

Debiased Machine Learning for Conformal Prediction of Counterfactual Outcomes Under Runtime Confounding

CLear 2026

Keith Barnatchez, Kevin P. Josey, Rachel C. Nethery, and Giovanni Parmigiani

April 7th, 2025



Collaborators



Rachel Nethery



Giovanni Parmigiani



Kevin Josey

Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

Methods

Numerical Experiments

Discussion

Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

Methods

Numerical Experiments

Discussion

Motivation: Counterfactual predictions and prediction intervals

Growing set of methods for predicting **counterfactual outcomes** $Y(a)$ under **different treatments** $a \in \mathcal{A}$, given a covariate profile X

$$\hat{E}[Y(a) | X]$$

X conditional outcome under tmt. $A=a$

Motivation: Counterfactual predictions and prediction intervals

Growing set of methods for predicting **counterfactual outcomes** $Y(a)$ under **different treatments** $a \in A$, given a covariate profile X

$$\hat{E}[Y(a) | X]$$

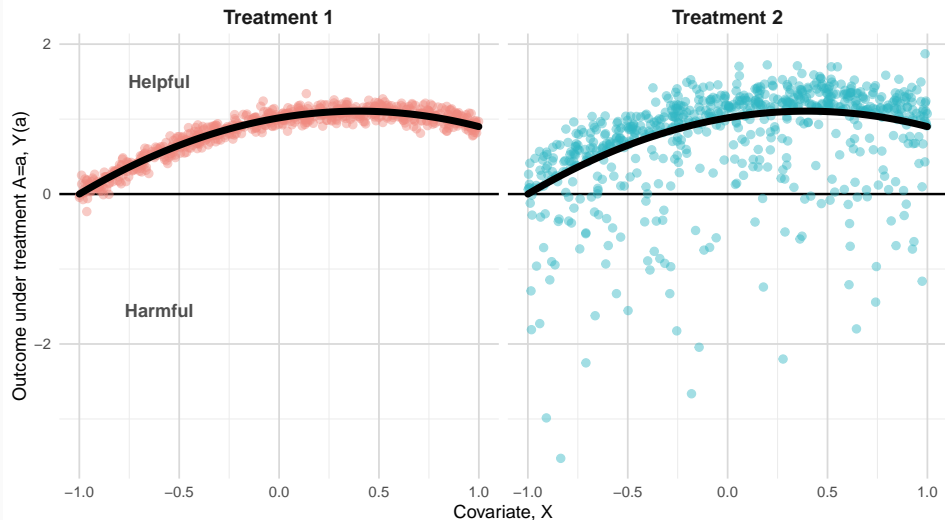
X conditional outcome under tmt. $A=a$

Along with point predictions, prediction intervals can serve as **key inputs** for decision making

Counterfactual predictions vs. prediction intervals

Two treatments with the same conditional mean response...

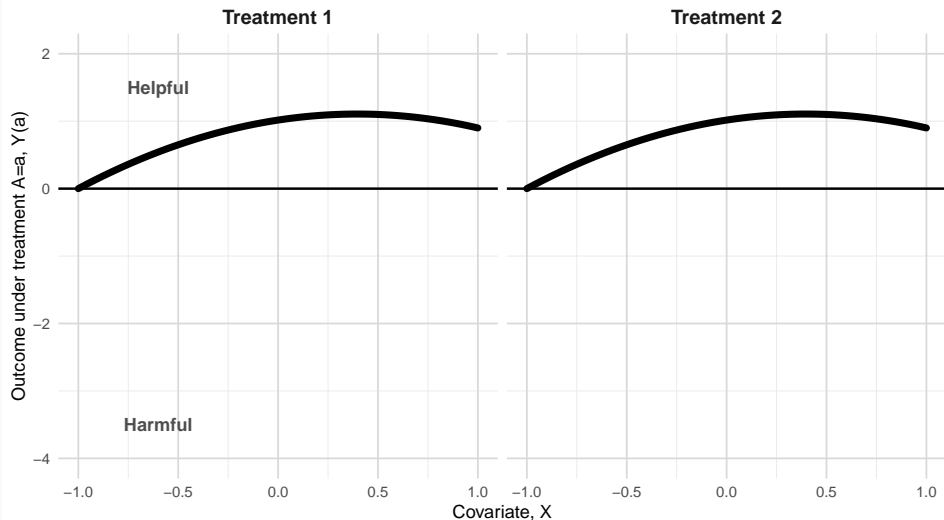
...but different conditional response distributions



Counterfactual predictions vs. prediction intervals

Two treatments with the same conditional mean response...

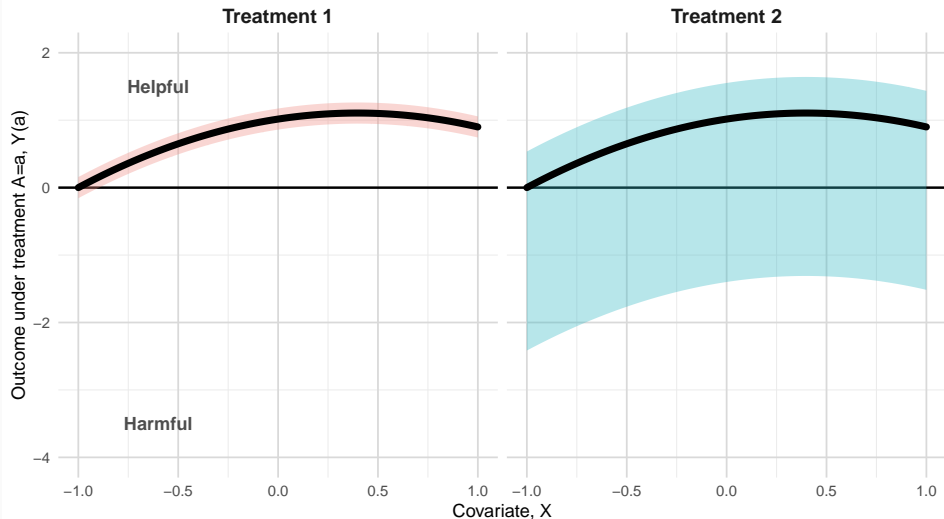
...but different conditional response distributions



Counterfactual predictions vs. prediction intervals

Two treatments with the same conditional mean response...

