Double/Debiased Machine Learning for Causal and Treatment Effects

July 16, 2017

This presentation is based on:

- "Double/De-biased Machine Learning for Causal and Treatment Effects"

  ArXiv 2016, with **Denis Chetverikov, Esther Duflo, Christian Hansen, Mert Demirer, Whitney Newey, James Robins**

Introduction

- Main goal: Estimate and construct confidence intervals for a low-dimensional parameter (θ_0) in the presence of high-dimensional nuisance parameter (η_0), where the latter may be estimated with the new generation of nonparametric statistical methods, branded as “machine learning” (ML) methods, such as
  - random forests,
  - boosted trees,
  - lasso,
  - ridge,
  - deep and standard neural nets,
  - gradient boosting,
  - their aggregations,
  - and cross-hybrids.

Introduction

- We build upon/extend the classic work in semi-parametric estimation which focused on "traditional" nonparametric methods for estimating η_0, e.g. Bickel, Klassen, Ritov, Wellner (1998), Andrews (1994), Linton (1996), Newey (1990, 1994), Robins and Rotnitzky (1995), Robinson (1988), Van der Vaart (1991), Van der Laan and Rubin (2008), many others.
- Theoretical analyses required the estimators $\widehat{\eta}$ of η_0 to take values in an entropically simple set – a Donsker set – which really rules out most of the new methods in the *high-dimensional* setting.

Literature

- Lots of recent work on inference based on lasso-type methods for estimating η_0
- Relatively little work on the use other ML methods in high-dimensional setting.

Two main points:

I. The ML methods seem remarkably effective in prediction contexts. However, good performance in prediction **does not necessarily translate** into good performance for estimation or inference about "causal" parameters. In fact, the performance **can be poor**.

Two main points:

I. The ML methods seem remarkably effective in prediction contexts. However, good performance in prediction **does not necessarily translate** into good performance for estimation or inference about “causal” parameters. In fact, the performance **can be poor**.

II. By doing **”double/di-biased” ML or “orthogonalized”** ML, and sample splitting, we can construct high quality point and interval estimates of ”causal” parameters.

Main Points via a Partially Linear Model

Illustrate the two main points in a canonical example:

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

- Y - outcome variable
- D - policy/treatment variable
- Z is a high-dimensional vector of other covariates, called "controls" or "confounders"
- θ_0 is the target parameter of interest

Z are confounders in the sense that

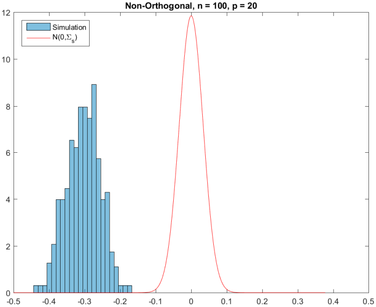
$$D = c + m_0(Z) + V, \quad \mathrm{E}[V \mid Z] = 0$$

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

Causal interpretation of θ_0: under conditional exogeneity/conditional random assignment of D given Z, θ_0 is the average causal effect of D on potential outcome.

Point I. “Naive” or Prediction-Based ML Approach is Bad

- 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 Random Forest of $Y - D\widehat{\theta}_0$ on Z to fit $\widehat{g}_0(Z)$ and the Ordinary Least Squares on $Y - \widehat{g}_0(Z)$ on D to get updated $\widehat{\theta}_0$; Repeat until convergence.
- Excellent prediction performance! BUT the distribution of $\widehat{\theta}_0 - \theta_0$ looks like this:



Point II. The “Double” ML Approach is Good

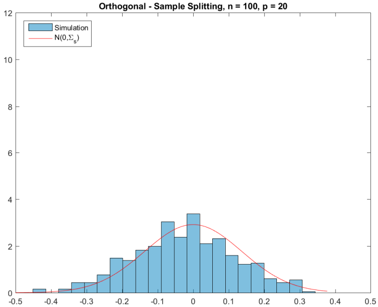
1. Predict Y and D using Z by

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

obtained using the Random Forest or other "best performing ML" tools.

