POLS 904 Final Project

Monte Carlo Simulation on Causal Forest

Jiacheng He

December 18, 2017

Introduction

In social science, researchers might be interested to estimate the effect of a binary treatment (either treated or not treated). In experimental setting, individuals are randomly assigned into a control group and a treated group. Formally, denote y_i as the outcome variable, W_i as the treatment assignment variable ($W_i = 1$ if in the treated group, $W_i = 0$ if in the control group), and X_i as a set of observed covariates (e.g. age, gender, race, eduction, etc). Then the researcher can estimate the classical linear model:

$$Y_i = \tau W_i + X_i\beta + \epsilon_i \tag{1}$$

Here $\tau = E[Y_i|W_i = 1] - E[Y_i|W_i = 0]$ is intepreted as the average treatment effect (ATE) across all individuals. X_i is included into the regression to make sure of unconfoundedness and to reduce the variance of the estimator $\hat{\tau}$.

But sometimes researchers might want to go beyond ATE, and try to further estimate heterogeneous treatment effect and identify the subgroup of the population who will benefit the most (or least) from the treatment. One approach is to estimate the conditional average treatment effect (CATE) $\tau(X_i) = E[Y_i|X_i, W_i = 1] - E[Y_i|X_i, W_i = 0]$, that is, express the treatment effect τ as a function of the observed covariates.

Causal forest developed by Wager and Athey (2017) aims to algorithmically search for the covariate space, identify the subspace where heterogeneity exists, and estimate the CATE in these subspaces. It is very similar to the popular random forest method. Wager and Athey (2017) also derived asymptotic distribution of the causal forest estimator so that statistical inference and hypothesis test become feasible when adopting this forest based method.

In this project, I run Monte Carlo simulation on the causal forest to examine its finite sample perfomance, such as the mean squared error (MSE) and the confidence interval coverage rate.

Model Framework

Consider a simple additive model. Define $\tau(X_i) = E[Y_i|X_i, W_i = 1] - E[Y_i|X_i, W_i = 0]$ We will have such relationship:

$$E[Y_i|X_i, W_i = 1] = E[Y_i|X_i, W_i = 0] + W_i \cdot \tau(X_i)$$

With some derivation, we will have such relationship:

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

where

$$\begin{aligned} \tau(X_i) &= E[Y_i|X_i, W_i = 1] - E[Y_i|X_i, W_i = 0] \\ m(X_i) &= E[Y_i|X_i] \\ e(X_i) &= E[W_i|X_i] \end{aligned}$$

Both $\tau(\cdot)$, $m(\cdot)$, and $e(\cdot)$ are nonparametric functions of the observed covariates X_i.

There are several challenges in nonparametrically estimate the CATE function $\tau(X_i)$. First, in real world application, we never observe the true individual treatment effect τ_i. At each moment, an individual is either in the treated status or in the non-treated status, so we never know what would have happened to the individual if the individual would have shifted his/her status. This is the fundamental problem in causal inference. As a result of the absence of the true τ_i, we can not perform cross validation, which is the routine in predictive machine learning.

Second, the existence of non-constant $m(\cdot)$ and $e(\cdot)$ will tend to confound our estimation, as I showed in the final presentation.

Causal Forest

The estimation algorithm of the causal forest is very closed to the random forest. There are two major divergence:

1. When growing each tree in the causal forest, we place the split at the point $\tilde{x_i}$, which maximizes the difference of $\hat{E}[Y_i|X_i = x_i, W_i = 1] - \hat{E}[Y_i|X_i = x_i, W_i = 0]$ ($\hat{\tau}$) across the two sides of $\tilde{x_i}$. While in the case of random forest we place the split based on $\hat{E}[Y_i|X_i = x_i]$ ($\hat{y}$).
2. When growing each tree, we use half of the training sample to do Step 1 above (placing split, identifying heterogenity covariate subspace), and use the other half of the training sample to calculate the $\hat{\tau}$ (estimation of the CATE in that subspace). Wager and Athey (2017) refer to this criterion as "honest splitting".

Also, similar to random forest, in causal forest algorithm it is not necessary to implement regularization or pruning.

Simulation Setup

In the Monte Carlo simulation experiment, I am interested to see how the algorithm performs as sample size and number of covariate change, under two different senarios: 1. constant treatment effect; 2. heterogeneous treatment effect. I set up two data generating processes (DGP) as:

DGP1 (constant τ)

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

where $f_{beta}^{2,4}(\cdot)$ is the density function of Beta distribution with shape parameters 2 and 4.

