Double/Debiased Machine Learning for Causal and Treatment Effects

May 2, 2018

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**
- "Double/De-biased Machine Learning with Regularized Riesz Representers"

  ArXiv 2018, with **Whitney Newey and James Robins**
- "Program Evaluation and Causal Inference with High-Dimensional Data",*ArXiv* 2013, Econometrica 2017

  with **Alexandre Belloni, I. Fernandez-Val, Christian Hansen**
- "Uniformly Valid Post-Regularization Confidence Regions for Many Functional Parameters in Z-Estimation Framework" *ArXiv* 2013, Annals of Statistics 2018+

  with **Alexandre Belloni, D. Chetverikov, Y. Wei**

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:

1. 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 Example 1: Partially Linear Model

Partially Linear Model

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

- 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.

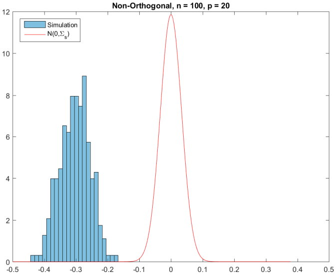
Causal interpretation of θ_0: under conditional exogeneity, or random assignment of D conditional on Z, θ_0 is the average causal effect of D on potential outcome.

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

- Predict Y using D and Z – and obtain

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

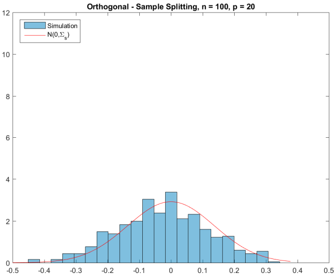
- For example, estimate by alternating minimization– given initial guess $\widehat{\eta}_0$, run Random Forest of $Y - D\widehat{\theta}_0$ on Z to fit $\widehat{g}_0(Z)$ and the Ordinary Least Squares on $Y - \widehat{g}_0(Z)$ on D to get updated $\widehat{\theta}_0$; Repeat until convergence.
- Excellent 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 be of order $o(1/\sqrt{n})$.

Heuristics: The Role of Sample Splitting = To Combat Overfitting

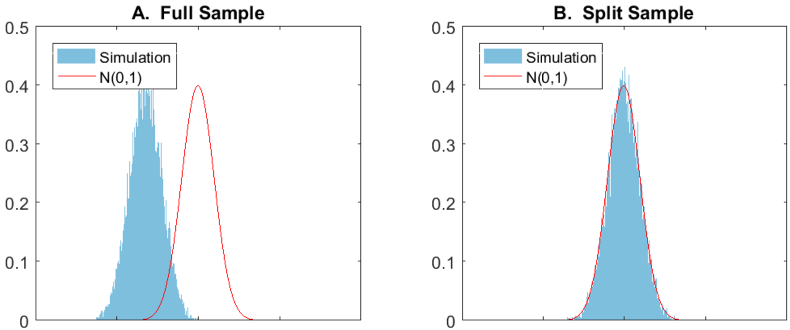
A contrived example with overfitting:

$$\widehat{\ell}(Z_i) = \ell(Z_i) + \underbrace{U_i/\sqrt{N}}_{\text{"overfitting"}}, \quad U_i \sim N(0,1)$$

This estimator of ℓ is excellent

$$\max_i \|\widehat{\ell}(Z_i) - \ell(Z_i)\| = O_p(\sqrt{logN/N}).$$

but without sample splitting, it creates a huge bias:



Heuristics: Overfitting = High Entropy

- 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

$$\mathbb{G}_n(\psi(W, \theta_0, \widehat{\eta}) - \mathbb{G}_n \psi(W, \theta_0, \eta_0) \to_P 0.$$

- **With Sample Splitting**: $\widehat{\eta}$ is based on the auxiliary sample, then this follows from $\|\widehat{\eta} - \eta_0\| \to 0$ and Chebyshev inequality.
- **Without Sample Splitting**: $\widehat{\eta}$ depends on the main sample, and $\widehat{\eta} \in \mathcal{M}_n$

$$\sup_{\eta \in \mathcal{M}_n} \left| \mathbb{G}_n \psi(W, \theta_0, \eta) - \mathbb{G}_n \psi(W, \theta_0, \eta_0) \right| \lesssim \sqrt{enropy(\mathcal{M}_n)} / \sqrt{n}$$

  Need to control the rate of entropy growth; see our Econometrica paper; this is hard to control in practice.

