Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments

Michael Lingzhi Li

Harvard University

June 16th, 2025

NCB Conference

Joint work with Kosuke Imai (Harvard University)

Motivation

- Two methodological revolutions over the past few decades

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning

- Causal machine learning (causal ML)

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning

- Causal machine learning (causal ML)
  1. estimation of heterogeneous treatment effects

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning

- Causal machine learning (causal ML)
  1. estimation of heterogeneous treatment effects
  2. development of individualized treatment rules

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning
- Causal machine learning (causal ML)
  1. estimation of heterogeneous treatment effects
  2. development of individualized treatment rules
- Experimental evaluation of causal ML

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning
- Causal machine learning (causal ML)
  1. estimation of heterogeneous treatment effects
  2. development of individualized treatment rules
- Experimental evaluation of causal ML
  1. ML algorithms may not work well in practice

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning
- Causal machine learning (causal ML)
  1. estimation of heterogeneous treatment effects
  2. development of individualized treatment rules
- Experimental evaluation of causal ML
  1. ML algorithms may not work well in practice
  2. assumption-free uncertainty quantification is essential

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning
- Causal machine learning (causal ML)
  1. estimation of heterogeneous treatment effects
  2. development of individualized treatment rules
- Experimental evaluation of causal ML
  1. ML algorithms may not work well in practice
  2. assumption-free uncertainty quantification is essential
- I will show how to experimentally evaluate heterogeneous treatment effects (HTEs) discovered by generic causal ML

Motivation

- Two methodological revolutions over the past few decades
  1. randomized experiments (field/lab/survey)
  2. machine learning
- Causal machine learning (causal ML)
  1. estimation of heterogeneous treatment effects
  2. development of individualized treatment rules
- Experimental evaluation of causal ML
  1. ML algorithms may not work well in practice
  2. assumption-free uncertainty quantification is essential
- I will show how to experimentally evaluate heterogeneous treatment effects (HTEs) discovered by generic causal ML
- An important step before trusting and utilizing estimated HTEs

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE
     - evaluate the group average treatment effect (GATES), for example, among those who are predicted by the ML algorithm to benefit from (or be harmed by) treatment the most

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE
     - evaluate the group average treatment effect (GATES), for example, among those who are predicted by the ML algorithm to benefit from (or be harmed by) treatment the most
2. Scenario II: Estimate and evaluate with the same experimental dataset

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE
     - evaluate the group average treatment effect (GATES), for example, among those who are predicted by the ML algorithm to benefit from (or be harmed by) treatment the most
2. Scenario II: Estimate and evaluate with the same experimental dataset
   - choose an ML algorithm

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE
     - evaluate the group average treatment effect (GATES), for example, among those who are predicted by the ML algorithm to benefit from (or be harmed by) treatment the most
2. Scenario II: Estimate and evaluate with the same experimental dataset
   - choose an ML algorithm
   - randomly split an experimental dataset into training and evaluation datasets

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE
     - evaluate the group average treatment effect (GATES), for example, among those who are predicted by the ML algorithm to benefit from (or be harmed by) treatment the most
2. Scenario II: Estimate and evaluate with the same experimental dataset
   - choose an ML algorithm
   - randomly split an experimental dataset into training and evaluation datasets
   - estimate CATE using the training dataset

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE
     - evaluate the group average treatment effect (GATES), for example, among those who are predicted by the ML algorithm to benefit from (or be harmed by) treatment the most
2. Scenario II: Estimate and evaluate with the same experimental dataset
   - choose an ML algorithm
   - randomly split an experimental dataset into training and evaluation datasets
   - estimate CATE using the training dataset
   - use the evaluation dataset and estimate the GATES

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE
     - evaluate the group average treatment effect (GATES), for example, among those who are predicted by the ML algorithm to benefit from (or be harmed by) treatment the most
2. Scenario II: Estimate and evaluate with the same experimental dataset
   - choose an ML algorithm
   - randomly split an experimental dataset into training and evaluation datasets
   - estimate CATE using the training dataset
   - use the evaluation dataset and estimate the GATES
   - flip the training and evaluation datasets and repeat

