Methodology

Assume $w(X) > 0$

Goal. $\mathbb{P}_{(X,Y)\sim\mathcal{P}_{\text{data}}\otimes\mathcal{P}_{\text{data}}}\left(Y(1) \in \hat{C}_\alpha(X)\right) \geq 1-\alpha$

with i.i.d samples (X, Y, T) from $\mathcal{P}_{(T|X)} \times \mathcal{P}_{(X,Y)}$

Note: Form of $\hat{Q}_\alpha$ depends on the estimand

ATE: $\hat{Q}_\alpha = \hat{Q}_\alpha$

ATT: $\hat{Q}_\alpha = \hat{Q}_\alpha(1)$

ATC: $\hat{Q}_\alpha = \hat{Q}_\alpha(0)$

Key idea.

1. Romano et al. (19') Conformal Quantile Regression (CQR)
2. Tibshirani et al. (19') Conformal prediction under covariate shift

Algorithm. Input α, data $(X, Y, T)_{\in I}$; I with $T=0$ is missing

1. Split data into training fold $\mathcal{Z}_{tr}$ (Suggest proportion 75%) calibration fold $\mathcal{Z}_{ca}$ (Suggest proportion 25%)
2. Use the training data to train 3 regression functions. $\hat{q}_{low}(x)$, $\hat{q}_{hi}(x)$, $\hat{\omega}(x) = \frac{\hat{p}(x=1)}{\hat{p}_{(T|X)}(x)}$
3. For i in calibration fold, compute score $U_i = \max\left\{\hat{q}_{low}(X_i) - Y_i,\ Y_i - \hat{q}_{hi}(X_i)\right\}$ and weight $W_i := \hat{\omega}(X_i)$
4. For a new covariate x^*, the interval $\hat{C}_\alpha(x^*)$ that satisfies $\mathbb{P}_{(X,Y)\sim\mathcal{P}_{data}\otimes\mathcal{P}_{data}}\left(Y(1) \in \hat{C}_\alpha(x^*)\right) \geq 1-\alpha$ is given by $\hat{C}_\alpha(x^*) = \left[\hat{q}_{low}(x^*) - \eta(x^*),\ \hat{q}_{hi}(x^*) + \eta(x^*)\right]$ where $\eta(x)$ is the quantile of the weighted empirical dist of $\{U_i : i \in \mathcal{Z}_{ca}\}$ dist. with weights $\propto W_i$ and $\hat{\omega}(x^*)$

A guarantee weighted conformal inference } Adapt to RS in the IW sense for bounds (and ellipse) } Adapt to CS in the same way

Example ATE: $\hat{Q}_\alpha = \hat{\tau}_\alpha$

$$\omega(x) = \frac{d\mathcal{P}_{(X,Y)}(x)}{d\mathcal{P}_{(T|X)}(x)} = \frac{P(x)}{P(x|T=1)} = \frac{P(x)}{\frac{P(T=1|x)P(x)}{P(T=1)}} = \frac{P(T=1)}{P(T=1|x)} \propto \frac{1}{\hat{P}(T=1|x)}$$

Simulation Study

Basic setting:

- The covariate vector $X = (X_1, \ldots, X_d)^T$ is such that $X_i = \Phi(X_i^*)$, where Φ denotes the cdf of the standard normal distribution and $(X_1^*, \ldots, X_d^*)$ is an equicorrelated multivariate Gaussian vector with mean zero and $\mathrm{Var}(X_j^*) = 1$, $\mathrm{Cor}(X_j^*, X_k^*) = \rho$ for $j \neq j'$. When $\rho = 0$, X is uniformly distributed on the unit cube. When $\rho > 0$, the variables are positively correlated.
- The potential outcome $Y(1)$ is generated as follows:

$$E[Y(1)|X] = f(X_1)I(X_2); \quad f(x) = \frac{2}{1 + \exp\left(-12(x-0.5)\right)}$$

which is the same as in Wager and Athey [2018], and

$$Y(1) = E[Y(1)|X] + \sigma(X)\epsilon, \quad \epsilon \sim N(0,1);$$

the homoscedastic case $\sigma(x)$ is σ^2 is considered in Wager and Athey [2018].

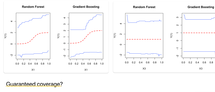
- The propensity score $e(x)$ is set as:

$$e(x) = \frac{1}{2}\left(1 + B_{\beta,\beta}(x)\right),$$

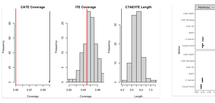
where $B_{\beta,\beta}$ is the cdf of the beta distribution with shape parameters $(2, 4)$. This ensures that $e(x) \in [0.25, 0.5]$, thereby providing sufficient overlap.

