Causal forests

A tutorial in high-dimensional causal inference

Ian Lundberg

General Exam
Frontiers of Causal Inference

12 October 2017

PC: Michael Schweppe via Wikimedia Commons

Introductory note for those finding these slides online

These slides were prepared for the Sociology Statistics Reading Group at Princeton. Everyone read the following paper in advance:

> Athey, Susan, and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360. [link]

The aim was to discuss this paper so that together the group would leave with a good understanding of its contribution. Along the way, these slides touch on other related literature that uses tree-based methods for causal inference. In addition, these slides counted as one of several requirements for a general exam I completed in causal inference.

Disclaimer that I invented nothing in these slides, and there are likely places where I misread the original papers.

Note: These slides assume randomized treatment assignment until the section labeled "confounding."

Causal inference: A missing data problem

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	1	1
2	High school	1	0	1	1
3	College	0	1	1	0
4	College	1	1	1	0

Causal inference: A missing data problem

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	1	1
2	High school	1	0	1	1
3	College	0	1	1	0
4	College	1	1	1	0

$$
\begin{aligned}
\bar{\tau} &= \bar{Y}_{i:W_i=1}(1) - \bar{Y}_{i:W_i=0}(0) \\
&= 1 - 0.5 \\
&= 0.5
\end{aligned}
$$

Causal inference: A missing data problem

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	1	?	1	?
3	College	0	1	?	?
4	College	1	?	1	?

Causal inference: A missing data problem

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	1	?	1	?
3	College	0	1	?	?
4	College	1	?	1	?

If $W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}$, then

$$\begin{aligned}\hat{\hat{\tau}} &= \bar{Y}_{i:W_i=1} - \bar{Y}_{i:W_i=0} \\ &= 1 - 0.5 \\ &= 0.5\end{aligned}$$

Causal inference: A missing data problem

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	1	?	1	?
3	College	0	1	?	?
4	College	1	?	1	?

What if we want to study $\tau_i = f(X_i)$?

$$\begin{aligned}\hat{\hat{\tau}}_{\text{High school}} &= \bar{Y}_{i:W_i=1,X_i=\text{High school}} \\ &\quad - \bar{Y}_{i:W_i=0,X_i=\text{High school}} \\ &= 1 - 0.5 \\ &= 0.5\end{aligned}$$

$$\begin{aligned}\hat{\hat{\tau}}_{\text{College}} &= \bar{Y}_{i:W_i=1,X_i=\text{College}} \\ &\quad - \bar{Y}_{i:W_i=0,X_i=\text{College}} \\ &= 1 - 1 \\ &= 0\end{aligned}$$

Causal inference: A missing data problem

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	1	?	1	?
3	College	0	1	?	?
4	College	1	?	1	?

What if there are dozens of X variables?

Causal inference: A missing data problem

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	1	?	1	?
3	College	0	1	?	?
4	College	1	?	1	?

What if there are dozens of X variables?
What if X is continuous?

Causal inference: A missing data problem

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	1	?	1	?
3	College	0	1	?	?
4	College	1	?	1	?

What if there are dozens of X variables?
What if X is continuous?

It's hard to know which subgroups of X
might show interesting effect heterogeneity

Start with a simpler prediction question.

Which subgroups of X have very different average outcomes?

Prediction: One tree

$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$

All observations

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_1)})^2$$

$$\begin{tikzcd}[column sep=small] & \text{All observations} \arrow[dl] \arrow[dr] & \\ X_i < k & & X_i \geq k \end{tikzcd}$$

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$

$$\begin{array}{ccc} & \text{All observations} & \\ \swarrow & & \searrow \\ X_i < k & & X_i \geq k \end{array}$$

Choose k to minimize MSE_1

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$

$$\begin{array}{ccc} & \text{All observations} & \\ \swarrow & & \searrow \\ X_i < 16 & & X_i \geq 16 \end{array}$$

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$

$$\text{MSE}_2 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_2)})^2$$

All observations

$X_i < 16$ $X_i \geq 16$

$X_i < k_1$ $X_i \geq k_1$ $X_i < k_2$ $X_i \geq k_2$

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$

$$\text{MSE}_2 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_2)})^2$$

$$\begin{tikzcd}[column sep=small]
 & & \text{All observations} \arrow[dl] \arrow[dr] & & \\
 & X_i < 16 \arrow[dl] \arrow[d] & & X_i \geq 16 \arrow[d] \arrow[dr] & \\
X_i < k_1 & X_i \geq k_1 & & X_i < k_2 & X_i \geq k_2
\end{tikzcd}$$

Choose k_1 or k_2 to minimize MSE_2

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$

$$\text{MSE}_2 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_2)})^2$$

$$\begin{tikzcd}
 & & \text{All observations} \arrow[dl] \arrow[dr] & \\
 & X_i < 16 \arrow[dl] \arrow[d] & & X_i \geq 16 \\
X_i < 12 & X_i \geq 12 & &
\end{tikzcd}$$

Prediction: One tree

$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$

$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$

$\text{MSE}_2 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_2)})^2$

MSE_3

$$
\begin{tikzcd}
 & & \text{All observations} \arrow[dl] \arrow[dr] & \\
 & X_i < 16 \arrow[dl] \arrow[d] & & X_i \geq 16 \\
X_i < 12 & X_i \geq 12 \arrow[dl] \arrow[dr] & & \\
Z_i = \text{White} & & Z_i \neq \text{White} &
\end{tikzcd}
$$

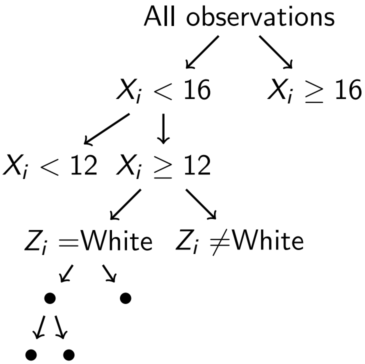
Prediction: One tree

$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$

$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_1)})^2$

$\text{MSE}_2 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_2)})^2$

MSE_3



Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_1)})^2$$

$$\text{MSE}_2 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_2)})^2$$

MSE_3

All observations

$X_i < 16$ $\quad X_i \geq 16$

$X_i < 12$ $\quad X_i \geq 12$

$Z_i =$White $\quad Z_i \neq$White

Could continue until all leaves had only one observation.

Unbiased but uselessly high variance!

Instead, regularize: keep only splits that improve MSE by more than c.

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$

$$\text{MSE}_2 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_2)})^2$$

$$\text{MSE}_3$$

All observations

- $X_i < 16$
  - $X_i < 12$
  - $X_i \geq 12$
    - $Z_i =$White
    - $Z_i \neq$White
- $X_i \geq 16$

Could continue until all leaves had only one observation.

Unbiased but uselessly high variance!

Instead, regularize: keep only splits that improve MSE by more than c.

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_1)})^2$$

$$\text{MSE}_2 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_2)})^2$$

All observations

$X_i < 16$ $\quad$ $X_i \geq 16$

$X_i < 12$ $\quad$ $X_i \geq 12$

Could continue until all leaves had only one observation.

Unbiased but uselessly high variance!

Instead, regularize: keep only splits that improve MSE by more than c.

