Double Machine Learning for Causal and Treatment Effects

Victor Chernozhukov

January 5, 2017

This presentation is based on the following papers:

- "Program Evaluation and Causal Inference with High-Dimensional Data", ArXiv 2013, Econometrica 2016+

  with **Alexandre Belloni, I. Fernandez-Val, Christian Hansen**

- "Double Machine Learning for Causal and Treatment Effects"

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

Introduction

- Main goal: Provide general framework for estimating and doing inference about a low-dimensional parameter (θ_0) in the presence of high-dimensional nuisance parameter (η_0) which 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 semiparametric 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 analysis here requires the estimators to take values in a Donsker set, which really rules out most of the new methods.

Literature

- Lots of recent work on inference based on lasso-type methods
  - e.g. Belloni, Chen, Chernozhukov, and Hansen (2012); Belloni, Chernozhukov, Fernández-Val, and Hansen (2015); Belloni, Chernozhukov, and Hansen (2010, 2014); Belloni, Chernozhukov, Hansen, and Kozbur (2015); Belloni, Chernozhukov, and Kato (2013a, 2013b); Belloni, Chernozhukov, and Wei (2013); Farrell (2015); Javanmard and Montanari (2014); Kozbur (2015); van de Geer, Bühulmann, Ritov, and Dezeure (2014); Zhang and Zhang (2014); Liu et al (2015, 2016).
- Little work on other ML methods, with exceptions, e.g., Chernozhukov, Hansen, and Spindler (2015) and Athey and Wager (2015);
  - Will build on the general framework in Chernozhukov, Hansen, and Spindler (2015)

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" 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
  - θ_0 is the target parameter of interest
- Z is a high-dimensional vector of other covariates, called "controls" or "confounders"

Z are confounders in the sense that

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

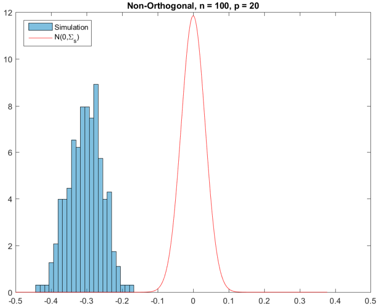
where $m_0 \neq 0$, as is typically the case in observational studies.

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 guesses, 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 fit $\widehat{\theta}_0$; Repeat until convergence.
- Excellent prediction performance! BUT the distribution of $\widehat{\theta}_0 - \theta_0$ looks like this:



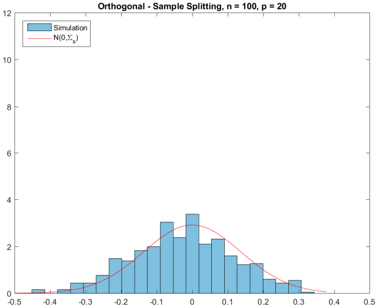
Point II. The "Double" ML Approach is Good

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:

$$\mathrm{E}[(Y - D\theta_0 - g_0(Z))D] = 0 \quad (1)$$

$$\mathrm{E}[(Y - D\theta_0)(D - E[D|Z])] = 0 \quad (2)$$

$$\mathrm{E}[((Y - E[Y|Z]) - (D - E[D|Z])\theta_0)(D - E[D|Z])] = 0 \quad (3)$$

- (1) - Regression adjustment;
- (2) - "propensity score adjustment"
- (3) - Neyman-orthogonal (semi-parametrically efficient score under homoscedasticity).

Both approaches generate estimators of θ_0 that solve the empirical analog of the moment conditions above, where instead of unknown nuisance functions

$$g_0(Z), \quad m_0(Z) := \mathrm{E}[D|Z], \quad \ell_0(Z) = \mathrm{E}[Y|Z]$$

we plug-in their ML-based estimators, obtained using a set-aside sample.

“Naive” or ”Prediction-focused” ML Estimation from (1)

Suppose we use (1) with an estimator $\widehat{g}_0(Z)$ to estimate θ_0:

$$\widehat{\theta}_0 = \left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - \widehat{g}_0(Z_i))$$

$$\sqrt{n}(\widehat{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\left(g_0(Z_i) - \widehat{g}_0(Z_i)\right)}_{:=b}$$

- $a \rightsquigarrow N(0, \bar{\Sigma})$ under standard conditions
- What about b?

Estimation Error in Nuisance Function

Will generally have $b \to \infty$:

$$b \approx (\mathrm{E}D^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(Z_i)\left(g_0(Z_i) - \widehat{g}_0(Z_i)\right)$$

$\left(g_0(Z_i) - \widehat{g}_0(Z_i)\right)$ error in estimating g_0

Heuristics:

- In nonparametric setting, the error is of order $n^{-\varphi}$ for $0 < \varphi < 1/2$
- b will then look like $\sqrt{n}n^{-\varphi} \to \infty$

The "naive" or prediction-focused ML estimator $\widehat{\theta}_0$ is not root-n consistent

Similar heuristics would apply to estimation with (2)

Orthogonalized or "Double ML" Formulation

Consider estimation based on (3)

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

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

Under mild conditions, can write

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

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

$$+ o_p(1)$$

Heuristic Convergence 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 $\ell(z)$
- $o(n^{-1/4})$ is often an attainable rate for estimating $m(z)$ and $\ell(z)$

The Double ML estimator $\check{\theta}_0$ is a $\sqrt{n}$ consistent and approximately centered normal quite generally.

Heuristic Convergence 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 $\ell(z)$
- $o(n^{-1/4})$ is often an attainable rate for estimating $m(z)$ and $\ell(z)$

The Double ML estimator $\check{\theta}_0$ is a $\sqrt{n}$ consistent and approximately centered normal quite generally.

