Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach

Ying Jin

Department of Statistics, Stanford University

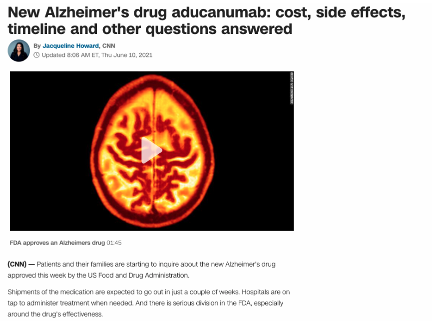
Stanford Causal Inference Group, January 25th, 2022

Joint work with

Zhimei Ren
UChicago Stats

Emmanuel Candès
Stanford Math & Stats

How would people react to a treatment?



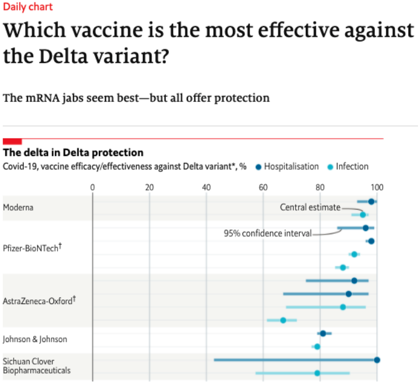
New Alzheimer's drug aducanumab: cost, side effects, timeline and other questions answered

By Jacqueline Howard, CNN
Updated 8:06 AM ET, Thu June 10, 2021

FDA approves an Alzheimers drug 01:45

(CNN) — Patients and their families are starting to inquire about the new Alzheimer's drug approved this week by the US Food and Drug Administration.

Shipments of the medication are expected to go out in just a couple of weeks. Hospitals are on tap to administer treatment when needed. And there is serious division in the FDA, especially around the drug's effectiveness.

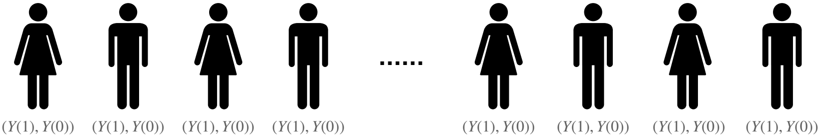


Daily chart

Which vaccine is the most effective against the Delta variant?

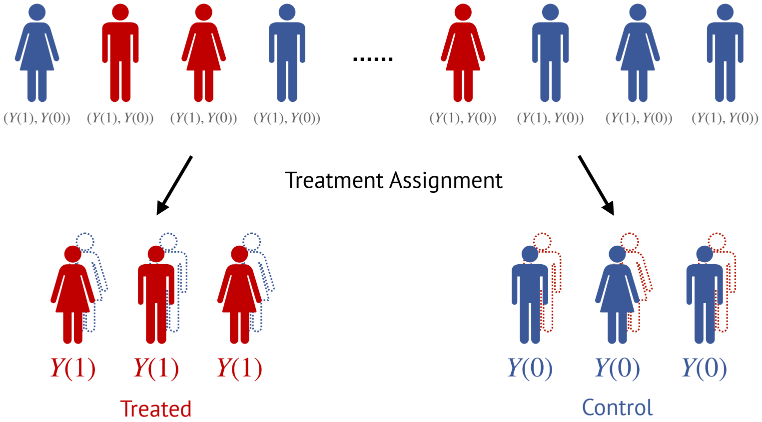
The mRNA jabs seem best—but all offer protection

Potential outcome framework

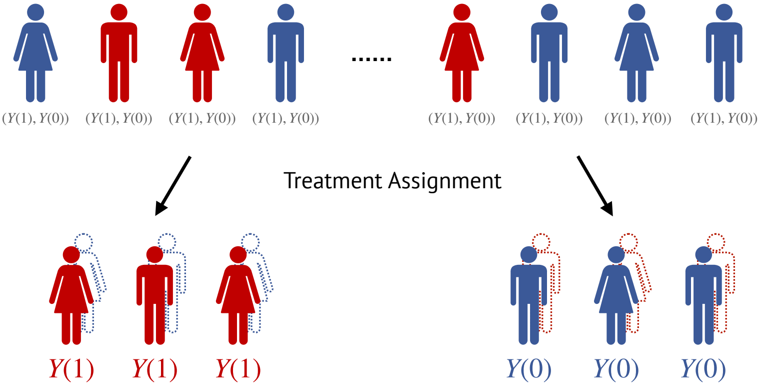


Sample from population

Potential outcome framework



Potential outcome framework



Treated

Control

Y(1)	✓	✓	✓
Y(0)	?	?	?

Y(1)	?	?	?
Y(0)	✓	✓	✓

Potential outcome framework

- Subjects $\big(X_i, Y_i(0), Y_i(1)\big) \overset{\text{i.i.d.}}{\sim} \mathcal{P}$
- Treatment assignment mechanism $T_i \sim \mathcal{T}$
- Population $(X_i, Y_i(1), Y_i(0), T_i) \sim \mathbb{P}$
- Partial observations: (X_i, T_i, Y_i), where $Y_i = Y(T_i)$ (SUTVA)

How would people react to a treatment?

Traditionally...

- Average treatment effect (ATE): $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

How would people react to a treatment?

Traditionally...

- Average treatment effect (ATE): $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

Heterogeneity? Individualization?

- Conditional average treatment effect (CATE): $\tau(x) = \mathbb{E}[Y(1) - Y(0) \,|\, X = x]$

How would people react to a treatment?

Traditionally...

- Average treatment effect (ATE): $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

Heterogeneity? Individualization?

- Conditional average treatment effect (CATE): $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$

But...

- Does an individual benefit from the treatment?

How would people react to a treatment?

Traditionally...

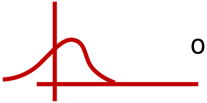
- Average treatment effect (ATE): $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

Heterogeneity? Individualization?

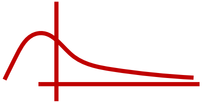
- Conditional average treatment effect (CATE): $\tau(x) = \mathbb{E}[Y(1) - Y(0) \,|\, X = x]$

But...

- Does an individual benefit from the treatment?
- Even if we know precisely that $\tau(x) = 1$, would $Y(1) - Y(0)$ of this individual be ?



or



How would people react to a treatment?

Traditionally...

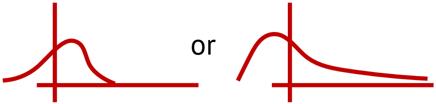
- Average treatment effect (ATE): $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

Heterogeneity? Individualization?

- Conditional average treatment effect (CATE): $\tau(x) = \mathbb{E}[Y(1) - Y(0) \,|\, X = x]$

But...

- Does an individual benefit from the treatment?
- Even if we know precisely that $\tau(x) = 1$, would $Y(1) - Y(0)$ of this individual be ?



- Inference of individual treatment effect (ITE) $Y(1) - Y(0)$.

Previous work

- Predictive inference of ITE [Lei and Candès '20]

For some unit $n+1$, construct predictive interval $\widehat{C}(X_{n+1})$ s.t.

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}(X_{n+1})\big) \geq 1 - \alpha.$$

Previous work

- Predictive inference of ITE [Lei and Candès '20]

For some unit $n+1$, construct predictive interval $\widehat{C}(X_{n+1})$ s.t.

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}(X_{n+1})\big) \geq 1 - \alpha.$$

- Main assumption: the strong ignorability condition

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \,|\, X_i$$

Previous work

- Predictive inference of ITE [Lei and Candès '20]

For some unit $n+1$, construct predictive interval $\widehat{C}(X_{n+1})$ s.t.

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}(X_{n+1})\big) \geq 1 - \alpha.$$

- Main assumption: the strong ignorability condition

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \,|\, X_i$$

- What happens if strong ignorability does not hold?

Previous work

- Predictive inference of ITE [Lei and Candès '20]

For some unit $n+1$, construct predictive interval $\widehat{C}(X_{n+1})$ s.t.

