Benchmarking Causal Estimators

Brady Neal

Context: many choices of causal estimators

Problem: no good way of choosing

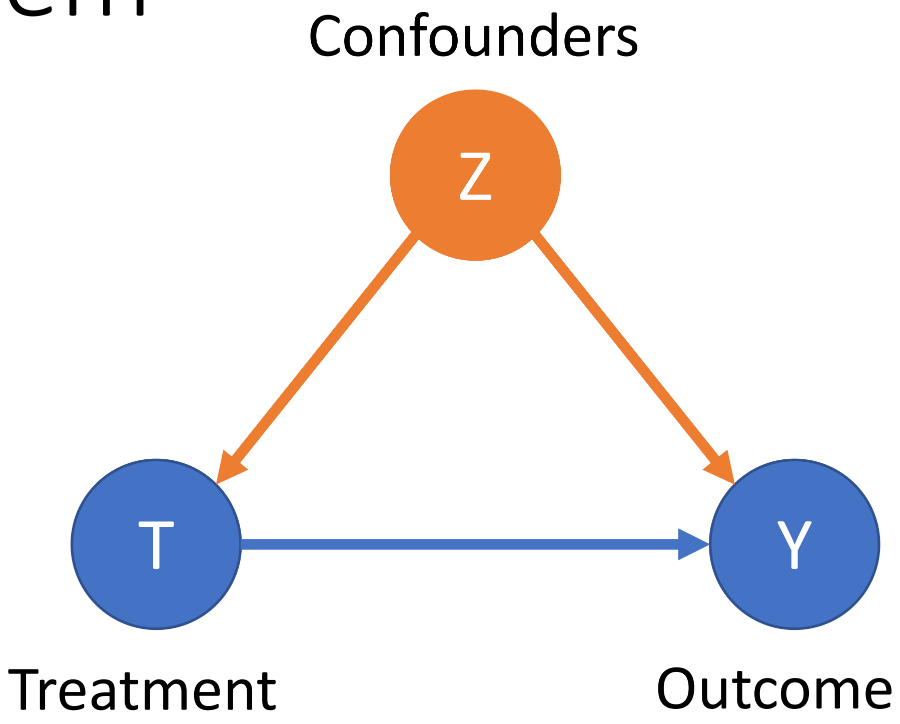
Solution: this project (ongoing)

Context: many choices of causal estimators

Problem: no good way of choosing

Solution: this project (ongoing)

Causal structure of basic observational causal inference problem



Conditional Ignorability: $\forall t \quad Y_t \perp\!\!\!\perp T \mid \underbrace{Z}_{\text{pre-treatment covariates}}$

Many causal estimators

Goal: Estimate average treatment effect (ATE) $\mathbb{E}[Y_1 - Y_0]$

Many different methods:

(non-exhaustive)

Adjustment formula: $\mathbb{E}[Y_t] = \sum_z \overbrace{\mathbb{E}[Y|T=t, Z=z]}^{\text{outcome model}} P(Z=z)$

(aka “g-formula”)

Inverse probability weighting: $\mathbb{E}[Y_t] = \mathbb{E}\left[\frac{I(T=t)Y}{\underbrace{P(T=t|Z=z)}_{\text{propensity score model}}}\right]$

$\underbrace{\text{Matching}}_{\text{distance metric (model), caliper, propensity score model}}$

Many models for each estimator

Linear models

$$\mathbb{E}[Y_t] = \sum_z \underline{\mathbb{E}[Y|T = t, Z = z]} P(Z = z)$$

Nonlinear models work better:

> "The most consistent conclusion was that methods that flexibly model the response surface perform better overall than methods that fail to do so." ~ Dorie et al. (2018)

Examples of flexible models: random forests, neural networks, etc.

Recap: many estimators and many models

Goal: Estimate average treatment effect (ATE) $\mathbb{E}[Y_1 - Y_0]$

Many different methods:

(non-exhaustive)

Adjustment formula: $$\mathbb{E}[Y_t] = \sum_z \overbrace{\mathbb{E}[Y|T=t, Z=z]}^{\text{outcome model}} P(Z=z)$$

(aka "g-formula")

Inverse probability weighting: $$\mathbb{E}[Y_t] = \mathbb{E}\left[\frac{I(T=t)Y}{\underbrace{P(T=t|Z=z)}}\right]$$

Matching

distance metric (model), caliper, propen

Linear models

$$\mathbb{E}[Y_t] = \sum_z \underline{\mathbb{E}[Y|T=t, Z=z]} P(Z=z)$$

Nonlinear models work better:

"The most consistent conclusion was that methods that flexibly model the response surface perform better overall than methods that fail to do so." ~ Dorie et al. (2018)

Examples of flexible models: random forests, neural networks, etc.

