

Meta-Learners for Estimation of Causal Effects: Finite Sample Cross-Fit Performance

EEA Annual Congress 2022

Gabriel Okasa

TIS-EPFL

Chair for Technology and Innovation Strategy
Swiss Federal Institute of Technology in Lausanne, Switzerland

Motivation

Motivation

Research Question:

- ▶ What is the finite sample performance of machine learning based meta-learners using cross-fitting for estimation of heterogeneous causal effects?

Motivation

Meta-Learners:

- ▶ flexibility in estimation of heterogeneous causal effects
- ▶ generality in the choice of the learning method (Künzel et al. 2019)
- ▶ lack of unifying simulation evidence for assessment of meta-learners

Cross-Fitting:

- ▶ overfitting bias due to estimation of nuisance functions (Chernozhukov et al. 2018)
- ▶ sample-splitting and cross-fitting to reduce bias and regain efficiency
- ▶ lack of simulation evidence for assessment of estimation procedures

Literature

Literature

- ▶ effect of job training on employment (Knaus 2020)
- ▶ effect of special education programs on academic performance (Sallin 2021)
- ▶ effect of waste pricing programs on pollution (Valente 2022)
- ▶ effect of quarantines on covid spread (Kristjanpoller et al. 2021)
- ▶ effect of blood pressure therapy on disease risk (Duan et al. 2019)
- ▶ effect of marketing campaigns on sales revenue (Gubela and Lessmann 2021)

Literature

- ▶ few simulation studies on machine learning estimation of heterogeneous causal effects (Knaus et al. 2020; Naghi and Wirths 2021)
- ▶ little evidence on the impact of sample-splitting and cross-fitting in finite samples (Jacob 2020; Zivich and Breskin 2021)
- ▶ limited results on the finite sample performance of meta-learners for estimation of causal effects (Curth and Schaar 2021)
- ▶ convergence performance of meta-learners based on cross-fitting unexplored so far

Notation

Notation

Data Inputs:

- ▶ treatment indicator $W_i \in \{0, 1\}$
- ▶ outcome variable Y_i
- ▶ covariates X_i

Nuisance Functions:

- ▶ propensity score function $e(x) = \mathbb{P}[W_i = 1 \mid X_i = x]$
- ▶ response function $\mu(x) = \mathbb{E}[Y_i \mid X_i = x]$

Meta-Learning:

- ▶ treatment effect function $\tau(x) = \zeta(W_i, X_i, Y_i, e(x), \mu(x))$

Identification

Identification

- ▶ Potential Outcomes Framework (Rubin 1974)
- ▶ potential outcome under treatment $Y_i(1)$ and under control $Y_i(0)$

Assumption (Conditional Independence)

$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i = x, \forall x \in \text{supp}(X_i)$.

Assumption (Common Support)

$0 < \mathbb{P}[W_i = 1 \mid X_i = x] < 1, \forall x \in \text{supp}(X_i)$.

Assumption (SUTVA)

$Y_i = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$.

Assumption (Exogeneity)

$X_i(0) = X_i(1)$.

Identification

Individual Treatment Effect (ITE):

$$\xi_i = Y_i(1) - Y_i(0).$$

Conditional Average Treatment Effect (CATE):

$$\begin{aligned}\tau(x) &= \mathbb{E}[\xi_i \mid X_i = x] \\ &= \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] \\ &= \mathbb{E}[Y_i(1) \mid X_i = x] - \mathbb{E}[Y_i(0) \mid X_i = x] \\ &= \mathbb{E}[Y_i(1) \mid X_i = x, W_i = 1] - \mathbb{E}[Y_i(0) \mid X_i = x, W_i = 0] \\ &= \mathbb{E}[Y_i \mid X_i = x, W_i = 1] - \mathbb{E}[Y_i \mid X_i = x, W_i = 0]\end{aligned}$$