Overview of the Proposed Methodology

1. Scenario I: Estimate and evaluate with separate datasets
   - choose an ML algorithm
   - estimate the conditional average treatment effect (CATE) using an external (possibly observational) dataset and treat it as fixed
   - with an experimental dataset
     - sort observations based on the estimated CATE
     - evaluate the group average treatment effect (GATES), for example, among those who are predicted by the ML algorithm to benefit from (or be harmed by) treatment the most
2. Scenario II: Estimate and evaluate with the same experimental dataset
   - choose an ML algorithm
   - randomly split an experimental dataset into training and evaluation datasets
   - estimate CATE using the training dataset
   - use the evaluation dataset and estimate the GATES
   - flip the training and evaluation datasets and repeat
   - average the results and account for uncertainty due to random splits

Setup

- Notation:
  - n experimental units
  - $T_i \in \{0, 1\}$: binary treatment
  - $Y_i(t)$ where $t \in \{0, 1\}$: potential outcomes
  - $Y_i = Y_i(T_i)$: observed outcome
  - X_i: moderator of interest

Setup

- Notation:
  - n experimental units
  - $T_i \in \{0, 1\}$: binary treatment
  - $Y_i(t)$ where $t \in \{0, 1\}$: potential outcomes
  - $Y_i = Y_i(T_i)$: observed outcome
  - X_i: moderator of interest

- Assumptions:
  1. no interference between units:

$$Y_i(T_1 = t_1, \ldots, T_n = t_n) = Y_i(T_i = t_i)$$

  2. randomization of treatment assignment:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$$

  3. random sampling of units:

$$\{Y_i(1), Y_i(0)\} \overset{\text{i.i.d.}}{\sim} \mathcal{P}$$

Exploration of Heterogeneous Treatment Effects

- Two commonly used treatment prioritization scores

Exploration of Heterogeneous Treatment Effects

- Two commonly used treatment prioritization scores
  1. Conditional average treatment effect (CATE):

$$\tau(x) = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)$$

Exploration of Heterogeneous Treatment Effects

- Two commonly used treatment prioritization scores
  1. Conditional average treatment effect (CATE):

$$\tau(x) \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)$$

  2. Baseline risk:

$$\lambda(x) \;=\; \mathbb{E}(Y_i(0) \mid X_i = x)$$

Exploration of Heterogeneous Treatment Effects

- Two commonly used treatment prioritization scores
  1. Conditional average treatment effect (CATE):

$$\tau(x) \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)$$

  2. Baseline risk:

$$\lambda(x) \;=\; \mathbb{E}(Y_i(0) \mid X_i = x)$$

- Estimate a score with ML algorithm using an external dataset

$$f : \mathcal{X} \longrightarrow \mathcal{S} \subset \mathbb{R}$$

Exploration of Heterogeneous Treatment Effects

- Two commonly used treatment prioritization scores
  1. Conditional average treatment effect (CATE):

$$\tau(\mathrm{x}) \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathrm{X}_i = \mathrm{x})$$

  2. Baseline risk:

$$\lambda(\mathrm{x}) \;=\; \mathbb{E}(Y_i(0) \mid \mathrm{X}_i = \mathrm{x})$$

- Estimate a score with ML algorithm using an external dataset

$$f : \mathcal{X} \longrightarrow \mathcal{S} \subset \mathbb{R}$$

- Group Average Treatment Effect (GATES; Chernozhukov et al. 2019)

$$\tau_k \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid p_{k-1} \leq S_i = f(X_i) < p_k)$$

for $k = 1, 2, \ldots, K$ where p_k is a quantile cutoff ($p_0 = -\infty$, $p_K = \infty$)

Statistical Inference for GATES

- How can we make valid statistical inference for GATES without assuming that the scores are correctly estimated by ML algorithm?

Statistical Inference for GATES

- How can we make valid statistical inference for GATES without assuming that the scores are correctly estimated by ML algorithm?