- 100 repeated experiments, sample size 1000, 10000 extra data for test, building 95% CI
- High dimensional (d=100), uncorrelated covariates, homoscedastic error

How do the intervals look like?

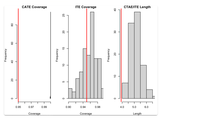


Guaranteed coverage?



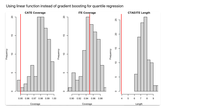
Why quantile regression?

The following result is from standard conformal inference, which provides quantiles of the mean.

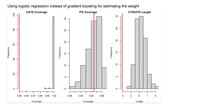


Doubly Robust?

Using linear function instead of gradient boosting for quantile regression



Using logistic regression instead of gradient boosting for estimating the weight

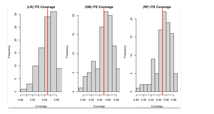


How to choose weight learner?

Lower bound provided by proposition 1 may be an option

Consider three learners for the weight function: logistic regression, gradient boosting, random forest

Corresponding lower bound, calculated empirically using Proposition 1, (LR) 0.463, (GB) 0.696, (RF) 0.750



Percentage with coverage rate greater than 95%, (LR) 0.56, (GB) 0.59, (RF) 0.62

ITE for subjects not in the study

Goal

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

Naive approach

For x^*, find $\hat{C}_1(x^*) = \left[\hat{q}^L_1(1, x^*),\ \hat{q}^U_1(1, x^*)\right]$ use only $\{X_i : T_i = 1\}$

and $\hat{C}_0(x^*) = \left[\hat{q}^L_0(0, x^*),\ \hat{q}^U_0(0, x^*)\right]$ use only $\{i : T_i = 0\}$

using one pass algorithm each

$$\hat{C}_{ITE} = \left[\hat{q}^L_1(1, x^*) - \hat{q}^U_0(0, x^*),\ \hat{q}^U_1(1, x^*) - \hat{q}^L_0(0, x^*)\right]$$

Separate inference, no borrow of information

Nested approach

Motivation: a purely inference method that leverage borrowing of info

Sketch of the procedure

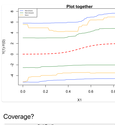
1. Split data into 2 folds $\left[\hat{q}^L_1, \hat{q}^U_1\right]$ $\left[\hat{q}^L_0, \hat{q}^U_0\right]$
2. Use data in the first fold, train $\hat{C}_1(x)$, $\hat{C}_0(x)$
3. Use the data in the second fold, create a new data set.

$\{(X_i, \hat{C}_i)\}$, where

$$\hat{C}_i = \begin{cases} \left[\hat{q}^L_1(x_i),\ x - \hat{q}^U_0(x_i)\right] & \text{if } T_i = 1 \\ \left[x - \hat{q}^U_1(x_i),\ x - \hat{q}^L_0(x_i)\right] & \text{if } T_i = 0 \end{cases}$$

Note that $\mathbb{P}\left(Y(1) - Y(0) \in \hat{C}_i\right) \geq 1-\alpha$

4. Use the new data set $\{(x_i, \hat{C}_i)\}$ to find a function $\hat{C}(x)$, and $\hat{C}(x)$ covers $\hat{C}_i$ well. Then $\hat{C}(x)$ is the interval we want



$\{(x_i, \hat{C}^{Inv}(x_i)), (x_i, \hat{C}^{Inv}(x_i)), \ldots\}$ Inverse quantile regression 60%

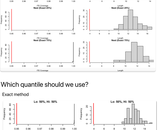
$\{(x_i, \hat{C}^{Exc}(x_i)), (x_i, \hat{C}^{Exc}(x_i)), \ldots\}$ Exact conformal inference 50%

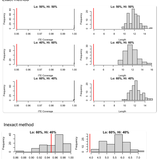
5. How to find $\hat{C}(x)$?
   - Inverse method: train separate model for the right/left end of $\hat{C}_i$
   - Exact method: Conformal inference

Simulation study

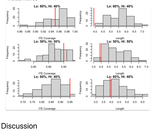
Same setting as before except $Y_i(0) \overset{iid}{\sim} N(0,1)$

How do the intervals look like?





Coverage?



Split proportion? Percentage in the title of the plots is the proportion used for the second fold.

Which quantile should we use?

Exact method

Inexact method

Discussion

1. This paper, and conformal inference in general, leverage the idea of split the data, and use the same the data for different purposes. Is this kind of counter-intuitive? Why this idea works well?

2. Any example in the Bayesian world that can this idea?