Double Machine Learning

August 2016

This presentation is based mostly on

- "Double Machine Learning for Improved Point and Interval Estimation of Treatment and Causal Parameters"

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

with some material from

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

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

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 "machine learning" (ML) methods, such as random forests, boosted trees, lasso, ridge, neural nets, gradient boosting, their aggregations and cross-hybrids.

- 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.

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)
- Little work on other ML methods, with exceptions, e.g., Chernozhukov, Hansen, and Spindler (2015), Athey and Wager (2015), Imbens and Athey (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” or “orthogonalized”** ML, we can construct high quality point and interval estimates of ”causal” parameters.

Key Points via a Partially Linear Model

Illustrate the two key 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 parameter of interest
- Z is a 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.

Moment conditions

Some possible moment conditions for identifying and estimating θ_0:

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

$$\mathrm{E}[(Y - D\theta_0)(D - m_0(Z))] = 0 \tag{2}$$

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

- (1) - Regression adjustment
- (2) - "propensity score" adjustment
- (3) - Neyman-orthogonal or doubly-robust (semi-parametrically efficient under homoscedasticity)

Form estimates for the nuisance functions

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

via ML method and plug them in to the empirical analogs of the moment condition to estimate θ_0.

“Naive” 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 $b \to \infty$:

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

$(g_0(Z_i) - \widehat{g}_0(Z_i))$ 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

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

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

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

$$+ o_p(1)$$

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 have conditions formulated only in terms of rates like $o(n^{-1/4})$ and not in terms of entropic complexity 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 $o_p(1)$.

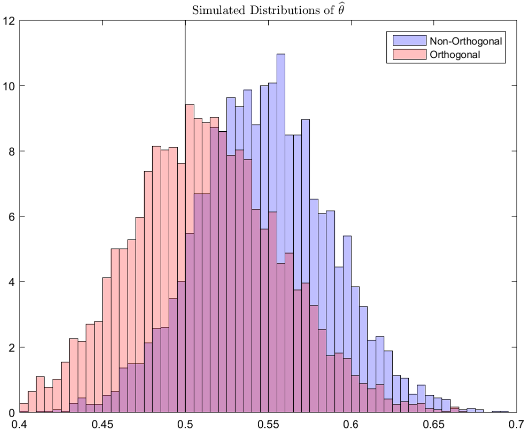
Without sample splitting, need maximal inequalities to control

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

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

- See our "Program Evaluation... " for results **without** sample splitting.

Simple Illustration with Random Forests



- nuisance functions fit with random forests (default settings)
- blue (non-orthogonal): estimator from (1); red (orthogonal): estimator from (3)

Neyman Orthogonality is Key Difference between (1) or (2) and (3)

Key difference between estimation based on (1) or (2) and estimation based on (3) is that (3) satisfies the Neyman orthogonality condition

Let $\eta_0 = (\ell_0, m_0) = (\mathrm{E}[Y|Z], \mathrm{E}[D|Z])$ and $\eta = (\ell, m)$. We have the Gateaux derivative operator with respect to η vanishes in this case:

$$\partial_\eta \mathrm{E}[((Y - \ell) - (D - m)\theta_0)(D - m)]_{\eta=\eta_0} = 0$$

The moment condition remains "valid" under "local" mistakes in the nuisance function.

In sharp contrast, this property generally does not hold for (1) for nuisance function g or for (2) for nuisance function m.

General Results

Moment conditions model:

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

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

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.

Sample Splitting

Results will make use of sample splitting:

- n observations with indices $i \in I \subset \{1, \ldots, N\}$ used to estimate θ_0
- Remaining $\pi n = N - n$ observations with indices $i \in I^c$ used to estimate η_0
  - π assumed bounded away from zero: $\pi \geqslant \pi_0 > 0$ for some fixed constant π_0
  - n is at least of the same order as πn
- Assume I and I^c form a random partition of the set $\{1, ..., N\}$
- Asymptotics take n increasing to ∞

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.

Regularity Conditions on Model

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
- (Orthogonality) ψ obeys the orthogonality condition for the set $\mathcal{T} \subset T$

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 $\eta \in \mathcal{T}$
  - $\mathrm{E}_P[\|\psi(W, \theta, \eta) - \psi(W, \theta_0, \eta_0)\|^2] \leqslant C_0(\|\theta - \theta_0\| \vee \|\eta - \eta_0\|_e)^{\omega}$
  - $\|\partial_r \mathrm{E}_P\left[\psi(W, \theta, \eta_0 + r(\eta - \eta_0))\right]\| \leqslant B_{1n}\|\eta - \eta_0\|_e$
  - $\|\partial_r^2 \mathrm{E}_P[\psi(W, \theta_0 + r(\theta - \theta_0), \eta_0 + r(\eta - \eta_0))]\| \leqslant B_{2n}(\|\theta - \theta_0\|^2 \vee \|\eta - \eta_0\|_e^2)$

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 $\eta \in \mathcal{T}$, $\|\eta - \eta_0\|_e \leqslant \tau_{\pi n}$

Conditions on Estimators of Nuisance Functions (Continued)

- (Pointwise Entropy) For each $\eta \in \mathcal{T}$, the function class $\mathcal{F}_{1,\eta} = \{\psi_j(\cdot, \theta, \eta) : j = 1, \ldots, d_\theta, \theta \in \Theta\}$ 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

Main Result

Let "Double ML" or "Orthogonalized ML" estimator $\check{\theta}_0$ 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 $\Sigma_0 := J_0^{-1} \mathrm{E}_P[\psi^2(W, \theta_0, \eta_0)](J_0^{-1})'$.