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_1)})^2$$

All observations

$X_i < 16$ $\quad$ $X_i \geq 16$

Could continue until all leaves had only one observation.

Unbiased but uselessly high variance!

Instead, regularize: keep only splits that improve MSE by more than c.

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j\in\ell(x_i|\Pi_1)})^2$$

All observations

$X_i < 16$ $\quad$ $X_i \geq 16$

Partition $\Pi \in \mathbb{P} \longrightarrow \left\{\ell_1 = \{x_i : x_i < 16\},\ \ell_2 = \{x_i : x_i \geq 16\}\right\}$

Leaves

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$

$$\begin{array}{ccc} & \text{All observations} & \\ \swarrow & & \searrow \\ X_i < 16 & & X_i \geq 16 \end{array}$$

Partition $\Pi \in \mathbb{P} \longrightarrow \left\{ \ell_1 = \{x_i : x_i < 16\},\ \ell_2 = \{x_i : x_i \geq 16\} \right\}$

Leaves

Prediction rule for new x:

$$\boxed{\hat{\mu}(x) = \bar{Y}_{j:x_j \in \ell(x_i|\Pi)}}$$

Prediction: One tree

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$

All observations

$X_i < 16$ $\quad$ $X_i \geq 16$

$$\text{Partition } \Pi \in \mathbb{P} \longrightarrow \left\{ \ell_1 = \{x_i : x_i < 16\},\ \ell_2 = \{x_i : x_i \geq 16\} \right\}$$

Leaves

Prediction rule for new x:

$$\boxed{\hat{\mu}(x) = \bar{Y}_{j:x_j \in \ell(x_i|\Pi)}}$$

Could we use this method to find causal effects $\hat{\tau}(x)$ that are heterogeneous between leaves?

Causal tree: What's different?

1. We do not observe the ground truth

Causal tree: What's different?

1. We do not observe the ground truth
2. Honest estimation:
    - One sample to choose partition
    - One sample to estimate leaf effects

Causal tree: What's different?

1. We do not observe the ground truth
2. Honest estimation:
   - One sample to choose partition
   - One sample to estimate leaf effects

Why is the split critical?

Fitting both on the training sample risks overfitting: Estimating many "heterogeneous effects" that are really just noise idiosyncratic to the sample.

Causal tree: What's different?

1. We do not observe the ground truth
2. Honest estimation:
    - One sample to choose partition
    - One sample to estimate leaf effects

Why is the split critical?

Fitting both on the training sample risks overfitting: Estimating many "heterogeneous effects" that are really just noise idiosyncratic to the sample.

We want to search for true heterogeneity, not noise.

Sample splitting

$$\text{MSE}_{\mu}(S^{\text{te}}, S^{\text{est}}, \Pi) \equiv \frac{1}{\#(S^{\text{te}})} \sum_{i \in S^{\text{te}}} \left\{ \overbrace{(Y_i - \hat{\mu}(X_i; S^{\text{est}}, \Pi))^2}^{\text{MSE criterion}} - \overbrace{Y_i^2}^{\text{Authors add}} \right\}$$

Note: The authors include the final Y_i^2 term to simplify the math; it just shifts the estimator by a constant.

Sample splitting

$$\mathrm{MSE}_\mu(S^{\mathrm{te}}, S^{\mathrm{est}}, \Pi) \equiv \frac{1}{\#(S^{\mathrm{te}})} \sum_{i \in S^{\mathrm{te}}} \left\{ \overbrace{(Y_i - \hat{\mu}(X_i; S^{\mathrm{est}}, \Pi))^2}^{\text{MSE criterion}} - \overbrace{Y_i^2}^{\text{Authors add}} \right\}$$

$$\mathrm{EMSE}_\mu(\Pi) \equiv \mathbb{E}_{S^{\mathrm{te}}, S^{\mathrm{est}}} \left[\mathrm{MSE}_\mu(S^{\mathrm{te}}, S^{\mathrm{est}}, \Pi) \right]$$

Note: The authors include the final Y_i^2 term to simplify the math; it just shifts the estimator by a constant.

Sample splitting

$$\mathrm{MSE}_\mu(S^{\mathrm{te}}, S^{\mathrm{est}}, \Pi) \equiv \frac{1}{\#(S^{\mathrm{te}})} \sum_{i \in S^{\mathrm{te}}} \left\{ \overbrace{(Y_i - \hat{\mu}(X_i; S^{\mathrm{est}}, \Pi))^2}^{\text{MSE criterion}} - \overbrace{Y_i^2}^{\text{Authors add}} \right\}$$

$$\mathrm{EMSE}_\mu(\Pi) \equiv \mathbb{E}_{S^{\mathrm{te}}, S^{\mathrm{est}}} \left[\mathrm{MSE}_\mu(S^{\mathrm{te}}, S^{\mathrm{est}}, \Pi) \right]$$

Honest criterion: Maximize

This is S^{tr} in the classical approach

$$Q^H(\pi) \equiv -\mathbb{E}_{S^{\mathrm{te}}, S^{\mathrm{est}}, S^{\mathrm{tr}}} \left[\mathrm{MSE}_\mu(S^{\mathrm{te}}, S^{\mathrm{est}}, \pi(S^{\mathrm{tr}})) \right]$$

where $\pi : \mathbb{R}^{p+1} \to \mathbb{P}$ is a function that takes a training sample $S^{\mathrm{tr}} \in \mathbb{R}^{p+1}$ and outputs a partition $\Pi \in \mathbb{P}$.

Note: The authors include the final Y_i^2 term to simplify the math; it just shifts the estimator by a constant.

Analytic estimator for $\text{EMSE}_\mu(\Pi)$ (p. 7356)

Goal: Estimate expected MSE using only the training sample.

This will be used to place splits when training a tree.

Analytic estimator for $\text{EMSE}_\mu(\Pi)$ (p. 7356)

$$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi) + \mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi)\right)^2 - Y_i^2\right]$$

$$- \mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2\right]$$

$$- \mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[2\left(Y_i - \mu(X_i \mid \Pi)\right)\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)\right]$$

Analytic estimator for $\text{EMSE}_\mu(\Pi)$ (p. 7356)

Expected mean squared error for a partition Π

$$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi) + \mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi)\right)^2 - Y_i^2\right]$$

$$- \mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2\right]$$

$$- \mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[2\left(Y_i - \mu(X_i \mid \Pi)\right)\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)\right]$$

Analytic estimator for $\text{EMSE}_\mu(\Pi)$ (p. 7356)

Expected mean squared error for a partition Π

$$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2 - Y_i^2\right]$$

Over estimation sets used to estimate the leaf-specific $\hat{\mu}$ and test sets to evaluate those

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi) + \mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi)\right)^2 - Y_i^2\right]$$

$$-\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2\right]$$

$$-\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[2\left(Y_i - \mu(X_i \mid \Pi)\right)\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)\right]$$

Analytic estimator for $\text{EMSE}_\mu(\Pi)$ (p. 7356)

Expected mean squared error for a partition Π

Prediction based on S^{est} from the leave $\ell(X_i)$ containing X_i

$$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2 - Y_i^2\right]$$

Over estimation sets used to estimate the leaf-specific $\hat{\mu}$ and test sets to evaluate those