General Results for Moment Condition Models

Moment conditions model:

$$\mathrm{E}[\psi(W, \theta_0, \eta_0)] = 0 \tag{3}$$

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

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.

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 $N - n$, 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: Main Theoretical Result

Let "Double ML" or "Orthogonalized ML" estimator

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

be such that

$$\frac{1}{n}\sum_{i\in I}\psi(W, \check{\theta}_0(I, I^c), \widehat{\eta}_0(I^c)) = 0$$

Lemma (Subsample Estimator)

Under assumptions stated in the paper, including the nuisance parameters estimated at $n^{-1/4}$ rate,

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

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

The subsample estimator does not attain the full efficiency, but can do the following.

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

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

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

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

DML Theory

Theorem (Full Efficiency of The DML Estimator)

Under assumptions stated in the paper, including the nuisance parameters estimated at $N^{-1/4}$ rate,

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

Building Orthogonal Scores in Parametric Setting

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

Building Orthogonal Scores in Parametric Setting

Can construct new estimating equation with desired orthogonality property:

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

Where "true" value (μ_0) is chosen such that

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

for the Hessian (Information Matrix):

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

- Nuisance parameter: $\eta = (\beta', \mathrm{vec}(\mu)')' \in T \times \mathcal{D} \subset \mathbb{R}^p, \quad p = p_0 + dp_0$
- Wie have $\mathrm{E}[\psi(W,\theta_0,\eta_0)] = 0$ for $\eta_0 = (\beta_0', \mathrm{vec}(\mu_0)')'$ and ψ obeys the orthogonality condition: $\partial_\eta \mathrm{E}[\psi(W,\theta_0,\eta)]\Big|_{\eta=\eta_0} = 0$
- The score ψ is the efficient score for inference about θ_0

⋆ See Chernozhukov, Hansen, Spindler (ARE, 2015) for parametric GMM case

Building 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))
- ⋆ See the DML paper for the Partially Linear Instrumental Variable Regression models.
- ⋆ See "Locally Robust Semi-parametric Estimation", with Newey et al, for applications to Dynamic Games/Dynamic Discrete Choices.
- ⋆ See "Program Evaluation ..." (Econometrica, 2016) for semi-parametric cases.

Building Orthogonal Scores For Linear Functionals of Conditional Expectation

- This is discussed in the ArXiv 2018 paper with W. Newey and J. Robins called **"DML with Regularized Riesz Representers"**.
- Linear functional of the conditional expectation $g = \mathrm{E}[Y \mid X]$

$$\theta_0 = \mathrm{E} m(X, g),$$

  where $g \mapsto \mathrm{E} m(X, g)$ is continuous, linear functional.
- Examples of Functionals:
  - ⋆ Average Treatment Effect,
  - ⋆ Policy Effects from Distributional Shifts of Covariates,
  - ⋆ Average Derivative,
  - ⋆ Average Welfare/Consumer Surplus.

- There exists a Riesz representer:

$$\mathrm{E}m(X, g) = \mathrm{E}\alpha(X)g(X).$$

- The Neyman-orthogonal score is given by

$$\psi(W, \theta, \eta) = \theta_0 - m(X, g) - \alpha(X)(Y - g(X)); \quad \eta = (g, \alpha).$$

- In many cases, Riesz representer is available in closed form.
- More generally, can construct an estimator of α_0 as a regularized solution to the empirical Riesz representer equations.
- Note that the more "natural" scores

$$\psi(W, \theta, g) = \theta_0 - m(X, g) \quad \text{and } \psi(W, \theta, g) = \alpha(X)(Y)$$

and not Neyman orthogonal.

Example 2. Average Effects in the Heterogeneous Model

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

$$Y = g_0(D, Z) + \zeta, \quad \mathrm{E}[\zeta \mid Z, D] = 0, \tag{4}$$
$$D = m_0(Z) + \nu, \quad \mathrm{E}[\nu \mid Z] = 0, \tag{5}$$

where the propensity score $m_0(Z)$ is bounded away from zero and one.