Meta-Learners

Meta-Learners

- ▶ decompose the causal problem into prediction problems (Künzel et al. 2019)
- ▶ generality in the choice of the learning method (Curth and Schaar 2021)
- ▶ do not modify the objective function, i.e. MSE minimization
- ▶ can be tuned and adapted to particular types of data, i.e. binary, sparse, etc.
- ▶ no restrictions in the choice of software libraries

Meta-Learners

- ▶ S-learner (Lo 2002): single response function
- ▶ T-learner (Hansotia and Rukstales 2002): two response functions
- ▶ X-learner (Künzel et al. 2019): two response functions and propensity score
- ▶ DR-learner (Kennedy 2020): two response functions and propensity score
- ▶ R-learner (Nie and Wager 2021): single response function and propensity score

Meta-Learners

Example

Algorithm 1: R-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_R(x) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTION;

estimate: $\mu(x) = E[Y_i \mid X_i = x]$ in $\{(X_i, Y_i)\}^T$;

PROPENSITY SCORE;

estimate: $e(x) = P[W_i = 1 \mid X_i = x]$ in $\{(X_i, W_i)\}^T$;

MODIFIED OUTCOME;

predict: $\hat{\phi}_i = \frac{(Y_i - \hat{\mu}(X_i))}{(W_i - \hat{e}(X_i))}$ in $\{(X_i, Y_i, W_i)\}^T$;

CATE FUNCTION;

estimate: $\tau_R(x) = E[\hat{\phi}_i \mid X_i = x]$ weighted by $(W_i - \hat{e}(X_i))^2$ in $\{(X_i, Y_i, W_i)\}^T$;

predict: $\hat{\tau}_R(X_i) = \hat{E}[\hat{\phi}_i \mid X_i = x]$ in $\{(X_i)\}^V$

end

Cross-Fitting

Cross-Fitting

- ▶ using the same data for estimation of the nuisance functions and the CATE function results in overfitting bias
- ▶ mitigate overfitting by estimating nuisance functions on one part of the data and the CATE function on the other one
- ▶ sample-splitting reduces the bias but increases variance
- ▶ cross-fitting regains the efficiency by swapping the samples and averaging the estimates (Chernozhukov et al. 2018)
- ▶ Newey and Robins (2018) propose *double* sample-splitting and *double* cross-fitting, where each nuisance function is estimated on separate part of the data

Cross-Fitting

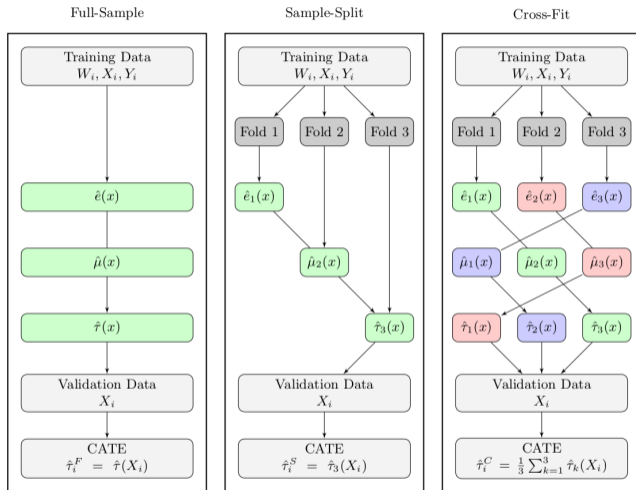


Figure 1: Illustration of the full-sample, sample-splitting and cross-fitting procedure.

