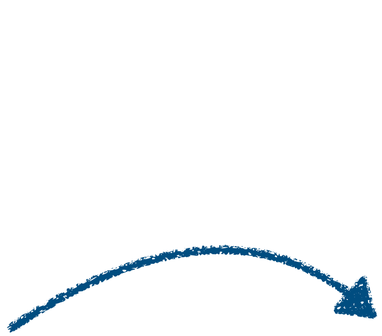
Conformal sensitivity analysis for individual treatment effects

Mingzhang Yin
University of Florida

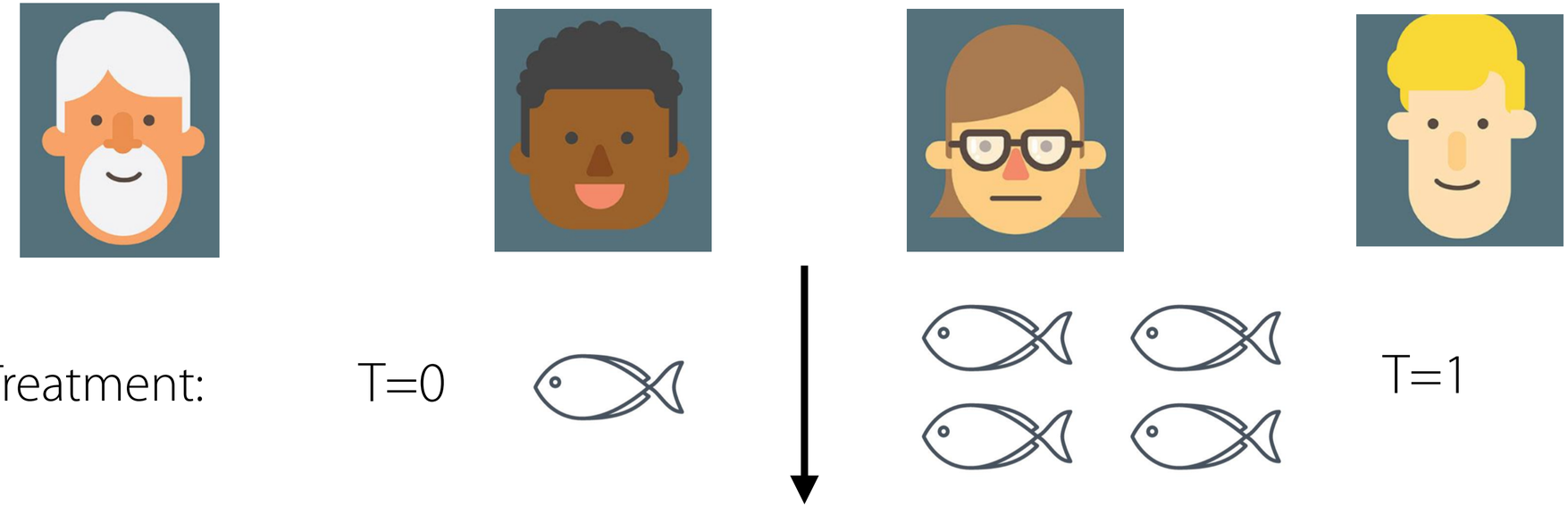
Joint work with Claudia Shi, Yixin Wang, and David M. Blei

An individual treatment effect

- What is the effect on my blood mercury level if I eat more sea fishes?
- If I took the vaccine, how much risk it would reduce for me?
- How much more spending would be, had Sam got a coupon?



Individual treatment effect (ITE)



Population distribution: $(X_i, T_i, Y_i(0), Y_i(1)) \sim P(X, T, Y(0), Y(1))$

Partial observation: (Covariates , Treatment , Outcome) = (X_i, T_i, Y_i)

Target estimand: The ITE $\tau_i = Y_i(1) - Y_i(0)$

Why ITE

- Popular alternative causal estimands

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

  - Conditional ATE (CATE): $\mathbb{E}[Y_i(1) - Y_i(0) \,|\, X_i]$

- ATE and CATE are coarse summary statistics

- ITE contains essential individual information for policy making

$= 1$

$\text{CATE} = 1$

$= 1.5$ $= -1$

$\text{CATE} = 1$

$= 25$ $= -5$

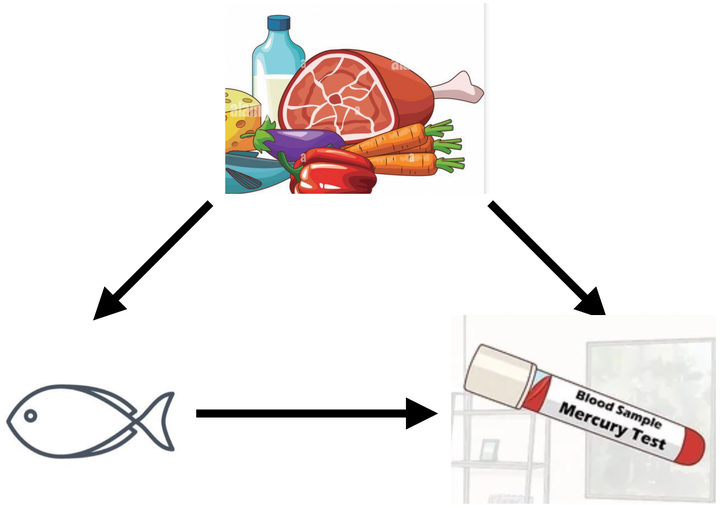
$\text{CATE} = 1$

Challenges in estimating the ITE

- Fundamental problem in causal inference (Holland 1986): for any individual unit, we can observe only one of Y(1) or Y(0); under SUTVA

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

- Unmeasured confounding: $(Y(0), Y(1)) \not\perp T \,|\, X.$



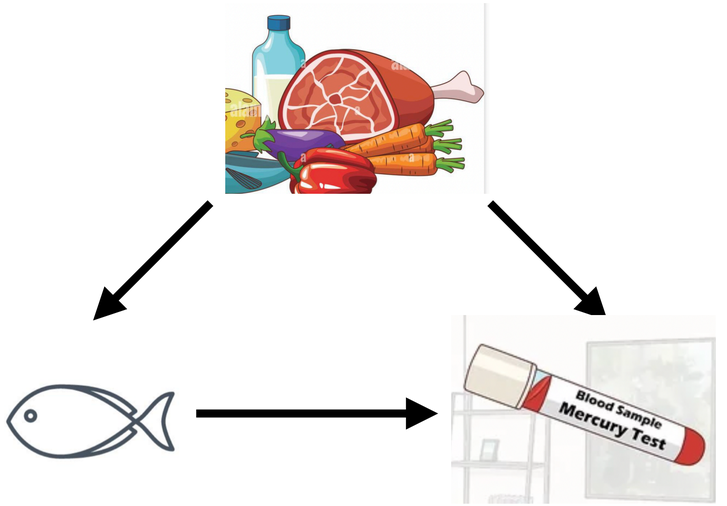
- Randomness of the ITE: the ITE $Y_i(1) - Y_i(0)$ is random even under a population distribution $P(X, T, Y(0), Y(1))$