- The average treatment effect (ATE) is

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

For causal interpretation assume conditional exogeneity of D given Z.

- The Riesz Representer is given by the Horwitz-Thompson transform:

$$\alpha_0(D, Z) = \frac{1(D = 1)}{m_0(Z)} - \frac{1(D = 0)}{1 - m_0(Z)}.$$

- Indeed, we can verify that

$$\mathrm{E}[g_0(1, Z) - g_0(0, Z)] = \mathrm{E}[g_0(D, Z)\alpha_0(D, Z)] = \mathrm{E}[Y\alpha_0(D, Z)].$$

The latter quantity provides the Horwitz-Thompson propensity-score based identification of θ_0: $\theta_0 = \mathrm{E}\alpha_0(D, Z)Y$.

Example 2 ctd. Average Effects in the Heterogeneous Model

- The Neyman-orthogonal score is generated by the formula

$$\psi(X, \theta, \eta) = \theta_0 - m(X, g) - \alpha(X)(Y - g(X)); \quad \eta = (g, \alpha).$$

  by setting:

$$m(X, g) = g(1, Z) - g(0, Z), \quad \alpha(X) = \frac{1(D = 1)}{m_0(Z)} - \frac{1(D = 0)}{1 - m_0(Z)}.$$

  where $\eta(Z) := (g(D, Z), m(Z))$ is the nuisance parameter.
- The score is the efficient score for inference on ATE. It is also known as the doubly robust score (Robins and Rotnizky, 1995).

Example 2: Wrong Scores for ATE High-Dimensional Problems

- The regression adjustment score

$$\varphi(Z; \theta, \eta) = \theta - (g(1, Z) - g(0, Z))$$

  is not Neyman-orthogonal.
- The propensity score-based score

$$\varphi(Y, Z; \theta, \eta) = \left(\frac{1(D = 1)}{m_0(Z)} - \frac{1(D = 0)}{1 - m_0(Z)} \right) Y$$

  is not Neyman-orthogonal.
- Our theory suggests that we should **not** use these scores in high-dimensional problems.

Example 3. Average Effects in Heterogeneous Treatment Effect Models with Endogenous Treatment

- Binary treatment D, binary instrument I, outcome Y, and regressors Z. The instrument I is randomly assigned, conditional on Z.
- Local Average Treatment Effect (LATE)

$$\theta_0 = \text{LATE} = \frac{\alpha_0}{\beta_0} = \frac{\text{ATE of } I \text{ on } Y}{\text{ATE of } I \text{ on } D},$$

so can use Example 2 to estimate top and bottom.

$$\psi(Y, D, I, Z; \theta, \eta) = \psi_1(D, I, W; \beta, \eta_1)\theta - \psi_2(Y, I, W; \alpha, \eta_2)$$

where ψ_j are ATE scores from Example 2.

Example 2 & 3 ctd: Distributional Effects

- By defining the outcome

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

  can study Distributional and Quantile Treatment Effects.

⋆ See "Program Evaluation ..." for worked out details.

Example 4: Average Derivative

- Here we have

$$Y = g(D, Z) + \epsilon, \quad \mathrm{E}[\epsilon | D, Z] = 0,$$

and the functional of interest is the average derivative with respect to D:

$$\theta_0 = \mathrm{E}[\partial_1 g(D, Z)].$$

- Using integration by parts we find:

$$\mathrm{E}[\partial_1 g(D, Z)] = \mathrm{E}[\alpha(D, Z) g(D, Z)], \quad \alpha(D, Z) = -\partial_1 \log f(D \mid Z)$$

- The Neyman-orthogonal score is generated by the formula

$$\psi(X, \theta, \eta) = \theta_0 - m(X, g) - \alpha(X)(Y - g(X)); \quad \eta = (g, \alpha),$$

by setting:

$$m(X, g) = \partial_1 g(D, Z), \quad \alpha(X) = -\partial_1 \log f(D \mid Z).$$

