Treatment Transfer Trees

Nandan Rao
Supervisor: Caterina Calsamiglia
Tutor: Francesc Trillas

2020

Outline

1. Motivation
2. Problem setup
3. Simulation setup
4. Model
5. Results

Motivation

Imagine you are a policymaker in Bosnia.

You recieve a proposal to fund a program providing micro-credit loans to entrepreneurs in your country. You are interested in increasing business revenues of small businesses in your country.

You can find one empirical study, from Mexico, which shows that the effects of a microcredit program on revenue were positive and significant.

What do you do?

Motivation

Now imagine you find 3 more studies, from Morroco, India, and Ethiopia, which show that the effects of microcredit programs on revenue were not significant, but with positive point estimates.

Now what do you do?

External validity

"…from the perspective of policy choice…What matters is the informativeness of a study for policy making, which depends jointly on internal and external validity." (Manski 2012)

"It is our belief that creating a rigorous framework for external validity is an important step in completing an ecosystem for social science field experiments, and a complement to many other aspects of experimentation." (Banjerjee, Chassang, Snowberg 2016)

Methods for joint validity

Qualitative methods (Shadish, Cook, Campbell 2001; Cartwright and Hardy 2013).

Bayesian hiearchical quantile effects (Meager 2018).

Lacking: a method that is quantitative and integrates covariates in the prediction context.

Setup

Individual i described by covariates $x_i \in \mathcal{X}$ and potential outcomes of interest $y_i^1, y_i^0 \in \mathcal{Y}$ and individual treatment effect: $\tau_i := y_i^1 - y_i^0$ for treatment $w_i \in \{0, 1\}$.

Individuals belong to separate **domains** $d \in \mathcal{D}$ with different population distributions $P_d(\tau, X)$.

Denote ATE for domain d as: $\tau(d) := \mathbb{E}_{P_d(\tau)}[\tau]$

Denote CATE for domain d as: $\tau(x, d) := \mathbb{E}_{P_d(\tau)}[\tau | X = x]$

Setup

We have unconfounded samples from $P_d(X, Y, W)$ for domains $\{d_1, \ldots, d_K\}$.

Unconfounding: $P_d(X, Y^w) = P_d(X, Y | W = w)$.

We have samples from $P_{d^*}(X)$ from domain of interest, $d^* \notin \{d_1, \ldots, d_K\}$.

Object of interest: $\tau(d^*) = \int \underbrace{\tau(X, d^*)}_{?} \underbrace{P_{d^*}(X)}_{\text{estimable}} dX$

Setup

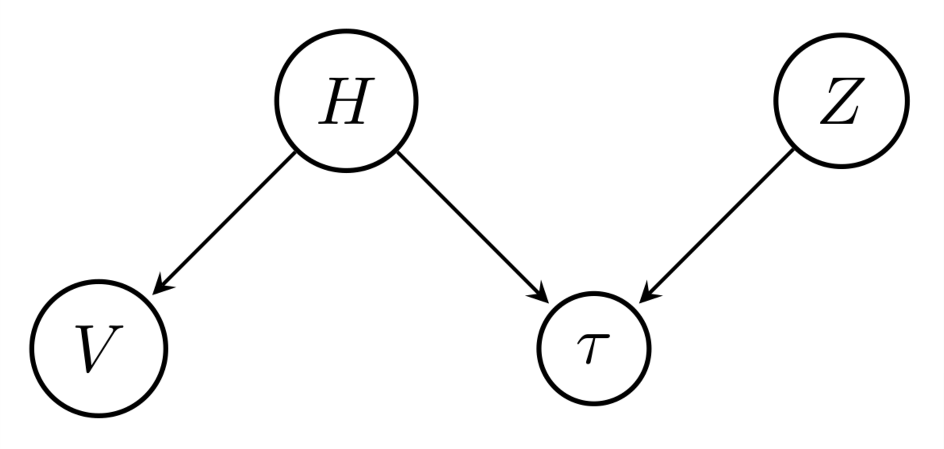
In general, $\tau(x, d^*) \overset{?}{=} \tau(x, d)$.

Goal is to find a feature-representation function $f : \mathcal{X} \to \mathbb{R}^M$ for which $\tau(f(x), d^*) \approx \tau(f(x), d)$.

Strategy: search for a $f(\cdot)$ such that $\tau(f(x), d)$ is stable for $d \in \{d_1, \ldots, d_K\}$.

Assumption: $\{d_1, \ldots, d_K\} \in \mathcal{D}, d^* \in \mathcal{D}$.