Challenges in estimating the ITE

- Fundamental problem in causal inference (Holland 1986): for any individual unit, we can observe only one of Y(1) or Y(0); under SUTVA

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

- Unmeasured confounding: $(Y(0), Y(1)) \not\perp T \,|\, X.$



- Randomness of the ITE: the ITE $Y_i(1) - Y_i(0)$ is random even under a population distribution $P(X, T, Y(0), Y(1))$

The ITE is not point-identifiable

Previous work: under unconfoundedness [Lei & Candès, JRSS-B, 2021]

Under the assumption of unconfoundedness, construct a predictive set $\hat{C}(X)$, s.t.

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

with a pre-specified mis-coverage rate $\alpha \in (0,1)$

How to infer the ITE under the violation of unconfoundedness?

Existing literature

	Average treatment effect (ATE)	Conditional average treatment effect (CATE)	Individual treatment effect (ITE)
Primary analysis under unconfoundedness	Rosenbaum & Rubin (1983) Imbens & Rubin (1997) Pearl (2000)	Chipman et al. (2010) Wager & Athey (2018) Künzel et al. (2019)	Lei & Candès (2021)
Sensitivity analysis of unconfoundedness	Rosenbaum (2002) Imbens (2003) Qingyuan, Small & Bhattacharya (2019)	Yadlowsky et al. (2018) Nathan, Mao & Zhou (2019) Jesson et al. (2020)	

Outline

- Review: conformal inference
- Method: conformal sensitivity analysis (CSA)
- Extension: conformalized sharp sensitivity analysis (CSSA)
- Practical considerations
- Numerical experiments

Conformal inference

- **Setting:** observe exchangeable (e.g., i.i.d.) data points $\{(X_i, Y_i)\}_{i=1}^{n}$ from an unknown distribution $P(X, Y)$.

- **Problem:** for a new target data point $(X, Y) \sim P(X, Y)$, observe X but not Y.

- **Goal:** construct a predictive band $\hat{C}(X)$ that covers the unknown Y with statistical guarantees and minimal assumptions

Recipe of (split) conformal inference

Step 1: model fit

- Randomly split the observed data to a training set $\mathcal{I}_{tr}$ and a calibration set $\mathcal{I}_{cal}$
- Learn a predictive function on the training set, e.g., $\min_{\theta} \sum_{i \in \mathcal{I}_{tr}} (y_i - \hat{y}_i(x_i; \theta))^2$

Recipe of (split) conformal inference

Step2: calibration and prediction

- Compute a nonconformity score for all calibration data $\{V(X_i, Y_i), \infty\}_{i \in \mathcal{I}_{cal}}$ e.g., $V(x, y) = |y - \hat{y}(x)|$
- By the exchangeability of the calibration data and a new target data (X, Y),

$$\mathbb{P}(V(X, Y) \leq Q_{1-\alpha}(\sum_{i=1}^{n_{cal}} \frac{1}{n_{cal}+1}\delta_{V_i} + \frac{1}{n_{cal}+1}\delta_{\infty})\}) \geq 1 - \alpha$$

- Set $\hat{C}(X) = \{y : V(X, y) \leq Q_{1-\alpha}\}$, then $\mathbb{P}(Y \in \hat{C}(X)) \geq 1 - \alpha$

Conformal inference is useful

- Distribution free: arbitrary $\hat{y}(x)$
- Finite sample coverage guarantees: no need for $n \to \infty$

Conformal inference is useful

- Distribution free: arbitrary $\hat{y}(x)$
- Finite sample coverage guarantees: no need for $n \to \infty$
- The (almost) only assumption is exchangeability (or i.i.d.) of the calibration and test data —> can still be too much!

Conformal inference for counterfactual prediction

$$\textit{Training}: \quad (X_i, Y_i(t)) \sim p(X \mid T = t) \cdot p(Y(t) \mid X, T = t), \quad i \in \{1, \cdots, n\};$$

$$\textit{Target}: \quad (X, Y(t)) \sim p(X) \cdot p(Y(t) \mid X)$$

Conformal inference for counterfactual prediction

$$Training: \quad (X_i, Y_i(t)) \sim p(X \mid T = t) \cdot p(Y(t) \mid X, T = t), \quad i \in \{1, \cdots, n\};$$

$$Target: \quad (X, Y(t)) \sim p(X) \cdot p(Y(t) \mid X)$$

Randomized controlled trials	$p(X) = p(X \mid T = t)$ $p(Y(t) \mid X) = p(Y(t) \mid X, T = t)$	Train/target are i.i.d

Conformal inference for counterfactual prediction

$$\textit{Training}: \quad (X_i, Y_i(t)) \sim p(X \mid T = t) \cdot p(Y(t) \mid X, T = t), \quad i \in \{1, \cdots, n\};$$

$$\textit{Target}: \quad (X, Y(t)) \sim p(X) \cdot p(Y(t) \mid X)$$

Randomized controlled trials	$p(X) = p(X \mid T = t)$ $p(Y(t) \mid X) = p(Y(t) \mid X, T = t)$	Train/target are i.i.d
Observational study under unconfoundedness	$p(X) \neq p(X \mid T = t)$ $p(Y(t) \mid X) = p(Y(t) \mid X, T = t)$	Covariate shift
Observational study w/o unconfoundedness	$p(X) \neq p(X \mid T = t)$ $p(Y(t) \mid X) \neq p(Y(t) \mid X, T = t)$	General distribution shift

Not exchangeable!