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi) + \mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi)\right)^2 - Y_i^2\right]$$

$$-\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2\right]$$

$$-\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[2\left(Y_i - \mu(X_i \mid \Pi)\right)\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)\right]$$

Analytic estimator for $\text{EMSE}_\mu(\Pi)$ (p. 7356)

$$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2 - Y_i^2\right]$$

Add a zero

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi) + \mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi)\right)^2 - Y_i^2\right]$$

$$- \mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)^2\right]$$

$$- \mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[2\left(Y_i - \mu(X_i \mid \Pi)\right)\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}},\Pi)\right)\right]$$

Analytic estimator for $\text{EMSE}_\mu(\Pi)$ (p. 7356)

$$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi) + \mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2 - Y_i^2\right]$$

First term^2

Second term^2

$$= -\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(Y_i - \mu(X_i \mid \Pi)\right)^2 - Y_i^2\right]$$

$$-\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)^2\right]$$

2(First term)(Second term)

$$-\mathbb{E}_{S^{\text{te}},S^{\text{est}}}\left[2\left(Y_i - \mu(X_i \mid \Pi)\right)\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right)\right]$$

Analytic estimator for $\text{EMSE}_\mu(\Pi)$ (p. 7356)

$$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{S^{te},S^{est}}\left[\left(Y_i - \hat{\mu}(X_i \mid S^{est},\Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{te},S^{est}}\left[\left(Y_i - \mu(X_i \mid \Pi) + \mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{est},\Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{te},S^{est}}\left[\left(Y_i - \mu(X_i \mid \Pi)\right)^2 - Y_i^2\right]$$

$E(A) = 0$ by assumption

$$-\mathbb{E}_{S^{te},S^{est}}\left[\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{est},\Pi)\right)^2\right]$$

$Cov(A, B) = 0$ because Y_i is from a sample independent of S^{est}

$$Cov(AB) = E(AB) - E(A)E(B)$$
$$0 = E(AB) - 0$$

$$-\mathbb{E}_{S^{te},S^{est}}\left[2\underbrace{\left(Y_i - \mu(X_i \mid \Pi)\right)}_{A}\underbrace{\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{est},\Pi)\right)}_{B}\right]$$

Analytic estimator for $\mathrm{EMSE}_\mu(\Pi)$ (p. 7356)

$$-\mathrm{EMSE}_\mu(\Pi) = -\mathbb{E}_{S^{te},S^{est}}\left[\left(Y_i - \hat{\mu}(X_i \mid S^{est}, \Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{te},S^{est}}\left[\left(Y_i - \mu(X_i \mid \Pi) + \mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{est}, \Pi)\right)^2 - Y_i^2\right]$$

$$= -\mathbb{E}_{S^{te},S^{est}}\left[\left(Y_i - \mu(X_i \mid \Pi)\right)^2 - Y_i^2\right]$$

$$-\mathbb{E}_{S^{te},S^{est}}\left[\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{est}, \Pi)\right)^2\right]$$

$$-\mathbb{E}_{S^{te},S^{est}}\left[2\overbrace{\left(Y_i - \mu(X_i \mid \Pi)\right)}^{A}\overbrace{\left(\mu(X_i \mid \Pi) - \hat{\mu}(X_i \mid S^{est}, \Pi)\right)}^{B}\right] = 0$$

$E(A) = 0$ by assumption

$Cov(A, B) = 0$ because Y_i is from a sample independent of S^{est}

$$Cov(AB) = E(AB) - E(A)E(B)$$
$$0 = E(AB) - 0$$

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[(Y_i - \mu(X_i \mid \Pi))^2 - Y_i^2\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[Y_i^2 + \mu^2(X_i \mid \Pi) - 2Y_i\mu(X_i \mid \Pi) - Y_i^2\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[\mu^2(X_i \mid \Pi) - 2\mu(X_i \mid \Pi)\mu(X_i \mid \Pi)\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}},X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

Athey & Imbens 2016, p. 7356

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[(Y_i - \mu(X_i \mid \Pi))^2 - Y_i^2\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

Y_i^2 terms cancel

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[Y_i^2 + \mu^2(X_i \mid \Pi) - 2Y_i\mu(X_i \mid \Pi) - Y_i^2\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[\mu^2(X_i \mid \Pi) - 2\mu(X_i \mid \Pi)\mu(X_i \mid \Pi)\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}},X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

Athey & Imbens 2016, p. 7356

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[(Y_i - \mu(X_i \mid \Pi))^2 - Y_i^2\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}(Y_i) = \mathbb{E}_{X_i,S^{\text{est}}}\mu(X_i \mid \Pi)$$

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[Y_i^2 + \mu^2(X_i \mid \Pi) - 2Y_i\mu(X_i \mid \Pi) - Y_i^2\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[\mu^2(X_i \mid \Pi) - 2\mu(X_i \mid \Pi)\mu(X_i \mid \Pi)\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}},X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

Athey & Imbens 2016, p. 7356

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[(Y_i - \mu(X_i \mid \Pi))^2 - Y_i^2\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[Y_i^2 + \mu^2(X_i \mid \Pi) - 2Y_i\mu(X_i \mid \Pi) - Y_i^2\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

$$= -\mathbb{E}_{(Y_i,X_i),S^{\text{est}}}\left[\mu^2(X_i \mid \Pi) - 2\mu(X_i \mid \Pi)\mu(X_i \mid \Pi)\right]$$
$$- \mathbb{E}_{X_i,S^{\text{est}}}\left[(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi) - \mu(X_i \mid \Pi))^2\right]$$

They have $\hat{\mu}^2$ here but I think they are wrong

I think

$$= \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}},X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

Athey & Imbens 2016, p. 7356

$$-\mathsf{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\mathrm{est}},X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\mathrm{est}}, \Pi))\right]$$

$$\text{Estimate with } \hat{\mathbb{V}}\left(\hat{\mu}(x \mid S^{\text{est}}, \Pi)\right) \equiv \frac{S^2_{S^{\text{tr}}}(\ell(x|\Pi))}{N^{\text{est}}(\ell(x|\Pi))}$$

$$-\text{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}},X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

$$\text{Estimate with } \hat{\mathbb{V}}\left(\hat{\mu}(x \mid S^{\text{est}}, \Pi)\right) \equiv \frac{S^2_{S^{\text{tr}}}(\ell(x|\Pi))}{N^{\text{est}}(\ell(x|\Pi))}$$

$$\hat{\mathbb{E}}_{X_i}\left[\hat{\mathbb{V}}_{S^{\text{est}}}\left(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right) \mid i \in S^{\text{te}}\right] = \sum_{\ell} p_\ell \frac{S^2_{S^{\text{tr}}}(\ell)}{N^{\text{est}}(\ell)}$$

$$-\text{EMSE}_\mu(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

$$\text{Estimate with } \hat{\mathbb{V}}\left(\hat{\mu}(x \mid S^{\text{est}}, \Pi)\right) \equiv \frac{S^2_{S^{\text{tr}}}(\ell(x|\Pi))}{N^{\text{est}}(\ell(x|\Pi))}$$

$$\hat{\mathbb{E}}_{X_i}\left[\hat{\mathbb{V}}_{S^{\text{est}}}\left(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right) \mid i \in S^{\text{te}}\right] = \sum_{\ell} p_\ell \frac{S^2_{S^{\text{tr}}}(\ell)}{N^{\text{est}}(\ell)}$$