- A natural difference-in-means estimator for GATES:

$$\hat{\tau}_k \;=\; \frac{K}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(X_i) - \frac{K}{n_0}\sum_{i=1}^{n} Y_i (1 - T_i)\hat{f}_k(X_i),$$

where $\hat{f}_k(\mathrm{X}_i) \;=\; 1\{S_i \geq \hat{p}_k(s)\} - 1\{S_i \geq \hat{p}_{k-1}\}$ is the group indicator

Statistical Inference for GATES

- How can we make valid statistical inference for GATES without assuming that the scores are correctly estimated by ML algorithm?
- A natural difference-in-means estimator for GATES:

$$\hat{\tau}_k \;=\; \frac{K}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(X_i) - \frac{K}{n_0}\sum_{i=1}^{n} Y_i (1-T_i)\hat{f}_k(X_i),$$

  where $\hat{f}_k(\mathrm{X}_i) \;=\; 1\{S_i \geq \hat{p}_k(s)\} - 1\{S_i \geq \hat{p}_{k-1}\}$ is the group indicator

- Bias bound and exact variance are derived, accounting for the estimation uncertainty of quantile cutoffs

Statistical Inference for GATES

- How can we make valid statistical inference for GATES without assuming that the scores are correctly estimated by ML algorithm?
- A natural difference-in-means estimator for GATES:

$$\hat{\tau}_k \;=\; \frac{K}{n_1}\sum_{i=1}^{n} Y_i T_i \hat{f}_k(X_i) - \frac{K}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\hat{f}_k(X_i),$$

  where $\hat{f}_k(\mathrm{X}_i) \;=\; 1\{S_i \geq \hat{p}_k(s)\} - 1\{S_i \geq \hat{p}_{k-1}\}$ is the group indicator

- Bias bound and exact variance are derived, accounting for the estimation uncertainty of quantile cutoffs
- Under mild regularity conditions (e.g., continuity of CATE at thresholds), the distribution of $\hat{\tau}_k$ is asymptotically normal

Statistical Hypothesis Tests for Subgroups

1. Nonparametric test of treatment effect homogeneity:

Statistical Hypothesis Tests for Subgroups

1. Nonparametric test of treatment effect homogeneity:
   - Null hypothesis:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_K.$$

Statistical Hypothesis Tests for Subgroups

1. Nonparametric test of treatment effect homogeneity:
   - Null hypothesis:
   $$H_0 : \tau_1 = \tau_2 = \cdots = \tau_K.$$
   - Test statistic:
   $$\hat{\boldsymbol{\tau}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}} \xrightarrow{d} \chi^2_K$$
   where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1 - \hat{\tau}, \cdots, \hat{\tau}_K - \hat{\tau})^\top$

Statistical Hypothesis Tests for Subgroups

1. Nonparametric test of treatment effect homogeneity:
   - Null hypothesis:
   $$H_0 : \tau_1 = \tau_2 = \cdots = \tau_K.$$
   - Test statistic:
   $$\hat{\boldsymbol{\tau}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}} \xrightarrow{d} \chi^2_K$$
   where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1 - \hat{\tau}, \cdots, \hat{\tau}_K - \hat{\tau})^\top$
2. Nonparametric test of rank-consistent treatment effect heterogeneity:

Statistical Hypothesis Tests for Subgroups

1. Nonparametric test of treatment effect homogeneity:
   - Null hypothesis:
   $$H_0 : \tau_1 = \tau_2 = \cdots = \tau_K.$$
   - Test statistic:
   $$\hat{\boldsymbol{\tau}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}} \xrightarrow{d} \chi^2_K$$
   where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1 - \hat{\tau}, \cdots, \hat{\tau}_K - \hat{\tau})^\top$
2. Nonparametric test of rank-consistent treatment effect heterogeneity:
   - Null hypothesis:
   $$H_0^* : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_K.$$

Statistical Hypothesis Tests for Subgroups

1. Nonparametric test of treatment effect homogeneity:
   - Null hypothesis:
   $$H_0 : \tau_1 = \tau_2 = \cdots = \tau_K.$$
   - Test statistic:
   $$\hat{\boldsymbol{\tau}}^\top \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\tau}} \xrightarrow{d} \chi^2_K$$
   where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1 - \hat{\tau}, \cdots, \hat{\tau}_K - \hat{\tau})^\top$
2. Nonparametric test of rank-consistent treatment effect heterogeneity:
   - Null hypothesis:
   $$H_0^* : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_K.$$
   - Test statistic:
   $$\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})\right)^\top \boldsymbol{\Sigma}^{-1} \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})\right) \xrightarrow{d} \bar{\chi}^2_K.$$
   where $\boldsymbol{\mu}^*(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{\mu}} \|\boldsymbol{\mu} - \boldsymbol{x}\|_2^2$ subject to $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_K$.

