CATE Lasso: Conditional Average Treatment Effect Estimation with High-Dimensional Linear Regressio

CATE Lasso: Conditional Average Treatment Effect Estimation with High-Dimensional Linear Regression (arXiv:2310.16819)
Triple/Debiased Lasso for Statistical Inference of Conditional Average Treatment Effects

Masahiro Kato and Masaaki Imaizumi

The University of Tokyo.

1. Introduction

- **High-dimensional linear Conditional Average Treatment Effect (CATE) estimation with Lasso-based regularization.**
  - Unlike the Lasso, we do not assume sparsity to parameters.
  - Implicit sparsity: impose sparsity on a difference of parameters.
- **Contributions**.
  - Least squares with Lasso specialized for CATE estimation.
  - Extend the Lasso to debiased Lasso.
  - Consistency and confidence intervals under implicit sparsity, which is a weaker assumption than the standard sparsity.

2. Problem Setting

- In empirical analysis, we are interested in treatment effects.
  - Treatment effects: how outcomes change after taking some action.
  - Examples: medicine, coupon, and personalized advertisement.
  - Treatment effects are counterfactual values.
- CATE: expected treatment effect conditioned on covariates.
- Binary treatments and potential outcomes.
  - Binary treatments $d \in \{1, 0\}$. Ex. treatment and control.
  - For $d \in \{1, 0\}$, there is a corresponding potential outcome $Y^d \in \mathbb{R}$.
  - p-dimensional covariates $X \in \mathbb{R}^p$.
- Linear models: $Y^d = X^\top \beta^d + \varepsilon^d,\ \mathbb{E}[\varepsilon^d] = 0.$
  - When X is random variable, we assume $\mathbb{E}[\varepsilon^d | X] = 0.$
- CATE: $\mathbb{E}[Y^1 - Y^0 | X = x] = x^\top(\beta^1 - \beta^0).$
- **Goal**: estimate the CATE from observations.
- **Observations**.
  - Each unit i takes an action D_i.
  - We observe $Y_i = DY_i^1 + (1 - D_i)Y_I^0.$
  - Each unit has p-dimensional covariates $X_i \in \mathcal{X} \subset \mathbb{R}^p.$
  - Observe n samples: $\{(Y_i, D_I, X_i)\}_{i=1}^n.$
- We consider high-dimensional X_i; that is, p takes a large value.
  - We do not assume sparsity on β^d directly.

3. Implicit Sparsity

- Separability of the parameter β^d.
- β^d is separated into an individual parameter α^d and common parameter γ:

$$\beta^d = (\alpha^{d\top}, \gamma^\top)^\top.$$

  - Individual parameter α^d: different value in linear models for Y_i^1 and Y_i^0.
  - Common parameter γ: same value in linear models for Y_i^1 and Y_i^0.
  - How to separate β^d is unknown.
- Assume that α^d is a s-dimensional parameter.
- Using $X = (W^\top, Z^\top)^\top$, we denote the linear model as

$$Y^d = X^\top\beta + \varepsilon^d = W^\top\alpha^d + Z^\top\gamma + \varepsilon^d.$$

- **Implicit sparsity: The dimension of CATE depends on α^d (s dimension):**

$$Y^1 - Y^0 = X^\top(\beta^1 - \beta^0) + \varepsilon^1 - \varepsilon^0 = W^\top(\alpha^1 - \alpha^0) + \varepsilon^1 - \varepsilon^0.$$

- We propose a novel Lasso estimator that utilizes the implicit sparsity.
- We estimate β_0^1 and β_0^0 as $(\hat{\beta}^1, \hat{\beta}^0) :=$

$$\arg\min_{\beta^1, \beta^0 \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n 1[D_i = 1]\left(Y_i - X_i^\top\beta^1\right)^2 + \frac{1}{n}\sum_{i=1}^n 1[D_i = 0]\left(Y_i - X_i^\top\beta^0\right)^2 + \lambda\|\beta^1 - \beta^0\|_1.$$

4. CATE LASSO

- We refer to $(\hat{\beta}^1, \hat{\beta}^0)$ as the **CATE Lasso estimator**.
  - In the standard Lasso, we penalize β^d itself.
  - CATE Lasso utilizes $\beta^1 - \beta^0 = \begin{pmatrix}\alpha^1 - \alpha^0 \\ \gamma - \gamma\end{pmatrix} = \begin{pmatrix}\alpha^1 - \alpha^0 \\ 0\end{pmatrix}.$

→ We predict the CATE by $x^\top\hat{\beta}$, where $\hat{\beta} = \hat{\beta}^1 - \hat{\beta}^0$.

