The Causal Tree Estimator for Heterogeneous Treatment Effects: Optimal Data Splitting Rules in Small Samples

Nomin Margad-Erdene

CEU

June 9, 2021

Outline

1. Introduction
   - Heterogeneous Treatment Effects
   - Causal Tree Setup
   - Honest Split
   - Modified Causal Tree: θ
2. Estimation Design
   - Data Generation Process
   - Monte-Carlo Simulations
   - Reported Statistics
3. Results
   - Robustness Checks
4. Conclusion

Heterogeneous Treatment Effects

- HTE estimation is applied in:
  - Drug R&D
  - Policy interventions
  - Marketing, advertisement, A/B tests[1]
- Identifying the groups:
  - High dimensional data
  - Avoid data mining to manipulate results
- Other tree-based supervised machine learning algorithms:
  - Bayesian Additive Regression Trees (Green and Kern, 2012)
  - Minimum Impurity Decision Assignment Trees (Laber and Zhao, 2015)
  - Decision Lists (Lakkaraju and Rudin, 2017)
  - Random Forests (Foster et. al, 2011)

[1] A/B testing: comparing the user response to two versions of a website or advertisement, by showing the variants to different people

Causal Tree Setup

Model introduced by Athey and Imbens (2016)

- N i.i.d observations indexed as $i = 1, 2, ..., N$
- Randomly assigned a binary treatment $D_i \in \{0, 1\}$
- Potential outcome model:

$$Y_i(D_i) = \begin{cases} Y_i(0) & \text{if } D_i = 0 \\ Y_i(1) & \text{if } D_i = 1 \end{cases} \tag{1}$$

- X_i is a $(N \times K)$ matrix of covariates or features
- Conditional Average Treatment Effect:

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$$

Causal Tree Setup

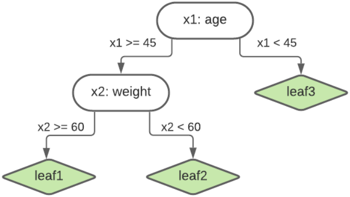

Figure 1: Example of a tree based method

- Decision Tree: conditional expectation
  - $\mu(x, \Pi) = \mathbb{E}[Y_i | X_i \in I(x, \Pi)]$
- Causal Tree: conditional average treatment effect
  - $\mu(D, x, \Pi) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i \in I(x, \Pi)]$

Causal Tree Setup: Honest Split

1. 2 mutually exclusive subsamples S_{Tr} with N^{Tr} observations and S_{Est} with N^{Est} observations
2. Use S_{Tr} to train the tree (find partitions)
3. Use S_{Est} to calculate the treatment effect.

Modified Causal Tree: θ

- So far: $N^{Tr} = N^{Est}$
- $\theta \in (0, 1)$ represents the share of observations allocated to the estimation subsample
- If $N^{Tr} + N^{Est} = 100$ and $\theta = 0.7$, $N^{Est} = 70$
- I test for $\theta \in [0.2, 0.8]$ with step size 0.1

Data Generation Process

- Three DGPs with 2, 4 and 8 distinct conditional average treatment effects respectively.
- $$\text{for D} = \{0{,}1\} \begin{cases} Pr(D_i = 1) = 0.5 \\ Pr(D_i = 0) = 0.5 \end{cases} \tag{2}$$
- Potential outcome structure

$$Y_i = D \cdot \gamma(X_i) + \eta(X_i) + \epsilon_i \tag{3}$$

- $\gamma(x)$: treatment effect, $\eta(x)$: mean effect.
- $X_i \sim \mathcal{N}(0, 1)$ is a $(N \times K)$ vector of covariates independent of ϵ_i.
- $\epsilon_i \sim \mathcal{N}(0, Var(e))$
- $Var(\epsilon) = [0.01, 1.0, 2.5]$, $N^{Tr} + N^{Est} = [500, 300, 100]$

Data Generation Process

- **DGP 1:** $Y_i = \underbrace{-1.5D + 3D \cdot \mathbb{I}_{\{x_1 \geq 0\}}}_{\text{treatment effect}} + \underbrace{\sum_{k=2}^{5} x_k}_{\text{mean effect}} + e_i$

	$x_1 \geq 0$	$x_1 < 0$
τ	1.5	-1.5

- **DGP 2:**

$$Y_i = \underbrace{-2D + 3D \cdot \mathbb{I}_{\{x_1 \geq 0\}} + D \cdot \mathbb{I}_{\{x_2 \geq 0\}} + D \cdot \mathbb{I}_{\{x_1 \geq 0 \text{ \& } x_2 \geq 0\}}}_{\text{treatment effect}} + \underbrace{\sum_{k=3}^{5} x_k}_{\text{mean effect}} + e_i$$

	$x_1 \geq 0$	$x_1 < 0$
$x_2 \geq 0$	3	-1
$x_2 < 0$	1	-2

Data Generation Process

- **DGP 3:**

$$Y_i = \underbrace{-5D + 6D \cdot \mathbb{I}_{\{x_1 \geq 0\}} + 2.5D \cdot \mathbb{I}_{\{x_2 \geq 0\}} + 1.5D \cdot \mathbb{I}_{\{x_3 \geq 0\}}}_{\text{treatment effect}} + \underbrace{\sum_{k=4}^{5} x_k}_{\text{mean effect}} + e_i$$

-

	$x_3 \geq 0$		$x_3 < 0$	
	$x_1 \geq 0$	$x_1 < 0$	$x_1 \geq 0$	$x_1 < 0$
$x_2 \geq 0$	5	-1	3.5	-2.5
$x_2 < 0$	2.5	-3.5	1	-5

