\sup_{s\in\mathcal{T}} \sum_{i=1}^n X_{i,s}^2)$ and $\sigma^2 = \sup_{s\in\mathcal{T}} \sum_{i=1}^n E(X_{i,s}^2)$. Then,*

$$\mathrm{var}(Z) \le \Sigma^2 + \sigma^2 \le 8E(Z) + 2\sigma^2. \tag{S1}$$

For $t \ge 0$,

$$\mathrm{pr}\{Z \ge E(Z) + t\} \le \exp\left(-\frac{t^2}{2\{2(\Sigma^2 + \sigma^2) + t\}}\right).$$

Proof of Lemma S1. See Boucheron et al. (2013), Theorem 11.8 and 12.2. □

LEMMA S2 (LEMMA S1 OF ZHANG & BRADIC (2022)). *Let $\{X_n\}$ and $\{Y_n\}$ be a sequence of random vectors. If for any $\epsilon > 0$, $\mathrm{pr}(\|X_n\| > \epsilon \mid Y_n) = o_p(1)$, then $\mathrm{pr}(\|X_n\| > \epsilon) \to 0$. In particular, this occurs if $E(\|X_n\|^q \mid Y_n) = o_p(1)$ for any $q \ge 1$, by Chebyshev's inequality.*

S1.2. Proofs of Theorem 1

We first introduce some notations. Recall $\gamma_n = m(n+m)^{-1}$ and $Pf = \int f\mathrm{d}P$ for a probability measure P. For a function class $\mathcal{F}$, the uniform norm for a map $z : \mathcal{F} \mapsto \mathbb{R}$ is defined as $\|z\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |z(f)|$, and $J_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)\}^{1/2}\mathrm{d}\epsilon$ is the bracketing integral. Let $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$ and $\mathbb{P}_m = m^{-1}\sum_{i=n+1}^{n+m} \delta_{X_i}$ be the empirical probability measures in labeled and unlabeled data, respectively, where δ_Z is the Dirac probability measure at the point Z. Let $u_k(y|X_i) = \hat{F}_k(y|X_i) - F_0(y|X_i)$ for convenience.

Now, by arranging terms of $\hat{F}_{\text{basd}}(y)$ (3), we have

$$n^{1/2}\{\hat{F}_{\text{basd}}(y) - F(y)\} = n^{1/2}\left\{\hat{F}_{\text{basd},0}(y) - F(y)\right\} + \gamma_n K^{-1/2}\sum_{k=1}^{K}\left[-n_K^{-1/2}\sum_{i\in\mathcal{I}_k}\{u_k(y|X_i) - E_X(u_k(y|X))\} + (n/m)^{1/2}m_K^{-1/2}\sum_{i\in\mathcal{J}_k}\{u_k(y|X_i) - E_X(u_k(y|X))\}\right], \tag{S2}$$

where $\hat{F}_{\text{basd},0}$ is defined as (1) with the non-random function $F_0(y|x)$, that is,

$$\hat{F}_{\text{basd},0}(y) = \frac{1}{n+m}\sum_{i=1}^{n+m} F_0(y|X_i) + \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{1}_y(Y_i) - F_0(y|X_i)\}. \tag{S3}$$

We first show some properties of $\hat{F}_{\text{basd},0}$, including the exponential tail bound (Lemma S3) and uniform weak convergence (Lemma S4).

LEMMA S3. *Suppose that the measurable function class $\mathcal{G} = \{F_0(y|X) : y \in \mathbb{R}\}$ satisfies that, for some $\eta \in (0, 2)$, $\log N_{[]}(\epsilon, \mathcal{G}, L_2(P_X)) \lesssim \epsilon^{-\eta}$ for every $\epsilon > 0$. Then, there exist some constants $C_0, c_0 > 0$ such that*

$$\text{pr}\left[\sup_{y\in\mathbb{R}}\left|n^{1/2}\left\{\hat{F}_{\text{basd},0}(y) - F(y)\right\}\right| \geq \overline{\sigma}(\mathcal{G})\delta\right] \leq C_0\exp(-c_0\delta^2),$$

where $\overline{\sigma}(\mathcal{G})$ is defined in Theorem 1.

Proof of Lemma S3. Let $\mathbb{F}_n(y) = n^{1/2}\{\hat{F}_{\text{basd},0}(y) - F(y)\}$ for convenience. By calculations, we have

$$\mathbb{F}_n(y) = n^{-1/2}\sum_{i=1}^{n}\left(\{\mathbb{1}_y(Y_i) - \gamma_n F_0(y|X_i)\} - [F(y) - \gamma_n E\{F_0(y|X)\}]\right) + n^{-1/2}\sum_{i=n+1}^{n+m}(1-\gamma_n)\left[F_0(y|X_i) - E\{F_0(y|X)\}\right], \tag{S4}$$

which implies that $\mathbb{F}_n(y)$ consists of independent sums.

We first establish the upper bound of $E\{\sup_{y\in\mathbb{R}}|\mathbb{F}_n(y)|\}$. We define the function class $\mathcal{F}_1 = \{\mathbb{1}_y(Y) - \gamma_n F_0(y|X) : y \in \mathbb{R}\}$. Then, we have

$$E\left\{\sup_{y\in\mathbb{R}}|\mathbb{F}_n(y)|\right\} \leq E\left\{\|n^{1/2}(\mathbb{P}_n - P)\|_{\mathcal{F}_1}\right\} + \{\gamma_n(1-\gamma_n)\}^{1/2}E\left\{\|m^{1/2}(\mathbb{P}_m - P)\|_{\mathcal{G}}\right\}. \tag{S5}$$

By the assumption on $\mathcal{G}$, we have $\log N_{[]}(\epsilon, \mathcal{G}, L_2(P_X)) \lesssim \epsilon^{-\eta}$ for every $\epsilon > 0$ and some $\eta \in (0, 2)$, and further $J_{[]}(\delta, \mathcal{G}, L_2(P)) \lesssim \delta^{1-\eta/2}$. It is well known that the bracketing number of $\{\mathbb{1}_y(Y) : y \in \mathbb{R}\}$ satisfies $N_{[]}(\epsilon, \{\mathbb{1}_y(Y) : y \in \mathbb{R}\}, L_2(P)) \lesssim \epsilon^{-2}$. Thus, for $\epsilon \in (0, 1)$,

$$\log N_{[]}(\epsilon, \mathcal{F}_1, L_2(P)) \lesssim \epsilon^{-\eta} + 2\log(1/\epsilon), \tag{S6}$$

and thus $J_{[]}(\delta, \mathcal{F}_1, L_2(P)) \lesssim \delta^{1-\eta/2}$. By Lemma 3.4.2 of van der Vaart & Wellner (1996), we have

$$E\{\|n^{1/2}(\mathbb{P}_n - P)\|_{\mathcal{F}_1}\} \lesssim J_{[]}(1, \mathcal{F}_1, L_2(P)) \Big(1 + n^{-1/2} J_{[]}(1, \mathcal{F}_1, L_2(P))\Big) \lesssim 1,$$
$$E\{\|m^{1/2}(\mathbb{P}_m - P)\|_{\mathcal{G}}\} \lesssim J_{[]}(1, \mathcal{G}, L_2(P)) \Big(1 + m^{-1/2} J_{[]}(1, \mathcal{G}, L_2(P))\Big) \lesssim 1.$$

Together with (S5), there exists a constant c_1 such that $E\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)|\} \le c_1$.

We now prove the conclusion of this lemma. We have

$$|\mathbb{F}_n(y)| \le 2n^{1/2} + mn^{1/2}/(n+m) \le 3n^{1/2},$$

where the first inequality holds by $0 \le F_0(y|X) \le 1$ and $|\mathbb{1}_y(Y_i) - \gamma_n F_0(y|X_i)| \le 1$. Then, for $\delta > 3n^{1/2}/\overline{\sigma}(\mathcal{G})$, $\text{pr}\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)| \ge \overline{\sigma}(\mathcal{G})\delta\} = 0$. For $2c_1/\overline{\sigma}(\mathcal{G}) \le \delta \le 3n^{1/2}/\overline{\sigma}(\mathcal{G})$, we have $\overline{\sigma}(\mathcal{G})\delta - E\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)|\} > 0$ due to $E\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)|\} \le c_1$. Thus, letting $t = n^{1/2}\overline{\sigma}(\mathcal{G})\delta/2 - n^{1/2}E\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)|\}/2$, with Lemma S1,

$$\begin{aligned}\text{pr}\left\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)| \ge \overline{\sigma}(\mathcal{G})\delta\right\} &\le 2\text{pr}\left[\sup_{y\in\mathbb{R}}\left\{n^{1/2}\mathbb{F}_n(y)/2\right\} - E\left[\sup_{y\in\mathbb{R}}\left\{n^{1/2}\mathbb{F}_n(y)/2\right\}\right] \ge t\right] \\ &\le 2\exp\left\{-\frac{t^2}{4(\Sigma^2+\sigma^2)+2t}\right\},\end{aligned}$$

where $\sigma^2 = n\overline{\sigma}^2(\mathcal{G})/4$ is the wimpy variance, and Σ^2 is the weak variance satisfying

$$\Sigma^2 \le 4n^{1/2}E\left\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)|\right\} + n\overline{\sigma}^2(\mathcal{G})/4 \le 4c_1 n^{1/2} + n\overline{\sigma}^2(\mathcal{G})/4$$

by (S1). So,

$$\text{pr}\left\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)| \ge \overline{\sigma}(\mathcal{G})\delta\right\} \le 2\exp\left\{-\frac{t^2}{2n\overline{\sigma}^2(\mathcal{G}) + 16c_1 n^{1/2} + 2t}\right\} \le C_2\exp\left(-c_2\delta^2\right),$$

where the last inequality holds by $n^{1/2}\overline{\sigma}(\mathcal{G})\delta/4 \le t \le 3n/2$ for some constants $C_2, c_2 > 0$. For $0 < \delta < 2c_1/\overline{\sigma}(\mathcal{G})$, $\text{pr}\{\sup_{y\in\mathbb{R}} |\mathbb{F}_n(y)| \ge \overline{\sigma}(\mathcal{G})\delta\} \le C_3\exp(-c_3\delta^2)$ holds for some constants $C_3, c_3 > 0$. Thus, the conclusion follows for some universal constants $C_0, c_0 > 0$. □

LEMMA S4. *Suppose that the measurable function class $\mathcal{G} = \{F_0(y|X) : y \in \mathbb{R}\}$ satisfies that, for some $\eta \in (0,2)$, $\log N_{[]}(\epsilon, \mathcal{G}, L_2(P_X)) \lesssim \epsilon^{-\eta}$ for every $\epsilon > 0$. Assuming $\gamma_n \to \gamma \in [0,1]$, we have*

$$n^{1/2}\{\hat{F}_{\text{basd},0}(y) - F(y)\} \rightsquigarrow \mathbb{F}(y; \mathcal{G}),$$

uniformly for $y \in \mathbb{R}$ as $n, m \to \infty$, where $\mathbb{F}(y;\mathcal{G}) = (1-\gamma)^{1/2}\mathbb{B}_1 \circ \mathbb{1}_y(Y) + \gamma^{1/2}\mathbb{B}_2 \circ \{\mathbb{1}_y(Y) - F_0(y|X)\}$, and $\mathbb{B}_1, \mathbb{B}_2$ are two independent Brownian bridges.