Outline

- Review: conformal inference
- Method: conformal sensitivity analysis (CSA)
- Extension: conformalized sharp sensitivity analysis (CSSA)
- Practical considerations
- Numerical experiments

Weighted conformal inference [Tibshirani et al. (2020)]

- Idea: recover the exchangeability by weighting
- The conformal weights are the density ratios of training and target distributions

$$w_t(x, y) = \frac{p(X = x)p(Y(t) = y \,|\, X = x)}{p(X = x \,|\, T = t)p(Y(t) = y \,|\, X = x, T = t)}$$

- Set $\{p_1^t, \cdots, p_{n+1}^t\}$ as the normalized $\{w_t(X_1, Y_1), \cdots, w_t(X_n, Y_n), w_t(X, y)\}$
- A valid predictive interval $\hat{C}_t(X) = \{y : V(X, y) \leq Q_{1-\alpha}(\sum_{i=1}^{n} p_i^t \delta_{V_i} + p_{n+1}^t \delta_\infty)\}$

Non-identifiability of weights under confounding

$$w_t(x,y) = \frac{p(X=x)\; p(Y(t)=y \mid X=x)}{p(X=x \mid T=t)\; p(Y(t)=y \mid X=x, T=t)}$$

- The red ratio is non-identifiable under unmeasured confounding

$$\frac{p(Y(t) \mid X)}{p(Y(t) \mid X, T=t)} = p(T=t \mid X) + \frac{p(Y(t) \mid X, T=1-t)}{p(Y(t) \mid X, T=t)} p(T=1-t \mid X)$$

- Weighted conformal prediction is not directly applicable to general distribution shift

- Need assumptions to quantify the non-identifiable ratio term

Sensitivity model

- Sensitivity model is a relaxation of unconfoundedness assumption
- Adopt the marginal sensitivity model (MSM) [Rosenbaum (2003), Tan (2006)]
- Confounding is the difference between selection score $s_t(x, y) = p(T = 1 \mid X = x, Y(t) = y$ propensity score $e(x) = p(T = 1 \mid X = x)$
- Assumption MSM(Γ): under the population distribution

$$1/\Gamma \leq OR(s_t(x, y), e(x)) = \frac{s_t(x, y)/(1 - s_t(x, y))}{e(x)/(1 - e(x))} \leq \Gamma$$

- Sensitivity parameter $\Gamma \geq 1$ measures the confounding strength

Connecting sensitivity model and conformal inference

- Connect the conformal weights to the MSM

$$\mathrm{OR}(s_t(x,y), e(x)) = \frac{p(Y(t)=y \mid X=x, T=1)}{p(Y(t)=y \mid X=x, T=0)}$$

- Derived from Tukey's factorization (Treatment assignment <-> Outcome)

  [Brook, 1964; Franks, Airoldi, Rubin, 2016]

$$p(T=1 \mid X, Y(t)) = \frac{p(Y(t) \mid X, T=1)p(T=1 \mid X)}{p(Y(t) \mid X)}$$
$$= 1/\left(1 + \frac{1-e(X)}{e(X)} \frac{p(Y(t) \mid X, T=0)}{p(Y(t) \mid X, T=1)}\right)$$

Connecting sensitivity model and conformal inference

- Connect the conformal weights to the MSM by Tukey's factorization [Brook, 1964]

$$\mathrm{OR}(s_t(x,y), e(x)) = \frac{p(Y(t)=y \,|\, X=x, T=1)}{p(Y(t)=y \,|\, X=x, T=0)}$$

- The confounding strength Γ is reflected in the range of conformal weights

Proposition (weights uncertainty)

$$\underbrace{\left(1+\frac{1}{\Gamma}\left(\frac{1-e(x)}{e(x)}\right)^{2t-1}\right)p(T=t)}_{w_{lo}^{\Gamma}(x)} \leq w_t(x,y) \leq \underbrace{\left(1+\Gamma\left(\frac{1-e(x)}{e(x)}\right)^{2t-1}\right)p(T=t)}_{w_{hi}^{\Gamma}(x)}$$

Connecting conformal weights to predictive set

- Find a valid predictive set for any $P(X, T, Y(0), Y(1))$ consistent with the MSM
- $\Leftrightarrow$ Find the union of predictive sets from each sensitivity model allowed by MSM
- $\Leftrightarrow$ Solve a constrained quantile optimization

$$\max_{w_{1:n+1}} \quad Q_{1-\alpha}\Big(\sum_{i=1}^{n} p_i \delta_{V_i} + p_{n+1}\delta_{\infty}\Big).$$

$$\text{subject to} \quad p_i = \frac{w_i}{\sum_{i=1}^{n+1} w_i}, \;\; 1 \le i \le n+1$$

$$w_{lo}^{\Gamma}(X_i) \le w_i \le w_{hi}^{\Gamma}(X_i), \;\; 1 \le i \le n, \;\; w_{lo}^{\Gamma}(X) \le w_{n+1} \le w_{hi}^{\Gamma}(X)$$

Validity of the predictive sets

- With optimal objective value $\hat{Q}^{\Gamma}$, the predictive set under MSM(Γ) is

$$\hat{C}_t^{\Gamma}(x) = \left\{ y \in \mathbb{R} : V(x, y) \leq \hat{Q}^{\Gamma} \right\}$$

Theorem (Yin, Shi, Wang, Blei, '22)

For any $P(X, T, Y(0), Y(1))$ consistent with the MSM with parameter Γ, under SUTVA and strong overlapping assumptions $\eta < e(X) < 1 - \eta,\ 0 < \eta < 0.5$

- For a known $e(X)$, the coverage $\in [1 - \alpha,\ 1 - \alpha + (\Gamma/\eta)/(n + \Gamma/\eta)]$
- For an estimated $\hat{e}(X)$, the coverage $\in [1 - \alpha - \Delta,\ 1 - \alpha + (\Gamma/\eta)/(n + \Gamma/\eta) + \Delta]$

$$\Delta = \frac{\Gamma}{2} p(T = t) \mathbb{E}_{x \sim p(X|T=t)} \left| \frac{1}{\hat{e}(x)^t (1 - \hat{e}(x))^{1-t}} - \frac{1}{e(x)^t (1 - e(x))^{1-t}} \right|$$

Solving quantile optimization

$$\max_{w_{1:n+1}} \quad Q_{1-\alpha}\Big(\sum_{i=1}^{n} p_i \delta_{V_i} + p_{n+1}\delta_{\infty}\Big).$$

