Conformal Inference of Counterfactuals and Individual Treatment Effects

Lihua Lei and Emmanuel J. Candès

Kiho Park
parkkiho@uchicago.edu

1 From Average Effects To Individual Effects

Estimating the average treatment effect (ATE) for a population of interest has been a main focus of causal inference, but it may be insufficient to validate an intervention. Thus, importance of individual treatment effect (ITE) is highlighted. The treatment effect heterogeneity is a great concern in a wide range of fields. The most existing works have targeted the conditional average treatment effects (CATE) instead of ATE.

The reliability of modern machine learning methods is typically under-studied. The confidence intervals for CATE or prediction intervals for ITE produced by frequently discussed methods have unsatisfactory or unacceptable coverage. We leverage ideas from conformal inference to construct valid prediction intervals for ITE under the potential outcome framework. We have two challenges:

(1) Construct intervals for ITE with reliable coverage for subjects in the study, for which one of the potential outcomes is missing

(2) Construct intervals for ITE with reliable coverage for subjects not in the study, and for which both potential outcomes are missing

2 From Point Estimates To Interval Estimates

2.1 Problem setup

Given n subjects, denote by $T_i \in \{0, 1\}$ the binary treatment indicator, by $(Y_i(1), Y_i(0))$ the pair of potential outcomes, and by X_i the vector of other covariates. We assume that

$$(Y_i(1), Y_i(0), T_i, X_i) \overset{i.i.d.}{\sim} (Y(1), Y(0), T, X)$$

where $(Y(1), Y(0), T, X)$ denotes a generic random vector. Under the stable unit treatment value assumption (SUTVA), the observed dataset comprises triple $(Y_i^{\text{obs}}, T_i, X_i)$ where $Y_i^{\text{obs}} = Y_i(1)$ if $T_i = 1$ and $Y_i^{\text{obs}} = Y_i(0)$ if $T_i = 0$. The individual treatment effect (ITE) is defined as

$$\tau_i = Y_i(1) - Y_i(0)$$

This ITE is unobserved and has to be inferred. We assume the strong ignorability:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$$

This assumption rules out any source of unmeasured confounders which affect both the treatment assignment and the potential outcomes.

2.2 Traditional inferential targets

CATE is defined as

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = m_1(x) - m_0(x)$$

where $m_1(x) = \mathbb{E}[Y(1) \mid X = x]$ and $m_0(x) = \mathbb{E}[Y(0) \mid X = x]$. Under the strong ignorability assumption,

$$m_1(x) = \mathbb{E}[Y^{\text{obs}} \mid X = x, T = 1] \text{ and } m_0(x) = \mathbb{E}[Y^{\text{obs}} \mid X = x, T = 0]$$

Thus, the estimation problem reduces to that of estimating certain conditional expectations. An alternative estimand is the conditional quantile treatment effect (CQTE), which is defined as the difference between the β-th quantiles of the distributions of $Y(1)$ and $Y(0)$.

2.3 Coverage of interval estimates

We take counterfactuals $(Y_i(1), Y_i(0))$'s and the ITE τ_i's as objects of inference. We want to construct intervals $\widehat{C}_1(x)$ and $\widehat{C}_0(x)$ obeying

$$\mathbb{P}\left(Y(t) \in \widehat{C}_t(X)\right) \geq 1 - \alpha, \quad (t = 0, 1)$$

Also, we seek $\widehat{C}_{\text{ITE}}(x)$ obeying marginal coverage

$$\mathbb{P}\left(Y(1) - Y(0) \in \widehat{C}_{\text{ITE}}(X)\right) \geq 1 - \alpha$$

Coverage would hold conditionally on $X = x$. Without modeling assumptions, it is known to be impossible to construct non-trivial prediction intervals with guaranteed conditional coverage.

2.4 General coverage criteria

In classical causal inference, it is often argued that the average treatment effect on the treated (ATT), defined as $\mathbb{E}[Y(1) - Y(0) \mid T = 1]$, is preferrable to the ATE. The identification of ATT is easier sine it requires a weaker assumption on the propensity scores. The criterion can be modified as

$$\mathbb{P}(Y(t) \in \widehat{C}_t(X) \mid T = 1) \geq 1 - \alpha, \quad (t = 0, 1)$$

Similarly we may consider the average treatment effect on the controls (ATC) by conditioning on $T = 0$. These considerations motivate the following general criterion:

$$\mathbb{P}_{(X, Y(t)) \sim Q_X \times P_{Y(t) \mid X}}(Y(t) \in \widehat{C}_t(X)) \geq 1 - \alpha, \quad (t = 0, 1)$$

Q_X can be chosen to be the covariate distribution in the target population, like $Q_X = P_{X \mid T=1}$ for ATT.

