POLS 904 Final Project

Simulation Study on Causal Forest

Jiacheng He

December 13, 2017

Introduction

Sometimes we are interested to estimate the heterogeneous treatment effect of a binary treatment on different subgroups.

e.g.
Suppose we run a randomized trial to test a new drug. We spilt the subjects into a control group and a treatment group. What's the effect of this drug for adult white males and teenager Asian females respectively?

Introduction

Wager and Athey (2017) developed causal forest method to predict heterogeneous treatment effect conditional on observed covariates.

In this project, I test the prediction accuracy and confidence interval coverage rate of causal forest via simulation.

Causal Forest

Model setup

Y_i: The outcome variable

W_i: $W_i = 1$ if individual i receives treatment, $W_i = 0$ if not treated

X_i: A vector of covariates

τ_i: Individual treatment effect. Never observed.

$$Y_i = m(X_i) + \frac{W_i}{2}\tau(X_i) + \frac{1-W_i}{2}\tau(X_i) + \epsilon_i$$

$m(X_i) = E[Y_i|X_i]$: The conditional mean of outcome

$\tau(X_i) = E[Y_i|X_i, W_i = 1] - E[Y_i|X_i, W_i = 0]$: The heterogeneous treatment effect (conditional on covarites X_i)

$e(X_i) = E[W_i|X_i]$: The treatment propensity

Causal Forest

Goal is to predict $\tau(X_i)$ (while random forest aims to predict $m(X_i)$)

Difficulty:
1. Disentangle $\tau(X_i)$ from $m(X_i)$ and $e(X_i)$
2. Cannot perform cross-validation, because we never observe the true τ_i (while in random forest we observe the true Y_i)

Algorithm

Similar to random forest

Place a split at point $\tilde{x}_i$ which maximize the difference of $\hat{E}[Y_i|X_i = x_i, W_i = 1] - \hat{E}[Y_i|X_i = x_i, W_i = 0]$ across the two sides of $\tilde{x}_i$

(while random forest maximize the difference of $\hat{E}[Y_i|X_i = x_i]$)

Simulation Setup

DGP1 (constant τ)

$$\tau(X_i) = 0$$
$$e(X_i) = (1 + f_{beta}^{2,4}(X_{1i}))/4$$
$$m(X_i) = 2X_{1i} - 1$$

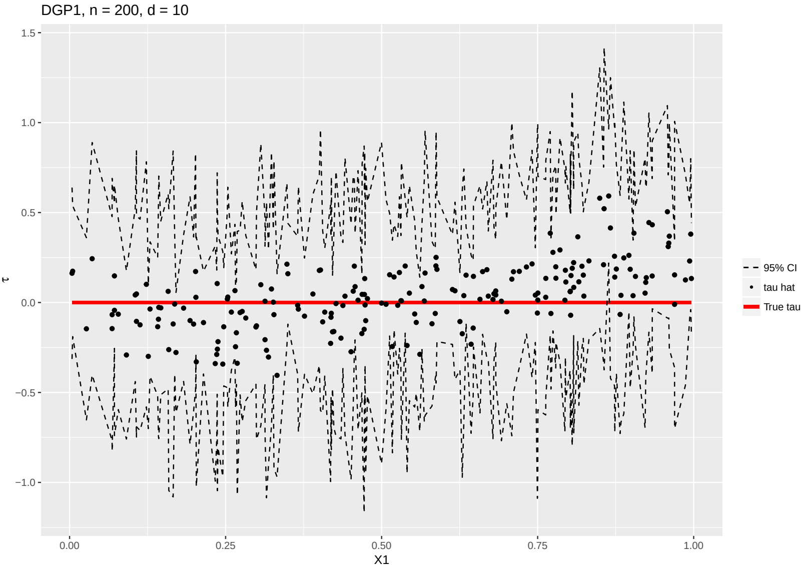
DGP2 (heterogeneous τ)

$$\tau(X_i) = 1 + \frac{1}{(1+e^{-20(X_{1i}-1/3)})(1+e^{-20(X_{2i}-1/3)})}$$
$$e(X_i) = 0.5$$
$$m(X_i) = 0$$

Simulation Setup

1. Draw $X_i \sim U(0,1)^d$, $W_i \sim binom(1, e(X_i))$, $\epsilon_i \sim N(0,1)$
2. Run the causal forest on a training set, then evaluate the trained model on a test set. ($n_{train} = n_{test}$)
3. For each senario, replicate it for 100 times. Then compute the average MSE and the 0.95 confidence interval coverage rate

An Example



DGP1, n = 200, d = 10
1.5
1.0
0.5
0.0
−0.5
−1.0
τ
0.00
0.25
0.50
0.75
1.00
X1
95% CI
tau hat
True tau