...but different conditional response distributions



Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

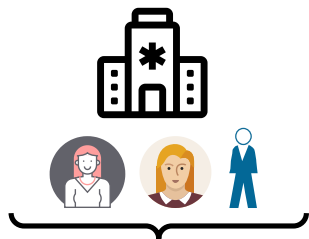
Methods

Numerical Experiments

Discussion



$$(Y_i, A_i, \mathbf{X}_i)_{i=1}^{n_1}$$



$$(\mathbf{V}_i)_{i=1}^{n_0}$$

where $\mathbf{V} \subset \mathbf{X}$



Goal: use info from source site to
inform decisions in target site



$$(Y_i, A_i, \mathbf{X}_i)_{i=1}^{n_1}$$



$$(\mathbf{V}_i)_{i=1}^{n_0}$$

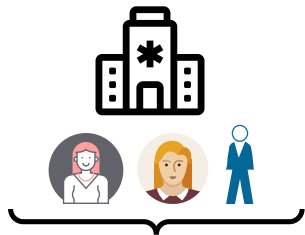
where $\mathbf{V} \subset \mathbf{X}$



$$(Y_i, A_i, \mathbf{X}_i)_{i=1}^{n_1}$$

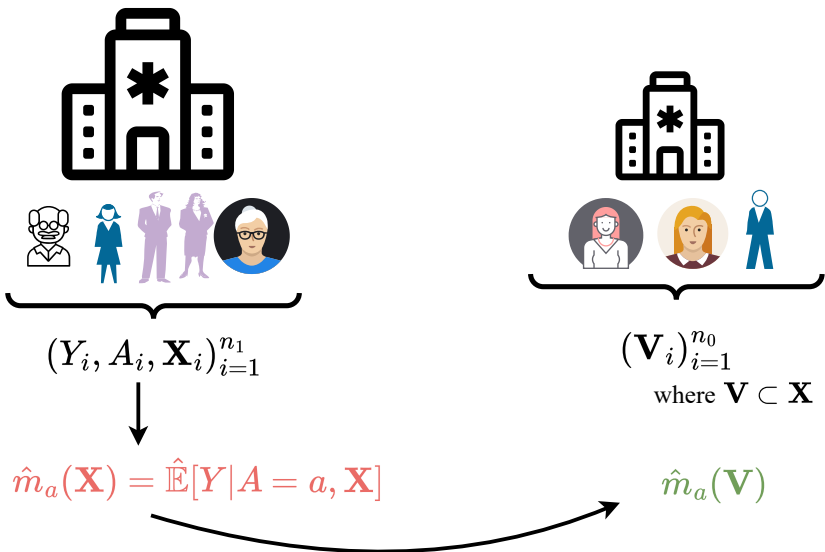


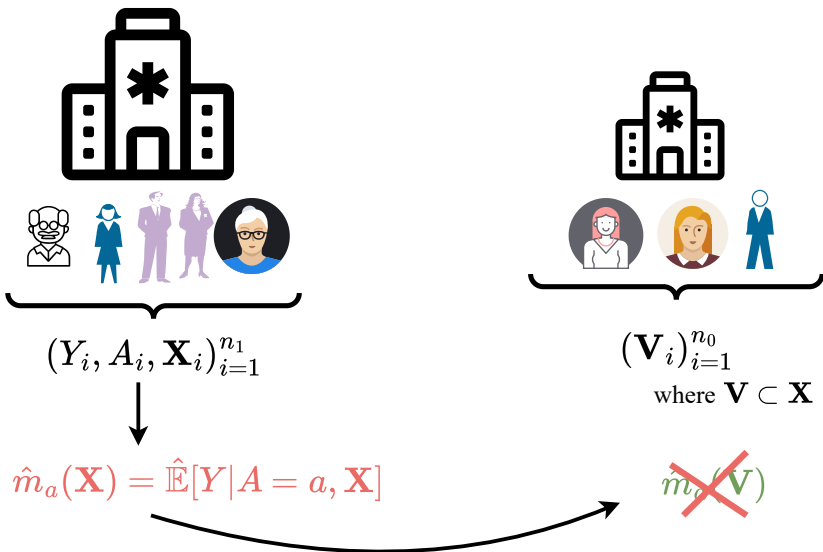
$$\hat{m}_a(\mathbf{X}) = \hat{\mathbb{E}}[Y|A = a, \mathbf{X}]$$



$$(\mathbf{V}_i)_{i=1}^{n_0}$$

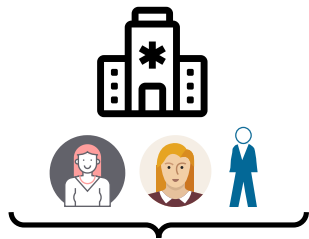
where $\mathbf{V} \subset \mathbf{X}$







$$(Y_i, A_i, \mathbf{X}_i)_{i=1}^{n_1}$$



$$(\mathbf{V}_i)_{i=1}^{n_0}$$

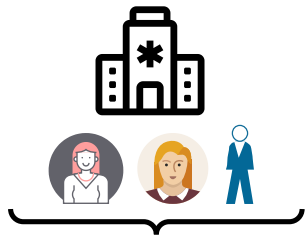
where $\mathbf{V} \subset \mathbf{X}$



$$(Y_i, A_i, \mathbf{X}_i)_{i=1}^{n_1}$$

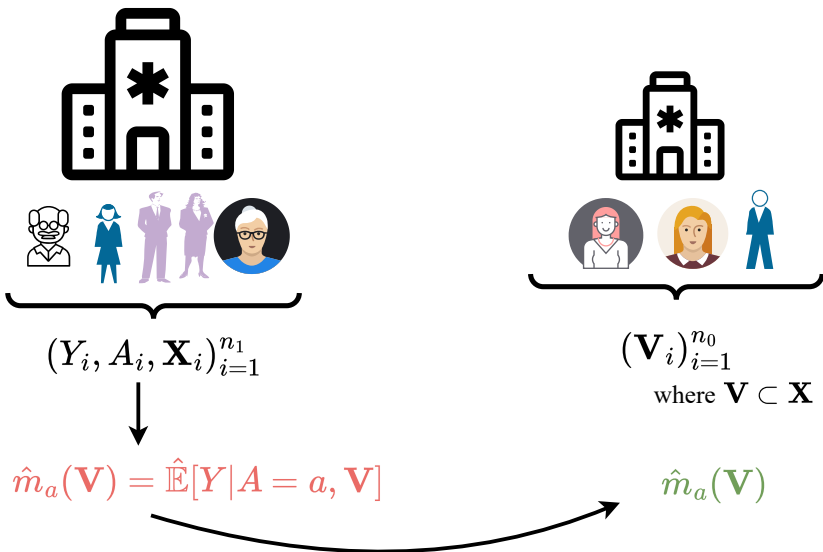


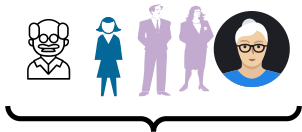
$$\hat{m}_a(\mathbf{V}) = \hat{\mathbb{E}}[Y | A = a, \mathbf{V}]$$



$$(\mathbf{V}_i)_{i=1}^{n_0}$$

where $\mathbf{V} \subset \mathbf{X}$



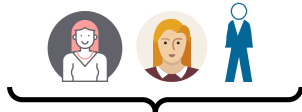


$$(Y_i, A_i, \mathbf{X}_i)_{i=1}^{n_1}$$



$$\hat{m}_a(\mathbf{V}) = \hat{\mathbb{E}}[Y|A = a, \mathbf{V}]$$

\mathbf{V} may not contain all
confounders

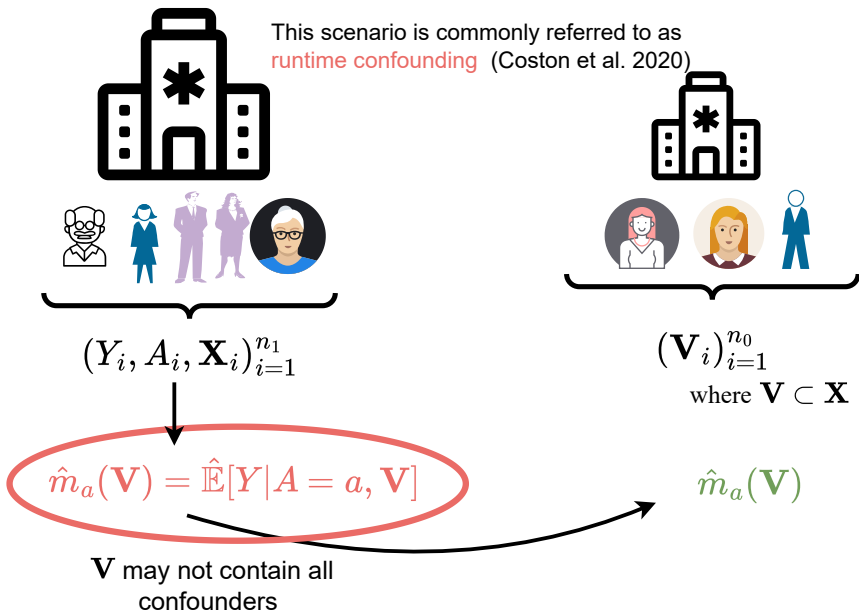


$$(\mathbf{V}_i)_{i=1}^{n_0}$$

where $\mathbf{V} \subset \mathbf{X}$

$$\hat{m}_a(\mathbf{V})$$

This scenario is commonly referred to as **runtime confounding** (Coston et al. 2020)



This paper...

- Proposes methods for constructing **prediction intervals** for counterfactual outcomes $Y(a)$ under runtime confounding
- We construct these intervals in a **model-agnostic** way that
 - Ensures their validity asymptotically, and
 - Relies on less data to obtain \approx valid coverage than commonly-taken approaches

Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

Methods

Numerical Experiments

Discussion

Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

Methods

Numerical Experiments

Discussion

Problem setting

Suppose we observe...

