Causal Inference in the Age of Big Data

Jasjeet Sekhon

UC Berkeley

November 14, 2017

Causal Inference and Big Data

- Measuring human activity has generated massive datasets with granular population data: e.g.,
  - Browsing, search, and purchase data from online platforms
  - Internet of things
  - Electronic medical records, genetic markers
  - Administrative data: schools, criminal justice, IRS
- Big in size and breadth: wide datasets
- Data can be used for personalization of treatments, creating markets, modeling behavior
- Many inferential issues: e.g., heterogeneity, targeting optimal treatments, interpretable results

ML Prediction versus Causal Inference

- Causal Inference is like a prediction problem: but predicting something we don't directly observe and possibly cannot estimate well in a given sample
- ML algorithms are good at prediction, but have issues with causal inference:
  - Interventions imply counterfactuals: response schedule versus model prediction
  - Validation requires estimation in the case of causal inference
  - Identification problems not solved by large data
  - Predicting the outcome mistaken for predicting the causal effect
    - targeting based on the lagged outcome

Classical Justifications Versus ML Pipelines

Two different justifications for statistical procedures:

1. (classical) statistical theory:
   it works because we have **relevant** theory that tells us it should
   Hopefully, this is not simply: "Assume that the data are generated by the following model …" (Brieman 2001)

2. Training/test loop:
   it works because we have validated against ground truth and it works

Classical Justifications Versus ML Pipelines

Two different justifications for statistical procedures:

1. (classical) statistical theory:
   it works because we have **relevant** theory that tells us it should
   Hopefully, this is not simply: "Assume that the data are generated by the following model ..." (Brieman 2001)

2. Training/test loop:
   it works because we have validated against ground truth and it works

On the **normal distribution**:

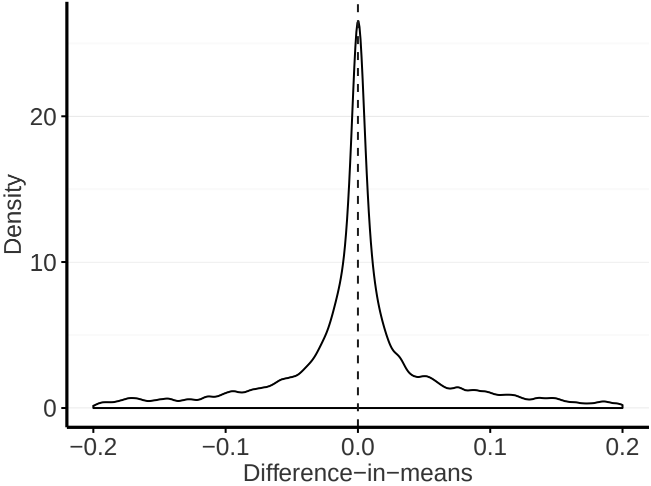
> *"Everyone believes in it: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact." — Henri Poincaré (quoted by de Finetti 1975)*

Even Classical Justifications Should be Validated

- Question: coverage for the population mean. Is $n = 1000$ enough?
- Sometimes, no. Not for many metrics, even when they are bounded
- For some metrics, asking for 95% CI results in only 60% coverage
- Data is very irregular
  Many zeros, IQR: 0

$$\frac{p100 - p99}{p99 - p50} > 10,000$$

Distribution of Treatment Effects



Sekhon and Shem-Tov (2017)

Classical Methods Meet Computing Challenges

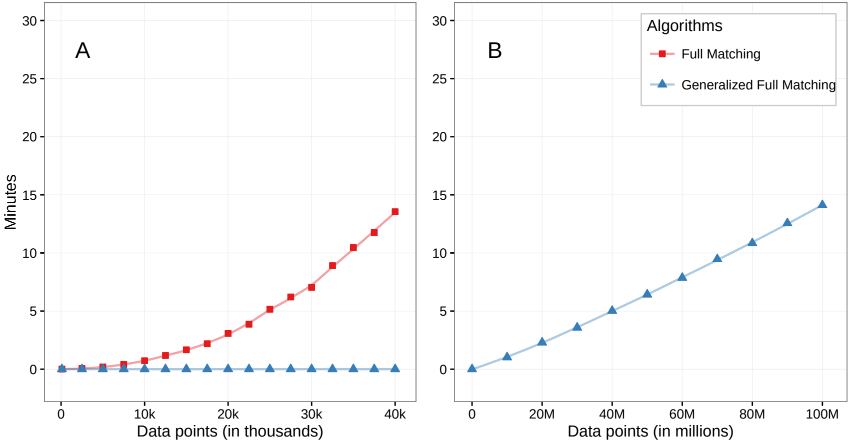
- With big data come small effect sizes: 1e-9
- Some traditional experimental design methods have become computationally infeasible—e.g., blocking, stratification
- Blocking: create strata and then randomize within strata
- Stratification: create strata after randomization
- Polynomial time solution not quick enough. Linearithmic is survivable.

Blocking/Post-Stratification

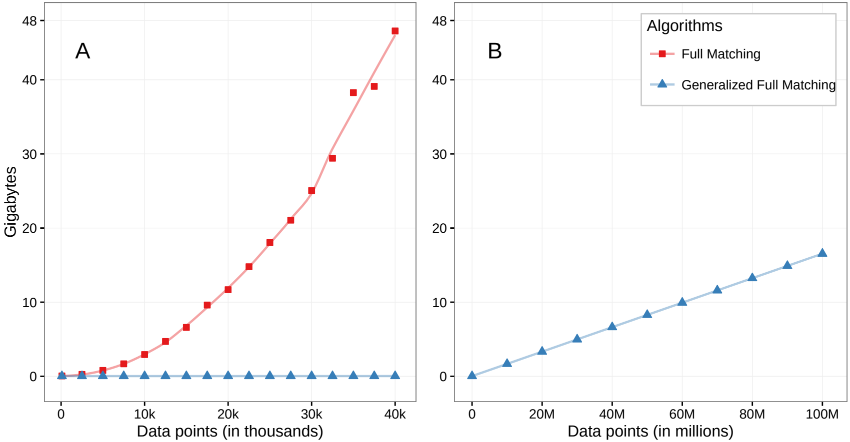
Minimizes the pair-wise Maximum Within-Block Distance: λ
(Higgins, Sävje, Sekhon 2016; Sävje, Higgins, Sekhon 2017)

- Any valid distance metric; triangle inequality
- We prove this is a NP-hard problem
- Ensures good covariate balance by design: approximately optimal: $\leq 4 \times \lambda$
- Works for any number of treatments and any minimum number of observations per block
- It is fast: $O(n \log n)$ expected time
- It is memory efficient: $O(n)$ storage
- Special cases
  1. with one covariate: λ
  2. with two covariates: $\leq 2 \times \lambda$

Time Complexity



Space Complexity



Why We Randomize?