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}(X_{n+1})\big) \geq 1-\alpha.$$

- Main assumption: the strong ignorability condition

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \,|\, X_i$$

- What happens if strong ignorability does not hold?
- How to reliably infer ITE in this situation?

Overview of this work

- Characterize the counterfactual distribution under unmeasured confounding
- Robust predictive inference of ITE:

$$\mathbb{P}(Y(1) - Y(0) \in \widehat{C}(X)) \geq 1 - \alpha$$

  if the observational data is confounded up to some extent
- Make robust causal conclusions: inverting a sequence of hypothesis testing

Building block: counterfactual prediction

For some unit $n+1$, construct predictive interval $\widehat{C}(X_{n+1})$ s.t.

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}(X_{n+1})\big) \geq 1 - \alpha.$$

Building block: counterfactual prediction

For some unit $n+1$, construct predictive interval $\widehat{C}(X_{n+1})$ s.t.

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}(X_{n+1})\big) \geq 1 - \alpha.$$

- One outcome missing: observe $(X_{n+1}, Y_{n+1}(t))$.
- Both outcomes missing: observe X_{n+1}.

Building block: counterfactual prediction

- One outcome missing:
  - e.g., unit comes from $\mathbb{P}_{X,Y(1),Y(0)\,|\,T=0}$ (control group)

Building block: counterfactual prediction

- One outcome missing:
  - e.g., unit comes from $\mathbb{P}_{X,Y(1),Y(0) \mid T=0}$ (control group)
  - reduces to counterfactual prediction of $Y(1)$

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}_1(X_{n+1}) - Y_{n+1}(0) \mid T_{n+1} = 0\big) = \mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}_1(X_{n+1}) \mid T_{n+1} = 0\big)$$

Building block: counterfactual prediction

- One outcome missing:
  - e.g., unit comes from $\mathbb{P}_{X,Y(1),Y(0)\,|\,T=0}$ (control group)
  - reduces to counterfactual prediction of $Y(1)$

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}_1(X_{n+1}) - Y_{n+1}(0) \,\big|\, T_{n+1} = 0\big) = \mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}_1(X_{n+1}) \,\big|\, T_{n+1} = 0\big)$$

- Both outcomes missing:
  - Unit comes from $\mathbb{P}_{X,Y(1),Y(0)}$ (super-population)
  - Combine two counterfactual prediction intervals

Building block: counterfactual prediction

- One outcome missing:
  - e.g., unit comes from $\mathbb{P}_{X,Y(1),Y(0)\,|\,T=0}$ (control group)
  - reduces to counterfactual prediction of $Y(1)$

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \widehat{C}_1(X_{n+1}) - Y_{n+1}(0) \,|\, T_{n+1} = 0\big) = \mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}_1(X_{n+1}) \,|\, T_{n+1} = 0\big)$$

- Both outcomes missing:
  - Unit comes from $\mathbb{P}_{X,Y(1),Y(0)}$ (super-population)
  - Combine two counterfactual prediction intervals

Counterfactual prediction: construct $\widehat{C}(X_{n+1})$ s.t.

$$\mathbb{P}\big(Y_{n+1}(1) \in \widehat{C}(X_{n+1}) \,\big|\, T_{n+1} = 0\big) \geq 1 - \alpha.$$

Robust predictive inference of counterfactuals

Robust prediction of ITE

Γ-value: making robust causal conclusions

Numerical experiments

Counterfactual prediction faces distributional shift

- The training distribution is $\mathbb{P}_{\text{train}} = (X_i, Y_i(1)) \sim \mathbb{P}_{X,Y(1) \mid T=1}$
- The target distribution is $\mathbb{P}_{\text{target}} = (X_{n+1}, Y_{n+1}(1)) \sim \mathbb{P}_{X,Y(1) \mid T=0}$

Counterfactual prediction faces distributional shift

- The training distribution is $\mathbb{P}_{\text{train}} = (X_i, Y_i(1)) \sim \mathbb{P}_{X,Y(1) \mid T=1}$
- The target distribution is $\mathbb{P}_{\text{target}} = (X_{n+1}, Y_{n+1}(1)) \sim \mathbb{P}_{X,Y(1) \mid T=0}$

- There is a distributional shift from $\mathbb{P}_{\text{train}}$ to $\mathbb{P}_{\text{target}}$:

$$\frac{\mathrm{d}\mathbb{P}_{\text{target}}}{\mathrm{d}\mathbb{P}_{\text{train}}} = \frac{\mathrm{d}\mathbb{P}_{X,Y(1) \mid T=0}}{\mathrm{d}\mathbb{P}_{X,Y(1) \mid T=1}} =: w(x, y)$$

Counterfactual prediction faces distributional shift

- The training distribution is $\mathbb{P}_{\text{train}} = (X_i, Y_i(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=1}$
- The target distribution is $\mathbb{P}_{\text{target}} = (X_{n+1}, Y_{n+1}(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=0}$
- There is a distributional shift from $\mathbb{P}_{\text{train}}$ to $\mathbb{P}_{\text{target}}$:

$$\frac{\mathrm{d}\mathbb{P}_{\text{target}}}{\mathrm{d}\mathbb{P}_{\text{train}}} = \frac{\mathrm{d}\mathbb{P}_{X,Y(1)\,|\,T=0}}{\mathrm{d}\mathbb{P}_{X,Y(1)\,|\,T=1}} =: w(x,y)$$

- How to do predictive inference under distributional shift?

Weighted (split) conformal prediction [Tibshirani et al., 2019]

- A general recipe to construct $\widehat{C}(X_{n+1})$ such that

$$\mathbb{P}\big(Y_{n+1} \in \widehat{C}(X_{n+1})\big) \geq 1 - \alpha$$

  even when there is a distributional shift $w(x, y)$ from $\mathbb{P}_{\text{train}}$ to $\mathbb{P}_{\text{target}}$

(Split) conformal inference framework [Vovk et al., 1999]

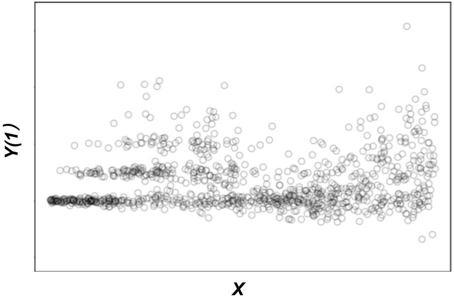
- Randomly split the $T_i = 1$ observations into two sets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{calib}}$ $(n = |\mathcal{D}_{\text{calib}}|)$
- Train any model on $\mathcal{D}_{\text{train}}$, and form a residual $V(x, y) \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$
- The prediction interval will take the form

$$\widehat{C}(x) = \left\{y \colon V(x, y) \leq \widehat{v}(y)\right\}$$

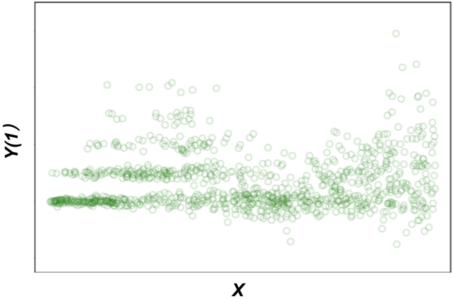
for some $\widehat{v}(y) \in \mathbb{R}$ calibrated by $\mathcal{D}_{\text{calib}}$

(Split) conformal inference framework [Vovk et al., 1999]

- Example: CQR (Conformalized Quantile Regression)
  - Sample splitting



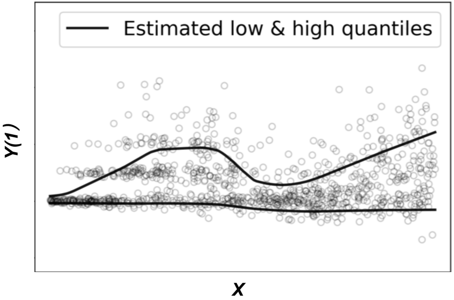
Training set



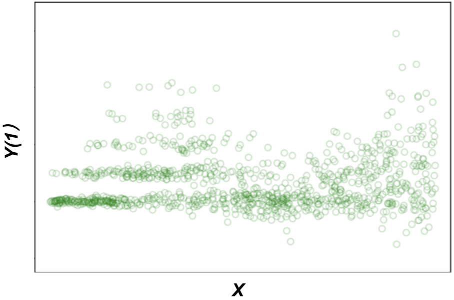
Calibration set

(Split) conformal inference framework [Vovk et al., 1999]