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:
  1. randomly split the data into L folds: $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:
  1. randomly split the data into L folds: $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$
  2. estimate the score using $L - 1$ folds: $\hat{f}_{-\ell}$

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:
  1. randomly split the data into L folds: $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$
  2. estimate the score using $L-1$ folds: $\hat{f}_{-\ell}$
  3. estimate GATES with the hold-out set: $\hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:
  1. randomly split the data into L folds: $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$
  2. estimate the score using $L-1$ folds: $\hat{f}_{-\ell}$
  3. estimate GATES with the hold-out set: $\hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$
  4. repeat the process for each ℓ and average

$$\hat{\tau}_k(F; n-m) \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$$

where $F : \mathcal{Z} \longrightarrow \mathcal{F}$ is a generic but stable ML algorithm with $\mathcal{Z}_{\text{train}} \in \mathcal{Z}$ and $\hat{f}_{\mathcal{Z}_{\text{train}}} = F(\mathcal{Z}_{\text{train}}) \in \mathcal{F}$

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:
  1. randomly split the data into L folds: $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$
  2. estimate the score using $L-1$ folds: $\hat{f}_{-\ell}$
  3. estimate GATES with the hold-out set: $\hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$
  4. repeat the process for each ℓ and average

$$\hat{\tau}_k(F; n-m) \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$$

  where $F : \mathcal{Z} \longrightarrow \mathcal{F}$ is a generic but stable ML algorithm with $\mathcal{Z}_{\text{train}} \in \mathcal{Z}$ and $\hat{f}_{\mathcal{Z}_{\text{train}}} = F(\mathcal{Z}_{\text{train}}) \in \mathcal{F}$

- Estimand: average performance of F

$$\begin{aligned} & \tau_k(F; n-m) \\ = \;& \mathbb{E}_{\mathcal{Z}_{\text{train}}^{n-m}}[\mathbb{E}\{Y_i(1) - Y_i(0) \mid p_{k-1}(\hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}}) \leq \hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}}(\mathrm{X}_i) < p_k(\hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}})\}]. \end{aligned}$$

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:
  1. randomly split the data into L folds: $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$
  2. estimate the score using $L-1$ folds: $\hat{f}_{-\ell}$
  3. estimate GATES with the hold-out set: $\hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$
  4. repeat the process for each ℓ and average

$$\hat{\tau}_k(F; n-m) \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$$

  where $F : \mathcal{Z} \longrightarrow \mathcal{F}$ is a generic but stable ML algorithm with $\mathcal{Z}_{\text{train}} \in \mathcal{Z}$ and $\hat{f}_{\mathcal{Z}_{\text{train}}} = F(\mathcal{Z}_{\text{train}}) \in \mathcal{F}$

- Estimand: average performance of F

$$\begin{aligned} & \tau_k(F; n-m) \\ = \;& \mathbb{E}_{\mathcal{Z}_{\text{train}}^{n-m}}[\mathbb{E}\{Y_i(1) - Y_i(0) \mid p_{k-1}(\hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}}) \leq \hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}}(\mathrm{X}_i) < p_k(\hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}})\}]. \end{aligned}$$

- Unbiasedness: $\mathbb{E}(\hat{\tau}_k(F; n-m)) = \tau_k(F; n-m)$

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:
  1. randomly split the data into L folds: $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$
  2. estimate the score using $L-1$ folds: $\hat{f}_{-\ell}$
  3. estimate GATES with the hold-out set: $\hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$
  4. repeat the process for each ℓ and average

$$\hat{\tau}_k(F; n-m) \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \hat{\tau}_k^{(\ell)}(\hat{f}_{-\ell})$$

  where $F : \mathcal{Z} \longrightarrow \mathcal{F}$ is a generic but stable ML algorithm with $\mathcal{Z}_{\text{train}} \in \mathcal{Z}$ and $\hat{f}_{\mathcal{Z}_{\text{train}}} = F(\mathcal{Z}_{\text{train}}) \in \mathcal{F}$