We also rely on sample splitting to get the third term to be $o_p(1)$, with only the rate restrictions like $o(n^{-1/4})$ on the performance of ML estimators.

This eliminates conditions on the entropic complexity of realizations of ML estimators.

Why Sample Splitting?

In the expansion

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = a^* + b^* + o_p(1)$$

the term $o_p(1)$ contains terms like

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

- *With sample splitting*, easy to control and claim $o_p(1)$.
- *Without sample splitting*, hard to control and claim $o_p(1)$.

- See "Program Evaluation... " (Econometrica, 2016+) for results **without** sample splitting.

Why Sample Splitting?

Technical Remark. Without sample splitting, need maximal inequalities to control

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

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 semi-parametric literature does not apply.**

Key Difference between (1) and (3) is Neyman Orthogonality

- Key difference between estimation based on (1) and estimation based on (3) is that (3) satisfies the Neyman orthogonality condition:

Key Difference between (1) and (3) is Neyman Orthogonality

- Key difference between estimation based on (1) and estimation based on (3) is that (3) satisfies the Neyman orthogonality condition:

  Let
  $$\eta_0 = (\ell_0, m_0) = (\mathrm{E}[Y|Z], \mathrm{E}[D|Z]) \quad , \quad \eta = (\ell, m).$$
  The Gateaux derivative operator of the moment condition (3) with respect to η vanishes:
  $$\partial_\eta \mathrm{E}[\underbrace{((Y - \ell(Z)) - (D - m(Z)))\theta_0)(D - m(Z))}_{\psi(W,\theta_0,\eta)}]\Big|_{\eta=\eta_0} = 0$$

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

- In sharp contrast, this property generally does not hold for the moment condition (1) for nuisance function g.

General Results for Moment Condition Models

Moment conditions model:

$$\mathrm{E}[\psi_j(W, \theta_0, \eta_0)] = 0, \quad j = 1, \dots, d_\theta \tag{4}$$

- $\psi = (\psi_1, \dots, \psi_{d_\theta})'$ 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:

$\psi = (\psi_1, \dots, \psi_{d_\theta})'$ 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 $j = 1, \dots, d_\theta$
- vanishes at $r = 0$: For all $\eta \in \mathcal{T}$ and $j = 1, \dots, d_\theta$,

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

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

Key Ingredient II: Sample Splitting

Results will make use of sample splitting:

- $\{1, ..., N\}$ = set of all observations;
- 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 (4), 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

$$\left\| \mathbb{E}_{n,I}[\psi(W, \check{\theta}_0, \widehat{\eta}_0) \right\| \leqslant \inf_{\theta \in \Theta} \left\| \mathbb{E}_{n,I}[\psi(W, \theta, \widehat{\eta}_0)] \right\| + \epsilon_n, \quad \epsilon_n = o(\delta_n n^{-1/2})$$

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

- full efficiency not obtained, but

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.

Details: Estimating Equations in Parametric Likelihood Example

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

Illustrate 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

Details: Orthogonal Estimating Equations 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) solves $J_{\theta\beta} - \mu J_{\beta\beta} = 0$ (i.e., $\mu_0 = J_{\theta\beta} J_{\beta\beta}^{-1}$) for

$$J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \partial_{(\theta', \beta')} \mathrm{E}\Big[\partial_{(\theta', \beta')'} \ell(W, \theta, \beta)\Big]\Big|_{\theta=\theta_0;\ \beta=\beta_0}$$

- 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

Details: General Construction of Orthogonal Moment Conditions/Estimating Equations

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

For example, can proceed by projecting 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))

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

Example 1. ATE in Partially Linear Model

Recall

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

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

$$Y = g_0(D, Z) + \zeta, \quad \mathrm{E}[\zeta \mid Z, D] = 0, \tag{5}$$

$$D = m_0(Z) + \nu, \quad \mathrm{E}[\nu \mid Z] = 0. \tag{6}$$

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

$$\psi(W,\theta,\eta) := \theta - \frac{D(Y-\eta_2(Z))}{\eta_3(Z)} - \frac{(1-D)(Y-\eta_1(Z)))}{1-\eta_3(Z)} - (\eta_1(Z)-\eta_2(Z)),$$
$$\eta_0(Z) := (g_0(0,Z), g_0(1,Z), m_0(Z))', \tag{7}$$

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

$$\psi(W,\theta,\eta) = \frac{D(Y-\eta_2(Z))}{\eta_4} - \frac{\eta_3(Z)(1-D)(Y-\eta_1(Z))}{(1-\eta_3(Z))\eta_4} + \frac{D(\eta_2(Z)-\eta_1(Z))}{\eta_4} - \theta\frac{D}{\eta_4},$$
$$\eta_0(Z) = (g_0(0,Z), g_0(1,Z), m_0(Z), \mathrm{E}[D])', \tag{8}$$

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:

$$\mathbb{E}_{n,I} \psi(W, \check{\theta}_0, \widehat{\eta}_0) = 0, \tag{9}$$

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

- 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 $Y^* = 1(Y \leqslant t)$ 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 Chernozhukov, Hansen, Spindler ARE, 2015 for parametric case
- See "Program Evaluation ..." (Econometrica, 2016) for semi-parametric case.

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 or total wealth
- 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.

Concluding Comments

Our results 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
   - 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.

- "Program Evaluation and Causal Inference with High-Dimensional Data", ArXiv 2013, Econometrica 2016+
  with **Alexandre Belloni, I. Fernandez-Val, Christian Hansen**

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