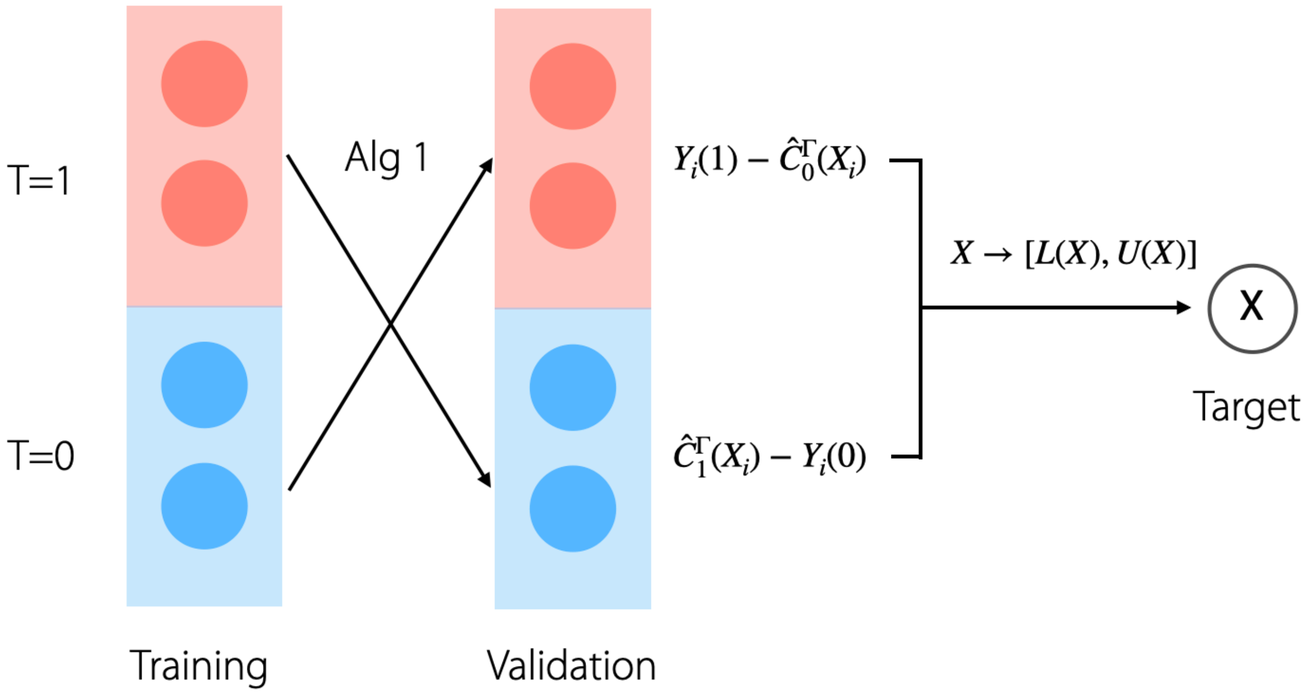
$$\text{subject to} \quad p_i = \frac{w_i}{\sum_{i=1}^{n+1} w_i}, \quad 1 \le i \le n+1$$

$$w_{lo}^{\Gamma}(X_i) \le w_i \le w_{hi}^{\Gamma}(X_i), \ 1 \le i \le n, \quad w_{lo}^{\Gamma}(X) \le w_{n+1} \le w_{hi}^{\Gamma}(X)$$

- Optimal objective value $\hat{Q}^{\Gamma} = V_{\hat{k}}$
- Proposition: $\hat{k} = \max\Big\{k : \textbf{for } k \le j \le n,\ w_j = w_{hi}^{\Gamma}(X_j);\ \textbf{for } j < k, w_j = w_{lo}^{\Gamma}(X_j);\ \sum_{j=k}^{n+1} p_j \ge \alpha\Big\}$
- Line search: sort V_i; initialize w_i as $w_{lo}^{\Gamma}(X_i)$, flip w_i from $w_{lo}^{\Gamma}(X_i)$ to $w_{hi}^{\Gamma}(X_i)$ for $i = n+1, n, \cdots, 1$, until $\sum_{i=k}^{n+1} p_i \ge \alpha$
- Computational cost is close to the optimal rate $\mathcal{O}(n)$ for relatively small α

Estimating ITE with no observed outcomes

- Consider the ITE for units with both potential outcomes unobserved
- Adopt a nested approach proposed by Lei & Candès (2021)



- Split data to training/validation sets
- Apply Alg. 1 to the group T=t in the training set and the group T=1-t in the validation set, for t=0,1
- Fit a predictor $X \to [L(X), U(X)]$ with data $(X_i, \hat{C}(X_i))$ in the validation set
- Construct predictive sets for the ITEs of the target units

Outline

- Review: conformal inference
- Method: conformal sensitivity analysis (CSA)
- Extension: conformalized sharp sensitivity analysis (CSSA)
- Practical considerations
- Numerical experiments

Sharpness of sensitivity models

- Sharpness: the sensitivity models should be data compatible
- Define sharp MSM(Γ)

$$\mathscr{E}_t^*(\Gamma) = \{s_t(x,y) : \text{OR}(s_t, e) \in [1/\Gamma, \Gamma], \int p^{(s_t)}(Y(t) = y \,|\, X = x, T = 1-t)dy = 1\}$$

- The integral constraint is equivalent to the constraint
  $\mathbb{E}[T/s_1(X, Y(1)) \,|\, X] = 1$ in Dorn & Guo [2021]

- Induce covariate balancing with user-specified covariates function $g(X_i)$

$$\mathbb{E}[\frac{g(X_i)T_i}{s_1(X_i, Y_i(1))}] = \mathbb{E}[\frac{g(X_i)(1-T_i)}{1-s_0(X_i, Y_i(0))}] = \mathbb{E}[g(X_i)]$$

  e.g. $g(X_i) = \hat{e}(X_i)$.