- $(Y_i; A_i; \mathbf{X}_i); i = 1; \dots; n_1$, from a source population
- $V_i; i = 1; \dots; n_0$ from a target population, where V is a subset of all the covariates collected in \mathbf{X} , where $\mathbf{X} = (V; U)$

Problem setting

Suppose we observe...

- $(Y_i; A_i; \mathbf{X}_i); i = 1; \dots; n_1$, from a source population
- $V_i; i = 1; \dots; n_0$ from a target population, where V is a subset of all the covariates collected in \mathbf{X} , where $\mathbf{X} = (V; U)$

Interest lies in constructing prediction intervals for counterfactual outcomes in the target population which attain a desired marginal coverage probability of $1 - \alpha$:

$$P(Y_i(a) \in \hat{C}_a(V_i) | \text{In target pop.}) = 1 - \alpha; \quad a \in A$$

Problem setting

Suppose we observe...

- $(Y_i; A_i; \mathbf{X}_i); i = 1; \dots; n_1$, from a source population
- $V_i; i = 1; \dots; n_0$ from a target population, where V is a subset of all the covariates collected in \mathbf{X} , where $\mathbf{X} = (V; U)$

Interest lies in constructing prediction intervals for counterfactual outcomes in the target population which attain a desired marginal coverage probability of 1 :

$$P(Y_i(a) \in \hat{C}_a(V_i) | \text{In target pop. } g = 1) = 1 \quad ; \quad a \in A$$

Y	A	V	U
Y_1	A_1	V_1	U_1
\vdots	\vdots	\vdots	\vdots
Y_{n_1}	A_{n_1}	V_{n_1}	U_{n_1}

Source sample

Y	A	V	U
NA	NA	V_{n_1+1}	NA
\vdots	\vdots	\vdots	\vdots
NA	NA	$V_{n_1+n_0}$	NA

Target sample

Problem setting

Suppose we observe...

- $(Y_i; A_i; \mathbf{X}_i); i = 1; \dots; n_1$, from a source population
- $\mathbf{V}_i; i = 1; \dots; n_0$ from a target population, where \mathbf{V} is a subset of all the covariates collected in \mathbf{X} , where $\mathbf{X} = (\mathbf{V}; \mathbf{U})$

Interest lies in constructing prediction intervals for counterfactual outcomes in the target population which attain a desired marginal coverage probability of 1 :

$$P(Y_i(a) \in \hat{C}_a(\mathbf{V}_i) | S = 0) = 1 \quad ; \quad a \in A$$

Y	A	\mathbf{V}	\mathbf{U}	S
Y_1	A_1	\mathbf{V}_1	\mathbf{U}_1	1
\vdots	\vdots	\vdots	\vdots	\vdots
Y_{n_1}	A_{n_1}	\mathbf{V}_{n_1}	\mathbf{U}_{n_1}	1
NA	NA	\mathbf{V}_{n_1+1}	NA	0
\vdots	\vdots	\vdots	\vdots	\vdots
NA	NA	$\mathbf{V}_{n_1+n_0}$	NA	0

Problem setting

Suppose we observe...

- $(Y_i; A_i; \mathbf{X}_i); i = 1; \dots; n_1$, from a source population
- $\mathbf{V}_i; i = 1; \dots; n_0$ from a target population, where \mathbf{V} is a subset of all the covariates collected in \mathbf{X} , where $\mathbf{X} = (\mathbf{V}; \mathbf{U})$

Interest lies in constructing prediction intervals for counterfactual outcomes in the target population which attain a desired marginal coverage probability of 1 - α :

$$P\{Y_i(a) \in \hat{C}_a(\mathbf{V}_i) \mid S = 0, g = 1\} = 1 - \alpha; \quad a \in A$$

This is an example of **runtime confounding**: the covariates we have access to at test time differ from those available to us at training time (Costin et al. 2020)

Counterfactual prediction intervals

Y	A	V	U	S
Y_1	A_1	V_1	U_1	1
\vdots	\vdots	\vdots	\vdots	\vdots
Y_{n_1}	A_{n_1}	V_{n_1}	U_{n_1}	1
NA	NA	V_{n_1+1}	NA	0
\vdots	\vdots	\vdots	\vdots	\vdots
NA	NA	$V_{n_1+n_0}$	NA	0

Fit model for $Y(a)$

$\hat{C}_a(V_{n_1+1})$

Interval for $Y(a)$

Assumptions

1. **(Positivity in the source study):** $0 < P(A = a | \mathbf{X}; S = 1) < 1$
2. **(Consistency):** $Y = \sum_{a \in \mathcal{A}} I(A = a) Y(a)$
3. **(Positivity of source membership):** $0 < P(S = 1 | \mathbf{V}) < 1$
4. **Runtime confounding in source:** $Y(a) \not\perp\!\!\!\perp A | \mathbf{X}; S = 1$, but $Y(a) \perp\!\!\!\perp A | \mathbf{V}; S = 1$
5. **Source exchangeability:** $Y(a) \perp\!\!\!\perp S | \mathbf{V}$

1. **(Positivity in the source study):** $0 < P(A = a | X; S = 1) < 1$
2. **(Consistency):** $Y = \sum_{a \in A} I(A = a) Y(a)$
3. **(Positivity of source membership):** $0 < P(S = 1 | V) < 1$
4. **Runtime confounding in source:** $Y(a) \not\perp\!\!\!\perp A | X; S = 1$, but $Y(a) \perp\!\!\!\perp A | V; S = 1$
5. **Source exchangeability:** $Y(a) \perp\!\!\!\perp S | V$

Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

Methods

Numerical Experiments

Discussion

Conformal prediction (Vovk et al., 2005) is a general framework that

- Takes an **arbitrary prediction function** $\hat{f}(\mathbf{X})$ for an outcome Y , and
- Outputs a prediction interval function $\hat{C}(\mathbf{X})$ for new observations such that

$$P(Y \in \hat{C}(\mathbf{X})) \geq 1 - \alpha$$

Conformal prediction: a tool for constructing valid prediction intervals

Conformal prediction (Vovk et al., 2005) is a general framework that

- Takes an **arbitrary prediction function** $\hat{f}(\mathbf{X})$ for an outcome Y , and
- Outputs a prediction interval function $\hat{C}(\mathbf{X})$ for new observations such that

$$P(Y \in \hat{C}(\mathbf{X})) \geq 1 - \alpha$$

When data are i.i.d. or exchangeable, above guarantee holds in **finite samples**

Conformal prediction: a tool for constructing valid prediction intervals

Conformal prediction (Vovk et al., 2005) is a general framework that

- Takes an **arbitrary prediction function** $\hat{f}(\mathbf{X})$ for an outcome Y , and
- Outputs a prediction interval function $\hat{C}(\mathbf{X})$ for new observations such that

$$P(Y \in \hat{C}(\mathbf{X})) \geq 1 - \alpha$$

When data are i.i.d. or exchangeable, above guarantee holds in **nite samples**

In broader settings (e.g. covariate shift between train + test data), nite sample coverage is generally **not guaranteed** (Bhattacharyya and Barber, 2026)

- The source of distribution shift needs to be estimated

The key ingredient in conformal prediction

The key idea behind conformal prediction (no covariate shift, everything i.i.d.) is to

1. Obtain absolute residuals $R(\mathbf{X}_i; Y_i) = |Y_i - \hat{f}(\mathbf{X}_i)|$ (example of a [conformity score](#))
2. Let $\hat{C}(\mathbf{X}_i) = \hat{f}(\mathbf{X}_i) \pm r$, where r is the 1 - α quantile of the abs. residuals
3. Notice this will imply for a new observation that $P(Y \in \hat{C}(\mathbf{X})) = 1 - \alpha$

The key ingredient in conformal prediction

The key idea behind conformal prediction (no covariate shift, everything i.i.d.) is to

1. Obtain absolute residuals $R(\mathbf{X}_i; Y_i) = |Y_i - \hat{f}(\mathbf{X}_i)|$ (example of a [conformity score](#))
2. Let $\hat{C}(\mathbf{X}_i) = \hat{f}(\mathbf{X}_i) + r$, where r is the 1 - α quantile of the abs. residuals
3. Notice this will imply for a new observation that $P(Y \in \hat{C}(\mathbf{X})) = 1 - \alpha$