- Estimand: average performance of F

$$\begin{aligned} & \tau_k(F; n-m) \\ = \;& \mathbb{E}_{\mathcal{Z}_{\text{train}}^{n-m}}[\mathbb{E}\{Y_i(1) - Y_i(0) \mid p_{k-1}(\hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}}) \leq \hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}}(\mathrm{X}_i) < p_k(\hat{f}_{\mathcal{Z}_{\text{train}}^{n-m}})\}]. \end{aligned}$$

- Unbiasedness: $\mathbb{E}(\hat{\tau}_k(F; n-m)) = \tau_k(F; n-m)$
- Finite-sample (conservative) variance estimator

Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition

Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition
  - 58 covariates (3 categorical, 5 binary, 27 counts, 13 continuous)

Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition
  - 58 covariates (3 categorical, 5 binary, 27 counts, 13 continuous)
  - sample size: $n = 4802$

Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition
  - 58 covariates (3 categorical, 5 binary, 27 counts, 13 continuous)
  - sample size: $n = 4802$
  - use empirical distribution of X_i as true distribution

Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition
  - 58 covariates (3 categorical, 5 binary, 27 counts, 13 continuous)
  - sample size: $n = 4802$
  - use empirical distribution of X_i as true distribution

- Machine learning algorithms

Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition
  - 58 covariates (3 categorical, 5 binary, 27 counts, 13 continuous)
  - sample size: $n = 4802$
  - use empirical distribution of X_i as true distribution

- Machine learning algorithms
  - Causal forest and Lasso

Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition
  - 58 covariates (3 categorical, 5 binary, 27 counts, 13 continuous)
  - sample size: $n = 4802$
  - use empirical distribution of X_i as true distribution

- Machine learning algorithms
  - Causal forest and Lasso
  - $L = 5$ and also use 5-fold cross validation for tuning

Simulation Study

- A highly nonlinear specification from the 2016 ACIC competition
  - 58 covariates (3 categorical, 5 binary, 27 counts, 13 continuous)
  - sample size: $n = 4802$
  - use empirical distribution of X_i as true distribution
- Machine learning algorithms
  - Causal forest and Lasso
  - $L = 5$ and also use 5-fold cross validation for tuning
- Fixed score (see the paper) and estimated one with cross-fitting

Simulation Results: Bias and Coverage

	$n = 100$			$n = 500$			$n = 2500$		
	bias	s.d.	coverage	bias	s.d.	coverage	bias	s.d.	coverage
Causal Forest									
$\hat{\tau}_1$	−0.05	2.97	94.0%	−0.01	1.57	95.6%	−0.01	0.59	97.7%
$\hat{\tau}_2$	−0.06	2.58	95.9	−0.04	1.08	98.2	0.01	0.54	98.6
$\hat{\tau}_3$	−0.01	2.56	96.7	−0.05	1.06	97.7	0.02	0.47	98.1
$\hat{\tau}_4$	−0.12	2.87	97.4	0.05	1.15	97.9	−0.01	0.51	98.6
$\hat{\tau}_5$	0.14	3.45	94.1	0.00	1.62	96.0	−0.01	0.62	98.3
LASSO									
$\hat{\tau}_1$	−0.13	3.20	97.6%	−0.03	1.49	96.0%	−0.00	0.67	96.0%
$\hat{\tau}_2$	0.04	2.28	97.5	−0.07	1.03	97.9	−0.02	0.59	98.9
$\hat{\tau}_3$	−0.13	2.35	96.6	−0.02	1.00	97.9	0.04	0.49	97.5
$\hat{\tau}_4$	−0.00	2.54	96.8	0.04	1.17	96.8	0.03	0.64	97.2
$\hat{\tau}_5$	0.11	3.62	96.2	0.05	1.81	95.0	0.02	0.70	95.3

- Reduction in standard errors compared with fixed F of the same evaluation size is more than 50% in some cases

Simulation Results: Size and Power of Tests

	$n = 100$		$n = 500$		$n = 2500$	
	rejection rate	median p-value	rejection rate	median p-value	rejection rate	median p-value
Causal Forest						
Homogeneity	1.4%	0.79	4.6%	0.71	51.4%	0.04
Rank-consistency	1.4%	0.70	0.8%	0.85	0.0%	0.98
LASSO						
Homogeneity	0.6%	0.88	1.8%	0.85	9.0%	0.66
Rank-consistency	1.0%	0.72	0.6%	0.77	0.2%	0.89

