Augmented Inverse Probability Weighting and DML for Treatment Effect Estimation

ML and Econometrics Term Project

Yu-Hsin Ho

June 1, 2023

Quick Recap of Motivation

- We want to estimate the average treatment effect (ATE) of a binary treatment D on an outcome Y
- Usually assuming SUTVA, or selection-on-observables: $\{Y(1), Y(0)\} \perp D|X$
- So we want to "control" for confounders X
- Usually this is done by linear regression
    - `reg Y D X, r`
- Problems:
    1. Relationship between Y and X is non-linear (specification error)
    2. We have more confidence on $D(X)$ instead of $Y(X)$ (e.g. experimental study)

Augmented Inverse Probability Weighting (AIPW)

- Proposed by Robins, Rotnitzky, and Zhao (1994, JASA)
- Propensity score: $m(X) = P(D = 1|X)$
- Response model: $g_d(X) = E[Y|X, D = d],\ d = 0, 1$
- **Doubly-robustness**: consistent if either $m(x)$ or $g_d(X)$ are correctly specified

$$
\begin{aligned}
\tau_{\text{AIPW}} = & \frac{1}{N}\sum_{i=1}^{N}\{g_1(X_i) - g_0(X_i) \\
& + \frac{D_i(Y_i - g_1(X_i))}{m(X_i)} - \frac{(1-D_i)(Y_i - g_0(X_i))}{1-m(X_i)}\}
\end{aligned}
$$

Simulation Study

1. Can AIPW really keep its promise?
2. The gains from using ML methods for nuisance function estimation?

Simulation Study

DGP

$$
\begin{aligned}
&Y = \tau D + X_1 X_2 + 4\sin(\pi X_3 X_4) + \exp(X_5) + \varepsilon \\
&\mathbb{P}(D = 1|X) = m(X) = \Phi(X_1 + X_3 + X_5 + X_1 X_3) \\
&D \sim \text{Bernoulli}(m(x)) \\
&X_p \sim N(1,1),\ p = 1, \cdots, 10; \ \ \varepsilon \sim N(0,1)
\end{aligned}
$$

- Treatment effect $\tau = 5$
- Confounders are X_1, X_3, X_5. Including them in the model used for estimation is sufficient to recover ATE

Simulation Study

Estimating Nuisance Functions

1. LASSO (`glmnet`)
    - `lambda`: tuned by CV
2. Random Forests (`ranger`)
    - `num.trees`: tuned by CV $\in [2000, 4000]$
    - `mtry`: tuned by CV
    - `sample.fraction` = 0.5
3. Boosting (`xgboost`)
    - `nrounds`: tuned by CV $\in [1, 6000]$
    - `max_depth` = 2,
    - `eta` = 0.01
    - `subsample` = 0.5

Simulation Study

Specifications

Spec	Predictors in $m(X)$	Predictors in $g(X)$
both	$X_1 \cdots X_{10}$	$X_1 \cdots X_{10}$
pscore	$X_1 \cdots X_{10}$	$X_6 \cdots X_{10}$
response	$X_6 \cdots X_{10}$	$X_1 \cdots X_{10}$

Simulation Study

Estimators

1. AIPW:
   $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}[g_1(X_i) - g_0(X_i)$
   $+ \frac{D_i(Y_i - g_1(X_i))}{m(X_i)} - \frac{(1-D_i)(Y_i - g_0(X_i))}{1-m(X_i)}]$
2. IPW: $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}[\frac{DY}{m(X)} - \frac{(1-D)Y}{1-m(X)}]$
3. OLS: $Y = \hat{\tau}D + X'\hat{\beta} + \hat{\varepsilon}$
4. PLS: $(Y - \hat{g}(X)) = \hat{\tau}(D - \hat{m}(X)) + \hat{\varepsilon}$