This is because

$$P(Y \in \hat{C}(\mathbf{X})) = P(\text{The residual } R_i \leq r) = P(\text{The residual } R_i \text{ falls below its } (1 - \alpha)\text{-quantile})$$

Notice the probability of the last event is $1 - \alpha$

The key ingredient in conformal prediction

The key idea behind conformal prediction (no covariate shift, everything i.i.d.) is to

1. Obtain absolute residuals $R(\mathbf{X}_i; Y_i) = |Y_i - \hat{f}(\mathbf{X}_i)|$ (example of a [conformity score](#))
2. Let $\hat{C}(\mathbf{X}_i) = \hat{f}(\mathbf{X}_i) \pm r$, where r is the 1 - α quantile of the abs. residuals
3. Notice this will imply for a new observation that $P(Y \in \hat{C}(\mathbf{X})) = 1 - \alpha$

The key ingredient in conformal prediction

The key idea behind conformal prediction (no covariate shift, everything i.i.d.) is to

1. Obtain absolute residuals $R(\mathbf{X}_i; Y_i) = |Y_i - \hat{f}(\mathbf{X}_i)|$ (example of a **conformity score**)
2. Let $\hat{C}(\mathbf{X}_i) = \hat{f}(\mathbf{X}_i) \pm r$, where r is the $1 - \alpha$ quantile of the abs. residuals
3. Notice this will imply for a new observation that $P(Y \in \hat{C}(\mathbf{X})) = 1 - \alpha$

For our **problem of interest**, this means we seek an interval $\hat{C}_a(\mathbf{V}) = \hat{f}_a(\mathbf{V}) \pm r_a$ such that

$$P(R(\mathbf{V}_i; Y_i(a)) \leq r_a | S = 0) = 1 - \alpha$$

The key ingredient in conformal prediction

The key idea behind conformal prediction (no covariate shift, everything i.i.d.) is to

1. Obtain absolute residuals $R(\mathbf{X}_i; Y_i) = |Y_i - \hat{f}(\mathbf{X}_i)|$ (example of a [conformity score](#))
2. Let $\hat{C}(\mathbf{X}_i) = \hat{f}(\mathbf{X}_i) \pm r$, where r is the 1 - α quantile of the abs. residuals
3. Notice this will imply for a new observation that $P(Y \in \hat{C}(\mathbf{X})) = 1 - \alpha$

For our [problem of interest](#), this means we seek an interval $\hat{C}_a(\mathbf{V}) = \hat{f}_a(\mathbf{V}) \pm r_a$ such that

$$P(R(\mathbf{V}_i; Y_i(a)) \leq r_a | S = 0) = 1 - \alpha$$

Key problem: We don't observe outcomes, treatments, or all of \mathbf{X} in the target pop, so we can't directly estimate r_a for our target population

- No way to compute residuals R_i , so there's no way to directly estimate r_a ;

Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

Methods

Numerical Experiments

Discussion

Quick aside: forming counterfactual predictions

While we remain **agnostic** about how predictions for $Y(a)$ are formed, it's useful to think about how to construct valid point predictions under runtime confounding

Quick aside: forming counterfactual predictions

While we remain **agnostic** about how predictions for $Y(a)$ are formed, it's useful to think about how to construct valid point predictions under runtime confounding

Coston et al. (2020) provide useful framework

1. First, $\hat{m}_a(\mathbf{X}) = \hat{E}(Y|A = a; \mathbf{X})$ with the training data
2. Next, regress the estimated $\hat{m}_a(\mathbf{X})$ on \mathbf{V} , yielding $\hat{a}(\mathbf{V}) = \hat{E}(\hat{m}_a(\mathbf{X})|\mathbf{V})$

Quick aside: forming counterfactual predictions

While we remain **agnostic** about how predictions for $Y(a)$ are formed, it's useful to think about how to construct valid point predictions under runtime confounding

Coston et al. (2020) provide useful framework

1. First, fit $\hat{m}_a(\mathbf{X}) = \hat{E}(Y|A = a; \mathbf{X})$ with the training data
2. Next, regress the estimated $\hat{m}_a(\mathbf{X})$ on \mathbf{V} , yielding $\hat{^a}(\mathbf{V}) = \hat{E}(\hat{m}_a(\mathbf{X})|\mathbf{V})$

Approach is motivated by the observation that

$$\begin{aligned} E(Y(a)|\mathbf{V}) &= E[E(Y(a)|\mathbf{V}, \mathbf{U})|\mathbf{V}] = E(E(Y|A = a; \mathbf{X})|\mathbf{V}) \\ &= E(m_a(\mathbf{X})|\mathbf{V}) \end{aligned}$$

Interval construction relies on estimation of r_a

Suppose we've fit a prediction model $\hat{\mu}_a(\mathbf{V})$ for $Y(a)$, let $R_a = Y(a) - \hat{\mu}_a(\mathbf{V})$, and recall our interest in the equation

$$P(R_a(\mathbf{V}_i; Y_i) - r_a \leq jS = 0) = 1$$

Interval construction relies on estimation of r_a

Suppose we've fit a prediction model $\hat{Y}_a(\mathbf{V})$ for $Y(a)$, let $R_a = Y(a) - \hat{Y}_a(\mathbf{V})$, and recall our interest in the equation

$$P(R_a(\mathbf{V}_i; Y_i) - r_a \leq 0) = 1$$

Notice r_a is the key quantity above { if we knew it, we could construct valid intervals for $Y(a)$ in the target population of the form $\hat{Y}_a(\mathbf{V}) - r_a$.

Interval construction relies on estimation of $r_{a;}$

Suppose we've fit a prediction model $\hat{a}_a(\mathbf{V})$ for $Y(a)$, let $R_a = jY(a) - \hat{a}_a(\mathbf{V})j$, and recall our interest in the equation

$$P(R_a(\mathbf{V}_i; Y_i) - r_{a;} \leq 0) = 1$$

Notice $r_{a;}$ is the key quantity above { if we knew it, we could construct valid intervals for $Y(a)$ in the target population of the form $\hat{a}_a(\mathbf{V}) \pm r_{a;}$,

- Our paper focuses on estimation of $r_{a;}$, with particular attention given to [semiparametric efficient estimation](#), following Yang et al. (2024)

IPW identification of r_a

Continuing to let r_a be the quantity which satisfies

$$P(R(V_i; Y_i(a)) \mid r_a; jS = 0) = 1$$

It can be shown that this same quantity *also* satisfies

$$E \left[\frac{P(S = 0 \mid V)}{P(A = a \mid X; S = 1)P(S = 1 \mid V)} \mid \underbrace{Z}_{\text{weights}} \right] \underbrace{SI(A = a) \mid f_{R_a}(V; Y, r_a; g)}_{\text{source pop. conformity scores}} = 1$$

IPW identification of r_a

Continuing to let r_a be the quantity which satisfies

$$P(R(V_i; Y_i(a)) \leq r_a | S = 0) = 1$$

It can be shown that this same quantity *also* satisfies

$$E \left[\frac{P(S=0 | V)}{P(A=a | X; S=1)P(S=1 | V)} \cdot \underbrace{SI(A=a) | f_{R_a(V; Y)}(r_a; g)}_{\substack{\text{source pop.} \\ \text{conformity scores}}} \right] = 1$$

weights

Intuition: (1) Take the conformity scores that you compute among those with $A = a$ in the source pop, and (2) find the 1st quantile of its reweighted distribution

Regression-based identification of r_a

Additionally, it can be shown r_a satisfies

$$E[m_a(r_a; \mathbf{V}) | S = 0] = 1;$$

where for generic r ,

$$q_a(r; \mathbf{X}) := P(R_a(\mathbf{V}; Y) \leq r | \mathbf{X}; A = a; S = 1);$$

$$m_a(r; \mathbf{V}) := E[q_a(r; \mathbf{X}) | \mathbf{V}; S = 1];$$

Regression-based identification of r_a

Additionally, it can be shown r_a satisfies

$$E[m_a(r_a; \mathbf{V}) | S = 0] = 1;$$

where for generic r ,

$$q_a(r; \mathbf{X}) := P(R_a(\mathbf{V}; Y) \leq r | \mathbf{X}; A = a; S = 1);$$

$$m_a(r; \mathbf{V}) := E[q_a(r; \mathbf{X}) | \mathbf{V}; S = 1];$$

Intuition:

1. For an arbitrary r , regress the \mathbf{X} -conditional conformity score source CDF $q_a(r; \mathbf{X})$...
2. ...on \mathbf{V} in the source population, yielding $m_a(r; \mathbf{V})$

r_a satisfies $E[m_a(r_a; \mathbf{V}) | S = 0] = 1$

Stepping back

- Remember that **our goal** is to construct intervals for $Y(a)$ in the target population so that

$$P(Y(a) \text{ is in the interval } j \text{ } \underbrace{\{ \bar{z}_j \}}_{\text{in the target pop.}}) = 1$$

Stepping back

- Remember that **our goal** is to construct intervals for $Y(a)$ in the target population so that

$$P(Y(a) \text{ is in the interval } j \text{ } \underbrace{\{ \bar{z}_j \}}_{\text{in the target pop.}}) = 1$$

- Conformal prediction** gives us a blueprint for constructing these intervals, while being agnostic to how we predict $Y(a)$

Stepping back

- Remember that **our goal** is to construct intervals for $Y(a)$ in the target population so that

$$P(Y(a) \text{ is in the interval } j \left. \begin{array}{c} \{ \underline{Z} \} \\ \text{in the target pop.} \end{array} \right) = 1$$

- Conformal prediction** gives us a blueprint for constructing these intervals, while being agnostic to how we predict $Y(a)$
- Once we have a prediction model for $Y(a)$ (takes in covariates $\mathbf{V} \in \mathcal{X}$, outputs $\hat{Y}(a)$), constructing valid intervals amounts to finding r_a satisfying

$$P\left(\underbrace{R(\mathbf{V}; Y(a))}_{\text{scores}} \left. \begin{array}{c} \{ \underline{Z} \} \end{array} \right) \leq r_a \mid S = 0\right) = 1$$

Stepping back

- Remember that **our goal** is to construct intervals for $Y(a)$ in the target population so that

$$P(Y(a) \text{ is in the interval } j \left. \begin{array}{c} \{ \underline{Z} \} \\ \text{in the target pop.} \end{array} \right) = 1$$

- Conformal prediction** gives us a blueprint for constructing these intervals, while being agnostic to how we predict $Y(a)$
- Once we have a prediction model for $Y(a)$ (takes in covariates $\mathbf{V} \in \mathcal{X}$, outputs $\hat{Y}(a)$), constructing valid intervals amounts to finding $r_{a; j}$ satisfying

$$P\left(\left. \begin{array}{c} R(\mathbf{V}; Y(a)) \\ \text{scores} \end{array} \right\{ \underline{Z} \} \right) r_{a; j} \leq 0 = 1$$

- So our goal boils down to estimating $r_{a; j}$ well

Slow convergence of the IPW plug-in estimator

The earlier weighting expression suggests we can estimate r_a via the estimating equation

$$\frac{1}{n} \sum_{i=1}^n \frac{\hat{P}(S_i = 0 | V_i)}{\hat{P}(A_i = a | \mathbf{X}_i; S_i = 1) \hat{P}(S_i = 1 | V_i)} \underbrace{S_i I(A_i = a) f_{R_a}(Y_i; V_i)}_{\text{source pop scores}} r_a - g \quad (1) = 0$$

Slow convergence of the IPW plug-in estimator

The earlier weighting expression suggests we can estimate r_a via the estimating equation

$$\frac{1}{n} \sum_{i=1}^n \frac{\hat{P}(S_i = 0 | V_i)}{\hat{P}(A_i = a | X_i; S_i = 1) \hat{P}(S_i = 1 | V_i)} \underbrace{S_i I(A_i = a) f R_a(Y_i; V_i)}_{\text{source pop scores}} r_a g \quad (1) = 0$$

But this suggests the rate at which we estimate r_a will be dictated by

$$\hat{r}_a - r_a = O_p\left(\frac{1}{\sqrt{n}} + \sqrt{\hat{g}_a + \hat{g}^j}\right)$$

where $\hat{g}_a(\mathbf{X}) = \hat{P}(A = a | S = 1; \mathbf{X})$ and $\hat{g}^j(\mathbf{V}) = \hat{P}(S = 1 | \mathbf{V})$

Slow convergence of the IPW plug-in estimator

The earlier weighting expression suggests we can estimate r_a via the estimating equation

$$\frac{1}{n} \sum_{i=1}^n \frac{\hat{P}(S_i = 0 | V_i)}{\hat{P}(A_i = a | X_i; S_i = 1) \hat{P}(S_i = 1 | V_i)} \underbrace{S_i I(A_i = a) f R_a(Y_i; V_i)}_{\text{source pop scores}} r_a g \quad (1) = 0$$

But this suggests the rate at which we estimate r_a will be dictated by

$$\hat{r}_a - r_a = O_P(1/\sqrt{n} + \sqrt{g_a} + \sqrt{g_a} + \sqrt{g_a})$$

where $\hat{g}_a(\mathbf{X}) = \hat{P}(A = a | S = 1; \mathbf{X})$ and $\hat{g}(\mathbf{V}) = \hat{P}(S = 1 | \mathbf{V})$

If we use exible models, this can be **notably slower** than ideal $1/\sqrt{n}$ rates

Debiased estimator

Much like in ATE estimation, we can improve estimation of $r_{a;}$ by basing estimation on its EIF. Letting \mathcal{F} collect all nuisance functions, we can choose $\hat{r}_{a;}$ so

$$\frac{1}{n} \sum_{i=1}^n \text{EIF}(\hat{r}_{a;}; \text{data}_i; \hat{\mathcal{F}}_{a;}) = 0$$

Debiased estimator

Much like in ATE estimation, we can improve estimation of $r_{a;}$ by basing estimation on its EIF. Letting $\hat{r}_{a;}$ collect all nuisance functions, we can choose $\hat{r}_{a;}$ so

$$\frac{1}{n} \sum_{i=1}^n \text{EIF}(\hat{r}_{a;}; \text{data}_i; \hat{r}_{a;}) = 0$$

Why? The population-level EIF for the true $r_{a;}$ is **mean zero**

Debiased estimator

Much like in ATE estimation, we can improve estimation of $r_{a;}$ by basing estimation on its EIF. Letting $\hat{r}_{a;}$ collect all nuisance functions, we can choose $\hat{r}_{a;}$ so

$$\frac{1}{n} \sum_{i=1}^n \text{EIF}(\hat{r}_{a;}; \text{data}_i; \hat{r}_{a;}) = 0$$

Why? The population-level EIF for the true $r_{a;}$ is **mean zero**

$$\begin{aligned} \text{EIF}(r_{a;}; \text{data}_i; (r_{a;})) &= (1 - S) f m(r_{a;}; \mathbf{X}) - (1 - S) g \\ &+ \frac{(1 - S)(V)}{(V)} S f q_a(r_{a;}; \mathbf{X}) - m_a(r_{a;}; V) g \\ &+ \frac{(1 - S)(V)}{g_a(\mathbf{X})(V)} \mathbb{1}(A = a) S f \mathbb{1}(R_a(Y; V) = r_{a;}) - q_a(r_{a;}; \mathbf{X}) g \end{aligned}$$

Debiased estimator

Much like in ATE estimation, we can improve estimation of r_a by basing estimation on its EIF. Letting \mathcal{F} collect all nuisance functions, we can choose \hat{r}_a so

$$\frac{1}{n} \sum_{i=1}^n \text{EIF}(\hat{r}_a; \text{data}_i; \hat{\mathcal{F}}) = 0$$