- Unbiased estimator by design
- Make probability statements: *"reasoned basis for inference"* (Fisher)
- Separate design from analysis (Cochran)

Conditional Average Treatment Effect (CATE)

Individual Treatment Effect (ITE): $D_i := Y_i(1) - Y_i(0)$

Let $\hat{\tau}_i$ be an estimator for D_i

$\tau(x_i)$ is the **CATE** for all units whose covariate vector is equal to x_i:

$$CATE := \tau(x_i) := \mathbb{E}\Big[D \Big| X = x_i\Big] = \mathbb{E}\Big[Y(t) - Y(c) \Big| X_i = x_i\Big]$$

Variance of Conditional Average Treatment Effect

$$\color{blue}{CATE} := \tau(x_i) := \mathbb{E}\Big[D\Big|X = x_i\Big] = \mathbb{E}\Big[Y(1) - Y(0)\Big|X_i = x_i\Big]$$

Decompose the MSE at x_i:

$$\mathbb{E}\left[(D_i - \hat{\tau}_i)^2 | X_i = x_i\right] =$$
$$\underbrace{\mathbb{E}\left[(D_i - \tau(x_i))^2 | X_i = x_i\right]}_{\text{Approximation Error}} + \underbrace{\mathbb{E}\left[(\tau(x_i) - \hat{\tau}_i)^2 | X_i = x_i\right]}_{\text{Estimation Error}}$$

- Since we cannot estimate D_i, we estimate the CATE at x_i
- But the error for the CATE is not the same as the error for the ITE

Supplementary

Neighbors mailing

30423-3

For more information: (517) 351-1975
email: etov@grebner.com
Practical Political Consulting
P. O. Box 6249
East Lansing, MI 48826

PRSRT STD
U.S. Postage
PAID
Lansing, MI
Permit # 444

ECRLOT **C050
THE JACKSON FAMILY
9999 MAPLE DR
FLINT MI 48507

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

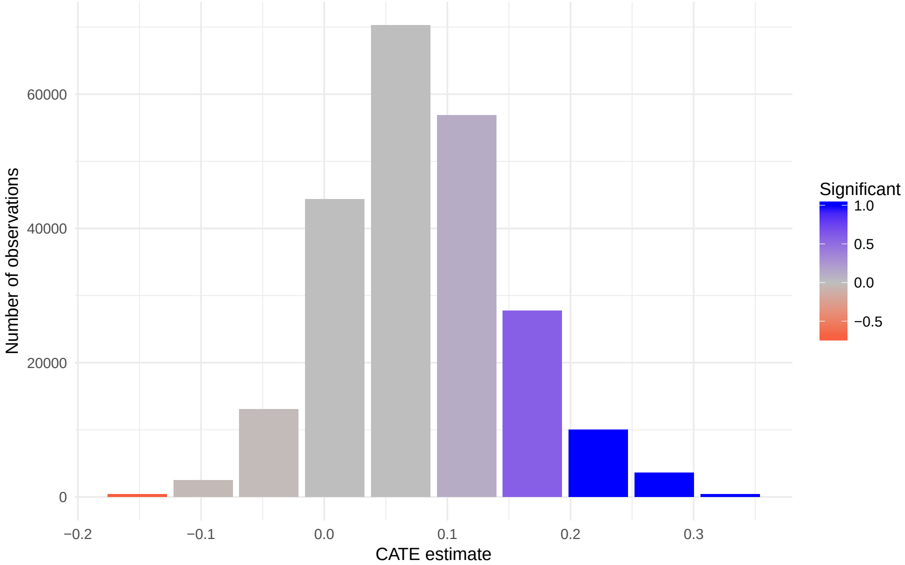
Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

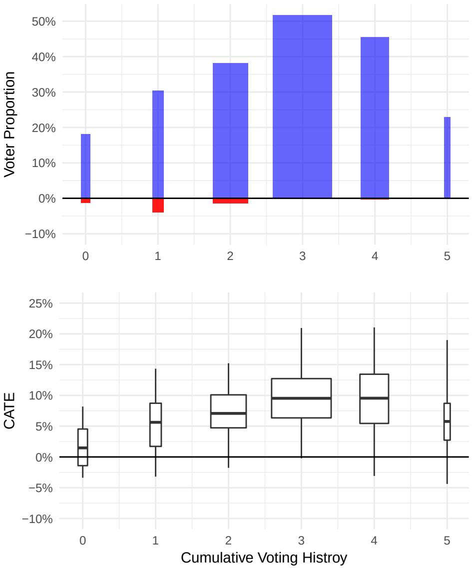
DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______
9999 BRIAN JOSEPH JACKSON		Voted	______
9991 JENNIFER KAY THOMPSON		Voted	______
9991 BOB R THOMPSON		Voted	______
9993 BILL S SMITH			______

GOTV: Social pressure (Künzel, Sekhon, Bickel, Yu 2017)



GOTV: Social pressure (Künzel, Sekhon, Bickel, Yu 2017)



How to estimate the CATE?

Meta–learners

A meta–learner decomposes the problem of estimating the CATE into several sub–regression problems. The estimator which solve those sub–problems are called **base–learners**

- Flexibility to choose base–learners which work well in a particular setting
- Deep Learning, (honest) Random Forests, BART, or other machine learning algorithms

Estimators for the CATE

$$
\begin{aligned}
\tau(x) &= \mathbb{E}[Y(1) - Y(0) | X = x] \\
&= \mathbb{E}[Y(1) | X = x] - \mathbb{E}[Y(0) | X = x]
\end{aligned}
$$

Estimators for the CATE

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\ &= \mu_1(x) - \mu_0(x)\end{aligned}$$

T–learner

1.) Split the data into control and treatment group,

2.) Estimate the response functions separately,

$$\begin{aligned}\hat{\mu}_1(x) &= \hat{\mathbb{E}}[Y^{obs}|X = x, W = 1] \\ \hat{\mu}_0(x) &= \hat{\mathbb{E}}[Y^{obs}|X = x, W = 0],\end{aligned}$$