Proof of Lemma S4. Recall from (S4) that we decompose $\mathbb{F}_n$ into two independent parts:

$$\mathbb{F}_n(y) = n^{1/2}(\mathbb{P}_n - P)\{\mathbb{1}_y(Y) - \gamma_n F_0(y|X)\} + \{\gamma_n(1-\gamma_n)\}^{1/2}\{m^{1/2}(\mathbb{P}_m - P)F_0(y|X)\}.$$

We define $\mathcal{F}_1 = \{\mathbb{1}_y(Y) - \gamma_n F_0(y|X) : y \in \mathbb{R}\}$ as in Lemma S3. By the assumption on $\mathcal{G}$ and (S6), both $\mathcal{F}_1$ and $\mathcal{G}$ are Donsker. Then, by the independence of $\mathbb{P}_n$ and $\mathbb{P}_m$, we have

$$\begin{aligned}&\Big(n^{1/2}(\mathbb{P}_n - P)\{\mathbb{1}_y(Y) - \gamma_n F_0(y|X)\}, m^{1/2}(\mathbb{P}_m - P)F_0(y|X)\Big)^{\mathrm{T}} \\ &\qquad\qquad \rightsquigarrow \left(\mathbb{B}_1 \circ \{\mathbb{1}_y(Y) - \gamma F_0(y|X)\}, \mathbb{B}_2 \circ \{F_0(y|X)\}\right)^{\mathrm{T}},\end{aligned}$$

uniformly for $y \in \mathbb{R}$. By the continuous mapping theorem, we have uniformly for $y \in \mathbb{R}$,

$$\begin{aligned}\mathbb{F}_n(y) &\rightsquigarrow \mathbb{B}_1 \circ \{\mathbb{1}_y(Y) - \gamma F_0(y|X)\} + \{\gamma(1-\gamma)\}^{1/2}\mathbb{B}_2 \circ \{F_0(y|X)\} \\ &\sim_d (1-\gamma)^{1/2}\mathbb{B}_1 \circ \mathbb{1}_y(Y) + \gamma^{1/2}\mathbb{B}_2 \circ \{\mathbb{1}_y(Y) - F_0(y|X)\}.\end{aligned}$$

The conclusion follows. □

Now, we turn back to the proof of Theorem 1. Recall the decomposition of $\hat{F}_{\text{basd}}$ in (S2). For the uniform exponential tail bound, the rest is to derive the exponential tail bound for $n_K^{1/2}(\mathbb{P}_{n,k} - P_X)u_k(y|X)$ and $m_K^{1/2}(\mathbb{P}_{m,k} - P_X)u_k(y|X)$ conditional on $\mathcal{L}_{-k}$, where $\mathbb{P}_{n,k} = {n_K}^{-1}\sum_{i\in\mathcal{I}_k}\delta_{X_i}$ and $\mathbb{P}_{m,k} = m_K^{-1}\sum_{i\in\mathcal{J}_k}\delta_{X_i}$ are the empirical probability measures in the kth folds of labeled and unlabeled data, respectively. We define the function class $\mathcal{F}_{2,k} = \{u_k(y|X) : y \in \mathbb{R}\}$. By Assumption 2 and the definition of bracketing numbers, we have $\log N_{[]}(\epsilon, \mathcal{F}_{2,k}, L_2(P_X)) \lesssim \epsilon^{-\eta}$ for every $\epsilon > 0$ and some $\eta \in (0,2)$. Then, by Theorem 2.14.2 of van der Vaart & Wellner (1996), denoting $\varrho_{n,k}^2 = E_X\{U_k(X)\}^2$, we have

$$\begin{aligned}E_X\left[\|n_K^{1/2}(\mathbb{P}_{n,k} - P_X)\|_{\mathcal{F}_{2,k}}\right] &\lesssim \varrho_{n,k}\int_0^1 \left\{1 + \log N_{[]}(\epsilon\varrho_{n,k}, \mathcal{F}_{2,k}, L_2(P_X))\right\}\mathrm{d}\epsilon \\ &\lesssim \varrho_{n,k}^{1-\eta/2}\int_0^1 \epsilon^{-\eta/2}\mathrm{d}\epsilon \le (1-\eta/2)^{-1}\varrho_n^{1-\eta/2}.\end{aligned} \tag{S7}$$

We have $E_X[\|m_K^{1/2}(\mathbb{P}_{m,k} - P_X)\|_{\mathcal{F}_{2,k}}] \lesssim (1-\eta/2)^{-1}\varrho_n^{1-\eta/2}$ similarly. Similar to the proof of Lemma S3, with Lemma S1, there exist constants $C_k, c_k > 0$ such that for $\delta \in (0, 2n_k^{1/2}\varrho_n^{(2-\eta)/4}]$

$$\mathrm{pr}\left\{\sup_{y\in\mathbb{R}}\left|n_K^{1/2}(\mathbb{P}_{n,k} - P_X)u_k(y|X)\right| \ge \varrho_n^{(2-\eta)/4}\delta \mid \mathcal{L}_{-k}\right\} \le C_k \exp\left(-c_k\delta^2\right),$$

and

$$\mathrm{pr}\left\{\sup_{y\in\mathbb{R}}\left|m_K^{1/2}(\mathbb{P}_{m,k} - P_X)u_k(y|X)\right| \ge \varrho_n^{(2-\eta)/4}\delta \mid \mathcal{L}_{-k}\right\} \le C_k \exp\left(-c_k\delta^2\right).$$

Then,

$$\begin{aligned}&\mathrm{pr}\left[\sup_{y\in\mathbb{R}}\left|-n_K^{1/2}(\mathbb{P}_{n,k} - P_X)u_k(y|X) + (n/m)^{1/2}m_K^{1/2}(\mathbb{P}_{m,k} - P_X)u_k(y|X)\right| \ge \left\{1 + (n/m)^{1/2}\right\}\varrho_n^{(2-\eta)/4}\delta\right] \\ &\le E\left[\mathrm{pr}\left\{\sup_{y\in\mathbb{R}}\left|n_K^{1/2}(\mathbb{P}_{n,k} - P_X)u_k(y|X)\right| \ge \varrho_n^{(2-\eta)/4}\delta \mid \mathcal{L}_{-k}\right\}\right] \\ &\quad + E\left[\mathrm{pr}\left\{\sup_{y\in\mathbb{R}}\left|m_K^{1/2}(\mathbb{P}_{m,k} - P_X)u_k(y|X)\right| \ge \varrho_n^{(2-\eta)/4}\delta \mid \mathcal{L}_{-k}\right\}\right] \\ &\le 4C_k \exp\left(-c_k\delta^2\right).\end{aligned} \tag{S8}$$

Thus, together with (S2) and Lemma S3, we have

$$\mathrm{pr}\left[\sup_{y\in\mathbb{R}}\left|n^{1/2}\left\{\hat{F}_{\text{basd}}(y) - F(y)\right\}\right| \ge \overline{\sigma}(\mathcal{G})\delta + K^{1/2}\gamma_n\left\{1 + (n/m)^{1/2}\right\}\varrho_n^{(2-\eta)/4}\delta\right] \le C\exp(-c\delta^2),$$

for some constants $C, c > 0$. The conclusion (4) in the theorem holds.

Next, we show the uniform weak convergence of $\hat{F}_{\mathrm{basd}}(y)$. By Lemma S4, we have

$$n^{1/2}\left\{\hat{F}_{\mathrm{basd},0}(y)-F(y)\right\} \rightsquigarrow (1-\gamma)^{1/2}\mathbb{B}_1\circ \mathbb{1}_y(Y)+\gamma^{1/2}\mathbb{B}_2\circ\{\mathbb{1}_y(Y)-F_0(y|X)\},$$

uniformly for $y\in\mathbb{R}$, where $\mathbb{B}_1$ and $\mathbb{B}_2$ are two independent Brownian bridges. With (S2), it suffices to show

$$\sup_y \frac{1}{n_K}\sum_{i\in\mathcal{I}_k}\{u_k(y|X_i)-E_X(u_k(y|X))\}=o_p(n_K^{-1/2}),$$
$$\sup_y \frac{1}{m_K}\sum_{i\in\mathcal{I}'_k}\{u_k(y|X_i)-E_X(u_k(y|X))\}=o_p(m_K^{-1/2}).$$

We have shown in (S7),

$$E_X\left(\sup_y\left|n_K^{-1/2}\sum_{i\in\mathcal{I}_k}[u_k(y|X_i)-E_X\{u_k(y|X)\}]\right|\right)\lesssim \varrho_n^{1-\eta/2}=o_p(1),$$

Then, by Lemma S2, the conclusion (5) in the theorem follows.

S1.3. Proof of other theoretical results

Proof of Proposition 1. From the proof of Theorem 1, we have that the asymptotic covariance of $\hat{F}_{\mathrm{basd}}$ is the same as that of $\hat{F}_{\mathrm{basd},0}$ defined in (S3). Thus, we consider the uniform weak convergence of $\hat{F}_{\mathrm{basd},0}$ in (S3) with $\mathcal{G}$ consisting of $F_0(y|X)=\mathrm{pr}\{Y\leq y\mid h(X)\}$ to give the asymptotic covariance. The function class $\mathcal{G}=\{F_0(y|X):y\in\mathbb{R}\}$ is a monotone process due to the form $\mathrm{pr}\{Y\leq y\mid h(X)\}$. Let $-\infty=\xi_0<\xi_1<\cdots<\xi_M=+\infty$ such that $E\{F_0(\xi_i-|X)\}-E\{F_0(\xi_{i-1}|X)\}\leq\epsilon^2$, where $M=1/\epsilon^2$ and $f(x_0-)=\lim_{x\uparrow x_0}f(x)$ is the left-sided limit as x approaches x_0 for a function f. Then, $[F_0(\xi_{i-1}|X),F_0(\xi_i-|X)]$ $(i=1,\ldots,M)$ constitute the ϵ-brackets of $\mathcal{G}$ in $L_2(P)$. Thus, $N_{[]}(\epsilon,\mathcal{G},L_2(P))\leq\epsilon^{-2}$, which satisfies the assumption of Lemma S4. So, we have $n^{1/2}\{\hat{F}_{\mathrm{basd},0}-F(y)\}\rightsquigarrow\mathbb{F}(y;\mathcal{G})$ uniformly over $y\in\mathbb{R}$, by Lemma S4.