Improving sharpness by covariate balancing

$$\max_{w_{1:n+1}} \quad Q_{1-\alpha}\Big(\sum_{i=1}^{n} p_i \delta_{V_i} + p_{n+1}\delta_{\infty}\Big).$$

$$\text{subject to} \quad p_i = \frac{w_i}{\sum_{i=1}^{n+1} w_i}, \;\; 1 \le i \le n+1$$

$$w_{lo}^{\Gamma}(X_i) \le w_i \le w_{hi}^{\Gamma}(X_i), \; 1 \le i \le n, \;\; w_{lo}^{\Gamma}(X) \le w_{n+1} \le w_{hi}^{\Gamma}(X)$$

$$\frac{1}{N_t}\sum_{i:T_i=t} g_k(X_i) w_i^{\Gamma} = \frac{1}{N}\sum_{i=1}^{N} \frac{T_i^t (1-T_i)^{1-t}}{\hat{e}(X_i)^t (1-\hat{e}(X_i))^{1-t}} g_k(X_i), \;\; 1 \le k \le K$$

- Incorporate covariate balancing conditions to the quantile optimization
- Obtain predictive sets with reduced size and tighter coverage
- Remaining challenge: the objective function is discontinuous

Implementation of CSSA

$$\max_{w_{1:n+1}} \frac{\sum_{i=J}^{n+1} w_i}{\sum_{i=1}^{n+1} w_i}$$

$$\text{subject to} \quad w_{lo}^{\Gamma}(X_i) \le w_i \le w_{hi}^{\Gamma}(X_i),\ 1 \le i \le n,\ \ w_{lo}^{\Gamma}(X) \le w_{n+1} \le w_{hi}^{\Gamma}(X)$$

$$\frac{1}{N_t} \sum_{i:T_i=t} g_k(X_i) w_i = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i^t (1-T_i)^{1-t}}{\hat{e}(X_i)^t (1-\hat{e}(X_i))^{1-t}} g_k(X_i),\ \ 1 \le k \le K$$

- Sort $V_1 \le V_2 \le , \cdots, \le V_{n+1}$
- Solve a sequence of nonlinear programming with linear constraints for $J = n+1, n, \cdots, 1$ (e.g. Python package scipy and R package nloptr)
- Stop at the first time when the objective > α; return V_J

Outline

- Review: conformal inference
- Method: conformal sensitivity analysis (CSA)
- Extension: conformalized sharp sensitivity analysis (CSSA)
- Practical considerations
- Numerical experiments

Calibration of the sensitivity parameter

- Sensitivity parameter Γ reflects the plausible range of OR$(s_t(X, Y(t)), e(X))$
- In general, knowing the magnitude of Γ requires domain knowledge
- Reference information can be obtained from data by regarding Y(t) as a covariate
- Compute the OR by leaving one covariate out

$$\Gamma_{ij} = OR(e(X_i), e((X_{\setminus j})_i))$$

Evaluating the ITE estimation

- Evaluating the coverage of a predictive set requires $Y \sim p(Y(t) \mid X, T = 1 - t)$
- Construct a counterfactual distribution by exponential tilting

$$p(Y(t) = y \mid X = x, T = 1 - t) = \frac{e^{f(y|x)}}{M(x)} p(Y(t) = y \mid X = x, T = t)$$

- The counterfactual distribution satisfies MSM if $e^{f(y|x)}/M(x) \in [1/\Gamma, \Gamma]$
- Use rejection sampling to sample from $p(Y(t) \mid X, T = 1 - t)$ with the proposal distribution $p(Y(t) \mid X, T = t)$

Outline

- Review: conformal inference
- Method: conformal sensitivity analysis (CSA)
- Extension: conformalized sharp sensitivity analysis (CSSA)
- Practical considerations
- Numerical experiments

Synthetic data

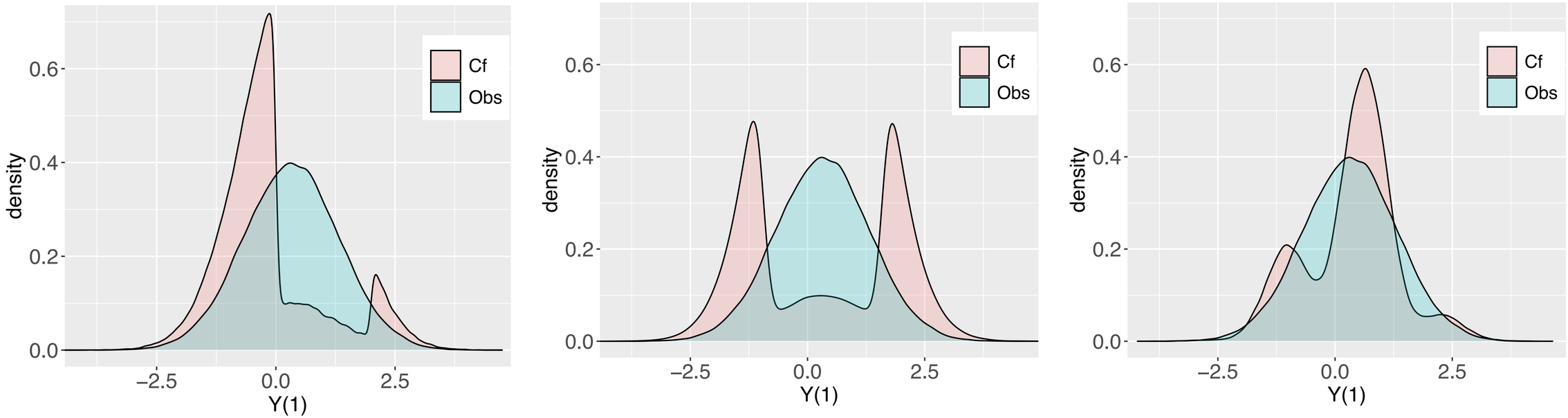
$$Y_i(1) = \mathbb{E}[Y_i(1) \,|\, T_i = 1, X_i] + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, \sigma^2);$$

$$\mathbb{E}[Y_i(1) \,|\, T_i = 1, X_i] = f(X_{i1}) f(X_{i2}), \;\; f(x) = \frac{2}{1 + \exp(-5(x - 0.5))}.$$

$$e(X_i) = \tfrac{1}{4}(1 + \beta_{2,4}(1 - X_{i1}))$$

- n = 3000, p = 20
- Homoscedastic case: σ = 1; Heteroscedastic case: σ ~ Unif(1/2, 3/2)
- Consider conformal inference with predictive model as the mean regression (CSA-M) and quantile regression (CSA-Q)