Data simulation



Z is a visible interacting variable
H is a latent interacting variable
V is a visible proxy for H

(Pearl and Bareinboim 2014)

$$
\begin{aligned}
y_i &= \tau_i w_i + \epsilon_i \\
\tau_i &= f_\tau(H_i, Z_i, \alpha, \delta) \\
Z_i &= \mathcal{N}(Q_d, R_d) \\
H_i &= \mathcal{N}(A_d, B_d) \\
V_i &= C_d H_i + \mathcal{U}(J_d, K_d)
\end{aligned}
$$

Stability of simulated data

Two candidate models if $f(\cdot)$ consisted of simple variable selection:

$$\mathbb{E}_{P_d(\tau)}[\tau|Z,V] = \int \mathbb{E}_{P_d(\tau)}[\tau|Z,H]\ P(H|V)\ dH$$

$$\mathbb{E}_{P_d(\tau)}[\tau|Z] = \int \mathbb{E}_{P_d(\tau)}[\tau|Z,H]\ P(H)\ dH$$

Given the DGP, stability of the models is determined by stability of $P(H)$ vs. stability of $P(H|V)$.

Model

A decision-tree model based on CART (Breiman et al. 1984) and Causal Trees (Athey and G. Imbens 2016).

A tree-based algorithm provides the following advantages:

1. It optimizes over the discrete space of covariates and their interactions with a local, greedy, forward-backward search.
2. It discretizes the continuous input space of covariates.

Model

Denote $\mathcal{S}$ the space of data samples $\{(x_1, y_1, w_1), \ldots, (x_N, y_N, w_N)\} \in \mathcal{S}$

Define a random sampling function $S : \mathcal{D} \to \mathcal{S}$. Notationally denote an expectation of a generic function $g : \mathcal{S} \to \mathbb{R}$ taken over the realizations of the sampling function as $\mathbb{E}_S[g(S(d))]$.

Model

Define $\mathbf{\Pi}$ as a set of discrete partitions of the covariate space $\{\Pi_1, \ldots, \Pi_P\} \in \mathcal{P}$.

Define a "leaf" function that finds the correct partition for a given observation:

$$\lambda(x, \mathbf{\Pi}) := \{\Pi_j \in \mathbf{\Pi} : x \in \Pi_j\}$$

Define a discretized CATE over a set of partitions $\mathbf{\Pi}$:

$$\tau(x, \mathbf{\Pi}, d) := \mathbb{E}_{P_d(\tau)}[\tau | X \in \lambda(x, \mathbf{\Pi}))]$$

Model

Define a "tree" function that finds the correct partition for a given observation and returns the sample data in that partition, $tree : (\mathcal{X}, \mathcal{P}, \mathcal{S}) \to \mathcal{S}$:

$$tree(x_i, \Pi, s) := \{(x, y, w) \in s : x \in \lambda(x_i, \Pi)\}$$

Let $\bar{y}^w : \mathcal{S} \to \mathbb{R}$ return the sample average of all outcomes where $w_i = w$.

Define the following estimator $\hat{\tau} : (\mathcal{X}, \mathcal{P}, \mathcal{S}) \to \mathbb{R}$:

$$\hat{\tau}(\cdot) = \bar{y}^1(tree(\cdot)) - \bar{y}^0(tree(\cdot))$$

Objective function

Oracle squared loss function is given by:

$$\ell(\hat{\tau}, \tau_i) = (\tau_i - \hat{\tau})^2$$

Expected oracle loss of a partitioning Π, given that the conditional treatment estimator is trained on single source domain d:

$$\mathbb{E}\ell(\Pi) = \mathbb{E}_{X,S}[(\tau_i - \hat{\tau}(x_i, \Pi, S(d)))^2]$$

Where $\tau_i, x_i \in d^*$.

Objective function

Proposition 1

The expected oracle loss of a partitioning, $\mathbb{E}\ell(\Pi)$, is estimable given the following assumptions:

1. $\{d_1, \ldots, d_K\} \in \mathcal{D}, d^* \in \mathcal{D}$.
2. $K \geq 2$

Objective function

$$\mathbb{E}\ell(\Pi) = \mathbb{E}_{X,S}[(\tau_i - \hat{\tau}(x_i, \Pi, S(d)))^2]$$

Expand the square.

$$\mathbb{E}\ell(\Pi) = \mathbb{E}_X[\tau_i^2] + \mathbb{E}_X[-2\mathbb{E}_S[\tau_i \hat{\tau}(x_i, \Pi, S(d))] + \mathbb{E}_S[\hat{\tau}(x_i, \Pi, S(d))^2]]$$

First term does not depend on Π.

Objective function

We can write the last term in terms of sample estimates:

$$
\begin{aligned}
&= \mathbb{E}_S[\hat{\tau}(x_i, \Pi, S(d))^2] \\
&= \mathbb{V}_S[\hat{\tau}(x_i, \Pi, S(d))] + \mathbb{E}_S[\hat{\tau}(x_i, \Pi, S(d))]^2 \\
&\approx \hat{\mathbb{V}}_S[\hat{\tau}(x_i, \Pi, S(d))] + \hat{\tau}(x_i, \Pi, S(d))^2
\end{aligned}
$$

Where $\hat{\mathbb{V}}_S := \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}$ is a conservative upper bound (Imbens and Rubin 2015).