2. Residualize $\widehat{W} = Y - \widehat{\mathrm{E}[Y|Z]}$ and $\widehat{V} = D - \widehat{\mathrm{E}[D|Z]}$
3. Regress $\widehat{W}$ on $\widehat{V}$ to get $\check{\theta}_0$.

- Frisch-Waugh-Lovell (1930s) style. The distribution of $\check{\theta}_0 - \theta_0$ looks like this:



Moment conditions

The two strategies rely on very different moment conditions for identifying and estimating θ_0:

$$\mathrm{E}[\psi(W, \theta_0, \eta_0)] = 0$$

$$\psi(W, \theta_0, \eta) = (Y - D\theta_0 - g_0(Z))D \tag{1}$$

$$\psi(W, \theta_0, \eta_0) = ((Y - E[Y|Z]) - (D - E[D|Z])\theta_0)(D - E[D|Z]) \tag{2}$$

- (1) - Regression adjustment score, with

$$\eta = g(Z), \quad \eta_0 = g_0(Z),$$

- (2) - Neyman-orthogonal score (Frisch-Waugh-Lovell), with

$$\eta = (\ell(Z), m(Z)), \quad \eta_0 = (\ell_0(Z), m_0(Z)) = (\mathrm{E}[Y \mid Z], \mathrm{E}[D \mid Z])$$

Both estimators solve the empirical analog of the moment conditions:

$$\frac{1}{n} \sum_{i=1}^{n} \psi(W_i, \theta, \widehat{\eta}_0) = 0,$$

where instead of unknown nuisance functions we plug-in their ML-based estimators, obtained using auxiliary (set-aside) sample.

Key Difference between (1) and (2) is Neyman Orthogonality

- The Neyman orthogonality condition:

$$\mathrm{D} = \partial_\eta \mathrm{E} \psi(W, \theta_0, \eta)|_{\eta=\eta_0} = \mathbf{0}$$

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

Key Difference between (1) and (2) is Neyman Orthogonality

- The Neyman orthogonality condition:

$$\mathrm{D} = \partial_\eta \mathrm{E} \psi(W, \theta_0, \eta)|_{\eta=\eta_0} = \mathbf{0}$$

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

- The condition *does hold* for the score (2) and *fails to hold* for the score (1),

Heuristics: The Role of Neyman Orthogonality

- We have expansion

$$J\sqrt{n}(\widehat{\theta}-\theta_0) = A_n + \sqrt{n}\mathrm{D}(\widehat{\eta}-\eta_0) + C\sqrt{n}O(\|\widehat{\eta}-\eta_0\|^2) + o_p(1),$$

  where the leading term A_n is well-behaved and approximately Gaussian under weak conditions, if sample-splitting is used and $\|\widehat{\eta}-\eta_0\| \to 0$.

Heuristics: The Role of Neyman Orthogonality

- We have expansion

$$J\sqrt{n}(\widehat{\theta} - \theta_0) = A_n + \sqrt{n}\mathrm{D}(\widehat{\eta} - \eta_0) + C\sqrt{n}O(\|\widehat{\eta} - \eta_0\|^2) + o_p(1),$$

  where the leading term A_n is well-behaved and approximately Gaussian under weak conditions, if sample-splitting is used and $\|\widehat{\eta} - \eta_0\| \to 0$.

- When $\mathrm{D} \neq 0$, since $\|\widehat{\eta} - \eta_0\| = O_P(n^{-\varphi})$, $0 < \varphi < 1/2$,

$$\sqrt{n}\mathrm{D}(\widehat{\eta} - \eta_0) \text{ is of order } \sqrt{n}n^{-\varphi} \to \infty.$$

  and the estimator without Neyman orthogonality is not root-n consistent.

Heuristics: The Role of Neyman Orthogonality?