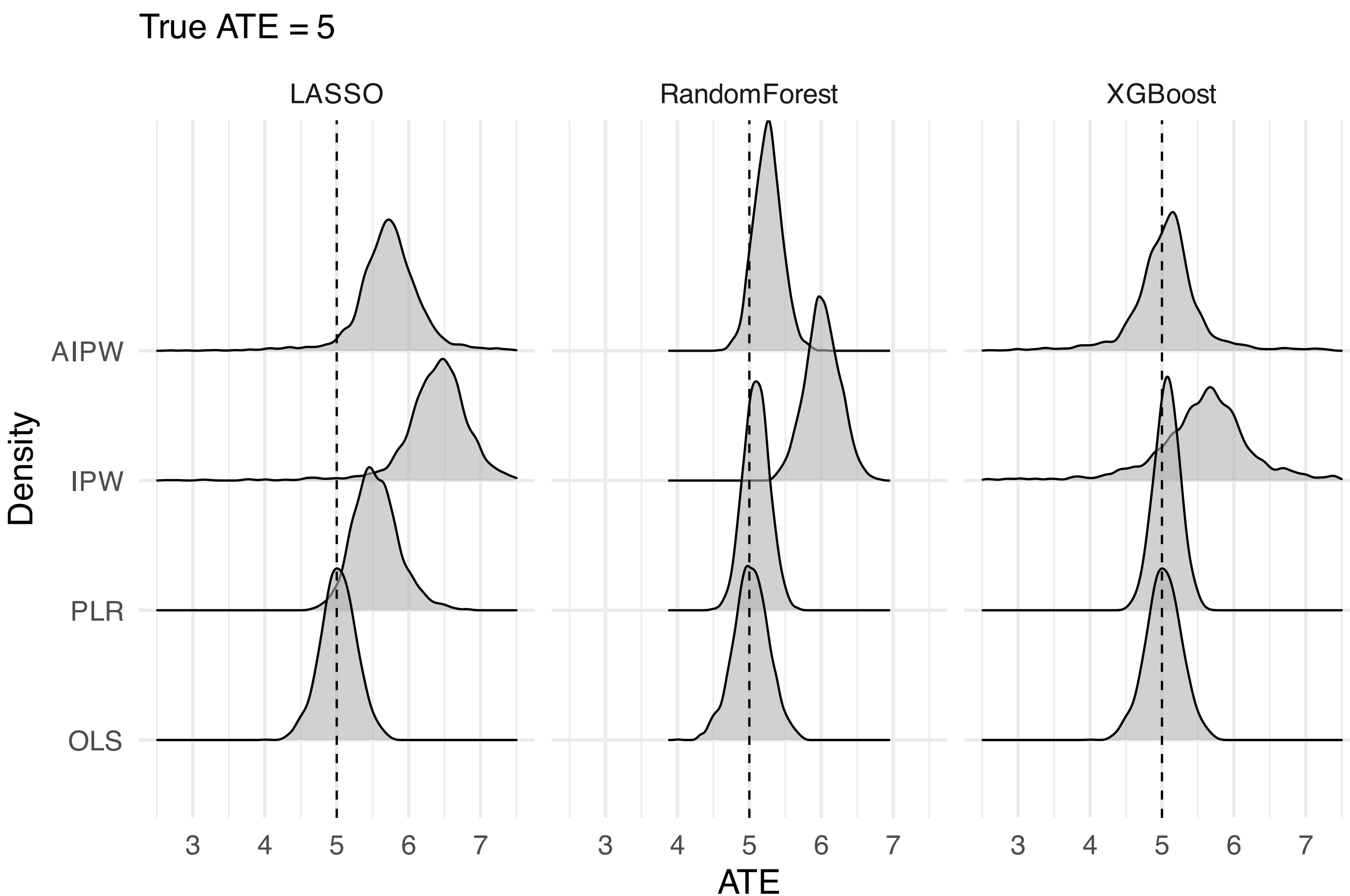
We get $3 \times 3 \times 4 = 36$ ATE estimates per iteration.