3 From Observables To Counterfactuals

3.1 Counterfactuals and covariate shift

We construct prediction intervals for $Y(1)$ and $Y(0)$ using the i.i.d. observations $(Y_i^{\text{obs}}, T_i, X_i)$, so only the sample in the treatment (resp. control) group are useful for constructing $\widehat{C}_1(x)$ (resp. $\widehat{C}_0(x)$).

Under SUTVA and strong ignorability assumption, the joint distribution of (X, Y^{obs}) of the observed treated samples is given by

$$P_{X|T=1} \times P_{Y(1)|X}$$

For interval estimates under covariate shifts between the sampling distribution and the target distribution $Q_X \times P_{Y(1)|X}$, we shall rely on weighted conformal inference developed by Tibshirani et al. (2019).

3.2 Weighted conformal inference

We extend the weighted conformal inference to CQR (conformal quantile regression). The weight is the likelihood ratio $w(x) = dQ_X(x)/dP_X(x)$ where P_X is the sampling distribution. Also, we use quantile estimates $\widehat{q}_{\alpha_{\mathrm{lo}}}$ and $\widehat{q}_{\alpha_{\mathrm{hi}}}$ of the α_{lo}-th and α_{hi}-th conditional quantiles of $Y \mid X = x$.

Algorithm 1 Weighted split-CQR

```
Input: level $\alpha$, data $\mathcal{Z} = (X_i, Y_i)_{i \in \mathcal{I}}$, testing point $x$, function $\widehat{q}_\beta(x; \mathcal{D})$ to fit $\beta$-th conditional quantile and function $\widehat{w}(x; \mathcal{D})$ to fit the weight function at $x$ using $\mathcal{D}$ as data
Procedure:
    1: Split $\mathcal{Z}$ into a training fold $\mathcal{Z}_{\mathrm{tr}} = (X_i, Y_i)_{i \in \mathcal{I}_{\mathrm{tr}}}$ and a calibration fold $\mathcal{Z}_{\mathrm{ca}} = (X_i, Y_i)_{i \in \mathcal{I}_{\mathrm{ca}}}$
    2: For each $i \in \mathcal{I}_{\mathrm{ca}}$, compute the score $V_i = \max\{\widehat{q}_{\alpha_{\mathrm{lo}}}(X_i; \mathcal{Z}_{\mathrm{tr}}) - Y_i, Y_i - \widehat{q}_{\alpha_{\mathrm{hi}}}(X_i; \mathcal{Z}_{\mathrm{tr}})\}$
    3: For each $i \in \mathcal{I}_{\mathrm{ca}}$, compute the weight $W_i = \widehat{w}(X_i; \mathcal{Z}_{\mathrm{tr}}) \in [0, \infty)$
    4: Compute the normalized weights $\widehat{p}_i(x) = \frac{W_i}{\sum_{i \in \mathcal{I}_{\mathrm{ca}}} W_i + \widehat{w}(x; \mathcal{Z}_{\mathrm{tr}})}$ and $\widehat{p}_i(x) = \frac{\widehat{w}(x; \mathcal{Z}_{\mathrm{tr}})}{\sum_{i \in \mathcal{I}_{\mathrm{ca}}} W_i + \widehat{w}(x; \mathcal{Z}_{\mathrm{tr}})}$
    5: Compute $\eta(x)$ as the $(1 - \alpha)$-th quantile of the distribution $\sum_{i \in \mathcal{I}_{\mathrm{ca}}} \widehat{p}_i(x) \delta_{V_i} + \widehat{p}_\infty \delta_\infty$
Output: $\widehat{C}(x) = [\widehat{q}_{\alpha_{\mathrm{lo}}}(x; \mathcal{Z}_{\mathrm{tr}}) - \eta(x), \widehat{q}_{\alpha_{\mathrm{hi}}}(x; \mathcal{Z}_{\mathrm{tr}}) + \eta(x)]$
```

