Direct estimation of conditional averaging treatment effect in high dimensions

Shota Katayama

Keio University, Japan

1 Introduction

The estimation of conditional average treatment effect (CATE) is a general and fundamental problem in observational studies. Such estimation problem is essential for policy evaluation, personalized medicine, offline or online marketing and advertising. Usually, to identify CATE, one requires the strong ignorability condition which says that outcomes and treatment assignment is independent conditional on covariates. In other words, only the covariates we collect affect both of outcomes and treatment assignment. If we fail to collect such a covariate, the strong ignorability does not hold. Clearly, a large number of covariates tends to meet the strong ignorability, although it is uncheckable condition from observations. With advances of information technology and database system, it would be plausible to consider the high dimensional covariates.

In this talk, we consider the estimation of CATE in high dimensions. Following the Neyman–Rubin's potential outcome framework (Rubin, 1974; Neyman et al., 1990), assume that there is a potential outcomes $(Y_i(0), Y_i(1))$ for each sample $i \in \{1, 2, \ldots, n\}$. Let $T_i \in \{0, 1\}$ be the assignment indicator. Then, $Y_i(0) \in \mathbb{R}$ is the potential outcome when the sample i is assigned to the control ($T_i = 0$) and $Y_i(1) \in \mathbb{R}$ is the potential outcome when it is assigned to the treatment ($T_i = 1$). Assume that we have n independent and identically distributed examples $\{(\boldsymbol{X}_i, T_i, Y_i(T_i))\}_{i=1}^n$ where $\boldsymbol{X}_i \in \mathbb{R}^p$ is the covariates with possibly high dimensions, that is, $p \gg n$. Our goal is to estimate the conditional average treatment effect (CATE) given by

$$\tau^*(\boldsymbol{x}) = \mathbb{E}\big\{Y_i(1) - Y_i(0) \,|\boldsymbol{X}_i = \boldsymbol{x}\big\}.$$

To identify the CATE, we assume the following strong ignorability condition.

Assumption 1. $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \,|\boldsymbol{X}_i$

Moreover, we assume the linearity for the potential outcomes.

Assumption 2. $\mathbb{E}\{Y_i(0)\,|\boldsymbol{X}_i = \boldsymbol{x}_i\} = \boldsymbol{x}_i^T\boldsymbol{\beta}_0^*$ and $\mathbb{E}\{Y_i(1)\,|\boldsymbol{X}_i = \boldsymbol{x}_i\} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1^*$.

From Assumption 2, we have $\tau^*(\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^*)$ and we can estimate $\boldsymbol{\beta}_1^*$ from the treated examples and can estimate $\boldsymbol{\beta}_0^*$ from the control examples under Assumption 1. Since the covariates $\boldsymbol{X}_i$ is high dimension, a natural approach would be applying the Lasso proposed by Tibshirani (1996) for each treated and control examples, i.e.,

$$\hat{\boldsymbol{\beta}}_t = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\sum_{T_i=t}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 + \lambda_t\|\boldsymbol{\beta}\|_1, \quad t = 0, 1,$$

where $Y_i = Y_i(T_i)$. Thus, we obtain the estimator of CATE as $\hat{\tau}(\boldsymbol{x}) = \boldsymbol{x}^T(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)$. However, such the procedure estimate $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_1^*$ separately. The treated (control) outcomes are predicted by treated (control) covariates. Hence, if $\boldsymbol{x}$ is coming from the distribution of $\boldsymbol{X}\,|T = 1$, then $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_1$ would be accurate but $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_0$ be not. Moreover, the non-zero elements of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$ usually do not imply zero elements of $\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0$ even when the corresponding elements of $\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^*$ are zero.