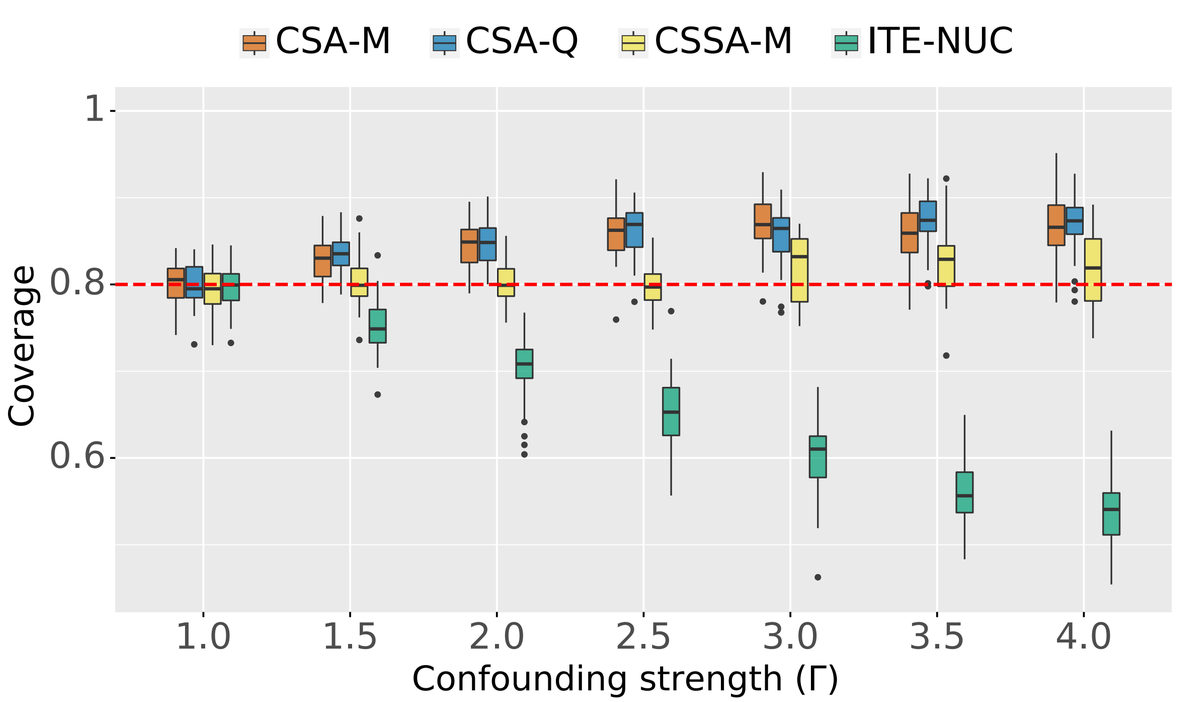
Synthetic data

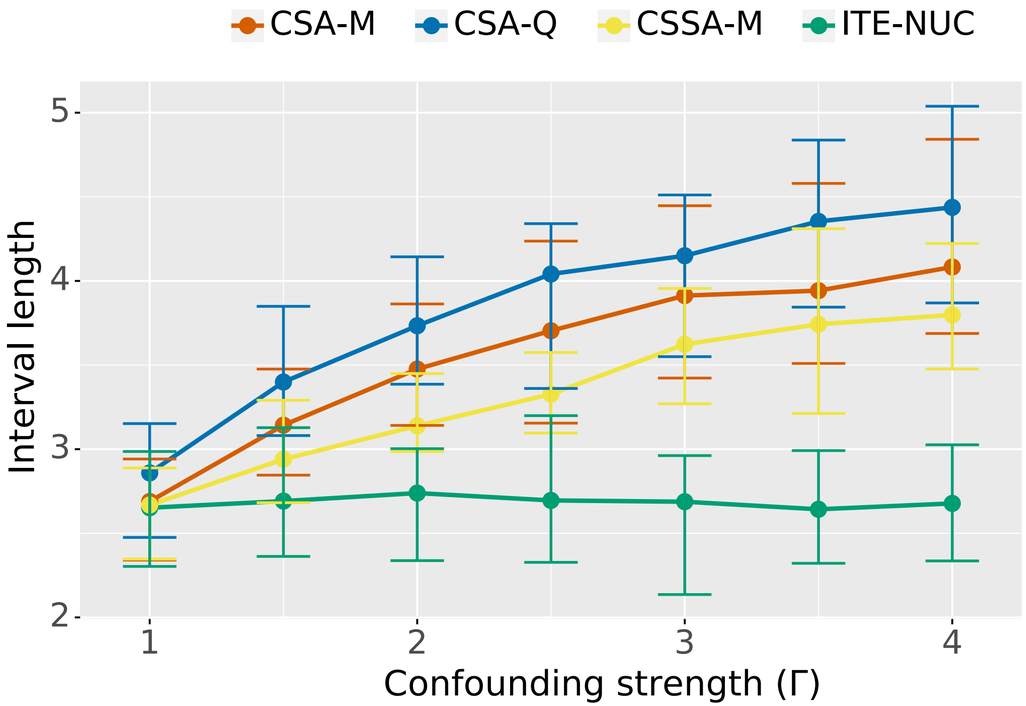


- Distribution of p(Y(1) | T=1, X) and plausible examples of p(Y(1) | T=0, X)
- p(Y(1) | T=0, X) is compatible with an MSM with Γ= 4
- The nonparametric nature of MSM has high functional flexibility

Synthetic data

Homoscedastic, target coverage rate 1- α = 0.8





- CSA and CSSA has valid coverage across confounding strength
- CSSA improves sharpness and has smaller predictive set

ITE-NUC assumes unconfoundedness [Lei & Candès 2021]