- Example: CQR (Conformalized Quantile Regression)
  - Train some conditional quantile function $\widehat{q}_{\alpha/2}(x)$ and $\widehat{q}_{1-\alpha/2}(x)$ on $\mathcal{D}_{\text{train}}$



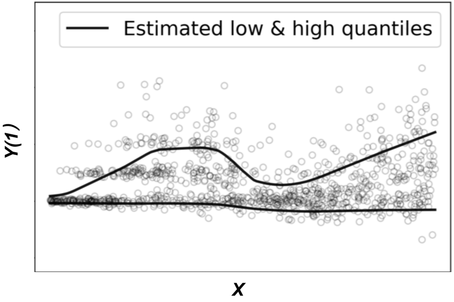
Training set



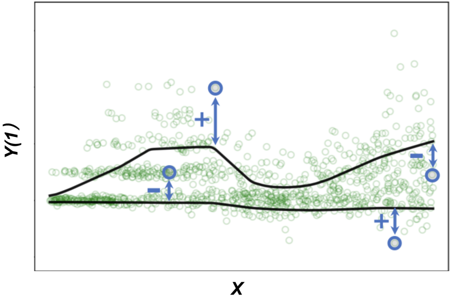
Calibration set

(Split) conformal inference framework [Vovk et al., 1999]

- Example: CQR (Conformalized Quantile Regression)
  - Train some conditional quantile function $\widehat{q}_{\alpha/2}(x)$ and $\widehat{q}_{1-\alpha/2}(x)$ on $\mathcal{D}_{\text{train}}$
  - Set $V(x, y) = \max\{\widehat{q}_{\alpha/2}(x) - y,\ y - \widehat{q}_{1-\alpha/2}(x)\}$
  - Choose some $\widehat{v} \in \mathbb{R}$ from $\mathcal{D}_{\text{calib}}$



Training set



Calibration set

(Split) conformal inference framework [Vovk et al., 1999]

- Example: CQR (Conformalized Quantile Regression)
  - Train some conditional quantile function $\widehat{q}_{\alpha/2}(x)$ and $\widehat{q}_{1-\alpha/2}(x)$ on $\mathcal{D}_{\text{train}}$
  - Set $V(x, y) = \max\{\widehat{q}_{\alpha/2}(x) - y,\ y - \widehat{q}_{1-\alpha/2}(x)\}$
  - Choose some $\widehat{v} \in \mathbb{R}$ from $\mathcal{D}_{\text{calib}}$
  - The predictive interval is

$$\widehat{C}(x) = \left[\widehat{q}_{\alpha/2}(x) - \widehat{v},\ \widehat{q}_{1-\alpha/2}(x) + \widehat{v}\right]$$

Weighted (split) conformal prediction [Tibshirani et al., 2019]

- Choose some $\widehat{v}(y) \in \mathbb{R}$ from $\mathcal{D}_{\text{calib}}$

Weighted (split) conformal prediction [Tibshirani et al., 2019]

- Choose some $\widehat{v}(y) \in \mathbb{R}$ from $\mathcal{D}_{\text{calib}}$
- When $\mathbb{P}_{\text{train}} = \mathbb{P}_{\text{target}}$

$$\widehat{v}(y) = \text{Quantile}\left(1 - \alpha; \sum_{i \in \mathcal{D}_{\text{calib}}} \frac{1}{n+1} \delta_{V(X_i, Y_i(1))}\right)$$

Weighted (split) conformal prediction [Tibshirani et al., 2019]

- Choose some $\widehat{v}(y) \in \mathbb{R}$ from $\mathcal{D}_{\text{calib}}$
- When $\mathbb{P}_{\text{train}} = \mathbb{P}_{\text{target}}$

$$\widehat{v}(y) = \text{Quantile}\left(1-\alpha; \sum_{i \in \mathcal{D}_{\text{calib}}} \frac{1}{n+1} \delta_{V(X_i, Y_i(1))}\right)$$

- When distributional shift $w(x,y) = \mathrm{d}\mathbb{P}_{\text{target}} / \mathrm{d}\mathbb{P}_{\text{train}}$

$$\widehat{v}(y) = \text{Quantile}\left(1-\alpha; \sum_{i \in \mathcal{D}_{\text{calib}}} p_w(i) \delta_{V(X_i, Y_i(1))}\right),$$

$$\text{where} \quad p_w(i) = \frac{w(X_i, Y_i)}{\sum_{j \in \mathcal{D}_{\text{calib}}} w(X_j, Y_j) + w(x,y)}$$

if $w(x,y) = w(x)$ is a covariate shift, $\widehat{v}(y)$ does not depend on y and is **computable**.

Weighted (split) conformal prediction [Tibshirani et al., 2019]

Conformal inference

- Any conditional distribution
- Any sample size
- Any training procedure

Without confounding, $w(x, y)$ is a covariate shift [Lei and Candès '20]

Counterfactual prediction without confounding [Lei and Candès '20]

- The training distribution is $\mathbb{P}_{\text{train}} = (X_i, Y_i(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=1}$
- The target distribution is $\mathbb{P}_{\text{target}} = (X_{n+1}, Y_{n+1}(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=0}$

Counterfactual prediction without confounding [Lei and Candès '20]

- The training distribution is $\mathbb{P}_{\text{train}} = (X_i, Y_i(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=1}$
- The target distribution is $\mathbb{P}_{\text{target}} = (X_{n+1}, Y_{n+1}(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=0}$
- Under the strong ignorability assumption (no confounding)

$$\mathbb{P}_{X,Y(1)\,|\,T=0} = \mathbb{P}_{X\,|\,T=0} \times \mathbb{P}_{Y(1)\,|\,X,T=0} = \underbrace{\mathbb{P}_{X\,|\,T=0}}_{\text{identifiable}} \times \underbrace{\mathbb{P}_{Y(1)\,|\,X,T=1}}_{\text{identifiable}}$$

Counterfactual prediction without confounding [Lei and Candès '20]

- The training distribution is $\mathbb{P}_{\text{train}} = (X_i, Y_i(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=1}$
- The target distribution is $\mathbb{P}_{\text{target}} = (X_{n+1}, Y_{n+1}(1)) \sim \mathbb{P}_{X,Y(1)\,|\,T=0}$

- Under the strong ignorability assumption (no confounding)

$$\mathbb{P}_{X,Y(1)\,|\,T=0} = \mathbb{P}_{X\,|\,T=0} \times \mathbb{P}_{Y(1)\,|\,X,T=0} = \underbrace{\mathbb{P}_{X\,|\,T=0}}_{\text{identifiable}} \times \underbrace{\mathbb{P}_{Y(1)\,|\,X,T=1}}_{\text{identifiable}}$$

- Predictive inference under covariate shift:

$$\frac{\mathrm{d}\mathbb{P}_{\text{target}}}{\mathrm{d}\mathbb{P}_{\text{train}}}(x,y) = \frac{\mathrm{d}\mathbb{P}_{X\,|\,T=1}}{\mathrm{d}\mathbb{P}_{X\,|\,T=0}}(x) =: w(x) \propto \frac{1-e(x)}{e(x)},$$

where $e(x) = \mathbb{P}(T = 1 \,|\, X = x)$ is the observed propensity score

Where we are now

Previous work:

- Counterfactual inference faces distributional shift
- A tool to conduct predictive inference
- Computable and identifiable under strong ignorability condition

Where we are now

Previous work:

- Counterfactual inference faces distributional shift
- A tool to conduct predictive inference
- Computable and identifiable under strong ignorability condition

- What if strong ignorability fails?
- The distributional shift is no longer a covariate shift, and is not identifiable / computable
- Is there a way to conduct reliable counterfactual inference?