3.) $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$

Estimators for the CATE

$$
\begin{aligned}
\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\
&= \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\
&= \mu_1(x) - \mu_0(x)
\end{aligned}
$$

T–learner

1.) Split the data into control and treatment group,

2.) Estimate the response functions separately,

$$
\begin{aligned}
\hat{\mu}_1(x) &= \hat{\mathbb{E}}[Y^{obs}|X = x, W = 1] \\
\hat{\mu}_0(x) &= \hat{\mathbb{E}}[Y^{obs}|X = x, W = 0],
\end{aligned}
$$

3.) $\hat{\tau}(x) := \hat{\mu}_1(x) - \hat{\mu}_0(x)$

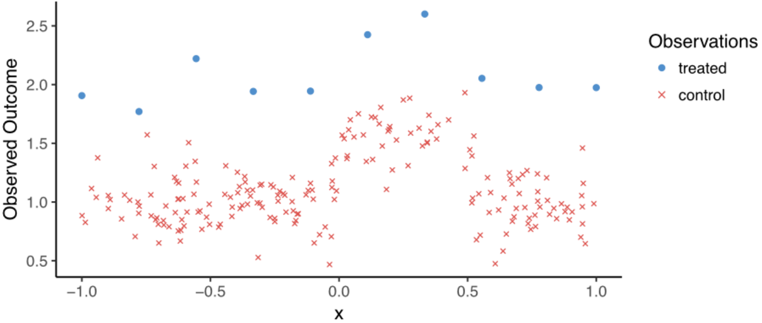
S–learner

1.) Use the treatment assignment as a usual variable without giving it any special role and estimate

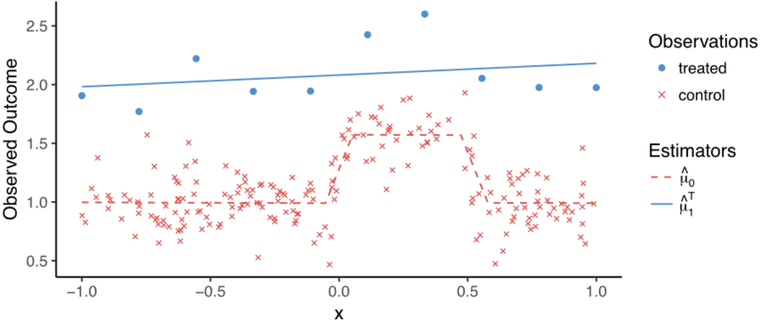
$$\hat{\mu}(x, w) = \hat{\mathbb{E}}[Y^{obs}|X = x, W = w]$$

2.) $\hat{\tau}(x) := \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$

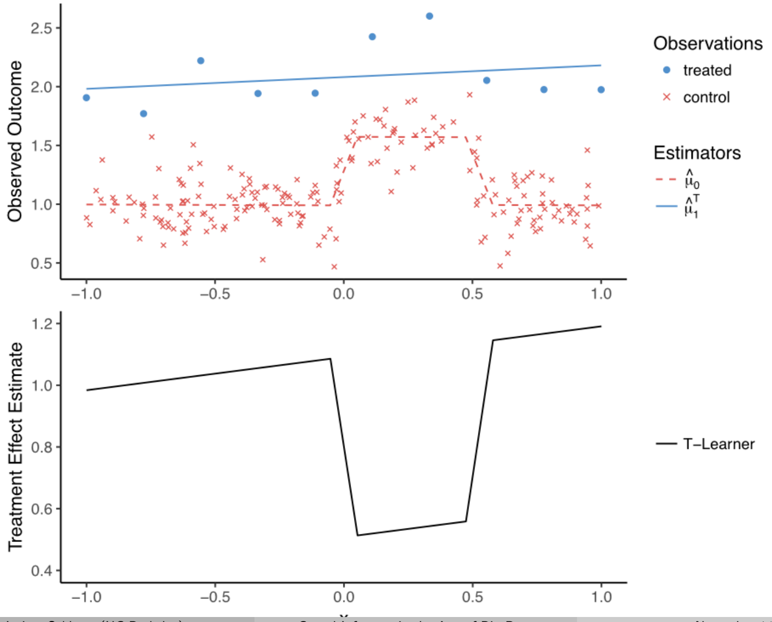
Motivating X



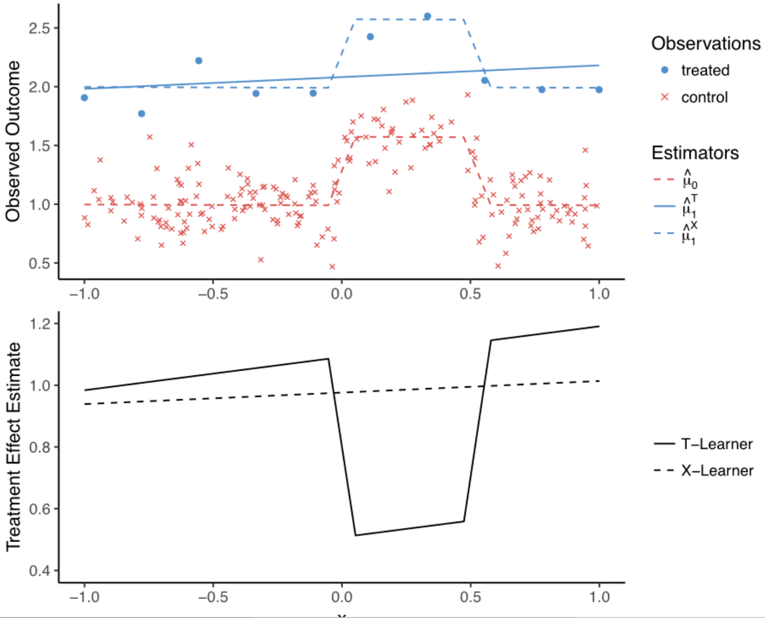
Motivating X



Motivating X



Motivating X



Definition of the X–learner

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y(1) - Y(0)|X = x] \\ &= \mathbb{E}[Y(1) - \mu_c(x)|X = x]\end{aligned}$$

with $\mu_c(x) = \mathbb{E}[Y(0)|X = x]$.