- Under Neyman orthogonality $\mathrm{D} = 0$, then

  $$\sqrt{n}\mathrm{D}(\widehat{\eta} - \eta) = 0,$$

  and for root-n consistency we only need,

  $$C\sqrt{n}O(\|\widehat{\eta} - \eta_0\|^2) \to 0,$$

  which requires $\|\widehat{\eta} - \eta_0\| = o_P(n^{-1/4})$ if $C \gg 0$.

Heuristics: The Role of Neyman Orthogonality?

- Under Neyman orthogonality $\mathrm{D} = 0$, then

  $$\sqrt{n}\mathrm{D}(\widehat{\eta} - \eta) = 0,$$

  and for root-n consistency we only need,

  $$C\sqrt{n}O(\|\widehat{\eta} - \eta_0\|^2) \to 0,$$

  which requires $\|\widehat{\eta} - \eta_0\| = o_P(n^{-1/4})$ if $C \gg 0$.
- This is attainable rate for many ML estimators, especially aggregated estimators.
- In some problems $C = 0$, like optimal IV problem in Belloni et al (2010) or when $m_0 = 0$ (as in the randomized control trials).
- In the partially linear model, the rate condition is finer, just requiring the product of rates to me of order $o(1/\sqrt{n})$.

Heuristics: The Role of Sample Splitting

- Need to show
$$A_n = \mathbb{G}_n(\psi(W, \theta_0, \widehat{\eta})) \rightsquigarrow N(0, \Omega),$$
where $\mathbb{G}_n$ is the empirical process:
$\mathbb{G}_n(f) = n^{-1/2} \sum_{i=1}^n (f(W_i) - \int f(w) dP(w))$.
- So we need
$$\mathbb{G}_n(\psi(W, \theta_0, \widehat{\eta}) - \mathbb{G}_n(\psi(W, \theta_0, \eta_0) \to_P 0.$$
- If $\widehat{\eta}$ is based on the auxiliary sample, not used in the main estimation, then this follows from $\|\widehat{\eta} - \eta_0\| \to 0$ and Chebyshev inequality.
- If $\widehat{\eta}$ is based on the main sample, need maximal inequalities to control
$$\sup_{\eta \in \mathcal{M}_n} \left| \mathbb{G}_n(\psi(W, \theta_0, \eta) - \mathbb{G}_n(\psi(W, \theta_0, \eta_0) \right|$$
We need to control the rate of entropy growth for $\mathcal{M}_n \ni \widehat{\eta}$...
- See our "Program Evaluation Paper.." in Econometrica for the rates at which entropy can grow. The condition is reasonable, but it might be hard to check for each new ML method...

General Results for Moment Condition Models

Moment conditions model:

$$\mathrm{E}[\psi_j(W, \theta_0, \eta_0)] = 0, \quad j = 1, \ldots, d_\theta \tag{3}$$

- $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ is a vector of known score functions
- W is a random element; 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)

Key Ingredient I: Neyman Orthogonality Condition

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

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

- 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 $j = 1, \ldots, d_\theta$,

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

Heuristically, small deviations in nuisance functions do not invalidate moment conditions.

How to Builds Orthogonal Scores

Can generally construct moment/score functions with desired orthogonality property building upon classic ideas of Neyman (1958, 1979)

Neyman's construction in parametric likelihood case.

Suppose log-likelihood function is given by $\ell(W, \theta, \beta)$

- θ d-dimensional parameter of interest
- β p_0-dimensional nuisance parameter

Under regularity, true parameter values satisfy

$$\mathrm{E}[\partial_\theta \ell(W, \theta_0, \beta_0)] = 0, \quad \mathrm{E}[\partial_\beta \ell(W, \theta_0, \beta_0)] = 0$$

$\varphi(W, \theta, \beta) = \partial_\theta \ell(W, \theta, \beta)$ in general does not possess the orthogonality property

How to Builds Orthogonal Scores: in Parametric Likelihood Model

Can construct new estimating equation with desired orthogonality property:

$$\psi(W, \theta, \eta) = \partial_\theta \ell(W, \theta, \beta) - \mu \partial_\beta \ell(W, \theta, \beta),$$

