Double/Debiased Machine Learning

Marcelo Ortiz-Villavicencio

EMORY UNIVERSITY

April 12, 2024

Econometrics Reading Group

Roadmap

Introduction

Introduction

- This presentation is based on the paper by Chernuzhukov et al. (2018, EJ).
- The DML method is nothing more than a practical recipe (framework) that incorporates ideas from the *semiparametric econometrics* literature and prediction methods from the modern *machine learning* literature to provide methods that are rigorous for statistical inference of causal treatment effects.

Why use ML for Causal Inference?

- When we estimate causal effect in observational studies we often rely on the *selection on observables*-type assumption: $Y(1), Y(0) \perp D \mid X$
- Typically we make *strong assumptions* about the function form of our model when we condition on *confounders*.
- Under misspecification of our functional form, we will end up with *biased estimates* of treatment effect even if we believe that we are in the absence of unmeasured confounding.
- Machine learning (ML) provides a systematic way to learn the form of the conditional expectation function from the data.
- However, we cannot apply these methods right away, and we should know under what conditions they are useful for causal inference problems!

Big Picture

- Provides a *general framework* to estimate treatment effects using ML methods.
- In particular, we can use any (preferably $n^{1/4}$-consistent) ML estimator with this approach.

Remark (Main Goal)

Estimate and construct confidence intervals for a low-dimensional parameter (θ_0) in the presence of high-dimensional nuisance parameters (η_0), where the latter may be estimated with ML methods, such as random forests, boosted trees, lasso, ridge, deep and standard neural nets, xgboost, etc.

Partially Linear Model as motivating example

- ML methods are remarkably good at prediction tasks but not for causal inference.
- However, via *Orthogonalization* and *Sample Splitting* we can construct high quality point and interval estimates of causal parameters.
- Let's consider the canonical example:

$$Y = D\theta_0 + g_0(Z) + U, \quad \mathrm{E}[U \mid Z, D] = 0$$

  where Y is the outcome variable, D is treatment variable, Z is a high-dimensional vector of confounders and θ_0 is the *target parameter of interest*.
- Z are confounders in the sense that

$$D = c + m_0(Z) + V, \quad \mathrm{E}[V \mid Z] = 0$$

  where $m_0 \neq 0$, as is typically the case in observational studies.

Comparing approaches: Naive and Orthogonal

- Naive:
  - Predict Y using D and Z and obtain

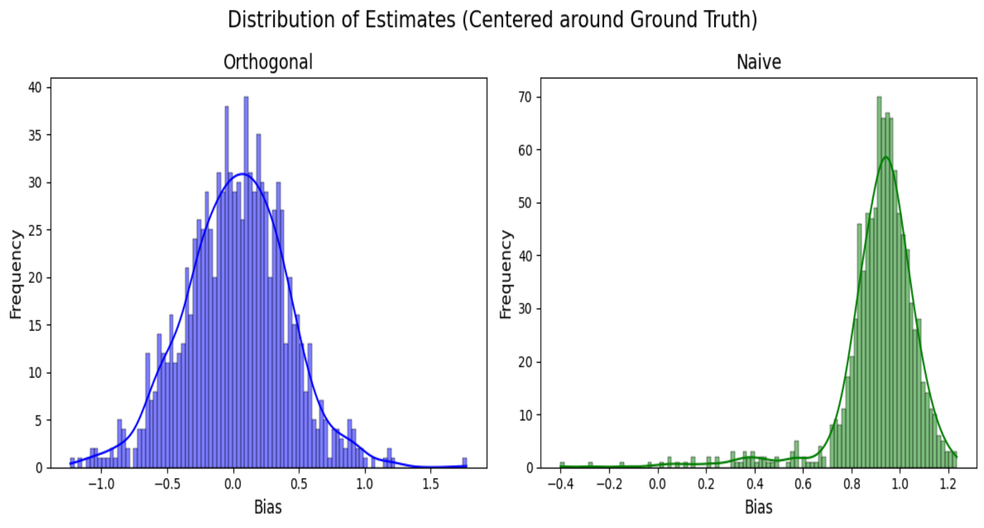
$$D\widehat{\theta}_0 + \widehat{g}_0(Z)$$