$$\text{(assuming } \approx \text{ equal leaf sizes)} \approx \sum_{\ell} \frac{1}{\#\ell} \frac{S^2_{S^{\text{tr}}}(\ell)}{N^{\text{est}}/\#\ell}$$

$$-\text{EMSE}_\mu(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

$$\text{Estimate with } \hat{\mathbb{V}}\left(\hat{\mu}(x \mid S^{\text{est}}, \Pi)\right) \equiv \frac{S^2_{S^{\text{tr}}}(\ell(x|\Pi))}{N^{\text{est}}(\ell(x|\Pi))}$$

$$\begin{aligned}
\hat{\mathbb{E}}_{X_i}\left[\hat{\mathbb{V}}_{S^{\text{est}}}\left(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi)\right) \mid i \in S^{\text{te}}\right] &= \sum_{\ell} p_\ell \frac{S^2_{S^{\text{tr}}}(\ell)}{N^{\text{est}}(\ell)} \\
\text{(assuming } \approx \text{ equal leaf sizes)} &\approx \sum_{\ell} \frac{1}{\#\ell} \frac{S^2_{S^{\text{tr}}}(\ell)}{N^{\text{est}}/\#\ell} \\
&= \frac{1}{N^{\text{est}}} \sum_{\ell \in \Pi} S^2_{S^{\text{tr}}}(\ell)
\end{aligned}$$

$$-\text{EMSE}_\mu(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

$$-\mathrm{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\mathrm{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\mathrm{est}}, \Pi))\right]$$

$$\mathbb{V}(\hat{\mu} \mid x, \Pi) = \mathbb{E}(\hat{\mu}^2 \mid x, \Pi) - \left[\mathbb{E}(\hat{\mu} \mid x, \Pi)\right]^2$$

$$-\mathrm{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\mathrm{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\mathrm{est}}, \Pi))\right]$$

$$\mathbb{V}(\hat{\mu} \mid x, \Pi) = \mathbb{E}(\hat{\mu}^2 \mid x, \Pi) - \left[\mathbb{E}(\hat{\mu} \mid x, \Pi)\right]^2$$

$$\frac{S^2_{S^{\text{tr}}}(\ell(x \mid \Pi))}{N^{\text{tr}}(\ell(x \mid \Pi))} \approx \hat{\mu}^2(x \mid S^{\text{tr}}\Pi) - \mu^2(x \mid \Pi)$$

$$-\text{EMSE}_\mu(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

$$\mathbb{V}(\hat{\mu} \mid x, \Pi) = \mathbb{E}(\hat{\mu}^2 \mid x, \Pi) - \left[\mathbb{E}(\hat{\mu} \mid x, \Pi)\right]^2$$

$$\frac{S^2_{S^{\text{tr}}}(\ell(x \mid \Pi))}{N^{\text{tr}}(\ell(x \mid \Pi))} \approx \hat{\mu}^2(x \mid S^{\text{tr}}\Pi) - \mu^2(x \mid \Pi)$$

$$\mu^2(x \mid \Pi) \approx \hat{\mu}^2(x \mid S^{\text{tr}}, \Pi) - \frac{S^2_{S^{\text{tr}}}(\ell(x \mid \Pi))}{N^{\text{tr}}(\ell(x \mid \Pi))}$$

$$-\text{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

$$\mathbb{V}(\hat{\mu} \mid x, \Pi) = \mathbb{E}(\hat{\mu}^2 \mid x, \Pi) - \left[\mathbb{E}(\hat{\mu} \mid x, \Pi)\right]^2$$

$$\frac{S^2_{S^{\text{tr}}}(\ell(x \mid \Pi))}{N^{\text{tr}}(\ell(x \mid \Pi))} \approx \hat{\mu}^2(x \mid S^{\text{tr}}\Pi) - \mu^2(x \mid \Pi)$$

$$\mu^2(x \mid \Pi) \approx \hat{\mu}^2(x \mid S^{\text{tr}}, \Pi) - \frac{S^2_{S^{\text{tr}}}(\ell(x \mid \Pi))}{N^{\text{tr}}(\ell(x \mid \Pi))}$$

$$\hat{\mathbb{E}}_{X_i}(\mu^2(X_i \mid \Pi)) \approx \frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\mu}^2(x_i \mid S^{\text{tr}}, \Pi) - \sum_{\ell} \frac{1}{\#\ell} \frac{S^2_{S^{\text{tr}}}(\ell)}{N^{\text{tr}}/\#\ell}$$

$$-\text{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

$$\mathbb{V}(\hat{\mu} \mid x, \Pi) = \mathbb{E}(\hat{\mu}^2 \mid x, \Pi) - \left[\mathbb{E}(\hat{\mu} \mid x, \Pi)\right]^2$$

$$\frac{S^2_{S^{\text{tr}}}(\ell(x \mid \Pi))}{N^{\text{tr}}(\ell(x \mid \Pi))} \approx \hat{\mu}^2(x \mid S^{\text{tr}}\Pi) - \mu^2(x \mid \Pi)$$

$$\mu^2(x \mid \Pi) \approx \hat{\mu}^2(x \mid S^{\text{tr}}, \Pi) - \frac{S^2_{S^{\text{tr}}}(\ell(x \mid \Pi))}{N^{\text{tr}}(\ell(x \mid \Pi))}$$

$$\hat{\mathbb{E}}_{X_i}(\mu^2(X_i \mid \Pi)) \approx \frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\mu}^2(x_i \mid S^{\text{tr}}, \Pi) - \sum_{\ell} \frac{1}{\#\ell} \frac{S^2_{S^{\text{tr}}}(\ell)}{N^{\text{tr}}/\#\ell}$$

$$= \frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\mu}^2(x_i \mid S^{\text{tr}}, \Pi) - \frac{1}{N^{\text{tr}}} \sum_{\ell} S^2_{S^{\text{tr}}}(\ell)$$

$$-\text{EMSE}_{\mu}(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi))\right]$$

$$-\widehat{\mathrm{EMSE}}_\mu(S^{\mathrm{tr}}, N^{\mathrm{est}}, \Pi) = \frac{1}{N^{\mathrm{tr}}} \sum_{i \in S^{\mathrm{tr}}} \hat{\mu}^2(X_i \mid S^{\mathrm{tr}}, \Pi) - \frac{1}{N^{\mathrm{tr}}} \sum_{\ell \in \Pi} S^2_{S^{\mathrm{tr}}}(\ell) - \frac{1}{N^{\mathrm{est}}} \sum_{\ell \in \Pi} S^2_{S^{\mathrm{tr}}}(\ell)$$

$$-\mathrm{EMSE}_\mu(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi)\right] - \mathbb{E}_{S^{\mathrm{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\mathrm{est}}, \Pi))\right]$$

$$
\begin{aligned}
-\widehat{\text{EMSE}}_\mu(S^{\text{tr}}, N^{\text{est}}, \Pi) &= \frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\mu}^2(X_i \mid S^{\text{tr}}, \Pi) - \frac{1}{N^{\text{tr}}} \sum_{\ell \in \Pi} S^2_{S^{\text{tr}}}(\ell) \\
&\quad - \frac{1}{N^{\text{est}}} \sum_{\ell \in \Pi} S^2_{S^{\text{tr}}}(\ell) \\
&= \underbrace{\frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\mu}^2(X_i \mid S^{\text{tr}}, \Pi)}_{\text{Conventional CART criterion}} - \underbrace{\left(\frac{1}{N^{\text{tr}}} + \frac{1}{N^{\text{est}}} \right) \sum_{\ell \in \Pi} S^2_{S^{\text{tr}}}(\ell)}_{\text{Uncertainty about leaf means}}
\end{aligned}
$$

$$
-\text{EMSE}_\mu(\Pi) = \mathbb{E}_{X_i}\left[\mu^2(X_i \mid \Pi) \right] - \mathbb{E}_{S^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\mu}(X_i \mid S^{\text{est}}, \Pi)) \right]
$$