- Nuisance parameter: $\eta = (\beta', \mathrm{vec}(\mu)')' \in T \times \mathcal{D} \subset \mathbb{R}^p, \quad p = p_0 + dp_0$
- μ is the $d \times p_0$ orthogonalization parameter matrix. True value (μ_0) is chosen such that

  $$J_{\theta\beta} - \mu J_{\beta\beta} = 0 \text{ (i.e., } \mu_0 = J_{\theta\beta} J_{\beta\beta}^{-1}\text{)}$$

  for the Hessian (Information Matrix):

  $$J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \partial_{(\theta', \beta')} \mathrm{E}\Big[\partial_{(\theta', \beta')'} \ell(W, \theta, \beta)\Big]\Big|_{\theta=\theta_0;\ \beta=\beta_0}$$
- Will have $\mathrm{E}[\psi(W, \theta_0, \eta_0)] = 0$ for $\eta_0 = (\beta_0', \mathrm{vec}(\mu_0)')'$ (provided μ_0 is well-defined)
- Importantly, ψ obeys the orthogonality condition: $\partial_\eta \mathrm{E}[\psi(W, \theta_0, \eta)]\Big|_{\eta=\eta_0} = 0$
- ψ is the efficient score for inference about θ_0

How to Builds Orthogonal Scores: in Moment Conditions Models

More generally, can construct orthogonal estimating equations as in the semiparametric estimation literature.

One key approach is to project the initial score/moment function onto orthocomplement of tangent space induced by nuisance function

- E.g. Chamberlain (1992), van der Vaart (1998), van der Vaart and Wellner (1996))

Many worked out examples, some follow later in the talk.

Orthogonal scores/moment functions will often have nuisance parameter η that is of higher dimension than "original" nuisance function β.

- Also see in partially linear model where nuisance parameter in orthogonal moment conditions involve two conditional expectations

Key Ingredient II: Sample Splitting

Results will make use of sample splitting:

- $\{1, ..., N\}$ = 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 $\{1, ..., N\}$

Use of sample splitting allows to get rid of "entropic" requirements and boil down requirements on ML estimators $\widehat{\eta}$ of η_0 to just rates.

Theory: Regularity Conditions for General Framework

Denote

$$J_0 := \partial_{\theta'} \Big\{ \mathrm{E}_P[\psi(W, \theta, \eta_0)] \Big\} \Big|_{\theta=\theta_0}$$

Let ω, c_0, and C_0 be strictly positive (and finite) constants, $n_0 \geqslant 3$ be a positive integer, and $(B_{1n})_{n \geqslant 1}$ and $(B_{2n})_{n \geqslant 1}$ be sequences of positive constants, possibly growing to infinity, with $B_{1n} \geqslant 1$ for all $n \geqslant 1$.

Assume for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$

- (Parameter not on boundary) θ_0 satisfies (3), and Θ contains a ball of radius $C_0 n^{-1/2} \log n$ centered at θ_0
- (Differentiability) The map $(\theta, \eta) \mapsto \mathrm{E}_P[\psi(W, \theta, \eta)]$ is twice continuously Gateaux-differentiable on $\Theta \times \mathcal{T}$
  - Does not require ψ to be differentiable
- (Neyman Orthogonality) ψ obeys the orthogonality condition for the set $\mathcal{T} \subset T$

Theory: Regularity Conditions on Model (Continued)

- (Identifiability) For all $\theta \in \Theta$, we have $\|\mathrm{E}_P[\psi(W, \theta, \eta_0)]\| \geqslant 2^{-1}\|J_0(\theta - \theta_0)\| \wedge c_0$ where the singular values of J_0 are between c_0 and C_0
- (Mild Smoothness) For all $r \in [0, 1)$, $\theta \in \Theta$, and $\eta \in \mathcal{T}$
  - $\mathrm{E}_P[\|\psi(W, \theta, \eta) - \psi(W, \theta_0, \eta_0)\|^2] \leqslant C_0(\|\theta - \theta_0\| \vee \|\eta - \eta_0\|_e)^\omega$
  - $\|\partial_r \mathrm{E}_P\left[\psi(W, \theta, \eta_0 + r(\eta - \eta_0))\right]\| \leqslant B_{1n}\|\eta - \eta_0\|_e$
  - $\|\partial_r^2 \mathrm{E}_P[\psi(W, \theta_0 + r(\theta - \theta_0), \eta_0 + r(\eta - \eta_0))]\| \leqslant B_{2n}(\|\theta - \theta_0\|^2 \vee \|\eta - \eta_0\|_e^2)$