The covariance of $\mathbb{F}(\cdot;\mathcal{G})$ is

$$\mathrm{cov}\begin{pmatrix}\mathbb{F}(s;\mathcal{G})\\ \mathbb{F}(t;\mathcal{G})\end{pmatrix}=(1-\gamma)\mathrm{cov}\begin{pmatrix}\mathbb{1}_s(Y)\\ \mathbb{1}_t(Y)\end{pmatrix}+\gamma\mathrm{cov}\begin{pmatrix}\mathbb{1}_s(Y)-F_0(s|X)\\ \mathbb{1}_t(Y)-F_0(t|X)\end{pmatrix}, \tag{S9}$$

for any $s,t\in\mathbb{R}$. By law of total expectation,

$$\begin{aligned}&\mathrm{cov}\left\{\begin{pmatrix}\mathbb{1}_s(Y)\\ \mathbb{1}_t(Y)\end{pmatrix},\begin{pmatrix}F_0(s|X)\\ F_0(t|X)\end{pmatrix}\right\}\\ &\quad=E\left\{\begin{pmatrix}\mathbb{1}_s(Y)-F(s)\\ \mathbb{1}_t(Y)-F(t)\end{pmatrix}\begin{pmatrix}F_0(s|X)-F(s)\\ F_0(t|X)-F(t)\end{pmatrix}^{\mathrm{T}}\right\}=\mathrm{cov}\begin{pmatrix}F_0(s|X)\\ F_0(t|X)\end{pmatrix}.\end{aligned}\tag{S10}$$

Combining (S9) and (S10), we have

$$\begin{aligned}\mathrm{cov}\begin{pmatrix}\mathbb{F}(s;\mathcal{G})\\ \mathbb{F}(t;\mathcal{G})\end{pmatrix}&=(1-\gamma)\mathrm{cov}\begin{pmatrix}\mathbb{1}_s(Y)\\ \mathbb{1}_t(Y)\end{pmatrix}+\gamma\left\{\mathrm{cov}\begin{pmatrix}\mathbb{1}_s(Y)\\ \mathbb{1}_t(Y)\end{pmatrix}-\mathrm{cov}\begin{pmatrix}F_0(s|X)\\ F_0(t|X)\end{pmatrix}\right\}\\ &=\mathrm{cov}\begin{pmatrix}\mathbb{1}_s(Y)\\ \mathbb{1}_t(Y)\end{pmatrix}-\gamma\mathrm{cov}\begin{pmatrix}F_0(s|X)\\ F_0(t|X)\end{pmatrix}\leqslant\mathrm{cov}\begin{pmatrix}\mathbb{1}_s(Y)\\ \mathbb{1}_t(Y)\end{pmatrix},\end{aligned}$$

where $M_1 \leqslant M_2$ for squared matrices M_1, M_2 means that $M_2 - M_1$ is a nonnegative definite matrix, and the last inequality holds due to the nonnegative definiteness of $\mathrm{cov}\{(F_0(s|X), F_0(t|X))^{\mathrm{T}}\}$. □

Proof of Corollary 1. The result is a direct application of the functional delta theorem (van der Vaart, 1998, Chapter 20). □

S2. Applications

S2.1. Conformal p-value

Conformal inference (Vovk et al., 2005) provides a powerful and flexible work to achieve distribution-free uncertainty quantification of predictors. Recently, there have been some works to build conformal p-values for testing certain hypotheses (Bates et al., 2023; Zhang et al., 2022). Jin & Candès (2023) specially investigated the prediction-oriented selection problem aiming to select samples whose unobserved outcomes exceed some specified values and proposed to construct conformal p-values of the predicted response values to implement the selection. For a test point X_0 and its corresponding unobserved Y_0, the corresponding hypothesis of interest is:

$$\mathbb{H}_0 : Y_0 \leq b_0 \quad \text{versus} \quad \mathbb{H}_1 : Y_0 > b_0, \tag{S11}$$

where $b_0 \in \mathbb{R}$ is a given constant.

A standard conformal test statistic and its p-value for (S11) can be constructed as follows. Denote $\hat{\mu}(x)$ as a pre-specified estimator of $E(Y \mid X = x)$ and $\hat{\mu}(X_0)$ as the predicted value of Y_0. By choosing a monotone function $V : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$, such as $V(x, y) = y - \hat{\mu}(x)$, one could obtain scores $\{V_i = V(X_i, Y_i)\}_{i \in \mathcal{L}}$ and $\hat{V}_0 = V(X_0, b_0)$. Then, the conformal p-value for this test (S11) is then computed as $\hat{p}_{\mathrm{CP}} = \{1 + \sum_{i=1}^{n} \mathbb{1}_{\hat{V}_0}(V_i)\}/(n+1)$. However, the small size n of labeled data often results in large variation in $\hat{p}_{\mathrm{CP}}$.

We observe that the conformal p-value $\hat{p}_{\mathrm{CP}}$ is actually derived from the empirical cumulative distribution function of V, that is, $\hat{p}_{\mathrm{CP}} = \{1 + n\hat{F}_{\mathrm{ecdf},V}(\hat{V}_0)\}/(n+1)$, where $\hat{F}_{\mathrm{ecdf},V}(v) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_v(V_i)$. In the heterogeneous cases (Romano et al., 2019), V is still related to covariate X, allowing us to apply the proposed procedure to observations $\{(V_i, X_i)\}_{i=1}^{n}$ and $\{X_i\}_{i=n+1}^{n+m}$ to reduce the variance in $\hat{p}_{\mathrm{CP}}$. Thus, the semi-supervised conformal p-value can be given by $\hat{p}_{\mathrm{basd}} = 1 + n\hat{F}_{\mathrm{basd},V}(\hat{V}_0)/(n+1)$, where $\hat{F}_{\mathrm{basd},V}(\cdot)$ is the estimated distribution function from Algorithm 1 of Section S3.1.

Corollary S1. *Let $F_V(v) = \mathrm{pr}(V \leq v \mid \hat{\mu})$ and $p_0 = F_V(\hat{V}_0)$ be the cumulative distribution function of V and the corresponding p-value at X_0. Provided that Theorem 1 holds for some $F_{0,V}(v|x)$, we have conditional on $\hat{\mu}$ and X_0,*

$$n^{1/2}\left(\hat{p}_{\mathrm{basd}} - p_0\right) \rightsquigarrow \mathcal{N}\left(0, (1-\gamma)\mathrm{var}\left\{\mathbb{1}_{\hat{V}_0}(V)\right\} + \gamma\mathrm{var}\left\{\mathbb{1}_{\hat{V}_0}(V) - F_{0,V}(\hat{V}_0|X)\right\}\right).$$

Proof of Corollary S1. This result directly follows from Theorem 1. □

In contrast, the standard conformal p-value satisfies $n^{1/2}\left(\hat{p}_{\mathrm{CP}} - p_0\right) \rightsquigarrow \mathcal{N}(0, \mathrm{var}\{\mathbb{1}_{\hat{V}_0}(V)\})$ given $\hat{\mu}$ and X_0. Similar arguments can exhibit that under some conditions, $\hat{p}_{\mathrm{basd}}$ achieves a smaller asymptotic variance compared to $\hat{p}_{\mathrm{CP}}$. Numerical results in Section S4.4 further illustrate the superiority of the semi-supervised conformal p-value.

S2.2. Local distributional treatment effects

In the past decades, casual inference has drawn significant attention and aims to reliably analyze the counterfactual quantities of an outcome variable with or without a treatment (Imbens & Rubin, 2015). Let Y_{0i} and Y_{1i} be the potential outcome for individual i without and with treatment, respectively. The observed outcome is the realized value $Y_i = (1 - D_i)Y_{0i} + D_iY_{1i}$, where $D_i \in \{0, 1\}$ is the treatment status. To address non-randomized treatment, instrumental variables are commonly employed (Angrist et al., 1996). With a binary instrumental variable Z_i for the ith individual, the analyst observes the realized treatment $D_i = D_{1i}Z_i + D_{0i}(1 - Z_i)$ instead of the potential treatment indicators (D_{0i} and D_{1i}) in practice. Together with additional control variables X_i, the observed data are $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^n$. The distributional treatment effects could be useful for policymakers who wish to take into account not only differences in average outcomes (Imbens & Rubin, 1997). The focus of distributional treatment effects is to test the difference between $F_{\mathrm{C}}^{(1)} = \mathrm{pr}(Y_{1i} \le y \mid D_{0i} = 0, D_{1i} = 1)$ and $F_{\mathrm{C}}^{(0)} = \mathrm{pr}(Y_{0i} \le y \mid D_{0i} = 0, D_{1i} = 1)$ (Abadie, 2002; Chernozhukov et al., 2013). For example, consider the null hypothesis $\mathbb{H}_0 : F_{\mathrm{C}}^{(1)} = F_{\mathrm{C}}^{(0)}$.

By introducing the instrumental variable Z_i and imposing certain identifiability assumptions, the local distributional treatment effect can be written as (Abadie, 2002),

$$F_{\mathrm{C}}^{(1)}(y) - F_{\mathrm{C}}^{(0)}(y) = \frac{\mathrm{pr}(Y_i \le y \mid Z_i = 1) - \mathrm{Pr}(Y_i \le y \mid Z_i = 0)}{\mathrm{pr}(D_i = 1 \mid Z_i = 1) - \mathrm{pr}(D_i = 1 \mid Z_i = 0)} = \kappa\{F^{(1)}(y) - F^{(0)}(y)\}, \tag{S12}$$

where $\kappa = 1/\{\mathrm{pr}(D_i = 1 \mid Z_i = 1) - \mathrm{pr}(D_i = 1 \mid Z_i = 0)\} < \infty$ is a constant. Testing the equality of $F_{\mathrm{C}}^{(1)}$ and $F_{\mathrm{C}}^{(0)}$ is therefore equivalent to testing $\mathbb{H}_0 : F^{(1)} = F^{(0)}$. Let $\mathcal{I}^{(1)} = \{i : Z_i = 1\}$ and $\mathcal{I}^{(0)} = \{i : Z_i = 0\}$ with sample sizes of n_1 and n_0 respectively. The conventional approach is to construct the Kolmogorov-Smirnov statistic based on $\hat{F}_{\mathrm{ecdf}}^{(1)}(y) - \hat{F}_{\mathrm{ecdf}}^{(0)}(y) := {n_1}^{-1} \sum_{i \in \mathcal{I}^{(1)}} \mathbb{1}_y(Y_i) - {n_0}^{-1} \sum_{i \in \mathcal{I}^{(0)}} \mathbb{1}_y(Y_i)$.