Honest inference for treatment effects

Note: We still assume
randomized
treatment assignment

Honest inference for treatment effects

Population-average potential outcomes within leaves:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi)\Big]$$

Honest inference for treatment effects

Population-average potential outcomes within leaves:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi)\Big]$$

Potential outcome for treatment w (heterogeneous by X_i)

Averaged over controls X_i in the leaf

Honest inference for treatment effects

Population-average potential outcomes within leaves:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi)\Big]$$

Average causal effect:

$$\tau(x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(1) - Y_i(0) \mid X_i \in \ell(x \mid \Pi)\Big] = \mu(1, x \mid \Pi) - \mu(0, x \mid \Pi)$$

Honest inference for treatment effects

Population-average potential outcomes within leaves:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi)\Big]$$

Average causal effect:

$$\tau(x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(1) - Y_i(0) \mid X_i \in \ell(x \mid \Pi)\Big] = \mu(1, x \mid \Pi) - \mu(0, x \mid \Pi)$$

Average effect evaluated at (potentially moderating) covariate value x

Honest inference for treatment effects

Population-average potential outcomes within leaves:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi)\Big]$$

Average causal effect:

$$\tau(x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(1) - Y_i(0) \mid X_i \in \ell(x \mid \Pi)\Big] = \mu(1, x \mid \Pi) - \mu(0, x \mid \Pi)$$

Difference in potential outcomes

Honest inference for treatment effects

Population-average potential outcomes within leaves:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi)\Big]$$

Average causal effect:

$$\tau(x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(1) - Y_i(0) \mid X_i \in \ell(x \mid \Pi)\Big] = \mu(1, x \mid \Pi) - \mu(0, x \mid \Pi)$$

Among observations in the leaf ℓ

Honest inference for treatment effects

Population-average potential outcomes within leaves:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi)\Big]$$

Average causal effect:

$$\tau(x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(1) - Y_i(0) \mid X_i \in \ell(x \mid \Pi)\Big] = \mu(1, x \mid \Pi) - \mu(0, x \mid \Pi)$$

Compact notation

Estimate:

$$\hat{\mu}(w, x \mid S, \Pi) \equiv \frac{1}{\#(\{i \in S_w : X_i \in \ell(x|\Pi)\})} \sum_{i \in S_w : X_i \in \ell(x|\Pi)} Y_i^{\mathrm{obs}}$$

Estimate:

$$\hat{\mu}(w, x \mid S, \Pi) \equiv \frac{1}{\#(\{i \in S_w : X_i \in \ell(x|\Pi)\})} \sum_{i \in S_w : X_i \in \ell(x|\Pi)} Y_i^{\text{obs}}$$

MSE for treatment effects:

$$\text{MSE}_\tau(S^{\text{te}}, S^{\text{est}}, \Pi) \equiv \frac{1}{\#(S^{\text{te}})} \sum_{i \in S^{\text{te}}} \left\{ \left(\tau_i - \hat{\tau}(X_i \mid S^{\text{est}}, \Pi) \right)^2 - \tau_i^2 \right\}$$

Estimate:

$$\hat{\mu}(w, x \mid S, \Pi) \equiv \frac{1}{\#(\{i \in S_w : X_i \in \ell(x|\Pi)\})} \sum_{i \in S_w : X_i \in \ell(x|\Pi)} Y_i^{\text{obs}}$$

MSE for treatment effects:

$$\text{MSE}_\tau(S^{\text{te}}, S^{\text{est}}, \Pi) \equiv \frac{1}{\#(S^{\text{te}})} \sum_{i \in S^{\text{te}}} \left\{ \left(\tau_i - \hat{\tau}(X_i \mid S^{\text{est}}, \Pi) \right)^2 - \tau_i^2 \right\}$$

Challenge! τ_i is *never* observed.

Adapt EMSE_μ to estimate EMSE_τ

$$-\widehat{\text{EMSE}}_\mu(S^{\text{tr}}, N^{\text{est}}, \Pi) = \underbrace{\frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\mu}^2(X_i \mid S^{\text{tr}}, \Pi)}_{\text{Conventional CART criterion}} - \underbrace{\left(\frac{1}{N^{\text{tr}}} + \frac{1}{N^{\text{est}}}\right) \sum_{\ell \in \Pi} S^2_{S^{\text{tr}}}(\ell)}_{\text{Uncertainty about leaf means}}$$

$$-\widehat{\text{EMSE}}_\tau(S^{\text{tr}}, N^{\text{est}}, \Pi) = \underbrace{\frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\tau}^2(X_i \mid S^{\text{tr}}, \Pi)}_{\substack{\text{Variance of treatment} \\ \text{effects across leaves}}}$$

$$- \underbrace{\left(\frac{1}{N^{\text{tr}}} + \frac{1}{N^{\text{est}}}\right) \sum_{\ell \in \Pi} \left(\frac{S^2_{S^{\text{tr}}_{\text{treat}}}(\ell)}{p} + \frac{S^2_{S^{\text{tr}}_{\text{control}}}(\ell)}{1-p}\right)}_{\text{Uncertainty about leaf treatment effects}}$$

Adapt EMSE_μ to estimate EMSE_τ

$$-\widehat{\text{EMSE}}_\mu(S^{\text{tr}}, N^{\text{est}}, \Pi) = \underbrace{\frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\mu}^2(X_i \mid S^{\text{tr}}, \Pi)}_{\text{Conventional CART criterion}} - \underbrace{\left(\frac{1}{N^{\text{tr}}} + \frac{1}{N^{\text{est}}}\right) \sum_{\ell \in \Pi} S^2_{S^{\text{tr}}}(\ell)}_{\text{Uncertainty about leaf means}}$$

$$-\widehat{\text{EMSE}}_\tau(S^{\text{tr}}, N^{\text{est}}, \Pi) = \underbrace{\frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\tau}^2(X_i \mid S^{\text{tr}}, \Pi)}_{\text{Variance of treatment effects across leaves}}$$

Prefers leaves with heterogeneous effects

$$\underbrace{- \left(\frac{1}{N^{\text{tr}}} + \frac{1}{N^{\text{est}}}\right) \sum_{\ell \in \Pi} \left(\frac{S^2_{S^{\text{tr}}_{\text{treat}}}(\ell)}{p} + \frac{S^2_{S^{\text{tr}}_{\text{control}}}(\ell)}{1-p} \right)}_{\text{Uncertainty about leaf treatment effects}}$$