Simulation Study

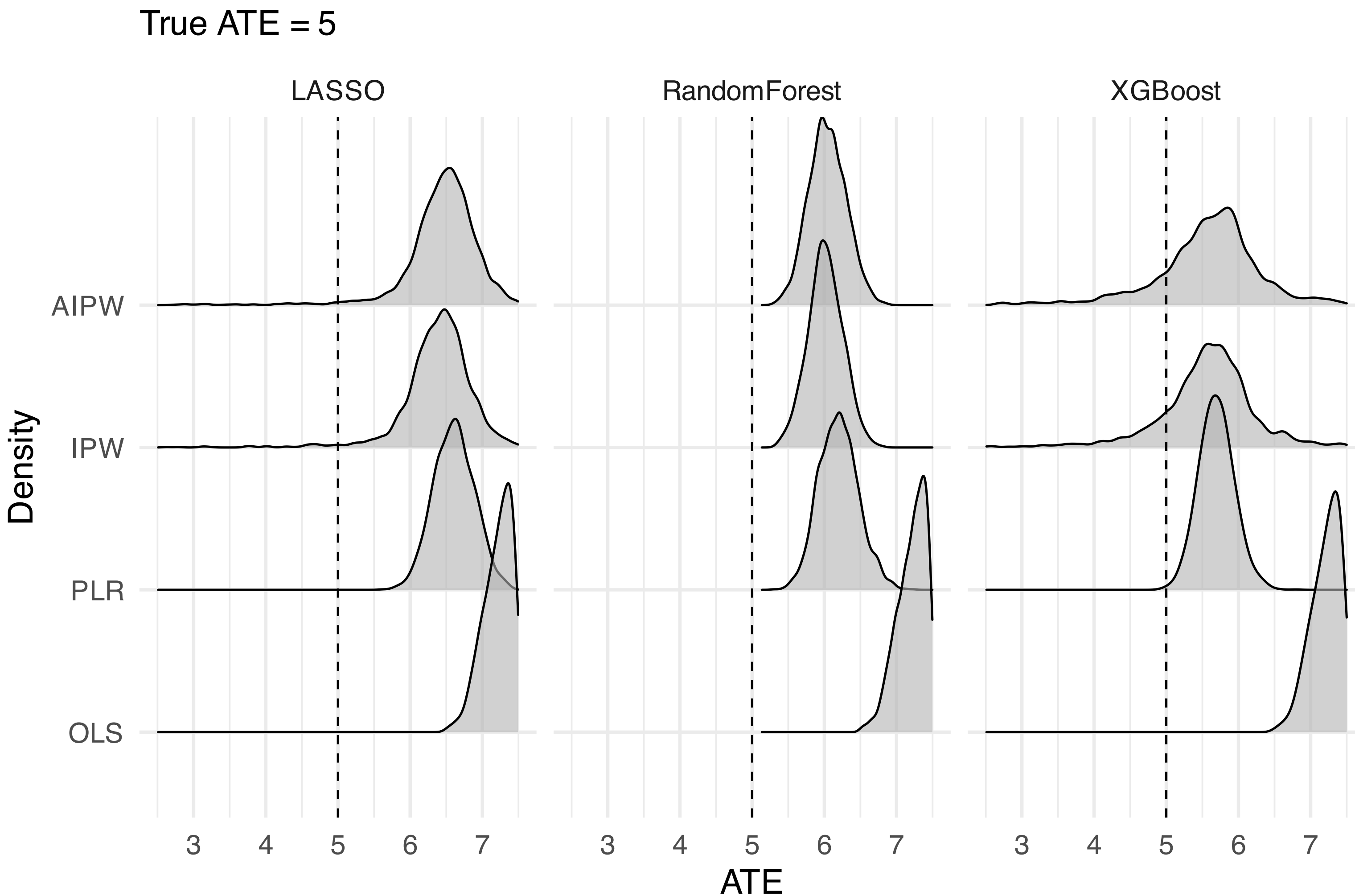
Procedures

1. Generate 2000 samples from DGP, each with 2000 observations
2. Use 1st sample to tune hyperparameters (10-fold CV)
3. Get ATE estimates with 2-fold crossfitting