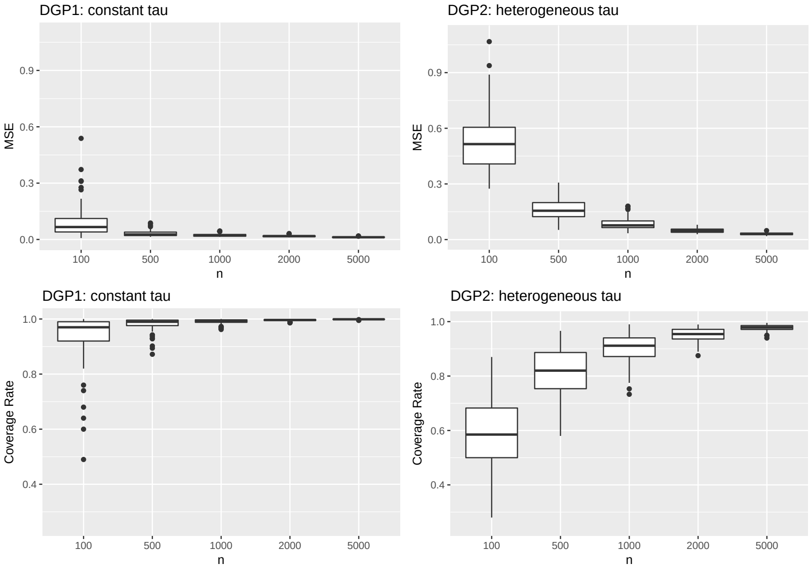
Sample Size and Covariate Size

Vary the sample size n and covariate size d to see what happens:

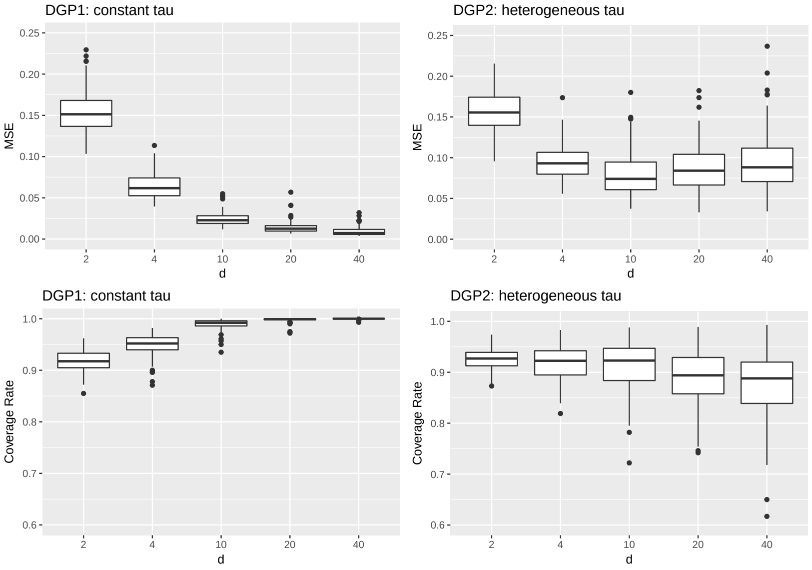
Fix $d = 10$, try $n = 100, 500, 1000, 2000, 5000$;

Fix $n = 1000$, try $d = 2, 4, 10, 20, 40$

Sample Size and Covariate Size



Sample Size and Covariate Size



Tuning Parameters

I try varying five tuning parameters, one at a time. I use DGP2 and fix $n = 1000, d = 10$

1. Sample fraction used in each tree training; (default 0.5)
2. Covariates used in each tree training; (default $\frac{2}{3}d$)
3. Number of trees; (default 2000)
4. Minimun # observations in each terminal node; (default NULL)
5. Regularization parameter λ; (default 0)

Tuning Parameters

1. Try sample fraction $s = 0.1, 0.2, 0.3, 0.4, 0.5$

s	MSE	coverage
0.1	0.2811	0.5075
0.2	0.1412	0.767
0.3	0.1067	0.8425
0.4	0.08107	0.9065
0.5	0.07753	0.914

Tuning Parameters

2. Try # covariates in each tree training $t = 4, 5, 6, 7, 8$

t	MSE	coverage
4	0.1157	0.833
5	0.09674	0.883
6	0.0898	0.89
7	0.07713	0.92
8	0.07511	0.917

Tuning Parameters

3. Try $\#$ trees $b = 500, 1000, 2000, 4000, 6000$

b	MSE	coverage
500	0.08462	0.96
1000	0.07933	0.9395
2000	0.08467	0.9
4000	0.07554	0.8915
6000	0.07713	0.8835

Tuning Parameters

4. Try minimun node size = 0, 10, 20, 40, 80

size	MSE	coverage
0	0.08151	0.902
10	0.0824	0.7995
20	0.0935	0.7225
40	0.08928	0.6915
80	0.1123	0.564

Tuning Parameters

5. Try $\lambda = 0.1, 1, 5, 10, 100$

lambda	MSE	coverage
0.1	0.08055	0.8975
1	0.08013	0.8995
5	0.09256	0.88
10	0.09358	0.883
100	0.1325	0.82