2 Proposed methododology

We construct a direct estimation procedure for $\boldsymbol{\theta}^* = \boldsymbol{\beta}_1^* - \boldsymbol{\beta}_0^*$ via the well-known consequence of the strong ignorability condition, given by

$$\tau^*(\boldsymbol{x}) = \mathbb{E}\left[Y_i\left\{\frac{T_i}{e(\boldsymbol{x})} - \frac{1 - T_i}{1 - e(\boldsymbol{x})}\right\}\middle|\boldsymbol{X}_i = \boldsymbol{x}\right],$$

where $e(\boldsymbol{x}) = \mathbb{P}(T_i = 1\,|\boldsymbol{X}_i = \boldsymbol{x})$ is the propensity score function at $\boldsymbol{x}$. Thus, $\boldsymbol{\theta}^*$ can be estimated by regressing the appropriately weighted outcomes on the covariates. The propensity score is unknown in most cases. An approach to estimate it in high dimensions may be generalized linear regression with sparse regularization (see, e.g., Fan and Li (2001) and Van de Geer (2008)), but it may lead to an biased estimator for $\boldsymbol{\theta}^*$ when the propensity score function is misspecified.

In this talk, inspired by Athey et al. (2018), a two-step estimation procedure of $\boldsymbol{\theta}^*$ is proposed. The first step obtains weightings for outcomes without specifying the propensity score and then Lasso is applied to the weighted outcomes. Let $\boldsymbol{Y}_0$ ($\boldsymbol{Y}_1$) be the vector of control (treated) outcomes and $\mathbb{X}_t$ ($t = 0, 1$) be the corresponding covariate matrix. Define

the Lasso for weighted outcomes as

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{D}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n}\|\boldsymbol{D}\boldsymbol{Y} - \mathbb{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1, \tag{1}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_0 \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} \mathbb{X}_1 \\ \mathbb{X}_0 \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{O} \\ \boldsymbol{O} & -\boldsymbol{D}_0 \end{pmatrix}$$

with diagonal matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$. Let $\hat{\boldsymbol{\Sigma}} = \mathbb{X}^T\mathbb{X}/n$ be the covariance matrix. Roughly speaking, the score function of $\|\boldsymbol{D}\boldsymbol{Y} - \mathbb{X}\boldsymbol{\theta}\|_2^2/(2n)$ at $\boldsymbol{\theta}^*$ is close to zero if

$$\frac{1}{n}\mathbb{X}_0^T\boldsymbol{D}_0\mathbb{X}_0 \approx \hat{\boldsymbol{\Sigma}} \quad \text{and} \quad \frac{1}{n}\mathbb{X}_1^T\boldsymbol{D}_1\mathbb{X}_1 \approx \hat{\boldsymbol{\Sigma}},$$

which means that $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ work in order to balance the two weighted covariance matrices of control and treated covariates through $\hat{\boldsymbol{\Sigma}}$. We consider computing the weight matrices as

$$\underset{\boldsymbol{D}_t}{\operatorname{argmin}} \|\boldsymbol{D}_t\|_{\max} \quad \text{subject to} \quad \|\mathbb{X}_t^T\boldsymbol{D}_t\mathbb{X}_t/n - \hat{\boldsymbol{\Sigma}}\|_{\max} \le \eta_t, \quad t = 0, 1, \tag{2}$$

where $\eta_t > 0$ $(t = 0, 1)$ and $\|\cdot\|_{\max}$ is the element-wise infinity norm. The minimization of $\|\boldsymbol{D}_t\|_{\max}$ is required to control the variance of the errors in linear models. After obtaining the weight matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ from (2), we can compute $\hat{\boldsymbol{\theta}}_{\boldsymbol{D}}$ through (1). To compute (2) efficiently, we apply the alternating direction method of multipliers (ADMM) after smoothing the objective function by the method in Nesterov (2005). The objective function $\|\boldsymbol{D}_t\|_{\max}$ in (2) is replaced by

$$\|\boldsymbol{D}_t\|_{\max}^{\mu_t} = \max_{\|\boldsymbol{z}_t\|_1\le 1} \boldsymbol{z}_t^T\boldsymbol{w}_t - \frac{\mu_t}{2}\|\boldsymbol{z}_t\|_2^2,$$

