Targeted minimum loss-based estimation and double/debiased machine learning

Konan Hara

University of Arizona

July 25, 2022

References

- “Targeted Maximum Likelihood Learning” by van der Laan & Rubin (2006)
- “CV-TMLE and double machine learning” in van der Laan’s website
- “Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning” by Díaz (2020)
- “Double/debiased machine learning for treatment and structural parameters” Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, & Robins (2018)
- “Locally Robust Semiparametric Inference with Debiased GMM” by Chernozhukov, Escanciano, Ichimura, Newey, & Robins (2022)

Regularization Bias

- Variables
  - Y: outcome
  - D: treatment indicator
  - X: covariates
- Observe $(y_i, d_i, x_i) \sim P_0$ i.i.d. for $i = 1, \dots, N$
- Easy to get $\sqrt{N}$-consistent θ_0 if

$$Y = D\theta_0 + X^\top \beta_0 + U, E[U|X, D] = 0, \beta_0 \in \mathbb{R}^p, \text{ where } p \text{ is small enough.}$$

- What happens if we apply lasso to the following model?

$$Y = D\theta_0 + X^\top \beta_0 + U, E[U|X, D] = 0, \beta_0 \in \mathbb{R}^p, \text{ where } p \text{ is very large.}$$

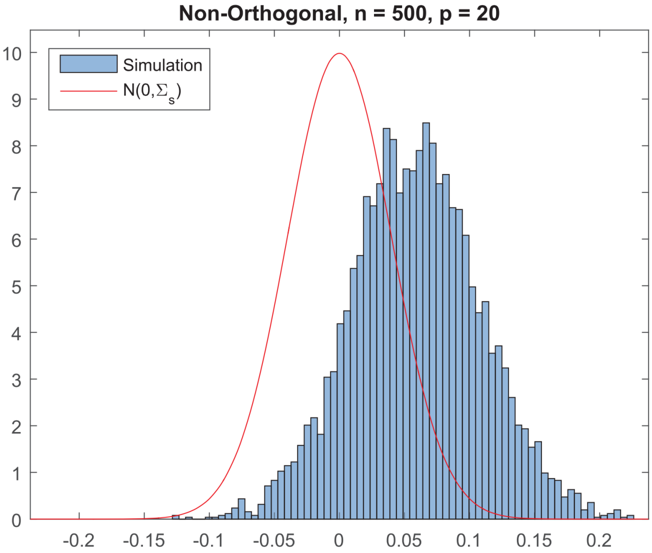
Regularization Bias

What happens if we apply ML prediction approach to the following model?

$$Y = D\theta_0 + g_0(X) + U, E[U|X, D] = 0.$$

1. Start from a guess of $\theta_0 \Rightarrow \hat{\theta}^0$
2. Apply ML to predict $Y - D\hat{\theta}^0$ using $X \Rightarrow \hat{g}^1(\cdot)$
3. Regress $Y - \hat{g}^1(X)$ on $D \Rightarrow \hat{\theta}^1$
4. Iterate until convergence $\Rightarrow \hat{\theta}_0$

Regularization Bias



Frish-Waugh-Lowell Theorem

Consider

$$Y = D\theta_0 + X^\top \beta_0 + U, E[U|X, D] = 0, \beta_0 \in \mathbb{R}^p, \text{ where } p \text{ is small enough.}$$

θ_0 can be consistently estimated by regressing

- residual of regression Y on X

on

- residual of regression D on X.

Double/Debiased Machine Learning Estimator

What happens if we apply FWL-style estimation to the following?

$$Y = D\theta_0 + X^\top \beta_0 + U, E[U|X, D] = 0, \beta_0 \in \mathbb{R}^p, \text{ where } p \text{ is very large.}$$

1. Apply lasso to predict D by X, and collect the residual $\Rightarrow \hat{V}$
2. Apply lasso to predict Y by X, and collect the residual $\Rightarrow \hat{W}$
3. Regress $\hat{W}$ on $\hat{V}$ $\Rightarrow$ DML estimator $\hat{\theta}_0$

