Statistical Performance Guarantee for Subgroup Identification with Generic Machine Learning

Kosuke Imai

Harvard University

September 16, 2024

The 4th Penn Conference on
Big Data in Biomedical and Population Health Sciences
University of Pennsylvania

Joint work with Michael Lingzhi Li (Harvard Business School)

Motivation

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

- Experimental evaluation of causal ML
  1. causal ML algorithms may not work well in practice
  2. need for assumption-lean evaluation with uncertainty quantification

- Today, I will show how to experimentally evaluate:
  1. heterogeneous treatment effects discovered by causal ML

     "Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments." *J. of Business & Econ. Stat.*

  2. subgroup of exceptional responders identified by causal ML

     "Statistical Performance Guarantee for Subgroup Identification with Generic Machine Learning." `https://arxiv.org/abs/2310.07973`

Standard Experimental Setup

- We will use experimental data to evaluate causal ML
- Notation: n experimental units
  1. $T_i \in \{0, 1\}$: binary treatment
  2. X_i: pre-treatment covariates
  3. $Y_i(t)$ where $t \in \{0, 1\}$: potential outcomes
  4. $Y_i = Y_i(T_i)$: observed outcome
- 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: $\{Y_i(1), Y_i(0)\} \overset{\text{i.i.d.}}{\sim} \mathcal{P}$

Evaluation of Heterogeneous Treatment Effects

- How can we statistically evaluate heterogeneous treatment effects discovered by a generic ML algorithm?
- Conditional Average Treatment Effect (CATE):

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

- CATE estimation based on ML algorithm

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

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

$$\tau_k \;=\; \mathbb{E}(Y_i(1) - Y_i(0) \mid c_{k-1} \leq s(X_i) < c_k)$$

for $k = 1, 2, \ldots, K$ where c_k is a *quantile cutoff* ($c_0 = -\infty,\ c_K = \infty$)

GATES Estimation

- A natural GATES estimator:

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

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

- Bias is small: finite-sample bound is derived
- Variance:

$$\mathbb{V}(\hat{\tau}_k) \;=\; K^2 \left[\frac{\mathbb{V}(\hat{f}_k(X_i)Y_i(1))}{n_1} + \frac{\mathbb{V}(\hat{f}_k(X_i)Y_i(0))}{n_0} + \underbrace{Cov(\hat{f}_k(\mathrm{X}_i)\tau_i, \hat{f}_k(\mathrm{X}_j)\tau_j)}_{\mathrm{Corr}(\hat{f}_k(X_i),\hat{f}_k(X_j))\neq 0} \right]$$

- Asymptotic normality

Estimation and Evaluation Using the Same Data

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

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

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

- Estimand: average performance of S

$$\tau_k(S) = \mathbb{E}_{\mathcal{Z}_{\text{train}}}[\mathbb{E}\{Y_i(1) - Y_i(0) \mid c_{k-1}(\hat{f}_{\mathcal{Z}_{\text{train}}}) \leq \hat{f}_{\mathcal{Z}_{\text{train}}}(\mathrm{X}_i) < c_k(\hat{s}_{\mathcal{Z}_{\text{train}}})\}]$$

- Inference without resampling

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

Evaluation Bias and Coverage under Cross-fitting

	$n = 100$			$n = 500$			$n = 2500$		
	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 S of the same evaluation size (see the paper) is more than 50% in some cases

Empirical Application

- National Supported Work Demonstration Program (LaLonde 1986)
- Temporary employment program to help disadvantaged workers by giving them a guaranteed job for 9 to 18 months

- Data
  - sample size: $n_1 = 297$ and $n_0 = 425$
  - outcome: annualized earnings in 1978 (36 months after the program)
  - 7 pre-treatment covariates: demographics and prior earnings

- Setup
  - ML algorithms: BART, Causal Forest, and LASSO
  - Sample-splitting: 2/3 of the data as training data
  - Cross-fitting: 3 folds

GATES Estimates (in 1,000 US Dollars)

	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$	$\hat{\tau}_5$
Sample-splitting					
BART	2.90	−0.73	−0.02	3.25	2.57
	[−2.25, 8.06]	[−5.05, 3.58]	[−3.47, 3.43]	[−1.53, 8.03]	[−3.82, 8.97]
Causal Forest	3.40	0.13	−0.85	−1.91	7.21
	[−1.29, 3.40]	[−5.37, 5.63]	[−5.22, 3.52]	[−5.16, 1.34]	[1.22, 13.19]
LASSO	1.86	2.62	−2.07	1.39	4.17
	[−3.59, 7.30]	[−1.69, 6.93]	[−5.39, 1.26]	[−2.95, 5.73]	[−2.30, 10.65]
Cross-fitting					
BART	0.40	−0.15	−0.40	2.52	2.19
	[−3.79, 4.59]	[−2.54, 2.23]	[−3.37, 2.56]	[−0.99, 6.03]	[−0.73, 5.11]
Causal Forest	−3.72	1.05	5.32	−2.64	4.55
	[−6.52, −0.93]	[−2.28, 4.37]	[2.63, 8.01]	[−5.07, −0.22]	[1.14, 7.96]
LASSO	0.65	0.45	−2.88	1.32	5.02
	[−3.65, 4.94]	[−3.28, 4.18]	[−5.38, −0.38]	[−1.83, 4.48]	[−0.14, 10.18]

Data-driven Subgroup Identification

- In GATES estimation, the percentile cutoffs are given
- Can we choose the cutoffs based on the data?
  1. those who benefit from treatment the most (exceptional responders)
  2. those who are harmed by treatment

- Challenges:
  1. sample size may not be large
  2. ML estimates of CATE may be biased and noisy
  3. proportion of exceptional responders may be small
- Can we provide a *statistical* guarantee?