Theory: Conditions on Estimators of Nuisance Functions

Second key condition is that nuisance functions are estimated "well-enough":

Let $(\Delta_n)_{n \geqslant 1}$ and $(\tau_{\pi n})_{n \geqslant 1}$ be some sequences of positive constants converging to zero, and let $a > 1$, $v > 0$, $K > 0$, and $q > 2$ be constants.

Assume for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$

- (Estimator and Truth) (i) w.p. at least $1 - \Delta_n$, $\widehat{\eta}_0 \in \mathcal{T}$ and (ii) $\eta_0 \in \mathcal{T}$.
  - Recall that "parameter space" for η is T
- (Convergence Rate) For all $\eta \in \mathcal{T}$, $\|\eta - \eta_0\|_e \leqslant \tau_{\pi n}$

Theory: Conditions on Estimators of Nuisance Functions (Continued)

- (Pointwise Entropy) For each $\eta \in \mathcal{T}$, the function class $\mathcal{F}_{1,\eta} = \{\psi_j(\cdot, \theta, \eta) : j = 1, ..., d_\theta, \theta \in \Theta\}$ is suitably measurable and its uniform entropy numbers obey

  $$\sup_Q \log N(\epsilon \| F_{1,\eta} \|_{Q,2}, \mathcal{F}_{1,\eta}, \| \cdot \|_{Q,2}) \leqslant v \log(a/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1$$

  where $F_{1,\eta}$ is a measurable envelope for $\mathcal{F}_{1,\eta}$ that satisfies $\| F_{1,\eta} \|_{P,q} \leqslant K$
- (Moments) For all $\eta \in \mathcal{T}$ and $f \in \mathcal{F}_{1,\eta}$, $c_0 \leqslant \| f \|_{P,2} \leqslant C_0$
- (Rates) $\tau_{\pi n}$ satisfies (a) $n^{-1/2} \leqslant C_0 \tau_{\pi n}$, (b) $(B_{1n} \tau_{\pi n})^{\omega/2} + n^{-1/2+1/q} \leqslant C_0 \delta_n$, and (c) $n^{1/2} B_{1n}^2 B_{2n} \tau_{\pi n}^2 \leqslant C_0 \delta_n$.

Rate of convergence is $\tau_{\pi n}$ - needs to be faster than $n^{-1/4}$

- Same as rate condition widely used in semiparametrics employing classical nonparametric estimators

Theory: Main Theoretical Result

Let "Double ML" or "Orthogonalized ML" estimator

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

be such that

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

Theorem (Main Result)

Under assumptions stated above, $\check{\theta}_0$ obeys

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

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

Theory: Attaining full efficiency by Cross-Fitting

- full efficiency not obtained, but can follow Belloni et al (2010,2012) to do the following:

Corollary

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

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

to gain full efficiency.

Corollary

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

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

to gain full efficiency.

Theory: Extensions to "Quasi" Splitting

- Given the split (I, I^c), it is tempting to use I^c to build a collection of ML estimators

$$\widehat{\eta}_m(I^c), \quad m = 1, ..., M$$

  for the nuisance parameters η, and then pick the winner $\widehat{\eta}_{m(I)}(I^c)$ based upon I. This does break the sample-splitting.

- The results still go through under the condition that the winning method has the rate $\tau_{\pi n}$ such that

$$\tau_{\pi n}\sqrt{\log M} \to 0.$$

- The entropy is back, but in a gentle, $\sqrt{\log M}$ way.

