Meta-learners for estimating heterogeneous treatment effects using machine learning

Künzel, Sekhon, Bickel, and Yu

Proceedings of the National Academy of Sciences - June 2017

Heterogeneous Treatment Effects

Generating model: $(Y_i(0), Y_i(1), X_i, W_i) \sim \mathcal{P}$
Observations: $\mathcal{D}_N = (Y_i(W_i), X_i, W_i)_{1 \leq i \leq N}$, $W_i \in \{0, 1\}$.

- Individual Treatment Effect: $D_i := Y_i(1) - Y_i(0)$, yet only one potential outcome is observed for each unit.
- Average Treatment Effect: ATE $:= \mathbb{E}[Y(1) - Y(0)]$.
- Conditional ATE: $\tau(x) := \mathbb{E}[Y(1) - Y(0) | X = x]$.

The responses under control/treatment will be useful:
$\mu_0(x) := \mathbb{E}[Y(0)|X = x]$ and $\mu_1(x) := \mathbb{E}[Y(1)|X = x]$.
Those are such that: $\tau(x) = \mu_1(x) - \mu_0(x)$.

Two simple existing methods for CATE estimation

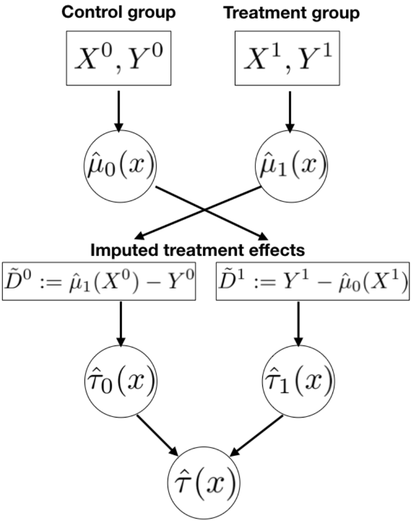
- **T-learner**: build estimators $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$, recall that $\tau(x) = \mu_1(x) - \mu_0(x)$ and use the estimator:

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

- **S-learner**: estimate the combined response function $\mu(x, w) := \mathbb{E}[Y^{obs}|X = x, W = w]$ using all observed data and all covariates including w (with no particular role) and build the estimator by shifting the value of w:

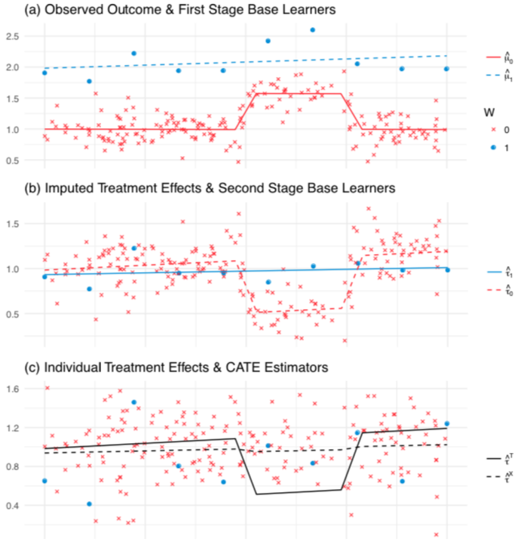
$$\hat{\tau}_S(x) := \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$$

The X-learner



1. Estimate μ_0, μ_1
2. Impute D using the other estimator
3. Estimate τ in each group
4. Combine them with a convex combination

X-learner with unbalanced design



Description of the simulations

Step 1:

$$X_i \sim \mathcal{N}(0, \Sigma) \ with \ i \in [1, d]$$

Step 2:

$$Y_i(1) = \mu_1(X_i) + \epsilon_i(1) \ with \ \epsilon_i(1) \sim \mathcal{N}(0, 1)$$
$$Y_i(0) = \mu_0(X_i) + \epsilon_i(0) \ with \ \epsilon_i(0) \sim \mathcal{N}(0, 1)$$