Next: adding robustness to conformal inference

Conformal inference

- Any conditional distribution
- Any sample size
- Any training procedure

+

- Validity under unknown distributional shift

Unmeasured confounding

Assume unobserved confounders U that influence both T and $(Y(0), Y(1))$.

Unmeasured confounding

Assume unobserved confounders U that influence both T and $(Y(0), Y(1))$.

- Subjects $\big(U_i, X_i, Y_i(0), Y_i(1)\big) \overset{\text{i.i.d.}}{\sim} \mathcal{P}$
- Treatment assignment $T_i \sim \mathcal{T}'$
- Observables: $(X_i, T_i, Y_i(T_i))$
- Assumption:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i \mid (X_i, U_i)$$

Unmeasured confounding

Assume unobserved confounders U that influence both T and $(Y(0), Y(1))$.

- Subjects $\big(U_i, X_i, Y_i(0), Y_i(1)\big) \overset{\text{i.i.d.}}{\sim} \mathcal{P}$
- Treatment assignment $T_i \sim \mathcal{T}'$
- Observables: $(X_i, T_i, Y_i(T_i))$
- Assumption:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i \mid (X_i, U_i)$$

Such confounders always exists: think about $U = (Y(1), Y(0))$

Modelling the confounding strength: the marginal Γ-selection condition

$$(Y_i(0), Y(1)) \perp\!\!\!\perp T_i \mid (X_i, U_i)$$

- Bounds on selection bias
- marginal Γ-selection condition [Tan '06, Zhao et al. '18]:

$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}(T=1 \mid X=x, U=u)}{\mathbb{P}(T=0 \mid X=x, U=u)} \cdot \frac{\mathbb{P}(T=0 \mid X=x)}{\mathbb{P}(T=1 \mid X=x)} \leq \Gamma$$

Distributional shift under the marginal sensitivity model

- Training/calibration distribution: $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{calib}} = (X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X,Y(1) \mid T=1}$
- Target distribution: $(X_{n+1}, Y_{n+1}) \sim \mathbb{P}_{X,Y(1) \mid T=0}$

Distributional shift under the marginal sensitivity model

- Training/calibration distribution: $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{calib}} = (X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X,Y(1)\,|\,T=1}$
- Target distribution: $(X_{n+1}, Y_{n+1}) \sim \mathbb{P}_{X,Y(1)\,|\,T=0}$

Under marginal Γ-selection,

$$\frac{1}{\Gamma} \cdot \frac{p}{1-p} \frac{1-e(x)}{e(x)} \leq \frac{d\mathbb{P}_{X,Y(1)\,|\,T=0}}{d\mathbb{P}_{X,Y(1)\,|\,T=1}} \leq \Gamma \cdot \frac{p}{1-p} \frac{1-e(x)}{e(x)},$$

where $e(x) = \mathbb{P}(T = 1 \,|\, X = x)$, $p = \mathbb{P}(T = 1)$.

- The distributional shift is bounded by identifiable functions of x

A more general problem: robust prediction

$$\text{Train/calibration distribution: } \mathbb{P}_{XY} = \mathbb{P}_{Y|X} \cdot \mathbb{P}_X$$
$$\text{Unknown target distribution: } \widetilde{\mathbb{P}}_{XY} = \widetilde{\mathbb{P}}_{Y|X} \cdot \widetilde{\mathbb{P}}_X$$

Assume for $\mathbb{P}$-almost all $x \in \mathcal{X}$,

$$\ell(x) \leq \frac{d\widetilde{\mathbb{P}}_{XY}}{d\mathbb{P}_{XY}}(x, y) \leq u(x).$$

Goal: construct a predictive interval $\widehat{C}(X_{n+1})$ such that

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1})\big) \geq 1 - \alpha.$$

Robust weighted conformal prediction

- Compute the residuals $\{V(X_i, Y_i) \colon i \in \mathcal{D}_{\text{calib}}\}$
- Let $\widehat{\ell}(x)$ and $\widehat{u}(x)$ be estimators of $\ell(\cdot), u(\cdot)$ from $\mathcal{D}_{\text{train}}$.
- Construct $\widehat{C}(X_{n+1}) = \{y : V(X_{n+1}, y) \leq \widehat{v}\}$, where

$$\widehat{v} = \sup_{\widehat{\ell}(X_i) \leq w(X_i, Y_i) \leq \widehat{u}(X_i)} \text{Quantile}\left(1-\alpha; \sum_{i \in \mathcal{D}_{\text{calib}}} p_w(i) \delta_{V(X_i, Y_i)}\right),$$

$$\text{where} \quad p_w(i) = \frac{w(X_i, Y_i)}{\sum_{j \in \mathcal{D}_{\text{calib}}} w(X_j, Y_j) + w(x, y)}$$

- A conservative estimate of what we should have if $w(x, y)$ is known

Robust weighted conformal prediction

- Let $V_{[1]} \leq V_{[2]} \leq \cdots V_{[n]}$ be the order statistics of $\{V(X_i, Y_i)\}_{i \in \mathcal{D}\text{calib}}$

Robust weighted conformal prediction

- Let $V_{[1]} \leq V_{[2]} \leq \cdots V_{[n]}$ be the order statistics of $\{V(X_i, Y_i)\}_{i \in \mathcal{D}\text{calib}}$
- An equivalent form of weighted conformal prediction: $\widehat{v} = V_{[k^*]}$,

$$k^* = \min \left\{ k \colon \frac{\sum_{i=1}^{k} w(X_{[i]}, Y_{[i]})}{\sum_{i=1}^{n} w(X_{[i]}, Y_{[i]}) + w(X_{n+1}, y)} \geq 1 - \alpha \right\}.$$

Robust weighted conformal prediction

- Let $V_{[1]} \leq V_{[2]} \leq \cdots V_{[n]}$ be the order statistics of $\{V(X_i, Y_i)\}_{i \in \mathcal{D}\text{calib}}$
- An equivalent form of weighted conformal prediction: $\widehat{v} = V_{[k^*]}$,

$$k^* = \min\left\{k : \frac{\sum_{i=1}^{k} w(X_{[i]}, Y_{[i]})}{\sum_{i=1}^{n} w(X_{[i]}, Y_{[i]}) + w(X_{n+1}, y)} \geq 1 - \alpha\right\}.$$

- Robust weighted conformal: $\widehat{v} = V_{[\widehat{k}^*]}$,

$$\widehat{k}^* = \min\left\{k : \frac{\sum_{i=1}^{k} \widehat{\ell}(X_{[i]})}{\sum_{i=1}^{k} \widehat{\ell}(X_{[i]}) + \sum_{i=k+1}^{n} \widehat{u}(X_{[i]}) + \widehat{u}(X_{n+1})} \geq 1 - \alpha\right\}.$$

Robust weighted conformal prediction

- Let $V_{[1]} \leq V_{[2]} \leq \cdots V_{[n]}$ be the order statistics of $\{V(X_i, Y_i)\}_{i \in \mathcal{D}\text{calib}}$
- An equivalent form of weighted conformal prediction: $\widehat{v} = V_{[k^*]}$,

$$k^* = \min\left\{k : \frac{\sum_{i=1}^{k} w(X_{[i]}, Y_{[i]})}{\sum_{i=1}^{n} w(X_{[i]}, Y_{[i]}) + w(X_{n+1}, y)} \geq 1 - \alpha\right\}.$$

- Robust weighted conformal: $\widehat{v} = V_{[\widehat{k}^*]}$,

$$\widehat{k}^* = \min\left\{k : \frac{\sum_{i=1}^{k} \widehat{\ell}(X_{[i]})}{\sum_{i=1}^{k} \widehat{\ell}(X_{[i]}) + \sum_{i=k+1}^{n} \widehat{u}(X_{[i]}) + \widehat{u}(X_{n+1})} \geq 1 - \alpha\right\}.$$

- Conservative estimate: $\widehat{k}^* \geq k^*$.