Cross-Fitting

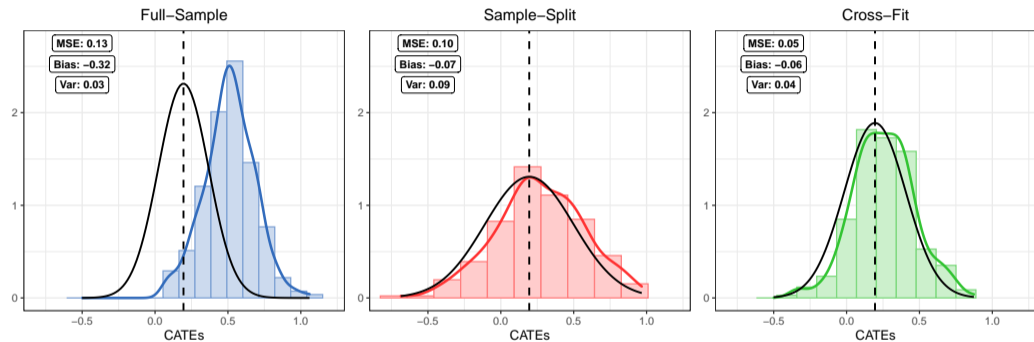


Figure 2: CATE distributions under full-sample, sample-splitting and cross-fitting estimation.

Analysis

Analysis

Framework:

- ▶ identification based on the selection-on-observables strategy
- ▶ implementations based on the full-sample, sample-splitting and cross-fitting
- ▶ meta-learners based on the random forest algorithm

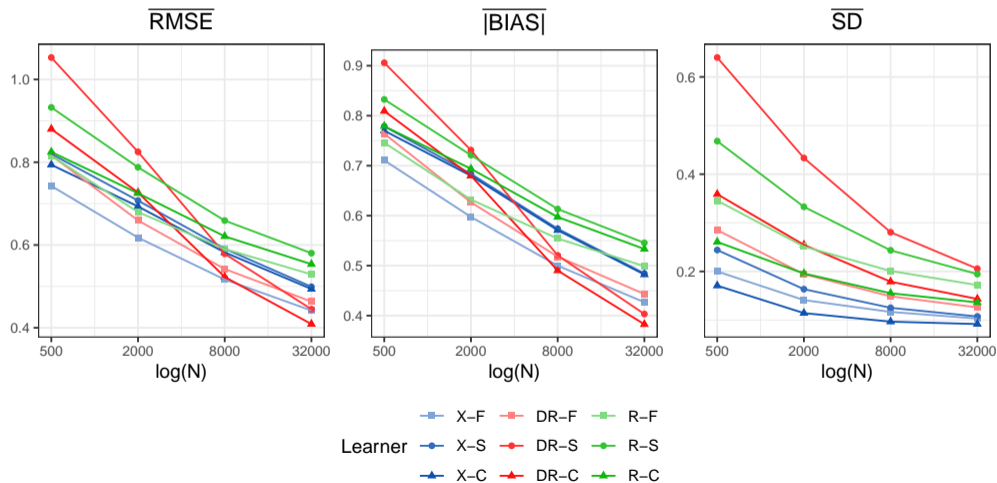
Simulation Study:

- ▶ synthetic and semi-synthetic simulations
- ▶ DGPs with unequal treatment shares, non-linearities and large-dimensions
- ▶ varying sample sizes up to 32'000 observations

Results

Results

Figure 3: Results for Main Simulation: unbalanced treatment and nonlinear CATE



Results

Estimation Procedures:

- ▶ sample-splitting effectively reduces the bias in large samples
- ▶ cross-fitting additionally regains the full sample size efficiency
- ▶ full-sample estimation preferable in small samples when using machine learning

Meta-Learners:

- ▶ varying impacts of the estimation procedures on the performance of meta-learners
- ▶ X-learner suitable for imbalanced treatment shares in any version and sample size
- ▶ DR-learner suitable for balanced treatment shares using cross-fitting in large samples

Conclusion

Conclusion

Takeaway:

- ▶ The performance of meta-learners varies greatly but the choice of the meta-learner and the estimation procedure can be guided by observable data characteristics.

Thank You for Your Attention!