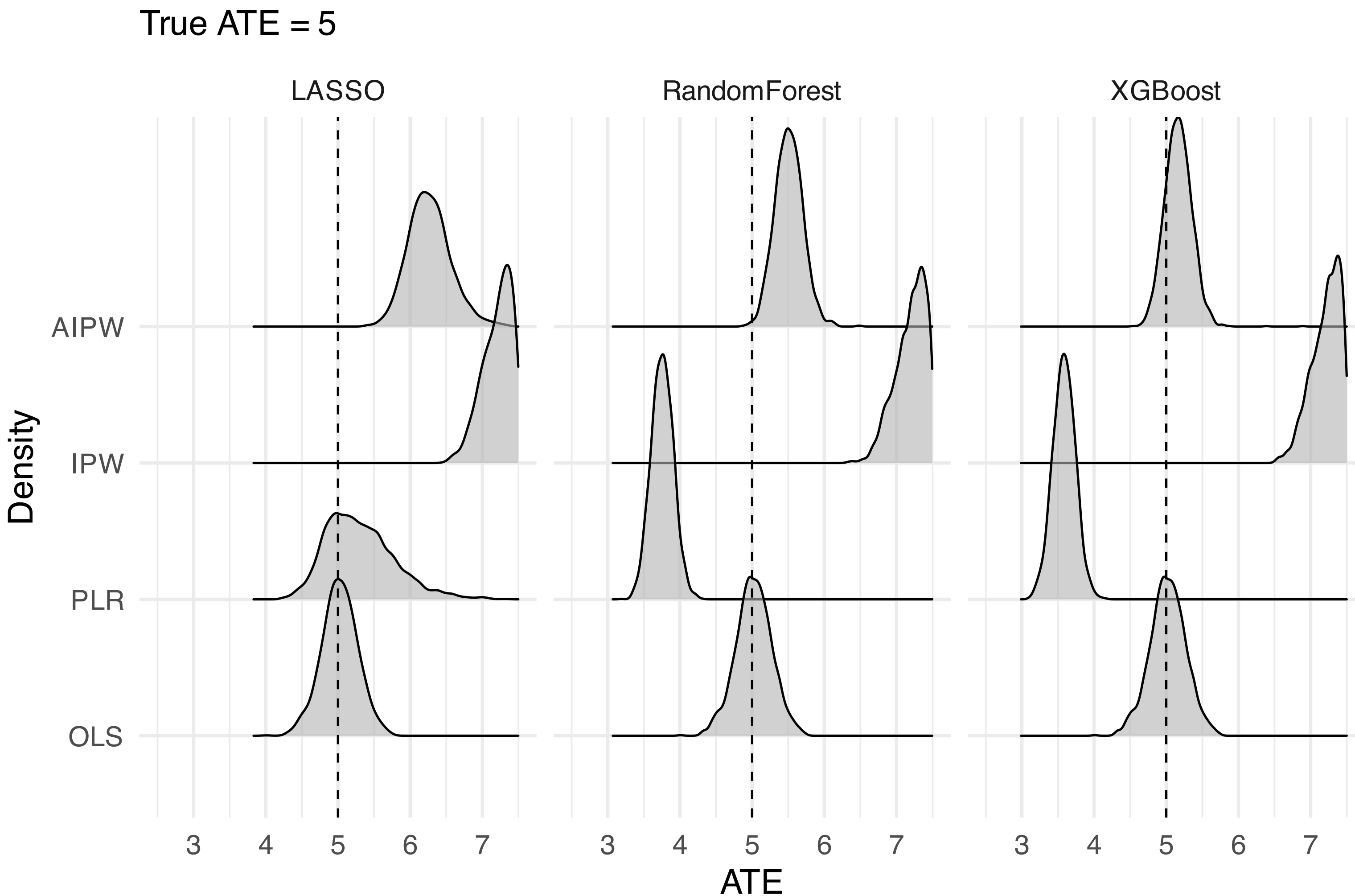
Results: Both specified correctly



Results: pscore specified correctly



Results: response specified correctly

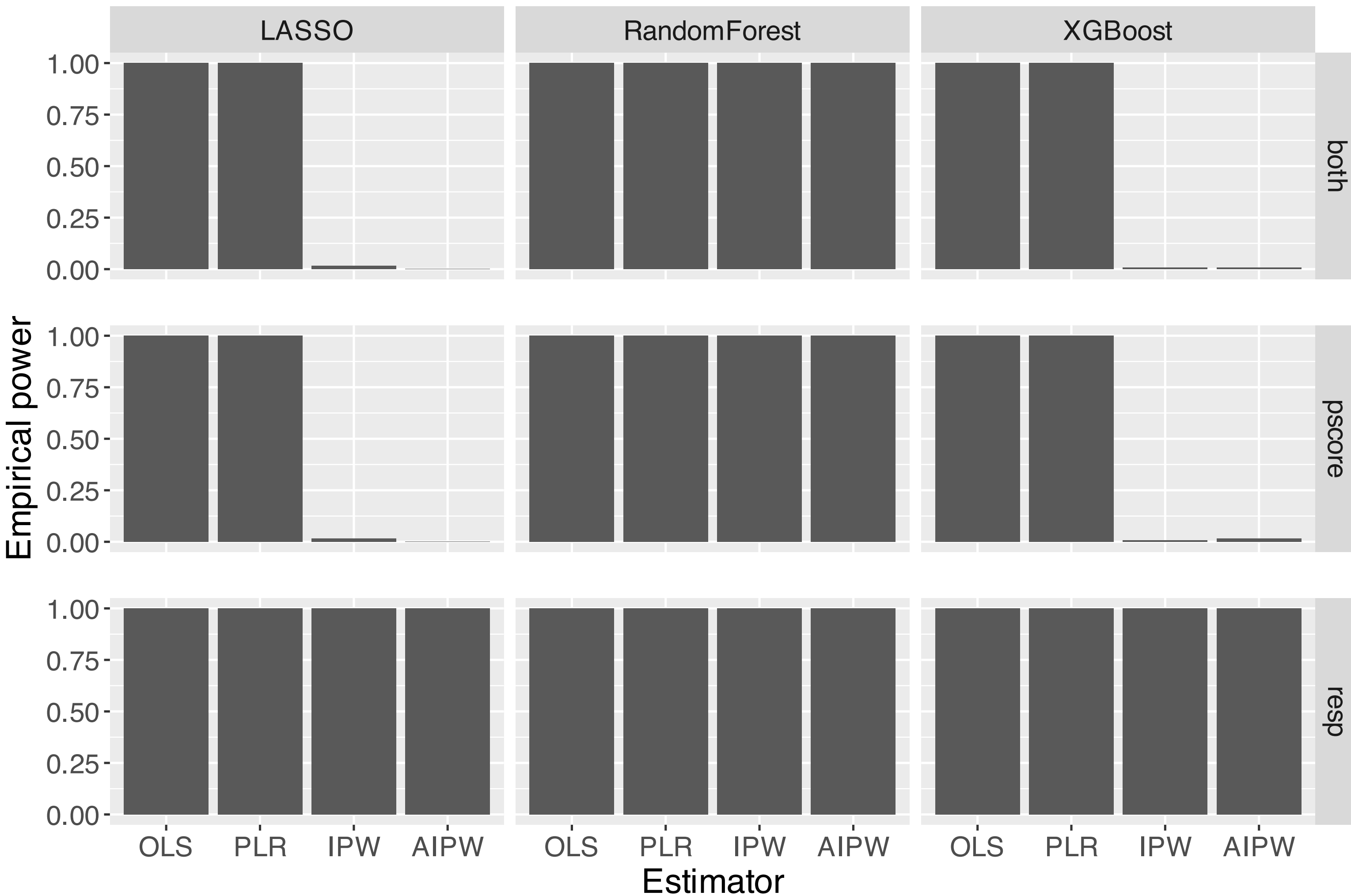


estimator	Bias	RMSE	S.D.
both - LASSO			
OLS	0.022	0.249	0.249
PLR	0.550	0.635	0.319
IPW	1.493	5.316	5.104
AIPW	1.227	21.434	21.405
both - RandomForest			
OLS	0.022	0.249	0.249
PLR	0.097	0.205	0.180
IPW	1.017	1.046	0.242
AIPW	0.255	0.320	0.194
both - XGBoost			
OLS	0.022	0.249	0.249
PLR	0.069	0.191	0.178
IPW	0.465	11.855	11.849
AIPW	−0.096	6.047	6.047