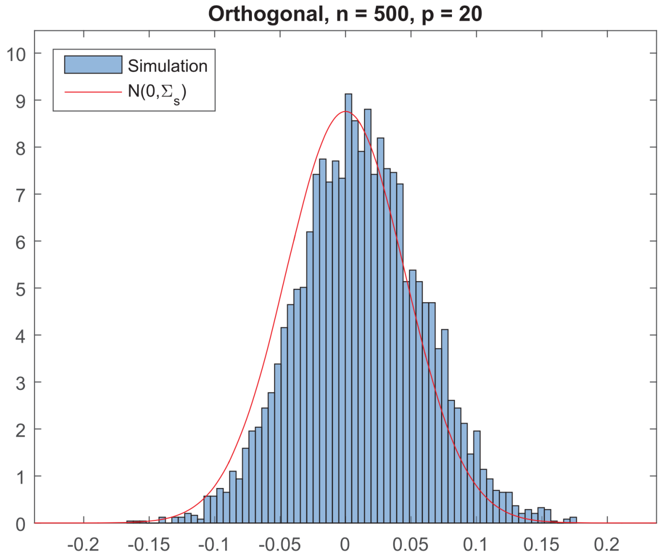
Double/Debiased Machine Learning Estimator

Consider a more general situation:

$$Y = D\theta_0 + g_0(X) + U, E[U|X, D] = 0.$$

1. Apply ML to predict D by X, and collect the residual $\Rightarrow \hat{V}$
2. Apply ML to predict Y by X, and collect the residual $\Rightarrow \hat{W}$
3. Regress $\hat{W}$ on $\hat{V}$ $\Rightarrow$ DML estimator $\hat{\theta}_0$

Double/Debiased Machine Learning Estimator



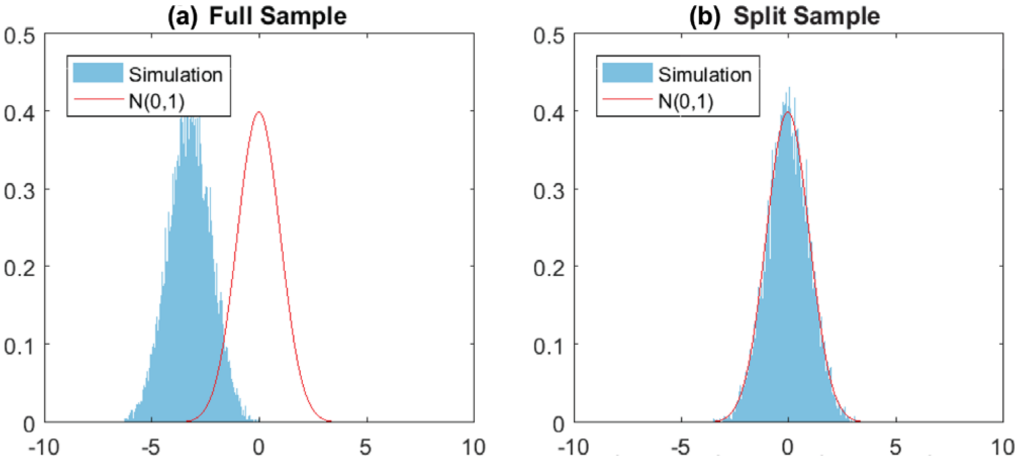
Split Sample

We need to use independent sample sets for implementing

1. Estimation for residuals $\hat{V}$ and $\hat{W}$
2. Regression $\hat{W}$ on $\hat{V}$

to get consistency.

Split Sample



Orthogonality: Non-linear Moment Condition

- General non-linear moment condition:

$$E[\psi(W;\theta_0,\eta_0)] = 0.$$

- In previous examples, $W = (Y, D, X)$ and $\eta_0 = (\beta_0, \gamma_0)$ or (g_0, m_0).
- Orthogonality condition:

$$\partial_\eta E[\psi(W;\theta_0,\eta_0)] = 0.$$

Example: ATE

- Consider

$$\begin{cases} Y = g_0(D, X) + U, & E[U|X, D] = 0 \\ D = m_0(X) + V, & E[V|X] = 0 \end{cases}.$$

- Want to estimate ATE:

$$\theta_0 = E[g_0(1, X) - g_0(0, X)].$$

- Score can be $\psi(W; \theta, \eta) =$

$$(g(1, X) - g(0, X)) + \frac{D(Y - g(1, X))}{m(X)} - \frac{(1 - D)(Y - g(0, X))}{1 - m(X)} - \theta,$$

where $\eta = (g, m)$.