`gabriel.okasa@epfl.ch`
`okasag.github.io`

References







References I

-  Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. “Double/debiased machine learning for treatment and structural parameters”. In: *Econometrics Journal* 21.1 (2018), pp. 1–68.
-  Curth, Alicia and Mihaela van der Schaar. “Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1810–1818.
-  Duan, Tony, Pranav Rajpurkar, Dillon Laird, Andrew Y. Ng, and Sanjay Basu. “Clinical Value of Predicting Individual Treatment Effects for Intensive Blood Pressure Therapy”. In: *Circulation: Cardiovascular Quality and Outcomes* 12.3 (2019).
-  Gubela, Robin M. and Stefan Lessmann. “Uplift modeling with value-driven evaluation metrics”. In: *Decision Support Systems* 150 (2021).
-  Hansotia, Behram and Brad Rukstales. “Incremental value modeling”. In: *Journal of Interactive Marketing* 16.3 (2002), pp. 35–46.



References II

-  Jacob, Daniel. “Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects”. In: *arXiv preprint arXiv:2007.02852* (2020).
-  Kennedy, Edward H. “Optimal doubly robust estimation of heterogeneous causal effects”. In: *arXiv preprint arXiv:2004.14497* (2020).
-  Knaus, Michael C, Michael Lechner, and Anthony Strittmatter. “Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence”. In: *The Econometrics Journal* (2020).
-  Knaus, Michael C. “Double Machine Learning based Program Evaluation under Unconfoundedness”. In: *arXiv preprint arXiv:2003.03191* (2020).
-  Kristjanpoller, Werner, Kevin Michell, and Marcel C. Minutolo. “A causal framework to determine the effectiveness of dynamic quarantine policy to mitigate COVID-19”. In: *Applied Soft Computing* 104 (2021).
-  Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. “Metalearners for estimating heterogeneous treatment effects using machine learning”. In: *Proceedings of the National Academy of Sciences* 116.10 (2019), pp. 4156–4165.

References III

-  Lo, Victor S.Y. “The true lift model: a novel data mining approach to response modeling in database marketing”. In: *ACM SIGKDD Explorations Newsletter* 4.2 (2002), pp. 78–86.
-  Naghi, Andrea A. and Christian P. Wirths. “Finite Sample Evaluation of Causal Machine Learning Methods: Guidelines for the Applied Researcher”. In: *SSRN Electronic Journal* (2021).
-  Newey, Whitney K. and James R. Robins. “Cross-fitting and fast remainder rates for semiparametric estimation”. In: *arXiv preprint arXiv:1801.09138* (2018).
-  Nie, X and S Wager. “Quasi-oracle estimation of heterogeneous treatment effects”. In: *Biometrika* 108.2 (2021), pp. 299–319.
-  Rubin, Donald B. “Estimating causal effects of treatment in randomized and nonrandomized studies”. In: *Journal of Educational Psychology* 66.5 (1974), pp. 688–701.
-  Sallin, Aurélien. “Estimating returns to special education: combining machine learning and text analysis to address confounding”. In: *arXiv preprint arXiv:2110.08807* (2021).

References IV

-  Valente, Marica. “Policy evaluation of waste pricing programs using heterogeneous causal effect estimation”. In: *arXiv preprint arXiv:2010.01105* (2022).
-  Zivich, Paul N. and Alexander Breskin. “Machine learning for causal inference: On the use of cross-fit estimators”. In: *Epidemiology* (2021), pp. 393–401.

Appendix

Algorithm 2: S-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_S(x) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTION;

estimate: $\mu(x, w) = E[Y_i \mid X_i = x, W_i = w]$ in $\{(X_i, Y_i, W_i)\}^T$;

CATE FUNCTION;

define: $\hat{\tau}_S(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$;

predict: $\hat{\tau}_S(X_i) = \hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0)$ in $\{(X_i)\}^V$

end

Algorithm 3: T-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_T(X_i) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTIONS;

estimate: $\mu(x, 1) = E[Y_i \mid X_i = x, W_i = 1]$ in $\{(X_i, Y_i)\}_{W_i=1}^T$;

estimate: $\mu(x, 0) = E[Y_i \mid X_i = x, W_i = 0]$ in $\{(X_i, Y_i)\}_{W_i=0}^T$;