Example 1. ATE in Partially Linear Model

Recall

$$
\begin{aligned}
&Y = D\theta_0 + g_0(Z) + \zeta, \qquad \mathrm{E}[\zeta \mid Z, D] = 0,\\
&D = m_0(Z) + V, \qquad \mathrm{E}[V \mid Z] = 0.
\end{aligned}
$$

Base estimation on orthogonal moment condition

$$\psi(W, \theta, \eta) = ((Y - \ell(Z) - \theta(D - m(Z)))(D - m(Z)), \quad \eta = (\ell, m).$$

Easy to see that

- θ_0 is a solution to $\mathrm{E}_P\psi(W, \theta_0, \eta_0) = 0$
- $\partial_\eta \mathrm{E}_P\psi(W, \theta_0, \eta)\Big|_{\eta=\eta_0} = 0$

Example 2. ATE and ATT in the Heterogeneous Treatment Effect Model

Consider a treatment $D \in \{0,1\}$. We consider vectors (Y, D, Z) such that

$$Y = g_0(D, Z) + \zeta, \quad \mathrm{E}[\zeta \mid Z, D] = 0, \tag{4}$$

$$D = m_0(Z) + \nu, \quad \mathrm{E}[\nu \mid Z] = 0. \tag{5}$$

The average treatment effect (ATE) is

$$\theta_0 = \mathrm{E}[g_0(1, Z) - g_0(0, Z)].$$

The the average treatment effect for the treated (ATT)

$$\theta_0 = \mathrm{E}[g_0(1, Z) - g_0(0, Z) | D = 1].$$

- The confounding factors Z affect the D via the propensity score $m(Z)$ and Y via the function $g_0(D, Z)$.
- Both of these functions are unknown and potentially complicated, and we can employ Machine Learning methods to learn them.

Example 2 Contuned. ATE and ATT in the Heterogeneous Treatment Effect Model

For estimation of the ATE, we employ

$$\begin{aligned} \psi(W,\theta,\eta) &:= \theta - \frac{D(Y-\eta_2(Z))}{\eta_3(Z)} - \frac{(1-D)(Y-\eta_1(Z)))}{1-\eta_3(Z)} - (\eta_1(Z)-\eta_2(Z)), \\ \eta_0(Z) &:= (g_0(0,Z), g_0(1,Z), m_0(Z))', \end{aligned} \tag{6}$$

where $\eta(Z) := (\eta_j(Z))_{j=1}^3$ is the nuisance parameter. The true value of this parameter is given above by $\eta_0(Z)$.

For estimation of ATT, we use the score

$$\begin{aligned} &\psi(W,\theta,\eta) = \frac{D(Y-\eta_2(Z))}{\eta_4} - \frac{\eta_3(Z)(1-D)(Y-\eta_1(Z))}{(1-\eta_3(Z))\eta_4} + \frac{D(\eta_2(Z)-\eta_1(Z))}{\eta_4} - \theta\frac{D}{\eta_4}, \\ &\eta_0(Z) = (g_0(0,Z), g_0(1,Z), m_0(Z), \mathrm{E}[D])', \end{aligned} \tag{7}$$

Example 2 Continued. ATE and ATT in the Heterogeneous Treatment Effect Model

It can be easily seen that true parameter values θ_0 for ATT and ATE obey

$$\mathrm{E}_P \psi(W, \theta_0, \eta_0) = 0,$$

for the respective scores and that the scores have the required orthogonality property:

$$\partial_\eta \mathrm{E}_P \psi(W, \theta_0, \eta) \Big|_{\eta = \eta_0} = 0.$$

We use ML methods to obtain:

$$\widehat{\eta}_0(Z) := (\widehat{g}_0(0, Z), \widehat{g}_0(1, Z), \widehat{m}_0(Z))',$$

$$\widehat{\eta}_0(Z) = (\widehat{g}_0(0, Z), \widehat{g}_0(1, Z), \widehat{m}_0(Z), \mathbb{E}_n[D]).$$

The resulting "double ML" estimator $\check{\theta}_0$ solves the empirical analog:

$$\mathbb{E}_{n,I} \psi(W, \check{\theta}_0, \widehat{\eta}_0) = 0, \tag{8}$$

and the solution $\check{\theta}_0$ can be given explicitly since the scores are affine with respect to θ.

