# Machine learning & nonparametric efficiency in causal inference

**Edward Kennedy**

Department of Statistics & Data Science
Carnegie Mellon University

ACIC, 23 May 2022

# Roadmap

Introduction & Setup
- Intro
- Target Parameter
- Identification

Efficiency Theory
- Background & Setup
- Nonparametric Efficiency Bounds
- Influence Functions

Nonparametric Estimation
- Intro & Setup
- Empirical Processes & Sample Splitting
- Second-Order Remainder

Software & References

arXiv > stat > arXiv:2203.06469

Search...

Help | Advance

**Statistics > Methodology**

*[Submitted on 12 Mar 2022]*

# Semiparametric doubly robust targeted double machine learning: a review

Edward H. Kennedy

In this review we cover the basics of efficient nonparametric parameter estimation (also called functional estimation), with a focus on parameters that arise in causal inference problems. We review both efficiency bounds (i.e., what is the best possible performance for estimating a given parameter?) and the analysis of particular estimators (i.e., what is this estimator's error, and does it attain the efficiency bound?) under weak assumptions. We emphasize minimax-style efficiency bounds, worked examples, and practical shortcuts for easing derivations. We gloss over most technical details, in the interest of highlighting important concepts and providing intuition for main ideas.

Subjects: **Methodology (stat.ME)**
Cite as: arXiv:2203.06469 **[stat.ME]**

# My disclaimers

1. A lot of the theory I'll talk about today has a long history
   - see: Andrews, Begun, Bickel, Pfanzagl, Pollard, Robins, Stein, Tsiatis, van der Laan, van der Vaart, Wellner...
2. Coverage driven by interests in causal/functional estimation
   - non/semiparametrics, empirical processes, etc. are huge fields
3. Some of what I say is my own philosophical perspective
   - reasonable people can disagree!

# What is the scientific goal?

An important first step in any scientific pursuit:

- have a **clearly defined goal**

In particular, for statistical estimation problems:

- we need a target parameter (estimand), which we'll call $\psi^*$

The *target parameter* $\psi^*$ is the main feature of interest

- e.g., what would happen if all vs. none were treated?
- called a **functional** if $\psi : \mathcal{P} \mapsto \mathbb{R}$ is some structured combination of parts of $\mathbb{P}$

I find it essential to be clear about our target parameter!

# Potential outcomes

Causal language lets us define target wrt. hypothetical experiments

- using, e.g., **potential outcomes**, structural equations, etc.

We use superscripts to denote what would have been observed under some intervention

- $Y^a$ denotes outcome $Y$ that would have been observed had we set treatment to $A = a$
- $Y^{\overline{a}_T}$ denotes outcome had we set treatment sequence $\overline{A}_T = (A_1, ..., A_T)$ to $\overline{a}_T = (a_1, ..., a_T)$
- $Y^G$ denotes outcome under stochastic intervention $G$ that assigns $A = 1$ with probability $g(x)$ depending on covariates $X$

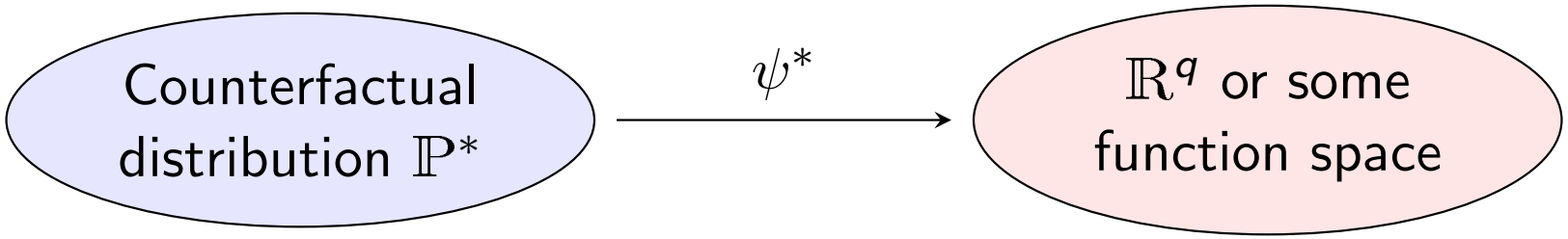
# Example causal parameters

| | |
|---|---|
| $\rightarrow$ ATE: | $\mathbb{E}(Y^1 - Y^0)$ |
| $\rightarrow$ conditional ATE: | $\mathbb{E}(Y^1 - Y^0 \mid V)$ |
| $\rightarrow$ local ATE: | $\mathbb{E}(Y^1 - Y^0 \mid A^1 > A^0)$ |
| $\rightarrow$ dose-response curve: | $\mathbb{E}(Y^a)$ |
| $\rightarrow$ optimal trt strategy: | $\arg\max_d \mathbb{E}(Y^{d(V)})$ |
| $\rightarrow$ MSM: | $\mathbb{E}(Y^{\bar{a}_T} \mid V)$ |
| $\rightarrow$ SNM: | $\mathbb{E}(Y^{\bar{a}_t,0} - Y^{\bar{a}_{t-1},0} \mid \overline{L}_t, \overline{A}_t)$ |

# Causal target parameter

Thus $\psi^*(\mathbb{P}^*)$ is a map from a counterfactual distribution $\mathbb{P}^*$

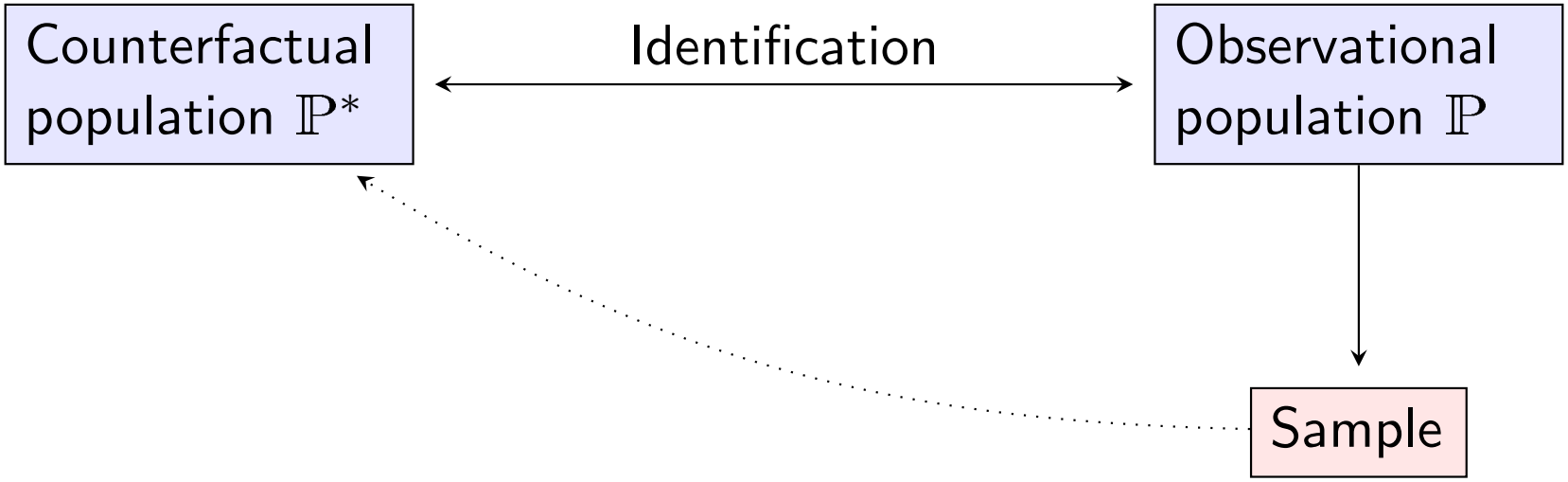
- can be a number, or function, or even more complex object



# Identification

After picking $\psi^*$, we need to link to observational distribution $\mathbb{P}$

- this is the enterprise of *identification*
- goal: express $\psi^*(\mathbb{P}^*) = \psi(\mathbb{P})$ for some mapping $\psi$



# Identification example: ATE

Assume:

1. Positivity: $\mathbb{P}\{\mathbb{P}(A = a \mid X) \geq \epsilon > 0\} = 1$
2. Consistency: $A = a \implies Y = Y^a$
3. Exchangeability: $A \perp\!\!\!\perp Y^a \mid X$

Then by 1 we can write:

$$\mathbb{E}(Y \mid X, A = a) \overset{2}{=} \mathbb{E}(Y^a \mid X, A = a) \overset{3}{=} \mathbb{E}(Y^a \mid X)$$

so that, e.g.: $\psi^* \equiv \mathbb{E}(Y^a) \;=\; \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\} \equiv \psi.$

# Identification example 2: g-formula

Let $\overline{X}_t = (X_1, ..., X_t)$, and $H_t = (\overline{X}_t, \overline{A}_{t-1})$ be history prior to $A_t$.