  - For example, estimate by alternating minimization: given initial guess $\widehat{\eta}_0$, run *xgboost* of $Y - D\widehat{\theta}_0$ on Z to fit $\widehat{g}_0(Z)$ and the *OLS* on $Y - \widehat{g}_0(Z)$ on D to get updated $\widehat{\theta}_0$; Repeat until convergence.
- Orthogonal:
  - Predict Y and D using Z by

$$\widehat{\mathrm{E}[Y \mid Z]} \text{ and } \widehat{\mathrm{E}[D \mid Z]},$$

    obtained using the *xgboost* or other well performing ML algorithm.
  - Residualize $\widehat{W} = Y - \widehat{\mathrm{E}[Y \mid Z]}$ and $\widehat{V} = D - \widehat{\mathrm{E}[D \mid Z]}$
  - Regress $\widehat{W}$ on $\widehat{V}$ to get $\check{\theta}_0$ (FWL on steroids!).

Comparing results



Key I: Neyman Orthogonality

Key Ingredients I

- DML estimation and inference are built on a *method-of-moments* estimator for a *low-dimensional* target parameter θ_0, using the empirical analog of the moment condition.

$$\mathrm{E}\psi\left(W; \theta_0, \eta_0\right) = 0$$

where ψ is the score function, W denotes a data vector, and η denotes nuisances parameters with true value η_0

- The first ingredient is using a score function $\psi(\cdot)$ such that

$$\mathrm{M}(\theta, \eta) = \mathrm{E}[\psi(W; \theta, \eta)]$$

identifies θ_0 when $\eta = \eta_0$

- That is, $\mathrm{M}\left(\theta, \eta_0\right) = 0$ if and only if $\theta = \theta_0$
- and the Neyman orthogonality condition is satisfied:

$$\partial_\eta \mathrm{M}\left(\theta_0, \eta\right)\big|_{\eta=\eta_0} = 0.$$

Neyman Orthogonality

Definition (Gateux Derivative)

The derivative ∂_η denotes the *pathwise (Gateaux) derivative* operator. Formally it is defined via usual derivatives taken in various directions: Given any "admissible" direction $\Delta = \eta - \eta_0$ and scalar deviation amount *t*, we have that

$$\partial_\eta \mathrm{M}(\theta, \eta)[\Delta] := \partial_t \mathrm{M}(\theta, \eta + t\Delta)|_{t=0}.$$

The statement

$$\partial_\eta \mathrm{M}(\theta_0, \eta_0) = 0$$

means that $\partial_\eta \mathrm{M}(\theta_0, \eta_0)[\Delta] = 0$ for any admissible direction Δ. The direction Δ is admissible if $\eta_0 + t\Delta$ is in the parameter space for η for all small values of *t*.

Intuition: Heuristically, the conditions says that the moment condition remains valid under local mistakes in the nuisance function.

Neyman Orthogonality

- The two strategies rely on different moment conditions for identifying and estimating θ_0:
  1. *Naive*:= $\psi(W, \theta_0, \eta) = (Y - D\theta_0 - g_0(Z))\, D$
     with $\eta = g(Z), \quad \eta_0 = g_0(Z)$
  2. *Orthogonal* := $\psi(W, \theta_0, \eta_0) = ((Y - E[Y \mid Z]) - (D - E[D \mid Z])\theta_0)\,(D - E[D \mid Z])$
     with $\eta = (\ell(Z), m(Z)), \quad \eta_0 = (\ell_0(Z), m_0(Z)) = (\mathrm{E}[Y \mid Z], \mathrm{E}[D \mid Z])$
- The Neyman Orthogonality condition does hold for the score *Orthogonal* and fails to hold for the score *Naive*.