Proposition 1. *Consider Algorithm 1 and assume* $(X_i, Y_i) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$. *First, consider the case where* $\widehat{w}(\cdot) = w(\cdot)$. *Then,*

$$\mathbb{P}_{(X,Y) \sim Q_X \times P_{Y|X}}(Y \in \widehat{C}(X)) \geq 1 - \alpha$$

holds without any further assumption. Further, if the non-conformity scores $\{V_i : i \in \mathcal{I}_{\mathrm{ca}}\}$ *have no ties almost surely,* Q_X *is absolutely continuous with respect to* P_X, *and* $(\mathbb{E}[\widehat{w}(X)^r])^{1/r} \leq M_r < \infty$, *then*

$$1 - \alpha \leq \mathbb{P}_{(X,Y) \sim Q_X \times P_{Y|X}}(Y \in \widehat{C}(X)) \leq 1 - \alpha + cn^{1/r-1}$$

where c *is a positive constant that only depends on* M_r *and* r.

In the general case where $\widehat{w}(\cdot) \neq w(\cdot)$, *set* $\Delta_w = (1/2)\mathbb{E}_{X \sim P_X}|\widehat{w}(X) - w(X)|$. *Then coverage is always lower bounded by* $1 - \alpha - \Delta_w$, *and upper bounded by* $1 - \alpha + \Delta_w + cn^{1/r-1}$.

Note that the propositon holds uniformly over all conditional distributions $P_{Y|X}$ and all procedures used to fit conditional quantiles. Also, it implies that weighted split-CQR has almost exact coverage when the calibration fold is large.

3.3 The role of the propensity score

Given a binary treatment, the propensity score is defined as the probability of getting treated given the covariate value:

$$e(x) = \mathbb{P}(T = 1 \mid X = x)$$

If the overlap condition $0 < e(X) < 1$ holds almost surely, ATE can be identified as follows:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[w_1(X)Y^{\mathrm{obs}}I(T=1) - w_0(X)Y^{\mathrm{obs}}I(T=0)\right]$$

where $w_1(x) = 1/e(x)$ and $w_0(x) = 1/(1-e(x))$ which are inverse probability weights (IPW). Similarly ATT can be identified with $w_1(x) = 1/\mathbb{P}(T=1)$ and $w_0(x) = e(x)/(1-e(x))\mathbb{P}(T=1)$ under the weaker overlap condition $e(X) < 1$ almost surely.

The weights in IPW estimators are similar in spirit to weighted conformal inference. In general, if Q_X is the covariate distribution in the target population,

$$w_1(x) = \frac{dQ_X(x)}{dP_{X|T=1}(x)} = \frac{dQ_X(x)}{dP_X(x)}\frac{\mathbb{P}(T=1)}{e(x)} \text{ and } w_0(x) = \frac{dQ_X(x)}{dP_{X|T=0}(x)} = \frac{dQ_X(x)}{dP_X(x)}\frac{\mathbb{P}(T=0)}{1-e(x)}$$

for conformal inference on $Y(1)$ and $Y(0)$. $Q_X = P_X$ for ATE-type conformal inference and $Q_X = P_{X|T=1}$ for ATT-type. Note that the weight is invariant up to any constant.

Table: Summary of weight functions for different inferential targets

Inferential type	ATE	ATT	ATC	General
$w_1(x)$	$1/e(x)$	1	$(1-e(x))/e(x)$	$(dQ/dP)(x)/e(x)$
$w_0(x)$	$1/(1-e(x))$	$e(x)/(1-e(x))$	1	$(dQ/dP)(x)/(1-e(x))$

The weighted conformal inference depends on propensity scores in the same way IPW estimation of average causal effects depends on these same scores.

3.4 Conformalized counterfactual inference is exact for randomized trials

For randomized trials with perfect compliance, the strong ignorability assumption is satisfied and the propensity score is known since it is designed by researchers. Weighting is not required in completely randomized experiments. One would apply weighted conformal inference for general stratified experiments such as blocking experiments. By Proposition 1, our method has almost exact coverage when the calibration fold is large and $\mathbb{E}[1/e(X)^2] < \infty$.

3.5 Conformalized counterfactual inference is doubly robust

For observational studies or randomized trials with imperfect compliance, the propensity score is unknown. Let $\widehat{e}(x)$ denote the estimate of $e(x)$.