Assume:

1. Positivity: $\mathbb{P}\{\mathbb{P}(A_t = a_t \mid H_t) \geq \epsilon > 0\} = 1$
2. Consistency: $\overline{A}_T = \overline{a}_T \implies Y = Y^{\overline{a}_T}$
3. Exchangeability: $A_t \perp\!\!\!\perp Y^{\overline{a}_T} \mid H_t$ for all $t$

Then

$$\mathbb{E}(Y^{\overline{a}_T}) = \int_{\mathcal{X}_1} ... \int_{\mathcal{X}_T} \mathbb{E}(Y \mid \overline{X}_T, \overline{A}_T = \overline{a}_T) \prod_{t=1}^{T} d\mathbb{P}(X_t \mid \overline{X}_{t-1}, \overline{A}_{t-1} = \overline{a}_{t-1})$$

# Identification example 3: LATE

Assume for $r \in \{0, 1\}$:

1. Positivity: $\mathbb{P}\{\mathbb{P}(R = r \mid X) \geq \epsilon > 0\} = 1$
2. Consistency: $A = A^R$ and $Y = Y^{RA}$
3. Exchangeability: $R \perp\!\!\!\perp (A^r, Y^r) \mid X$
4. Exclusion: $Y^{ra} = Y^a$ for all $a \in \{0, 1\}$
5. Instrumentation: $\mathbb{P}(A^1 > A^0) \geq \epsilon > 0$
6. Monotonicity: $\mathbb{P}(A^1 \geq A^0) = 1$

Then:

$$\mathbb{E}(Y^1 - Y^0 \mid A^1 > A^0) = \frac{\mathbb{E}\{\mathbb{E}(Y \mid X, R = 1) - \mathbb{E}(Y \mid X, R = 0)\}}{\mathbb{E}\{\mathbb{E}(A \mid X, R = 1) - \mathbb{E}(A \mid X, R = 0)\}}$$

# Identification example 4: ATE bounds

Assume:

1. Positivity: $\mathbb{P}\{\mathbb{P}(A = a \mid X) \geq \epsilon > 0\} = 1$
2. Consistency: $A = a \implies Y = Y^a$
3. $\delta$-Exchangeability: $|\mathbb{E}(Y^a \mid X, A = a) - \mathbb{E}(Y^a \mid X, A \neq a)| \leq \delta$

Then $\psi = \mathbb{E}(Y^a)$ is not identified but can be bounded as

$$\psi \in [\psi_\ell, \psi_u] = \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\} \pm \delta \mathbb{P}(A \neq a)$$

Now we can just treat the bounds as our target parameter

# Causal inference is over

I find it v. helpful to keep causal & statistical issues separate

- Causal stuff (our target, what we think abt confounding, etc.) only tells us **what** we should be estimating
- not **how** to estimate it

After we identify the causal target parameter $\psi^*(\mathbb{P}^*)$

- the role of causal inference is over
- now we have a pure **functional estimation** problem

There are also many *non-causal* functional estimation problems

- expected density: $\psi = \int p(x)^2 \, dx$
- entropy: $\psi = -\int p(x) \log p(x) \, dx$
- mutual information: $\psi = \int \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, dx \, dy$

$\rightarrow$ the methods we discuss are equally useful for these problems

# Causal inference is over

Now we have a purely statistical task:

- observe iid sample $Z_1, ..., Z_n$ with $Z \sim \mathbb{P}$, assuming $\mathbb{P} \in \mathcal{P}$ for statistical model $\mathcal{P}$ (i.e., set of possible distributions)
- we want to construct a 'good' estimator $\hat{\psi}$ of $\psi = \psi(\mathbb{P})$

In theory we can construct $\hat{\psi}$ using any preferred approach:

1. parametric Bayes or MLE
2. nonparametric MLE / plug-in
3. **nonparametric influence function-based**

I will discuss and argue for the last approach

# Aside: Estimation/inference basics

An estimator is just a map from the data to, e.g., a number

- in math: $\hat{\psi} : (Z_1, ..., Z_n) \mapsto \mathbb{R}$

Estimators often take the form of a sample average, e.g.,

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_i) \equiv \mathbb{P}_n\{\varphi(Z)\}$$

at least asymptotically, where

- $\mathbb{P}_n(\cdot)$ is short-hand for sample average

# Aside: Big-oh and little-oh notation, etc

$X_n = O_{\mathbb{P}}(1)$ means $X_n$ stays bounded as $n \to \infty$, i.e.,

$$\mathbb{P}(|X_n| > M) < \epsilon \quad \text{for any } \epsilon, \text{ if } n \text{ large enough}$$

Similarly $X_n = o_{\mathbb{P}}(1)$ means $X_n \to 0$ as $n \to \infty$, i.e., for all $\epsilon$

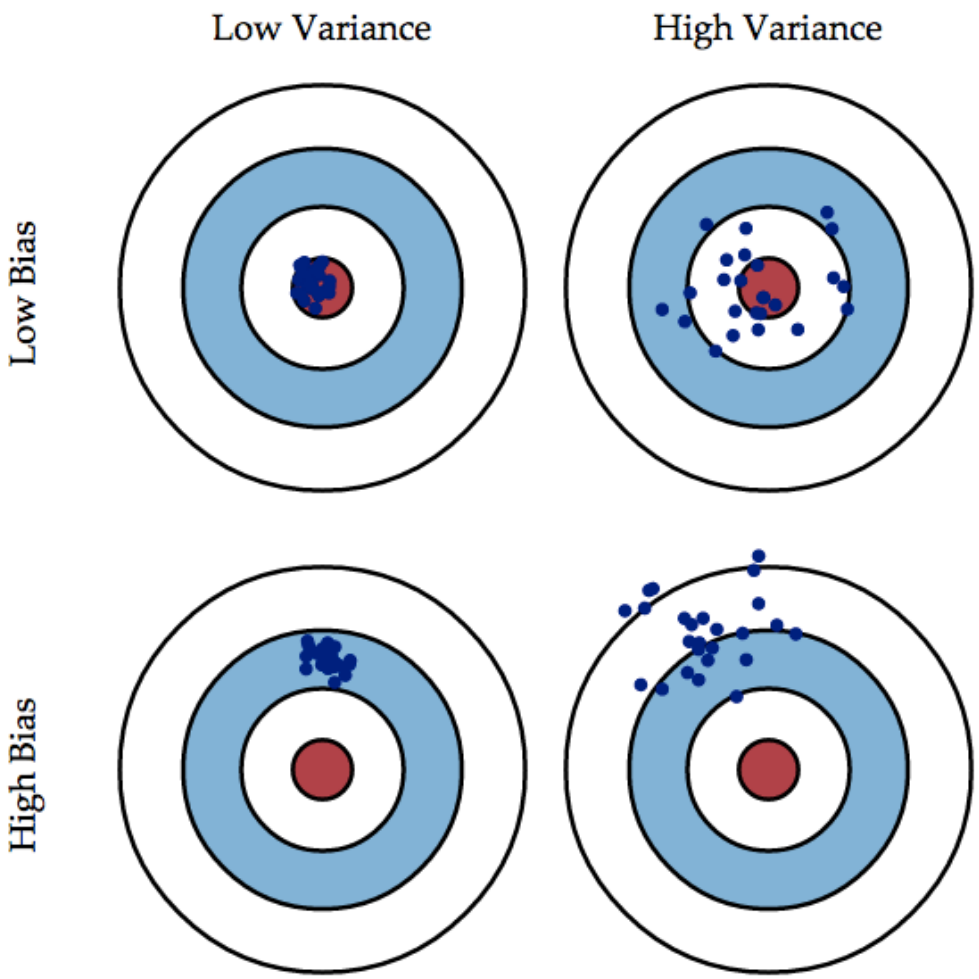
$$\mathbb{P}(|X_n| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty$$

An estimator is root-n consistent and asymptotically normal if

$$\sqrt{n}(\hat{\psi} - \psi) = \sqrt{n}\mathbb{P}_n(\varphi) + o_{\mathbb{P}}(1) \rightsquigarrow N(0, \text{var}(\varphi))$$

and then we say $\widehat{\psi}$ has influence function $\varphi$

# Aside: Not all estimators are created equal



# Aside: Rates of convergence

Some estimators hit their targets more precisely than others

An estimator $\hat{\psi}$ has rate of convergence $r_n \to \infty$ if

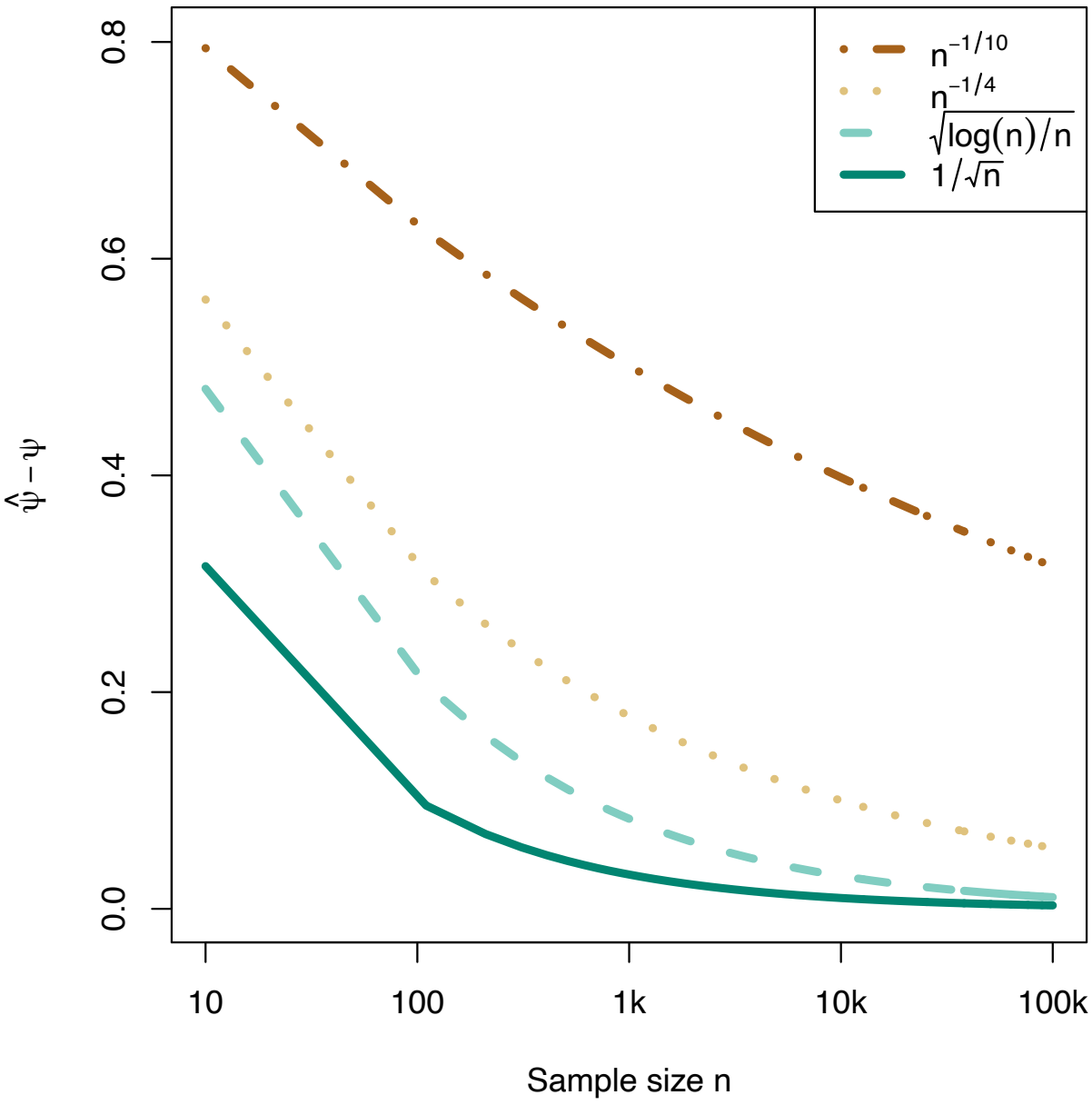
$$r_n(\hat{\psi} - \psi) = O_{\mathbb{P}}(1) \quad \Longleftrightarrow \quad \hat{\psi} - \psi = O_{\mathbb{P}}(1/r_n)$$

