Introduction of Conformal Inference

by Weihao LI

12th April 2023

Outline

Objective of conformal inference

- Informally, we want to build confidence interval without any distributional assumption.

Objective of conformal inference

- Informally, we want to build confidence interval without any distributional assumption.
- Consider i.i.d. regression data

$$Z_1, \ldots, Z_n \sim P,$$

where each $Z_i = (X_i, Y_i)$ is a random variable in $\mathbb{R}^d \times \mathbb{R}$, comprised of a response variable Y_i and a d-dimensional vector of features (or predictors, or covariates) $X_i = (X_i(1), \ldots, X_i(d))$.

Objective of conformal inference

- Informally, we want to build confidence interval without any distributional assumption.
- Consider i.i.d. regression data

$$Z_1, \ldots, Z_n \sim P,$$

where each $Z_i = (X_i, Y_i)$ is a random variable in $\mathbb{R}^d \times \mathbb{R}$, comprised of a response variable Y_i and a d-dimensional vector of features (or predictors, or covariates) $X_i = (X_i(1), \ldots, X_i(d))$.
- Constructing a prediction interval $C \subseteq \mathbb{R}^d \times \mathbb{R}$ based on $Z_1, \ldots, Z_n$ with the property that

$$\mathbb{P}\left(Y_{n+1} \in C\left(X_{n+1}\right)\right) \geq 1 - \alpha \quad (1)$$

A simple exercise

Exercise:

Suppose we have positive i.i.d random variables $R_1, \cdots, R_n, R_{n+1}$. Let $Q_{1-\alpha}$ denote the empirical $1-\alpha$ quantile for $\{R_1, \cdots, R_n\}$, what is approximate value of $\mathbb{P}(R_{n+1} \leq Q_{1-\alpha})$?

A simple exercise

Exercise:

Suppose we have positive i.i.d random variables $R_1, \cdots, R_n, R_{n+1}$. Let $Q_{1-\alpha}$ denote the empirical $1-\alpha$ quantile for $\{R_1, \cdots, R_n\}$, what is approximate value of $\mathbb{P}(R_{n+1} \leq Q_{1-\alpha})$?

Answer: Consider ordered statisitics for $n+1$ R_i

$$R_{(1)}, R_{(2)}, R_{(3)}, \cdots, R_{(n)}, R_{(n+1)}$$

$$\mathbb{P}(R_{n+1} \leq Q_{1-\alpha}) \approx \mathbb{P}(R_{n+1} \text{ rank lower than } (1-\alpha)(n+1)) \approx 1-\alpha \quad (2)$$

A simple exercise

Exercise:

Suppose we have positive i.i.d random variables $R_1, \cdots, R_n, R_{n+1}$. Let $Q_{1-\alpha}$ denote the empirical $1-\alpha$ quantile for $\{R_1, \cdots, R_n\}$, what is approximate value of $\mathbb{P}(R_{n+1} \leq Q_{1-\alpha})$?
Answer: Consider ordered statisitics for $n+1$ R_i

$$R_{(1)}, R_{(2)}, R_{(3)}, \cdots, R_{(n)}, R_{(n+1)}$$

$$\mathbb{P}(R_{n+1} \leq Q_{1-\alpha}) \approx \mathbb{P}(R_{n+1} \text{ rank lower than } (1-\alpha)(n+1)) \approx 1-\alpha \quad (2)$$

Inverse the empirical CDF: $\text{Quantile}\left(1-\alpha; \frac{1}{n}\sum_{i=1}^{n} \delta_{R_i}\right)$
Remark: relax i.i.d to exchangeable.

Construction of conformal prediction set

- Suppose we have estimator $\hat{\mu} : \mathcal{X} \to \mathcal{Y}$ independent of our data $(X_1, Y_1), \cdots (X_n, Y_n), (X_{n+1}, Y_{n+1}$ unknown)

Construction of conformal prediction set

- Suppose we have estimator $\hat{\mu} : \mathcal{X} \to \mathcal{Y}$ independent of our data $(X_1, Y_1), \cdots (X_n, Y_n), (X_{n+1}, Y_{n+1}$ unknown)
- Apply $\hat{\mu}$ to $n+1$ data points, $R_i := |Y_i - \hat{\mu}(X_i)|$, then we know those $R_1 \cdots R_n, R_{n+1}$ are i.i.d.

Construction of conformal prediction set

- Suppose we have estimator $\hat{\mu} : \mathcal{X} \to \mathcal{Y}$ independent of our data $(X_1, Y_1), \cdots (X_n, Y_n), (X_{n+1}, Y_{n+1}$ unknown)
- Apply $\hat{\mu}$ to $n+1$ data points, $R_i := |Y_i - \hat{\mu}(X_i)|$, then we know those $R_1 \cdots R_n, R_{n+1}$ are i.i.d.
- By simple exercise: Let $Q_{1-\alpha}$ denote $1-\alpha$ quantile for $\{R_1, \cdots, R_n\}$

$$1-\alpha \approx \mathbb{P}(R_{n+1} \leq Q_{1-\alpha}) = \mathbb{P}(|Y_{n+1} - \hat{\mu}(X_{n+1})| \leq Q_{1-\alpha})$$

$$\Rightarrow \mathbb{P}(Y_{n+1} \in [\hat{\mu}(X_{n+1}) - Q_{1-\alpha}, \hat{\mu}(X_{n+1}) + Q_{1-\alpha}]) \approx 1-\alpha$$