DGP2 (heterogeneous τ)

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

Training causal forests also requires setting up tuning parameters, the same as when we train random forests. In this project, I also try to vary different tuning paramters and evaluate the performance of these trained models. The tuning parameters I try are as follows: (1) sample fraction used in growing each tree; (2) covariates used in growing each tree; (3) Number of trees to build the forest; (4) Minimun number of observations in each terminal leaf; (5) Regularization parameter λ.

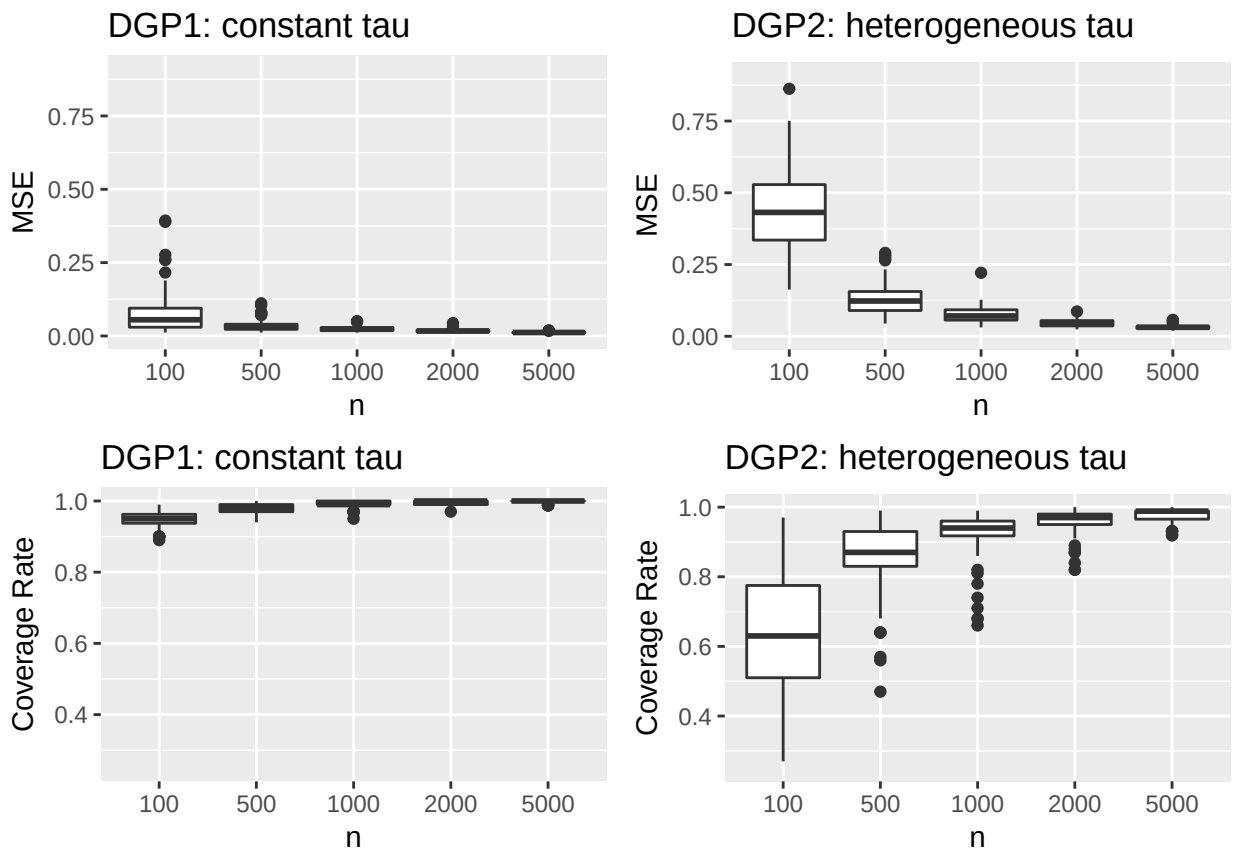


Figure 1: This is a caption

Please keep in mind that cross validation is feasible in treatment effect estimation. So in practice there is no general guidance to select these tuning parameters in the training stage. Evaluation of choices of tuning parameters is only possible when we assume a data generating process and hence know the true τ_i in Monte Carlo simulation. And we can only implement the evaluation in the test set.