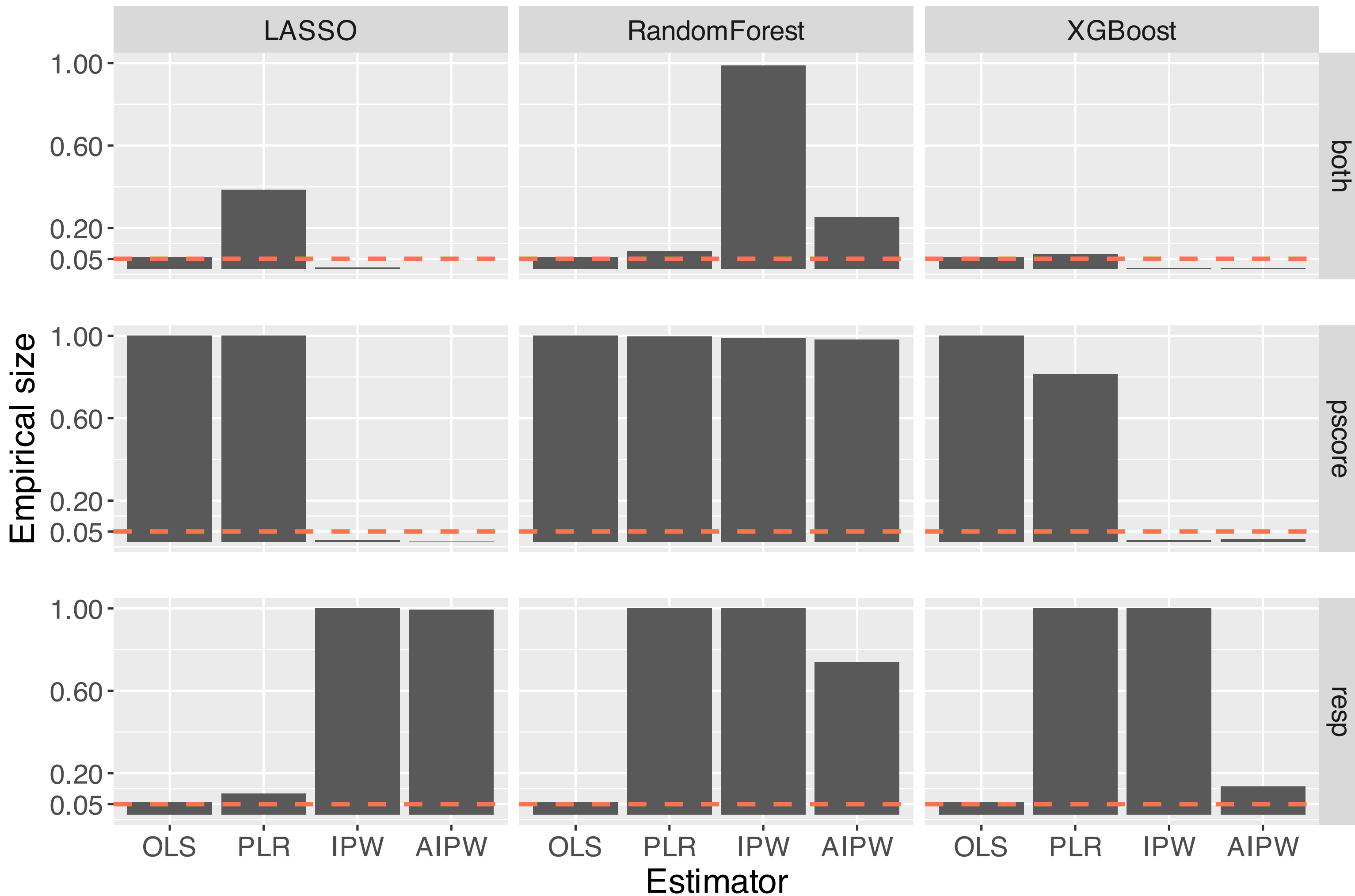
estimator	Bias	RMSE	S.D.
pscore - LASSO			
OLS	2.357	2.376	0.295
PLR	1.605	1.629	0.281
IPW	1.493	5.316	5.104
AIPW	2.020	21.586	21.497
pscore - RandomForest			
OLS	2.357	2.376	0.295
PLR	1.197	1.228	0.273
IPW	1.017	1.046	0.242
AIPW	1.043	1.074	0.258
pscore - XGBoost			
OLS	2.357	2.376	0.295
PLR	0.695	0.736	0.244
IPW	0.695	15.178	15.166
AIPW	0.813	7.971	7.931

estimator	Bias	RMSE	S.D.
resp - LASSO			
OLS	0.022	0.249	0.249
PLR	0.309	0.566	0.475
IPW	2.357	2.375	0.298
AIPW	1.266	1.300	0.296
resp - RandomForest			
OLS	0.022	0.249	0.249
PLR	−1.245	1.254	0.152
IPW	2.357	2.378	0.315
AIPW	0.509	0.545	0.194
resp - XGBoost			
OLS	0.022	0.249	0.249
PLR	−1.415	1.424	0.156
IPW	2.358	2.376	0.296
AIPW	0.166	0.251	0.188

Empirical Power ($\tau = 5, H_0 = 0$)