Objective function

Middle term: conditional on the value being within a leaf, τ_i and $\hat{\tau}(x_i, \Pi, S(d))$ become independent:

$$-2\mathbb{E}_{X,S}[\underbrace{\mathbb{E}_X[\tau_i | x_i \in \Pi_j]}_{\approx \hat{\tau}(x_i, \Pi, S(d^*))} \underbrace{\mathbb{E}_S[\hat{\tau}(x_i, \Pi, S(d)) | x_i \in \Pi_j]}_{\approx \hat{\tau}(x_i, \Pi, S(d))}]$$

Pooled estimator

Pooled prediction: $\frac{1}{K}\sum_k \hat{\tau}(x_i, \Pi, S(d_k))$

Objective function:

$$\sum_j P(x_i \in \Pi_j)\Bigg[-2 \min_{k\in\{1,\ldots,P\}} \left(\hat{\tau}(\Pi_j, S(d_k))\frac{1}{P-1}\sum_{m\neq k}^{P}\hat{\tau}(\Pi_j, S(d_m))\right) \; + \; \frac{1}{K^2}\sum_k \hat{\mathbb{V}}_S[\hat{\tau}(x_i, \Pi, S(d_k))] + \left[\frac{1}{K}\sum_k \hat{\tau}(x_i, \Pi, S(d_k))\right]^2\Bigg]$$

Results

$$\begin{aligned}
f_{\tau}^{linear} &= \alpha V_i + \delta H_i \\
f_{\tau}^{nl-hard} &= \alpha V_i \cdot 1\{V_i > 0\} + \delta H_i \cdot 1\{H_i > 0 \\
f_{\tau}^{nl-simple} &= \alpha \cdot 1\{V_i > 0\} + \delta \cdot 1\{H_i > 0\} \\
f_{\tau}^{nl-stacked} &= \sum_{j=-2}^{2} \alpha \cdot 1\{V_i > j\} + \sum_{j=-2}^{2} \delta \cdot 1\{H_i > j\}
\end{aligned}$$

Results

Transfer R-squared:

$$R^2_{tr} := 1 - \frac{MSE}{\frac{1}{N}\sum_i (\tau_i - \bar{\tau}_{\{d_1,\ldots,d_K\}})^2}$$

Where $\bar{\tau}_{\{d_1,\ldots,d_K\}}$ is the oracle average treatment effect on the pooled source domains.

Results - UP (K=4)

measure	model	linear	nl-hard	nl-simple	nl-stacked
r2t-0.1	CT	-1.07	-0.63	-0.28	-1.07
r2t-0.1	TTT-M	-0.45	-0.04	0.14	-0.19
r2t-0.5	CT	-0.02	0.08	0.33	-0.02
r2t-0.5	TTT-M	0.13	0.16	0.43	0.10
CI-coverage	CT	0.17	0.23	0.16	0.18
CI-coverage	TTT-M	0.82	0.91	0.93	0.86

Results - FS (K=4)

measure	model	linear	nl-hard	nl-simple	nl-stacked
r2t-0.1	CT	0.49	0.42	0.48	0.40
r2t-0.1	TTT-M	0.43	0.37	0.46	0.24
r2t-0.5	CT	0.52	0.45	0.51	0.44
r2t-0.5	TTT-M	0.48	0.41	0.49	0.33
CI-coverage	CT	0.76	0.75	0.72	0.74
CI-coverage	TTT-M	1.00	1.00	1.00	1.00

Results - UP (K=3)

measure	model	linear	nl-hard	nl-simple	nl-stacked
r2t-0.1	CT	-1.47	-0.70	-0.25	-1.05
r2t-0.1	TTT-M	-1.03	-0.35	0.09	-0.49
r2t-0.5	CT	-0.10	0.06	0.29	-0.03
r2t-0.5	TTT-M	0.10	0.15	0.43	0.08
CI-coverage	CT	0.17	0.23	0.18	0.17
CI-coverage	TTT-M	0.73	0.87	0.87	0.76

Results - UP (K=2)

measure	model	linear	nl-hard	nl-simple	nl-stacked
r2t-0.1	CT	-1.90	-1.15	-0.53	-1.66
r2t-0.1	TTT-M	-1.53	-0.88	-0.30	-1.30
r2t-0.5	CT	-0.26	-0.00	0.20	-0.25
r2t-0.5	TTT-M	0.01	0.10	0.32	-0.02
CI-coverage	CT	0.12	0.20	0.17	0.13
CI-coverage	TTT-M	0.58	0.77	0.73	0.58