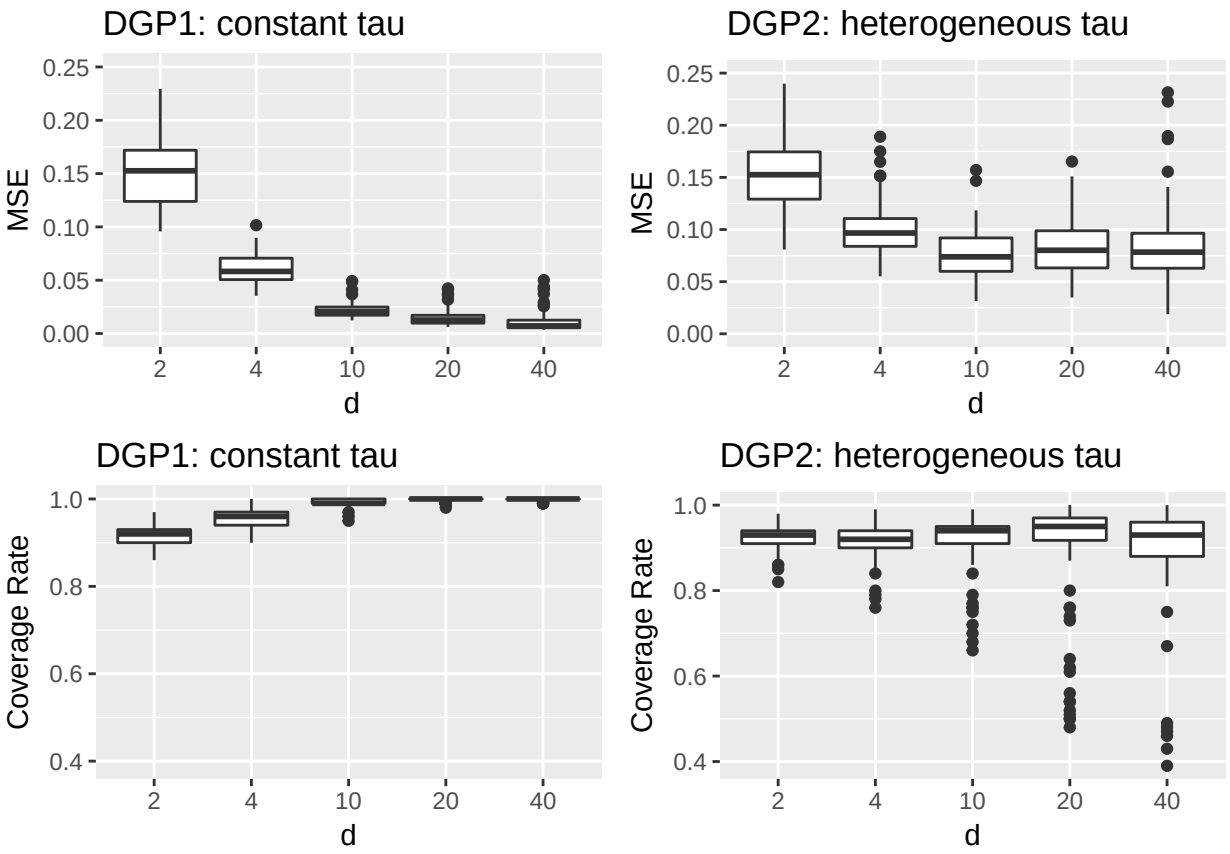
I draw $X_i \sim U(0,1)^d$, $W_i \sim binom(1, e(X_i))$, $\epsilon_i \sim N(0,1)$. (d is the number of covariates). Then I train the causal forest model on a training set, and evaluate the trained model on a test set with 100 data points. For each senario, I replicate it for 100 times. For each replication I generate a new training set, while the test set is invariant for all replications. Then I plot the box plot of the MSE and 95% confidence interval coverage rate.

Result

Sample Size and Number of Covariates

First, I will look at how the performance of the causal forest respond to the change of sample size n and number of covariates d:

1. Fix $d = 10$, try $n = 100, 500, 1000, 2000, 5000$;
2. Fix $n = 1000$, try $d = 2, 4, 10, 20, 40$;



In both data generating processes, only the first two covariates $X1$, $X2$ contribute to the $\tau(\cdot)$ function. Therefore, adding extra covariates is purely adding "noise" to the causal forest algorithm. We should expect the variance of the heterogenous effect estimates increases as d increases.

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

s	MSE	coverage
0.1	0.242	0.535
0.2	0.123	0.87
0.3	0.082	0.92
0.4	0.073	0.94
0.5	0.075	0.94

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

t	MSE	coverage
4	0.102	0.865
5	0.082	0.92
6	0.079	0.92
7	0.07	0.94
8	0.074	0.94

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

b	MSE	coverage
500	0.088	0.97
1000	0.079	0.96
2000	0.076	0.95
4000	0.077	0.94
6000	0.075	0.915

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

size	MSE	coverage
0	0.068	0.94
10	0.066	0.89
20	0.066	0.9
40	0.063	0.915
80	0.076	0.905

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

lambda	MSE	coverage
0	0.067	0.95
0.1	0.069	0.94
1	0.066	0.94
5	0.076	0.925
10	0.079	0.92