Construction of conformal prediction set

- Suppose we have estimator $\hat{\mu} : \mathcal{X} \to \mathcal{Y}$ independent of our data $(X_1, Y_1), \cdots (X_n, Y_n), (X_{n+1}, Y_{n+1}$ unknown)
- Apply $\hat{\mu}$ to $n+1$ data points, $R_i := |Y_i - \hat{\mu}(X_i)|$, then we know those $R_1 \cdots R_n, R_{n+1}$ are i.i.d.
- By simple exercise: Let $Q_{1-\alpha}$ denote $1-\alpha$ quantile for $\{R_1, \cdots, R_n\}$

$$1-\alpha \approx \mathbb{P}(R_{n+1} \leq Q_{1-\alpha}) = \mathbb{P}(|Y_{n+1} - \hat{\mu}(X_{n+1})| \leq Q_{1-\alpha})$$

$$\Rightarrow \mathbb{P}(Y_{n+1} \in [\hat{\mu}(X_{n+1}) - Q_{1-\alpha}, \hat{\mu}(X_{n+1}) + Q_{1-\alpha}]) \approx 1-\alpha$$

- [LGR+18] Define $C_{\text{split}}(x) = [\widehat{\mu}(x) - Q_{1-\alpha}, \widehat{\mu}(x) + Q_{1-\alpha}]$

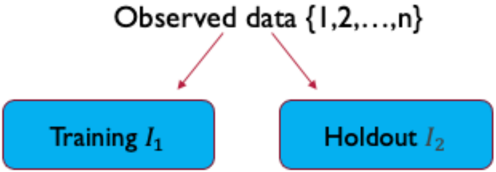
$$1-\alpha \leq \mathbb{P}\left(Y_{n+1} \in C_{\text{split}}(X_{n+1})\right) \leq 1-\alpha + \frac{2}{n+2}$$

Remark on conformal prediction set

- How to get $\hat{\mu}$:

Remark on conformal prediction set

- How to get $\hat{\mu}$:



For **any** regression algorithm $\mathcal{A}$

$\widehat{\mu} = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$

$R_i = \left|Y_i - \widehat{\mu}(X_i)\right|, i \in \mathcal{I}_2$

$Q_{1-\alpha}$ is $(1-\alpha)(|\mathcal{I}_2| + 1)$ quantile of $\{R_i : i \in \mathcal{I}_2\}$

Return $C_{\text{split}}(x) = [\widehat{\mu}(x) - Q_{1-\alpha}, \widehat{\mu}(x) + Q_{1-\alpha}]$, for all $x \in \mathbb{R}^d$

Remark on conformal prediction set

$$C_{\text{split}}(X_{n+1}) = [\widehat{\mu}(X_{n+1}) - Q_{1-\alpha}, \widehat{\mu}(X_{n+1}) + Q_{1-\alpha}]$$

- Marginal coverage: $\mathbb{P}$ is over joint distribution of (X, Y),i.e. $\mathbb{P}_{(X,Y)}$

$$1 - \alpha \leq \mathbb{P}_{(X,Y)}\left(Y_{n+1} \in C_{\text{split}}(X_{n+1})\right) \leq 1 - \alpha + \frac{2}{|\mathcal{I}_2| + 2}$$

Remark on conformal prediction set

$$C_{\text{split}}(X_{n+1}) = [\widehat{\mu}(X_{n+1}) - Q_{1-\alpha}, \widehat{\mu}(X_{n+1}) + Q_{1-\alpha}]$$

- Marginal coverage: $\mathbb{P}$ is over joint distribution of (X, Y),i.e. $\mathbb{P}_{(X,Y)}$

$$1 - \alpha \leq \mathbb{P}_{(X,Y)}\left(Y_{n+1} \in C_{\text{split}}(X_{n+1})\right) \leq 1 - \alpha + \frac{2}{|\mathcal{I}_2| + 2}$$

Much stronger property

$$\mathbb{P}(Y_{n+1} \in C(x) \mid X_{n+1} = x) \geq 1 - \alpha \text{ for all } x \in \mathbb{R}^d$$

Remark on conformal prediction set

$$C_{\text{split}}(X_{n+1}) = [\widehat{\mu}(X_{n+1}) - Q_{1-\alpha}, \widehat{\mu}(X_{n+1}) + Q_{1-\alpha}]$$

- Marginal coverage: $\mathbb{P}$ is over joint distribution of (X, Y),i.e. $\mathbb{P}_{(X,Y)}$

$$1 - \alpha \leq \mathbb{P}_{(X,Y)}\left(Y_{n+1} \in C_{\text{split}}(X_{n+1})\right) \leq 1 - \alpha + \frac{2}{|\mathcal{I}_2| + 2}$$

  Much stronger property

$$\mathbb{P}(Y_{n+1} \in C(x) \mid X_{n+1} = x) \geq 1 - \alpha \text{ for all } x \in \mathbb{R}^d$$

- Problem of C_{split}: Constant CI width equal to $2Q_{1-\alpha}$ for any X_{n+1},

Remark on conformal prediction set

$$C_{\text{split}}(X_{n+1}) = [\widehat{\mu}(X_{n+1}) - Q_{1-\alpha}, \widehat{\mu}(X_{n+1}) + Q_{1-\alpha}]$$