CATE FUNCTION;

define: $\hat{\tau}_T(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$;

predict: $\hat{\tau}_T(X_i) = \hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0)$ in $\{(X_i)\}^V$

end

Algorithm 4: X-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_X(X_i) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTIONS;

estimate: $\mu(x, 1) = E[Y_i \mid X_i = x, W_i = 1]$ in $\{(X_i, Y_i)\}_{W_i=1}^T$;

estimate: $\mu(x, 0) = E[Y_i \mid X_i = x, W_i = 0]$ in $\{(X_i, Y_i)\}_{W_i=0}^T$;

IMPUTED EFFECTS;

predict: $\tilde{\xi}_i^1 = Y_i - \hat{\mu}(X_i, 0)$ in $\{(X_i, Y_i)\}_{W_i=1}^T$;

predict: $\tilde{\xi}_i^0 = Y_i - \hat{\mu}(X_i, 1)$ in $\{(X_i, Y_i)\}_{W_i=0}^T$;

TREATMENT EFFECTS;

estimate: $\tau(x, 1) = E[\tilde{\xi}_i^1 \mid X_i = x, W_i = 1]$ in $\{(X_i, Y_i)\}_{W_i=1}^T$;

estimate: $\tau(x, 0) = E[\tilde{\xi}_i^0 \mid X_i = x, W_i = 0]$ in $\{(X_i, Y_i)\}_{W_i=0}^T$;

PROPENSITY SCORE;

estimate: $e(x) = P[W_i = 1 \mid X_i = x]$ in $\{(X_i, W_i)\}^T$;

CATE FUNCTION;

define: $\hat{\tau}_X(x) = \hat{e}(x) \cdot \hat{\tau}(x, 0) + (1 - \hat{e}(x)) \cdot \hat{\tau}(x, 1)$;

predict: $\hat{\tau}_X(X_i) = \hat{e}(X_i) \cdot \hat{\tau}(X_i, 0) + (1 - \hat{e}(X_i)) \cdot \hat{\tau}(X_i, 1)$ in $\{(X_i)\}^V$

end

Meta-Learners

X-learner

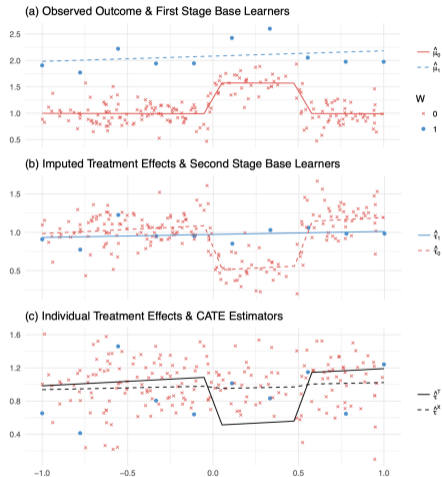


Figure 4: CATE Estimation via T-Learner vs. X-Learner (Künzel et al. 2019)

Algorithm 5: DR-LEARNER

Input: Training Data: $\{(X_i, Y_i, W_i)\}^T$, Validation Data: $\{(X_i)\}^V$

Output: CATE: $\hat{\tau}_{DR}(x) = \hat{E}[Y_i(1) - Y_i(0) \mid X_i = x]$

begin

RESPONSE FUNCTIONS;

estimate: $\mu(x, 1) = E[Y_i \mid X_i = x, W_i = 1]$ in $\{(X_i, Y_i)\}_{W_i=1}^T$;

estimate: $\mu(x, 0) = E[Y_i \mid X_i = x, W_i = 0]$ in $\{(X_i, Y_i)\}_{W_i=0}^T$;

PROBABILITY SCORE;

estimate: $e(x) = P[W_i = 1 \mid X_i = x]$ in $\{(X_i, W_i)\}^T$;