- For $\hat{\beta} = \hat{\beta}^1 - \hat{\beta}^0$, we obtain estimators $(\hat{\beta}, \hat{\beta}^0)$ by the following two step:

1. **Interpolating estimator of β^0:**

   First estimate β^0 as $\hat{\beta}^0 := \left(\sum_{i=1}^n 1[D_i = 0]\, X_i X_i^\top\right)^+ \sum_{i=1}^n 1[D_i = 0] X_i\, Y_i.$

   $\left(\sum_{i=1}^n 1[D_i = 0] X_i^\top\, X_i\right)^+$ is the pseudo-inverse of $\sum_{i=1}^n 1[D_i = 0] X_i^\top\, X_i.$

2. **Lasso estimator of β:**

   Estimate β as $\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n 1[D_i = 1]\left(Y_i - X_i^\top(\beta + \hat{\beta}^0)\right)^2 + \lambda\|\beta\|_1$

- **$\widehat{\beta} - \beta = o_P(1)$ as $n \to \infty$ under the standard assumptions of Lasso!**
  - We do not identify β^1 and β^0 themselves but identify β.

➢ Construction of confidence intervals is difficult...

5. DR CATE LASSO

- Consider another estimator for CATE estimator.

➢ DR CATE Lasso: Employ the doubly robust (DR)-estimator.

- Let $\hat{\mu}^d(x)$ and $\hat{\pi}(d \mid x)$ be estimators of $\mu^d(x) := \mathbb{E}[Y^d(1) \mid X = x]$ and $\pi(d \mid x) := \mathbb{E}[D = d \mid X = x]$ that satisfy

$$\left\|\hat{\mu}^d(X) - \mu^d(X)\right\|_2 = o_p(1), \qquad \|\hat{\pi}(d \mid X) - \pi(d \mid X)\|_2 = o_p(1),$$

$$\left\|\hat{\mu}^d(X) - \mu^d(X)\right\|_2 \|\hat{\pi}(d \mid X) - \pi(d \mid X)\|_2 = o_p\left(n^{-1/2}\right).$$

- Construct the DR estimator of $Y^1 - Y^0$ as

$$Q_i := \frac{1[D_i = 1]\left(Y_i - \hat{\mu}^1(X_i)\right)}{\hat{\pi}(1 \mid X_i)} - \frac{1[D_i = 0]\left(Y_i - \hat{\mu}^0(X_i)\right)}{\hat{\pi}(0 \mid X_i)}.$$

- Estimate β as $\hat{\beta} := \arg\min_{\beta, \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left(\mathrm{Q}_i - X_i^\top\beta\right)^2 + \lambda\|\beta\|_1.$

- The Lasso estimator is biased → apply the debiased Lasso technique.

➢ **Debiased DR CATE Lasso estimator** is defined as $\hat{b} = \hat{\beta} + \widehat{\Theta}\lambda\hat{\kappa}.$

  - $\widehat{\Theta}$: an approximate of $\left(\sum_{i=1}^n X_i X_i^\top\right)^{-1}$ using the node-wise regression.
  - $\lambda\hat{\kappa} = \sum_{i=1}^n X_i\left(Y_i - X_i^\top\hat{\beta}\right)/n.$

- We can show the $\sqrt{n}$-consistency and confidence interval for $\hat{b}$.

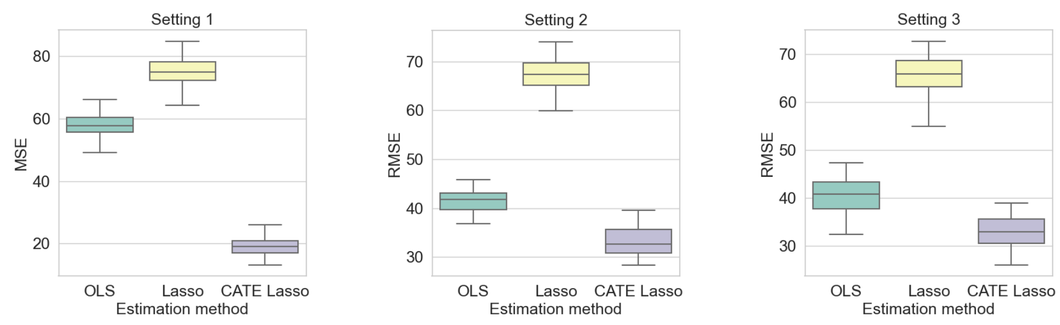
↔ Require more complicated procedures and assumptions than CATE Lasso.

6. Experiments

➢ Simulation studies to investigate the performances of the CATE Lasso.

$$Y^d = 1 + W^\top\alpha^d + Z^\top\gamma + \varepsilon^d, \qquad \alpha^d, \gamma \sim \mathrm{Uniform}[-10, 10].$$

- Compare the CATE Lasso estimator with the OLS and Lasso estimators.
- Root MSE: $\sqrt{\mathbb{E}\left[\left(\mathbb{E}[Y^1 - Y^0 | X] - X^\top\hat{\beta}\right)^2\right]}.$
- Sample size $n = 500$. Dimension $p = 300$
- Setting 1: $s = 10$. Setting 2: $s = 25$. Setting 3: $s = 50$.



Reference

Bühlmann, P. and van de Geer, S. *Statistics for high-dimensional data*.

van de Geer, S., Bu¨hlmann, P., Ritov, Y., and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models.