- Heterogeneous but rank-consistent effects

Simulation Results: Size and Power of Tests

	$n = 100$		$n = 500$		$n = 2500$	
	rejection rate	median p-value	rejection rate	median p-value	rejection rate	median p-value
Causal Forest						
Homogeneity	1.4%	0.79	4.6%	0.71	51.4%	0.04
Rank-consistency	1.4%	0.70	0.8%	0.85	0.0%	0.98
LASSO						
Homogeneity	0.6%	0.88	1.8%	0.85	9.0%	0.66
Rank-consistency	1.0%	0.72	0.6%	0.77	0.2%	0.89

- Heterogeneous but rank-consistent effects
- More conservative and lower power than fixed case

Simulation Results: Size and Power of Tests

	$n = 100$		$n = 500$		$n = 2500$	
	rejection rate	median p-value	rejection rate	median p-value	rejection rate	median p-value
Causal Forest						
Homogeneity	1.4%	0.79	4.6%	0.71	51.4%	0.04
Rank-consistency	1.4%	0.70	0.8%	0.85	0.0%	0.98
LASSO						
Homogeneity	0.6%	0.88	1.8%	0.85	9.0%	0.66
Rank-consistency	1.0%	0.72	0.6%	0.77	0.2%	0.89

- Heterogeneous but rank-consistent effects
- More conservative and lower power than fixed case
- When sample size is large, cross-fitting yields higher power

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys
- Data

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys
- Data
  - Sample size: $n_1 = 302$, $n_0 = 305$ (total $n = 607$)

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys
- Data
  - Sample size: $n_1 = 302$, $n_0 = 305$ (total $n = 607$)
  - Outcome: recurrence of UTIs at 2-year follow-up; we use $-Y$ so positive effect = fewer UTIs

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys
- Data
  - Sample size: $n_1 = 302$, $n_0 = 305$ (total $n = 607$)
  - Outcome: recurrence of UTIs at 2-year follow-up; we use $-Y$ so positive effect = fewer UTIs
  - 7 pre-treatment covariates: demographics, tests, and prior conditions

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys
- Data
  - Sample size: $n_1 = 302$, $n_0 = 305$ (total $n = 607$)
  - Outcome: recurrence of UTIs at 2-year follow-up; we use $-Y$ so positive effect = fewer UTIs
  - 7 pre-treatment covariates: demographics, tests, and prior conditions
- Setup

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys
- Data
  - Sample size: $n_1 = 302$, $n_0 = 305$ (total $n = 607$)
  - Outcome: recurrence of UTIs at 2-year follow-up; we use $-Y$ so positive effect = fewer UTIs
  - 7 pre-treatment covariates: demographics, tests, and prior conditions
- Setup
  - ML algorithms: Causal Forest, BART, and LASSO

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys
- Data
  - Sample size: $n_1 = 302$, $n_0 = 305$ (total $n = 607$)
  - Outcome: recurrence of UTIs at 2-year follow-up; we use $-Y$ so positive effect = fewer UTIs
  - 7 pre-treatment covariates: demographics, tests, and prior conditions
- Setup
  - ML algorithms: Causal Forest, BART, and LASSO
  - Sample-splitting: 67% training, 33% evaluation

Empirical Application

- Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) Trial
- Double-blind RCT evaluating whether daily antimicrobial prophylaxis prevents recurrence of UTIs
- UTIs: urine flows backward from bladder to ureters/kidneys
- Data
  - Sample size: $n_1 = 302$, $n_0 = 305$ (total $n = 607$)
  - Outcome: recurrence of UTIs at 2-year follow-up; we use $-Y$ so positive effect = fewer UTIs
  - 7 pre-treatment covariates: demographics, tests, and prior conditions
- Setup
  - ML algorithms: Causal Forest, BART, and LASSO
  - Sample-splitting: 67% training, 33% evaluation
  - Cross-fitting: 3 folds; tuning via 5-fold CV within training sets

GATES Estimates (in % Decrease in UTI Recurrence)

