Fisher-Schultz Lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India

Victor Chernozhukov et al. (2025), Econometrica (forthcoming).

Naoki Eguchi

2025.7.9 ミクロ計量経済学

Faculty of Medicine, Kyoto University

Introduction

Motivation

- When conducting RCT (Randomized Controlled Trails), reseachers and policy makers are often curious about not only ATE but also HTE (Heterogeneous Treatment Effects).
- In such cases, ML methods are good for estimating CATE but inconsistent in most cases especially high-dimension ($d > \log N$)
- Also, we have a difficulty of getting uniformly valid inference.
- Lasso-based methods are not a magic bullet in that regularization bias and untestable assumption occurs.
- Consider the conditional unconfoundedness setting such that $D \perp\!\!\!\perp (Y(1), Y(0)) | Z$ and the propensity score $p(Z) = P[D = 1|Z] \in (0, 1)$ is known.

Proposed estimator

- Let (M, A) denote a random partition of $\{1, ..., N\}$.
- Stage 1 : From the auxilirary samlple A, we obtain ML estimators of BCA and CATE $z \mapsto B(z)$ and $z \mapsto S(z)$.
    - These estimators are (of course) biased and noisy. (but it is okay !)
- Stage 2 : From the main sample M, we focus on the feature of CATE.
    - Best Linear Preditor (BLP) of the CATE
    - Sorted Group Average Treatment Effects (GATES) : average of CATE in group
    - Classification Analysis (CLAN) : compare the most and least affected group

Main identification results and estimation strategies

BLP (Best Linear Predictor)

- Firstly, we obtain the estimator $S(Z)$ for CATE $s_0(Z)$ by some ML method using the auxilirary samples $\mathcal{A}$.
- BLP is defined as the projection of $s_0(Z)$ on the linear span of 1 and $S(Z)$ in $L^2(P)$.

$$\text{BLP}[s_0(Z)|S(Z)] = \arg \min_{f(Z) \in \text{Span}(1, S(Z))} \mathbb{E}[(s_0(Z) - f(Z))^2]$$

  - This equals to the solution of $\arg \min_{b_1, b_2} \mathbb{E}[(s_0(Z) - b_1 - b_2 S(Z))^2]$.

    $\rightarrow \beta_1 = \mathbb{E}[s_0(Z)], \beta_2 = \dfrac{\text{Cov}(s_0(Z), S(Z))}{\text{Var}(S(Z))}$

First strategy : Weighted Residual BLP

- Consider the regresson model with the moment condition as follows.

$$Y = \alpha' X_1 + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S(Z) - \mathbb{E}[S(Z)]) + \epsilon, \mathbb{E}[w(Z)\epsilon X] = 0$$

$$\text{where } w(Z) = \frac{1}{p(Z)(1 - p(Z))}, X = (X_1, X_2),$$

$$X_1 = (1, B(Z)), X_2 = (D - p(Z), (D - p(Z)(S(Z) - \mathbb{E}[S(Z)]))).$$

Theorem 1

- Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. (known function)
- Assume that Y and X have finite second moments, $\mathbb{E}[XX']$ is full rank, and $\mathrm{Var}(S(Z)) > 0$.
- Then, $(\beta_1, \beta_2)' = \arg\min_{b_1, b_2} \mathbb{E}[(s_0(Z) - b_1 - b_2 S(Z))^2]$ (Identified)

Second strategy : Horvitz-Thompson BLP

- Horvitz-Thompson transformed response YH such that $H = \dfrac{D - p(Z)}{p(Z)(1 - p(Z))}$ provides an unbiased signal about CATE : $\mathbb{E}[YH|Z] = s_0(Z)$

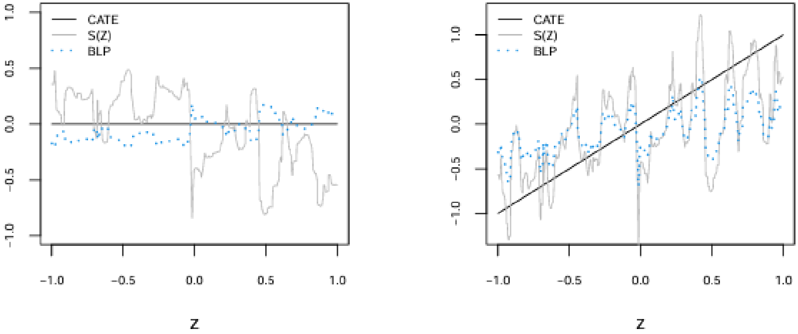
$$YH = \mu' X_1 H + \beta_1 + \beta_2(S(Z) - \mathbb{E}[S(Z)]) + \epsilon, \ \mathbb{E}[\epsilon \tilde{X}] = 0$$

$$\text{where } X_1 = (1, B(Z), p(Z)S(Z))' \ \tilde{X} = (X_1' H, 1, S(Z) - \mathbb{E}[S(Z)])$$