Example: ATE

1. Apply ML to predict Y by (D, X), and get $\hat{g}(D, X)$
   - Caution: naive estimator $\hat{\theta} = \sum_{i=1}^{N} \{\hat{g}(1, x_i) - \hat{g}(0, x_i)\}$ will be biased!
2. Apply ML to predict D by X, and get $\hat{m}(X)$
3. DML estimator $\hat{\theta} =$

$$\sum_{i=1}^{N} \left\{ (\hat{g}(1, x_i) - \hat{g}(0, x_i)) + \frac{d_i(y_i - \hat{g}(1, x_i))}{\hat{m}(x_i)} - \frac{(1 - d_i)(y_i - \hat{g}(0, x_i))}{1 - \hat{m}(x_i)} \right\}$$

   is unbiased!!
   - DML estimator coincides with the "doubly robust" estimator for ATE.

Targeted minimum loss-based estimation (TMLE)

- Variables
  - Y: outcome
  - D: treatment indicator
  - X: covariates
- Observe $(y_i, d_i, x_i) \sim P_0$ i.i.d. for $i = 1, \dots, N$
- Statistical model $\mathcal{M}$ contains P_0
- Want to estimate $\Psi(P)$: e.g, ATE

$$\Psi(P) = E_P[E_P(Y|D=1, X) - E_P(Y|D=0, X)]$$

- Idea of TMLE:
  1. Estimate P_0 by MLE with a correction term $\epsilon \Rightarrow P_{\hat{\epsilon}}$
  2. Plug $P_{\hat{\epsilon}}$ into estimator of $\Psi \Rightarrow$ TMLE estimator $\hat{\Psi}(P_{\hat{\epsilon}})$
- Also use sample splitting.

Example: ATE

1. Set statistical model: $P(Y|D,X) \sim \mathcal{N}(m(D,X), \sigma^2(D,X))$
2. Obtain correction formula: $P_\epsilon(Y|D,X) \sim \mathcal{N}(m(D,X) + \epsilon h(P)(D,X), \sigma^2(D,X))$

$$\text{where } h(P)(D,X) = \left(\frac{\mathbb{1}(D=1)}{E_P[D=1|X]} - \frac{\mathbb{1}(D=0)}{E_P[D=0|X]} \right) \sigma^2(D,X)$$

   - Correction is made to satisfy the "efficient influence curve" condition
3. Estimate $m(D,X), \sigma^2(D,X), \epsilon$ with MLE by fitting $P_\epsilon(Y|D,X)$ to data $\Rightarrow \hat{m}(D,X), \hat{h}(D,X), \hat{\epsilon}$
4. $\hat{m}(D,X) + \hat{\epsilon}\hat{h}(D,X)$ is the corrected estimator for $E_P(Y|D,X)$
5. TMLE estimator

$$\hat{\Psi}(P_{\hat{\epsilon}}) = \sum_{i=1}^{N} \left[\left\{ \hat{m}(1,X) + \hat{\epsilon}\hat{h}(1,X) \right\} - \left\{ \hat{m}(0,X) + \hat{\epsilon}\hat{h}(0,X) \right\} \right]$$

Difference between TMLE and DML?

- van der Laan and Díaz argue TMLE is more general framework
  $\Rightarrow$ No, they ignore Chernozhukov et al. (2022), which provides a general framework of Chernozhukov et al. (2018)
- Both are coming from the semiparametric idea of first stage estimator can be high-dimensional when the second stage parameter of interest is low dimension.
- They are just following different contexts:
  - Biostatistics, statistical model
  - Econometrics, GMM