This means the difference $\hat{\psi} - \psi$ behaves like $1/r_n \to 0$

- in other words: $\hat{\psi}$ is clustering around $\psi$ (i.e., consistent)

The rate $r_n$ tells us how fast/precisely $\hat{\psi}$ clusters around $\psi$

- why does this matter? $\to$ fast rates mean more information (e.g., tighter CIs) with smaller sample!


$n^{-1/10}$
$n^{-1/4}$
$\sqrt{\log(n)/n}$
$1/\sqrt{n}$
0.0
0.2
0.4
0.6
0.8
$\hat{\psi} - \psi$
10
100
1k
10k
100k
Sample size n

# Aside: The curse of dimensionality

The curse of dimensionality says roughly that statistical methods must degrade as we include more and more covariates

e.g., for $d$-dimensional regression function $\mu(X) = \mathbb{E}(Y \mid X)$ that is $\beta$-smooth, best possible convergence rate for any $\hat{\mu}$ is

$$\inf_{\hat{\mu}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \|\hat{\mu} - \mu_{\mathbb{P}}\| \geq C n^{\frac{-1}{2+d/\beta}}$$

In most optimistic case, where $d = 1$ and $\beta = 1000$, still $< \sqrt{n}$

- if $d = 20$ and $\beta = 1$, this gives very slow $n^{-1/22}$ rate

# Why not parametric approach?

Could assume $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}^d\}$ indexed by finite-dimensional $\theta$

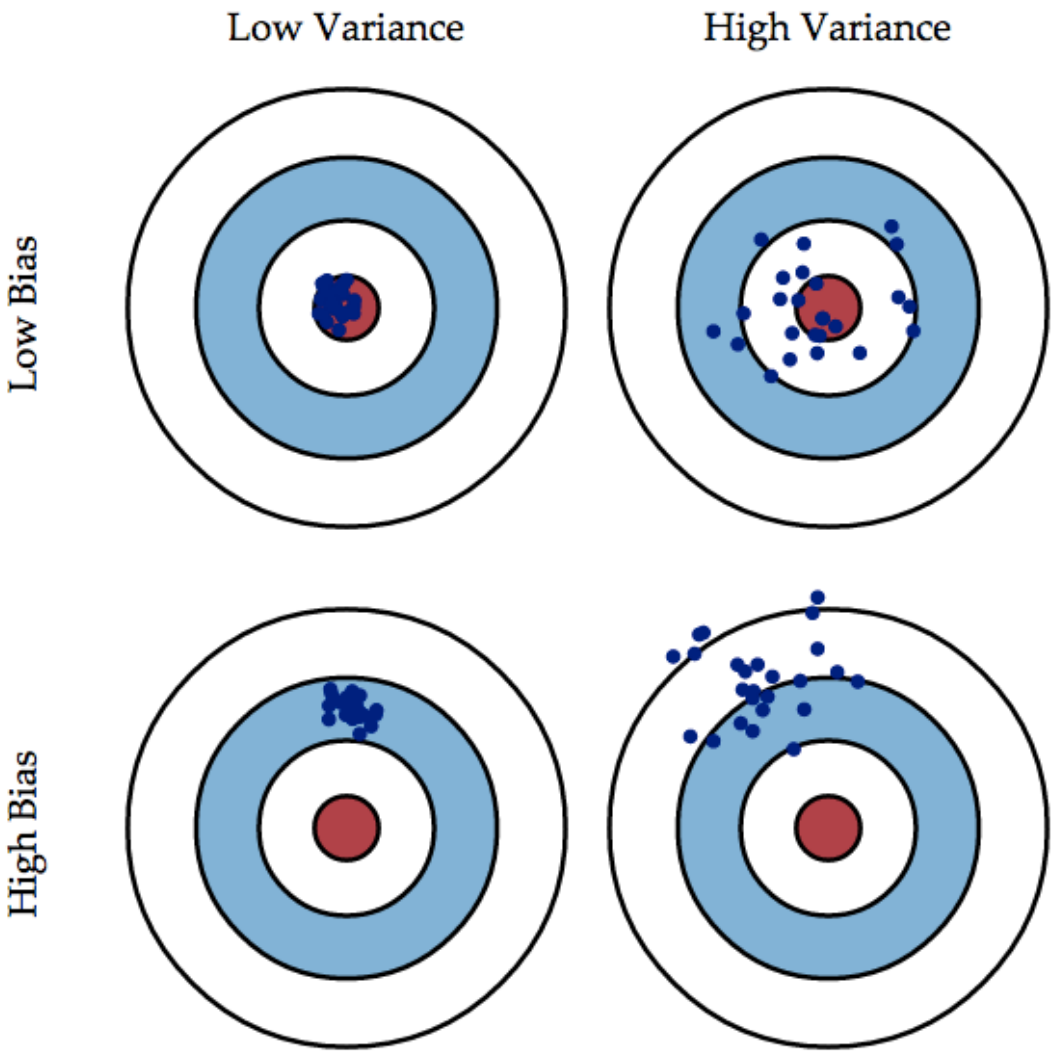
- then, for $\hat{\theta}$ the MLE, classical theory says $\psi(\mathbb{P}_{\hat{\theta}})$ is asymptotically efficient (minimax optimal) under smoothness

But the assumption $\mathbb{P} = \mathbb{P}_\theta$ is often too restrictive

- can we really know the exact form of $\mathbb{P}$ up to finite-dim $\theta$? even when $Z$ contains many (continuous) components?

Further this approach might discard known structure

- propensity score is known in trial, but factors out of likelihood


Low Variance
High Variance
Low Bias
High Bias

# Nonparametric plug-in

So parametric approach is likely misspecified and thus also biased

- suggests using more flexible nonparametric approach
- non/semiparametric $= \mathcal{P}$ not indexed by finite-dim. parameter

The most natural nonparametric approach is a plug-in $\psi(\hat{\mathbb{P}})$

- where $\hat{\mathbb{P}}$ is some initial estimator of $\mathbb{P}$ or relevant components

Such estimators are generally not $\sqrt{n}$-consistent, and further do not yield CIs without impractical undersmoothing

- few special cases, which require specific estimators $\hat{\mathbb{P}}$, strong smoothness assumptions, and undersmoothing