- Marginal coverage: $\mathbb{P}$ is over joint distribution of (X, Y),i.e. $\mathbb{P}_{(X,Y)}$

$$1-\alpha \leq \mathbb{P}_{(X,Y)}\left(Y_{n+1} \in C_{\text{split}}(X_{n+1})\right) \leq 1-\alpha + \frac{2}{|\mathcal{I}_2| + 2}$$

Much stronger property

$$\mathbb{P}(Y_{n+1} \in C(x) \mid X_{n+1} = x) \geq 1-\alpha \text{ for all } x \in \mathbb{R}^d$$

- Problem of C_{split}: Constant CI width equal to $2Q_{1-\alpha}$ for any X_{n+1}, solution: use training data to fit an local variability $\hat{\rho}(x)$

$$R_{n+1} = \frac{\left|Y_{n+1} - \widehat{\mu}(X_{n+1})\right|}{\widehat{\rho}(X_{n+1})}, \quad R_i = \frac{\left|Y_i - \widehat{\mu}(X_i)\right|}{\widehat{\rho}(X_i)}, i \in \mathcal{I}_2$$

$$C_{\text{split}}^{\text{local}}(X_{n+1}) = [\widehat{\mu}(X_{n+1}) - \widehat{\rho}(X_{n+1})Q_{1-\alpha}, \widehat{\mu}(X_{n+1}) + \widehat{\rho}(X_{n+1})Q_{1-\alpha}]$$

Alternative method: Conformal quantile regression [RPC19]

- Use $q_\alpha(\cdot)$ denote quantile function. Natruallly, $[q_\alpha(Y), q_{1-\alpha}(Y)]$ is best $1-\alpha$ CI for Y.

Alternative method: Conformal quantile regression [RPC19]

- Use $q_\alpha(\cdot)$ denote quantile function. Natruallly, $[q_\alpha(Y), q_{1-\alpha}(Y)]$ is best $1-\alpha$ CI for Y.
- Training $\mathcal{I}_1$, holdout $\mathcal{I}_2$. For **any quantile** regression algorithm $\mathcal{A}$
  $\widehat{q} = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$
  $R_i^{\mathrm{CQR}} = \max\left\{\hat{q}_{\alpha/2}(X_i) - Y_i, Y_i - \hat{q}_{1-\alpha/2}(X_i)\right\}, i \in \mathcal{I}_2$
  $Q_{1-\alpha}$ is $(1-\alpha)(|\mathcal{I}_2|+1)$ quantile of $\left\{R_i^{\mathrm{CQR}} : i \in \mathcal{I}_2\right\}$
  Return $C_{\mathrm{split}}^{\mathrm{CQR}}(x) = [\hat{q}_{\alpha/2}(x) - Q_{1-\alpha}, \hat{q}_{1-\alpha/2}(x) + Q_{1-\alpha}]$
  $(\max\{q_1 - Y, Y - q_2\} \le Q \Rightarrow Y \in [q_1 - Q, q_2 + Q])$

Alternative method: Conformal quantile regression [RPC19]

- Use $q_\alpha(\cdot)$ denote quantile function. Natruallly, $[q_\alpha(Y), q_{1-\alpha}(Y)]$ is best $1-\alpha$ CI for Y.
- Training $\mathcal{I}_1$, holdout $\mathcal{I}_2$. For **any quantile** regression algorithm $\mathcal{A}$
  $\widehat{q} = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$
  $R_i^{\mathrm{CQR}} = \max\left\{\hat{q}_{\alpha/2}(X_i) - Y_i, Y_i - \hat{q}_{1-\alpha/2}(X_i)\right\}, i \in \mathcal{I}_2$
  $Q_{1-\alpha}$ is $(1-\alpha)(|\mathcal{I}_2|+1)$ quantile of $\left\{R_i^{\mathrm{CQR}} : i \in \mathcal{I}_2\right\}$
  Return $C_{\text{split}}^{\mathrm{CQR}}(x) = [\hat{q}_{\alpha/2}(x) - Q_{1-\alpha}, \hat{q}_{1-\alpha/2}(x) + Q_{1-\alpha}]$
  $(\max\{q_1 - Y, Y - q_2\} \leq Q \Rightarrow Y \in [q_1 - Q, q_2 + Q])$
- Marginal coverage:

$$1-\alpha \leq \mathbb{P}\left(Y_{n+1} \in C_{\text{split}}^{\mathrm{CQR}}(X_{n+1})\right) \leq 1-\alpha + \frac{2}{|\mathcal{I}_2|+2}$$

Alternative method: Conformal quantile regression [RPC19]

- Use $q_\alpha(\cdot)$ denote quantile function. Natruallly, $[q_\alpha(Y), q_{1-\alpha}(Y)]$ is best $1-\alpha$ CI for Y.
- Training $\mathcal{I}_1$, holdout $\mathcal{I}_2$. For **any quantile** regression algorithm $\mathcal{A}$
  $\widehat{q} = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$
  $R_i^{\mathrm{CQR}} = \max\left\{\hat{q}_{\alpha/2}(X_i) - Y_i, Y_i - \hat{q}_{1-\alpha/2}(X_i)\right\}, i \in \mathcal{I}_2$
  $Q_{1-\alpha}$ is $(1-\alpha)(|\mathcal{I}_2|+1)$ quantile of $\left\{R_i^{\mathrm{CQR}} : i \in \mathcal{I}_2\right\}$
  Return $C_{\mathrm{split}}^{\mathrm{CQR}}(x) = [\hat{q}_{\alpha/2}(x) - Q_{1-\alpha}, \hat{q}_{1-\alpha/2}(x) + Q_{1-\alpha}]$
  $(\max\{q_1 - Y, Y - q_2\} \leq Q \Rightarrow Y \in [q_1 - Q, q_2 + Q])$
- Marginal coverage:

$$1-\alpha \leq \mathbb{P}\left(Y_{n+1} \in C_{\mathrm{split}}^{\mathrm{CQR}}(X_{n+1})\right) \leq 1-\alpha + \frac{2}{|\mathcal{I}_2|+2}$$

- Benefit: adapt to local variability.