Example 3. LATE and LATTE in Heterogeneous Treatment Effect Models with Endogenous Treatment

- LATE can be written as a ratio of ATE of a binary instrument on D and Y, so can use Example 2 to estimate each piece.
- Similar construction works for LATTE.
- By defining

$$\tilde{Y}_t = 1(Y \leqslant t)$$

  can study Distributional and Quantile Treatment Effects.
- See "Program Evaluation ..." paper for details.

Example 4. Moment Condition Models

Very common framework in structural econometrics.

- See the paper for the partially linear IV models.
- See Chernozhukov, Hansen, Spindler ARE, 2015 for parametric GMM case
- See "Program Evaluation ..." (Econometrica, 2016) for semi-parametric case.
- See the paper with Whitney on "Locally Robust Semi-parametric Estimation", with applications to dynamic games.

Empirical Example: 401(k) Pension Plan

Follow Poterba et al (97), Abadie (03). Data from 1991 SIPP, $n = 9,915$

- Y is net total financial assets
- D is indicator for working at a firm that offers a 401(k) pension plan
- Z includes age, income, family size, education, and indicators for married, two-earner, defined benefit pension, IRA participation, and home ownership

D is plausibly exogenous at the time when 401(k) was introduced

Controlling for Z is important due to 401(k) mostly offered by firms employing mostly workers from middle and above middle class (Poterba, Venti, and Wise 94, 95, 96, 01)

Empirical Example: 401(k)

Table: Estimated ATE of 401(k) Eligibility on Net Financial Assets

	RForest	PLasso	B-Trees	Nnet	BestML
A. Part. Linear Model					
ATE	8845	8984	8612	9319	8922
	(1317)	(1406)	(1338)	(1352)	(1203)
B. Interactive Model					
ATE	8133	8734	8405	7526	8295
	(1483)	(1168)	(1193)	(1327)	(1162)

Estimated ATE and heteroscedasticity robust standard errors (in parentheses) from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Further details about the methods are provided in the main text.

Application to Ghana Data (Duflo et al, 2017) with 2000 controls

- Study effect of secondary education.
- Ground truth: experimental estimates of the effect of secondary education.
- Try to recover experimental estimates from observational/non-experimental data using **2,000** controls.

Returns To Secondary School Completion for Males

Outcome	Experimental	Observ.: OLS (5 controls)	Observ.: DML
Standardized Score	0.502	0.595	**0.486**
	(0.205)	(0.069)	(0.066)
Wage Worker	0.057	0.091	**0.082**
	(0.109)	(0.036)	(0.037)
Log Earnings	-0.195	**-0.094**	-0.064
	(0.245)	(0.087)	(0.088)
Partner pregnant	-0.089	-0.167	**-0.120**
	(0.093)	(0.032)	(0.030)

Concluding Comments

We provide a general set of results that allow $\sqrt{n}$-consistent estimation and provably valid (asymptotic) inference for causal parameters, using a wide class of flexible (ML, nonparametric) methods to fit the nuisance parameters.

Three key elements:

1. Neyman-Orthogonal estimating equations
2. Fast enough convergence of estimators of nuisance quantities
3. Sample splitting allows a wide Class of ML estimators.
    - Really eliminates requirements on the entropic complexity on the realizations of $\widehat{\eta}$
    - Allows establishment of results using only rate conditions, not exploiting specific structure of ML estimators (as in, e.g., results for inference following lasso-type estimation in full-sample)

Thank you!
References.

- "Double Machine Learning for Causal and Treatment Effects" ArXiv 2016, with **Denis Chetverikov, Esther Duflo, Christian Hansen, Mert Demirer, Whitney Newey, James Robins**