X–learner

1.) Estimate the control response function:

$$\hat{\mu}_c(x) = \hat{\mathbb{E}}[Y(0)|X = x],$$

2.) Define the **imputed ITE**:

$$\tilde{D}^1_i := Y_i(1) - \hat{\mu}_c(X_i(1)),$$

3.) Estimate the CATE:

$$\hat{\tau}(x) = \hat{\mathbb{E}}[\tilde{D}^1|X = x].$$

Definition of the X–learner

Algorithm 1 X–learner

procedure X–LEARNER(X, Y, W)

$\hat{\mu}_c = M_1(Y^0 \sim X^0)$ ▷ Estimate response function

$\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_c(X_i^1)$ ▷ Impute ITE

$\hat{\tau}_1 = M_3(\tilde{D}^1 \sim X^1)$ ▷ Estimate CATE

end procedure

Definition of the X–learner

Algorithm 2 X–learner

procedure $\text{X–L}\textsc{earner}(X, Y, W)$

$\hat{\mu}_c = M_1(Y^0 \sim X^0)$ ▷ Estimate response function
$\hat{\mu}_t = M_2(Y^1 \sim X^1)$

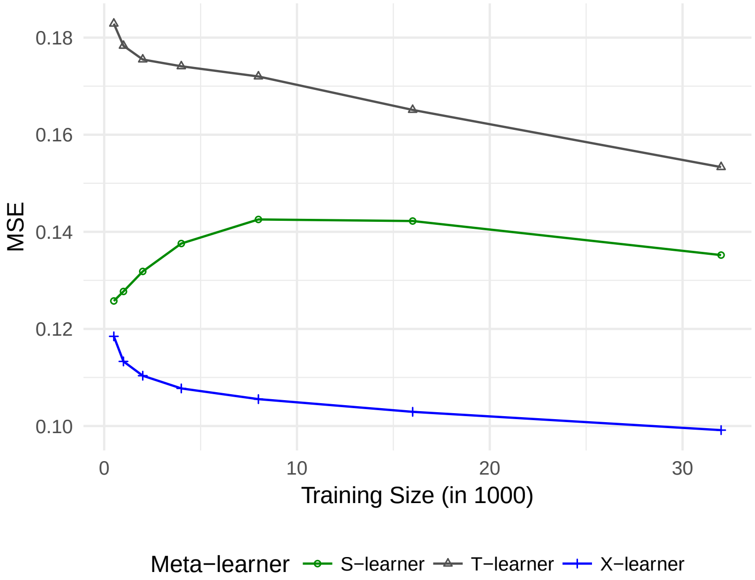
$\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_c(X_i^1)$ ▷ Impute ITE
$\tilde{D}_i^0 := \hat{\mu}_t(X_i^0) - Y_i^0$

$\hat{\tau}_1 = M_3(\tilde{D}^1 \sim X^1)$ ▷ Estimate CATE
$\hat{\tau}_0 = M_4(\tilde{D}^0 \sim X^0)$

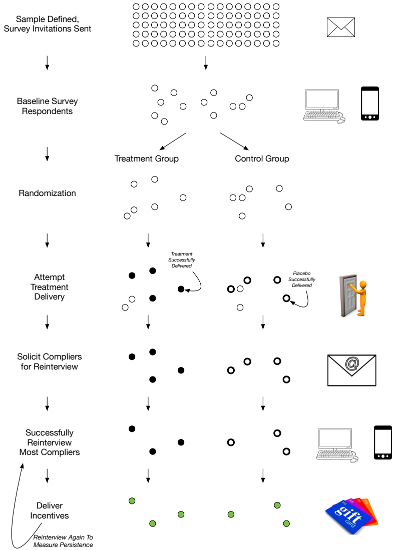
$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$ ▷ Average

end procedure

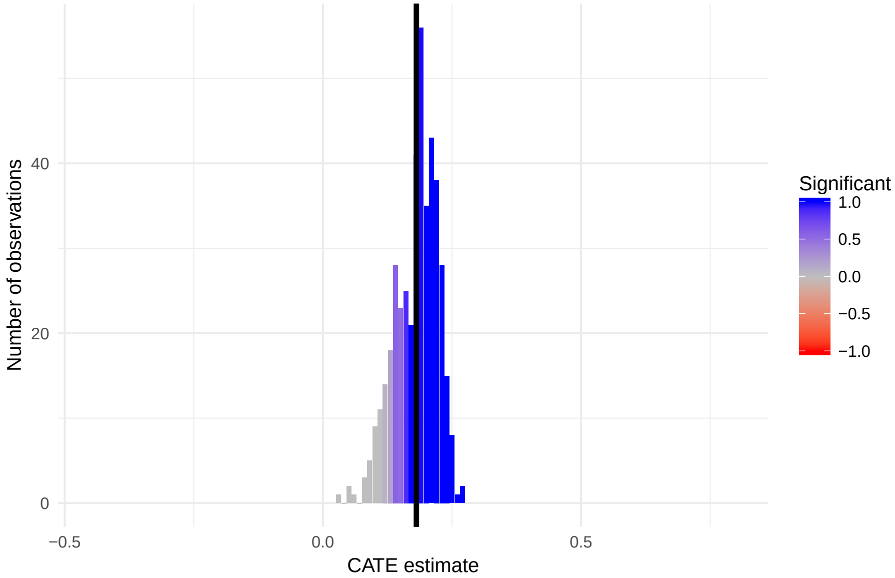
Data Simulation: Social pressure and Voter Turnout



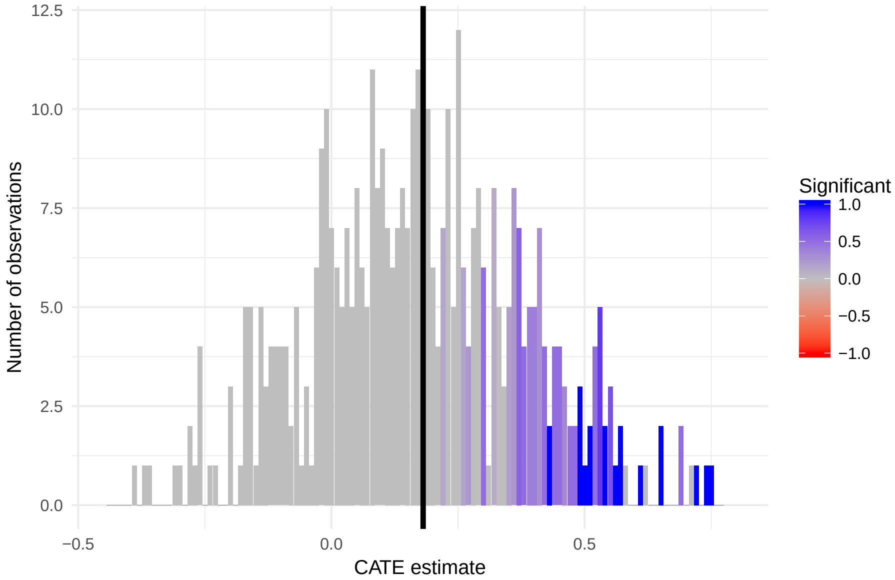
Design for Persuasion Experiments (Broockman, Kalla, Sekhon 2016)