Conformal prediction under covariate shift: weighted conformal inference [TFBCR19]

- New data may not i.i.d with previous data

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X}, i = 1, \ldots, n,$$

$$(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}, \text{ independently.}$$

Assume $w(X_i) = \mathrm{d}\widetilde{P}_X(X_i) / \mathrm{d}P_X(X_i)$ is known

Conformal prediction under covariate shift: weighted conformal inference [TFBCR19]

- New data may not i.i.d with previous data

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X}, i = 1, \ldots, n,$$
$$(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}, \text{ independently.}$$

  Assume $w(X_i) = \mathrm{d}\widetilde{P}_X(X_i) / \mathrm{d}P_X(X_i)$ is known
- no covariate shift: $1 - \alpha$ quantile of $\frac{1}{n}\sum_{i=1}^{n} \delta_{R_i}$

Conformal prediction under covariate shift: weighted conformal inference [TFBCR19]

- New data may not i.i.d with previous data

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X}, i = 1, \ldots, n,$$
$$(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}, \text{ independently.}$$

  Assume $w(X_i) = d\widetilde{P}_X(X_i) / dP_X(X_i)$ is known
- no covariate shift: $1 - \alpha$ quantile of $\frac{1}{n}\sum_{i=1}^{n} \delta_{R_i}$
  covariate shift: $1 - \alpha$ quantile of $\frac{1}{n}\sum_{i=1}^{n} p_i^w \delta_{R_i}$

$$p_i^w = \frac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}, i = 1, \ldots, n$$

Conformal prediction under covariate shift: weighted conformal inference [TFBCR19]

- New data may not i.i.d with previous data

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X}, i = 1, \dots, n,$$

$$(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}, \text{ independently.}$$

Assume $w(X_i) = \mathrm{d}\widetilde{P}_X(X_i) / \mathrm{d}P_X(X_i)$ is known

- no covariate shift: $1 - \alpha$ quantile of $\frac{1}{n}\sum_{i=1}^{n} \delta_{R_i}$
  covariate shift: $1 - \alpha$ quantile of $\frac{1}{n}\sum_{i=1}^{n} p_i^w \delta_{R_i}$

$$p_i^w = \frac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}, i = 1, \dots, n$$

- Construction: $C_{\text{split}}^w(x) = [\widehat{\mu}(x) - Q_{1-\alpha}^w, \widehat{\mu}(x) + Q_{1-\alpha}^w]$

$$\mathbb{P}\left(Y_{n+1} \in C_{\text{split}}^w(X_{n+1})\right) \geq 1 - \alpha$$

Inference of counterfactuals? Potential outcomes

Unit	X_i	T_i	$Y_i(1)$	$Y_i(0)$	Y_i^{obs}
Treatment Group					
1	✓	1	✓	x	$Y_1(1)$
2	✓	1	✓	x	$Y_2(1)$
3	✓	1	✓	x	$Y_3(1)$
4	✓	1	✓	x	$Y_4(1)$
5	✓	1	✓	x	$Y_5(1)$
Control Group					
6	✓	0	x	✓	$Y_6(0)$
7	✓	0	x	✓	$Y_7(0)$
8	✓	0	x	✓	$Y_8(0)$
9	✓	0	x	✓	$Y_9(0)$
10	✓	0	x	✓	$Y_{10}(0)$

Inference of counterfactuals? Potential outcomes [LC20]

Assumption

- stable unit treatment values (SUTVA)
- (i.i.d.)
- unconfoundedness $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$

- Individual treatment effect(ITE) τ_i is defined as

$$\tau_i \triangleq Y_i(1) - Y_i(0).$$

Inference of counterfactuals? Potential outcomes [LC20]

Assumption

- stable unit treatment values (SUTVA)
- (i.i.d.)
- unconfoundedness $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$

- Individual treatment effect(ITE) τ_i is defined as

$$\tau_i \triangleq Y_i(1) - Y_i(0).$$

  τ_i never observed.
- Traditional target: CATE, $\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]$

Inference of counterfactuals? Potential outcomes [LC20]

Assumption

- stable unit treatment values (SUTVA)
- (i.i.d.)
- unconfoundedness $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$

- Individual treatment effect(ITE) τ_i is defined as

$$\tau_i \triangleq Y_i(1) - Y_i(0).$$

  τ_i never observed.
- Traditional target: CATE, $\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]$
- **Goal**: find interval estimate $\hat{C}_t(X)$, s.t.,