How do we choose?

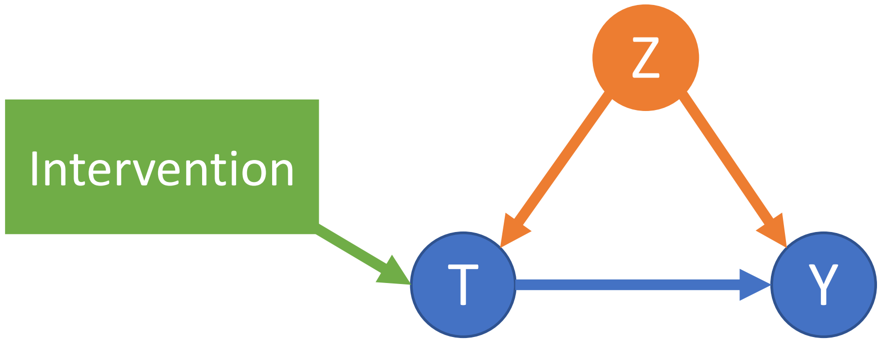
Context: many choices of causal estimators

Problem: no good way of choosing

Solution: this project (ongoing)

No ground-truth causal effects

There is no ground-truth for $\mathbb{E}[Y_1 - Y_0]$ because we do not have access to interventional (experimental) data



Two* choices:

1. Run simulations, where we intervene on T by changing the code of the simulation, allowing us to get Y_t
2. Don’t worry about ground-truth and just use real data (which may have unobserved confounding)

*and a couple other choices, which we don’t have time to get into

Benchmark desiderata recap

1. Realistic data
2. Access to ground-truth causal effects

Simulations have #2 but not #1.
Real data has #1 but not #2.

Solution: modern generative models fit to real data

Context: many choices of causal estimators

Problem: no good way of choosing

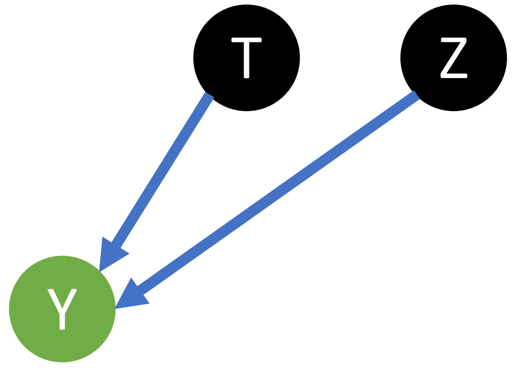
Solution: modern generative models

Modern generative models

Feed them samples from distributions such as $p(z,t,y)$, $p(t,y \mid z)$, or $p(y|z,t)$, and they learn to accurately model those distributions

Example: fit a model such as a probabilistic neural network to $p(y|z,t)$, and it will be able to generate samples of $\underline{Y_t}$ by taking t and z as inputs

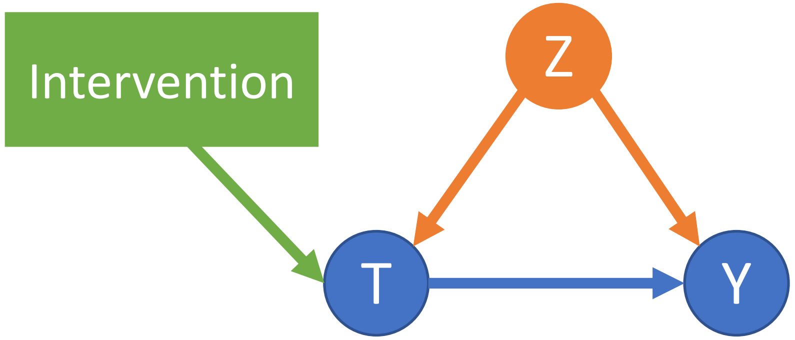
Because z are the only covariates, there is no unobserved confounding