Theorem 2

- Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. (known function)
- Assume that Y has finite second moments, $\mathbb{E}[\tilde{X}\tilde{X}']$ is finite and full rank, and $\text{Var}(S(Z)) > 0$.
- Then, $(\beta_1, \beta_2)' = \arg\min_{b_1, b_2} \mathbb{E}[(s_0(Z) - b_1 - b_2 S(Z))^2]$ (Identified)

FIGURE 1. BLP Using ML Proxy vs the ML Proxy



Risk reduction: 68%

Risk reduction: 5%

NOTES: The CATE is plotted with the solid black line; the proxy predictor $S(Z)$, produced by Random Forest, is plotted with the solid grey (light) line; and the BLP is plotted with the dotted blue line. The left panel corresponds to the no heterogeneity example, $s_0(z) = 0$ and the right panel to the strong heterogeneity example, $s_0(z) = z$. In both panels, the BLP is less noisy than the ML proxy reducing the RMSE by 68% and 5%.

GATES (sorted Group Average Treatment Effects)

- Firstly, we build the groups by the estimated value $S(Z)$ of $s_0(Z)$.

$$G_k = \{S(Z) \in I_k\}, k = 1, ..., K, I_k = [l_{k-1}, l_k) : \text{disjoint}$$

- The estimand ”GATES” is defined as $\mathbb{E}[s_0(Z)|G_k]$ for $k = 1, ..., K$.
    - WR GATES

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k (D - p(Z)) \mathbf{1}_{G_k} + \nu, \ \mathbb{E}[w(Z)\nu W] = 0$$

$$\text{where } W = (X_1, W_2')', \ W_2 = \{(D - p(Z))\mathbf{1}_{G_k}\}_{k=1}^{K}.$$

    - HT GATES

$$YH = \mu_0' X_1 H + \sum_{k=1}^{K} \gamma_k \mathbf{1}_{G_k} + \nu, \ \mathbb{E}[\nu \tilde{W}] = 0$$

$$\text{where } \tilde{W} = (X_1' H, \tilde{W}_2')', \ \tilde{W}_2 = \{\mathbf{1}_{G_k}\}_{k=1}^{K}$$

Identification of GATES

Theorem 3

- Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. (known function)
- Assume that Y has finite second moments and both $\mathbb{E}[WW']$ and $\mathbb{E}[\tilde{W}\tilde{W}']$ are finite and full rank.
- Then, $\gamma = \{\gamma_k\}_{k=1}^{K}$ defined in two different ways are equivalent and identified.

$$\gamma_k = \mathbb{E}[s_0(Z)|G_k]$$

CLAN (Classification Analysis)

- When BLP and GATES reveal substantial heterogeneity, it is interesting to know the properties of the subpopulations that are most and least affected.
  - We focus on the ”least affected group” G_1 and ”most affected group” G_K.
- Let $g(Y, Z)$ be a vector of characteristics of an observational unit. The estimands are

$$\delta_1 = \mathbb{E}[g(Y, Z)|G_1] \quad \text{and} \quad \delta_K = \mathbb{E}[g(Y, Z)|G_K]$$

- These parameters are identified with no assumption because they are just average of observed variables.
- We compare δ_1 and δ_K to detect (single out) the covariates which causes the heterogeneity.
  - We can extend the comparison of not only averages but also variances or distributions.

”Variational” estimation and inference methods

Uncertainty

- Let θ denote a generic target parameter such as BLP β_2 or GATE γ_k.
- There are two principal sources of sampling uncertainty.
  - Estimation uncertainty regarding the parameter θ, conditional on the data subpopulations
  - Uncertainty or ”variation” induced by the data splitting
- Actually, estimation uncertainty is a standard topic, so, as usual, we can solve this problem by the Gaussian approximation to construct a confidence interval.
- On the other hands, data-splitting uncertainty is a novel topic, which is solved by taking a median of any estimators in permuated splitting.