# Ex: integrated density squared

Consider $\psi = \int p(x)^2 \, dx = \mathbb{E}\{p(X)\}$. A natural plug-in is

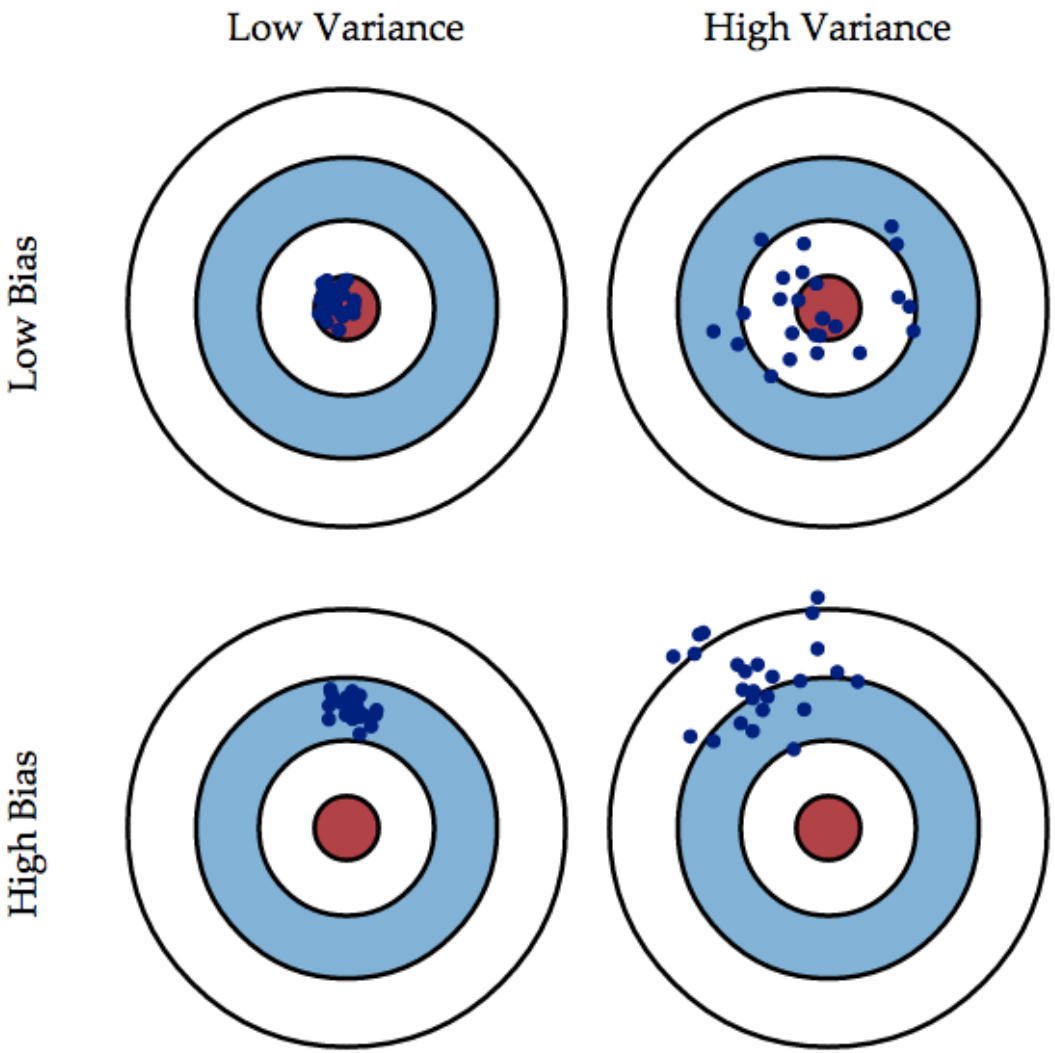
$$\hat{\psi} = \mathbb{P}_n\{\hat{p}(X)\}.$$

Using tools from next half of tutorial, we can show

$$\hat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})p + \mathbb{P}(\hat{p} - p) + o_{\mathbb{P}}(1/\sqrt{n})$$

The first term on RHS is $O_{\mathbb{P}}(1/\sqrt{n})$ by CLT

- but second is $|\mathbb{P}(\hat{p} - p)| \leq \|\hat{p} - p\|$ by Cauchy-Schwarz, and minimax lower bound for this is $O_{\mathbb{P}}(n^{-1/(2+d/\beta)})$ for Hölder($\beta$)


Low Variance
High Variance
Low Bias
High Bias

# Ex: integrated density squared

However if $p$ is Hölder($\beta$) and we use $\hat{\psi}$ with kernel estimator

$$\hat{p}(t) = \mathbb{P}_n \left\{ \frac{1}{h} K \left( \frac{X - t}{h} \right) \right\}$$

It is possible to show that

$$\mathbb{P}(\hat{p} - p) = (\mathbb{P}_n - \mathbb{P})p + o_{\mathbb{P}}(1/\sqrt{n}) = O_{\mathbb{P}}(1/\sqrt{n})$$

IF: 1. $K$ is higher-order kernel and $h \sim n^{-1/(\beta+d)}$ (undersmoothing)

2. $\beta > d$ (strong smoothness assumption)

# Nonparametric plug-ins

Similar results hold for ATE functional $\mathbb{E}\{\mathbb{E}(Y \mid X, A = 1)\}$

- Hahn (1998), Abadie & Imbens (2006)

This discussion begs several pressing questions:

- could we have done better than plug-in when $\beta > d$?
- is there any hope under less stringent smoothness ($\beta \leq d$)?
- what if we want to rely on structure beyond smoothness?
- what if we want to use other generic nuisance estimators?

$\rightarrow$ leads to nonparametric efficiency theory & influence functions

# What is a semiparametric model?

First let's consider some examples of semiparametric models

Statistical model: a set $\mathcal{P}$ of possible distributions containing $\mathbb{P}$

- for parametric models we assume $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}^d\}$, e.g., $Z \sim N(\mu, \Sigma)$ so that $\theta = (\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{R}^{p \times p}$

**Semiparametric models** $\mathcal{P}$ have infinite-dim. component

- i.e., *cannot be indexed by only finitely many real parameters*

# Examples

The simplest example of a semiparametric model is...

- a nonparametric model, where $\mathcal{P}$ includes all distn.s

Thus semiparametrics *includes nonparametrics as a special case*

- hence the first semiparametric models were called 'parametric-nonparametric models' by Begun et al. (1983)

Other examples:

- GEE, Cox model, experiment w/known propensity score, etc.

# Nonparametric causal models

Common to know/assume structure on treatment process

Often we may not know anything about *outcome OR treatment*

- then a nonparametric model makes most sense
- this is the perspective I often take in observational studies

Remarkably, can often still make progress in nonparametric models

- e.g., $\sqrt{n}$-rates of convergence, valid CIs, even when incorporating machine learning
- but we need influence functions and empirical process theory / sample splitting

# Lower bounds

Given a target parameter $\psi$ and model $\mathcal{P}$ (nonparametric)

- now turn to: lower bds/benchmarks for estimation error
- i.e., how well can we hope to estimate $\psi$ over the model $\mathcal{P}$?

A natural way to characterize optimality is via minimax rates

$$R_n = \inf_{\widehat{\tau}} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\widehat{\psi} - \psi(P)|$$

i.e., the best possible (worst-case) error, *across all estimators*

Lower bounds have crucial implications, theoretical & practical

- precisely illustrates fundamental limits / statistical difficulty
- gives benchmark for best possible performance

# Classical lower bounds

First consider classical parametric setup

- i.e., Cramer-Rao lower bound

C-R says for "smooth" $\{\mathbb{P}_\theta : \theta \in \mathbb{R}^d\}$ & *any* unbiased $\widehat{\psi}$

$$\text{var}(\widehat{\psi}) \geq \psi'(\theta)^2 / \text{var}\{s_\theta(Z)\}$$

where $s_{\theta^*}(z) = \frac{\partial}{\partial\theta} \log p_\theta(z)|_{\theta=\theta^*}$ is the score

This is also a lower bound in local minimax sense: for any $\hat{\psi}$

$$\inf_{\delta>0} \liminf_{n\to\infty} \sup_{\|\theta'-\theta\|<\delta} \mathbb{E}_{\theta'} \left[ n \left\{ \hat{\psi} - \psi(\theta') \right\}^2 \right] \geq \frac{\psi'(\theta)^2}{\text{var}\{s_\theta(Z)\}}$$

# Parametric submodels

How to exploit C-R to find lower bound for nonparametric $\mathcal{P}$?

Let parametric submodel $\mathcal{P}_\epsilon = \{\mathbb{P}_\epsilon : \epsilon \in \mathbb{R}\} \subseteq \mathcal{P}$ with $\mathbb{P} = \mathbb{P}_0$

- thus $\mathcal{P}_\epsilon$ respects the model and contains the truth
- note: this is technical device, not for use with real data

A common choice of submodel is, for a mean-zero $h$,

$$p_\epsilon(z) = d\mathbb{P}(z)\{1 + \epsilon h(z)\}$$

where $\|h\|_\infty \leq M$ and $\epsilon < 1/M$ so that $p_\epsilon(z) \geq 0$

Now any lower bound for $\mathcal{P}_\epsilon$ is also a lower bound for $\mathcal{P}$

- always easier to estimate $\psi$ under smaller $\mathcal{P}_\epsilon$ than larger $\mathcal{P}$

# Stein (1956)

## EFFICIENT NONPARAMETRIC TESTING AND ESTIMATION

CHARLES STEIN
STANFORD UNIVERSITY

### 1. Introduction

It is customary to treat nonparametric statistical theory as a subject completely different from parametric theory. In this paper, I try to study one of the more obvious connections between the two subjects. Clearly a nonparametric problem is at least as difficult as any of the parametric problems obtained by assuming we have enough knowledge of the unknown state of nature to restrict it to a finite-dimensional set. For a problem in which one wants to estimate a single real-valued function of the unknown state of nature it frequently happens that (in a sense made somewhat more precise in section 2 and, for special cases, in later sections) there is, through each state of nature, a one-dimensional problem which is, for large samples, at least as difficult (to a first approximation) as any other finite-dimensional problem at that point. If a procedure does essentially as well, for large samples, as one could do for each such one-dimensional problem, one is justified in considering the procedure efficient for large samples. If there is no such procedure, one may be forced to adopt a less severe definition of efficiency, as suggested by Wolfowitz [1].

# Nonparametric lower bounds

Since any lower bound for $\mathcal{P}_\epsilon$ is also one for $\mathcal{P}$, the best and most informative is the greatest such lower bound

- $\mathcal{P}_\epsilon$ is smooth & finite-dim, so C-R tells us how to compute!

The Cramer-Rao bound for $\mathcal{P}_\epsilon$ is

$$\psi'(\mathbb{P}_\epsilon)^2/\text{var}(s_\epsilon)$$

where $\psi'(\mathbb{P}_\epsilon) = \frac{\partial}{\partial\epsilon}\psi(\mathbb{P}_\epsilon)|_{\epsilon=0}$ and $s_\epsilon = s_\epsilon(z) = \frac{\partial}{\partial\epsilon}\log p_\epsilon(z)|_{\epsilon=0}$ is the submodel score function

- e.g., $s_\epsilon(z) = h(z)$ for previous common submodel

# Distributional Taylor expansion

What can we say about the derivative in the numerator?