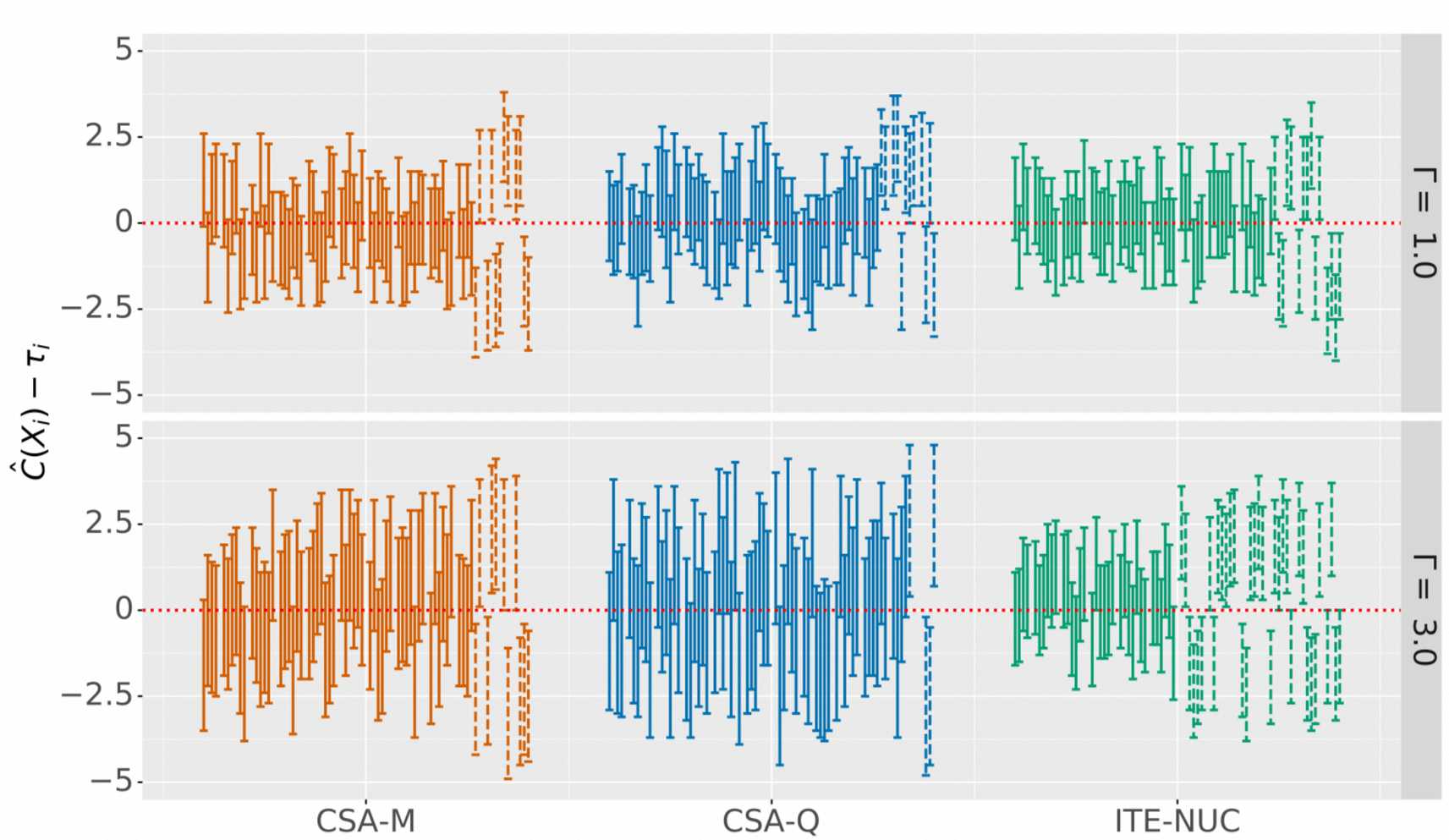
Synthetic data

Homoscedastic, target coverage rate 1- α = 0.8

	Γ	CSA	CSSA	Boot.Sens.	BART	ITE-NUC
Homosc.	1.0	0.80	0.80	0.03	0.62	0.80
	1.5	0.83	0.80	0.16	0.58	0.75
	2.0	0.85	0.80	0.24	0.54	0.69
	3.0	0.87	0.82	0.32	0.49	0.60
	4.0	0.87	0.83	0.37	0.46	0.54

- Bootstrap sensitivity analysis [Zhao et. al. 2018] and BART [Chipman et. al. 2010] are designed for the ATE and CATE estimation respectively
- ITE estimation and sensitivity analysis are necessary

Synthetic data

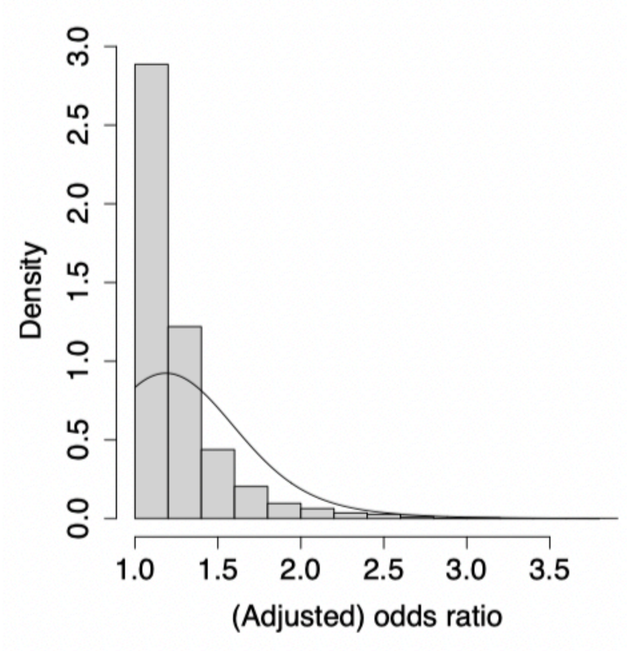


Each panel contains the predictive set $\hat{C}^{\Gamma}(X_i) - \tau_i$ for 70 randomly sampled individuals

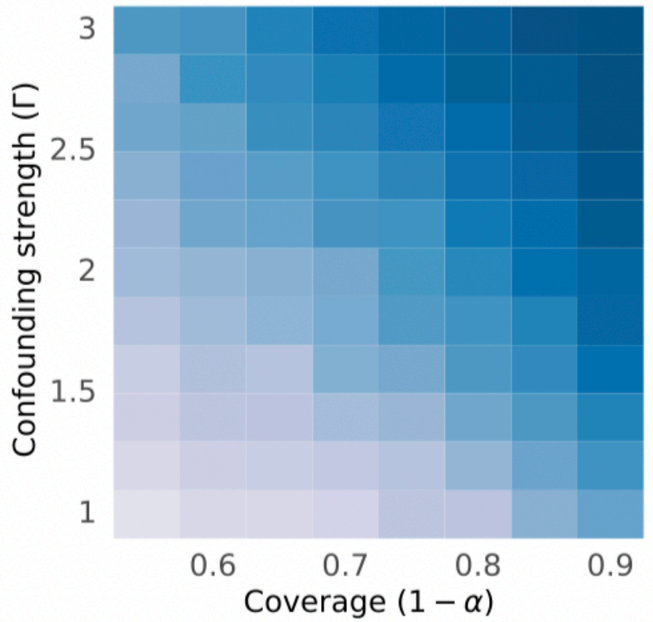
National Health and Nutrition Examination Survey (2013-2014)

- Observed data $\{X_i, Y_i, T_i\}_{i=1'}^{N}$ N = 1107
- Covariates X: demographics and health conditions
- Treatment T: high fish consumption (T=1, >= 12 servings of fish or shellfish in a previous month)
- Outcome Y: blood mercury level (ug/L) in logarithmic scale