PSEUDO OUTCOME;

predict: $\hat{\psi}_i = \frac{W_i(Y_i - \hat{\mu}(X_i, 1))}{\hat{e}(X_i)} - \frac{(1 - W_i)(Y_i - \hat{\mu}(X_i, 0))}{1 - \hat{e}(X_i)} + \hat{\mu}(X_i, 1) - \hat{\mu}(X_i, 0)$ in $\{(X_i, Y_i, W_i)\}^T$;

CATE FUNCTION;

estimate: $\tau_{DR}(x) = E[\hat{\psi}_i \mid X_i = x]$ in $\{(X_i, Y_i, W_i)\}^T$;

predict: $\hat{\tau}_{DR}(X_i) = \hat{E}[\hat{\psi}_i \mid X_i = x]$ in $\{(X_i)\}^V$

end

Meta-Learners

DR-learner

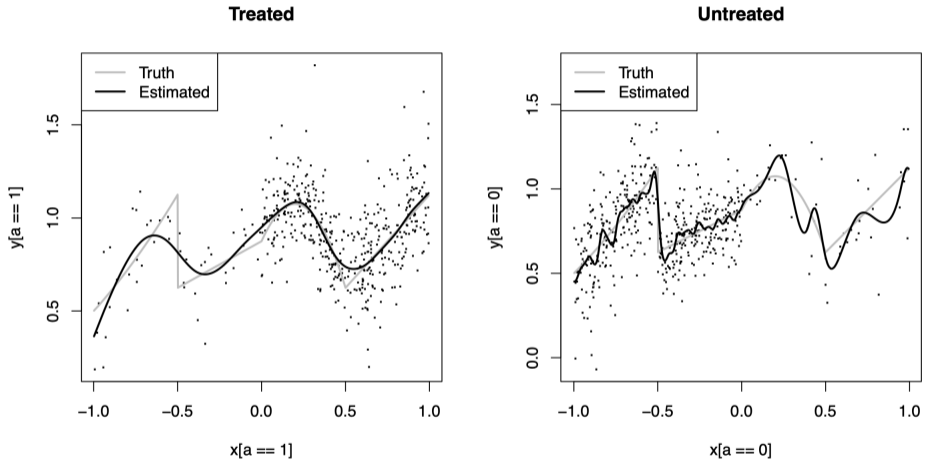


Figure 5: Smoothing Spline Estimation of the Response Functions (Kennedy 2020)

Meta-Learners

DR-learner

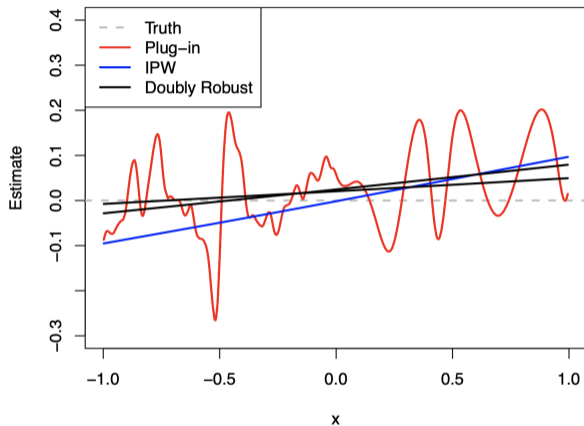
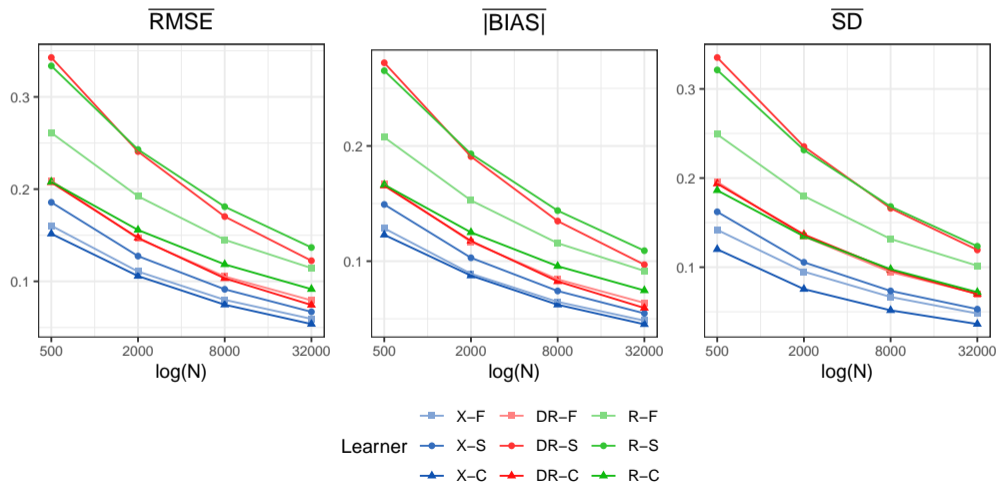