Suppose $\psi$ is smooth & satisfies distributional Taylor expansion

$$\psi(\overline{P}) - \psi(P) = \int \varphi(\overline{P})\ d(\overline{P} - P) + R_2(\overline{P}, P)$$

for mean-zero & finite-var $\varphi(z; P)$, and $R_2$ a 2nd-order remainder

- this is sometimes called von Mises expansion

$R_2$ being 2nd-order means it only depends on *products* or *squares* of differences between $P$ and $\overline{P}$

# von Mises (1947)

## ON THE ASYMPTOTIC DISTRIBUTION OF DIFFERENTIABLE STATISTICAL FUNCTIONS

By R. v. Mises

*Harvard University*

Table of Contents

PAGE

Introduction ..... 309

*Part I. Preliminary Theorems.*

1. Asymptotically Equal Distributions ..... 311
2. Special Class of Statistical Functions: Quantics ..... 312
3. Asymptotic Expectation of Excess-Power Products ..... 314
4. Asymptotic Expectation and Variance of Quantics ..... 317
5. Final Statement on the Limit of Expectation of Quantics ..... 320
6. Theorem on Products of n Functions ..... 322

*Part II. Differentiable Statistical Functions.*

1. Definitions ..... 323
2. Taylor Development ..... 325
3. General Theorem ..... 327
4. Illustrations ..... 329

*Part III. Second-Type Asymptotic Distribution.*

1. Statement of the Problem ..... 331
2. Characteristic Function ..... 332
3. Asymptotic Value of $Q_n(u)$ ..... 335

# Pfanzagl (1982)

Let $\kappa: \mathfrak{P} \to \mathbb{R}$ be a functional. For asymptotic theory, the *local properties* of this functional are essential, i.e., its behavior in contiguous neighborhoods. These local properties determine how good optimal tests and estimators can be asymptotically. The mathematical construct suitable for this purpose is the gradient.

<u>4.1.1. Definition</u>. A function $\kappa^{\bullet}(\cdot,P) \in \mathscr{L}_*(P)$ is a *gradient* of $\kappa$ at $P$ for $\mathfrak{P}$ if for every $g \in T(P,\mathfrak{P})$ and every path $P_{t,g}$, $t \downarrow 0$, in $\mathfrak{P}$ with derivative $g$,

$$\kappa(P_{t,g}) - \kappa(P) = tP(\kappa^{\bullet}(\cdot,P)g) + o(t) .$$

If $\kappa$ admits a gradient at $P$ we call $\kappa$ *differentiable* at $P$.

# Distributional Taylor expansion

The von Mises expansion is just an infinite-dimensional / distributional analog of Taylor expansion

- $\varphi(z; Q)$ acts as a usual derivative term
- describes how $\psi$ changes locally, when moving from $P$ to $\overline{P}$

When $Z \in \{1, ..., k\}$ is discrete so $P = \{p_1, ..., p_k\}$, the von Mises expansion reduces to usual Taylor expansion with $R_2(\overline{P}, P)$ equal to

$$\psi(\overline{p}_1, ..., \overline{p}_k) - \psi(p_1, ..., p_k) - \sum_{j=1}^{k} \frac{\partial}{\partial t_j} \psi(t_1, ..., t_k) \Big|_{t=\overline{p}} (\overline{p}_j - p_j)$$

# Pathwise differentiability

A related notion of smoothness is pathwise differentiability:

$$\frac{\partial}{\partial \epsilon}\psi(\mathbb{P}_\epsilon)\Big|_{\epsilon=0} = \int \varphi(z;\mathbb{P})s_\epsilon(z)\ d\mathbb{P}(z)$$

for every smooth submodel $P_\epsilon$.

This is implied by the von Mises expansion, under regularity conditions, by taking $(P, Q) = (P_\epsilon, P)$, differentiating both sides, and noting that $R_2$ being second order means

$$\frac{\partial}{\partial \epsilon}R_2(P, P_\epsilon)\Big|_{\epsilon=0} = 0.$$

which is also essentially Neyman orthogonality

# Pathwise differentiability is important

The pathwise differentiability of $\psi$ as a map from $\mathcal{P} \to \mathbb{R}$, i.e., that

$$\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) = \int \varphi(\overline{\mathbb{P}}) \, d(\overline{\mathbb{P}} - \mathbb{P}) + R_2(\overline{\mathbb{P}}, \mathbb{P})$$

for mean-zero/finite-variance $\varphi$, is really key

- will see later that this not only gives lower bounds, but also suggests how to bias-correct a plug-in

Typically say any $\varphi$ satisfying above is an influence function for $\psi$

- or influence curve or gradient
- however, don't confuse with an IF for *an estimator* $\widehat{\psi}$

# Efficient influence function

If $\psi$ is pathwise differentiable, then the greatest lower bound is

$$\sup_{\mathcal{P}_\epsilon} \frac{\psi'(\mathbb{P}_\epsilon)^2}{\mathbb{E}(s_\epsilon^2)} = \sup_h \frac{\mathbb{P}(\varphi h)^2}{\mathbb{P}(h^2)} \leq \mathbb{P}(\varphi^2)$$

inequality by Cauchy-Schwarz, equality if $\varphi$ is valid submodel score

- valid score $\iff$ $\varphi$ is in tangent space (closure of score space)

Therefore $\mathbb{P}(\varphi^2) = \text{var}(\varphi)$ is nonparametric analog of CR bound!

- we call $\varphi$ the efficient influence function

This is hugely important - implies no estimator can beat

$$\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N(0, \text{var}(\varphi))$$

in a (local) asymptotic minimax sense!

# Local minimaxity

**Theorem (Corollary 2.6, van der Vaart (2002))**

*Let $\psi : \mathcal{P} \mapsto \mathbb{R}$ be pathwise differentiable with efficient influence function $\varphi$. Assume the model is nonparametric or the tangent space is a convex cone. Then*

$$\inf_{\delta>0} \liminf_{n\to\infty} \sup_{TV(P,Q)<\delta} n\, \mathbb{E}_Q\left[\{\widehat{\psi} - \psi(Q)\}^2\right] \geq var\{\varphi(Z;P)\}$$

*for any estimator sequence $\widehat{\psi} = \widehat{\psi}_n$.*

# Some history from Bickel, Klaasen, Ritov, Wellner (1993)

Chapter 3 contains our basic answers to questions A and B. The material in section 3.2 on tangent spaces is basic to the rest of the monograph. In section 3.3 we give an approach to information bounds for semiparametric models which derives from a more nonparametric view involving pathwise (Hadamard) differentiation of Euclidean-valued functions $\nu$. It stems from figure 2.4.2 and propositions 2.4.1.B, 2.4.2, and 2.4.3. This approach began in the work of Levit (1978) and Koshevnik and Levit (1976), and has been further developed by Pfanzagl and Wefelmeyer (1982), (1985), Pfanzagl (1990), and Van der Vaart (1988a). A key element of our approach is the emphasis on the asymptotic linearity of efficient procedures and identification of the linear approximation—what we call the *efficient influence function*. An alternative approach to information bounds for semiparametric models based on scores and score operators developed by Begun, Hall, Huang, and Wellner (1983) is outlined in section 3.4. This approach, which is based on what we call the *efficient score function*, continues the theme begun in proposition 2.4.1.A and figure 2.4.1 of section 2.4.

These same themes, but with different emphases and with greater generality, have been developed in the fundamental works of Le Cam (1986), Le Cam and Yang (1990), and Ibragimov and Has'minskii (1981).

# Taking a step back

Let's review what we've learned here

We have a lower bound, indicating that no estimator can be more efficient than

$$\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N(0, \mathrm{var}(\varphi))$$

in the asymptotic minimax sense, where $\varphi$ is the efficient influence function, i.e., a mean-zero finite-variance function satisfying

$$\frac{\partial}{\partial \epsilon}\psi(\mathbb{P}_\epsilon)\Big|_{\epsilon=0} = \int \varphi(z;\mathbb{P})s_\epsilon(z)\ d\mathbb{P}(z)$$

for all submodels and scores, and which is itself a score

- begs the question: is the bound sharp? can it be attained?

# Influence functions

Before we figure out whether/how the bound can be attained, let's look at some examples of influence functions $\varphi$

- it turns out if we know $\varphi$ then we can often construct efficient estimators under weak conditions (next part of tutorial)

There are 3 main ways I know of to derive IFs

1. most general: compute derivative $\psi'(\mathbb{P}_\epsilon)$ and solve for $\varphi$
2. often easier to pretend data are discrete and compute Gateaux derivative in direction of point mass contamination
3. mine: use chain/product rules w/ simple IFs as building blocks

# Influence function for mean

The simplest IF is for the mean $\psi = \mathbb{E}(Z) = \int z \, d\mathbb{P}(z)$:

$$\begin{aligned}\psi'(\mathbb{P}_\epsilon) &= \frac{\partial}{\partial \epsilon} \int z \, d\mathbb{P}_\epsilon(z) \Big|_{\epsilon=0} = \int z \frac{\partial}{\partial \epsilon} d\mathbb{P}_\epsilon(z) \Big|_{\epsilon=0} \\ &= \int z \left( \frac{\partial}{\partial \epsilon} \log d\mathbb{P}_\epsilon(z) \right) \, d\mathbb{P}_\epsilon(z) \Big|_{\epsilon=0} \\ &= \int (z - \psi) \left( \frac{\partial}{\partial \epsilon} \log d\mathbb{P}_\epsilon(z) \right) \Big|_{\epsilon=0} d\mathbb{P}(z)\end{aligned}$$

Thus EIF is $\varphi(z; \mathbb{P}) = z - \psi$. Also von Mises holds with $R_2 = 0$.