How to get ground-truth

1. Get $p(z)$ from real data
2. Fit a generative model for $p(t, y|z)$ to *any* real data
3. Recover Y_t using the generative model just like one would use a simulation
4. Benchmark all the different estimators using the generatively modeled data and the corresponding ground-truth $\mathbb{E}[Y_1 - Y_0]$

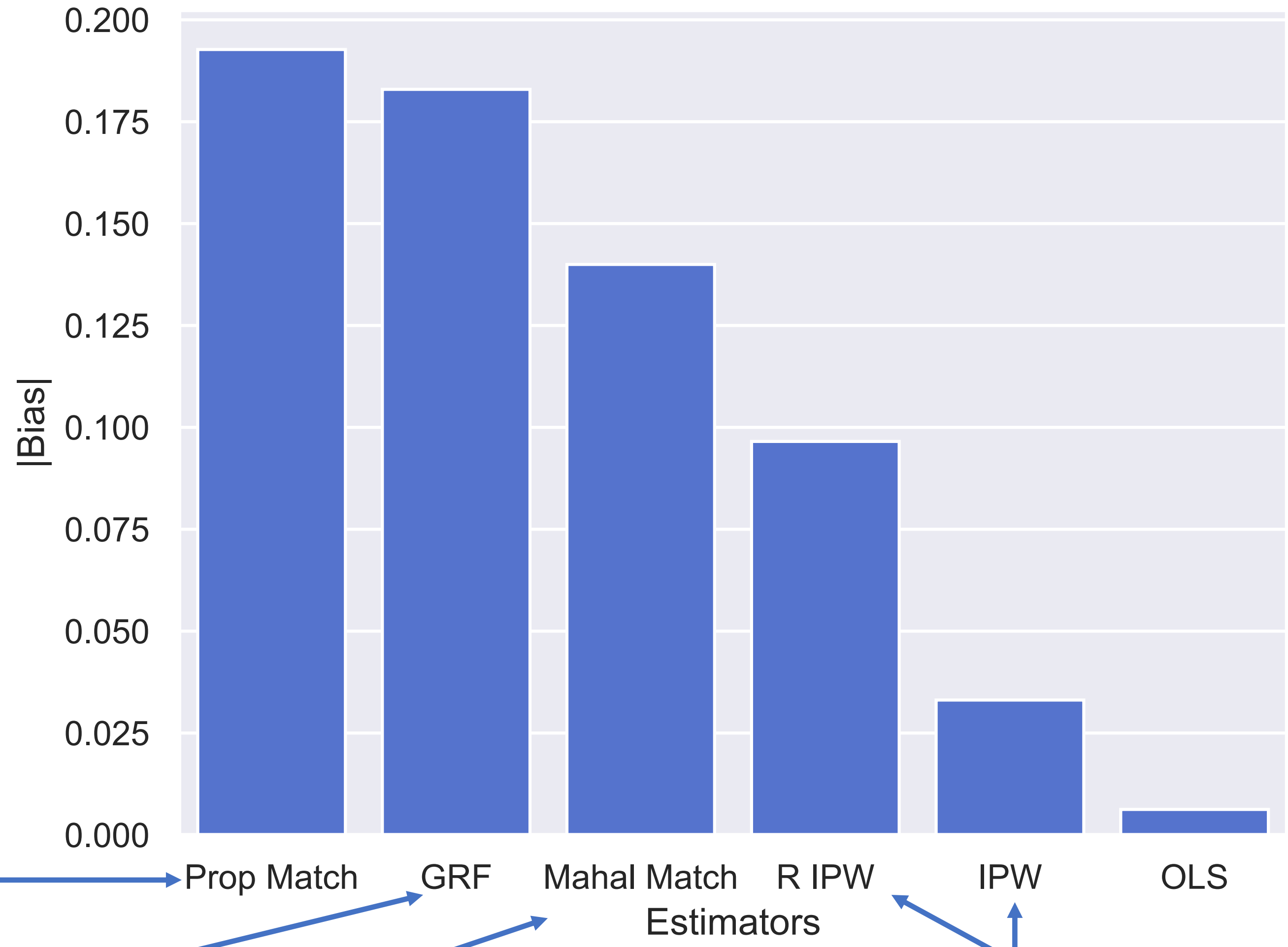