Theorem 1. *Let $N = |\mathcal{Z}_{\mathrm{tr}}|$ and $n = |\mathcal{Z}_{\mathrm{ca}}|$. Further, let $\widehat{q}_{\beta,N}(x) = \widehat{q}_{\beta,N}(x;\mathcal{Z}_{\mathrm{tr}})$ be an estimate of the β-th conditional quantile $q_\beta(x)$ of $Y(1)$ given $X = x$, $\widehat{e}_N(x) = \widehat{e}_N(x;\mathcal{Z}_{\mathrm{tr}})$ be an estimate of $e(x)$, and $\widehat{C}_{N,n}(x)$ be the resulting interval from Algorithm 1. Assume that $\mathbb{E}[1/\widehat{e}_N(X) \mid \mathcal{Z}_{\mathrm{tr}}] < \infty$ and $\mathbb{E}[1/e(X)] < \infty$. Assume that one of the following holds:* **A1** $\lim_{N\to\infty}\mathbb{E}\left|\frac{1}{\widehat{e}_N(X)} - \frac{1}{e(X)}\right| = 0$; **A2** *(1) $\alpha_{\mathrm{hi}} - \alpha_{\mathrm{lo}} = 1-\alpha$, (2) there exists $r, b_1, b_2 > 0$ such that $\mathbb{P}(Y(1) = y \mid X = x) \in [b_1, b_2]$ uniformly over all (x,y) with $y \in [q_{\alpha_{\mathrm{lo}}}(x) - r, q_{\alpha_{\mathrm{lo}}}(x) + r] \cup [q_{\alpha_{\mathrm{hi}}}(x) - r, q_{\alpha_{\mathrm{hi}}}(x) + r]$, (3) there exists $\delta > 0$ such that*

$$\limsup_{N\to\infty}\mathbb{E}\left[\frac{1}{\widehat{e}_N(X)^{1+\delta}}\right] < \infty, \quad \lim_{N\to\infty}\mathbb{E}\left[\frac{H_N(X)}{\widehat{e}_N(X)}\right] = \lim_{N\to\infty}\mathbb{E}\left[\frac{H_N(X)}{e(X)}\right] = 0$$

where $H_N(x) = \max\{|\widehat{q}_{\alpha_{\text{lo}},N}(x) - q_{\alpha_{\text{lo}}}(x)|, |\widehat{q}_{\alpha_{\text{hi}},N}(x) - q_{\alpha_{\text{hi}}}(x)|\}$. Then under SUTVA and the strong ignorability assumption,

$$\lim_{N,n\to\infty} \mathbb{P}_{(X,Y(1))\sim P_X \times P_{Y(1)|X}}(Y(1) \in \widehat{C}_{N,n}(X)) \geq 1 - \alpha$$

Furthermore, if **A2** *holds, then for any $\epsilon > 0$,*

$$\lim_{N,n\to\infty} \mathbb{P}_{X\sim P_X} \left(\mathbb{P}(Y(1) \in \widehat{C}_{N,n}(X) \mid X) \leq 1 - \alpha - \epsilon\right) = 0$$

The property is analogous to the doubly robust point estimation of ATE, which yields consistent estimators if either the propensity score or the conditional mean of potential outcomes are consistent.

3.6 Numerical experiments

We demonstrate the performance of our methods via simulation studies. We present comparisons with three competing methods offering qualitatively different approaches to uncertainty quantification: Causal Forest, X-learner, and Bayesian Additive Regression Trees (BART). For wegithed split-CQR, we estimate the propensity score via the gradient boosting algorithm. We further estimate the conditional quantiles in three different ways: quantile random forest, gradient boosting, and BART.

We compute the empirical coverage of CATE and ITE as $(1/n_{\text{test}}) \sum_{i=1}^{n_{\text{test}}} I\left(\tau(X_i) \in \widehat{C}_{\text{ITE}}(X_i)\right)$ and $(1/n_{\text{test}}) \sum_{i=1}^{n_{\text{test}}} I\left(Y_i(1) - Y_i(0) \in \widehat{C}_{\text{ITE}}(X_i)\right)$ respectively. Note that we should distinguish the confidence intervals for CATE and prediction intervals for ITE. Causal Forest and X-learner are only guaranteed to cover CATE theoretically, whereas weighted split-CQR is only guaranteed to cover ITE. BART has guarantees to cover both CATE and ITE.