Consider estimation based on (2)

$$\check{\theta}_0 = \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{V}_i^2\right)^{-1}\frac{1}{n}\sum_{i=1}^{N}\widehat{V}_i\widehat{W}_i$$

where $\widehat{V} = D - \widehat{m}_0(Z), \widehat{W} = Y - \widehat{\ell}_0(Z)$,

Under mild conditions, can write

$$\sqrt{n}\left(\check{\theta}_0 - \theta_0\right) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}V_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_iU_i}_{:=a^*}$$

$$+ \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}V_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m_0\left(Z_i\right) - \hat{m}_0\left(Z_i\right)\right)\left(\ell_0\left(Z_i\right) - \widehat{\ell}_0\left(Z_i\right)\right)}_{:=b^*}$$

$$+ o_p(1)$$

Converging Properties

- $a^* \rightsquigarrow N(0, \Sigma)$ under standard conditions
- b^* now depends on product of estimation errors in both nuisance functions
- b^* will look like $\sqrt{n}n^{-(\varphi_m+\varphi_\ell)}$ where $n^{-\varphi_m}$ and $n^{-\varphi_\ell}$ are respectively appropriate convergence rates of estimators for $m(z)$ and $\ell(z)$
- $o\left(n^{-1/4}\right)$ is often an attainable rate for estimating $m(z)$ and $\ell(z)$

Remark

A key input is the use of high-quality machine learning estimators of the nuisance parameters. A sufficient condition in the examples given includes the requirement

$$n^{1/4} \|\hat{\eta} - \eta_0\|_{L^2} \approx 0$$

- Fortunately, there are performance guarantees for most of these ML methods that make it possible to satisfy the conditions stated above.

Key II: Sample Splitting

Key Ingredients II

- The second key ingredient is to use a form of *sample splitting* at the stage of producing the estimator of the main parameter θ_0, which allows to avoid biases arising from **overfitting**.
- Technically, we rely on *sample splitting* to get the third term of the DML estimator to be $o_p(1)$ with only the rate restriction of $o(n^{-1/4})$ on the performance of the ML algorithm.
- This eliminates conditions on the *entropic complexity* of the realization of ML estimators (very difficult to check in practice).

Sample Splitting

- In the expansion $\sqrt{n}\left(\check{\theta}_0 - \theta_0\right) = a^* + b^* + o_p(1)$ the term $o_p(1)$ contains terms like

$$\left(\frac{1}{n}\sum_{i=1}^{n} V_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\left(m_0\left(Z_i\right) - \hat{m}\left(Z_i\right)\right)$$

- With sample splitting, easy to control and claim $o_p(1)$.
- Without sample splitting, it is difficult to control and claim $o_p(1)$.

Remark

Without sample splitting, need maximal inequalities to control

$$\sup_{m \in \mathcal{M}_n} \left| \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\left(m_0\left(Z_i\right) - m\left(Z_i\right)\right)\right|$$

where $\mathcal{M}_n \ni \widehat{m}$ with probability going to 1, and need to control the entropy of $\mathcal{M}_n$, which typically grows in modern high-dimensional applications. In particular, the assumption that $\mathcal{M}_n$ is P -Donsker used in semiparametric literature does not apply.

General Results from Moment Condition Models

General Results: Neyman Orthogonality

Moment conditions model:

$$\mathrm{E}\left[\psi_j\left(W, \theta_0, \eta_0\right)\right] = 0, \quad j = 1, \ldots, d_\theta$$

- $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ is a vector of known score functions
- W is a random element; we observe random sample $(W_i)_{i=1}^N$ from the distribution of W
- θ_0 is the low-dimensional parameter of interest
- η_0 is the true value of the nuisance parameter $\eta \in T$ for some convex set T equipped with a norm $\|\cdot\|_e$ (can be a function or vector of functions)

General Results: Neyman Orthogonality

Key orthogonality condition: $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ obeys the orthogonality condition with respect to $\mathcal{T} \subset T$ if the *Gateaux derivative* map

$$\mathrm{D}_{r,j}[\eta - \eta_0] := \partial_r \left\{ \mathrm{E}_P \left[\psi_j \left(W, \theta_0, \eta_0 + r(\eta - \eta_0) \right) \right] \right\}$$

- exists for all $r \in [0, 1), \eta \in \mathcal{T}$, and $j = 1, \ldots, d_\theta$
- vanishes at $r = 0$: For all $\eta \in \mathcal{T}$ and $j = 1, \ldots, d_\theta$,

$$\partial_\eta \mathrm{E}_P \psi_j (W, \theta_0, \eta) \big|_{\eta = \eta_0} [\eta - \eta_0] := \mathrm{D}_{0,j}[\eta - \eta_0] = 0$$