Why? The population-level EIF for the true r_a is **mean zero**

$$\begin{aligned} \text{EIF}(r_a; \text{data}_i; (r_a; \mathcal{F})) &= (1 - S) f m(r_a; \mathbf{X}) - (1 - S) g \\ &+ \frac{(1 - S) (V)}{(V)} S f q_a(r_a; \mathbf{X}) - m_a(r_a; \mathbf{V}) g \\ &+ \frac{(1 - S) (V)}{g_a(\mathbf{X}) (V)} \mathbb{1}(A = a) S f \mathbb{1}(R_a(Y; \mathbf{V}) = r_a) - q_a(r_a; \mathbf{X}) g \end{aligned}$$

Challenge: Solving above estimating equation requires repeatedly estimating $q_a(r; \mathbf{X})$ and $m_a(r; \mathbf{V})$ at different r

Debiased estimator

Much like in ATE estimation, we can improve estimation of r_a by basing estimation on its EIF. Letting \mathcal{F} collect all nuisance functions, we can choose \hat{r}_a so

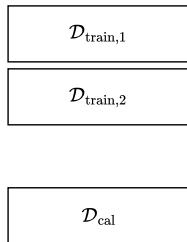
$$\frac{1}{n} \sum_{i=1}^n \text{EIF}(\hat{r}_a; \text{data}_i; \hat{\mathcal{F}}) = 0$$

Why? The population-level EIF for the true r_a is **mean zero**

$$\begin{aligned} \text{EIF}(r_a; \text{data}_i; (r_a; \mathcal{F})) &= (1 - S) f m(r_a; \mathbf{X}) - (1 - S) g \\ &+ \frac{(1 - S) (V)}{(V)} S f q_a(r_a; \mathbf{X}) - m_a(r_a; \mathbf{V}) g \\ &+ \frac{(1 - S) (V)}{g_a(\mathbf{X}) (V)} \mathbb{1}(A = a) S f \mathbb{1}(R_a(Y; \mathbf{V}) = r_a) - q_a(r_a; \mathbf{X}) g \end{aligned}$$

Can avoid estimating $q_a(r; \mathbf{X})$ and $m_a(r; \mathbf{V})$ at different r by adopting the localized DML framework from Kallus et al. (2024)

Localized DML Implementation



To avoid estimating $q_a(r; \mathbf{X})$ and $m_a(r; \mathbf{V})$ at different r , we adopt the localized DML framework from Kallus et al. (2024)

Localized DML Implementation



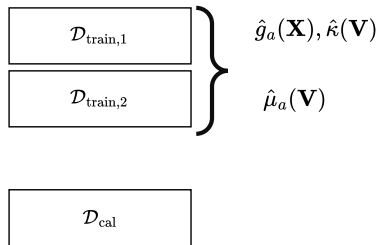
$\mathcal{D}_{\text{train},1}$

$\mathcal{D}_{\text{train},2}$

\mathcal{D}_{cal}

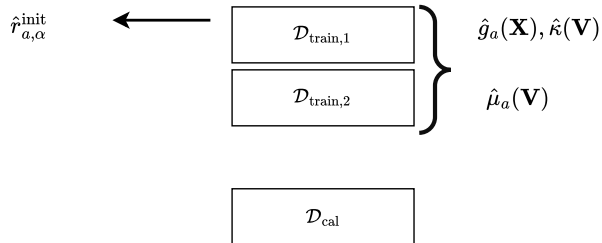
Idea: split data into training and calibration folds

Localized DML Implementation



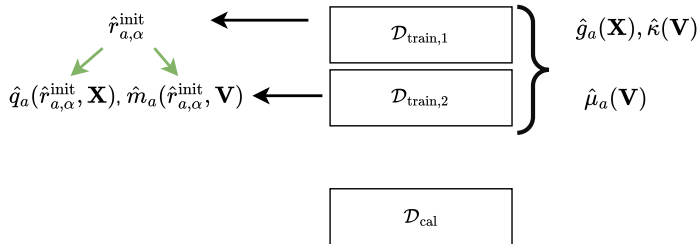
Using all of the training data, t nuisances that don't depend on r

Localized DML Implementation



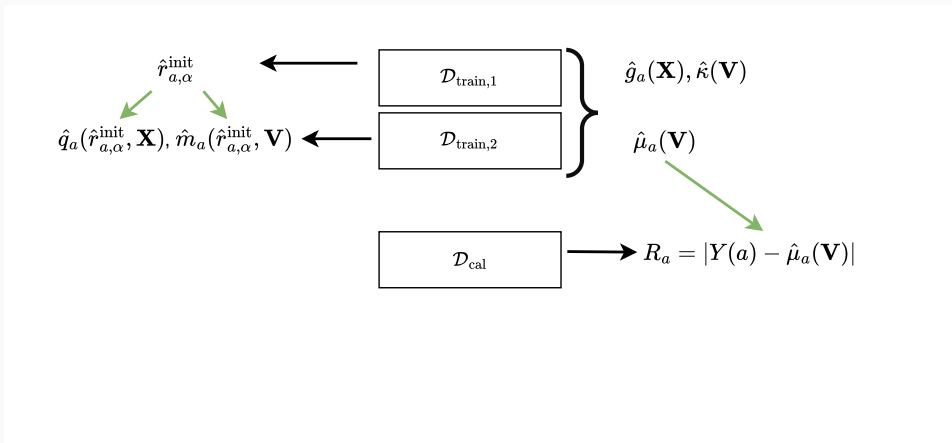
Obtain an initial estimate $\hat{r}_{a,\alpha}^{\text{init}}$ on first fold of the training data, $\mathcal{D}_{\text{train},1}$

Localized DML Implementation



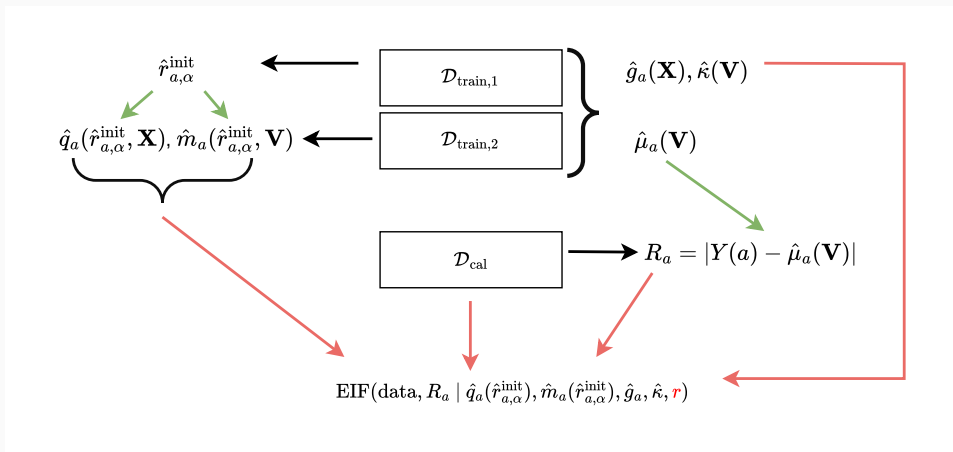
Train $\hat{q}_a(r; \mathbf{X})$ and $\hat{m}_a(r; \mathbf{V})$ in other training fold, $\mathcal{D}_{\text{train},2}$ at $\hat{r}_{a;}$

Localized DML Implementation



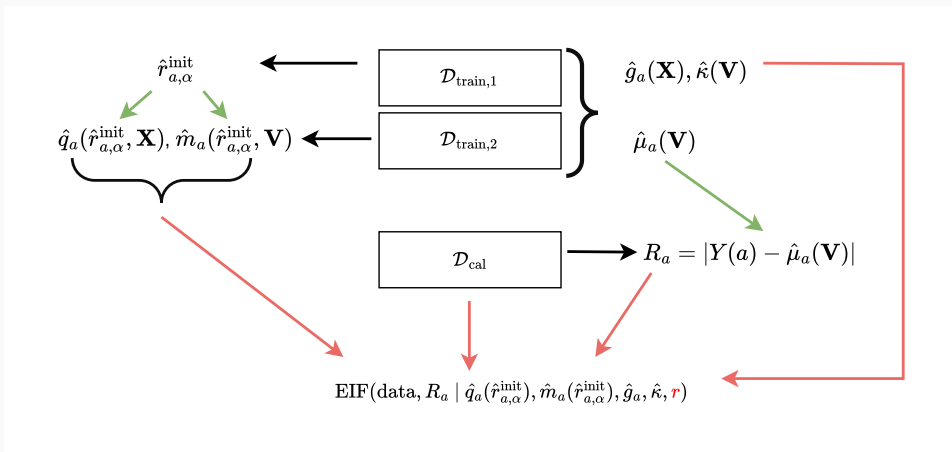
Using $\hat{\mu}_a(\mathbf{V})$ from training folds, construct conformity scores in the calibration set