Estimation uncertainty in single split

- Consider a sample split $\{(a, m)\}$ of $\{1, ..., N\}$ with $|a| = N - n, |m| = n$.
- All estimators θ_a satisfies the sufficient conditions for being approxmately Gaussian, conditionally on $Data_a$.

$$P(\frac{\hat{\theta_a} - \theta_a}{\hat{\sigma_a}} < z|Data_a) \to \Phi(z) \text{ for } z \in \mathbb{R}, \text{ as } N \text{ and } n \to \infty$$

- Therefore, the confidence interval represents

$$[L_a, U_a] = [\hat{\theta}_a \pm \Phi^{-1}(1 - \frac{\alpha}{2})\hat{\sigma}_a]$$

- We have straightforward inference conditional on a sigle data split.

Splitting uncertainty in multiple splits

- For each data split $\{(a, m)\}$ such that $a \in \mathcal{A}$, we obtain estimators $\{\hat{\theta}_a | a \in \mathcal{A}\}$.
    - Then, we take the median of it : $\hat{\theta} = M[\hat{\theta}_a | Data]$
    - Also, the β-quantile confidence interval is

$$[L, U] \text{ where } L = Q_\beta(L_a | Data), U = Q_{1-\beta}(U_a | Data)$$

- Median and β-quantile achieves the concentration property.

$$\mathbb{E}[|\hat{\theta} - \theta_0|] \leq \mathbb{E}[|\hat{\theta}_a - \theta_0|] \text{ for any } \hat{\theta}_a$$

$$\max\{\mathbb{E}[|U - \theta_0|], \mathbb{E}[|L - \theta_0|]\} \leq \max\{\mathbb{E}[|U_a - \theta_0|], \mathbb{E}[|L_a - \theta_0|]\} \text{ for any } U_a, L_a$$

$$|U - L| \leq \sup_{a \in \mathcal{A}} |U_a - L_a|$$

- By taking the median or quantile, we have a kind of robustness.

Causal machines that learn CATE better

Causal learners for Stage 1

- In Stage 1, using the auxilirary sample A, we estimate
  - BCA (Baseline Conditional Average) : $b_0(Z) = \mathbb{E}[Y(0)|Z]$
  - CATE (Conditional Average Treatment Effect) : $s_0(Z) = \mathbb{E}[Y(1) - Y(0)|Z]$

→ solve either of Weighted Residual (WR) learner or Horvitz-Thompson (HT) learner.

$$(B, S) \in \arg\min_{b \in \mathcal{B}, s \in \mathcal{S}} \sum_{i \in A} \frac{1}{p(Z_i)(1 - p(Z_i))} \{Y_i - b(Z_i) - (D_i - p(Z_i))s(Z_i)\}^2$$

$$(B, S) \in \arg\min_{b \in \mathcal{B}, s \in \mathcal{S}} \sum_{i \in A} \{\frac{D_i - p(Z_i)}{p(Z_i)(1 - p(Z_i))}(Y_i - b(Z_i)) - s(Z_i)\}^2$$

where $\mathcal{B}$ and $\mathcal{S}$ are functional parameter spaces

- 以後は簡単のため, $w(Z) = \dfrac{1}{p(Z)(1 - p(Z))}, H = \dfrac{D - p(Z)}{p(Z)(1 - p(Z))}$ と表記する.

Oracle properties of the population objective functions

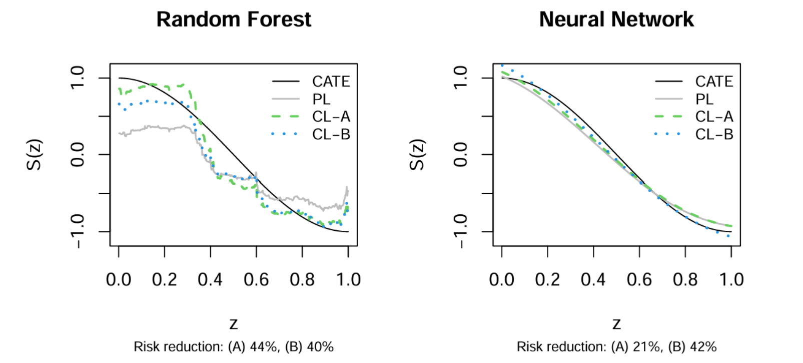
Theorem 4

- Suppose $Y, b(Z), s(Z), w(Z) \in L^2$ (2 乗可積分).
- Then, the expectation of the loss functions can be decomposed