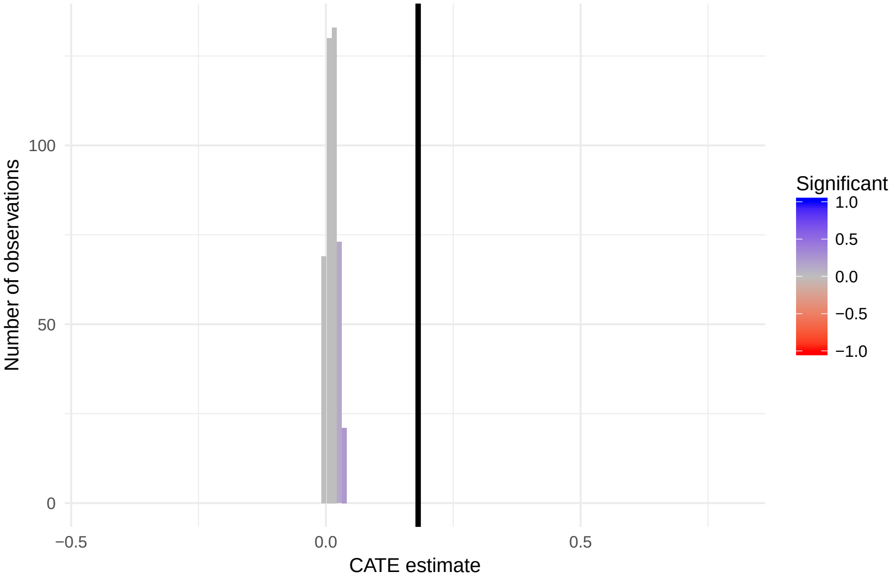
Reducing Transphobia: X–RF



Reducing Transphobia: T–RF



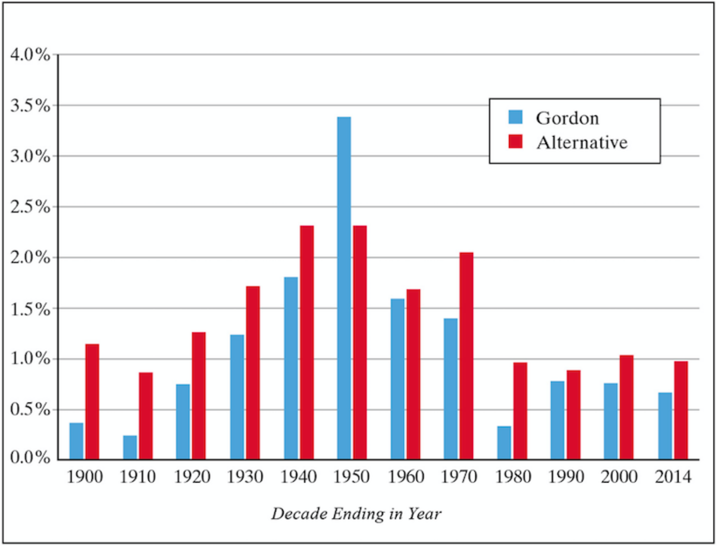
Reducing Transphobia: S–RF



Thoughts About Policy

- Algorithmic Pass and innovation
- Productivity, growth, and regulation/capture
- Adverse selection into policy discussions

Productivity Growth (Nordhaus 2016)



My Collaborators

- Peter Bickel
- David Broockman
- Michael Higgins
- Joshua Kalla
- Sören Künzel
- Fredrik Sävje
- Yotam Shem-Tov
- Bin Yu

Properties of the X–learner: Setup for Theory

A model for estimating the CATE

$$X \sim \lambda$$
$$W \sim \mathrm{Bern}(e(X))$$
$$Y(0) = \mu_0(X) + \varepsilon(0)$$
$$Y(1) = \mu_1(X) + \varepsilon(1)$$

- If τ satisfies some regularity conditions (e.g. sparsity or smoothness), it can be directly exploited in the second base–learner
- This effect is in particular strong when μ_0 can be estimated very well
- Or when the error when estimating $\mu_0(x_i)$ is uncorrelated from the error when estimating $\mu_0(x_j)$ for $i \neq j$

Theorem 1

Künzel, Sekhon, Bickel, Yu 2017

Assume we observe m control and n treatment units,

1.) Ignorability holds: $(Y(0), Y(1)) \perp W | X$

2.) The treatment effect is linear, $\tau(x) = x^T \beta$

3.) There exists an estimator $\hat{\mu}_0$ with $\mathbb{E}[(\mu_0(x) - \hat{\mu}_0(x))^2] \leq C_x^0 m^{-a}$

Then the X-learner with $\hat{\mu}_0$ in the first stage, OLS in the second stage, achieves the parametric rate in n,

$$\mathbb{E}\left[\|\tau(x) - \hat{\tau}_X(x)\|^2\right] \leq C_x^1 m^{-a} + C_x^2 n^{-1}$$

If there are a many control units, such that $m \asymp n^{1/a}$, then

$$\mathbb{E}\left[\|\tau(x) - \hat{\tau}_X(x)\|^2\right] \leq 2C_x^1 n^{-1}$$

Conjecture

Conjecture about the Minimax rates of the X–learner

If the response functions can be estimated at a particular rate a_μ, the CATE can be estimated at a rate of a_τ, the right choice of base learners, and some additional assumptions, then the two parts of the X–learner will achieve the rates of:

$$\hat{\tau}_0 \in \mathcal{O}(m^{-a_\tau} + n^{-a_\mu})$$

$$\hat{\tau}_1 \in \mathcal{O}(m^{-a_\mu} + n^{-a_\tau})$$

Theorem 2

Theorem covers the case when estimating the CATE function is not beneficial

Künzel, Sekhon, Bickel, Yu 2017

X–learner is minimax optimal for a class of estimators using KNN as the base leaner.
Assume:

- Outcome functions are Lipschitz continuous
- CATE function has no simplification
- Features are uniformly distributed $[0, 1]^d$

The fastest possible rate of convergence for this class of problems is:

$$\mathcal{O}\left(\min(n_0, n_1)^{-\frac{1}{2+d}}\right)$$

- The speed of convergence is dominated by the size of the smaller assignment group
- In the worst case, there is nothing to learn from the other assignment group

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 =$$

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 = \sigma^2 = \alpha$$

With one data point?