Localized DML Implementation



Construct $\text{EIF}(r; \text{data}_i, j(r_{a;}))$ for all calibration set units

Localized DML Implementation



$$\text{Solve } \prod_{i \in \mathcal{D}_{\text{cal}}} \text{EIF}(r; \text{data}_i | \hat{r}_{a_i}^{\text{init}}) = 0$$

Theorem When $\hat{C}_a(\mathbf{V})$ is constructed according to our proposed DML algorithm, then

$$P(Y(a) \in \hat{C}_a(\mathbf{V}) | S = 0) = 1 + O_p(1) = \frac{p}{n} + R_n; \text{ where}$$

$$R_n = \sup_r |q_a(r) - \hat{q}_a(r)| + \sup_r |m_a(r) - \hat{m}_a(r)|$$

Theorem When $\hat{C}_a(\mathbf{V})$ is constructed according to our proposed DML algorithm, then

$$P(Y(a) \geq \hat{C}_a(\mathbf{V}) | S = 0) = 1 + O_p(1) = \rho_{\bar{n}} + R_n; \text{ where}$$

$$R_n = \sup_r \left(\frac{1}{n} \sum_{i=1}^n \hat{q}_a(r; \mathbf{X}_i) - q_a(r; \mathbf{X}) \right) + \sup_r \left(\frac{1}{n} \sum_{i=1}^n \hat{m}_a(r; \mathbf{X}_i) - m_a(r; \mathbf{X}) \right)$$

Bias decomposition implies **coverage bias** will shrink at ideal $n^{-1/2}$ rates so long as

1. $\sup_r \left(\frac{1}{n} \sum_{i=1}^n \hat{q}_a(r; \mathbf{X}_i) - q_a(r; \mathbf{X}) \right) = o_p(1) = \rho_{\bar{n}}$, and
2. $\sup_r \left(\frac{1}{n} \sum_{i=1}^n \hat{m}_a(r; \mathbf{X}_i) - m_a(r; \mathbf{X}) \right) = o_p(1) = \rho_{\bar{n}}$

Attainable if all nuisances are estimated at, e.g. $n^{-1/4}$ rates

Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

Methods

Numerical Experiments

Discussion

Compared our proposed method to

- [Weighted](#) conformal prediction (Tibshirani et al., 2019) based on earlier IPW expression
- A DML method based on Yang et al. (2024) that *ignores* runtime confounding, pretending V collects all confounders

Snapshot of synthetic data experiments

Compared our proposed method to

- **Weighted** conformal prediction (Tibshirani et al., 2019) based on earlier IPW expression
- A DML method based on Yang et al. (2024) that *ignores* runtime confounding, pretending V collects all confounders

Simulated data corresponding to a **runtime confounding** scenario, varying the sample size

Snapshot of synthetic data experiments

Compared our proposed method to

- **Weighted** conformal prediction (Tibshirani et al., 2019) based on earlier IPW expression
- A DML method based on Yang et al. (2024) that *ignores* runtime confounding, pretending V collects all confounders

Simulated data corresponding to a **runtime confounding** scenario, varying the sample size

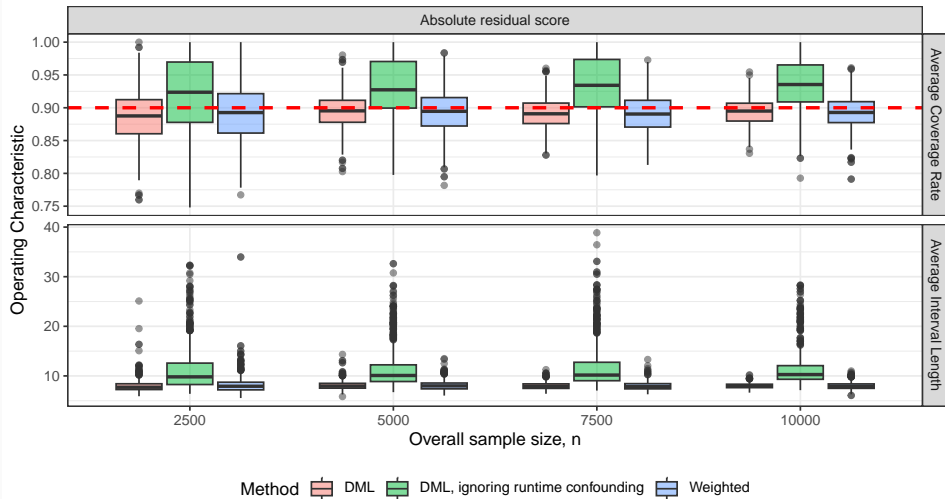
For simplicity, focused on producing 90% prediction intervals for $Y(1)$ and $Y(0)$ in target data. Across 1,000 iterations for each sample size considered, we compared

- The pooled average coverage rate of $Y(1)$ and $Y(0)$ in target data
- The pooled average interval width

Snapshot of synthetic experiments: varying n

Conformal prediction of counterfactuals in target population

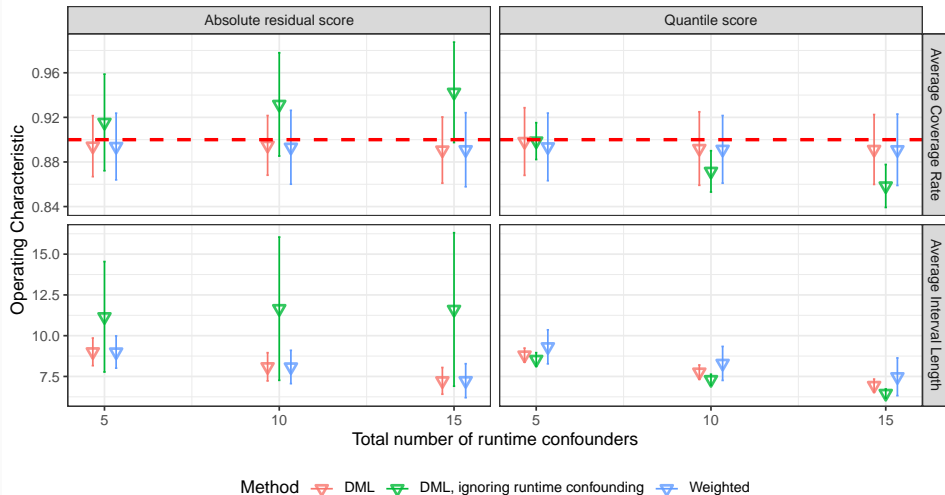
Average coverage and length of prediction intervals, 10 runtime confounders



Snapshot of synthetic experiments: varying number of runtime confounders

Conformal prediction of counterfactuals in target population

Average coverage and length of prediction intervals, fixing $n=5000$ and varying number of runtime confounders



Roadmap

Background

Counterfactual Prediction Intervals

Runtime Confounding and Multi-Source Data

Problem setting

Data Structure

Conformal Prediction Primer

Methods

Numerical Experiments

Discussion

Runtime confounding is a common challenge that can complicate the construction of valid prediction intervals for counterfactual outcomes

- ^ Our proposed method enables construction of asymptotically valid prediction intervals for counterfactual outcomes under runtime confounding
- ^ Theory + experiments demonstrate ignoring runtime confounding can result in intervals which considerably miscover

Many avenues for future work, including

- ^ Methods which address instances of **poor overlap** between the source and target populations
- ^ **Sensitivity analyses** to partially relax source independence assumption

Thank you!