When control covariates X_i independent of Z_i are available, the proposed procedure could help improve the estimation efficiency of $F^{(1)} - F^{(0)}$. We can treat $\{(X_i, Y_i)\}_{i \in \mathcal{I}^{(1)}}$ and $\{X_i\}_{i \in \mathcal{I}^{(0)}}$ as labeled and unlabeled data, and use the proposed method to obtain semi-supervised estimator $\hat{F}_{\mathrm{basd}}^{(1)}(y)$. Concretely, we do partitions $\mathcal{I}^{(1)} = \mathcal{I}_1^{(1)} \cup \cdots \cup \mathcal{I}_K^{(1)}$ and $\mathcal{I}^{(0)} = \mathcal{I}_1^{(0)} \cup \cdots \cup \mathcal{I}_K^{(0)}$, and obtain the conditional cumulative distribution function estimator $\{\hat{F}_k^{(1)}(y|x) : y \in \mathbb{R}\}$ based on $\{(X_i, Y_i)\}_{i \in \mathcal{I}_{-k}^{(1)}}$ for each $k \in \{1, \ldots, K\}$. Then,

$$\hat{F}_{\mathrm{basd}}^{(1)}(y) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k^{(1)} \cup \mathcal{I}_k^{(0)}} \hat{F}_k^{(1)}(y|X_i) + \frac{1}{n_1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k^{(1)}} \left\{ \mathbb{1}_y(Y_i) - \hat{F}_k^{(1)}(y|X_i) \right\}.$$

With similar notations $\{\hat{F}_k^{(0)}(y|x) : y \in \mathbb{R}\}$, we have

$$\hat{F}_{\mathrm{basd}}^{(0)}(y) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k^{(1)} \cup \mathcal{I}_k^{(0)}} \hat{F}_k^{(0)}(y|X_i) + \frac{1}{n_0} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k^{(0)}} \left\{ \mathbb{1}_y(Y_i) - \hat{F}_k^{(0)}(y|X_i) \right\}.$$

The Kolmogorov-Smirnov statistic is accordingly constructed with

$$\hat{F}_{\mathrm{basd}}^{(1)}(y) - \hat{F}_{\mathrm{basd}}^{(0)}(y) = \frac{1}{n_1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k^{(1)}} \left\{ \mathbb{1}_y(Y_i) - \hat{F}_k(y|X_i) \right\} - \frac{1}{n_0} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k^{(0)}} \left\{ \mathbb{1}_y(Y_i) - \hat{F}_k(y|X_i) \right\}, \tag{S13}$$

where $\hat{F}_k(y|x) = n_0\hat{F}_k^{(1)}(y|x)/n + n_1\hat{F}_k^{(0)}(y|x)/n$. By similar assumptions to Assumption 1–2 on $\hat{F}_k^{(1)}(y|x)$ and $\hat{F}_k^{(0)}(y|x)$ for some $F_0^{(1)}(y|x)$ and $F_0^{(0)}(y|x)$, respectively, we have

$$\hat{F}^{(1)}_{\text{basd}}(y) - \hat{F}^{(0)}_{\text{basd}}(y) = \frac{1}{n_1}\sum_{i\in\mathcal{I}^{(1)}}\{\mathbb{1}_y(Y_i) - F_0(y|X_i)\} - \frac{1}{n_0}\sum_{i\in\mathcal{I}^{(0)}}\{\mathbb{1}_y(Y_i) - F_0(y|X_i)\} + o_p(n^{-1/2}), \tag{S14}$$

where $F_0(y|x) = n_0F_0^{(1)}(y|x)/n + n_1F_0^{(0)}(y|x)/n$.

Next, we introduce the theory of $\hat{F}^{(1)}_{\text{basd}}(y) - \hat{F}^{(0)}_{\text{basd}}(y)$ based on (S14).

Assumption S1. (i) Independence of the instrumental variable: $(Y_{0i}, Y_{1i}, D_{0i}, D_{1i})$ is independent of Z_i. (ii) $0 < \text{pr}(Z_i = 1) = \lambda < 1$ and $\text{pr}(D_{1i} = 1) > \text{pr}(D_{0i} = 1)$. (iii) Monotonicity: $\text{pr}(D_{1i} \geq D_{0i}) = 1$. (iv) Z_i is also independent of X_i.

Assumption S1(i)–(iii) are widely-used identifying assumptions (Abadie, 2002). Assumption S1(iv) is the requirement for the semi-supervised setting and is valid in many empirical studies such as those in Angrist (1990) and Angrist & Krueger (1991), where a purely random quantity, like draft lottery or quarter of birth, serves as the instrumental variable.

PROPOSITION S1. *Suppose that* (S14) *holds, and* $\mathcal{G}^{(j)} = \{F_0^{(j)}(y|X) : y \in \mathbb{R}\}$ *satisfies that* $\log N_{[]}(\epsilon, \mathcal{G}^{(j)}, L_2(P_X)) \lesssim \epsilon^{-\eta}$ *for every* $\epsilon > 0$*, some* $\eta \in (0, 2)$*, and each* $j \in \{0, 1\}$*. Then, provided with Assumption S1, we have*

$$(n_1n_0/n)^{1/2}\left\{\hat{F}^{(1)}_{\text{basd}}(y) - \hat{F}^{(0)}_{\text{basd}}(y)\right\} \rightsquigarrow \mathbb{B}\circ[\mathbb{1}_y(Y) - \{\lambda F_0^{(0)}(y|X) + (1-\lambda)F_0^{(1)}(y|X)\}]$$

uniformly over $y \in \mathbb{R}$ *under* $\mathbb{H}_0 : F^{(1)} = F^{(0)}$*, where* $\mathbb{B}$ *is a Brownian bridge.*

Proof of Proposition S1. We prove the conclusions following Chapter 3.7 of van der Vaart & Wellner (1996). By the assumptions on $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(0)}$, it is easy to show that the function class $\mathcal{F} = \{\mathbb{1}_y(Y) - F_0(y|X) : y \in \mathbb{R}\}$ with $F_0(y|X) = \{n_1F_0^{(0)}(y|X) + n_0F_0^{(1)}(y|X)\}/n$ is Donsker. Define the empirical measures

$$\mathbb{P}_{1,n_1} = \frac{1}{n_1}\sum_{i\in\mathcal{I}^{(1)}}\delta_{(Y_i,X_i)}, \quad \mathbb{P}_{0,n_0} = \frac{1}{n_0}\sum_{i\in\mathcal{I}^{(0)}}\delta_{(Y_i,X_i)}, \tag{S15}$$

where $\delta_{(Y,X)}$ is the Dirac probability measure at the point (X, Y). By Theorem 3.5.1 in van der Vaart & Wellner (1996), we have

$$n_1^{1/2}(\mathbb{P}_{1,n_1} - P_1) \rightsquigarrow \mathbb{B}_{1,P_1}, \quad n_0^{1/2}(\mathbb{P}_{0,n_0} - P_0) \rightsquigarrow \mathbb{B}_{2,P_0}$$

in $\ell^\infty(\mathcal{F})$, where $\ell^\infty(\mathcal{F})$ is the set of uniformly bounded real functions on $\mathcal{F}$ and $\mathbb{B}_{1,P}, \mathbb{B}_{2,P}$ are two independent P-Brownian bridges. We have

$$\begin{aligned}&(n_1n_0/n)^{1/2}(\mathbb{P}_{1,n_1} - \mathbb{P}_{0,n_0})\\ &\quad = (n_0/n)^{1/2}\left\{n_1^{1/2}(\mathbb{P}_{1,n_1} - P_1)\right\} - (n_1/n)^{1/2}\left\{n_0^{1/2}(\mathbb{P}_{0,n_0} - P_0)\right\} + (n_0n_1/n)^{1/2}(P_1 - P_0).\end{aligned}$$

Under $P_1 = P_0 = P$, $(n_1n_0/n)^{1/2}(\mathbb{P}_{1,n_1} - \mathbb{P}_{0,n_0})$ converges weakly to $(1-\gamma)^{1/2}\mathbb{B}_{1,P} - \gamma^{1/2}\mathbb{B}_{2,P}$, which is the same in distribution as $\mathbb{B}_{1,P}$. The conclusion follows. □

Abadie (2002) showed that $(n_1n_0/n)^{1/2}\{\hat{F}^{(1)}_{\text{ecdf}}(y) - \hat{F}^{(0)}_{\text{ecdf}}(y)\} \rightsquigarrow \mathbb{B}\circ\mathbb{1}_y(Y)$. It can be seen that under similar conditions to Proposition 1, $\hat{F}^{(1)}_{\text{basd}} - \hat{F}^{(0)}_{\text{basd}}$ is asymptotically more efficient than $\hat{F}^{(1)}_{\text{ecdf}}(y) - \hat{F}^{(0)}_{\text{ecdf}}$.

Next, we introduce some tests on the counterfactual distributions $F_{\text{C}}^{(1)}$ and $F_{\text{C}}^{(0)}$. We have shown in (S12) that $F_{\text{C}}^{(1)} - F_{\text{C}}^{(0)} = \kappa\{F^{(1)} - F^{(0)}\}$, where κ is a constant and $F^{(j)}(y) = \text{pr}\{Y \leq y \mid Z = j\}$. Through the Kolmogorov-Smirnov statistic and the modified Kolmogorov-Smirnov statistic (McFadden, 1989), we can test the equality and the first-order stochastic dominance by using $\hat{F}^{(1)}_{\text{basd}} - \hat{F}^{(0)}_{\text{basd}}$. Concretely, we test the equality (i.e., $\mathbb{H}_0^{\text{eq}} : F_{\text{C}}^{(1)} = F_{\text{C}}^{(0)}$) by

$$T^{\text{eq}}_{\text{basd}} = (n_1 n_0/n)^{1/2} \sup_{y\in\mathbb{R}} \left| \hat{F}^{(0)}_{\text{basd}}(y) - \hat{F}^{(1)}_{\text{basd}}(y) \right|,$$

and the first-order stochastic dominance (i.e., $\mathbb{H}_0^{\text{fsd}} : F_{\text{C}}^{(1)}(y) \geq F_{\text{C}}^{(0)}(y)$ for all $y \in \mathbb{R}$) by

$$T^{\text{fsd}}_{\text{basd}} = (n_1 n_0/n)^{1/2} \sup_{y\in\mathbb{R}} \left\{ \hat{F}^{(0)}_{\text{basd}}(y) - \hat{F}^{(1)}_{\text{basd}}(y) \right\}.$$

Let P_1, P_0 be the probability distribution of Y conditional on $Z = 1$ and $Z = 0$, respectively. Let $\mathcal{F} = \{\mathbb{1}_y(Y) - F_0(y|X) : y \in \mathbb{R}\}$ and $D_n = (n_1 n_0/n)^{1/2}(\mathbb{P}_{1,n_1} - \mathbb{P}_{0,n_0})$, where $\mathbb{P}_{1,n_1}$ and $\mathbb{P}_{0,n_0}$ are empirical measures defined in (S15). From Proposition S1, we can see that under $P_1 = P_0 = P$,

$$D_n = (n_0/n)^{1/2} \left\{ n_1^{1/2}(\mathbb{P}_{1,n_1} - P_1) \right\} - (n_1/n)^{1/2} \left\{ n_0^{1/2}(\mathbb{P}_{0,n_0} - P_0) \right\} \rightsquigarrow \mathbb{B}_P, \text{ in } \ell^\infty(\mathcal{F}),$$

where $\mathbb{B}_P$ is a P-Brownian bridge. For $z \in \ell^\infty(\mathcal{F})$, we can define two operators $T^{\text{eq}}(z) = \sup_{f\in\mathcal{F}} |z(f)|$ and $T^{\text{fsd}}(z) = \sup_{f\in\mathcal{F}} z(f)$, which are continuous on z. Thus, by continuous mapping theorem, $T^{\text{eq}}(D_n) \rightsquigarrow T^{\text{eq}}(\mathbb{B}_P)$ and $T^{\text{fsd}}(D_n) \rightsquigarrow T^{\text{fsd}}(\mathbb{B}_P)$, respectively. We use the least favorable case ($P_1 = P_0 = P$) to derive the asymptotic null distribution. It can be seen the statistics go into infinity under any fixed alternative.