Here it is easy to see the bound is attained. For $\hat{\psi} = \mathbb{P}_n(Z)$

$$\text{var}\left\{ \sqrt{n}(\hat{\psi} - \psi) \right\} = \text{var}(Z) = \text{var}(\varphi)$$

# Influence function for integrated square density

For $\psi = \int p(z)^2 \, dz$ we have

$$\begin{aligned}
\psi'(\mathbb{P}_\epsilon) &= \frac{\partial}{\partial \epsilon} \int p_\epsilon(z)^2 \, dz \Big|_{\epsilon=0} = \int \frac{\partial}{\partial \epsilon} p_\epsilon(z)^2 \, dz \Big|_{\epsilon=0} \\
&= \int 2 p_\epsilon(z) \left( \frac{\partial}{\partial \epsilon} \log p_\epsilon(z) \right) p_\epsilon(z) \, dz \Big|_{\epsilon=0} \\
&= \int 2 \Big\{ p(z) - \psi \Big\} \left( \frac{\partial}{\partial \epsilon} \log p_\epsilon(z) \right) \Big|_{\epsilon=0} p(z) \, dz
\end{aligned}$$

so EIF is $\varphi = 2(p - \psi)$. Here the von Mises remainder is

$$R_2(\mathbb{P}, \mathbb{Q}) = - \int \Big\{ p(z) - q(z) \Big\}^2 dz$$

# Gateaux derivative approach

For more complicated functionals other techniques can be quicker

von Mises and Hampel in the 1940s/1970s used Gateaux derivative of $\psi$ at $\mathbb{P}(z')$ in direction of point mass $(\delta_z - \mathbb{P}(z'))$

$$\frac{\partial}{\partial \epsilon} \psi \Big\{ (1-\epsilon)\mathbb{P}(z') + \epsilon \delta_z \Big\} \Big|_{\epsilon=0} = \varphi(z)$$

This is technically *only valid for discrete Z*

- but can usually pretend discrete, and general EIF will be clear

Why does this work? Note LHS is just pathwise derivative for particular submodel with $h = \delta_z$ ... $\implies \int \varphi s_\epsilon \, d\mathbb{P} = \varphi(z)$

# Influence function for conditional mean

Let $Z = (X, Y)$ with $X$ discrete, and suppose $\psi = \mathbb{E}(Y \mid X = x)$.

One can similarly show that in a nonparametric model

$$\varphi(Z; \mathbb{P}) = \frac{\mathbb{1}(X = x)}{\mathbb{P}(X = x)} \Big\{ Y - \mathbb{E}(Y \mid X = x) \Big\}$$

A quick way to see this is to note $\psi = \frac{\mathbb{E}\{Y \mathbb{1}(X=x)\}}{\mathbb{E}\{\mathbb{1}(X=x)\}}$

- we know EIF for numerator and denominator, now can use fact that EIF is a derivative to justify quotient rule
- this trick can also work for more complicated parameters: many are structured combinations of conditional means

# The strategy I use

TRICK 1. Pretend the data are discrete.

TRICK 2. Treat IFs as derivatives, allowing use of differentiation rules.

For example, let $\mathbb{IF} : \Psi \to L_2(\mathbb{P})$ map functional $\psi : \mathcal{P} \to \mathbb{R}$ to its IF $\varphi(z) \in L_2(\mathbb{P})$ in a nonparametric model.

Then:

TRICK 2a. *(product rule)* $\mathbb{IF}(\psi_1\psi_2) = \mathbb{IF}(\psi_1)\psi_2 + \psi_1\mathbb{IF}(\psi_2)$

TRICK 2b. *(chain rule)* $\mathbb{IF}(f(\psi)) = f'(\psi)\mathbb{IF}(\psi)$

TRICK 3. Use influence function building blocks, e.g.,
$\mathbb{IF}(\mathbb{E}(Y \mid X = x)) = \frac{\mathbb{1}(X=x)}{\mathbb{P}(X=x)}\{Y - \mathbb{E}(Y \mid X = x)\}$.

# Influence function for ATE

For $\psi = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1)\}$ under nonparametric model, EIF is

$$\varphi(Z; \mathbb{P}) = \frac{A}{\pi(X)}\Big\{Y - \mu(X)\Big\} + \mu(X) - \psi$$

where $\pi(X) = \mathbb{P}(A = 1 \mid X)$ and $\mu(X) = \mathbb{E}(Y \mid X, A = 1)$.

The second-order von Mises remainder is

$$R_2(\overline{\mathbb{P}}, \mathbb{P}) = \int \frac{1}{\overline{\pi}(x)}\Big\{\pi(x) - \overline{\pi}(x)\Big\}\Big\{\mu(x) - \overline{\mu}(x)\Big\} d\mathbb{P}(x)$$

$\rightarrow$ *foreshadowing*: double robustness...

# ATE proof using Strategies 1-2

## A Appendix

### A.1 Two Derivations of the Influence Function of the ATE

#### A.1.1 Integral Equation Approach

To derive the influence function via the pathwise differentiability condition, one needs to first calculate the derivative of the parameter

$$\frac{\partial}{\partial \epsilon}\psi(\mathbb{P}_\epsilon)\Big|_{\epsilon=0} \tag{15}$$

on any smooth parametric submodel $\mathbb{P}_\epsilon$ (satisfying $\mathbb{P}_0 = \mathbb{P}$), set it equal to the inner product (i.e., covariance)

$$\int \varphi(z)\left(\frac{\partial}{\partial \epsilon}\log d\mathbb{P}_\epsilon(z)\right)\Big|_{\epsilon=0} d\mathbb{P}(z) \tag{16}$$

and solve this integral equation for the influence function $\varphi(z)$. In what follows we tackle these two tasks separately, for the average treatment effect parameter.

**Derivative of parameter**

First we will compute the derivative of the parameter in (15) for $\mathbb{P}_\epsilon$ a generic smooth parametric submodel satisfying $\mathbb{P}_0 = \mathbb{P}$. Note that the submodel score decomposes as a sum

$$\begin{aligned} s_\epsilon(z) \equiv \frac{\partial}{\partial \epsilon}\log d\mathbb{P}_\epsilon(z) &= \frac{\partial}{\partial \epsilon}\log d\mathbb{P}_\epsilon(y \mid x, a) + \frac{\partial}{\partial \epsilon}\log d\mathbb{P}_\epsilon(a \mid x) + \frac{\partial}{\partial \epsilon}\log d\mathbb{P}_\epsilon(x) \\ &\equiv s_\epsilon(y \mid x, a) + s_\epsilon(a \mid x) + s_\epsilon(x). \end{aligned}$$

Therefore, using the definition of the parameter, its derivative equals

$$\begin{aligned} \frac{\partial}{\partial \epsilon}\psi(\mathbb{P}_\epsilon) &= \frac{\partial}{\partial \epsilon}\int\int y \; d\mathbb{P}_\epsilon(y \mid x, 1) \; d\mathbb{P}_\epsilon(x) \\ &= \int\int y \; \frac{\partial}{\partial \epsilon} d\mathbb{P}_\epsilon(y \mid x, 1) \; d\mathbb{P}_\epsilon(x) + \int\int y \; d\mathbb{P}_\epsilon(y \mid x, 1) \; \frac{\partial}{\partial \epsilon} d\mathbb{P}_\epsilon(x) \\ &= \int\int y \; s_\epsilon(y \mid x, 1) \; d\mathbb{P}_\epsilon(y \mid x, 1) \; d\mathbb{P}_\epsilon(x) + \int\int y \; d\mathbb{P}_\epsilon(y \mid x, 1) \; s_\epsilon(x) \; d\mathbb{P}_\epsilon(x), \end{aligned}$$

where in the third equality we used the derivative of a logarithm, i.e., that $\frac{\partial}{\partial \epsilon}\log f_\epsilon(z) = \frac{\partial}{\partial \epsilon} f_\epsilon(z)/f_\epsilon(z)$. Therefore the derivative of the ATE on a submodel at $\epsilon = 0$ is given by

$$\frac{\partial}{\partial \epsilon}\psi(\mathbb{P}_\epsilon)\Big|_{\epsilon=0} = \int\int \Big\{y \; s_0(y \mid x, 1) + \mu(x)s_0(x)\Big\} \; d\mathbb{P}(y \mid x, 1) \; d\mathbb{P}(x). \tag{17}$$

Now our task is to write this derivative in inner product form as in (16).

**Expressing as inner product**

Specifically here we aim to write the derivative (17) in the form

$$\int \varphi(z)\Big\{s_0(y \mid x, a) + s_0(a \mid x) + s_0(x)\Big\} \; d\mathbb{P}(z)$$



for a mean zero function $\varphi$, i.e., we need to solve the integral equation

$$\int \Big\{\int y \; s_0(y \mid x, 1) \; d\mathbb{P}(y \mid x, 1) + \mu(x)s_0(x)\Big\} \; d\mathbb{P}(x) = \int \varphi(z)s_0(z) \; d\mathbb{P}(z) \tag{18}$$

for $\varphi$. This does not have an obvious solution if the correct form for $\varphi$ is not already known (in which case one can just check that the equation holds). A potentially time-consuming and error-bound path is to keep conjecturing candidates for $\varphi$, until one works.