<https://kbarnatchez.com>

[Link to paper](#)

References

- Bhattacharyya, A. and Barber, R. F. (2026). Group-weighted conformal prediction. *Electronic Journal of Statistics*, 20(1):1171{1199.
- Coston, A., Kennedy, E., and Chouldechova, A. (2020). Counterfactual predictions under runtime confounding. *Advances in neural information processing systems*33:4150{4162.
- Kallus, N., Mao, X., and Uehara, M. (2024). Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*25(16):1{59.
- Tibshirani, R. J., Foygel Barber, R., Candès, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*32.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.
- Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2024). Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae009.

Appendix [Back to start](#)

Theorems

Identification of τ_a ;

EIF for τ_a ;

Robustness Theorem

Assumptions

Counterfactual Quantile Regression

For compactness, let $R_a := R_a(Y; V)$ and notice

$$\begin{aligned}
 P(R_a \leq r_{a;} | S = 0) &= E(I(R_a \leq r_{a;}) | S = 0) \\
 &= E[E(I(R_a \leq r_{a;}) | X; S = 0) | S = 0] \\
 &= E[E(I(R_a \leq r_{a;}) | X; S = 1) | S = 0] \\
 &= E[E[E(I(R_a \leq r_{a;}) | X; A = a; S = 1) | V; S = 1] | S = 0] \\
 &= E[E[q_a(r_{a;}; X) | V; S = 1] | S = 0] \\
 &= E[m_a(r; V) | S = 0]
 \end{aligned}$$

Theorem

Let $q_a(r) := (q_a(r); m_a(r); g_a)$ and suppose \mathbb{P} is proportional to P . The efficient influence curve for r_a in a nonparametric model for \mathbb{P} is proportional to

$$\begin{aligned} & q_a(r_a; \mathbb{O}; q_a(r_a; \cdot)) := \\ & (1 - S)(m_a(r_a; V) - q_a(r_a; X)) + \frac{S(1 - (V))}{(V)} f(q_a(r_a; X) - m_a(r_a; V))g \\ & + w_a(\mathbb{O})f(R_a(Y; V) - r_a) - q_a(r_a; X)g; \end{aligned} \quad (1)$$

where $q_a(r; X) = P(R_a = r | X; A = a; S = 1)$, $m_a(r; X) = E[q_a(r; X) | V; S = 1]$, $(V) = P(S = 1 | V)$, and $g_a(X) = P(A = a | X; S = 1)$

Robustness Theorem

When $\hat{C}_a(V)$ is constructed according to our proposed debiased machine learning algorithm, then

$$P(Y(a) \in \hat{C}_a(V) | S = 0) = 1 + O_p\left(\frac{p}{\bar{n}} + R_n\right); \text{ where}$$

$$R_n = \sup_r \| \hat{q}_a(r; \cdot) - q_a(r; \cdot) \| + \sup_r \| \hat{m}_a(r; \cdot) - m_a(r; \cdot) \| + \dots$$

Theorems

Identification of r_a ;

EIF for r_a ;

Robustness Theorem

Assumptions

Counterfactual Quantile Regression

Assumptions

Assumption (Positivity)

$0 < P(A = a | X = x) < 1$ for all x with positive support

Assumption (Consistency)

$Y = I(A = a) Y(a)$

Assumption (Unconfoundedness)

$Y(a) \perp A | X$

Assumption (Source exchangeability)

$Y(a) \perp S | V$

Assumption (Source positivity)

$0 < P(S = 1 | V = v) < 1$ for all v with positive support

Revisiting Assumption 4

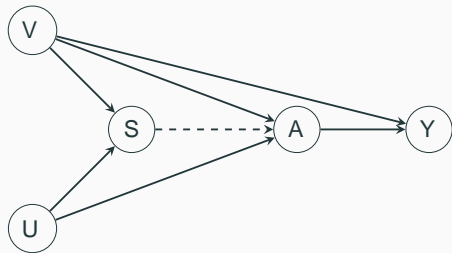
Can be difficult to assess when Assumption 4, $Y(a) \geq S_j V$, will hold

Often easier to instead think through the following two Assumptions, which imply Assumption 4

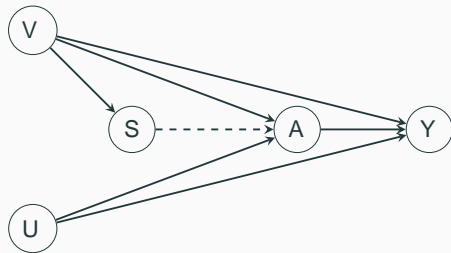
$$^{\wedge} Y(a) \geq S_j X$$

$$^{\wedge} U \geq S_j V$$

Some DAGs consistent with independence assumptions

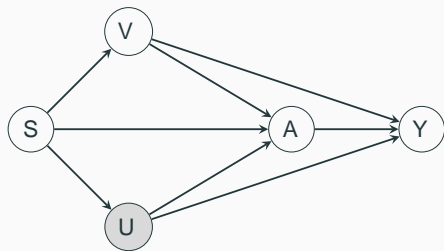


(a)

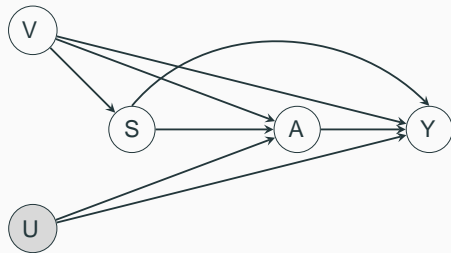


(b)

DAGs implying Assumption 4 is violated



(a)



(b)

Theorems

Identification of τ_a ;

EIF for τ_a ;

Robustness Theorem

Assumptions

Counterfactual Quantile Regression

Quantile Regression

Earlier, we discussed simple way for constructing intervals:

1. First, $\hat{m}_a(\mathbf{X}) = \hat{E}(Y | A = a; \mathbf{X})$ with the training data
2. Next, regress the estimated $\hat{m}_a(\mathbf{X})$ on \mathbf{V} , yielding $\hat{\mu}_a(\mathbf{V}) = \hat{E}(\hat{m}_a(\mathbf{X}) | \mathbf{V})$

Drawback of this approach is that interval will be **constant width** { won't adapt to local variability in $Y(a)$ around values of \mathbf{V}

Common approach in conformal prediction is to instead construct intervals based on a quantile conformity score

Quick aside: forming counterfactual predictions

Another common approach to form prediction intervals is to

- Predict $Q_{a; \tau_2}(\mathbf{V}) := \text{Quantile}_{\tau_2}(Y|\mathbf{V})$ and $Q_{a;1-\tau_2}(\mathbf{V}) := \text{Quantile}_{1-\tau_2}(Y|\mathbf{V})$
- Use tools from conformal to adjust naive intervals $(\hat{Q}_{a; \tau_2}(\mathbf{V}); \hat{Q}_{a;1-\tau_2}(\mathbf{V}))$

Challenge: $E(Q_{a; \tau_2}(\mathbf{X})|\mathbf{V}) \neq \text{Quantile}_{\tau_2}(Y(a)|\mathbf{V})$ due to nonlinearity of the quantile function

- But we need to use all of \mathbf{X} since \mathbf{V} doesn't contain all confounders

In our paper, we derive a valid loss function for estimating $Q_{a; \tau_2}(\mathbf{X})$ under runtime confounding settings

Quantile weighted loss function

If $Q_{a; \cdot}(\mathbf{v})$ satisfies $P(Y(a) \leq Q_{a; \cdot}(\mathbf{v}) | \mathbf{V} = \mathbf{v}; S = 0) = 1 - \alpha$ for all \mathbf{v} with positive support, then it additionally satisfies

$$Q_{a; \cdot}(\mathbf{V}) = \arg \min_{Q_{a, \alpha}} E [w_a(\mathbf{O}) | Y = Q_{a; \cdot}(\mathbf{V})]; \quad (2)$$

where $\rho(x) = |x| \ln|x| + (1 - |x|)$ is the pinball loss function and

$$w_a(\mathbf{O}) := \frac{I(A = a)S(1 - \alpha | \mathbf{V})}{g_a(\mathbf{X} | \mathbf{V})};$$