Problem of the Standard Approach

- The problem is trivial if we had an infinite amount of data

$$p^* = \underset{p\in[0,1]}{\operatorname{argmax}}\, \Psi(p) \quad \text{where } \Psi(p) \;=\; \mathbb{E}[\underbrace{Y_i(1) - Y_i(0)}_{:=\psi_i} \mid F(s(X_i)) \geq 1-p],$$

- Standard method suffers from multiple testing problem:

$$\hat{p}_n \;=\; \underset{p\in[0,1]}{\operatorname{argmax}}\, \widehat{\Psi}_n(p) \quad \text{where } \widehat{\Psi}_n(p) \;=\; \frac{1}{np}\sum_{i=1}^{\lfloor np\rfloor} \hat{\psi}_{[n,i]}$$

where $s(X_{[n,1]}) \geq \cdots \geq s(X_{[n,n]})$ and

$$\hat{\psi}_{[n,i]} = \frac{T_{[n,i]}Y_{[n,i]}}{n_1/n} - \frac{(1 - T_{[n,i]})Y_{[n,i]}}{n_0/n}$$

Providing a Statistical Performance Guarantee

- (one-sided) Uniform confidence band:

$$\mathbb{P}\left(\forall p \in [0,1],\ \Psi(p) \geq \widehat{\Psi}_n(p) - C_n(p,\alpha)\right) \geq 1-\alpha.$$

- Safe identification of exceptional responders:

$$\underline{\hat{p}}_n \ = \ \operatorname*{argmax}_{p\in[0,1]} \widehat{\Psi}_n(p) - C_n(p,\alpha),$$

implying

$$\mathbb{P}\left(\Psi(p^*) \geq \widehat{\Psi}_n(\underline{\hat{p}}_n) - C_n(\underline{\hat{p}}_n,\alpha)\right) \geq \ \mathbb{P}\left(\Psi(\underline{\hat{p}}_n) \geq \widehat{\Psi}_n(\underline{\hat{p}}_n) - C_n(\underline{\hat{p}}_n,\alpha)\right)$$
$$\geq \ 1-\alpha.$$

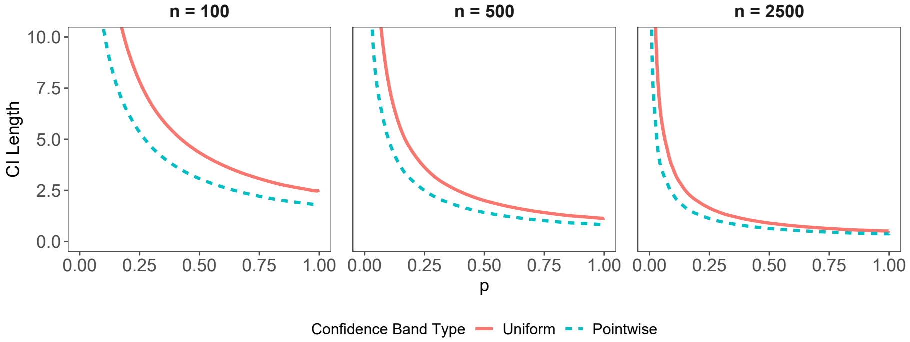
- Other data-driven selection of p is possible: e.g., for a given c

$$\text{estimate } \ \underline{\hat{p}}_n(c) \ = \ \sup\{p\in[0,1] : \widehat{\Psi}_n(p) - C_n(p,\alpha) \geq c\},$$
$$\text{to target } \ p^*(c) \ = \ \sup\{p\in[0,1] : \Psi(p) \geq c\}$$

Simulation Studies

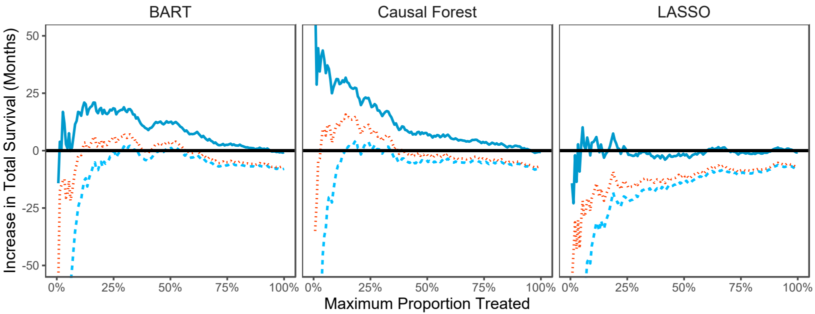
- A data generating process from the ACIC

ML algorithm	Uniform			Pointwise		
	$n = 100$	$n = 500$	$n = 2500$	$n = 100$	$n = 500$	$n = 2500$
BART	96.1%	96.0%	95.2%	87.2%	76.5%	70.3%
Causal Forest	96.0%	95.3%	95.7%	83.7%	77.1%	71.9%
LASSO	95.8%	95.6%	95.6%	84.1%	76.0%	69.8%



Empirical Application

- Clinical trial data on late-stage prostate cancer ($n_1 = 125$, $n_0 = 127$)
- Outcome: total survival in months, Treatment: estrogen
- Sample-split (40% train., 60% eval.), ATE estimate -0.3 month



ML algorithm	Estimated proportion of exceptional responders	Estimated GATES	90% uniform confidence band
Causal Forest	18.8%	27.2	$(4.45, \infty)$
BART	32.2%	18.1	$(2.12, \infty)$
LASSO	91.2%	1.35	$(-6.26, \infty)$

Concluding Remarks

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

- Safe deployment of causal ML requires uncertainty quantification
- Subgroup identification with statistical performance guarantees
  - Does not assume that ML algorithms are accurate
  - Computationally efficient (no resampling)
  - 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://imai.fas.harvard.edu/research/