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

Kosuke Imai

Harvard University

January 8, 2024

JSQPS Winter Meeting

Joint work with Michael Lingzhi Li (Harvard Business School)

Motivation and Overview

- Two methodological revolutions over the past two 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 discovered by generic causal ML

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
  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 data set

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

- 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 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?

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

- Bias bound and exact variance are derived, accounting for the estimation uncertainty of 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:
   - 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:
   $$(\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}}))^\top \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\tau}} - \boldsymbol{\mu}^*(\hat{\boldsymbol{\tau}})) \xrightarrow{d} \bar{\chi}^2_K.$$
   where $\boldsymbol{\mu}^*(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{\mu}} \|\boldsymbol{\mu} - \boldsymbol{x}\|_2^2$ subject to $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_K$.

Estimation and Evaluation Using the Same Data

- Cross-fitting procedure:
  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

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

- Unbiasedness: $\mathbb{E}(\hat{\tau}_k(F; n-m)) = \tau_k(F; n-m)$
- Finite-sample (conservative) variance estimator (Imai and Li, *JASA*, 2023)

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
- More conservative and lower power than fixed case
- When sample size is large, cross-fitting yields higher power

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: Causal Forest, BART, 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]

- Greater statistical power with cross-fitting
- ML algorithms are not necessarily reliable

Results of Hypothesis Tests

	Causal Forest		BART		LASSO	
	stat	*p*-value	stat	*p*-value	stat	*p*-value
Sample-splitting						
Homogeneity	9.78	0.08	2.76	0.74	5.26	0.36
Rank-consistency	3.07	0.32	1.13	0.66	3.14	0.30
Cross-fitting						
Homogeneity	30.29	0.00	2.32	0.80	10.79	0.06
Rank-consistency	0.06	0.69	0.04	0.89	0.45	0.71

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://imai.fas.harvard.edu/research/`