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

Counterfactual inference

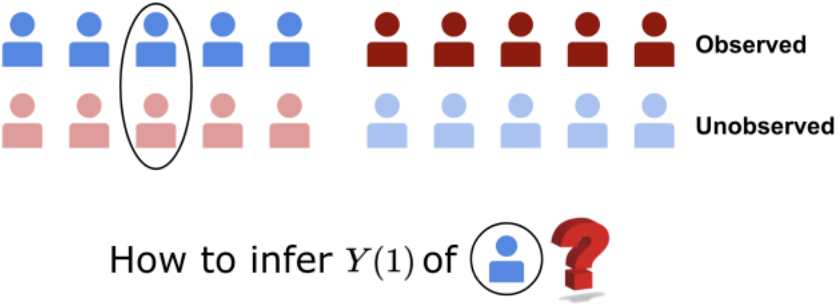
Assign treatment by a coin toss for each subject based on the propensity score e(x)

$$\mathbb{P}(\text{treated} \mid X = x) = e(x)$$

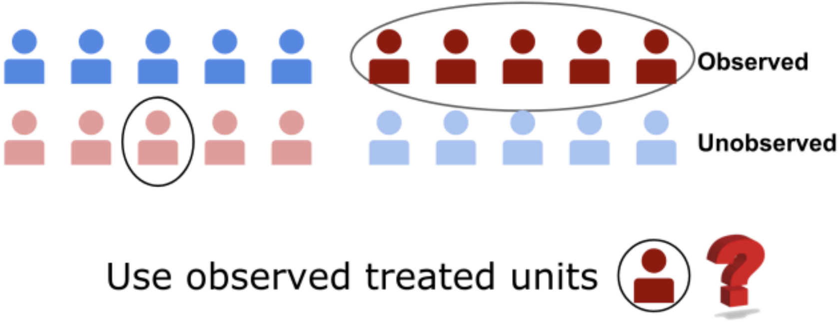
$$\mathbb{P}(\text{control} \mid X = x) = 1 - e(x)$$

Each subject has potential outcomes $(Y(1), Y(0))$ and the observed outcome Y^{obs}

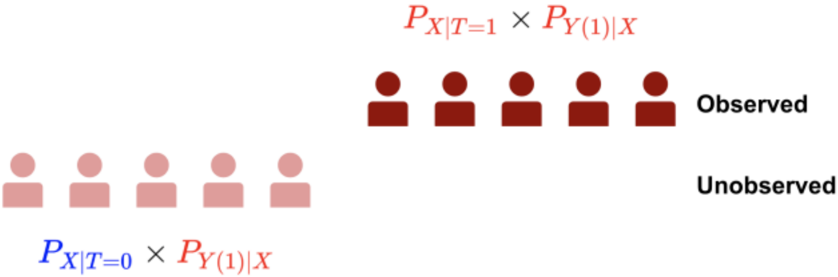
Counterfactual inference



Counterfactual inference



Covariate shift under unconfoundedness $Y(1) \perp\!\!\!\perp T \mid X$



Counterfactual inference

$$(X_i, Y_i^{\text{obs}}) \overset{i.i.d.}{\sim} P_{X|T=0} \times P_{Y(0)|X}$$

$$(X_i, Y_i^{\text{obs}}) \overset{i.i.d.}{\sim} P_{X|T=1} \times P_{Y(1)|X}$$

Use i.i.d. samples (observed treated units) from $P_{X|T=1} \times P_{Y(1)|X}$ to construct $\hat{C}_1(X)$ with

$$\mathbb{P}\Big(Y(1) \in \hat{C}_1(X)\Big) \geq 90\% \text{ under } P_{X|T=0} \times P_{Y(1)|X}$$

$$\text{Covariate shift } w(x) \triangleq \frac{dP_{X|T=0}}{dP_{X|T=1}}(x)$$

Counterfactual inference

$$(X_i, Y_i^{\text{obs}}) \overset{i.i.d.}{\sim} P_{X|T=0} \times P_{Y(0)|X}$$

$$(X_i, Y_i^{\text{obs}}) \overset{i.i.d.}{\sim} P_{X|T=1} \times P_{Y(1)|X}$$

Use i.i.d. samples (observed treated units) from $P_{X|T=1} \times P_{Y(1)|X}$ to construct $\hat{C}_1(X)$ with

$$\mathbb{P}\left(Y(1) \in \hat{C}_1(X)\right) \geq 90\% \text{ under } P_{X|T=0} \times P_{Y(1)|X}$$

$$\text{Covariate shift } w(x) \triangleq \frac{dP_{X|T=0}}{dP_{X|T=1}}(x) \propto \frac{1 - e(x)}{e(x)}$$

Survival data

- Observations: $\{(X_i, \tilde{T}_i, \Delta_i)\}_{i=1}^n$ i.i.d.

Survival data

- Observations: $\{(X_i, \tilde{T}_i, \Delta_i)\}_{i=1}^n$ i.i.d.
- Event indicator $\Delta_i = I(T_i < C_i)$:

$$\tilde{T}_i = \begin{cases} T_i & \text{if } \Delta_i = 1 \\ C_i & \text{if } \Delta_i = 0 \end{cases} \Rightarrow \tilde{T}_i = \min(T_i, C_i)$$

- Conditional independent censoring: $T \perp\!\!\!\perp C \mid X$

Survival data

- Observations: $\{(X_i, \tilde{T}_i, \Delta_i)\}_{i=1}^n$ i.i.d.
- Event indicator $\Delta_i = I(T_i < C_i)$:

$$\tilde{T}_i = \begin{cases} T_i & \text{if } \Delta_i = 1 \\ C_i & \text{if } \Delta_i = 0 \end{cases} \Rightarrow \tilde{T}_i = \min(T_i, C_i)$$