Prefers leaves with good fit (leaf-specific effects estimated precisely)

Four partitioning estimators

1. Causal trees

Split by

$$-\widehat{\text{EMSE}}_\tau(S^{\text{tr}}, N^{\text{est}}, \Pi) = \underbrace{\frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\tau}^2(X_i \mid S^{\text{tr}}, \Pi)}_{\substack{\text{Variance of treatment} \\ \text{effects across leaves}}}$$

$$\underbrace{- \left(\frac{1}{N^{\text{tr}}} + \frac{1}{N^{\text{est}}}\right) \sum_{\ell \in \Pi} \left(\frac{S^2_{S^{\text{tr}}_{\text{treat}}}(\ell)}{p} + \frac{S^2_{S^{\text{tr}}_{\text{control}}}(\ell)}{1-p} \right)}_{\text{Uncertainty about leaf treatment effects}}$$

Prefers leaves with heterogeneous effects

Prefers leaves with good fit (leaf-specific effects estimated precisely)

- Benefit: Prioritizes heterogeneity ($\hat{\tau}$ varies a lot) and fit (within-leaf precision)
- Drawback: Cannot be done with off-the-shelf CART methods

2. Transformed outcome trees

Transform the outcome

$$Y_i^* = Y_i \frac{W_i - p}{p(1-p)} \to \mathbb{E}(Y_i^* \mid X_i = x) = \tau(x)$$

$$\begin{aligned}
\mathbb{E}(Y_i^*) &= \mathbb{E}\left[Y_i \frac{W_i - p}{p(1-p)}\right] \\
&= \mathbb{E}\left[Y_i \frac{W_i}{p(1-p)}\right] - \mathbb{E}\left[Y_i \frac{p}{p(1-p)}\right] \\
&= \mathbb{E}\left[Y_i(1) \frac{W_i}{p(1-p)}\right] - \mathbb{E}\left[\Big(Y_i(1)W_i + Y_i(0)(1-W_i)\Big)\frac{p}{p(1-p)}\right] \\
&= Y_i(1)\frac{1}{p(1-p)}\mathbb{E}[W_i] - Y_i(1)\frac{p}{p(1-p)}\mathbb{E}[W_i] - Y_i(0)\frac{p}{p(1-p)}\mathbb{E}[1-W_i] \\
&= Y_i(1)\frac{1-p}{p(1-p)}\mathbb{E}[W_i] - Y_i(0)\frac{p}{p(1-p)}\mathbb{E}[1-W_i] \\
&= Y_i(1)\frac{p(1-p)}{p(1-p)} - Y_i(0)\frac{p(1-p)}{p(1-p)} \\
&= Y_i(1) - Y_i(0) = \tau_i
\end{aligned}$$

2. Transformed outcome trees

- Benefit: Can use off-the-shelf CART methods for prediction
- Drawbacks: Inefficient. Treatment is ignored after transforming outcome.
  If within a leaf $\bar{W} \neq p$ (by chance), then sample average within leaf is a poor estimator of $\hat{\tau}$.

3. Fit-based trees

Replace

$$\mathrm{MSE}_{\mu}(S^{\mathrm{te}}, S^{\mathrm{est}}, \Pi) \equiv \frac{1}{\#(S^{\mathrm{te}})} \sum_{i \in S^{\mathrm{te}}} \left\{ (Y_i - \hat{\mu}(X_i; S^{\mathrm{est}}, \Pi))^2 - Y_i^2 \right\}$$

with the fit-based split rule

$$\mathrm{MSE}_{\mu,W}(S^{\mathrm{te}}, S^{\mathrm{est}}, \Pi) \equiv \sum_{i \in S^{\mathrm{te}}} \left\{ (Y_i - \hat{\mu}_w(W_i X_i; S^{\mathrm{est}}, \Pi))^2 - Y_i^2 \right\}$$

which loss by model fit within each leaf: the difference from the expected value for the treatment group of observation i.

Benefit: Prefers splits that lead to better fit.

Drawback: Does not prefer splits that lead to variation in treatment effects.

Zeileis et al. 2008

4. Squared T-statistic trees

Split based on:

$\hat{\tau}$ in left leaf in right leaf

$$T^2 \equiv N \frac{(\bar{Y}_L - \bar{Y}_R)^2}{S^2/N_L + S^2/N_R}$$

Benefit: Prefers splits that lead to variation in treatment effects.

Drawback: Missed opportunity to improve fit: ignores useful splits between leaves with similar treatment effects but very different average values.

Su et al. 2009

From trees to forests: Double-sample trees

An individual tree can be noisy. Instead, we might fit a forest.

1. Draw a sample of size s
2. Split into an $\mathcal{I}$ and $\mathcal{J}$ sample.
3. Grow a tree on the $\mathcal{J}$ sample
4. Estimate leaf-specific $\hat{\tau}_\ell$ using the $\mathcal{I}$ sample

Repeat many times.

Advantages of forests:

- Consistent for true $\tau(x)$
- Asymptotic normality
- Asymptotic variance is estimable

Why *double-sample* forests:

- Advantage: Trees search for heterogeneous effects
- Disadvantage: Requires sample splitting

Wager & Athey 2017

From trees to forests: Propensity trees

An individual tree can be noisy. Instead, we might fit a forest.

1. Draw a sample of size s
2. Grow a tree on the $\mathcal{J}$ sample to predict W
   - Each leaf must have at least k observations of each treatment class
3. Estimate $\hat{\tau}_\ell$ on each leaf

Repeat many times.

Advantages of forests:

- Consistent for true $\tau(x)$
- Asymptotic normality
- Asymptotic variance is estimable

Why *propensity* forests:

- Advantage: Can use full sample
- Disadvantage: Does not search for heterogeneous effects

Wager & Athey 2017

Summary of causal trees and forests

- There is no ground truth: We never observe τ_i

Summary of causal trees and forests

- There is no ground truth: We never observe τ_i
- Causal trees search for leaves with
  - heterogeneous effects across leaves
  - precisely-estimated leaf effects

Summary of causal trees and forests

- There is no ground truth: We never observe τ_i
- Causal trees search for leaves with
  - heterogeneous effects across leaves
  - precisely-estimated leaf effects
- Require extra sample splitting

Summary of causal trees and forests

- There is no ground truth: We never observe τ_i
- Causal trees search for leaves with
  - heterogeneous effects across leaves
  - precisely-estimated leaf effects
- Require extra sample splitting
- Work well with randomized treatments.

Summary of causal trees and forests

- There is no ground truth: We never observe τ_i
- Causal trees search for leaves with
  - heterogeneous effects across leaves
  - precisely-estimated leaf effects
- Require extra sample splitting
- Work well with randomized treatments.
- With selection on observables, the general recommendation is propensity forests
  - Maximizes the goal of addressing confounding by ignoring heterogeneous effects when choosing splits

Summary of causal trees and forests

- There is no ground truth: We never observe τ_i
- Causal trees search for leaves with
  - heterogeneous effects across leaves
  - precisely-estimated leaf effects
- Require extra sample splitting
- Work well with randomized treatments.
- With selection on observables, the general recommendation is propensity forests
  - Maximizes the goal of addressing confounding by ignoring heterogeneous effects when choosing splits
  - Generalized random forests also perform well (Athey, Tibshirani, & Wager 2017)

Summary of causal trees and forests

- There is no ground truth: We never observe τ_i
- Causal trees search for leaves with
    - heterogeneous effects across leaves
    - precisely-estimated leaf effects
- Require extra sample splitting
- Work well with randomized treatments.
- With selection on observables, the general recommendation is propensity forests
    - Maximizes the goal of addressing confounding by ignoring heterogeneous effects when choosing splits
    - Generalized random forests also perform well (Athey, Tibshirani, & Wager 2017)
    - But "the challenge in using adaptive methods... is that selection bias can be difficult to quantify" (Wager & Athey p. 24).