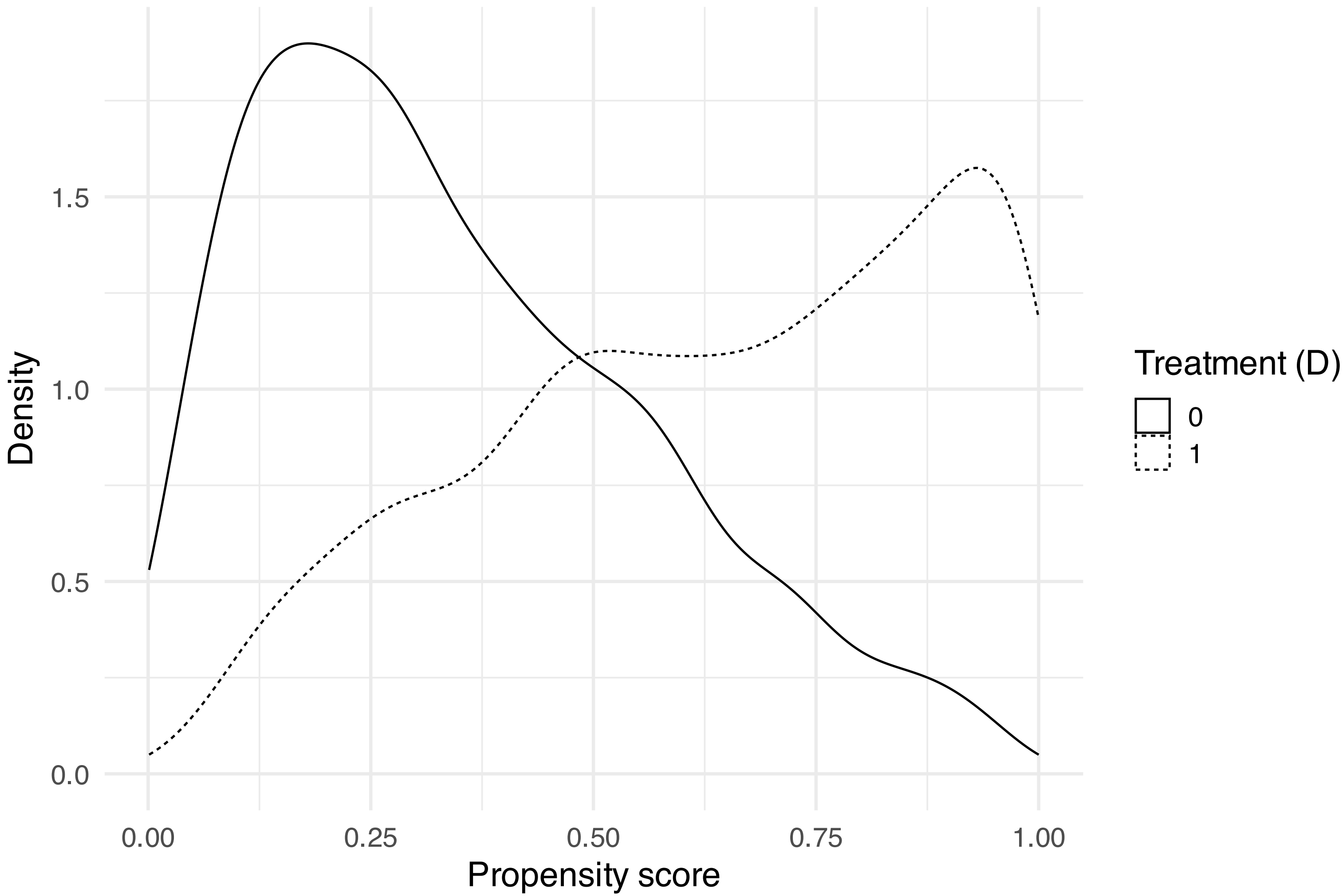
Empirical Size ($\tau = 5, H_0 = 5$)



Promises AIPW can/cannot keep

- AIPW is indeed doubly robust
- It works well when propensity score is not extreme
  - E.g. the spec. that only response is correctly specified
  - Higher efficiency than OLS
- But the curse is that inverse-weighting based estimators suffer from sensitivity to extreme propensity scores
  - It'll explode the estimate when $m(X)$ is close to 0 or 1
  - High variance, lacks of power

Distribution of Propensity Score



Conclusions

Surprisingly, OLS (containing only 1st-order term) is not bad when relevent variables are included

Still, it does not contain treatment assignment information which we sometimes are more confident with

Conclusions

What we want is doubly-robustness but stable to extreme propensity scores

- Key: Prevent extreme weighting
- Some refinements are done:
  - Normalized AIPW (Rostami, and Saarela 2021)
  - Overlap weighting (Li, Morgan, and Zaslavsky 2018, JASA)
    - off-the-shelf function implemented in `grf` R library
    - fast, and works quite well