Individual Treatment Effects: Information Theory Bound

$Y_u \sim P = N(\mu, \sigma^2)$, and we want to predict a new Y_i.
Our expected risk with infinite data is:

$$\mathbb{E}(\mu - Y_i)^2 = \sigma^2 = \alpha$$

With one data point?

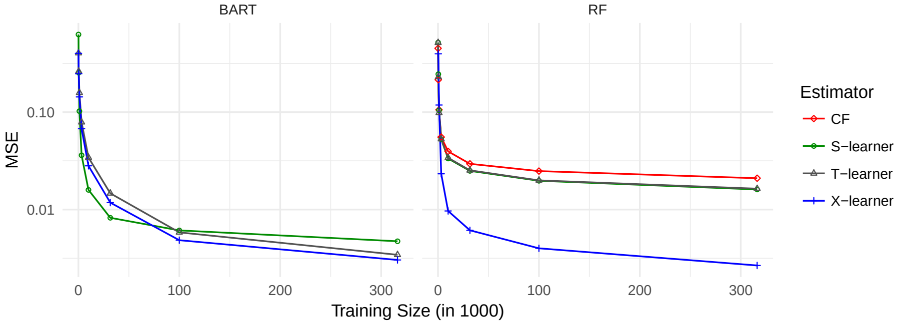
$$\begin{aligned} E(Y_i - Y_u)^2 &= E(Y_i - \mu + Y_u - \mu)^2 \\ &= E(Y_i - \mu)^2 + E(Y_u - \mu)^2 \\ &= 2\sigma^2 \\ &= 2\alpha \end{aligned}$$

General results for Cover-Hart class, which is a convex cone (Gneiting, 2012)
Back to CATE

Reducing Transphobia: Simulation

Algorithm	RMSE	Bias
X–RF	1.102	0.0122
T–RF	1.090	0.0110
S–RF	1.207	-0.1073

Complex Treatment Effect



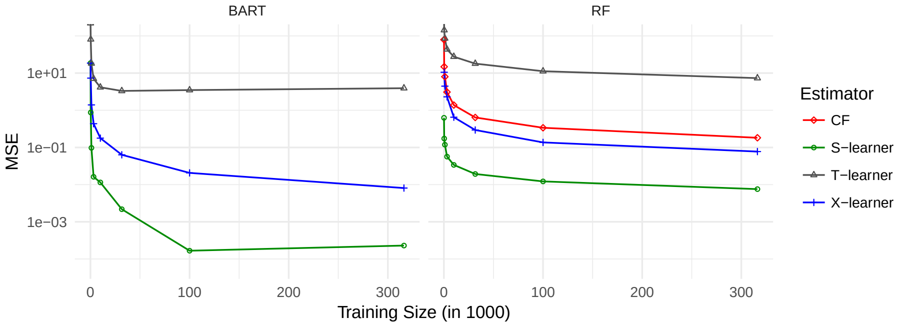
Complex Setting (WA, 2)

$$\mu_1(x) = \frac{1}{2}\eta(x_1)\eta(x_2) \text{ with } \eta(x) = 1 + \frac{1}{1 + e^{-20(x-1/3)}}$$

$$\mu_0(x) = -\mu_1(x)$$

$$e(x) = 0.5$$

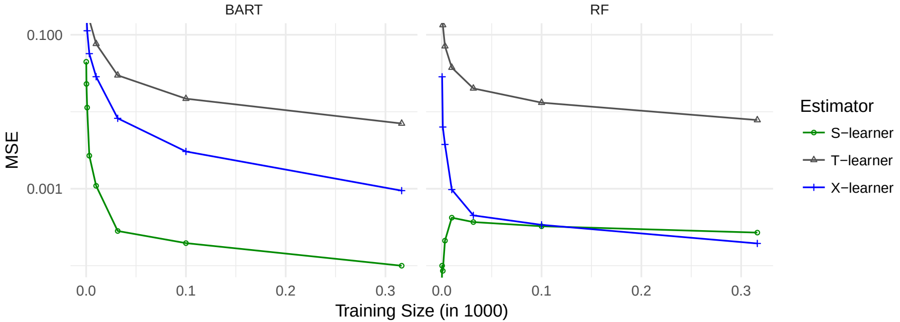
No Treatment Effect



Simple Setting

$$\mu_1(x) = x^T\beta, \text{ with } \beta \sim \text{Unif}([1,30]^d)$$
$$\mu_0(x) = \mu_1(x)$$
$$e(x) = 0.5$$

Resisting Confounding

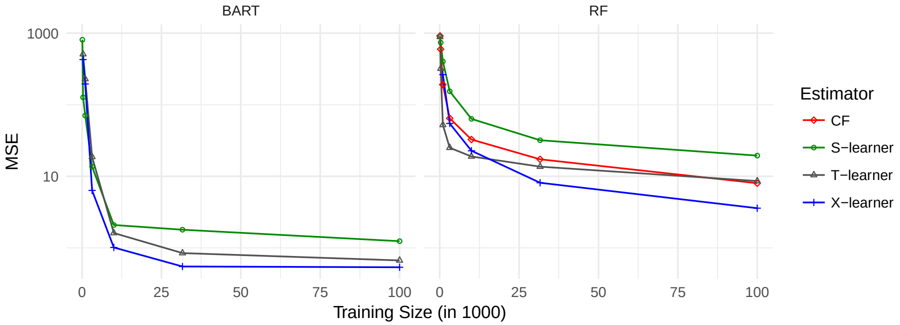


Confounded without TE (WA, 1)

$$\mu_1(x) = 2x_1 - 1,$$
$$\mu_0(x) = 2x_1 - 1,$$
$$e(x) = \frac{1}{4}(1 + \beta_{2,4}(x_1))$$