The asymptotic null distribution depends on the underlying distribution P. So, we give a sampling strategy to approximate the null distribution. Letting $\zeta_i(y) = \zeta(y; X_i, Y_i) = \mathbb{1}_y(Y_i) - F_0(y|X_i)$, we first define the sample covariance of $\mathbb{B}_P$. Under $P_0 = P_1 = P$, the underlying distributions of (X, Y) conditional on $Z = 1$ and $Z = 0$ are the same, so we can set $\hat{\Sigma}(s,t) = n^{-1}\sum_{i=1}^n \{\zeta_i(s) - \bar{\zeta}(s)\}\{\zeta_i(t) - \bar{\zeta}(t)\}$, where $s, t \in \mathbb{R}$ and $\bar{\zeta}(s) = n^{-1}\sum_{i=1}^n \zeta_i(s)$. Then, we draw L realizations $B_1, \ldots, B_L$ from the centered Gaussian process with covariance function $\hat{\Sigma}$ and calculate $g_\ell = \sup_{y\in\mathbb{R}} |B_\ell(y)|$ for testing equality or $g_\ell = \sup_{y\in\mathbb{R}} B_\ell(y)$ for testing first-order stochastic dominance for each realization B_ℓ. Finally, we reject the null hypothesis (equality or first-order stochastic dominance) if the corresponding statistic exceeds the threshold c_n, where c_n is the $(1-\alpha)$th sample quantile of $g_1, \ldots, g_L$.

S3. Auxiliary materials

S3.1. Algorithm

We summarize the proposed method in Algorithm 1, including estimation and inference.

S3.2. Monotonization via rearrangement

It is important to address a technical concern regarding the potential non-monotonicity of $\hat{F}_{\text{basd}}(y)$, which stems from the subtraction term in the definition (1) of the proposed estimator. While non-monotonicity might not pose problems in certain applications due to uniform convergence, it can be visually apparent and affect interpretation.

We address this issue with the rearrangement operation, a general strategy for monotonizing an initial estimate of an unknown monotonic function (Chernozhukov et al., 2009, 2010). The procedure involves an increasing bijective mapping $\varphi : \mathbb{R} \mapsto [0, 1]$, such as the cumulative distribution function of a standard Gaussian random variable. The increasing rearrangement of

Algorithm 1. Bias-Amended Semi-supervised Distribution (BASD)

```
Input: Observed data $\mathcal{L} = \{(Y_i, X_i)\}_{i=1}^{n}$ and $\mathcal{U} = \{X_i\}_{i=n+1}^{n+m}$, conditional distribution
       estimation algorithm $\mathcal{A}$, number of folds $K$, number of sampling for inference $B$.
Discrete index grid $-\infty < y_0 < \cdots < y_Q < +\infty$.
Randomly partition $\mathcal{L}$ into $K$ nearly equal-sized disjoint subsets $\mathcal{L}_1, \dots, \mathcal{L}_K$.
Randomly partition $\mathcal{U}$ into $K$ nearly equal-sized disjoint subsets $\mathcal{U}_1, \dots, \mathcal{U}_K$.
/* Estimation                                                          */
for $k = 1, \dots, K$ do
    Obtain $\{\hat{F}_k(y_q|x) : 0 \le q \le Q\}$ with learning algorithm $\mathcal{A}$ based on $\mathcal{L}_{-k}$.
    Get the estimator in the $k$th fold $\hat{F}_{k,\mathrm{B}}$ by (2), numerically approximated on grid.
Summarize the $K$ estimators by averaging to get $\hat{F}_{\mathrm{basd}}$ by (3).
/* Inference                                                           */
for $k = 1, \dots, K$ do
    Calculate $\mathcal{V}_{1,k} = \{(\mathbb{1}_{y_0}(Y_i), \dots, \mathbb{1}_{y_Q}(Y_i))^{\mathrm{T}}\}_{Y_i \in \mathcal{L}_k}$.
    Calculate $\mathcal{V}_{2,k} = \{(\mathbb{1}_{y_0}(Y_i) - \hat{F}_k(y_0|X_i), \dots, \mathbb{1}_{y_Q}(Y_i) - \hat{F}_k(y_Q|X_i))^{\mathrm{T}}\}_{(Y_i,X_i) \in \mathcal{L}_k}$.
Get the sample covariance $\hat{\Sigma}_1$ of $\mathbb{B}_1 \circ \mathbb{1}_y(Y)$ using $\mathcal{V}_{1,1} \cup \cdots \cup \mathcal{V}_{1,K}$.
Get the sample covariance $\hat{\Sigma}_2$ of $\mathbb{B}_2 \circ \{\mathbb{1}_y(Y) - F_0(y|X)\}$ using $\mathcal{V}_{2,1} \cup \cdots \cup \mathcal{V}_{2,K}$.
Calculate the sample covariance $\hat{\Sigma} = (n\hat{\Sigma}_1 + m\hat{\Sigma}_2)/(n+m)$ of $\mathbb{F}(y; \mathcal{G})$.
Generate $B$ realizations on grid $\mathbb{F}_1, \dots, \mathbb{F}_B \sim \mathcal{N}(0, \hat{\Sigma})$.
Calculate $g_b = \max_{0 \le q \le Q} |\mathbb{F}_b(y_q)|$ for $b = 1, \dots, B$.
Get the upper $(1-\alpha)$th sample quantile of $\{g_1, \dots, g_B\}$ as the threshold $L$.
Output: The cross-fitted cumulative distribution function estimator $\hat{F}_{\mathrm{basd}}$ and the
        confidence band $\hat{F}_{\mathrm{basd}} \pm Ln^{-1/2}$ numerically approximated on grid.
```

$\hat{F}_{\mathrm{basd}}(y)$, denoted as $\hat{F}^{\dagger}_{\mathrm{basd}}(y)$, is defined as $\hat{F}^{\dagger}_{\mathrm{basd}}(y) = Q^{\dagger} \circ \varphi(y)$, where

$$Q^{\dagger}(u) = \inf\left\{ y \in \mathbb{R} : \int_0^1 \mathbb{1}_y(Q(x))\mathrm{d}x \ge u \right\} \quad \text{and} \quad Q \equiv \hat{F}_{\mathrm{basd}} \circ \varphi^{-1}.$$

The key lies in leveraging the fact that rearrangement can monotonize functions from $[0,1]$ to $[0,1]$. For the cumulative distribution function estimator from $\mathbb{R}$ to $[0,1]$, we first use a bijection φ to obtain $Q = \hat{F}_{\mathrm{basd}} \circ \varphi^{-1} : [0,1] \mapsto [0,1]$. Then, Q is rearranged to yield the monotonized $Q^{\dagger}$, which is the quantile function of the random variable $Q(U)$ with $U \sim \mathrm{Unif}[0,1]$. Finally, to get the rearrangement of $\hat{F}_{\mathrm{basd}}$, we compose $Q^{\dagger}$ with the bijection φ, resulting $\hat{F}_{\mathrm{basd}}(y) = Q^{\dagger} \circ \varphi(y)$.

It is crucial that the rearrangement is a deterministic mathematical operation, ensuring a mapping from $\hat{F}_{\mathrm{basd}}(y)$ to $\hat{F}^{\dagger}_{\mathrm{basd}}(y)$ is non-random. With the conclusions on rearrangement (Chernozhukov et al., 2009, 2010), we have the following proposition for the rearrangement $\hat{F}^{\dagger}_{\mathrm{basd}}(y)$.

PROPOSITION S2. *Suppose $F(y)$ has $\nabla F(y) > 0$ for each y. If Theorem 1 holds for $\hat{F}_{\mathrm{basd}}(y)$, it still holds for the corresponding rearrangement $\hat{F}^{\dagger}_{\mathrm{basd}}(y)$.*

Proof of Proposition S2. The conclusions follow from Proposition 1 in Chernozhukov et al. (2009) and Corollary 3 in Chernozhukov et al. (2010). □

It is also worth emphasizing that by applying rearrangement to the conditional cumulative distribution function estimator $\hat{F}_k(y|x)$, Assumption 2 is naturally satisfied by the rearranged $\hat{F}_k(y|x)$. In particular, according to Proposition 5 of Chernozhukov et al. (2010), the rearrangement $\hat{F}_k^{\dagger}(y|x)$ of $\hat{F}_k(y|x)$ converges to the rearrangement $F_0^{\dagger}(y|x)$ of $F_0(y|x)$. Both $\hat{F}_k^{\dagger}(y|x)$ and $F_0^{\dagger}(y|x)$ are monotone with respect to y, so Assumption 2 can be verified.

S3.3. Parameter inference

In this section, we supplement some materials for parameter inference derived from the proposed estimator. As the main text, we denote the parameter of interest as $\theta(F)$, where $\theta(F)$ is Hadamard differentiable at F with derivative $\nabla\theta_F(\cdot)$.

We have shown in the main text that the asymptotic distribution of the plug-in estimator $\theta(\hat{F}_{\text{basd}})$ for a broader class of parameters, such as one sample U-statistics (Example 1), can be derived through Corollary 1. Here, we present that the parameter inference derived from the proposed distribution estimator aligns with some existing results in literature (Example S1–S2).

Example S1 *(Mean).* Let $\theta(F) = \int y \mathrm{d}F(y)$ be the mean of Y. Then,

$$\nabla\theta_F\{\mathbb{F}(\cdot;\mathcal{G})\} \overset{d}{\sim} \mathcal{N}\left(0, (1-\gamma)\mathrm{var}(Y) + \gamma\mathrm{var}\left\{Y - \int y \mathrm{d}F_0(y|X)\right\}\right),$$

which coincides with the results in Zhang et al. (2019) and Zhang & Bradic (2022), where they established the $\sqrt{n}$-consistency and asymptotic normality on semi-supervised mean inference. This alignment arises because $\int y\mathrm{d}F_0(y|X)$ can be taken as an approximation of the conditional mean function $E(Y \mid X) = \int y\mathrm{d}F(y|X)$.

Example S2 *(Quantile).* Let $\theta(F) = \inf\{y \in \mathbb{R} : F(y) \geq \tau\} = q_\tau$ be the τth quantile. Then,

$$\nabla\theta_F\{\mathbb{F}(\cdot;\mathcal{G})\} \overset{d}{\sim} \mathcal{N}\left(0, \frac{(1-\gamma)\mathrm{var}\{\mathbb{1}_{q_\tau}(Y)\} + \gamma\mathrm{var}\left\{\mathbb{1}_{q_\tau}(Y) - \int \mathbb{1}_{q_\tau}(y)\mathrm{d}F_0(y|X)\right\}}{\{\nabla F(q_\tau)\}^2}\right),$$

which aligns with the conclusion on semi-supervised quantile inference in Chakrabortty et al. (2022) when we specify $F_0(y|X) = \mathrm{pr}(Y \leq y \mid M^{\mathrm{T}}X)$ for some matrix $M \in \mathbb{R}^{p\times q}$.