When we see the empirical marginal coverage of CATE and ITE, Causal Forest and X-learner have poor coverage in all scenarios. BART has better coverage than them, but it has poor coverage when the covariates are correlated and the errors are heteroscedastic. Our method weighted split-CQR is the only method achieving valid coverage in all scenarios.

When we see the interval lengths for ITE, Causal Forest, X-learner, and BART have short intervals. Weighted split-CQR, with BART as the learner of conditional quantiles, produces intervals that are almost as narrow as those produced with BART in homoscedastic settings, because $\eta(x) \approx 0$. In heteroscedastic settings, it has much larger variability in interval lengths because BART fails to estimate the conditional quantiles well. In all cases, we observe that the interval lengths of weighted split-CQR with gradient boosting and random forest are reasonably short.

For the conditional coverage of ITE, Causal Forest, X-learner, and BART all have decreasing conditional coverage as $\sigma^2(x)$ increases. In contrast, weighted split-CQR with quantile random forest or quantile gradient boosting maintains conditional coverage.

We emphasize that our comparisons do not carry over to point estimates of CATE. In general, the accuracy of a point estimate does not inform coverage of an interval estimate. Also, an interval estimate does not have to be derived from a point estimate.

4 From Counterfactuals To Treatment Effects

We finally turn our attention to intervals for ITE for subjects not in the study, and for which both potential outcomes are missing.

4.1 naive approach

For a testing point x, we can construct a pair of prediction intervals at level $1-\alpha/2$, $[\widehat{Y}^L(1;x), \widehat{Y}^R(1;x)]$ for $Y(1)$ and $[\widehat{Y}^L(0;x), \widehat{Y}^R(0;x)]$ for $Y(0)$. Then, we can obtain an interval for ITE at the level $1-\alpha$:

$$\widehat{C}_{\text{ITE}} = [\widehat{Y}^L(1;x) - \widehat{Y}^R(0;x),\ \widehat{Y}^R(1;x) - \widehat{Y}^L(0;x)]$$

4.2 A nested approach

We propose another strategy, "nested approach". This procedure splits the data into two folds and train $\widehat{C}_1(x)$ and $\widehat{C}_0(x)$ on the first fold by applying counterfactual inference. We can obtain an interval for ITE defined as

$$\widehat{C}_{\text{ITE}}(x;t,y^{\text{obs}}) = \begin{cases} y^{\text{obs}} - \widehat{C}_0(x) & \text{if } t=1 \\ \widehat{C}_1(x) - y^{\text{obs}} & \text{if } t=0 \end{cases}$$

Put $\widehat{C}_i = \widehat{C}_{\text{ITE}}(X_i;T_i,Y_i^{\text{obs}})$ for each i in the second fold. If

$$\mathbb{P}(Y_i(0) \in \widehat{C}_0(X_i) \mid T_i = 1) \geq 1-\alpha, \quad \mathbb{P}(Y_i(1) \in \widehat{C}_1(X_i) \mid T_i = 0) \geq 1-\alpha \tag{1}$$

then

$$\mathbb{P}\left(Y_i(1) - Y_i(0) \in \widehat{C}_i\right) \geq 1-\alpha$$

For randomized experiments with known propensity score $e(x)$, 1 is satisfied. For observational studies, we can use $\hat{e}(x)$ and 1 is satisfied approximately if the conditions of Theorem 1 hold.

The nested procedure creates an i.i.d. dataset $(X_i, \widehat{C}_i)$, conditional on the first fold. If we can fit a model $\tilde{C}_{\text{ITE}}(x)$ of $\widehat{C}_i$ on X_i, then the intervals can be generalized to subjects with both potential outcomes missing. The nested method is fitting an observed uncertainty measurement $\widehat{C}_i$, instead of estimating the unobserved uncertainty.

4.3 An inexact and an exact method under the nested framework

The function $\tilde{C}_{\text{ITE}}(x)$ can be obtained by training a model of the left- and right-end point of $\widehat{C}_i$ on X_i separately. Then, $\tilde{C}_{\text{ITE}}$ does not have same theoretical guarantee, so we refer to it as an "inexact method".

We propose a secondary conformal inference procedure on the induced dataset $(X_i, \widehat{C}_i)$. We call it as "exact method". Given a generic observation (X, T, Y^{obs}), let $C = \widehat{C}_{\text{ITE}}(X, T, Y^{\text{obs}})$ denote the induced interval.