Robust weighted conformal prediction

- Let $V_{[1]} \leq V_{[2]} \leq \cdots V_{[n]}$ be the order statistics of $\{V(X_i, Y_i)\}_{i \in \mathcal{D}\text{calib}}$
- An equivalent form of weighted conformal prediction: $\widehat{v} = V_{[k^*]}$,

$$k^* = \min\left\{k : \frac{\sum_{i=1}^{k} w(X_{[i]}, Y_{[i]})}{\sum_{i=1}^{n} w(X_{[i]}, Y_{[i]}) + w(X_{n+1}, y)} \geq 1 - \alpha\right\}.$$

- Robust weighted conformal: $\widehat{v} = V_{[\widehat{k}^*]}$,

$$\widehat{k}^* = \min\left\{k : \frac{\sum_{i=1}^{k} \widehat{\ell}(X_{[i]})}{\sum_{i=1}^{k} \widehat{\ell}(X_{[i]}) + \sum_{i=k+1}^{n} \widehat{u}(X_{[i]}) + \widehat{u}(X_{n+1})} \geq 1 - \alpha\right\}.$$

- Conservative estimate: $\widehat{k}^* \geq k^*$.
- Optimal under the constraints on unknown weights

$$\widehat{\ell}(X_i) \leq w(X_i, Y_i) \leq \widehat{u}(X_i).$$

Marginally valid predictive interval: coverage guarantee

Theorem

If the super-population satisfies marginal Γ-selection, then

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1})\big) \geq 1 - \alpha - \widehat{\Delta} \cdot \big\|1/\widehat{\ell}(X_i)\big\|_q,$$

where the gap is

$$\widehat{\Delta} = \Big\|[\widehat{\ell}(X_i) - w(X_i, Y_i)]_+\Big\|_p + \Big\|[\widehat{u}(X_i) - w(X_i, Y_i)]_-\Big\|_p + \frac{1}{n}\Big\|w(X_i, Y_i)^{1/p} \cdot [\widehat{u}(X_i) - w(X_i, Y_i)]_-\Big\|_p.$$

The expectation is over all training data and the test sample.

- When using estimated bounds, coverage depends on how **aggressive** the estimators are.

Where we are now

- Counterfactual inference under unmeasured confounding
  - Characterize the distributional shift under unmeasured confounding
  - Robust conformal inference by conservative quantile
  - Model free, finite sample, robustness to estimation error

Where we are now

- Counterfactual inference under unmeasured confounding
  - Characterize the distributional shift under unmeasured confounding
  - Robust conformal inference by conservative quantile
  - Model free, finite sample, robustness to estimation error

- Too conservative in face of unknown distributional shift?

  Did not use the constraint that $w(x, y)$ is a likelihood ratio function.

Where we are now

- Counterfactual inference under unmeasured confounding
  - Characterize the distributional shift under unmeasured confounding
  - Robust conformal inference by conservative quantile
  - Model free, finite sample, robustness to estimation error

- Too conservative in face of unknown distributional shift?

  Did not use the constraint that $w(x, y)$ is a likelihood ratio function.

- Marginalized guarantee? Can we have confidence given the current dataset?

Next: PAC-type predictive inference

Conformal inference

- Any conditional distribution
- Any sample size
- Any training procedure

✚

- Validity under unknown distributional shift

✚

- Training-conditional validity
- Sharpness

PAC-type prediction interval

- For prediction interval $\widehat{C}(x) = \{y \colon V(x,y) \leq \widehat{v}\}$,

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1}) \,\big|\, \mathcal{D}_{\text{calib}}\big) = \widetilde{\mathbb{P}}\big(V(X_{n+1}, Y_{n+1}) \leq \widehat{v} \,|\, \mathcal{D}_{\text{calib}}\big) = \widetilde{F}(\widehat{v}),$$

where $\widetilde{F}$ is the distribution function of $V(X,Y)$ if $(X,Y) \sim \widetilde{\mathbb{P}}$

PAC-type prediction interval

- For prediction interval $\widehat{C}(x) = \{y \colon V(x,y) \le \widehat{v}\}$,

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1}) \,\big|\, \mathcal{D}_{\mathrm{calib}}\big) = \widetilde{\mathbb{P}}\big(V(X_{n+1}, Y_{n+1}) \le \widehat{v} \,|\, \mathcal{D}_{\mathrm{calib}}\big) = \widetilde{F}(\widehat{v}),$$

  where $\widetilde{F}$ is the distribution function of $V(X, Y)$ if $(X, Y) \sim \widetilde{\mathbb{P}}$

- Idea: construct $\widehat{v}$ as a conservative estimate of $\widetilde{F}^{-1}(1-\alpha)$

PAC-type prediction interval

- Idea: let $\widehat{v}$ be a conservative estimate of $\widetilde{F}^{-1}(1-\alpha)$ for all $\widetilde{F}$

PAC-type prediction interval

- Idea: let $\widehat{v}$ be a conservative estimate of $\widetilde{F}^{-1}(1-\alpha)$ for all $\widetilde{F}$

- Construct envelope function such that

$$G(t) \leq \inf_{\widetilde{F}} \widetilde{F}(t), \quad \text{where} \quad \widetilde{F}(t) = \widetilde{\mathbb{P}}(V(X,Y) \leq t).$$

We set $G(t) = \max\left\{\mathbb{E}[\mathbb{1}_{\{V(X,Y)\leq t\}}\,\widehat{\ell}(X)], 1 - \mathbb{E}[\mathbb{1}_{\{V(X,Y)>t\}}\,\widehat{u}(X)]\right\}$

PAC-type prediction interval

- Idea: let $\widehat{v}$ be a conservative estimate of $\widetilde{F}^{-1}(1-\alpha)$ for all $\widetilde{F}$
- Construct envelope function such that

$$G(t) \leq \inf_{\widetilde{F}} \widetilde{F}(t), \quad \text{where} \quad \widetilde{F}(t) = \widetilde{\mathbb{P}}(V(X,Y) \leq t).$$

We set $G(t) = \max\left\{\mathbb{E}[\mathbb{1}_{\{V(X,Y)\leq t\}}\,\widehat{\ell}(X)], 1 - \mathbb{E}[\mathbb{1}_{\{V(X,Y)>t\}}\,\widehat{u}(X)]\right\}$

- Construct a lower confidence bound $\widehat{G}(t)$ for $G(t)$ such that for any fixed t,

$$\mathbb{P}_{\mathcal{D}_{\text{calib}}}\left(\widehat{G}(t) \leq G(t)\right) \geq 1 - \delta$$

PAC-type prediction interval

- Idea: let $\widehat{v}$ be a conservative estimate of $\widetilde{F}^{-1}(1-\alpha)$ for all $\widetilde{F}$

- Construct envelope function such that

$$G(t) \leq \inf_{\widetilde{F}} \widetilde{F}(t), \quad \text{where} \quad \widetilde{F}(t) = \widetilde{\mathbb{P}}(V(X,Y) \leq t).$$

We set $G(t) = \max\left\{\mathbb{E}[\mathbb{1}_{\{V(X,Y)\leq t\}}\,\widehat{\ell}(X)], 1 - \mathbb{E}[\mathbb{1}_{\{V(X,Y)>t\}}\,\widehat{u}(X)]\right\}$

- Construct a lower confidence bound $\widehat{G}(t)$ for $G(t)$ such that for any fixed t,

$$\mathbb{P}_{\mathcal{D}_{\text{calib}}}\left(\widehat{G}(t) \leq G(t)\right) \geq 1 - \delta$$

- Construct the prediction interval

$$\widehat{C}(X_{n+1}) = \left\{y : V(X_{n+1}, y) \leq \inf\left\{t : \widehat{G}_n(t) \geq 1-\alpha\right\}\right\}$$

PAC-type predictive interval: estimated bounds

Theorem

If the super-population satisfies marginal Γ-selection, then

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \widehat{C}(X_{n+1}) \,\big|\, \mathcal{D}\big) \geq 1 - \alpha - \widehat{\Delta}$$

with probability at least $1 - \delta$ with respect to $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{calib}$, where

$$\widehat{\Delta} = \max\Big\{\mathbb{E}\big[(\widehat{\ell}(X_i) - w(X_i, Y_i)_+\big],\ \mathbb{E}\big[(\widehat{u}(X_i) - w(X_i, Y_i))_-\big]\Big\}.$$

- The procedure is tight for some classes of problems
- In the paper: an exact characterization of the range of $\widetilde{\mathbb{P}}$

Robust predictive inference of counterfactuals

Robust prediction of ITE

Γ-value: making robust causal conclusions

Numerical experiments

Construct robust prediction of ITE

- Now construct predictive intervals for $\Delta = Y(1) - Y(0)$ such that

$$\mathbb{P}\big(\Delta \in \widehat{C}(X, \Gamma)\big) \geq 1 - \alpha$$

when the observational data satisfies marginal Γ selection.