If treatment is not randomized

Causal trees find heterogeneous effects but cannot guarantee that confounding is addressed.

Next we focus on why high-dimensional confounding is hard

Why aren't causal trees guaranteed to address confounding?

Plan

1. What does address confounding? Standardization
2. Why is tree-based standardization biased? Regularization
3. Is there anything we can do? Chernozhukov et al.

What works: Nonparametric standardization

What if $\{Y_i(0), Y_i(1)\} \not\perp\!\!\!\perp W_i$ but $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$?

What works: Nonparametric standardization

What if $\{Y_i(0), Y_i(1)\} \not\perp\!\!\!\perp W_i$ but $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$?

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	1	1
2	High school	0	0	1	1
3	High school	1	0	1	1
4	College	0	1	1	0
5	College	1	1	1	0
6	College	1	1	1	0

What works: Nonparametric standardization

What if $\{Y_i(0), Y_i(1)\} \not\perp\!\!\!\perp W_i$ but $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$?

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	0	0	?	?
3	High school	1	?	1	?
4	College	0	1	?	?
5	College	1	?	1	?
6	College	1	?	1	?

What works: Nonparametric standardization

What if $\{Y_i(0), Y_i(1)\} \not\perp\!\!\!\perp W_i$ but $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$?

We need to estimate $\hat{\tau}$ within each level of X_i.

			Potential employment		
ID	Education X_i	Treated W_i	No job training $Y_i(0)$	Job training $Y_i(1)$	Treatment effect $\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	0	0	?	?
3	High school	1	?	1	?
4	College	0	1	?	?
5	College	1	?	1	?
6	College	1	?	1	?

What works: Nonparametric standardization

$$\hat{\hat{\tau}} = \sum_{x \in \text{Support of } X} \mathbb{P}(X = x)\left(\bar{Y}_{i:W_i=1,X_i=x} - \bar{Y}_{i:W_i=0,X_i=x}\right)$$

$$= \mathbb{P}(X_i = \text{High school})\left(\bar{Y}_{i:W_i=1,X_i=\text{High school}} - \bar{Y}_{i:W_i=0,X_i=\text{High school}}\right)$$

$$+ \mathbb{P}(X_i = \text{College})\left(\bar{Y}_{i:W_i=1,X_i=\text{College}} - \bar{Y}_{i:W_i=0,X_i=\text{College}}\right)$$

$$= \frac{1}{2}(1-0) + \frac{1}{2}(1-1) = 0.5 + 0 =$$

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	0	0	?	?
3	High school	1	?	1	?
4	College	0	1	?	?
5	College	1	?	1	?
6	College	1	?	1	?

What works: Nonparametric standardization

$$\hat{\bar{\tau}} = \sum_{x \in \text{Support of } X} \mathbb{P}(X = x)\left(\bar{Y}_{i:W_i=1,X_i=x} - \bar{Y}_{i:W_i=0,X_i=x}\right)$$

$$= \mathbb{P}(X_i = \text{High school})\left(\bar{Y}_{i:W_i=1,X_i=\text{High school}} - \bar{Y}_{i:W_i=0,X_i=\text{High school}}\right)$$

$$+ \mathbb{P}(X_i = \text{College})\left(\bar{Y}_{i:W_i=1,X_i=\text{College}} - \bar{Y}_{i:W_i=0,X_i=\text{College}}\right)$$

$$= \frac{1}{2}(1-0) + \frac{1}{2}(1-1) = 0.5 + 0 =$$

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	0	0	?	?
3	High school	1	?	1	?
4	College	0	1	?	?
5	College	1	?	1	?
6	College	1	?	1	?

What works: Nonparametric standardization

$$\hat{\hat{\tau}} = \sum_{x \in \text{Support of } X} \mathbb{P}(X = x)\left(\bar{Y}_{i:W_i=1,X_i=x} - \bar{Y}_{i:W_i=0,X_i=x}\right)$$

$$= \mathbb{P}(X_i = \text{High school})\left(\bar{Y}_{i:W_i=1,X_i=\text{High school}} - \bar{Y}_{i:W_i=0,X_i=\text{High school}}\right)$$

$$+ \mathbb{P}(X_i = \text{College})\left(\bar{Y}_{i:W_i=1,X_i=\text{College}} - \bar{Y}_{i:W_i=0,X_i=\text{College}}\right)$$

$$= \frac{1}{2}(1-0) + \frac{1}{2}(1-1) = 0.5 + 0 =$$

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	0	0	?	?
3	High school	1	?	1	?
4	College	0	1	?	?
5	College	1	?	1	?
6	College	1	?	1	?

What works: Nonparametric standardization

$$
\begin{aligned}
\hat{\hat{\tau}} &= \sum_{x \in \text{Support of } X} \mathbb{P}(X = x)\left(\bar{Y}_{i:W_i=1,X_i=x} - \bar{Y}_{i:W_i=0,X_i=x}\right) \\
&= \mathbb{P}(X_i = \text{High school})\left(\bar{Y}_{i:W_i=1,X_i=\text{High school}} - \bar{Y}_{i:W_i=0,X_i=\text{High school}}\right) \\
&\quad + \mathbb{P}(X_i = \text{College})\left(\bar{Y}_{i:W_i=1,X_i=\text{College}} - \bar{Y}_{i:W_i=0,X_i=\text{College}}\right) \\
&= \frac{1}{2}(1-0) + \frac{1}{2}(1-1) = 0.5 + 0 = 0.5
\end{aligned}
$$

			Potential employment		
	Education	Treated	No job training	Job training	Treatment effect
ID	X_i	W_i	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
1	High school	0	0	?	?
2	High school	0	0	?	?
3	High school	1	?	1	?
4	College	0	1	?	?
5	College	1	?	1	?
6	College	1	?	1	?

What works: Nonparametric standardization

But when there are many cells of the covariates X_i,

nonparametric standardization is impossible!

Why is tree-based standardization biased? Regularization

With no regularization, a tree would grow until each leaf was completely homogenous in X_i.

But this tree would be very noisy! We prune our trees so that leaves contain more observations.