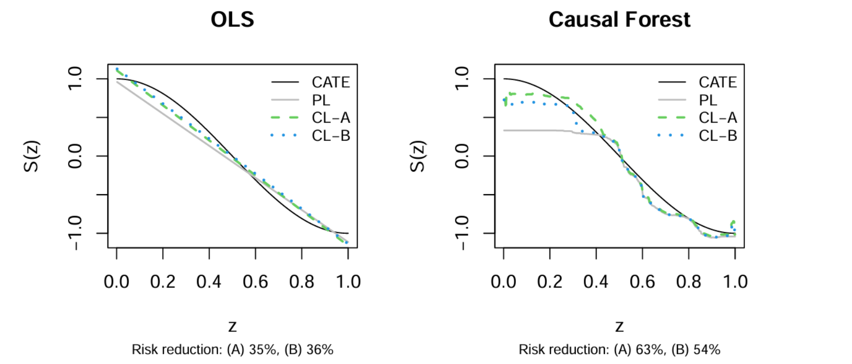
$$\begin{aligned} &\mathbb{E}[w(Z)\{Y - b(Z) - (D - p(Z))s(Z)\}^2] = \mathbb{E}[(s_0(Z) - s(Z))^2] + C_{ib} \\ &\mathbb{E}[(H(Y - b(Z) - s(Z)))^2] = \mathbb{E}[(s_0(Z) - s(Z))^2] + C_{2b} \end{aligned}$$

where
$C_{1b} = \mathbb{E}[w(Z)(\tilde{b}_0(Z) - b(Z))^2] + C_1, C_{1b} = \mathbb{E}[w(Z)(\bar{b}_0(Z) - b(Z))^2] + C_2$

- This theorem shows that the minimizers provide the best approxmation for $s_0(Z)$ in the sense of mean-squared error in the class $\mathcal{S}$.
- Moreover, this occurs even though we do not know $s_0(Z)$. (oracle!)



- We compare the CATE learners derived from
  - the standard preditive Random Forest (RF) and Neural Network (NN)
  - Causal Learners (CL) from RF and NN that solve the objective function
- We find that the causal learners (CL) are better approximating the CATE function.



- We can improve the standard predictive OLS by the causal OLS taht solves the objective function.
- Also, improve the causal forest by a causal boosting step that solves the objective function.

Implementation Details

Inference algorithm

1. Split the sample into the main sample M and the auxilirary sample A.
2. Using A, train each (optional) ML method and output prediction B (BCA) and S (CATE) for M.
3. Estimate BLP, GATES and CLAN using M.
4. If the winning ML methods were not chosen, we chose the best-of-fit in median-aggregated estimator $\hat{\theta}$. (e.g. cross-varidation)
5. Compute and report quantile-aggregated point-estimate, p-values, and confidence intervals.

Final marks

Summary

- We focus the estimation of HTE, which is usually biased and inconsistent.
- Thus, we use ML method for proxying CATE, then, feature just best linear predictor, which is easy to interpret.
- This agnostic approach enables us to be valid in high-dimension, not to make strong assumption, and to avoid over-fitting.
- For sample splitting, we take a median for robustness.

References

References

- Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val (2025), *Fisher-Schultz Lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India*, Econometrica (forthcoming).
- Xinkun Nie, Stefan Wager (2021), *Quasi-Oracle Estimation of Heterogeneous Treatment Effects*, Biometrika.
- Susan Athey, Guido Imbens (2016), *Recursive partitioning for heterogeneous causal effects*, Proceedings of the National Academy of Sciences.
- Kosuke Imai and Michael Lingzhi Li (2025), *Statistical Inference for Heterogeneous Treatment Effects Discovered by Generic Machine Learning in Randomized Experiments*, Journal of Business and Economic Statistics.
- 金本拓 (2024), 因果推論ー基礎からの機械学習・時系列解析・因果探索を用いた意思決定のアプローチー, オーム社. (主に5章)