where $\boldsymbol{w}_t$ is the vector of the diagonal elements of $\boldsymbol{D}_t$ and $\mu_t \ge 0$ is the smoothing parameter. When $\mu_t = 0$, $\|\boldsymbol{D}_t\|_{\max}^{\mu_t} = \|\boldsymbol{D}_t\|_{\max}$ and $\|\boldsymbol{D}_t\|_{\max}^{\mu_t}$ is differentiable when $\mu_t > 0$. The optimization problem (2) with $\|\boldsymbol{D}_t\|_{\max}^{\mu_t}$ is equivalent to

$$\underset{\boldsymbol{D}_t,\boldsymbol{\Theta}_t}{\operatorname{argmin}} \|\boldsymbol{D}_t\|_{\max}^{\mu_t} \quad \text{subject to} \quad \boldsymbol{\Theta}_t = \mathbb{X}_t^T\boldsymbol{D}_t\mathbb{X}_t/n - \hat{\boldsymbol{\Sigma}} \quad \text{and} \quad \|\boldsymbol{\Theta}_t\|_{\max} \le \eta_t$$

whose Lagrangian function is given by

$$L(\boldsymbol{D}_t, \boldsymbol{\Theta}_t, \boldsymbol{\Lambda}_t) = \|\boldsymbol{D}_t\|_{\max}^{\mu_t} + \operatorname{tr}\boldsymbol{\Lambda}_t\{\boldsymbol{\Theta}_t - (\mathbb{X}_t^T\boldsymbol{D}_t\mathbb{X}_t/n - \hat{\boldsymbol{\Sigma}})\} + \frac{\rho_t}{2}\|\boldsymbol{\Theta} - (\mathbb{X}_t^T\boldsymbol{D}_t\mathbb{X}_t/n - \hat{\boldsymbol{\Sigma}})\|_{\mathrm{F}}^2,$$

where $\boldsymbol{\Lambda}_t$ is the Lagrange multiplier and $\rho_t > 0$ is controlling the step size. Them, the ADMM algorithm iteratively updates each parameter as

$$\boldsymbol{D}_t^{(k+1)} \leftarrow \underset{\boldsymbol{D}_t}{\operatorname{argmin}}\, L\big(\boldsymbol{D}_t, \boldsymbol{\Theta}^{(k)}, \boldsymbol{\Lambda}_t^{(k)}\big), \tag{3}$$

$$\boldsymbol{\Theta}_t^{(k+1)} \leftarrow \underset{\boldsymbol{\Theta}_t}{\operatorname{argmin}}\, L\big(\boldsymbol{D}_t^{(k+1)}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}_t^{(k)}\big), \tag{4}$$

$$\boldsymbol{\Lambda}_t^{(k+1)} \leftarrow \boldsymbol{\Lambda}_t^{(k)} + \rho_t\big\{\boldsymbol{\Theta}^{(k+1)} - (\mathbb{X}_t^T \boldsymbol{D}_t^{(k+1)} \mathbb{X}_t/n - \hat{\boldsymbol{\Sigma}})\big\}. \tag{5}$$

The optimization in (3) can be computed by the gradient descent for instance since $\|\boldsymbol{D}_t\|_{\max}^{\mu_t}$ is differentiable for $\mu_t > 0$. The update (4) is given by

$$\boldsymbol{\Theta}_t^{(k+1)} \leftarrow \operatorname{sgn}(\bar{\boldsymbol{\Theta}}_t) \min\big(|\bar{\boldsymbol{\Theta}}_t|, \eta_t\big), \quad \bar{\boldsymbol{\Theta}}_t = \mathbb{X}_t^T \boldsymbol{D}_t^{(k+1)} \mathbb{X}_t/n - \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Lambda}^{(k)}/\rho_t,$$

where all of $\operatorname{sgn}(\cdot)$, min and $|\cdot|$ are element-wise operator.

References

Athey, S., Imbens, G.W. and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B*, **80**(4), 597–623.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456), 1348–1360.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, **66**(5), 688.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical programming*, **103**(1), 127–152.

Splawa-Neyman, J., Dabrowska, D.M., and Speed, T.P. (1990). On the application of probability theory to agricultural experiments. *Essay on principles. Section 9. Statistical Science*, 465–472.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58**(1), 267–288.

Van de Geer, S.A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, **36**(2), 614–645.