- Conditional independent censoring: $T \perp\!\!\!\perp C \mid X$
- Objective:

$$\mathbb{P}(T_{n+1} \geq \hat{L}(X_{n+1})) \geq 1 - \alpha$$

Survival data

- Observations: $\{(X_i, \tilde{T}_i, \Delta_i)\}_{i=1}^n$ i.i.d.
- Event indicator $\Delta_i = I(T_i < C_i)$:

$$\tilde{T}_i = \begin{cases} T_i & \text{if } \Delta_i = 1 \\ C_i & \text{if } \Delta_i = 0 \end{cases} \Rightarrow \tilde{T}_i = \min(T_i, C_i)$$

- Conditional independent censoring: $T \perp\!\!\!\perp C \mid X$
- Objective:

$$\mathbb{P}(T_{n+1} \geq \hat{L}(X_{n+1})) \geq 1 - \alpha$$

- Naive solution: $T_{n+1} \geq \tilde{T}_{n+1}$

$$\mathbb{P}(T_{n+1} \geq \hat{L}(X_{n+1})) \geq \mathbb{P}(\tilde{T}_{n+1} \geq \hat{L}(X_{n+1})) \geq 1 - \alpha$$

Survival data

- Observations: $\{(X_i, \tilde{T}_i, \Delta_i)\}_{i=1}^n$ i.i.d.
- Event indicator $\Delta_i = I(T_i < C_i)$:

$$\tilde{T}_i = \begin{cases} T_i & \text{if } \Delta_i = 1 \\ C_i & \text{if } \Delta_i = 0 \end{cases} \quad \Rightarrow \tilde{T}_i = \min(T_i, C_i)$$

- Conditional independent censoring: $T \perp\!\!\!\perp C \mid X$
- Objective:

$$\mathbb{P}(T_{n+1} \geq \hat{L}(X_{n+1})) \geq 1 - \alpha$$

- Naive solution: $T_{n+1} \geq \tilde{T}_{n+1}$

$$\mathbb{P}(T_{n+1} \geq \hat{L}(X_{n+1})) \geq \mathbb{P}(\tilde{T}_{n+1} \geq \hat{L}(X_{n+1})) \geq 1 - \alpha$$

Bad even if we have oracle quantile $\tilde{q}_\alpha(X)$, $q_\alpha(x)$

$$\mathbb{P}\left(T \geq q_\alpha(x) \mid X = x\right) = 1 - \alpha = \mathbb{P}\left(\widetilde{T} \geq \tilde{q}_\alpha(x) \mid X = x\right)$$

$$= \mathbb{P}\left(T \geq \tilde{q}_\alpha(x) \mid X = x\right) \mathbb{P}\left(C \geq \tilde{q}_\alpha(x) \mid X = x\right)$$

Survival data

- Observations: $\{(X_i, \tilde{T}_i, \Delta_i)\}_{i=1}^n$ i.i.d.
- Event indicator $\Delta_i = I(T_i < C_i)$:

$$\tilde{T}_i = \begin{cases} T_i & \text{if } \Delta_i = 1 \\ C_i & \text{if } \Delta_i = 0 \end{cases} \Rightarrow \tilde{T}_i = \min(T_i, C_i)$$

- Conditional independent censoring: $T \perp\!\!\!\perp C \mid X$
- Objective:

$$\mathbb{P}(T_{n+1} \geq \hat{L}(X_{n+1})) \geq 1 - \alpha$$

- Naive solution: $T_{n+1} \geq \tilde{T}_{n+1}$

$$\mathbb{P}(T_{n+1} \geq \hat{L}(X_{n+1})) \geq \mathbb{P}(\tilde{T}_{n+1} \geq \hat{L}(X_{n+1})) \geq 1 - \alpha$$

Bad even if we have oracle quantile $\tilde{q}_\alpha(X)$, $q_\alpha(x)$

$$\mathbb{P}\left(T \geq q_\alpha(x) \mid X = x\right) = 1 - \alpha = \mathbb{P}\left(\widetilde{T} \geq \tilde{q}_\alpha(x) \mid X = x\right)$$

$$= \mathbb{P}\left(T \geq \tilde{q}_\alpha(x) \mid X = x\right) \mathbb{P}\left(C \geq \tilde{q}_\alpha(x) \mid X = x\right)$$

small censoring time is bad, conservative.

Treat T as a "potential outcome" under the "treatment" $\Delta = 1$?

- Event indicator $\Delta_i = I(T_i < C_i)$:

$$\tilde{T}_i = \begin{cases} T_i & \text{if } \Delta_i = 1 \\ C_i & \text{if } \Delta_i = 0 \end{cases}$$

Treat T as a "potential outcome" under the "treatment" $\Delta = 1$?

- Event indicator $\Delta_i = I(T_i < C_i)$:

$$\tilde{T}_i = \begin{cases} T_i & \text{if } \Delta_i = 1 \\ C_i & \text{if } \Delta_i = 0 \end{cases}$$

- Invalid because "unconfoundedness" does not hold:

$$(T, C) \not\perp I(T < C) \mid X$$

$(X_i, T_i)_{\Delta_i = 1}$ has shifts in both the covariate distribution and conditional distribution

- What group should we condition on?