Algorithm 2 (Unweighted) conformal inference for interval outcomes

```
Input: level $\gamma$, data $\mathcal{Z} = (X_i, C_i)_{i \in \mathcal{I}}$ where $C_i = [C_i^L, C_i^R]$, testing point $x$, functions $\widehat{m}^L(x;\mathcal{D})$, $\widehat{m}^R(x;\mathcal{D})$ to fit the conditional mean/median of $C^L$, $C^R$
Procedure:
  1: Split $\mathcal{Z}$ into a trainng fold $\mathcal{Z}_{\text{tr}} = (X_i, C_i)_{i \in \mathcal{I}_{\text{tr}}}$ and a calibration fold $\mathcal{Z}_{\text{ca}} = (X_i, C_i)_{i \in \mathcal{I}_{\text{ca}}}$
  2: For each $i \in \mathcal{I}_{\text{ca}}$, compute score $V_i = \max\{\widehat{m}^L(X_i;\mathcal{Z}_{\text{tr}}) - C_i^L, C_i^R - \widehat{m}^R(X_i;\mathcal{Z}_{\text{tr}})\}$
  3: Compute $\eta$ as the $(1-\gamma)(1+1/|\mathcal{Z}_{\text{tr}}|)$ quantile of the empirical distribution of $\{V_i : i \in \mathcal{I}_{\text{ca}}\}$
Output: $\widehat{\mathcal{C}}_{\text{ITE}}(x) = [\widehat{m}^L(X_i;\mathcal{Z}_{\text{tr}}) - \eta, \widehat{m}^R(X_i;\mathcal{Z}_{\text{tr}}) + \eta]$
```

Theorem 2. *Consider Algorithm 2 and assume* $(X_i, C_i) \overset{i.i.d.}{\sim} (X, C)$*. Then,* $\mathbb{P}\left(C \subset \widehat{\mathcal{C}}_{\text{ITE}}(X)\right) \geq 1 - \gamma$.

With this theorem,

$$\mathbb{P}(Y(1) - Y(0) \notin \widehat{\mathcal{C}}_{\text{ITE}}(X)) \leq \mathbb{P}(Y(1) - Y(0) \notin C) + \mathbb{P}(C \not\subset \widehat{\mathcal{C}}_{\text{ITE}}(X)) \leq \alpha + \gamma$$

Then, we can obtain prediction intervals for ITE with a theoretical guarantee of coverage. The following algorithm summarizes the nested approach.

Algorithm 3 Nested approach for interval estimates of ITE

```
Input: level $\alpha$, level $\gamma$ (only for exact version), data $\mathcal{Z} = (X_i, Y_i, T_i)_{i=1}^n$, testing point $x$
Step I: data splitting
  1: Split the data into two folds $\mathcal{Z}_1$ and $\mathcal{Z}_2$
  2: Estimate propensity score $\widehat{e}(x)$ on $\mathcal{Z}_1$
Step II: counterfactual inference on $\mathcal{Z}_2$
for $i$ in $\mathcal{Z}_2$ with $T_i = 1$ do
    Compute $[\widehat{Y}_i^L(0), \widehat{Y}_i^R(0)]$ in Algorithm 1 on $\mathcal{Z}_1$ with level $\alpha$ and $w_0(x) = \widehat{e}(x)/(1 - \widehat{e}(x))$
    Compute $\widehat{C}_i = [Y_i(1) - \widehat{Y}_i^R(0), Y_i(1) - \widehat{Y}_i^L(0)]$
end for
for $i$ in $\mathcal{Z}_2$ with $T_i = 0$ do
    Compute $[\widehat{Y}_i^L(1), \widehat{Y}_i^R(1)]$ in Algorithm 1 on $\mathcal{Z}_1$ with level $\alpha$ and $w_1(x) = (1 - \widehat{e}(x))/\widehat{e}(x)$
    Compute $\widehat{C}_i = [\widehat{Y}_i^L(1) - Y_i(0), \widehat{Y}_i^R(1) - Y_i(0)]$
end for
Step III: Interval of ITE on the testing point
  1: (Exact version) Apply Algorithm 2 on $(X_i, \widehat{C}_i)_{i \in \mathcal{Z}_2}$ with level $\gamma$, yielding an interval $\widehat{\mathcal{C}}_{\text{ITE}}(x)$
  2: (Inexact version) Apply Fit conditional quantiles of $\widehat{C}^L$ and $\widehat{C}^R$, yielding an interval $\widehat{\mathcal{C}}_{\text{ITE}}(x)$
Output: $\widehat{\mathcal{C}}_{\text{ITE}}(x)$
```