Next, we give a concrete example to demonstrate Remark 4 for parameters under the framework of M-estimation or Z-estimation concretely. Define the population loss function $\ell(\beta) = E|(Y_1 - X_1^{\mathrm{T}}\beta) - (Y_2 - X_2^{\mathrm{T}}\beta)|$, where (Y_1, X_1) and (Y_2, X_2) are independent and identically distributed copies. This loss is closely related to Jaeckel's dispersion function with Wilcoxon scores (Wang et al., 2020). Denoting $f(z_1, z_2) = |z_1 - z_2|$ and $Z(\beta) = Y - X^{\mathrm{T}}\beta$, we target at estimating $\ell(\beta) = E\{f(Z_1(\beta), Z_2(\beta))\}$ as Example 1. We then take $Z(\beta)$ and X as the response and covariates, respectively, and the proposed estimator $\hat{F}_{\text{basd},Z}(\cdot;\beta)$ for the distribution of $Z(\beta)$ can be similarly obtained. Consequently, a semi-supervised empirical loss function would be $\hat{\ell}_{\text{basd}}(\beta) = \iint f(z_1, z_2)\mathrm{d}\hat{F}_{\text{basd},Z}(z_1;\beta)\mathrm{d}\hat{F}_{\text{basd},Z}(z_2;\beta)$. Minimizing such a loss could yield a semi-supervised estimator for the parameter β. We anticipate that the semi-supervised empirical loss function has a smaller asymptotic variance than its supervised counterpart. See Song et al. (2024) for some similar discussions.

Remark S1. Regarding the practical implementation of $\theta(\hat{F}_{\text{basd}})$, we give a computation method through discretization. In particular, we first monotonize the proposed $\hat{F}_{\text{basd}}$ through rearrangement (Chernozhukov et al., 2009), which could provide a monotonized distribution estimator without altering conclusions in Theorem 1 (see Remark 1 and Section S3.2). We consider the proposed distribution estimator after monotonization as $\hat{F}_{\text{basd}}$ in the following statements. A fine grid $q_1 < \cdots < q_N$ is employed such that $q_i = \min\{y : \hat{F}_{\text{basd}}(y) \geq i/(N+1)\}$, where $N \in \mathbb{Z}_+$ is a large integer. We consider the step function $\check{F}_{\text{basd}}(y) = N^{-1}\sum_{i=1}^N \mathbb{1}(q_i \leq y)$ as an approximation of $\hat{F}_{\text{basd}}$, and it can be easily derived that $\sup_y |\hat{F}_{\text{basd}}(y) - \check{F}_{\text{basd}}(y)| \leq (N+1)^{-1}$. The parameter based on the step function $\theta(\check{F}_{\text{basd}})$ can be easily calculated. Thus,

Table S1. *Comparison of the proposed $\hat{F}_{\mathrm{basd}}$ for different choices of the number of cross-fitting folds K. The overall mean squared error of the empirical distribution function is denoted by* MSE_0. *The level of the confidence band is set as $\alpha = 0.1$*

	$\mathrm{MSE}/\mathrm{MSE}_0$				Coverage				Length×10^2			
K	2	5	10	20	2	5	10	20	2	5	10	20
$p = 100$												
$\hat{F}_{\mathrm{basd}}$-GAMLSS	0.54	0.51	0.51	0.51	0.88	0.90	0.90	0.89	5.63	5.58	5.57	5.56
$\hat{F}_{\mathrm{basd}}$-Engression	0.78	0.87	0.82	–	0.90	0.88	0.88	–	7.10	7.14	7.16	–
$\hat{F}_{\mathrm{basd}}$-DRF	0.93	0.90	0.90	0.90	0.90	0.90	0.90	0.90	7.16	7.06	7.05	7.03
$\hat{F}_{\mathrm{ecdf}}$	1.00	1.00	1.00	1.00	0.90	0.90	0.90	0.90	7.36	7.36	7.36	7.36
$p = 500$												
$\hat{F}_{\mathrm{basd}}$-GAMLSS	0.92	0.90	0.88	0.87	0.88	0.88	0.88	0.88	7.14	7.06	7.04	7.03
$\hat{F}_{\mathrm{basd}}$-Engression	0.94	0.83	0.82	–	0.89	0.90	0.88	–	8.08	7.66	7.57	–
$\hat{F}_{\mathrm{basd}}$-DRF	0.99	0.99	0.99	0.99	0.87	0.87	0.86	0.86	7.34	7.35	7.35	7.33
$\hat{F}_{\mathrm{ecdf}}$	1.00	1.00	1.00	1.00	0.87	0.87	0.87	0.87	7.37	7.37	7.37	7.37

$\hat{F}_{\mathrm{basd}}$-GAMLSS, $\hat{F}_{\mathrm{basd}}$-Engression and $\hat{F}_{\mathrm{basd}}$-DRF represent the $\hat{F}_{\mathrm{basd}}$ estimators with $\mathcal{A}$ specified as the boosting method for fitting generalized additive models (Hofner et al., 2016), the engression estimator (Shen & Meinshausen, 2024) and distributional random forests (Ćevid et al., 2022), respectively. "–" means that the experiment is stopped before finishing due to a long running time.

$\theta(\hat{F}_{\mathrm{basd}}) \approx \theta(\check{F}_{\mathrm{basd}})$, and the numerical error $|\theta(\hat{F}_{\mathrm{basd}}) - \theta(\check{F}_{\mathrm{basd}})|$ depends on both N and $\theta(\cdot)$.

S4. Additional Numerical Studies

S4.1. Implementation details of the proposed method

The implementation of conditional cumulative distribution function estimation for the proposed distribution estimators is outlined below. For the boosting method to fit generalized additive models for location, scale, and shape (Hofner et al., 2016), we employ the R package `gamboostLSS`, as described in their article. The R package `drf` is used for distributional random forests (Ćevid et al., 2022). For the neural network-based engression method (Shen & Meinshausen, 2024), we utilize the R package `engression`. In the generating model $Y = \mu(X) + \sigma(X)\varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$, the true conditional cumulative distribution function used in the BASD* is explicitly given by $\Phi[\{y - \mu(X)\}/\sigma(X)]$. The number of replications in the main text is 200.

S4.2. Additional simulations for distribution inference

In the main text, we considered 10 cross-fitting folds $K = 10$ for the simulations. Here, we present results for other choices of K and provide guidance on selecting K in practice. In Table S1, we replicate the simulation settings from Table 1, and evaluate $K \in \{2, 5, 10, 20\}$. The results indicate that efficiency generally improves as K increases, reflecting the fact that larger K values allow for more samples to be used in training the conditional distribution models. However, a larger K also results in greater computational costs, as the conditional distribution models must be fitted more frequently. Thus, there is a trade-off between estimation accuracy and computational cost. We find that the proposed estimators with $K \geq 5$ yields similar performance for fitting conditional distribution models. Hence, we set $K = 10$ in Section 3 and recommend choosing $K \geq 5$ in practice.

Table S2. *Comparison of the proposed $\hat{F}_{\mathrm{basd}}$ and the empirical cumulative distribution function $\hat{F}_{\mathrm{ecdf}}$ in the same settings as Table 1 except under different p's. The overall mean squared error of the empirical cumulative distribution function is denoted by MSE_0. The level of the confidence band is set as $\alpha = 0.1$*

	$p=10$			$p=1000$		
	$\mathrm{MSE}/\mathrm{MSE}_0$	Coverage	Length$\times 10^2$	$\mathrm{MSE}/\mathrm{MSE}_0$	Coverage	Length$\times 10^2$
$\tilde{F}^*_{\mathrm{basd}}$	0.47	0.88	5.52	0.48	0.90	5.53
$\hat{F}_{\mathrm{basd}}$-GAMLSS	0.46	0.90	5.53	0.97	0.90	7.30
$\hat{F}_{\mathrm{basd}}$-Engression	0.48	0.90	5.60	1.04	0.92	8.29
$\hat{F}_{\mathrm{basd}}$-DRF	0.48	0.88	5.56	1.00	0.89	7.37
$\hat{F}_{\mathrm{ecdf}}$	1.00	0.88	7.37	1.00	0.90	7.35

$\tilde{F}^*_{\mathrm{basd}}$, $\tilde{F}_{\mathrm{basd}}$ with known $F(y|x)$; $\hat{F}_{\mathrm{basd}}$-GAMLSS, $\hat{F}_{\mathrm{basd}}$-Engression and $\hat{F}_{\mathrm{basd}}$-DRF represent the $\hat{F}_{\mathrm{basd}}$ estimators with $\mathcal{A}$ specified as the boosting method for fitting generalized additive models (Hofner et al., 2016), the engression estimator (Shen & Meinshausen, 2024) and distributional random forests (Ćevid et al., 2022), respectively.

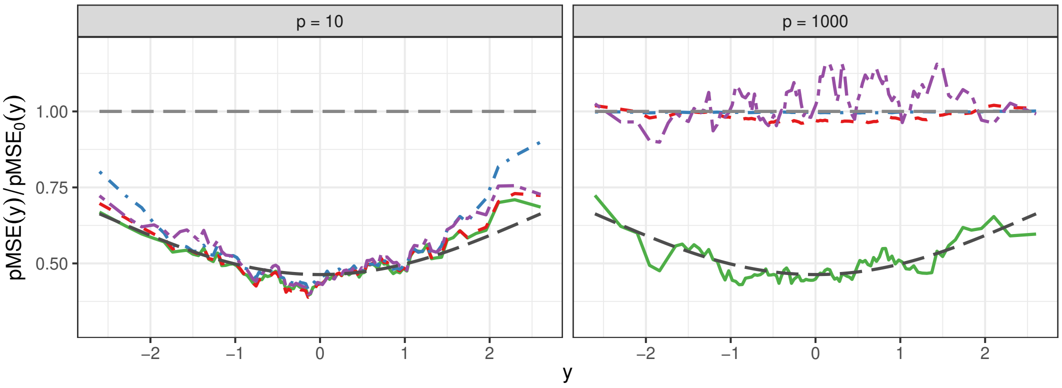

Fig. S1. The ratio of pointwise mean squared errors of $\hat{F}_{\mathrm{basd}}$ in the same settings as Table S2. The pointwise mean squared error of $\hat{F}_{\mathrm{ecdf}}$ is denoted by $\mathrm{pMSE}_0(\mathrm{y})$. The plot includes $\tilde{F}^*_{\mathrm{basd}}$ (green, solid), $\hat{F}_{\mathrm{basd}}$-GAMLSS (red, dashed), $\hat{F}_{\mathrm{basd}}$-Engression (purple, twodash), and $\hat{F}_{\mathrm{basd}}$-DRF (blue, dotdash). The curve (darkgray, longdash) is the theoretical relative semi-parametric efficiency lower bound. See the footnotes below Table S2 for the concrete explanation of methods.