Instead it can be helpful to start by decomposing the influence function as $\varphi(z) = \varphi_y(y, x, a) + \varphi_a(a, x) + \varphi_x(x)$ where

$$\begin{aligned} \int \varphi_y(y, x, a) \; d\mathbb{P}(y \mid x, a) &= 0 \\ \int \varphi_a(a, x) \; d\mathbb{P}(a \mid x) &= 0 \\ \int \varphi_x(x) \; d\mathbb{P}(x) &= 0. \end{aligned} \tag{19}$$

Note that this decomposition holds for any random variable $\varphi = \varphi(Z)$ by simply centering appropriately, i.e., defining $\varphi_y(Y, X, A) = \varphi - \mathbb{E}(\varphi \mid X, A)$, $\varphi_a(A, X) = \mathbb{E}(\varphi \mid X, A) - \mathbb{E}(\varphi \mid X)$, and $\varphi_x(X) = \mathbb{E}(\varphi \mid X) - \mathbb{E}(\varphi)$. Importantly, with this decomposition, the inner product on the right-hand side of (18) simplifies to

$$\begin{aligned} &\int \Big\{\varphi_y(y, x, a) + \varphi_a(a, x) + \varphi_x(x)\Big\}\Big\{s_0(y \mid x, a) + s_0(a \mid x) + s_0(x)\Big\} \; d\mathbb{P}(z) \\ &\quad = \int \Big\{\varphi_y(y, x, a)s_0(y \mid x, a) + \varphi_a(a, x)s_0(a \mid x) + \varphi_x(x)s_0(x)\Big\} \; d\mathbb{P}(z) \end{aligned}$$

by virtue of the restrictions in (19), and the fact that $s_0(y \mid x, a)$, $s_0(a \mid x)$, and $s_0(x)$ are score functions and so similarly have conditional mean zero, i.e.,

$$\begin{aligned} \int s_0(y \mid x, a) \; d\mathbb{P}(y \mid x, a) &= 0 \\ \int s_0(a \mid x) \; d\mathbb{P}(a \mid x) &= 0 \\ \int s_0(x) \; d\mathbb{P}(x) &= 0. \end{aligned}$$

By using the decomposition (19) we have essentially transformed our problem from solving one big integral equation to solving three smaller and easier integral equations. Specifically we now need to solve for $(\varphi_y, \varphi_a, \varphi_x)$ in

$$\begin{aligned} &\int\int \Big\{y \; s_0(y \mid x, 1) + \mu(x)s_0(x)\Big\} \; d\mathbb{P}(y \mid x, 1) \; d\mathbb{P}(x) \\ &\quad = \int \Big\{\varphi_y(y, x, a)s_0(y \mid x, a) + \varphi_a(a, x)s_0(a \mid x) + \varphi_x(x)s_0(x)\Big\} \; d\mathbb{P}(z) \end{aligned}$$

This immediately leads us to the choices $\varphi_a(a, x) = 0$ and $\varphi_x(x) = \mu(x) - \mathbb{E}\{\mu(X)\} = \mu(x) - \psi$.

# ATE proof using Strategies 1-2

The remaining task is to find $\varphi_y(y, x, a)$, which based on the above amounts to solving the integral equation

$$\int\int y\ s_0(y \mid x, 1)\ d\mathbb{P}(y \mid x, 1)\ d\mathbb{P}(x) = \int \varphi_y(y, x, a) s_0(y \mid x, a)\ d\mathbb{P}(z)$$

Note the derivative term on the left-hand side above equals

$$\begin{aligned}\int\int y\ s_0(y \mid x, 1)\ d\mathbb{P}(y \mid x, 1)\ d\mathbb{P}(x) &= \int\int\int y\ s_0(y \mid x, 1)\ d\mathbb{P}(y \mid x, 1)\ d\mathbb{P}(a \mid x)\ d\mathbb{P}(x) \\ &= \int\int\int \frac{ay}{\pi(x)}\ s_0(y \mid x, a)\ d\mathbb{P}(y \mid x, a)\ d\mathbb{P}(a \mid x)\ d\mathbb{P}(x)\end{aligned}$$

where in the first equality we simply introduced the treatment distribution $d\mathbb{P}(a \mid x) = a\pi(x) + (1-a)(1-\pi(x))$, and in the second we multiplied by the indicator $a$ to pick out the required $s_0(y \mid x, 1)$ term, and then divided by $\pi(x)$ to cancel out the additional $\pi(x)$ term that appears due to averaging the indicator $a$. We note this logic is perhaps still a bit mysterious for those who have not used it before. The term $ay/\pi(x)$ has the right mean, but needs to be centered so that it has conditional mean zero and thus satisfies the first line of (19). This leads to the choice

$$\varphi_y(y, a, x) = \frac{ay}{\pi(x)} - \int \frac{ay}{\pi(x)}\ d\mathbb{P}(y \mid x, a) = \frac{a}{\pi(x)}\Big\{y - \mu(x)\Big\}.$$

Combining with $\varphi_a(a, x)$ and $\varphi_x(x)$ above gives the overall influence function

$$\varphi(Z) = \frac{A}{\pi(X)}\Big\{Y - \mu(X)\Big\} + \mu(X) - \psi. \tag{20}$$

Since the work above shows that this satisfies the pathwise differentiability condition for any sufficiently smooth parametric submodel, it is the influence function.

### A.1.2 Gateaux Derivative Approach

The Gateaux derivative approach can be viewed as a special case of the above, where one uses a particular choice of parametric submodel, for which the pathwise derivative is actually equal to the influence function, rather than an integral equation that needs to be solved. This is accomplished by using a submodel whose score is a point mass, as described in the main text.

One simple such submodel is given by $\mathbb{P}^*_\epsilon(z) = (1-\epsilon)\mathbb{P}(z) + \epsilon\delta_Z$, where $\delta_Z$ is the Dirac measure at $z = Z$. Since $z$ is discrete, we can just work with the mass function $p^*_\epsilon(z) = (1-\epsilon)p(z) + \epsilon\mathbb{1}(z = Z)$. First note that for the submodel $p^*_\epsilon(z)$ we have

$$\begin{aligned}p^*_\epsilon(y \mid x, a) &= \frac{p^*_\epsilon(z)}{p^*_\epsilon(a, x)} = \frac{(1-\epsilon)p(z) + \epsilon\mathbb{1}(z = Z)}{(1-\epsilon)p(a, x) + \epsilon\mathbb{1}(a = A, x = X)} \\ p^*_\epsilon(a \mid x) &= \frac{p^*_\epsilon(a, x)}{p^*_\epsilon(x)} = \frac{(1-\epsilon)p(a, x) + \epsilon\mathbb{1}(a = A, x = X)}{(1-\epsilon)p(x) + \epsilon\mathbb{1}(x = X)} \\ p^*_\epsilon(x) &= (1-\epsilon)p(x) + \epsilon\mathbb{1}(x = X)\end{aligned}$$



and

$$\begin{aligned}\frac{\partial}{\partial\epsilon} p^*_\epsilon(y \mid x, a)\Big|_{\epsilon=0} &= \frac{\mathbb{1}(z = Z) - p(z)}{(1-\epsilon)p(a, x) + \epsilon\mathbb{1}(a = A, x = X)}\Big|_{\epsilon=0} \\ &\quad - p^*_\epsilon(y \mid x, a)\frac{\mathbb{1}(a = A, x = X) - p(a, x)}{(1-\epsilon)p(a, x) + \epsilon\mathbb{1}(a = A, x = X)}\Big|_{\epsilon=0} \\ &= \frac{\mathbb{1}(z = Z) - p(z)}{p(a, x)} - p(y \mid x, a)\frac{\mathbb{1}(a = A, x = X) - p(a, x)}{p(a, x)} \\ &= \mathbb{1}(a = A, x = X)\left\{\frac{\mathbb{1}(y = Y) - p(y \mid x, a)}{p(a, x)}\right\}\end{aligned} \tag{21}$$

where the first equality follows from the chain rule, and the rest by just rearranging.

Now we evaluate the parameter on the submodel, differentiate, and set $\epsilon = 0$, which gives

$$\begin{aligned}\frac{\partial}{\partial\epsilon}\psi(p^*_\epsilon)\Big|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\sum_{x,y} y\ p^*_\epsilon(y \mid x, 1)\ p^*_\epsilon(x)\Big|_{\epsilon=0} \\ &= \sum_{x,y} y\left\{\frac{\partial}{\partial\epsilon}p^*_\epsilon(y \mid x, 1)\ p^*_\epsilon(x) + p^*_\epsilon(y \mid x, 1)\ \frac{\partial}{\partial\epsilon}p^*_\epsilon(x)\right\}\Big|_{\epsilon=0} \\ &= \sum_{x,y} y\left[\mathbb{1}(1 = A, x = X)\left\{\frac{\mathbb{1}(y = Y) - p(y \mid x, 1)}{p(1, x)}\right\}p(x) + p(y \mid x, 1)\Big\{\mathbb{1}(x = X) - p(x)\Big\}\right] \\ &= \frac{A}{\pi(X)}\Big\{Y - \mu(X)\Big\} + \mu(X) - \psi\end{aligned}$$

where the second equality follows by the chain rule, the third by substituting in the expression in (21), and the fourth rearranging.

Therefore the Gateaux derivative approach gives the same influence function as the more involved integral equation approach, though even the Gateaux derivative required more than a page of calculations, and some care with the submodel derivatives. Note also that the influence function we arrived at is perfectly well-defined outside of the discrete setup, as long as the regression functions $\pi$ and $\mu$ are well-defined.