General Results: Sample Splitting

Results will make use of sample splitting:

- $\{1, \ldots, N\}$ = set of all observation names;
- I = **main sample** = set of observation numbers, of size n, is used to estimate θ_0
- I^c = **auxilliary sample** = set of observations, of size $\pi n = N - n$, is used to estimate η_0;
- I and I^c form a random partition of the set $\{1, \ldots, N\}$

Main Theoretical Result

Under regularity conditions (See paper), let Double ML estimator

$$\check{\theta}_0 = \check{\theta}_0 \left(I, I^c\right)$$

be such that

$$\left\| \frac{1}{n} \sum_{i \in I} \psi \left(W, \check{\theta}_0, \widehat{\eta}_0\right) \right\| \leqslant \epsilon_n, \quad \epsilon_n = o\left(\delta_n n^{-1/2}\right)$$

Theorem

$\check{\theta}_0$ obeys

$$\sqrt{n} \Sigma_0^{-1/2} \left(\check{\theta}_0 - \theta_0\right) = \frac{1}{\sqrt{n}} \sum_{i \in I} \bar{\psi} \left(W_i\right) + O_P\left(\delta_n\right) \rightsquigarrow N(0, I),$$

uniformly over $P \in \mathcal{P}_n$, where $\bar{\psi}(\cdot) := -\Sigma_0^{-1/2} J_0^{-1} \psi\left(\cdot, \theta_0, \eta_0\right)$ and $\Sigma_0 := J_0^{-1} \mathrm{E}_P\left[\psi^2\left(W, \theta_0, \eta_0\right)\right]\left(J_0^{-1}\right)'$ and $J_0 := \partial_{\theta'} \left\{\mathrm{E}_P\left[\psi\left(W, \theta, \eta_0\right)\right]\right\}\big|_{\theta=\theta_0}$.

EMORY UNIVERSITY

Corollary (2-fold cross-validation)

Can do a random 2-way split with $\pi = 1$*, obtain estimates* $\check{\theta}_0 \left(I, I^c\right)$ *and* $\check{\theta}_0 \left(I^c, I\right)$ *and average them*

$$\check{\theta}_0 = \frac{1}{2}\check{\theta}_0 \left(I, I^c\right) + \frac{1}{2}\check{\theta}_0 \left(I^c, I\right)$$

to gain full efficiency.

Corollary (k-fold cross-validation)

Can do also a random K-way split $(I_1, \ldots, I_K)$ *of* $\{1, \ldots, N\}$*, so that* $\pi = (K - 1)$*, obtain estimates* $\check{\theta}_0 \left(I_k, I_k^c\right)$*, for* $k = 1, \ldots, K$*, and average them*

$$\check{\theta} = \frac{1}{K}\sum_{k=1}^{K} \check{\theta}_0 \left(I_k, I_k^c\right)$$

to gain full efficiency.

The DML inference method

1. **Inputs:** Provide the data frame $(W_i)_{i=1}^n$, the Neyman orthogonal score/moment function $\psi(W, \theta, \eta)$ that identifies the statistical parameter of interest, and the name and model for ML estimation method(s) for η.
2. **Train ML Predictors on Folds:** Take a *K-fold random partition* $(I_k)_{k=1}^K$ of observation indices $\{1, \ldots, n\}$ such that the size of each fold is about the same. For each $k \in \{1, \ldots, K\}$, construct a high-quality machine learning estimator $\hat{\eta}_{[k]}$ that depends only on a subset of data $(X_i)_{i \notin I_k}$ that excludes the k-th fold.
3. **Estimate Moments:** Letting $k(i) = \{k : i \in I_k\}$, construct the moment equation estimate

$$\hat{\mathrm{M}}(\theta, \hat{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \psi\left(W_i; \theta, \hat{\eta}_{[k(i)]}\right)$$

The DML inference method

4. **Compute the Estimator:** Set the estimator $\hat{\theta}$ as the solution to the equation.