Figure 6: CATE Estimation via T-Learner vs. DR-Learner (Kennedy 2020)

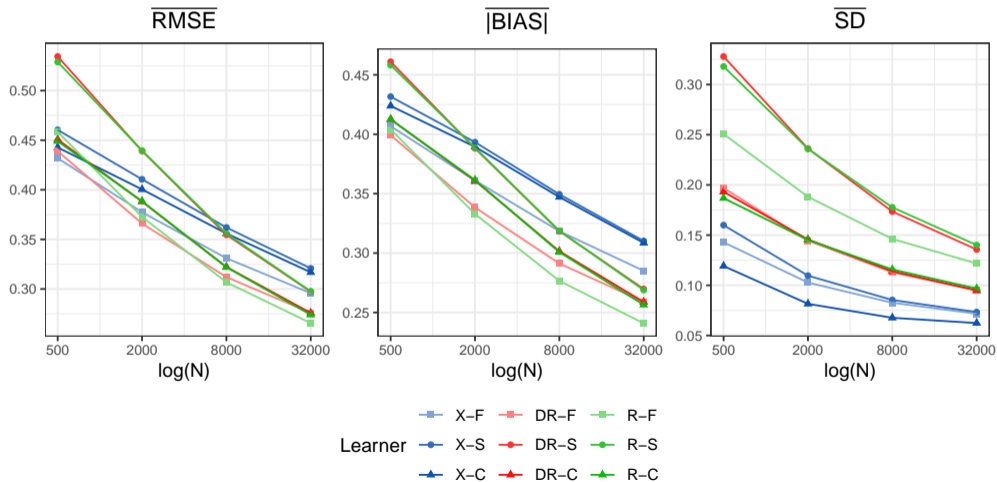
Results

Figure 7: Results for Simulation 1: balanced treatment and constant zero CATE



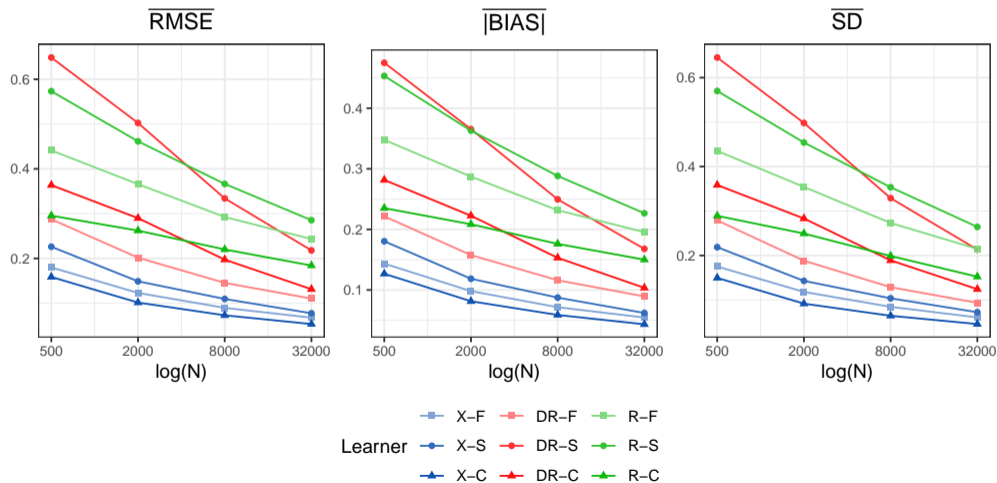
Results

Figure 8: Results for Simulation 2: balanced treatment and complex nonlinear CATE