Construct robust prediction of ITE

- Now construct predictive intervals for $\Delta = Y(1) - Y(0)$ such that

$$\mathbb{P}\big(\Delta \in \widehat{C}(X, \Gamma)\big) \geq 1 - \alpha$$

  when the observational data satisfies marginal Γ selection.

- When one outcome is missing: use prediction of the missing outcome.

- When both outcomes are missing: two prediction intervals with Bonferroni correction.

Robust predictive inference of counterfactuals

Robust prediction of ITE

Γ-value: making robust causal conclusions

Numerical experiments

Testing a sequence of hypotheses

- Sensitivity analysis gauges the robustness of causal conclusion against confounding

Testing a sequence of hypotheses

- Sensitivity analysis gauges the robustness of causal conclusion against confounding
- Suppose Γ^* is the unknown true confounding level

Testing a sequence of hypotheses

- Sensitivity analysis gauges the robustness of causal conclusion against confounding
- Suppose Γ^* is the unknown true confounding level
- Simultaneously test the hypotheses

$$H_0(\Gamma): \quad \Delta = Y(1) - Y(0) \in C, \quad \text{and} \quad \Gamma^* \le \Gamma.$$

Here C is a pre-specified set, for example, $(-\infty, 0]$

Testing a sequence of hypotheses

- Sensitivity analysis gauges the robustness of causal conclusion against confounding
- Suppose Γ^* is the unknown true confounding level
- Simultaneously test the hypotheses

$$H_0(\Gamma): \quad \Delta = Y(1) - Y(0) \in C, \quad \text{and} \quad \Gamma^* \leq \Gamma.$$

  Here C is a pre-specified set, for example, $(-\infty, 0]$

- Rejecting $H_0(\Gamma)$ means
  - either the causal conclusion is non-null
  - either the confounding strength is larger than Γ

Testing a sequence of hypotheses

- Simultaneously test the hypotheses

$$H_0(\Gamma): \quad \Delta = Y(1) - Y(0) \in C, \quad \text{and} \quad \Gamma^* \leq \Gamma.$$

Testing a sequence of hypotheses

- Simultaneously test the hypotheses

$$H_0(\Gamma)\colon \quad \Delta = Y(1) - Y(0) \in C, \quad \text{and} \quad \Gamma^* \leq \Gamma.$$

Our testing procedure:

$$\mathcal{R} = \left\{\Gamma\colon C \cap \widehat{C}(X, \Gamma) = \varnothing\right\}$$

The probability that we make any mistake on any true $H_0(\Gamma)$ is controlled:

$$\mathbb{P}\left(\mathcal{R} \cap \mathcal{H}_0 \neq \varnothing\right) \leq \alpha,$$

where $\mathcal{H}_0$ is the set of true hypotheses.

Γ-value and robust causal conclusion

- Using our procedure, $\widehat{C}(X, \Gamma)$ is nested, so $\mathcal{R} = (1, \widehat{\Gamma})$. "ITE is not in C unless $\Gamma^* \geq \widehat{\Gamma}$"
- If we set $C = (-\infty, 0]$, we test for positive ITE. Then we guarantee

$$\mathbb{P}\big(\mathcal{R} \cap \mathcal{H}_0 \neq \varnothing\big) = \mathbb{P}\big(Y(1) \leq Y(0),\ \widehat{\Gamma} \geq \Gamma^*\big) \leq \alpha.$$

- $\widehat{\Gamma}$ is a LCB for Γ^* in the case of a deterministic null

Robust predictive inference of counterfactuals

Robust prediction of ITE

Γ-value: making robust causal conclusions

Numerical experiments

Counterfactual prediction: simulation settings

- Confounded simulation setting adapted from [Yadlowsky et al.'18]
- Task: counterfactual prediction of $Y(1)$ for super population

Counterfactual prediction: simulation settings

- Confounded simulation setting adapted from [Yadlowsky et al.'18]
- Task: counterfactual prediction of $Y(1)$ for super population
- Identical population model
  - Covariates $X \sim \mathrm{Unif}[0,1]^p$, $p = 4, 20$.
  - Confounder $U \mid_X \sim N((0, 1 + \frac{1}{2}(2.5X_1)^2)$, $Y(1) = \beta^\top X + U$,
  - Sample size $n_{\mathrm{train}} = n_{\mathrm{calib}} \in \{500, 2000, 5000\}$,

Counterfactual prediction: simulation settings

- Confounded simulation setting adapted from [Yadlowsky et al.'18]
- Task: counterfactual prediction of $Y(1)$ for super population
- Identical population model
  - Covariates $X \sim \text{Unif}[0,1]^p$, $p = 4, 20$.
  - Confounder $U \,|\, _X \sim N((0, 1 + \frac{1}{2}(2.5X_1)^2),\ Y(1) = \beta^\top X + U$,
  - Sample size $n_{\text{train}} = n_{\text{calib}} \in \{500, 2000, 5000\}$,
- Different treatment mechanism for different $\Gamma \Rightarrow$ different observed distributions
  - Confounding level $\Gamma \in \{1.5, 2, 2.5, 3, 5\}$.
  - $e(x) = \text{logit}(x^\top \beta)$, $\beta = (-0.531, 0.126, -0.312, 0.018, 0, \cdots, 0) \in \mathbb{R}^p$.
  - $e(x,u) = \ell(x)\, \mathbb{1}\{|U| > t(x)\} + u(x)\, \mathbb{1}\{|U| \le t(x)\}$, satisfies marginal Γ-selection.
  - Treatment assignment $T_i \,|\, _{X_i, U_i} \sim \text{Bern}(e(X, U))$

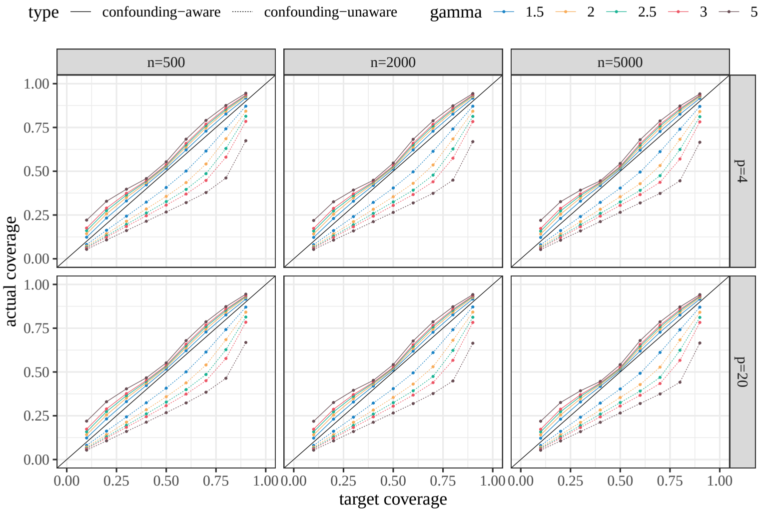
Counterfactual prediction: simulation settings