Leveraging the censoring mechanism[CLR23]

- Small censoring time is bad → condition on group with larger censoring time? E.g. $C \geq c_0$
- Obviously $(X, C, T) \overset{d}{\neq} (X, C, T) \mid C \geq c_0$.

$$P_{(X,\widetilde{T})|C\geq c_0} = P_{X|C\geq c_0} \times P_{\widetilde{T}|X,C\geq c_0}$$

shift in both distribution.

Leveraging the censoring mechanism[CLR23]

- Small censoring time is bad $\rightarrow$ condition on group with larger censoring time? E.g. $C \geq c_0$
- Obviously $(X, C, T) \overset{d}{\neq} (X, C, T) \mid C \geq c_0$.

$$P_{(X,\widetilde{T})|C\geq c_0} = P_{X|C\geq c_0} \times P_{\widetilde{T}|X,C\geq c_0}$$

  shift in both distribution.
- Consider new censored time $\widetilde{T} \wedge c_0$, where $a \wedge b = \min\{a, b\}$

Leveraging the censoring mechanism[CLR23]

- Small censoring time is bad → condition on group with larger censoring time? E.g. $C \geq c_0$
- Obviously $(X, C, T) \overset{d}{\neq} (X, C, T) \mid C \geq c_0$.

$$P_{(X,\widetilde{T})|C\geq c_0} = P_{X|C\geq c_0} \times P_{\widetilde{T}|X,C\geq c_0}$$

shift in both distribution.

- Consider new censored time $\widetilde{T} \wedge c_0$, where $a \wedge b = \min\{a, b\}$

$$P_{(X,\widetilde{T}\wedge c_0)|C\geq c_0} = P_{X|C\geq c_0} \times P_{\widetilde{T}\wedge c_0|X,C\geq c_0} \overset{(a)}{=} P_{X|C\geq c_0} \times P_{T\wedge c_0|X,C\geq c_0}$$

$$\overset{(b)}{=} P_{X|C\geq c_0} \times P_{T\wedge c_0|X}$$

Leveraging the censoring mechanism[CLR23]

- Small censoring time is bad → condition on group with larger censoring time? E.g. $C \geq c_0$
- Obviously $(X, C, T) \overset{d}{\neq} (X, C, T) \mid C \geq c_0$.

$$P_{(X,\widetilde{T})|C\geq c_0} = P_{X|C\geq c_0} \times P_{\widetilde{T}|X,C\geq c_0}$$

  shift in both distribution.
- Consider new censored time $\widetilde{T} \wedge c_0$, where $a \wedge b = \min\{a, b\}$

$$P_{\left(X,\widetilde{T}\wedge c_0\right)|C\geq c_0} = P_{X|C\geq c_0} \times P_{\widetilde{T}\wedge c_0|X,C\geq c_0} \overset{(a)}{=} P_{X|C\geq c_0} \times P_{T\wedge c_0|X,C\geq c_0} \overset{(b)}{=} P_{X|C\geq c_0} \times P_{T\wedge c_0|X}$$

- We can build a lower bound for $T_{n+1} \wedge c_0$ via weighted conformal

$$\frac{dP_X}{dP_{X|C\geq c_0}}(x) = \frac{\mathbb{P}(C \geq c_0)}{\mathbb{P}(C \geq c_0 \mid X = x)}$$

The essence of conformal method

- Non-conformity score(S): how well a sample Z conforms to rest of data, if S is large, we say that Z is non-conforming or ”strange”.

The essence of conformal method

- Non-conformity score(S): how well a sample Z conforms to rest of data, if S is large, we say that Z is non-conforming or "strange".
    - E.g.: $S_{n+1} = |Y_{n+1} - \hat{\mu}(X_{n+1})|$, large residual→ strange →

The essence of conformal method

- Non-conformity score(S): how well a sample Z conforms to rest of data, if S is large, we say that Z is non-conforming or ”strange”.
  - E.g.: $S_{n+1} = |Y_{n+1} - \hat{\mu}(X_{n+1})|$, large residual→ strange → small p-value in hypothesis test.

The essence of conformal method

- Non-conformity score(S): how well a sample Z conforms to rest of data, if S is large, we say that Z is non-conforming or "strange".
  - E.g.: $S_{n+1} = |Y_{n+1} - \hat{\mu}(X_{n+1})|$, large residual$\rightarrow$ strange $\rightarrow$ small p-value in hypothesis test.

Conformity score $\longrightarrow$ Hypothesis test $\longrightarrow$ Conf Interval

The essence of conformal method

- Non-conformity score(S): how well a sample Z conforms to rest of data, if S is large, we say that Z is non-conforming or ”strange”.
  - E.g.: $S_{n+1} = |Y_{n+1} - \hat{\mu}(X_{n+1})|$, large residual→ strange → small p-value in hypothesis test.

    Conformity score → Hypothesis test → Conf Interval
- Novel conformity score→ different task

The essence of conformal method

- Non-conformity score(S): how well a sample Z conforms to rest of data, if S is large, we say that Z is non-conforming or ”strange”.
    - E.g.: $S_{n+1} = |Y_{n+1} - \hat{\mu}(X_{n+1})|$, large residual→ strange → small p-value in hypothesis test.

  Conformity score ⟶ Hypothesis test ⟶ Conf Interval

- Novel conformity score→ different task
- Hypothesis test
    - H_0: X_{n+1} follow same distribution with observed data[BCL+21]

The essence of conformal method

- Non-conformity score(S): how well a sample Z conforms to rest of data, if S is large, we say that Z is non-conforming or ”strange”.
  - E.g.: $S_{n+1} = |Y_{n+1} - \hat{\mu}(X_{n+1})|$, large residual→ strange → small p-value in hypothesis test.