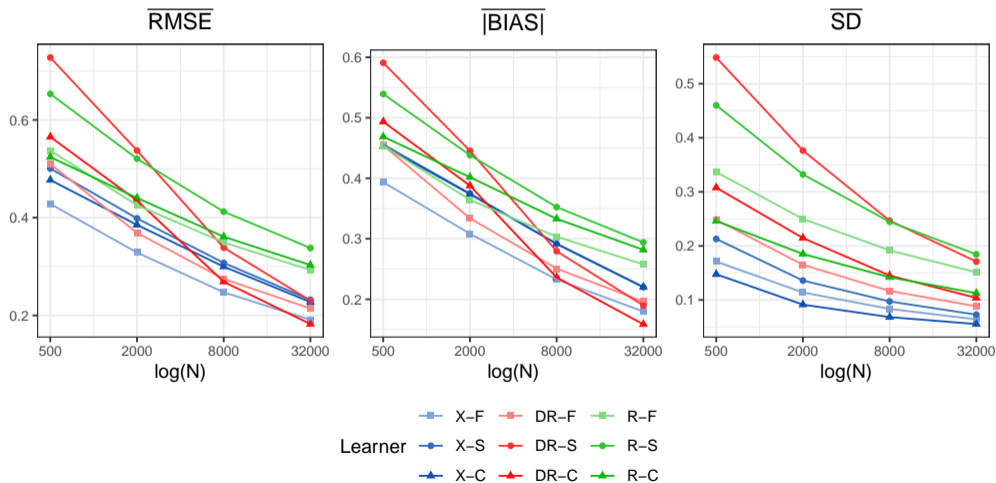
Results

Figure 9: Results for Simulation 3: highly unbalanced treatment and constant non-zero CATE



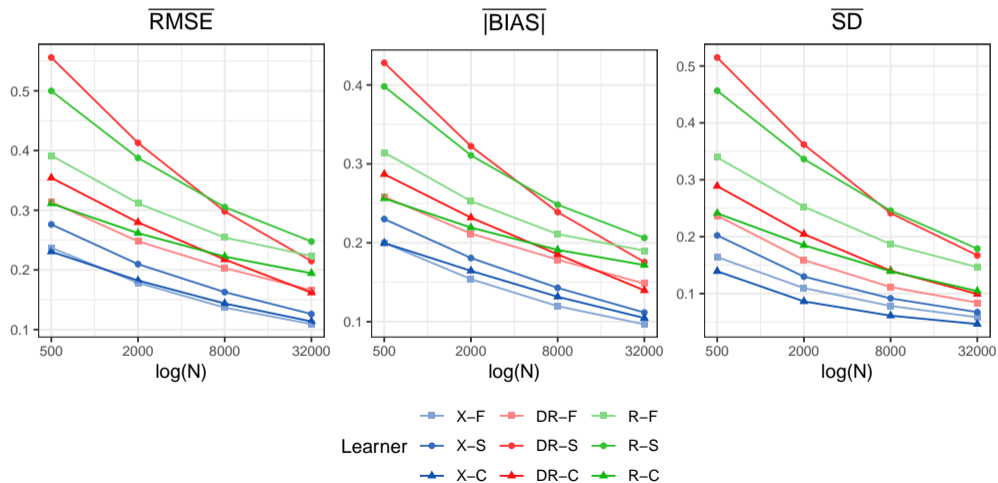
Results

Figure 10: Results for Simulation 4: unbalanced treatment and simple CATE



Results

Figure 11: Results for Simulation 5: unbalanced treatment and linear CATE



Results

Figure 12: Results for Semi-synthetic Simulation