	$\hat{\tau}_1$ (%)	$\hat{\tau}_2$ (%)	$\hat{\tau}_3$ (%)	$\hat{\tau}_4$ (%)	$\hat{\tau}_5$ (%)
Sample-splitting					
Causal Forest	-0.1	-0.0	9.9	9.9	9.9
	$[-19.7, 19.5]$	$[-14.0, 13.9]$	$[-9.3, 29.2]$	$[-13.8, 33.6]$	$[-13.7, 33.5]$
BART	-5.2	20.0	5.0	-5.1	14.8
	$[-26.9, 16.6]$	$[1.4, 38.6]$	$[-4.7, 14.7]$	$[-22.1, 11.9]$	$[-13.2, 42.9]$
LASSO	0.0	9.9	-0.1	9.9	9.9
	$[-14.0, 13.9]$	$[-9.4, 29.3]$	$[-19.7, 19.5]$	$[-13.9, 33.7]$	$[-13.0, 32.8]$
Cross-fitting					
Causal Forest	3.2	-5.1	-3.4	14.7	26.2
	$[-8.7, 15.1]$	$[-26.5, 16.4]$	$[-22.2, 15.4]$	$[0.1, 29.2]$	$[7.2, 45.1]$
BART	-1.8	-5.0	11.4	9.8	21.2
	$[-15.5, 11.9]$	$[-13.4, 3.5]$	$[-10.3, 33.0]$	$[-5.7, 25.2]$	$[3.8, 38.6]$
LASSO	-1.7	-1.5	3.2	11.4	21.2
	$[-13.9, 10.4]$	$[-23.4, 26.4]$	$[-11.6, 17.9]$	$[-4.8, 27.7]$	$[-4.6, 47.0]$

- Stat. significant effects found only with cross-fitting

GATES Estimates (in % Decrease in UTI Recurrence)

	$\hat{\tau}_1$ (%)	$\hat{\tau}_2$ (%)	$\hat{\tau}_3$ (%)	$\hat{\tau}_4$ (%)	$\hat{\tau}_5$ (%)
Sample-splitting					
Causal Forest	-0.1	-0.0	9.9	9.9	9.9
	$[-19.7, 19.5]$	$[-14.0, 13.9]$	$[-9.3, 29.2]$	$[-13.8, 33.6]$	$[-13.7, 33.5]$
BART	-5.2	20.0	5.0	-5.1	14.8
	$[-26.9, 16.6]$	$[1.4, 38.6]$	$[-4.7, 14.7]$	$[-22.1, 11.9]$	$[-13.2, 42.9]$
LASSO	0.0	9.9	-0.1	9.9	9.9
	$[-14.0, 13.9]$	$[-9.4, 29.3]$	$[-19.7, 19.5]$	$[-13.9, 33.7]$	$[-13.0, 32.8]$
Cross-fitting					
Causal Forest	3.2	-5.1	-3.4	14.7	26.2
	$[-8.7, 15.1]$	$[-26.5, 16.4]$	$[-22.2, 15.4]$	$[0.1, 29.2]$	$[7.2, 45.1]$
BART	-1.8	-5.0	11.4	9.8	21.2
	$[-15.5, 11.9]$	$[-13.4, 3.5]$	$[-10.3, 33.0]$	$[-5.7, 25.2]$	$[3.8, 38.6]$
LASSO	-1.7	-1.5	3.2	11.4	21.2
	$[-13.9, 10.4]$	$[-23.4, 26.4]$	$[-11.6, 17.9]$	$[-4.8, 27.7]$	$[-4.6, 47.0]$

- Stat. significant effects found only with cross-fitting
- Causal Forest: 40% of patients benefit; BART: 20%

GATES Estimates (in % Decrease in UTI Recurrence)