Step 3:

$$W_i \sim Bern(e(X_i))$$

Examples of the simulations

Simulation with "simple" τ and "complex" μ:

$\mu_0(x) = x^T\beta + 5 \times \mathbf{1}_{x_1>0.5}\ with\ \beta \in \mathbb{R}^{20}$
$\mu_1(x) = \mu_0(x) + 8 \times \mathbf{1}_{x_2<0.1}$
$\tau(x) = 8 \times \mathbf{1}_{x_2<0.1}$

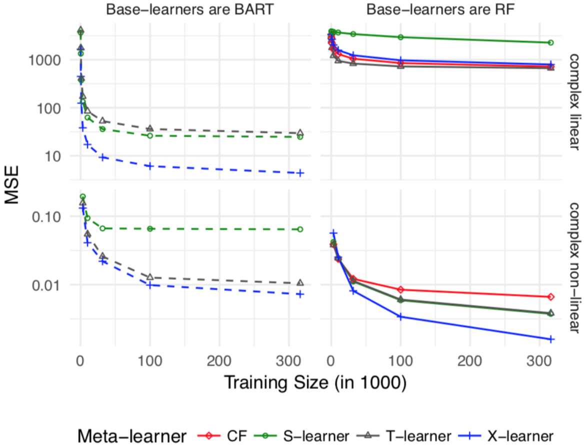
Simulation with the same effect:

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

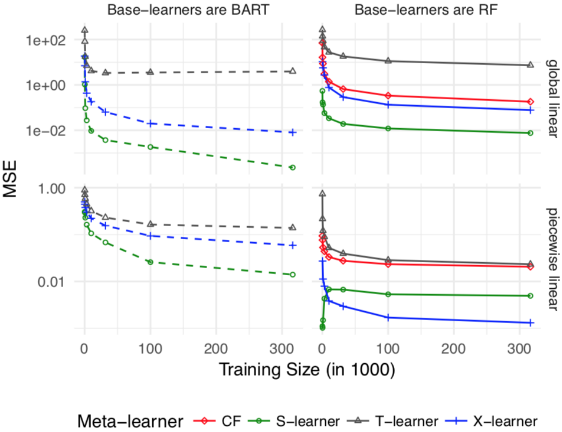
Simulation non linear:

$\mu_0(x) = \frac{1}{2}\zeta(x_1)\zeta(x_2)$
$\mu_1(x) = -\frac{1}{2}\zeta(x_1)\zeta(x_2)$
$\zeta(x) = \frac{2}{1+\exp^{-12(x-\frac{1}{2})}}$

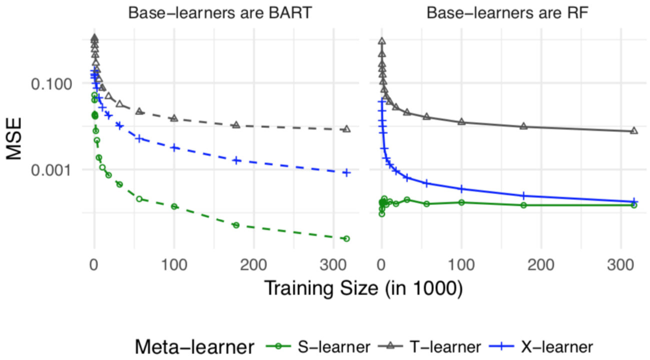
Simulation with "simple" τ and "complex" μ results