4.4 Empirical performance

We demonstrate the performance of our methods by generating a synthetic dataset $(X_i, T_i, Y_i(1), Y_i(0))$ based on the National Study of Learning Mindsets (NSLM). Then, we compare our method (naive, exact, inexact) with Causal Forest, X-learner, and BART (naive, inexact). Note that neither the naive nor the nested approaches are limited to CQR.

We present the coverage of ITE and the average length of intervals estimated on the testing set. Causal Forest and X-learner have poor coverage and short intervals. The naive methods with both CQR and BART are conservative, and the exact nested method with CQR is also conservative. In contrast, the inexact nested with CQR or BART are less conservative. The average length of intervals of inexact-CQR is just slightly above that of inexact-BART while significantly lower than that of either the naive or the exact nested methods. Therefore, we see that the inexact weighted split-CQR produces reasonably short intervals.

We investigate the conditional coverage as a function of the conditional variance $\sigma^2(x)$ and the CATE $\tau(x)$ respectively. We can see that inexact-CQR has desirable and relatively even conditional coverage while inexact-BART performs worse.

4.5 Re-analysing NSLM data

We apply inexact-CQR with BART as the learner to re-analyze the NSLM data. The figures show the fractions of intervals that only cover positive and negative values respectively. We see some evidence of positive ITE when α is above 0.25 while no evidence of any negative ITE even when $\alpha = 0.5$. Then, we can decide to assign treatment to a patient if the lower prediction bound of her ITE interval is positive.

5 From Potential Outcomes to Other Causal Frameworks

We proposed a method based on weighted conformal inference which produces interval estimates of counterfactuals and individual treatment effects under the potential outcome framework. For randomized experiments with perfect compliance, our method has guaranteed coverage in finite samples. For randomized experiments with ignorable compliance or observational studies under the strong ignorability assumption, our method is doubly robust in the sense that the coverage is asymptotically guaranteed. In contrast, existing methods may suffer from a significant coverage deficit even in simple models.

5.1 Causal diagram framework

A general framework of causal inference based on graphical models defines casual effects through the "do" operator, which modifies the observed distribution by removing the causal paths that directly point to the intervention variable. We discuss the case where T is the intervention variable, Y is the outcome variable and X is a set of variables that satisfies the back-door criterion. In this case, $P_{(X,Y)|\mathrm{do}(T=t)} = P_X \times P_{Y|X,T=t}$ is the target distribution to be inferred. In contrast, the observed distribution of (X,Y) given $T = t$ is $P_{(X,Y)|T=t} = P_{X|T=t} \times P_{Y|X,T=t}$. Since this has the same structure as in the potential outcome framework, the weighted split-CQR can be applied.

5.2 Invariant prediction framework

Invariant prediction is another framework which is powerful when there are multiple data sources under different interventions. Consider an outcome variable Y, a set of interventions or covariates X and an environment variable E that indicates the source of data. This framework assumes that $Y \perp\!\!\!\perp E \mid X$ while $X \mid E \sim P_X^E$. The goal is to predict the outcome under a new environment. The observed distribution of (X,Y) is $P_X^{e_j} \times P_{Y|X}$ while the target distribution to be inferred is $P_X^{e_0} \times P_{Y|X}$ where $e_1, \cdots, e_J, e_0$ are each environment. If $J = 1$, weighted split-CQR with weight function $dP_X^{e_0}(x)/dP_X^{e_1}(x)$ produces doubly robust intervals of Y under environment e_0. If $J > 1$, the weight function becomes much more complicated. An alternative approach is to create a weighted population $\left(\sum_{j=1}^J q_j P_X^{e_j}\right) \times P_{Y|X}$ and apply Algorithm 1 on this pseudo dataset with weight $w(x) = 1/\sum_{j=1}^J q_j(dP_X^{e_j}/dP_X^{e_0})(x)$.

References

Tibshirani, R. J., Barber R. F., Candès, E., & Ramdas, A. (2019). *Conformal prediction under covariate shift.* In Advances in Neural Information Processing Systems, p. 2530-2540.