In addition to the results for $p = 100$ and $p = 500$ presented in the main text, we also provide results for $p = 10$ and $p = 1000$ under the same setting in Table S2 and Fig. S1, using $K = 5$. 315 We can see that, our proposed framework still works as $\hat{F}_{\mathrm{basd}}$ with true conditional distribution function (i.e., BASD*) still achieves the semi-parametric efficiency lower bound, but the preformance of conditional distribution estimators (and further the proposed distribution estimators) deteriorates. Therefore, we recommend our proposed framework to be used when conditional distribution function could be well estimated, which typically happens for the case with a relative small p to n. 320

Finally, we consider a complex model $Y_i = \mu(X_i) + \sigma(X_i)\varepsilon_i$ $(i = 1, \ldots, n+m)$ with $\mu(X_i) = 3\cos(\sum_{j=1}^{3} X_{ij})$ and $\sigma(X_i) = \exp(\sum_{j=4}^{7} X_{ij}/2)/3$, where $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}} \sim \mathcal{N}_p(0_p, \mathrm{I}_p)$ and $\varepsilon_i \sim \mathcal{N}(0,1)$ are independent. We set $(n, p, K) = (1000, 50, 5)$ and $m \in \{0.5n, n, 5n, 10n\}$, and list the results in Table S3 and Fig. S2. The results indicate that as m increases, the performance of the proposed estimators improves, which coincides with our theory. 325 Also, it can be observed that our distribution estimators perform well in this setting.

Table S3. *Comparison of the proposed $\hat{F}_{\mathrm{basd}}$ and the empirical cumulative distribution function $\hat{F}_{\mathrm{ecdf}}$ in the nonlinear setting. The overall mean squared error of the empirical distribution function is denoted by* MSE_0. *The level of the confidence band is set as* $\alpha = 0.1$

	$\mathrm{MSE}/\mathrm{MSE}_0$				Coverage				Length$\times 10^2$			
m/n	0.5	1	5	10	0.5	1	5	10	0.5	1	5	10
$\tilde{F}^*_{\mathrm{basd}}$	0.73	0.58	0.33	0.28	0.90	0.92	0.92	0.92	6.43	5.93	4.74	4.44
$\hat{F}_{\mathrm{basd}}$-GAMLSS	0.98	0.95	0.89	0.88	0.91	0.90	0.90	0.90	7.25	7.21	7.10	7.09
$\hat{F}_{\mathrm{basd}}$-Engression	0.90	0.89	0.82	0.75	0.90	0.91	0.89	0.92	7.17	7.10	6.90	6.86
$\hat{F}_{\mathrm{basd}}$-DRF	0.98	0.97	0.95	0.95	0.90	0.90	0.92	0.90	7.30	7.27	7.24	7.21
$\hat{F}_{\mathrm{ecdf}}$	1.00	1.00	1.00	1.00	0.91	0.91	0.91	0.91	7.36	7.36	7.36	7.36

$\tilde{F}^*_{\mathrm{basd}}$, $\tilde{F}_{\mathrm{basd}}$ with known $F(y|x)$; $\hat{F}_{\mathrm{basd}}$-GAMLSS, $\hat{F}_{\mathrm{basd}}$-Engression and $\hat{F}_{\mathrm{basd}}$-DRF represent the $\hat{F}_{\mathrm{basd}}$ estimators with $\mathcal{A}$ specified as the boosting method for fitting generalized additive models (Hofner et al., 2016), the engression estimator (Shen & Meinshausen, 2024) and distributional random forests (Ćevid et al., 2022), respectively.

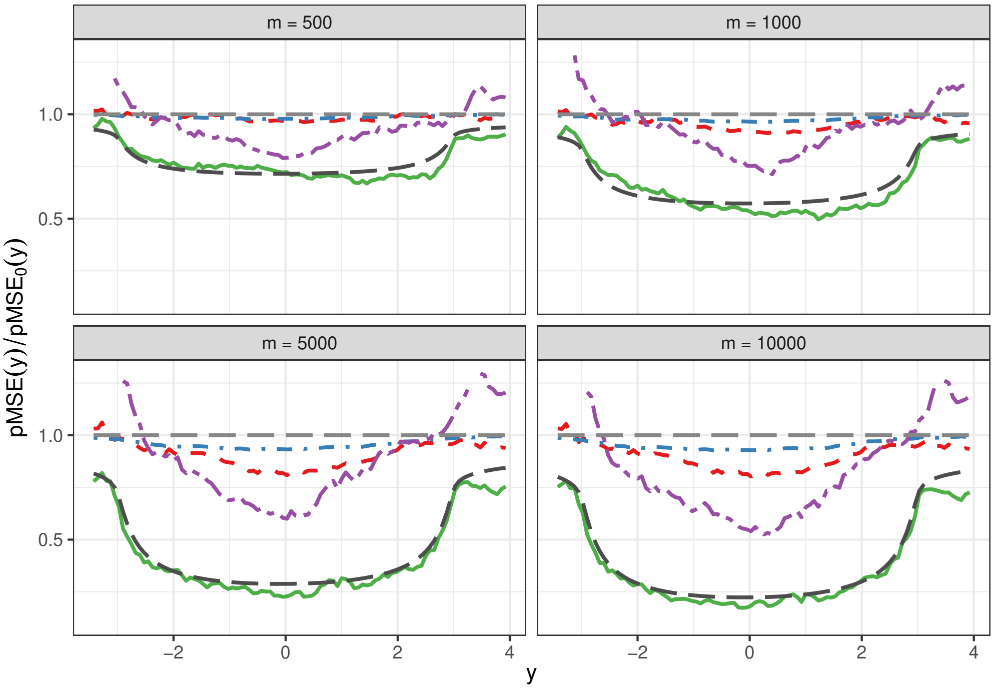

Fig. S2. The ratio of pointwise mean squared errors of $\hat{F}_{\mathrm{basd}}$ in the nonlinear setting. The pointwise mean squared error of $\hat{F}_{\mathrm{ecdf}}$ is denoted by $\mathrm{pMSE}_0(\mathrm{y})$. The plot includes $\tilde{F}^*_{\mathrm{basd}}$ (green, solid), $\hat{F}_{\mathrm{basd}}$-GAMLSS (red, dashed), $\hat{F}_{\mathrm{basd}}$-Engression (purple, twodash), and $\hat{F}_{\mathrm{basd}}$-DRF (blue, dotdash). The curve (darkgray, longdash) is the theoretical relative semi-parametric efficiency lower bound. See the footnotes below Table S3 for the concrete explanation of methods.

S4.3. Mean inference

We proceed to show the performance of the proposed framework applied to parameter inference (see Section 2.3). Specifically, we consider the mean inference of Y, i.e. $\theta = \int y \mathrm{d}F(y)$. The data is generated the same as that in Section 3. We compare our procedure with several popular methods: the sample mean calculated by the labeled data, the semi-supervised mean estimator with least-squares in Zhang et al. (2019) and the one incorporating random forests in Zhang &

Table S4. *Comparison of ratios of mean squared errors of semi-supervised mean estimators to supervised sample mean*

	Plug-in Estimators $\theta(\hat{F})$				Literature	
	$\tilde{F}^*_{\mathrm{basd}}$	$\hat{F}_{\mathrm{basd}}$-GAMLSS	$\hat{F}_{\mathrm{basd}}$-Engression	$\hat{F}_{\mathrm{basd}}$-DRF	Zhang et al. (2019)	Zhang & Bradic (2022)
$p = 10$	0.30	0.30	0.32	0.33	0.29	0.34
$p = 100$	0.30	0.32	0.51	0.84	0.30	0.68
$p = 500$	0.31	0.82	0.46	0.98	0.44	0.90
$p = 1000$	0.33	0.96	0.68	0.99	–	0.97

$\tilde{F}^*_{\mathrm{basd}}$, $\tilde{F}_{\mathrm{basd}}$ with known $F(y|x)$; $\hat{F}_{\mathrm{basd}}$-GAMLSS, $\hat{F}_{\mathrm{basd}}$-Engression and $\hat{F}_{\mathrm{basd}}$-DRF represent the $\hat{F}_{\mathrm{basd}}$ estimators with $\mathcal{A}$ specified as the boosting method for fitting generalized additive models (Hofner et al., 2016), the engression estimator (Shen & Meinshausen, 2024) and distributional random forests (Ćevid et al., 2022), respectively. "–" indicates the method is not applicable in the current setting.

Table S5. *Mean squared error ratio $\hat{E}(\hat{p}_{\mathrm{basd}} - p_0)^2/\hat{E}(\hat{p}_{\mathrm{CP}} - p_0)^2$ in different settings*

	$X_0 = (3,1)$				$X_0 = (4,4)$			
b_0	p_0	m			p_0	m		
		500	1000	5000		500	1000	5000
10	0.531	0.996	0.993	0.990	0.075	0.966	0.963	0.963
15	0.707	0.891	0.883	0.878	0.097	0.957	0.952	0.950
20	0.774	0.884	0.872	0.864	0.125	0.960	0.958	0.954
30	0.848	0.881	0.867	0.862	0.228	0.966	0.961	0.960

Bradic (2022). We compare the ratio of mean squared errors of the semi-supervised estimator to the supervised sample mean in Table S4. The value less than one means that the corresponding semi-supervised mean estimator is more efficient than the sample mean. We can see that the semi-supervised mean estimators all works, and our proposed estimators are comparable to those in the literature.

S4.4. Conformal p-value

We now compare the semi-supervised conformal p-values with the conformal p-values computed in Jin & Candès (2023). The model is generated by $Y_i = \mu(X_i) + \sigma(X_i)\varepsilon_i$, where $\mu(X) = 2\sum_{j=1}^{2}(X_j^2 - 3)$, $\sigma(X) = 5.5 - |\mu(X)|$ and $\varepsilon_i \sim \mathcal{N}(0,1)$ is the i.i.d. random noise independent of X_i. Two-dimensional covariates are independently generated as $X_i \sim \mathrm{Unif}([0,5]^2)$. We set $n = 100$, $m \in \{500, 1000, 5000\}$ and use random forests to estimate model $\hat{\mu}(\cdot)$ with another n training data. Our proposed method employs the distributional random forests (Ćevid et al., 2022) as learning algorithm $\mathcal{A}$.

We consider two test points of interest $X_0 \in \{(3,1), (4,4)\}$ and the hypothesis $\mathbb{H}_0 : Y_0 \leq b_0$ in (S11) with $b_0 \in \{10, 15, 20, 30\}$, ensuring a diverse range of corresponding p-values. The mean squared error ratio $\hat{E}(\hat{p}_{\mathrm{basd}} - p_0)^2/\hat{E}(\hat{p}_{\mathrm{CP}} - p_0)^2$ is listed in Table S5. For both test settings, all ratio values are less than 1 and the ratio decreases as the number of unlabeled data increases.