Example 5: Policy Effect from a Distribution Shift

- We have that
$$Y = g(X) + \epsilon, \quad \mathrm{E}[\epsilon \mid X] = 0.$$
Consider a policy that shifts the distribution of X from F_0 to a new distribution F_1, through a mapping $x \mapsto T(x)$. Suppose that g is not affected under the policy.
- The effect from a counterfactual change of covariate distribution from F_0 to F_1:
$$\theta_0 = \mathrm{E}g_0(T(X)) - \mathrm{E}g_0(X) = \int g_0(x)[d(F_1(x) - F_0(x))/dF_0(x)]dF_0(x).$$
- The Neyman-orthogonal score is generated by the formula
$$\psi(X, \theta, \eta) = \theta_0 - m(X, g) - \alpha(X)(Y - g(X)); \quad \eta = (g, \alpha),$$
by setting:
$$m(X, g) = g(T(X)) - g(X), \quad \alpha(X) = [d(F_1(x) - F_0(x))/dF_0(x)].$$

Empirical Example I: Bonus Experiment

- Reanalysis of the Pennsylvania Reemployment Bonus experiment conducted by the US Department of Labor in the 1980s to test the incentive effects of unemployment insurance
- Claimants were randomly assigned either to control and treatment group
- Individuals in the treatment groups were offered a cash bonus if they found a job
- Our treatment variable, D, is an indicator variable for being assigned treatment
- The outcome variable, Y, is the log of duration of unemployment for the UI claimants. The vector of covariates, X, consists of age group dummies, gender, race, the number of dependents, quarter of the experiment, location within the state, existence of recall expectations, and type of occupation.

Empirical Results

Table: Estimated Effect of Cash Bonus on Unemployment Duration

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Regression Model							
ATE (2 fold)	-0.081	-0.084	-0.074	-0.079	-0.073	-0.079	-0.078
	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
ATE (5 fold)	-0.081	-0.085	-0.074	-0.077	-0.073	-0.078	-0.077
	[0.036]	[0.036]	[0.036]	[0.035]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
B. Partially Linear Regression Model							
ATE (2 fold)	-0.080	-0.084	-0.077	-0.076	-0.074	-0.075	-0.075
	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)
ATE (5 fold)	-0.080	-0.084	-0.077	-0.074	-0.073	-0.075	-0.074
	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.037)	(0.036)	(0.035)	(0.036)	(0.035)	(0.035)

Note: Results are based on 100 splits with point estimates calculated the median method. The median standard error across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses.

Empirical Example II : 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 Results: ATE for 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.

Empirical Results: LATE for 401(k)

Table: Estimated LATE Effect of 401(k) Participation on Net Financial Assets

	Lasso	Forest	Boosting	Neural Net.	Best
A. Interactive Regression Model					
LATE (2 fold)	8978	11384	11329	11094	10952
	[2192]	[1749]	[1666]	[1903]	[1657]
LATE (5 fold)	8944	11764	11133	11186	11113
	[2259]	[1788]	[1661]	[1795]	[1645]

Note: Estimated LATE based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard error across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

The Effect of Institutions on Economic Growth

- Instrumental variable estimation using DML
- Effect of institutions on aggregate output following the work of Acemoglu et al.(2001)
- The outcome variable, Y, is the logarithm of GDP per capita and D, is a proxy for the strength of institutions
- To deal with endogeneity, we use an instrumental variable Z, which is mortality rates for early European settlers.
- Our raw set of control variables, X, include distance from the equator and dummy variables for Africa, Asia, North America, and South America.
- The results are qualitatively similar to Acemoglu et al.(2001)'s results.

Results

Table: Estimated Effect of Institutions on Output

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
2 fold	0.85	0.81	0.84	0.77	0.94	0.8	0.83
	[0.28]	[0.42]	[0.38]	[0.33]	[0.32]	[0.35]	[0.34]
	(0.22)	(0.29)	(0.3)	(0.27)	(0.28)	(0.3)	(0.29)
5 fold	0.77	0.95	0.9	0.73	1.00	0.83	0.88
	[0.24]	[0.46]	[0.41]	[0.33]	[0.33]	[0.37]	[0.41]
	(0.17)	(0.45)	(0.4)	(0.27)	(0.3)	(0.34)	(0.39)

Note: Results are based on 100 splits with point estimates calculated the median method. The median standard error across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses.

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)