More Estimators

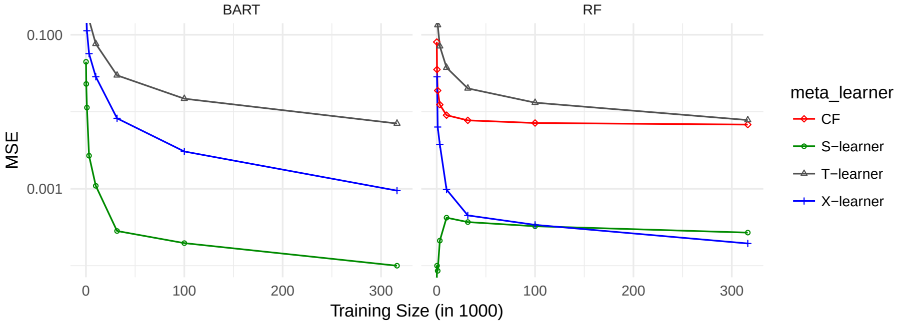
Flexibility of Base Learners is Needed



Complicated Setting

$$\mu_1(x) = x^T\beta_1, \text{ with } \beta_1 \sim \text{Unif}([1,30]^d)$$
$$\mu_0(x) = x^T\beta_0, \text{ with } \beta_0 \sim \text{Unif}([1,30]^d)$$
$$e(x) = .5$$

Resisting Confounding: different base learners, same effect

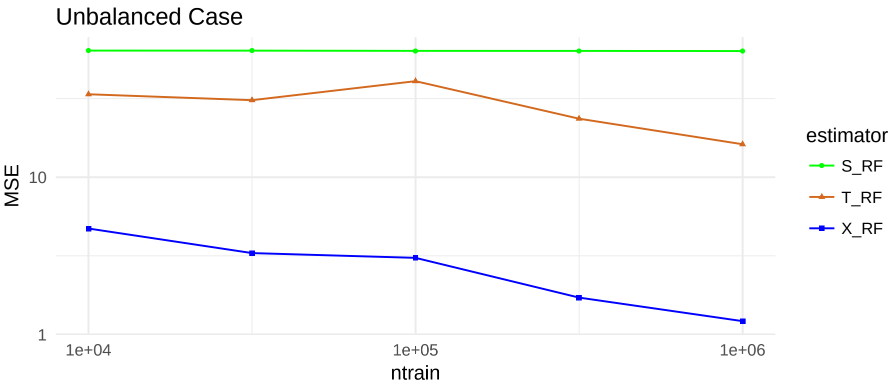


Confounded without TE (WA, 1)

$$\mu_1(x) = 2x_1 - 1,$$
$$\mu_0(x) = 2x_1 - 1,$$
$$e(x) = \frac{1}{4}(1 + \beta_{2,4}(x_1))$$

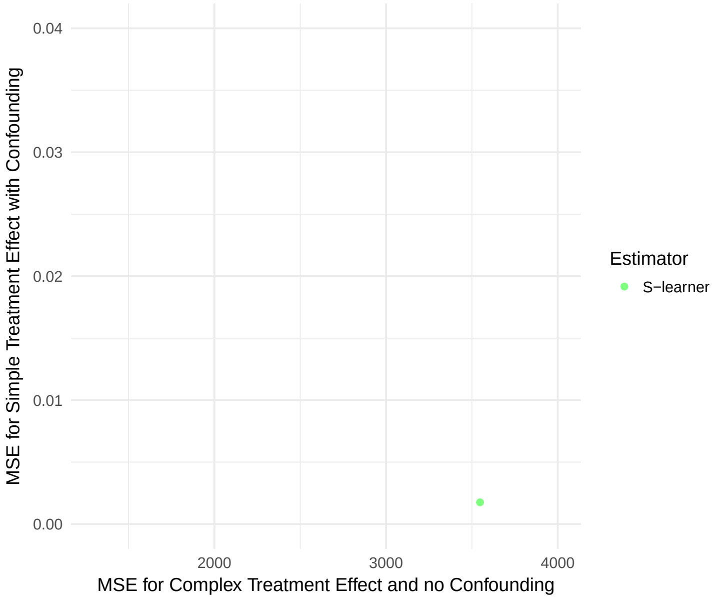
Back

The Unbalanced Case



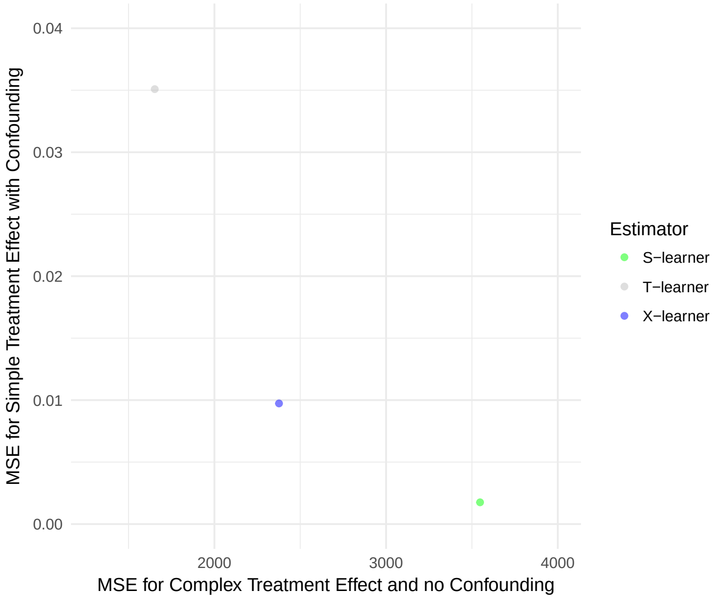
$$\mu_0(x) = x^T\beta + 5 * 1(x1 > .5), \quad \text{with} \quad \beta \sim \text{Unif}\left([1,5]^d\right)$$
$$\mu_1(x) = \mu_1(x) + 8$$
$$e(x) = 0.01$$

Adaptivity

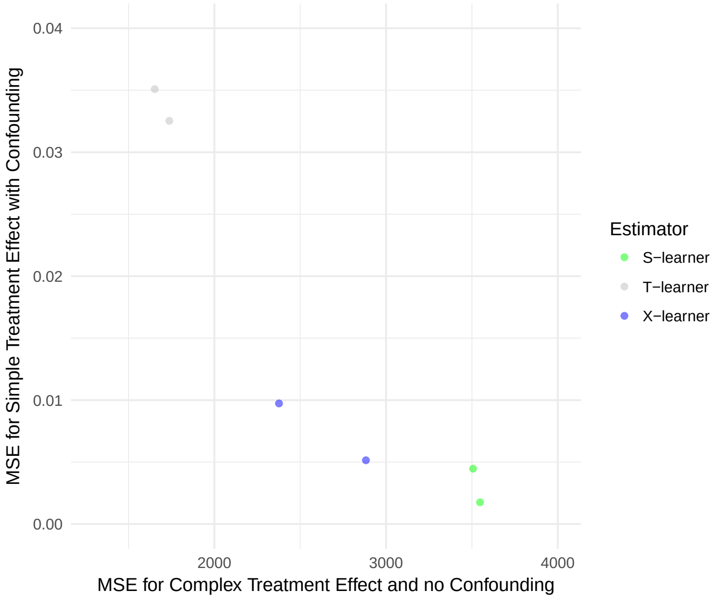


0.04
0.03
0.02
0.01
0.00
MSE for Simple Treatment Effect with Confounding
2000
3000
4000
MSE for Complex Treatment Effect and no Confounding
Estimator
S−learner