- Confounded simulation setting adapted from [Yadlowsky et al.'18]
- Task: counterfactual prediction of $Y(1)$ for super population
- Identical population model
  - Covariates $X \sim \mathrm{Unif}[0,1]^p$, $p = 4, 20$.
  - Confounder $U \mid_X \sim N((0, 1 + \frac{1}{2}(2.5X_1)^2)$, $Y(1) = \beta^\top X + U$,
  - Sample size $n_{\text{train}} = n_{\text{calib}} \in \{500, 2000, 5000\}$,
- Different treatment mechanism for different $\Gamma \Rightarrow$ different observed distributions
  - Confounding level $\Gamma \in \{1.5, 2, 2.5, 3, 5\}$.
  - $e(x) = \mathrm{logit}(x^\top \beta)$, $\beta = (-0.531, 0.126, -0.312, 0.018, 0, \cdots, 0) \in \mathbb{R}^p$.
  - $e(x,u) = \ell(x)\,\mathbb{1}\{|U| > t(x)\} + u(x)\,\mathbb{1}\{|U| \leq t(x)\}$, satisfies marginal Γ-selection.
  - Treatment assignment $T_i \mid_{X_i, U_i} \sim \mathrm{Bern}(e(X, U))$
- The setting tries to be 'adversarial'.

Evaluated Methods

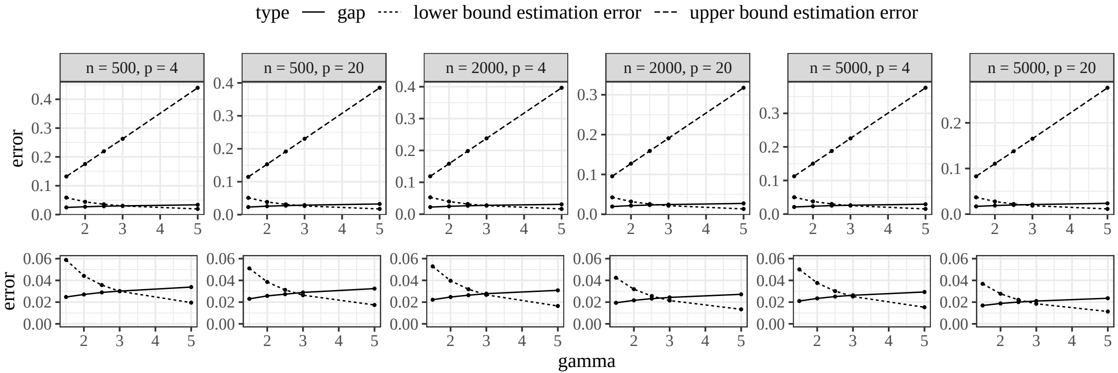
- Run $\alpha \in \{0.1, \ldots, 0.9\}$ and $\delta = 0.05$.
- Marginally valid procedure with $\widehat{e}(x)$ estimated from random forest on $\mathcal{D}_{\text{train}}$.
  - Coverage of target should be over $1 - \alpha$ averaged over all runs.
- PAC-type procedure with $\widehat{e}(x)$ from random forest on $\mathcal{D}_{\text{train}}$, lower bound by W-S-R inequality.
  - The 0.05 quantile of $N = 1000$ runs of empirical coverage on test samples should be over $1 - \alpha$.

Marginally valid procedure with estimated $\widehat{e}(x)$

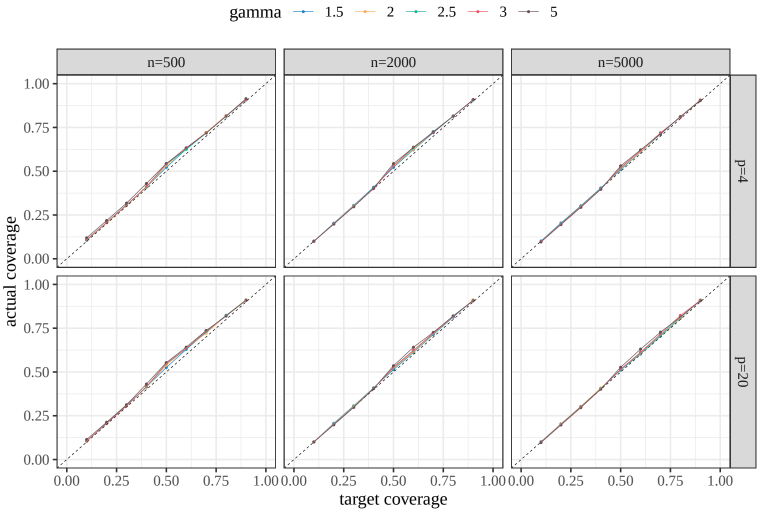


Marginally valid procedure with estimated $\widehat{e}(x)$

- Robustness to estimation error

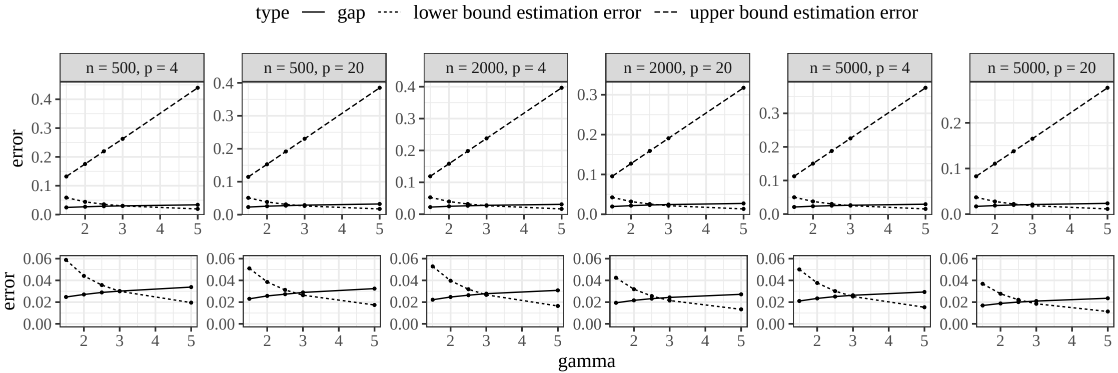


High-probability valid procedure with estimated $\widehat{e}(x)$



High-probability valid procedure with estimated $\widehat{e}(x)$

- Robustness to estimation error

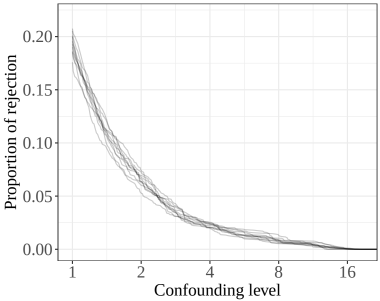


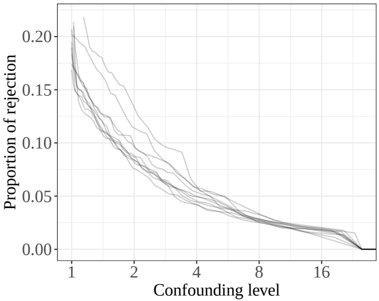
Sensitivity analysis on real dataset

- NSLM data from the ACIC data challenge.
- ITE on the treated: test sample is from $T = 1$.
- Test $H_0^-(\Gamma)$: $Y(1) \leq Y(0), \Gamma^* \leq \Gamma$ (find positive ITEs) and $H_0^+(\Gamma)$: $Y(1) \geq Y(0), \Gamma^* \leq \Gamma$ (find negative ITEs)
- Fix $\alpha = 0.1$ and $\delta = 0.05$.