$$\hat{M}(\hat{\theta}, \hat{\eta}) = 0.$$

5. **Estimate Its Variance:** Estimate the *asymptotic variance* of $\hat{\theta}$ by

$$\hat{\mathrm{V}} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\varphi}(W_i) \, \hat{\varphi}(W_i)' \right] - \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\varphi}(W_i) \right] \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\varphi}(W_i) \right]',$$

where

$$\hat{\varphi}(W_i) = -\hat{J}_0^{-1} \psi \left(W_i; \hat{\theta}, \hat{\eta}_{[k(i)]} \right)$$

and

$$\hat{J}_0 := \partial_\theta \frac{1}{n} \sum_{i=1}^{n} \psi \left(W_i; \hat{\theta}, \hat{\eta}_{[k(i)]} \right).$$

The DML inference method

6. **Confidence Intervals:** Form an approximate $(1-\alpha)\%$ *confidence interval* for any functional $\ell'\theta_0$, where ℓ is a vector of constants, as

$$\left[\ell'\hat{\theta} \pm c\sqrt{\ell'\hat{\mathrm{V}}\ell/n}\right]$$

where c is the $(1-\alpha/2)$ quantile of $N(0,1)$.

Application: Debiased machine learning of conditional average treatment effects and other causal functions

Big Picture

- Semenova & Chernozhukov (2021, EJ) provides *estimation* and *inference* methods on a nonparametric function $g(x)$ that summarizes *heterogeneous/causal/structural effects* conditional on a *small set* of covariates X.
- Represent this structural function as a *conditional expectation* of an *unbiased signal* that depends on a nuisance parameter estimated by ML methods.
- Other papers study a specific feature of CATE. This paper operates in a classical *observational setting*, with many potential controls, and targets the *true* CATE function.
- Procedure:
  1. Adjust the signal to make it *Neyman-orthogonal* with respect to the *first-stage regularization bias*.
  2. Project the signal onto a set of basis functions to get the *best linear predictor* of the structural function.
  3. Simultaneous inference on all parameters of the best linear predictor by *Gaussian bootstrap*.

Setup

- Consider a function $g(x)$ which can be represented as a *conditional expectation function*

$$g(x) = \mathbb{E}\left[Y(\eta_0) \mid X = x\right]$$

  where $Y(\eta_0)$ is refer as *signal*, and depends on a *nuisance function* $\eta_0(z)$ of a control vector Z.
- Examples of signals include the *Conditional Average Treatment Effect* (CATE), *Continuous Treatment Effects* (CTEs), etc.
- Examples of nuisance functions include the *propensity score*, the *conditional density*, and the *regression function*, among others.
- Keep in mind: *dim*(Z) is high; *dim*(X) is low.

Orthogonality Property

- Focus on signals $Y(\eta_0)$ that have the orthogonality property.
- Formally, we require the *pathwise derivative* of the conditional expectation to be *zero conditional* on X:

$$\partial_r \mathbb{E}\left[Y(\eta_0 + r(\eta - \eta_0)) \mid X = x\right]\big|_{r=0} = 0, \quad \text{for all } x \text{ and } \eta$$

- If the signal $Y(\eta)$ is *orthogonal*, its plug-in estimate $Y(\widehat{\eta})$ is *insensitive to bias* in the estimation of $\widehat{\eta}$ (i.e., regularization bias), which results from applying ML methods in high dimensions.
- Under mild conditions, $Y(\widehat{\eta})$ delivers a high-quality estimator of the target function $g(x)$.

Orthogonality Property for CTEs

- Let $X \in \mathbb{R}$ be a one-dimensional *continuous treatment*.
- Let Y^x be the potential outcome corresponding to the subject's response after receiving x units of treatment
- $V = (X, Z, Y)$ consists of the treatment X, the control vector Z, and the observed outcome $Y = Y^X$.
- If potential outcomes $\{Y^x, x \in \mathbb{R}\}$ are *independent of treatment* X *conditional on controls* Z, the average potential outcome is identified as

$$\mathbb{E}\left[Y^x\right] = \mathbb{E}\mu_0(x, Z) = \int \mu_0(x, z) dP_Z(z),$$

  where $\mu_0(x, z) = \mathbb{E}[Y \mid X = x, Z = z]$ is the regression function of the observed outcome.
- Since Z is *high dimensional*, it is necessary to estimate the regression function $\mu_0(x, z)$ with some *regularized* technique to achieve convergence.