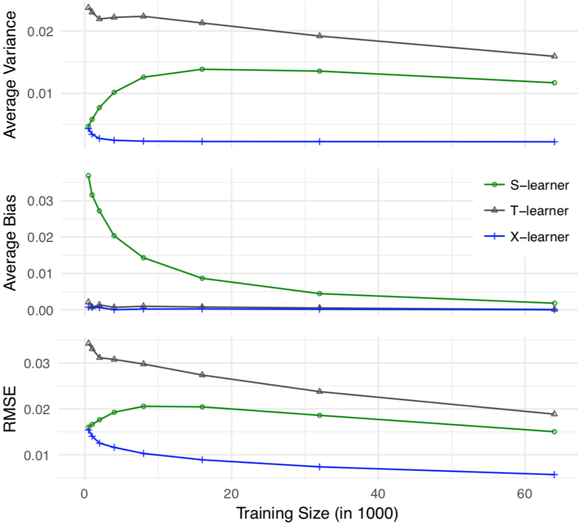
Simulation with the same effect results



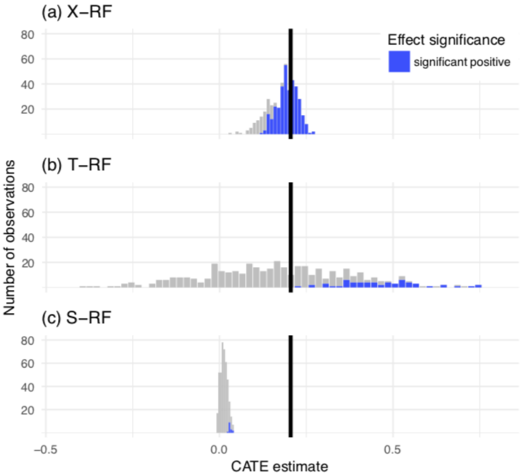
Simulation with confound results



Application : Social pressure and voter turnout



Application : Reducing transphobia



Minimax rates of estimation

- EMSE of estimator $\hat{\mu}_N$ for $\mathcal{P}$: $\mathbb{E}_{(\mathcal{D}_N, \mathcal{X})}\left[\left(\hat{\mu}_N(\mathcal{X}) - \mu(\mathcal{X})\right)^2\right]$
- Family with bounded minimax rate a: family F of distributions $\mathcal{P}$ s.t. $\min_{\hat{\mu}_N} \max_{\mathcal{P} \in F} \mathrm{EMSE}(\mathcal{P}, \hat{\mu}_N) = O(N^{-a})$

Theorem 1: Minimax rates of the T-learner

Suppose we can estimate μ_0 and μ_1 at rate a_μ. Then some T-learner can estimate τ with an EMSE of order $O(m^{-a_\mu} + n^{-a_\mu})$.

Conjecture 1: Minimax rates of the X-learner

Suppose we can estimate μ_0 and μ_1 at rate a_μ, and the imputed treatment effects at rate a_τ. Then some X-learner can estimate τ_1 with an EMSE of order $O(m^{-a_\mu} + n^{-a_\tau})$.

Two situations where the conjecture holds

Theorem 2: Minimax rates of the X-learner, linear case

Assume that:

- μ_0 can be estimated at rate a_μ
- The treatment effect is linear. This implies $a_\tau = 1$

Some X-learner can estimate τ_1 with error $O(m^{-a_\mu} + n^{-1})$

Theorem 7: Minimax rates of the X-learner, smooth case

Assume that:

- The features are $\mathcal{U}([0,1]^d)$ and the noise $\mathcal{N}(0, \sigma^2)$
- μ_0 and μ_1 are L-Lipschitz. This implies $a_\mu = a_\tau = 2/(2+d)$

Some X-learner can estimate τ_1 with error $O\left(m^{-\frac{2}{2+d}} + n^{-\frac{2}{2+d}}\right)$ and this rate is optimal for any estimator.

Pros & cons: the cons

- Unjustified assumptions to create situations where the X-learner outperforms the T-learner:
  - Simple τ but complicated μ_i, even though $\tau = \mu_1 - \mu_0$
  - τ and μ_i depending on separate feature subsets
- No rule or heuristic for choosing the weights of $\hat{\tau}_0$ and $\hat{\tau}_1$
- Base estimators (BART, RF) only tree-based, probably chosen because the authors had implemented them
- X-learner has high bias and low coverage of the bootstrap CI

Pros & cons: the pros

- X-learner can exploit class imbalance in observational data
- X-learner can exploit structure in the CATE function
- Partial theoretical justification for performance
- Empirical testing on simulated and field data