Sensitivity analysis on real dataset

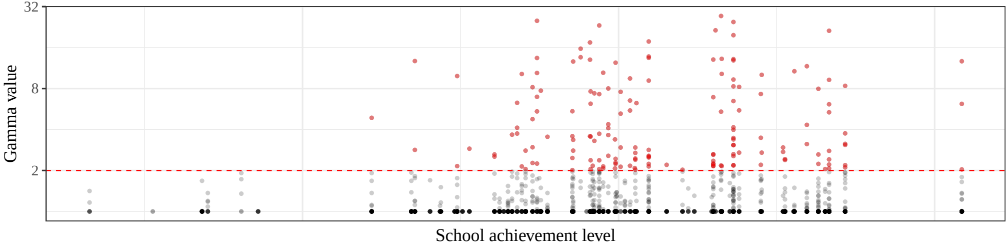
- Testing $H_0^-(\Gamma)$: $Y(1) \leq Y(0), \Gamma^* \leq \Gamma$ (find positive ITEs):





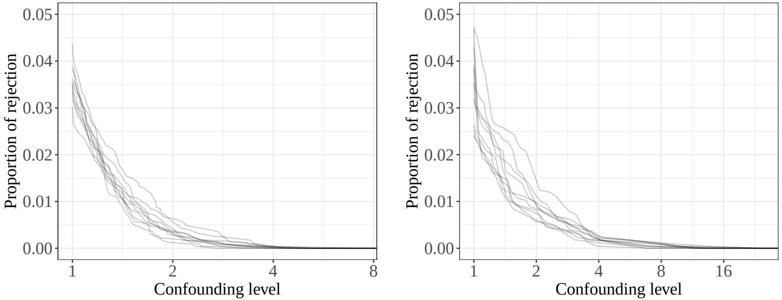
- about 20% of test treated units are identified as positive ITE at $\Gamma = 1$.
- some samples have large values of $\widehat{\Gamma}$.
- PAC-type (right) procedure has slightly higher power but less stable.

Γ-value versus school achievement level for positive ITE:



Sensitivity analysis on real dataset

- Testing $H_0^+(\Gamma)$: $Y(1) \geq Y(0), \Gamma^* \leq \Gamma$ (find negative ITEs):



- about 3.5% of test treated units are identified as negative ITE at $\Gamma = 1$.
- values of $\widehat{\Gamma}$ are much smaller than those for positive ITE.

https://arxiv.org/abs/2111.12161

Thanks!

Worst-case distribution function

- In the general robust prediction problem, the range of distributions is

$$\mathcal{P}(\mathbb{P}, \ell, u) = \left\{ \widetilde{\mathbb{P}} \colon \ell(x) \le \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(x, y) \le u(x) \ \ \mathbb{P}\text{-almost surely} \right\}.$$

Proposition

The worst-case c.d.f. for $\mathcal{P}(\mathbb{P}, \ell, u)$ is

$$F\big(t\,; \mathcal{P}(\mathbb{P}, \ell, u)\big) = \max\left\{ \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \le t\}}\,\ell(X)\big],\ 1 - \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) > t\}}\,u(X)\big] \right\},$$

which equals $G(t)$, so PAC-type procedure is tight for this identification set.

Worst-case distribution function: counterfactual inference

- For a super-population to be meaningful, it should agree with the observed: $\mathbb{P}^{\text{sup}}_{X,Y,T} = \mathbb{P}^{\text{obs}}_{X,Y,T}$.
- For example, to predict $Y(1) \mid T = 0$, the identification set is

$$\mathcal{P} = \left\{ \widetilde{\mathbb{P}} = \mathbb{P}^{\text{sup}}_{X,Y(1)\,|\,T=0} \colon\ \mathbb{P}^{\text{sup}}_{X,Y,T} = \mathbb{P}^{\text{obs}}_{X,Y,T} \text{ and satisfies } \Gamma \text{ selection} \right\}. \tag{4.1}$$

Proposition

The identification set $\mathcal{P}$ in (4.1) is equivalently characterized by

$$\mathcal{P}(\mathbb{P}, f, \ell_0, u_0) = \left\{ \widetilde{\mathbb{P}} \colon \frac{\mathrm{d}\widetilde{\mathbb{P}}_X}{\mathrm{d}\mathbb{P}_X}(x) = f(x),\ \ell_0(x) \le \frac{\mathrm{d}\widetilde{\mathbb{P}}_{Y(1)\,|\,X}}{\mathrm{d}\mathbb{P}_{Y(1)\,|\,X}}(y \mid x) \le u_0(x) \ \ \mathbb{P}\text{-almost surely} \right\}. \tag{4.2}$$

Writing $e(x) = \mathbb{P}^{\text{obs}}(T = 1 \mid X = x)$, $p_0 = \mathbb{P}^{\text{obs}}(T = 0)$ and $p_1 = \mathbb{P}^{\text{obs}}(T = 1)$, the bound functions are given by $f(x) = p_1(1 - e(x))/[p_0 \cdot e(x)]$, $\ell_0(x) = 1/\Gamma$ and $u_0(x) = \Gamma$.

- The X-distributional shift is identifiable, the $Y \mid X$ shift is bounded.

Worst-case distribution function: counterfactual inference

- We have the following characterization of worst-case distribution function.

Proposition

For each $t \in \mathbb{R}$, the worst-case distribution function of (4.2) is

$$F\big(t\,;\mathcal{P}(\mathbb{P}, f, \ell_0, u_0)\big) = \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\, w^*(X, Y)\big], \tag{4.3}$$

where the expectation is with respect to generic random variables $(X, Y) \sim \mathbb{P}$, and

$$w^*(x, y) = f(x) \cdot \big[\ell_0(x)\,\mathbb{1}_{\{V(x,y)<q(\tau(x)\,;x,\mathbb{P})\}} + \gamma_0(x)\,\mathbb{1}_{\{V(x,y)=q(\tau(x)\,;x,\mathbb{P})\}} + u_0(x)\,\mathbb{1}_{\{V(x,y)>q(\tau(x)\,;x,\mathbb{P})\}}\big].$$

Here $\tau(x) = (u_0(x) - 1)/(u_0(x) - \ell_0(x))$ and $\gamma_0(x)$ is chosen such that $\mathbb{E}[w^*(x, Y) \,|\, X = x] = f(x)$.

- In all counterfactual prediction problems, we have $\tau(x) = \tau$ for some constant τ.
- If $q(\tau; x, \mathbb{P})$ can be consistently estimated, the coverage can be tight.

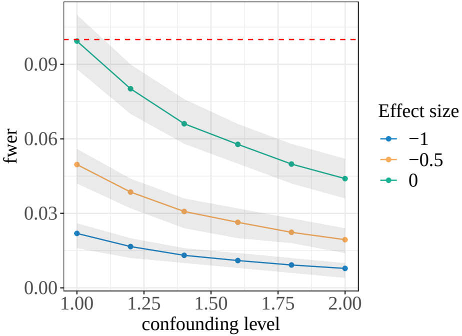
Simulation on sensitivity analysis

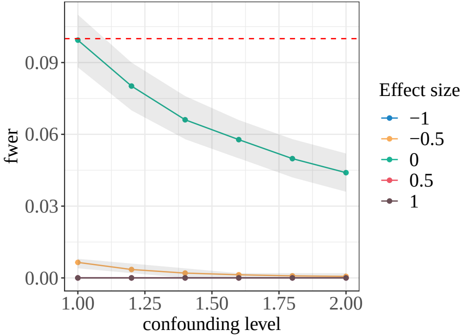
- Study ITE on the treated
- Data generating process
  - The sample super-population, $X \sim N(0, I_4)$, $U \sim N(0, 1)$ independent.
  - $Y(0) = X^\top \beta + U$, $T \sim \text{Bernoulli}(e(X, U))$, confounded.

  - Observe smaller $Y(0)$ in the control group than target.
  - Predictive interval for $Y(0)$ is shifted left by confounding, creating 'fake' effects.
- Actual ITE with sizes $a \in \{-1, -0.5, 0, 0.5, 1\}$
  - Case 1: fixed ITE $Y(1) = Y(0) + a$,
  - Case 2: confounded random ITE $Y(1) = Y(0) + a \cdot U$,

Simulation results: error control

- The proportion of $\widehat{\Gamma} \geq \Gamma^*$ and $Y(1) - Y(0) \leq 0$:





- Error rate is controlled.
- If ITE is random, it's conservative.