Causal Graph



Example benchmarking on **linear** data

True ATE:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

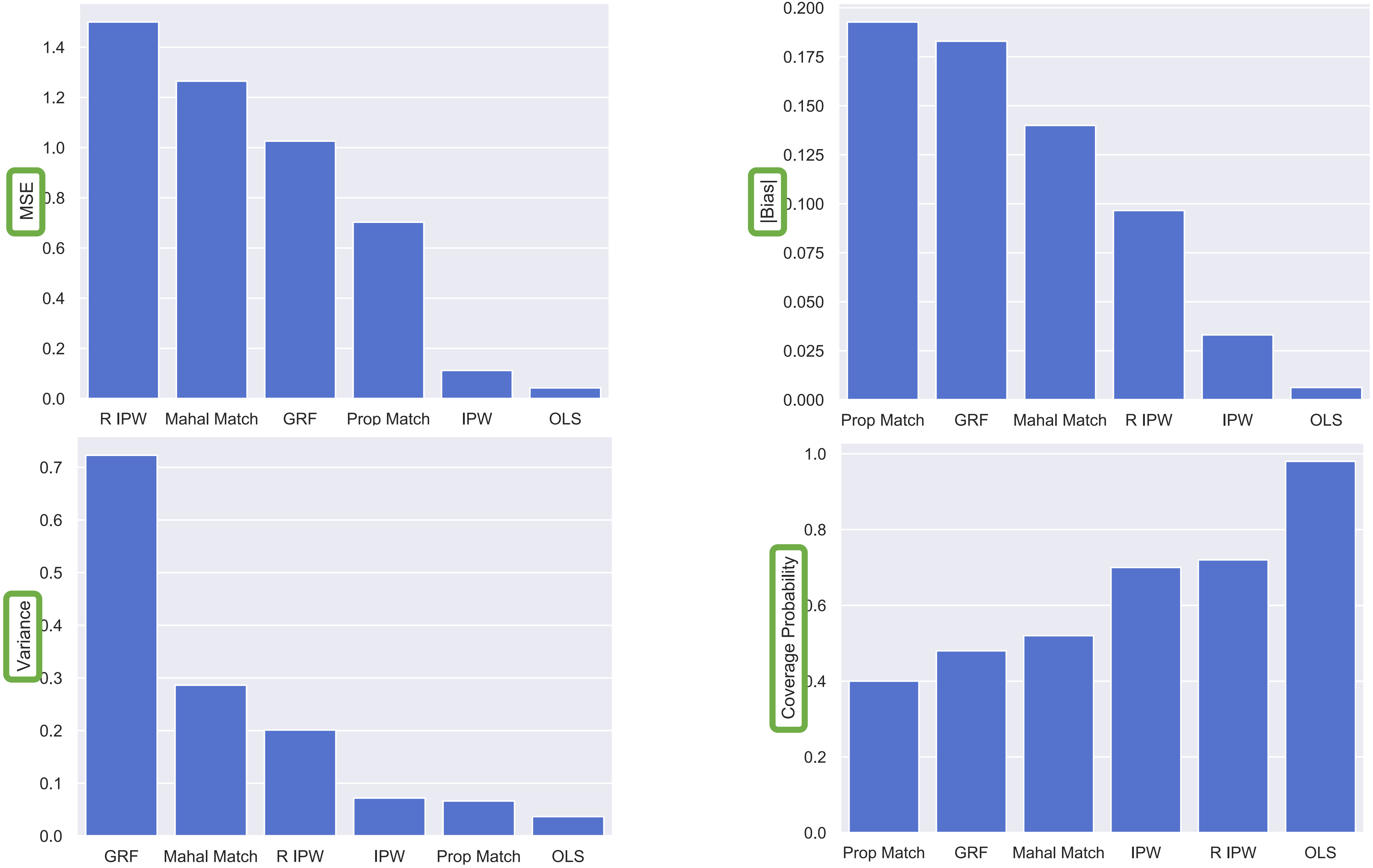


Propensity Score Matching

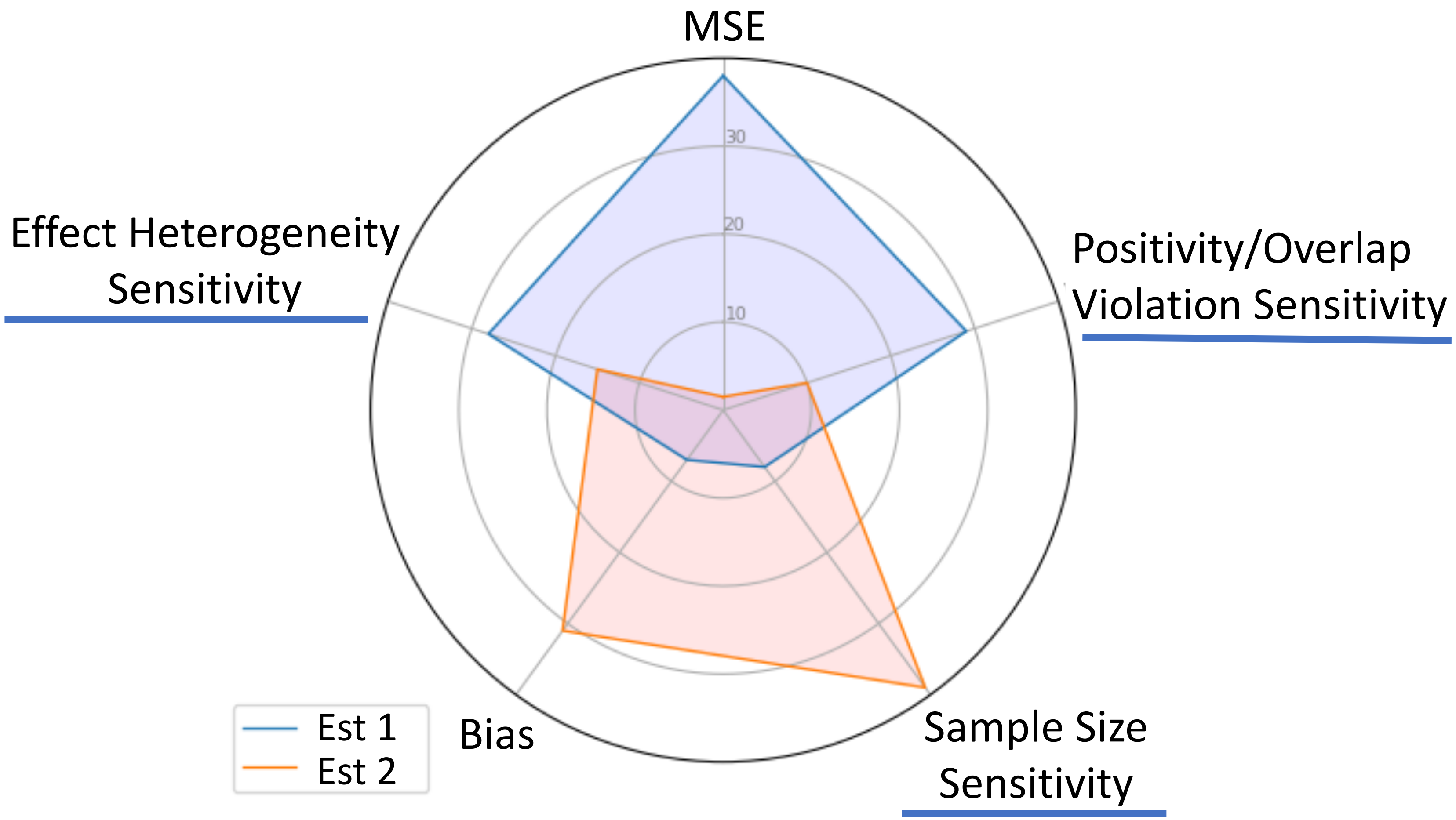
Generalized Random Forests

Mahalanobis Matching

Inverse Probability Weighting


MSE
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
R IPW
Mahal Match
GRF
Prop Match
IPW
OLS
|Bias|
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Prop Match
GRF
Mahal Match
R IPW
IPW
OLS
Variance
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GRF
Mahal Match
R IPW
IPW
Prop Match
OLS
Coverage Probability
1.0
0.8
0.6
0.4
0.2
0.0
Prop Match
GRF
Mahal Match
IPW
R IPW
OLS