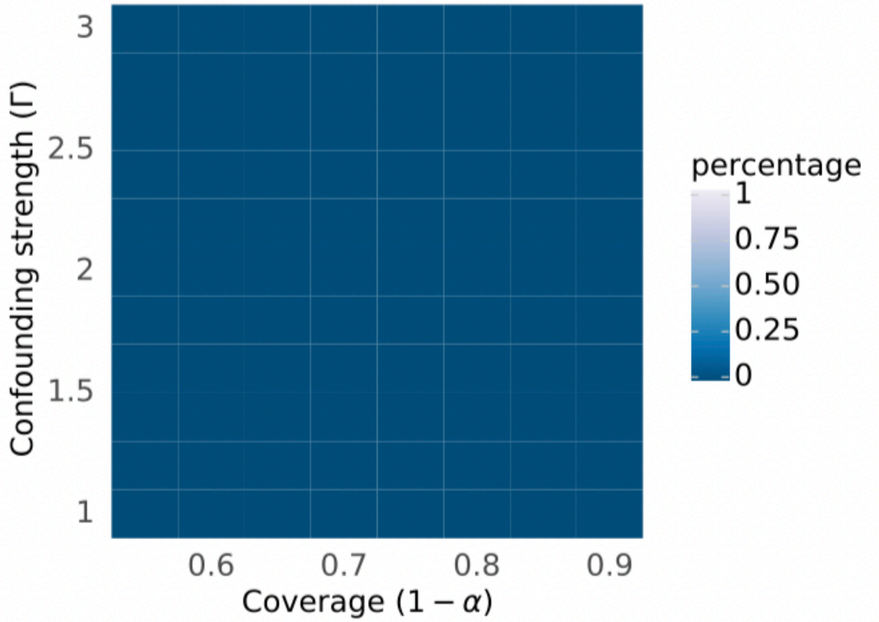
National Health and Nutrition Examination Survey



(a) Reference information for Γ

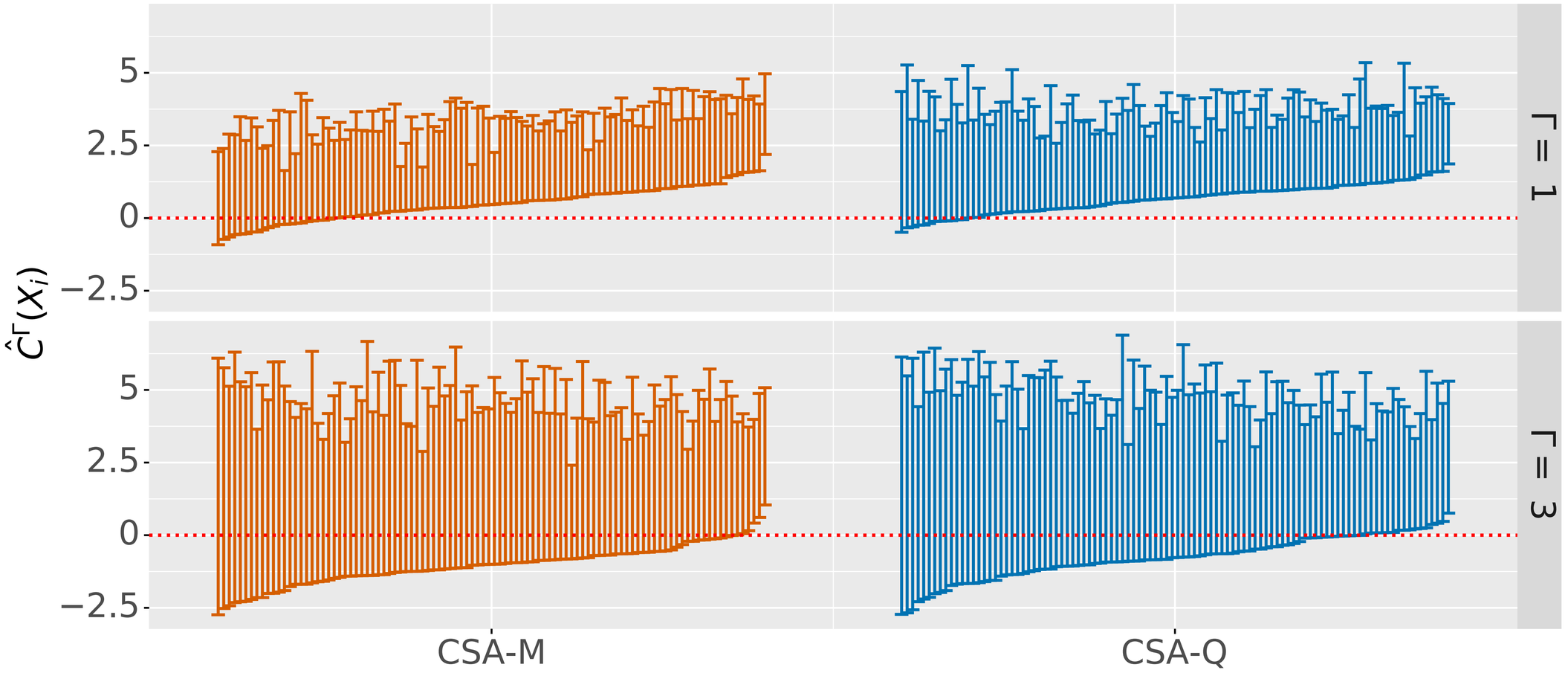


(b) Fraction of positive intervals (L(X) > 0)



(c) Fraction of negative intervals (U(X)<0)

National Health and Nutrition Examination Survey (2013-2014)



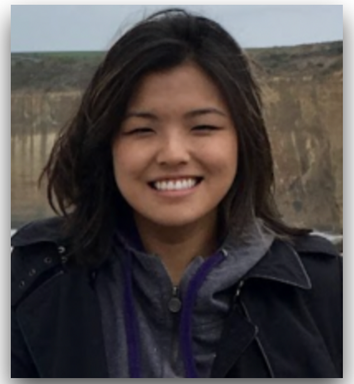
The ITE estimation under two confounding strengths. Γ is the level of robustness.

Takeaways and future direction

- CSA and CSSA estimate the ITE under unmeasured confounding
- The coverage guarantees are distribution-free, nonparametric and finite-sample
- Application in personalized medicine with individual heterogeneity?
- Make it actionable: policy evaluation and learning, personalization and targeting, causal recommendation?

Thank you!

Collaborators



Paper: **Yin**, Shi, Wang & Blei, "Conformal Sensitivity Analysis for Individual Treatment Effects", Journal of the American Statistical Association, 2022.

Contact: m.yin@ufl.edu