Adaptivity

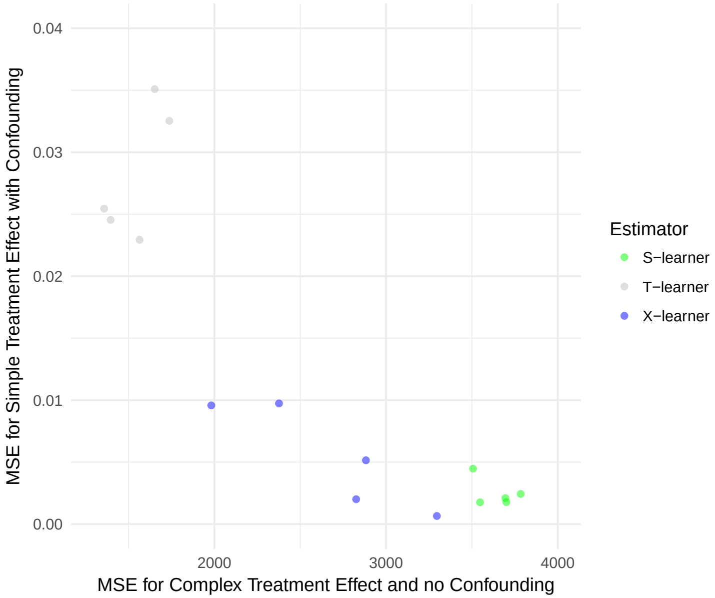

0.04
0.03
0.02
0.01
0.00
2000
3000
4000
MSE for Simple Treatment Effect with Confounding
MSE for Complex Treatment Effect and no Confounding
Estimator
S−learner
T−learner
X−learner

Adaptivity

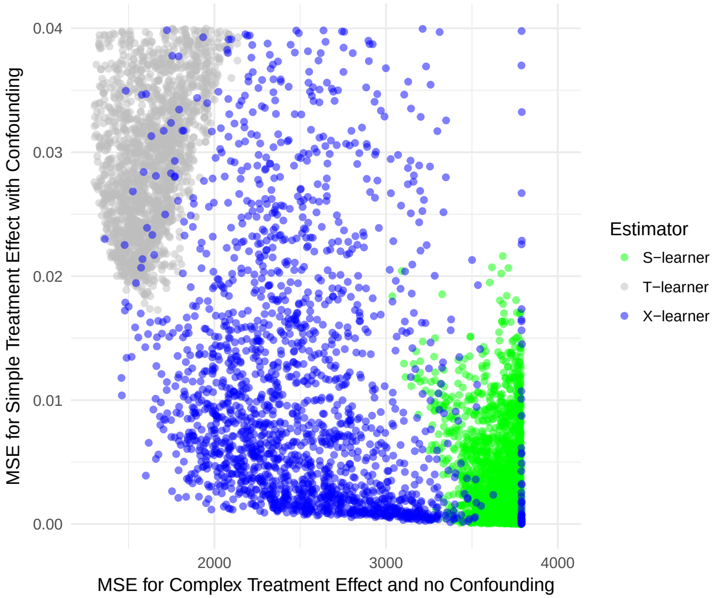


0.04
0.03
0.02
0.01
0.00
MSE for Simple Treatment Effect with Confounding
2000
3000
4000
MSE for Complex Treatment Effect and no Confounding
Estimator
S−learner
T−learner
X−learner

Adaptivity


0.04
0.03
0.02
0.01
0.00
MSE for Simple Treatment Effect with Confounding
2000
3000
4000
MSE for Complex Treatment Effect and no Confounding
Estimator
S−learner
T−learner
X−learner

Adaptivity



MSE for Simple Treatment Effect with Confounding
0.04
0.03
0.02
0.01
0.00
2000
3000
4000
MSE for Complex Treatment Effect and no Confounding
Estimator
S−learner
T−learner
X−learner

Tuning

All meta–learners can be separated into several small regression problems, and we tune them separately using tuning methods which are specific for each of the learner

We have implemented a package combining the X–learner with honest Random Forests and it currently implements three tuning methods:

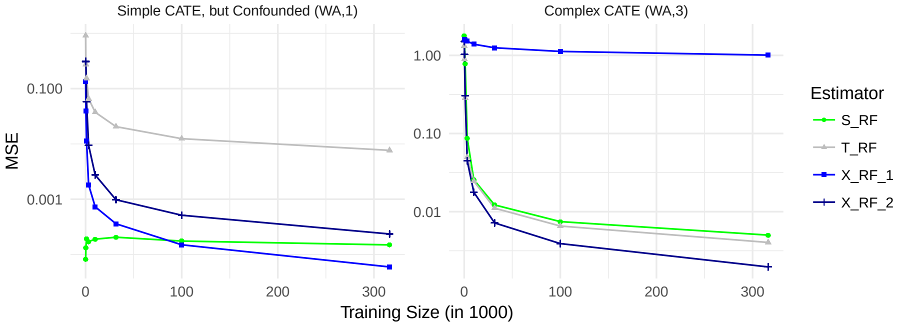
1.) Pre-specified tuning

2.) Gaussian Process

3.) Hyperband

Supplementary

Tuning



$$\mu_1(x) = 2x_1 - 1,$$
$$\mu_0(x) = 2x_1 - 1,$$
$$e(x) = \frac{1}{4}(1 + \beta_{2,4}(X_1))$$

$$\mu_1(x) = \zeta(X_1)\zeta(X_2),$$
$$\mu_0(x) = -\zeta(X_1)\zeta(X_2),$$
$$e(x) = 0.5,$$
$$\zeta(x) = \frac{2}{1 + e^{-12(x-1/2)}}$$