	$\hat{\tau}_1$ (%)	$\hat{\tau}_2$ (%)	$\hat{\tau}_3$ (%)	$\hat{\tau}_4$ (%)	$\hat{\tau}_5$ (%)
Sample-splitting					
Causal Forest	-0.1	-0.0	9.9	9.9	9.9
	$[-19.7, 19.5]$	$[-14.0, 13.9]$	$[-9.3, 29.2]$	$[-13.8, 33.6]$	$[-13.7, 33.5]$
BART	-5.2	20.0	5.0	-5.1	14.8
	$[-26.9, 16.6]$	$[1.4, 38.6]$	$[-4.7, 14.7]$	$[-22.1, 11.9]$	$[-13.2, 42.9]$
LASSO	0.0	9.9	-0.1	9.9	9.9
	$[-14.0, 13.9]$	$[-9.4, 29.3]$	$[-19.7, 19.5]$	$[-13.9, 33.7]$	$[-13.0, 32.8]$
Cross-fitting					
Causal Forest	3.2	-5.1	-3.4	14.7	26.2
	$[-8.7, 15.1]$	$[-26.5, 16.4]$	$[-22.2, 15.4]$	$[0.1, 29.2]$	$[7.2, 45.1]$
BART	-1.8	-5.0	11.4	9.8	21.2
	$[-15.5, 11.9]$	$[-13.4, 3.5]$	$[-10.3, 33.0]$	$[-5.7, 25.2]$	$[3.8, 38.6]$
LASSO	-1.7	-1.5	3.2	11.4	21.2
	$[-13.9, 10.4]$	$[-23.4, 26.4]$	$[-11.6, 17.9]$	$[-4.8, 27.7]$	$[-4.6, 47.0]$

- Stat. significant effects found only with cross-fitting
- Causal Forest: 40% of patients benefit; BART: 20%
- LASSO fails to identify any group with significant benefit

Results of Hypothesis Tests

	Causal Forest		BART		LASSO	
	stat	p-value	stat	p-value	stat	p-value
Sample-splitting						
Homogeneous Treatment Effects	1.45	0.918	5.16	0.397	1.45	0.918
Rank-consistent Treatment Effects	0.00	0.990	3.78	0.222	0.511	0.845
Cross-fitting						
Homogeneous Treatment Effects	12.5	0.029	13.7	0.020	6.38	0.271
Rank-consistent Treatment Effects	0.97	0.727	0.17	0.920	0.01	0.993

- Causal Forest and BART reject homogeneity under cross-fitting

Results of Hypothesis Tests

	Causal Forest		BART		LASSO	
	stat	p-value	stat	p-value	stat	p-value
Sample-splitting						
Homogeneous Treatment Effects	1.45	0.918	5.16	0.397	1.45	0.918
Rank-consistent Treatment Effects	0.00	0.990	3.78	0.222	0.511	0.845
Cross-fitting						
Homogeneous Treatment Effects	12.5	0.029	13.7	0.020	6.38	0.271
Rank-consistent Treatment Effects	0.97	0.727	0.17	0.920	0.01	0.993

- Causal Forest and BART reject homogeneity under cross-fitting
- No algorithm rejects rank-consistency

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification
  - experimental evaluation of HTEs and ITRs

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification
  - experimental evaluation of HTEs and ITRs
  - no modeling assumption

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification
  - experimental evaluation of HTEs and ITRs
  - no modeling assumption
  - no resampling (computationally efficient)

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification
  - experimental evaluation of HTEs and ITRs
  - no modeling assumption
  - no resampling (computationally efficient)
  - applicable to any complex causal ML algorithms

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification
  - experimental evaluation of HTEs and ITRs
  - no modeling assumption
  - no resampling (computationally efficient)
  - applicable to any complex causal ML algorithms
  - good small sample performance

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification
  - experimental evaluation of HTEs and ITRs
  - no modeling assumption
  - no resampling (computationally efficient)
  - applicable to any complex causal ML algorithms
  - good small sample performance
- Open source software: evalITR: Evaluating Individualized Treatment Rules at CRAN `https://CRAN.R-project.org/package=evalITR`

Concluding Remarks

- Causal machine learning (ML) is rapidly becoming popular
  - estimation of heterogeneous treatment effects (HTEs)
  - development of individualized treatment rules (ITRs)
- Safe deployment of causal ML requires uncertainty quantification
  - experimental evaluation of HTEs and ITRs
  - no modeling assumption
  - no resampling (computationally efficient)
  - applicable to any complex causal ML algorithms
  - good small sample performance
- Open source software: evalITR: Evaluating Individualized Treatment Rules at CRAN `https://CRAN.R-project.org/package=evalITR`
- More information:
  `https://www.michaellz.com/machine-learning-inference`