Naive approach: problematic?

- To estimate $\mathbb{E}[Y^x]$ we can consider the sample analog

$$\widetilde{g}(x) = \int \widehat{\mu}(x,z)d\widehat{P}_Z(z),$$

where $\widehat{\mu}(x,z)$ is a *regularized estimate* of $\mu_0(x,Z)$, and $\widehat{P}_Z(z)$ the empirical analog of P_Z

- Problem: This approach results in a *biased estimate*, and the *bias of estimation error* $\widehat{\mu}(x,Z) - \mu_0(x,Z)$ does not vanish faster than $N^{1/2}$
- The plug-in estimator inherits this first-order bias because the moment equation is *not orthogonal to perturbations* of μ
- This bias implies that the plug-in estimator $\widetilde{g}(x)$ *will not converge* at the optimal rate.

First stage: Doubly Robust Signal

- Let $g(x) = \mathbb{E}[Y^x]$.
- We choose $Y(\eta)$ to be a *doubly robust signal* in the sense of Kennedy et. al. (2017)

$$Y(\eta) := \frac{Y - \mu(X, Z)}{s(X \mid Z)} w(X) + \int \mu(X, z) dP_Z(z)$$

- Nuisance parameter

$$\eta_0(x, z) = \left\{ s_0(x \mid z), \mu_0(x, z), w_0(x) \right\}$$

consist in the regression function, conditional density of $X \mid Z$, and marginal treatment density.

First stage: Doubly Robust Signal

- The previous procedure is more costly because the nuisance parameter includes two more functions: $s_0(x \mid z)$, and $w_0(x)$
- However the *signal* is *conditional orthogonal* with respect to each nuisance function in $\eta_0(x, z)$

$$\mathbb{E}\left[\begin{matrix} -\int_{z\in\mathcal{Z}}(\mu(X,z)-\mu_0(X,z))\,dP_Z(z)+\int_{z\in\mathcal{Z}}(\mu(x,z)-\mu_0(x,z))\,dP_Z(z) \\ \frac{\mu_0(X,Z)-Y}{s_0^2(X|Z)}(s(X\mid Z)-s_0(X\mid Z)) \\ \frac{Y-\mu_0(X,Z)}{s_0(X|Z)}(W(X)-w_0(X)) \end{matrix} \;\middle|\; X=x\right]=0$$

- This guarantees the *bias of the estimation error* $\widehat{\eta}(x, Z) - \eta_0(x, Z)$ **does not** create *first-order bias* in the estimated signal $Y(\hat{\eta})$ and affects only its higher-order bias.
- Therefore, the estimate of the target function based on $Y(\hat{\eta})$ is *high quality* under plausible conditions.

Second stage: Linear projection onto basis function

- Consider a *linear projection* of an orthogonal signal $Y(\eta)$ onto a vector of *basis functions* $p(X)$,

$$\beta := \arg\min_{b \in \mathbb{R}^d} \mathbb{E}\left(Y(\eta) - p(X)'b\right)^2 .$$

- The choice of basis functions depends on the *desired interpretation* of the linear approximation.

Example

Consider partitioning the support of X into d *mutually exclusive groups* $\{G_k\}_{k=1}^d$. Setting

$$p_k(x) = \mathbf{1}\{x \in G_k\}, \quad k \in \{1, 2, \ldots, d\}$$

implies that $p(x)'\beta_0$ is a *group average treatment effect* for group k such that $x \in G_k$.

- Our inference will target this parameter, allowing the *number of groups to increase* at some rate.

Example: CTE

- Let $X \in \mathbb{R}$ be a *continuous treatment variable*, Z be a vector of the *controls*, Y^x stand for the *potential outcomes* corresponding to the subject's response after receiving x units of treatment. $Y = Y^X$ be the *observed outcome*.
- For a given value x, the target function is the average potential outcome

$$g(x) = \mathbb{E}\left[Y^x\right]$$

- Unconfoundedness: Suppose all of the potential outcomes $\{Y^x, x \in \mathbb{R}\}$ are independent of $X \mid Z$

$$\{Y^x, x \in \mathbb{R}\} \perp X \mid Z.$$

- Then $g(x)$ is identified as

$$g(x) = \mathbb{E}\mu_0(x, Z)$$

- Doubly Robust signal is *conditionally orthogonal* with respect to the nuisance parameter consisting of the *generalized propensity score*, *regression function* of Y on X, Z, and the *marginal treatment density*.