- Treatment effects are more precisely estimated
- But treatment effects are biased if there is confounding within leaves

Is there anything we can do? Chernozhukov et al.

$$\overbrace{Y = D\theta_0 + g_0(X) + U}^{\text{Outcome equation}} \qquad \overbrace{D = m_0(X) + V}^{\text{Treatment assignment}}$$

One might be tempted to estimate $\hat{g}_0(X)$ by machine learning and then state:

$$\hat{\theta}_0 = \frac{\frac{1}{n}\sum_{i\in\mathcal{I}} D_i(Y_i - \hat{g}_0(X_i))}{\frac{1}{n}\sum_{i\in\mathcal{I}} D_i^2}$$

Is there anything we can do? Chernozhukov et al.

$$\overbrace{Y = D\theta_0 + g_0(X) + U}^{\text{Outcome equation}} \qquad \overbrace{D = m_0(X) + V}^{\text{Treatment assignment}}$$

One might be tempted to estimate $\hat{g}_0(X)$ by machine learning and then state:

$$\hat{\theta}_0 = \frac{\frac{1}{n}\sum_{i\in\mathcal{I}} D_i(Y_i - \hat{g}_0(X_i))}{\frac{1}{n}\sum_{i\in\mathcal{I}} D_i^2}$$

This will be biased because the estimator $\hat{g}_0$ is regularized.

$$b = \frac{1}{\mathbb{E}(D_i^2)}\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}} \overbrace{\Big(m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i)\Big)}^{\text{Does not have mean 0}} + o_P(1)$$

Is there anything we can do? Chernozhukov et al.

$$\overbrace{Y = D\theta_0 + g_0(X) + U}^{\text{Outcome equation}} \qquad \overbrace{D = m_0(X) + V}^{\text{Treatment assignment}}$$

One might be tempted to estimate $\hat{g}_0(X)$ by machine learning and then state:

$$\hat{\theta}_0 = \frac{\frac{1}{n}\sum_{i\in\mathcal{I}} D_i(Y_i - \hat{g}_0(X_i))}{\frac{1}{n}\sum_{i\in\mathcal{I}} D_i^2}$$

This will be biased because the estimator $\hat{g}_0$ is regularized.

$$b = \frac{1}{\mathbb{E}(D_i^2)}\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}} \overbrace{\left(m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i)\right)}^{\text{Does not have mean 0}} + o_P(1)$$

Key: D_i is centered at $m_0(X) \neq 0$. We should recenter D_i.

Is there anything we can do? Chernozhukov et al.

$$\overbrace{Y = D\theta_0 + g_0(X) + U}^{\text{Outcome equation}} \qquad \overbrace{D = m_0(X) + V}^{\text{Treatment assignment}}$$

1. Split the sample into $\mathcal{I}$ and $\mathcal{J}$
2. Estimate $\hat{g}_0(X)$ using sample $\mathcal{J}$
3. Estimate $\hat{m}_0(X)$ using sample $\mathcal{J}$
4. Orthogonalize D on X (approximately)

$$\hat{V} = D - \hat{m}_0(X)$$

5. Estimate the treatment effect

Biased

$$\hat{\theta}_0 = \frac{\frac{1}{n}\sum_{i \in \mathcal{I}} D_i (Y_i - \hat{g}_0(X_i))}{\frac{1}{n}\sum_{i \in \mathcal{I}} D_i^2}$$

De-biased

$$\hat{\theta}_0 = \frac{\frac{1}{n}\sum_{i \in \mathcal{I}} \hat{V}_i (Y_i - \hat{g}_0(X_i))}{\frac{1}{n}\sum_{i \in \mathcal{I}} \hat{V}_i D_i}$$

Chernozhukov et al. 2016

Bias remaining in de-biased estimator (Chernozhukov et al.)

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

Bias remaining in de-biased estimator (Chernozhukov et al.)

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

$$a^* = \frac{1}{\mathbb{E}(V^2)} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}} V_i U_i \rightarrow N(0, \Sigma)$$

Because a^* converges to mean 0, we don't worry about it.

Bias remaining in de-biased estimator (Chernozhukov et al.)

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

Regularization bias:

$$b^* = \frac{1}{\mathbb{E}(V^2)} \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}} \Big(\hat{m}_0(X_i) - m_0(X_i)\Big)\Big(\hat{g}_0(X_i) - g_0(X_i)\Big)$$

Vanishes "under a broad range of data-generating processes."

Bounded above by

$$\sqrt{n}\, n^{-\psi_m}\, n^{-\psi_g}$$

Rate of convergence of $\hat{m}_0 \to m$

Rate of convergence of $\hat{g}_0 \to g$

Bias remaining in de-biased estimator (Chernozhukov et al.)

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

An example of the third term in the partially linear model:

$$c^* = \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}} V_i \Big(\hat{g}_0(X_i) - g_0(X_i) \Big)$$

If $\hat{g}_0$ is estimated on an auxiliary sample $\mathcal{J}$, then V_i and $\hat{g}_0(X_i)$ will be uncorrelated and $\mathbb{E}(c^*) = 0$.

BART: Bayesian Additive Regression Trees

Differs from random forests:

- Fixed number of trees
- Backfits repeatedly over the fixed number of trees
- Strong prior encourages shallow trees
- Uncertainty comes automatically from posterior samples

Chipman, George, & McCulloch 2010

BART model

$$Y = \sum_{j=1}^{m} g_j(x \mid T_j, M_j) + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

T_j prior

$$P(\underbrace{D_j = d}_{\text{Tree depth}}) = \alpha(1+d)^{-\beta}$$

$$\text{Split variable} \sim \text{Uniform(Available variables)}$$

$$\text{Split value} \sim \text{Uniform(Available split values)}$$

$\mu_{ij} \mid T_j$ prior

$$\underbrace{\mu_{ij}}_{\text{Tree } i \text{ leaf } j} \sim N\left(\underbrace{\mu_m, \sigma_\mu^2}_{\substack{\text{Chosen so that} \\ \text{high probability of} \\ E(Y|x) \in (y_{\min}, y_{\max})}} \right)$$

σ prior

$$\sigma \sim \frac{\nu\lambda}{\chi^2_\nu} \text{ (inverse chi-square)}$$

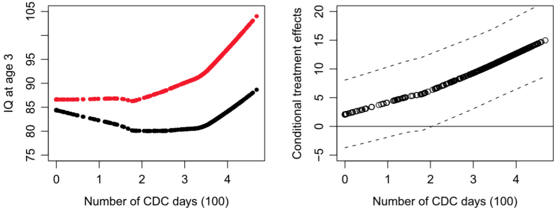
They recommend $\{\alpha = .95, \beta = 2\} \rightarrow$ 97% of prior probability is on 4 or fewer terminal nodes.

Chipman, George, & McCulloch 2010

BART for causal inference

Goal: Model the response surface as a function of treatment and pre-treatment covariates

1. Fit a flexible model for $Y = f(X, W)$
2. Set $W = 0$ to predict $\hat{Y}_i(0)$ for all i
3. Set $W = 1$ to predict $\hat{Y}_i(1)$ for all i
4. Difference to estimate $\hat{\tau}_i$
5. Plot effects



Hill 2011

BART: Benefits and drawbacks

Benefits

- Less researcher discretion for tuning parameters
- Automatic posterior uncertainty estimates

Drawbacks

- Not guaranteed to address confounding due to regularization
- No theoretical guarantees of centering over truth
- Splitting is based on prediction and is not explicitly optimized for causal inference within leaves

Summary

- Causal trees can detect high-dimensional covariate-based treatment effect heterogeneity
- Work well with high-order interactions
- Causal forests give theoretically valid confidence intervals
- Bayesian approaches (BART) are less theoretically verified but give easy uncertainty
- With high-dimensional confounding, all methods are biased but can be designed to be consistent.