Monte-Carlo Simulations

FLOWCHART

Table 1: Input Parameters

Name	Param.	Values	Description
$N^{Tr} + N^{Est}$	**n**	500, 300, 100	number of observations in $N^{Tr} + N^{Est}$ sample
N^{Te}	**n_test**	5000	number of observations in test sample
θ	**est_size**	[0.2 : 0.1 : 0.8]	share of **n** devoted to the estimation subsample
R	**reps**	500	Monte-Carlo repetitions
$Var(e)$	**var_e**	0.01, 1.0, 2.5	variance of error term in DGP

Monte-Carlo and Modified Causal Tree Scripts **here**

Reported Statistics

- MSE, Bias and Variance of Conditional Average Treatment Effects (CATE):

$$\widehat{BIAS}_{CATE} = \frac{1}{R}\sum_{r=1}^{R}(\widehat{CATE}_r(\mathbf{X}) - CATE_{True})$$

$$\widehat{VAR}_{CATE} = \frac{1}{R}\sum_{r=1}^{R}(\widehat{CATE}_r(\mathbf{X}) - \overline{\widehat{CATE}})^2$$

$$\widehat{MSE}_{CATE} = \widehat{BIAS}^2_{CATE} + \widehat{VAR}_{CATE}$$

Reported Statistics

- Total MSE, Toal Bias and Total Variance of Individual Treatment Effects:

$$\widehat{BIAS}^2_T = \frac{1}{N^{Te}} \sum_{i \in S_{Te}} (\bar{\hat{\tau}}(X_i) - \tau_i(X_i))^2$$

where:

$$\bar{\hat{\tau}}(X_i) = \frac{\sum_{r=1}^{R} \hat{\tau}_{ir}}{R} \quad \forall \;\; i \in S_{Te}$$

$$\widehat{VAR}_T = \frac{1}{N^{Te}} \sum_{i \in S_{Te}} \widehat{V(\tau)}_i(X_i)$$

where:

$$\widehat{V(\tau)}_i(X_i) = \frac{\sum_{r=1}^{R} (\hat{\tau}_{ir} - \bar{\hat{\tau}}_i)^2}{R} \quad \forall \;\; i \in S_{Te}$$

Results

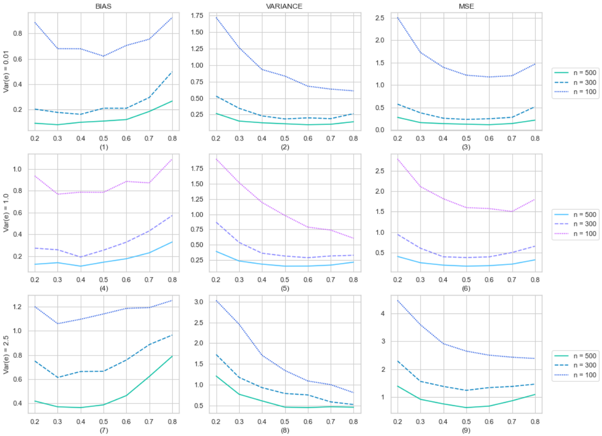

Figure 2: CATE bias, variance and MSE for DGP 1

Results

Table 2: Summary of results (Design 1 with 2 ATEs)

N^{Tr+Est}	500			300			100		
$Var(e)$	0.01	1.00	2.50	0.01	1.00	2.50	0.01	1.00	2.50
Minimum MSE of ATE	0.113	0.170	0.618	0.230	0.380	1.237	1.180	1.503	2.380
θ where MSE is minimzed	0.6	0.5	0.5	0.5	0.5	0.5	0.6	0.7	0.8
θ where variance is minimzed	0.6	0.5	0.6	0.5	0.6	0.8	0.8	0.8	0.8
θ where bias is minimzed	0.3	0.4	0.4	0.4	0.4	0.3	0.5	0.3	0.3
SD of MSE ($\theta \in [0.2 : 0.8]$)	0.059	0.087	0.268	0.139	0.203	0.354	0.473	0.446	0.773
SD of MSE ($\theta \in [0.3 : 0.7]$)	0.018	0.033	0.125	0.059	0.097	0.117	0.230	0.247	0.472

Robustness Checks

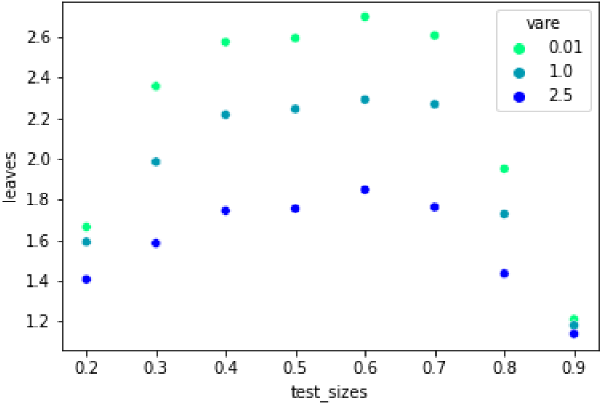

Figure 3: Average number of estimated leaves for cases when $N^{Tr+Est} = 100$

Conclusion

- In large samples: $\theta \in [0.3, 0.7]$
- In small samples and in data sets with noise: $\theta \in [0.5, 0.7]$
- Optimal to set $\theta = 0.6$ in small samples

Limitations:

- Adjustment to the existing estimation method
- Only 3 DGPs of same type (may not be generalizable)