Example: CATE

- Let Y_1 and Y_0 be the potential outcomes
- Let $D = 1$ be a dummy for whether a subject is treated.
- The object of interest is the CATE

$$g(x) := \mathbb{E}\left[Y_1 - Y_0 \mid X = x\right]$$

- Unconfoundedness: $Y_1, Y_0 \perp D \mid Z$
- One can define a *orthogonal signal* with respect to the nuisance parameter $\eta_0(z) := \{s_0(z), \mu_0(1, z), \mu_0(0, z)\}$ such that

$$Y(\eta) := \mu(1, Z) - \mu(0, Z) + \frac{D[Y - \mu(1, Z)]}{s(Z)} - \frac{(1 - D)[Y - \mu(0, Z)]}{1 - s(Z)}$$

Example: Conditional Average Partial Derivative

EMORY UNIVERSITY

- Let $D \in \mathbb{R}$ be a *continuous treatment variable*, Z be a vector of the *controls*, Y^d stand for the *potential outcomes* after receiving d units of treatment and X be a *subvector of controls* Z.
- The target function is the *average partial derivative* conditional on a covariate vector X

$$g(x) = \partial_d \mathbb{E}\left[Y^D \mid X = x\right].$$

- Unconfoundedness: $\left\{Y^d, d \in \mathbb{R}\right\} \perp D \mid Z$
- $g(x)$ is identified as

$$g(x) = \mathbb{E}[\partial_d \underbrace{\mu_0(D, Z)}_{\mathbb{E}[Y \mid D = d, Z = z]} \mid X = x]$$

- Using the following signal *orthogonal* with respect to the nuisance parameter $\eta_0(d, z) = \{\mu_0(d, z), s_0(d \mid z)\}$:

$$Y(\eta) := -\partial_d \log s(D \mid Z)[Y - \mu(D, Z)] + \partial_d \mu(D, Z)$$

Orthogonal estimator: Two-stages

EMORY UNIVERSITY

1. Construct an estimate $\hat{\eta}$ of the nuisance parameter $\hat{\eta}_0$, using an ML model capable of dealing with the high-dimensional covariate vector Z.
2. Construct $\hat{Y}_i := Y_i(\hat{\eta})$ and run OLS of $\hat{Y}_i$ on $p(X_i)$.

Definition (Cross-fitting)

(1) For a random sample of size N, denote a K-fold random partition of the sample indices $[N] = \{1, 2, \ldots, N\}$ by $(J_k)_{k=1}^{K}$, where K is the number of partitions, and the sample size of each fold is $n = N/K$. For each $k \in [K] = \{1, 2, \ldots, K\}$ define $J_k^c = \{1, 2, \ldots, N\} \setminus J_k$.
(2) For each $k \in [K]$, construct an estimator $\widehat{\eta}_k = \widehat{\eta}\left(V_{i \in J_k^c}\right)$ of the nuisance parameter η_0 by using only the data $\left\{V_j : j \in J_k^c\right\}$. For any observation $i \in J_k$, define $\widehat{Y}_i := Y_i\left(\widehat{\eta}_k\right)$.

Definition (Orthogonal Estimator)

Given $\left(\widehat{Y}_i\right)_{i=1}^{N}$, define $\widehat{\beta} := \left(\frac{1}{N}\sum_{i=1}^{N} p\left(X_i\right) p\left(X_i\right)'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} p\left(X_i\right)\widehat{Y}_i$

Thanks!

marcelo.ortiz@emory.edu
marcelortiz.com
@marcelortizv

References

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters*. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097
- Semenova, V., & Chernozhukov, V. (2021). *Debiased machine learning of conditional average treatment effects and other causal functions*. The Econometrics Journal, 24(2), 264-289. https://doi.org/10.1093/ectj/utaa027