Radar chart: visualize many metrics



Appendix

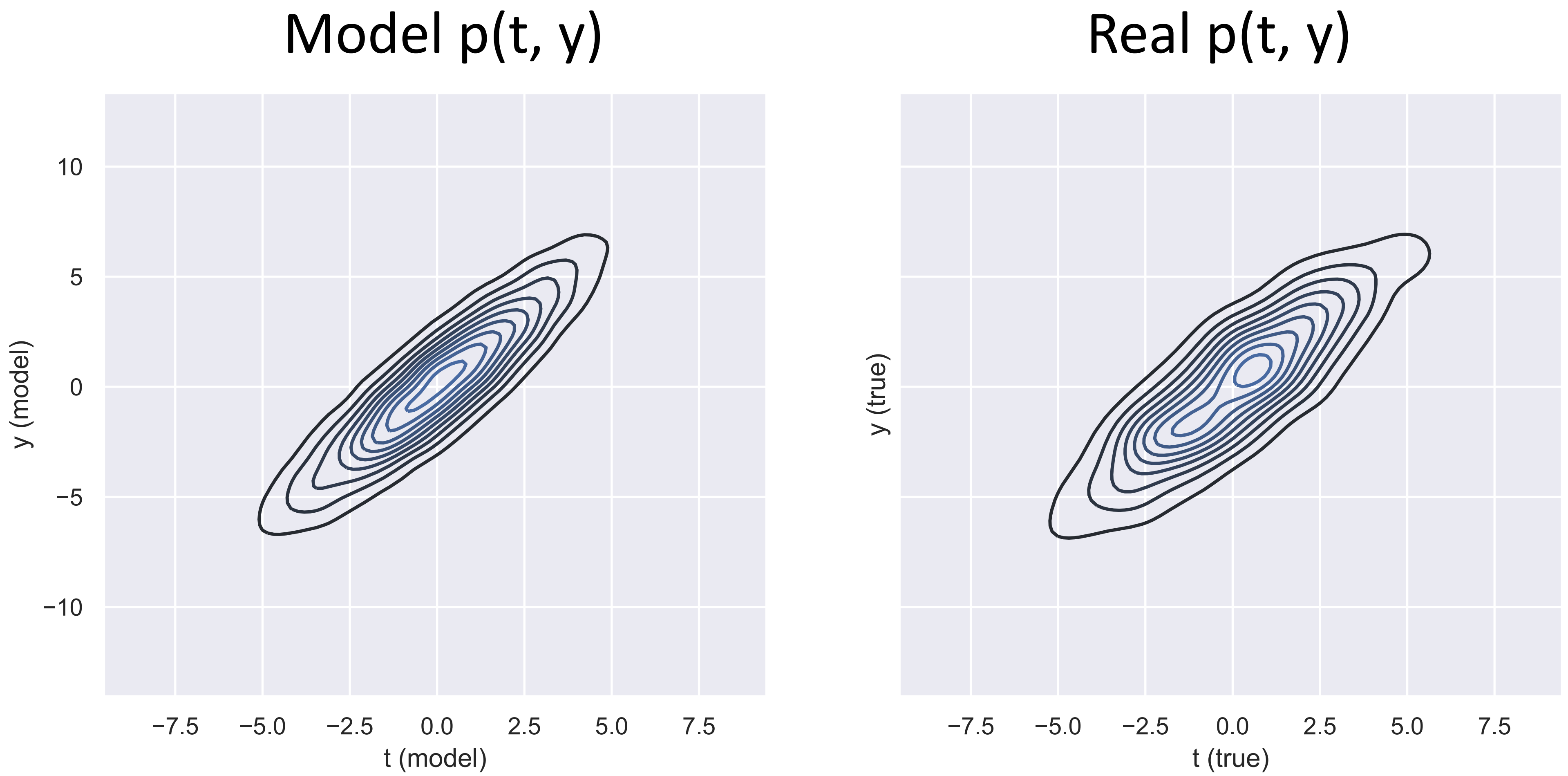
Other choices: constructed observational studies (e.g. LaLonde (1986))

1. Take RCT, so there is a ground-truth causal effect
2. Replace control group with observational data (introduces confounding)

Problems with constructed observational study:

1. Never know if there is unobserved confounding (which no estimator can correct for, making the RCT ground-truth unachievable)
2. The observational data is not necessarily the same population as the RCT data, so we don't know if the ground-truth is correct

How to know if generative model works?



And quantitative metrics such as KS test as Earth Mover’s Distance