S4.5. Local distributional treatment effect

We illustrate the proposed method for testing the local distributional treatment effect using observed data $\{(Y_i, D_i, Z_i, X_i)\}_{i=1}^{n}$, where $Z_i \sim \mathrm{Bernoulli}(0.4)$ and $D_i = D_{0i} + Z_i(D_{1i} - D_{0i})$ with paired (D_{0i}, D_{1i}) generated independently such that $\mathrm{pr}(D_{0i} = 0, D_{1i} = 0) = \mathrm{pr}(D_{0i} =$

Table S6. *Empirical sizes ($\Delta = 0$) and powers ($\Delta > 0$) for testing local distributional treatment effects when $\alpha = 0.1$, $p = 5$ and $n \in \{200, 1000\}$*

	$\mathbb{H}_0^{\mathrm{eq}} : F_{\mathrm{C}}^{(1)} = F_{\mathrm{C}}^{(0)}$				$\mathbb{H}_0^{\mathrm{fsd}} : F_{\mathrm{C}}^{(1)} \geq F_{\mathrm{C}}^{(0)}$			
	$\Delta = 0$	0.2	0.4	0.6	$\Delta = 0$	0.2	0.4	0.6
$n = 200$								
BASD	0.087	0.173	0.422	0.708	0.108	0.286	0.586	0.844
Abadie (2002)	0.101	0.123	0.184	0.267	0.112	0.193	0.288	0.398
$n = 1000$								
BASD	0.098	0.560	0.985	1.000	0.086	0.709	0.998	1.000
Abadie (2002)	0.098	0.213	0.448	0.733	0.108	0.299	0.603	0.848

BASD, the test based on $\hat{F}_{\mathrm{basd}}^{(0)}(y) - \hat{F}_{\mathrm{basd}}^{(1)}(y)$ in (S13).

$1, D_{1i} = 1) = 1/5$ and $\mathrm{pr}(D_{0i} = 0, D_{1i} = 1) = 3/5$. For generating $Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i})$, we consider $Y_{0i} = X_i^{\mathrm{T}}\beta + \varepsilon_i$ with $\beta = (2, -1, 0.5, 0, \ldots, 0)^{\mathrm{T}}$, $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $X_i \sim \mathcal{N}(0, \mathrm{I}_p)$, and $Y_{1i} = Y_{0i} + \Delta$, where Δ is the signal magnitude with $\Delta \in \{0, 0.2, 0.4, 0.6\}$. The goal is to test the hypotheses $\mathbb{H}_0^{\mathrm{eq}} : F_{\mathrm{C}}^{(1)} = F_{\mathrm{C}}^{(0)}$ and $\mathbb{H}_0^{\mathrm{fsd}} : F_{\mathrm{C}}^{(1)} \geq F_{\mathrm{C}}^{(0)}$, as aforementioned in Section S2.2.

We compare the empirical sizes and powers of the test equipped with our proposed method and the test proposed in Abadie (2002). Our proposed method chooses the distributional single-index model (Henzi et al., 2021) as the learning algorithm $\mathcal{A}$. The nominal level is set as $\alpha = 0.1$ and the critical value is obtained from 2,000 resamplings. Table S6 reports the results on the equality and first-order stochastic dominance tests under $p = 5$ and $n \in \{200, 1000\}$. The proposed method has empirical sizes close to the nominal level, and it also performs more powerful than Abadie (2002) as the signal magnitude increases. This improvement is expected as our method properly uses more information from covariates.

S4.6. Real data analysis

We consider an application of estimating homeless people in Los Angeles Country. Homelessness has been a public issue for America since near a century ago (Rossi, 1991). A key question for the demographers is to estimate the number of homeless in a specific region, as this information can help stakeholders (e.g., homeless service advocates and selected government agencies) determine the required social resources. However, it is challenging because the homeless are often dispersed (Rossi, 1991). Although visiting homeless shelters can provide some data, many homeless individuals remain uncounted, as they may not utilize these services.

We use data from a study conducted by Los Angeles Homeless Services Authority in 2004–2005, which was also analysed by Zhang et al. (2019). Los Angeles County spans over 4000 square miles and includes 2054 census tracts, making a full street survey prohibitively expensive. Therefore, a stratified spatial sampling of census tracts was employed. First, 244 tracts believed to have large numbers of homeless were visited. Next, for the rest of the tracts, 265 of them were randomly selected and visited, leaving 1545 unvisited tracts. In addition to homeless counts, some covariates known to correlate with the response were available for all 2054 census tracts (Kriegler & Berk, 2010), and seven of these covariates were included in our analysis (see Table S7).

The total number of homeless in Los Angeles can be calculated through estimating the average number of homeless per tract in all 1810 non-preselected tracts. To do this, we apply the proposed framework for mean inference to these 1810 samples, which include 265 labeled and 1545 unlabeled samples. The conditional cumulative distribution function estimation methods are as described in Section 3. We compare the semi-supervised mean estimators proposed by

Table S7. *Covariate names in the homeless data*

Name	Description
Perc.Industrial	% of land used for industrial purposes
Perc.Residential	% of land used for residential purposes
Perc.Vacant	% of land that is vacant
Perc.Commercial	% of land used for commercial purposes
Perc.OwnerOcc	% of owner-occupied housing units
Perc.Minority	% of population that is non-Caucasian
MedianHouseholdIncome	Median household income

Table S8. *Estimated average number of homeless per tract in all 1810 non-preselected tracts. The Length refers to the length of 95% confidence interval*

	Plug-in Estimators $\int y\mathrm{d}\hat{F}(y)$			Literature		Sample Mean
	$\hat{F}_{\mathrm{basd}}$-GAMLSS	$\hat{F}_{\mathrm{basd}}$-DRF	$\hat{F}_{\mathrm{basd}}$-Engression	Zhang et al. (2019)	Zhang & Bradic (2022)	
Estimate	21.89	22.45	22.75	22.38	22.35	21.61
Length	7.70	7.46	8.15	7.40	7.77	7.75

$\hat{F}_{\mathrm{basd}}$-GAMLSS, $\hat{F}_{\mathrm{basd}}$-Engression and $\hat{F}_{\mathrm{basd}}$-DRF represent the $\hat{F}_{\mathrm{basd}}$ estimators with $\mathcal{A}$ specified as the boosting method for fitting generalized additive models (Hofner et al., 2016), the engression estimator (Shen & Meinshausen, 2024) and distributional random forests (Ćevid et al., 2022), respectively.

Table S9. *Estimated τth quantiles of the homeless street count in all 1810 non-preselected tracts*

	Plug-in Estimators $\inf\{y \in \mathbb{N} : \hat{F}(y) \geq \tau\}$			Sample Quantile
	$\hat{F}_{\mathrm{basd}}$-GAMLSS	$\hat{F}_{\mathrm{basd}}$-DRF	$\hat{F}_{\mathrm{basd}}$-Engression	
$\tau = 0.1$	2	2	2	2
$\tau = 0.3$	5	5	5	5
$\tau = 0.5$	12	12	13	12
$\tau = 0.7$	24	25	26	24
$\tau = 0.9$	47	47	48	47

See the footnotes below Table S8 for the concrete explanation of methods.

Zhang et al. (2019) using least squares and by Zhang & Bradic (2022) using random forest implementation. The results are summarized in Table S8.

According to Zhang et al. (2019), it is reasonable to obtain estimates higher than the sample mean, and the semi-supervised estimates align with this analysis. We also observe that, the proposed estimate using the conditional distribution estimator "engression" (Shen & Meinshausen, 2024), yields a longer confidence interval, which can be attributed to the unstable fitting process of the neural network. Overall, our estimates are comparable to the specialized methods reported in the literature.

In addition to estimating the average number of homeless per tract, our framework can provide more detailed information. The distribution of homeless counts is highly skewed: 75 percent of the observed counts are fewer than 28 people, while 22 of the 265 tracts have at least 50 homeless individuals. To ensure adequate resource allocation, we recommend using the 70th percentile estimate for budgeting purposes. Our semi-supervised quantile estimates suggest that higher values than those obtained from the sample quantile should be considered.

REFERENCES

ABADIE, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *J. Am. Statist. Assoc.* **97**, 284–292.

ANGRIST, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *Am. Econ. Rev.* **80**, 313–336.

ANGRIST, J. D., IMBENS, G. W. & RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–455.

ANGRIST, J. D. & KRUEGER, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.* **106**, 979–1014.

BATES, S., CANDÈS, E., LEI, L., ROMANO, Y. & SESIA, M. (2023). Testing for outliers with conformal p-values. *Ann. Statist.* **51**, 149–178.

BOUCHERON, S., LUGOSI, G. & MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford University Press.

ĆEVID, D., MICHEL, L., NÄF, J., BÜHLMANN, P. & MEINSHAUSEN, N. (2022). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *J. Mach. Learn. Res.* **23**, 1–79.

CHAKRABORTTY, A., DAI, G. & CARROLL, R. J. (2022). Semi-supervised quantile estimation: Robust and efficient inference in high dimensional settings. *arXiv*: 2201.10208 .

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. & GALICHON, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96**, 559–575.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. & GALICHON, A. (2010). Quantile and probability curves without crossing. *Econometrica* **78**, 1093–1125.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. & MELLY, B. (2013). Inference on counterfactual distributions. *Econometrica* **81**, 2205–2268.

HENZI, A., KLEGER, G.-R. & ZIEGEL, J. F. (2021). Distributional (single) index models. *J. Am. Statist. Assoc.* **118**, 489–503.

HOFNER, B., MAYR, A. & SCHMID, M. (2016). gamboostlss: An r package for model building and variable selection in the gamlss framework. *J. Stat. Softw.* **74**, 1–31.

IMBENS, G. W. & RUBIN, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econom. Stud.* **64**, 555–574.

IMBENS, G. W. & RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

JIN, Y. & CANDÈS, E. J. (2023). Selection by prediction with conformal p-values. *J. Mach. Learn. Res.* **24**, 1–41.

KRIEGLER, B. & BERK, R. (2010). Small area estimation of the homeless in Los Angeles: an application of cost-sensitive stochastic gradient boosting. *Ann. Appl. Stat.* **4**, 1234–1255.

MCFADDEN, D. (1989). Testing for stochastic dominance. In *Studies in the Economics of Uncertainty: In Honor of Josef Hadar*. New York: Springer, pp. 113–134.

ROMANO, Y., PATTERSON, E. & CANDÈS, E. (2019). Conformalized quantile regression. In *Proc. 33rd Int. Conf. Neural Info. Proces. Syst.*, vol. 32.

ROSSI, P. H. (1991). Strategies for homeless research in the 1990s. *Hous. Policy Debate* **2**, 1027–1055.

SHEN, X. & MEINSHAUSEN, N. (2024). Engression: Extrapolation through the lens of distributional regression. *arXiv*: 2307.00835 .

SONG, S., LIN, Y. & ZHOU, Y. (2024). A general M-estimation theory in semi-supervised framework. *J. Amer. Statist. Assoc.* **119**, 1065–1075.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

VAN DER VAART, A. W. & WELLNER, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer Science & Business Media.

VOVK, V., GAMMERMAN, A. & SHAFER, G. (2005). *Algorithmic Learning in a Random World*. New York: Springer Science & Business Media.

WANG, L., PENG, B., BRADIC, J., LI, R. & WU, Y. (2020). A tuning-free robust and efficient approach to high-dimensional regression. *J. Am. Statist. Assoc.* **115**, 1700–1714.

ZHANG, A., BROWN, L. D. & CAI, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *Ann. Statist.* **47**, 2538–2566.

ZHANG, Y. & BRADIC, J. (2022). High-dimensional semi-supervised learning: In search of optimal inference of the mean. *Biometrika* **109**, 387–403.

ZHANG, Y., JIANG, H., REN, H., ZOU, C. & DOU, D. (2022). AutoMS: Automatic model selection for novelty detection with error rate control. In *Proc. 35rd Int. Conf. Neural Info. Proces. Syst.*, vol. 35.