- full efficiency not obtained, but

Corollary

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

Extensions: "Quasi" Splitting

- Given the split (I, I^c), it is tempting to use I^c to estimate $\widehat{\eta}$ by a collection of ML methods, $m = 1, ..., M$, and then pick the winner based upon I. This breaks 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.

Partially Linear Model Again

Specialize general results to partially linear model as illustration.

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

Partially Linear Model Assumptions

Let $(\delta_n)_{n=1}^{\infty}$ and $(\Delta_n)_{n=1}^{\infty}$ be sequences of positive constants approaching 0, c and C be fixed positive constants, and $q > 4$.

Let $\mathcal{P}$ be the collection of probability laws P for the triple (Y, D, Z) which follow the partially linear model such that

- (bounded true parameter) $|\theta_0| \leqslant C$
- (moments) $\|X\|_{P,q} \leqslant C$ for $X = Y, D, g_0(Z), m_0(Z)$, $\|V\|_{P,2} \geqslant c$, and $\|\zeta\|_{P,2} \geqslant c$
- (rates for nuisance functions) for all $n \geqslant 1$ with probability no less than $1 - \Delta_n$, $\|\widehat{\ell}_0(Z) - \ell(Z)\|_{P,2} + \|\widehat{m}_0(Z) - m_0(Z)\|_{P,2} + n^{1/2}\|\widehat{\ell}_0(Z) - \ell(Z)\|_{P,2} \cdot \|\widehat{m}_0(Z) - m_0(Z)\|_{P,2} \leqslant \delta_n$

Only requires rate conditions on estimators of nuisance functions.

- opens up possibility to use wide range of estimators
- achieved by sample splitting
- somewhat weaker than conditions for lasso-based approaches in this setting (cost is sample-splitting)
- with additional splitting $n^{1/2}\|\widehat{\ell}_0(Z) - \ell(Z)\|_{P,2} \cdot \|\widehat{m}_0(Z) - m_0(Z)\|_{P,2} \to 0$ can be replaced by a similar condition on biases.

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 likelihood function 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

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

General Construction of Orthogonal 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

401(k) Example

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) plan
- Z includes age, income, family size, education, and indicators for married, two-earner, defined benefit pension, IRA participation, and home ownership

Individuals with high unobserved preference for saving more prone to participate in 401(k)

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

Important to control for income to capture unobserved heterogeneity in employment decisions (Poterba, Venti, and Wise 94, 95, 96, 01)

Some Results from Partially Linear Model

- Baseline 1 (Y on D): 19,559 (1413)
- Baseline 2 (Y on D and Z - linear): 5896 (1523)
- Baseline 3 (Y on D, income, age, educ cat dummies): 8645 (1350)
    - Essentially PVW specification
- Tree (cv to choose depth): 8851 (1451)
    - Y on Z: $RMSE_{A \to B} = 56230$ $RMSE_{B \to A} = 64224$
    - D on Z: $RMSE_{A \to B} = .450$ $RMSE_{B \to A} = .446$
- Random Forest (Default implementation): 9656 (1308)
    - Y on Z: $RMSE_{A \to B} = 51895$ $RMSE_{B \to A} = 58715$
    - D on Z: $RMSE_{A \to B} = .454$ $RMSE_{B \to A} = .450$
- Boosted Trees (cv to choose depth and stopping): 9752 (1353)
    - Y on Z: $RMSE_{A \to B} = 53127$ $RMSE_{B \to A} = 58694$
    - D on Z: $RMSE_{A \to B} = .445$ $RMSE_{B \to A} = .442$

Some Results from Partially Linear Model

- Ridge (cv-min penalty choice): 10103 (1318)
  - Y on Z: $RMSE_{A \to B} = 50745$ $RMSE_{B \to A} = 56846$
  - D on Z: $RMSE_{A \to B} = .446$ $RMSE_{B \to A} = .447$
- Lasso (cv-min penalty choice): 10176 (1336)
  - Y on Z: $RMSE_{A \to B} = 51289$ $RMSE_{B \to A} = 57436$
  - D on Z: $RMSE_{A \to B} = .442$ $RMSE_{B \to A} = .446$
- Post-lasso (plug-in penalty choice, explicitly allow heteroscedasticity): 9411 (1327)
  - Y on Z: $RMSE_{A \to B} = 51489$ $RMSE_{B \to A} = 58371$
  - D on Z: $RMSE_{A \to B} = .446$ $RMSE_{B \to A} = .446$
- Lava (cv-min penalty choice): 9628 (1296)
  - Y on Z: $RMSE_{A \to B} = 51715$ $RMSE_{B \to A} = 57323$
  - D on Z: $RMSE_{A \to B} = .446$ $RMSE_{B \to A} = .447$
- Use RMSE min for each object: 9272 (1284)

All of these methods use 299 variables constructed from a quadratic spline in each of income, age, and educ, main effects in other variables, and all first order interactions

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)