Conformity score → Hypothesis test → Conf Interval

- Novel conformity score→ different task
- Hypothesis test
  - H_0: X_{n+1} follow same distribution with observed data[BCL+21]
  - H_0: Two-sample conditional distribution are equal [HL23]

Testing for outliers with conformal p-value [BCL+21]

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{cal}}, \quad |\mathcal{D}^{\text{train}}| = |\mathcal{D}^{\text{cal}}| = n$$

- Clean i.i.d observed data from P_X, given many testing data want to test $\mathcal{H}_{0,i} : X_i \sim P_X$, for any $X_i \in \mathcal{D}^{\text{test}}$

Testing for outliers with conformal p-value [BCL+21]

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{cal}}, \quad |\mathcal{D}^{\text{train}}| = |\mathcal{D}^{\text{cal}}| = n$$

- Clean i.i.d observed data from P_X, given many testing data want to test $\mathcal{H}_{0,i} : X_i \sim P_X$, for any $X_i \in \mathcal{D}^{\text{test}}$
- Machine learning tool give a score: $\hat{s}(X_i)$, p-value is given by

$$\hat{u}^{(\text{marg})}(X_i) = \frac{1 + \left|\left\{j \in \mathcal{D}^{\text{cal}} : \hat{s}\left(X_j\right) \leq \hat{s}(X_i)\right\}\right|}{n+1}$$

with

$$\mathbb{P}\left[\hat{u}^{(\text{marg})}(X_i) \leq t\right] \leq t \quad \text{under } \mathcal{H}_{0,i}$$

Testing for outliers with conformal p-value [BCL+21]

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{cal}}, \quad |\mathcal{D}^{\text{train}}| = |\mathcal{D}^{\text{cal}}| = n$$

- Clean i.i.d observed data from P_X, given many testing data want to test $\mathcal{H}_{0,i} : X_i \sim P_X$, for any $X_i \in \mathcal{D}^{\text{test}}$
- Machine learning tool give a score: $\hat{s}(X_i)$, p-value is given by

$$\hat{u}^{(\text{marg})}(X_i) = \frac{1 + \left|\left\{j \in \mathcal{D}^{\text{cal}} : \hat{s}\left(X_j\right) \leq \hat{s}(X_i)\right\}\right|}{n+1}$$

with

$$\mathbb{P}\left[\hat{u}^{(\text{marg})}(X_i) \leq t\right] \leq t \quad \text{under } \mathcal{H}_{0,i}$$

  - Negative result: global testing can fail:

$$H_0 : X_{2n+1}, \ldots, X_{2n+m} \overset{\text{i.i.d.}}{\sim} P_X$$

Testing for outliers with conformal p-value [BCL+21]

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{cal}}, \quad |\mathcal{D}^{\text{train}}| = |\mathcal{D}^{\text{cal}}| = n$$

- Clean i.i.d observed data from P_X, given many testing data want to test $\mathcal{H}_{0,i} : X_i \sim P_X$, for any $X_i \in \mathcal{D}^{\text{test}}$
- Machine learning tool give a score: $\hat{s}(X_i)$, p-value is given by

$$\hat{u}^{(\text{marg})}(X_i) = \frac{1 + \left|\left\{j \in \mathcal{D}^{\text{cal}} : \hat{s}\left(X_j\right) \leq \hat{s}(X_i)\right\}\right|}{n+1}$$

with

$$\mathbb{P}\left[\hat{u}^{(\text{marg})}(X_i) \leq t\right] \leq t \quad \text{under } \mathcal{H}_{0,i}$$

  - Negative result: global testing can fail:

$$H_0 : X_{2n+1}, \ldots, X_{2n+m} \overset{\text{i.i.d.}}{\sim} P_X$$

  - Positive result: $\{\hat{u}^{(\text{marg})}(X_{2n+1}), \cdots \hat{u}^{(\text{marg})}(X_{2n+m})\}$ are PRDS.

End of presentation

All models are wrong, but some are (hopefully) useful

End of presentation

All models are wrong, but some are (hopefully) useful

All models are wrong, but **conformal** can make them safe and useful!

Thanks!

Reference I

- Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia, Testing for outliers with conformal p-values, arXiv preprint arXiv:2104.08279 (2021).
- Emmanuel Candès, Lihua Lei, and Zhimei Ren, Conformalized survival analysis, Journal of the Royal Statistical Society Series B: Statistical Methodology **85** (2023), no. 1, 24–45.
- Xiaoyu Hu and Jing Lei, A two-sample conditional distribution test using conformal prediction and weighted rank sum, Journal of the American Statistical Association (2023), 1–19.
- Lihua Lei and Emmanuel J Candès, Conformal inference of counterfactuals and individual treatment effects, arXiv preprint arXiv:2006.06138 (2020).

Reference II

- Jing Lei, Max GSell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman, Distribution-free predictive inference for regression, Journal of the American Statistical Association **113** (2018), no. 523, 1094–1111.
- Yaniv Romano, Evan Patterson, and Emmanuel Candes, Conformalized quantile regression, Advances in neural information processing systems **32** (2019).
- Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas, Conformal prediction under covariate shift, Advances in neural information processing systems **32** (2019).