# ATE proof using Strategy 3

The influence function $\mathbb{IF}(\psi)$ for the ATE $\psi = \mathbb{E}\{\mu(X)\}$ is

$$\mathbb{IF}\left\{\sum_x \mu(x)p(x)\right\} = \sum_x \left[\mathbb{IF}\{\mu(x)\}p(x) + \mu(x)\mathbb{IF}\{p(x)\}\right]$$
$$= \sum_x \left[\frac{\mathbb{1}(X = x, A = 1)}{p(1, x)}\left\{Y - \mu(x)\right\}p(x) + \mu(x)\{\mathbb{1}(x = X) - p(x)\}\right]$$
$$= \frac{A}{\pi(X)}\left\{Y - \mu(X)\right\} + \mu(X) - \psi$$

where the first equality follows by Trick 1, the second by Trick 2a, the third by Trick 3, and the fourth by rearranging.

# Proper semiparametric models

We have focused on nonparametric models: $\mathcal{P}$ = all distributions

What if $\mathcal{P}$ is restricted in some way?

- i.e., we may know the propensity score $\pi(x)$, or may want to assume $\mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0) = \psi$

Then there are more/many influence functions

- why? we are reducing the set of submodel scores
- so the condition that $\psi'(\mathbb{P}_\epsilon) = \text{cov}(\varphi, s_\epsilon)$ is less stringent
- still only one efficient IF: the EIF that is valid submodel score, i.e., in tangent space (i.e., space of scores + limit pts)

# What happens w/ATE when PS is known?

If $\pi$ known, influence functions take the form

$$\varphi_g(Z;\mathbb{P}) = \frac{A}{\pi(X)}\Big\{Y - g(X)\Big\} + g(X) - \psi$$

for any $g$. The EIF is same as nonparametric case ($g = \mu$).

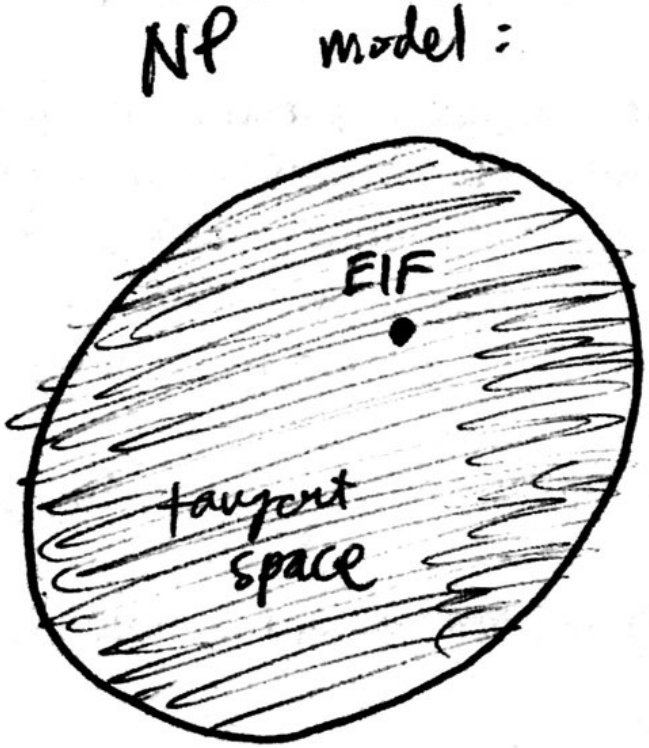
Why are these IFs not IFs in nonparametric model? Scores differ:

$$p_\epsilon(y \mid x, a)\pi(x)p_\epsilon(x) \quad \text{vs.} \quad p_\epsilon(y \mid x, a)p_\epsilon(a \mid x)p_\epsilon(x)$$

$$\implies s_\epsilon(y \mid x, a) + 0 + s_\epsilon(x) \quad \text{vs.} \quad s_\epsilon(y \mid x, a) + s_\epsilon(a \mid x) + s_\epsilon(x)$$

$\implies$ for $g \neq \mu$, we have $\text{cov}(\varphi_g, s_\epsilon) \neq \psi'(\mathbb{P}_\epsilon)$ for $s_\epsilon$ the NP score

- $\pi$ unknown $\implies$ $\varphi_g$ is valid submodel score but not valid IF
- $\pi$ known $\implies$ $\varphi_g$ is valid IF but not valid submodel score



# Aside: semiparametrics vs. projections

To me there is a distinct difference between semiparametric models where, e.g., $\pi$ is known vs. where parametric structure is assumed

- one can avoid the latter by using projections

For example it is common to assume $\gamma(X) = g(X; \psi)$ where $\gamma(X) = \mathbb{E}(Y \mid X, A = 1) - \mathbb{E}(Y \mid X, A = 0)$ is the CATE

Instead you could use $g$ nonparametrically, only for projection, e.g.,

$$\psi = \arg\min_{\psi^*} \mathbb{E}\Big[w(X)\{\gamma(X) - g(X; \psi^*)\}^2\Big]$$

This might yield small loss in efficiency (constants), but many advantages: big gains in interpretation, math, & honesty (imo)

# Recap

Where are we?

- we have a powerful nonparametric lower bound from the EIF, and know how to construct it in general cases

Now we need upper bounds!

- how should we construct estimators?
- under what conditions (if any) are they efficient?

Ideally we want estimators that allow general machine learning methods for estimating $\mathbb{P}$

- next section!

# Some notation

Throughout will use

$$\mathbb{P}(f) = \mathbb{P}\{f(Z)\} = \int f(z)\, d\mathbb{P}(z)$$

for expectations over new observation $Z$ (treating $f$ as fixed)

Thus $\mathbb{P}(\hat{f}) = \mathbb{E}(\hat{f} \mid Z_1, ..., Z_n)$ is random when $\hat{f}$ is (e.g., when estimated from sample)

- contrast with $\mathbb{E}\{\hat{f}(Z)\}$, which *averages over both* $\hat{f}$ *and* $Z$
- $\mathbb{E}\{\hat{f}(Z)\} \neq \mathbb{P}(\hat{f})$ unless $\hat{f} = f$ is a fixed function

# Plug-in bias

We're considering $\psi$ satisfying von Mises/Taylor expansion

$$\psi(\overline{P}) - \psi(P) = \int \varphi(\overline{P})\ d(\overline{P} - P) + R_2(\overline{P}, P)$$

for IF $\varphi(z; \mathbb{P})$ and $R_2$ a 2nd-order remainder.

This suggests plug-in estimators will typically have 1st order bias:

$$\psi(\hat{\mathbb{P}}) - \psi(\mathbb{P}) = -\int \varphi(\hat{\mathbb{P}})\ d\mathbb{P} + R_2(\hat{\mathbb{P}}, \mathbb{P})$$

(since $\int \varphi(\overline{P})\ d\overline{P} = 0$)

$\rightarrow$ Plug-ins are simple & intuitive, but typically suboptimal

# Plug-in bias

A plug-in estimator for the ATE is

$$\widehat{\psi}_{pi} = \mathbb{P}_n\{\widehat{\mu}(X)\}$$

If $\widehat{\mu}$ estimated on split independent of $\mathbb{P}_n$, the bias is

$$\mathbb{E}(\widehat{\psi}_{pi} - \psi) = \int \mathbb{E}\{\widehat{\mu}(x) - \mu(x)\} \, d\mathbb{P}(x),$$

which is just integrated bias of $\widehat{\mu}$ itself

- would typically be of same order as pointwise bias
- w/standard tuning, larger than $1/\sqrt{n}$ in nonparametric models

Intuitively the plug-in essentially makes parameter estimation as hard as regression estimation

# Bias correction

... any ideas about how to move forward? $\rightarrow$ estimate the bias!

From before, the bias is $-\int \varphi(z; \widehat{\mathbb{P}})\, d\mathbb{P}(z)$: just a mean!

So a natural bias-corrected estimator is

$$\widehat{\psi} = \psi(\widehat{\mathbb{P}}) + \mathbb{P}_n\{\varphi(Z; \widehat{\mathbb{P}})\}$$

sometimes called one-step estimator

- has a pretty long history, going back to the 1970s
- coincides in one way or another with many of your favorite semiparametric / doubly robust / TMLE / DML estimators

# Pfanzagl (1982)

<u>11.4.1. Improvement procedure</u>. Assume that $\kappa: \mathfrak{P} \to \mathbb{R}$ is differentiable. For $P \in \mathfrak{P}$ let $\kappa^{\bullet}(\cdot, P) \in T(P, \mathfrak{P})$ denote its canonical gradient. Let $P_n$, $n \in \mathbb{N}$, be an estimator-sequence attaining its values in $\mathfrak{P}$. Then the *improved estimator-sequence* $\kappa^n$, $n \in \mathbb{N}$, is defined by

$$(11.4.2) \quad \kappa^n(\underline{x}) := \kappa(P_n(\underline{x}, \cdot)) + n^{-1} \sum_{\nu=1}^{n} \kappa^{\bullet}(x_\nu, P_n(\underline{x}, \cdot)), \quad \underline{x} \in X^n, \ n \in \mathbb{N}.$$

The reader familiar with parametric statistical theory will easily recognize that (11.4.2) generalizes the improvement procedure based on the Newton-Raphson approximation to the solution of the likelihood equation (see, e.g., LeCam, 1956, p. 139).

# Some examples

One-step estimator for the expected density:

$$\widehat{\psi} = 2\mathbb{P}_n\{\widehat{p}(Z)\} - \int \widehat{p}(z)^2 \, dz.$$

One-step estimator for the ATE:

$$\widehat{\psi} = \mathbb{P}_n \left[ \widehat{\mu}(X) + \frac{A\{Y - \widehat{\mu}(X)\}}{\widehat{\pi}(X)} \right]$$

One-step estimator for the expected conditional covariance:

$$\widehat{\psi} = \mathbb{P}_n \left[ \left\{ A - \widehat{\pi}(X) \right\} \left\{ Y - \widehat{\mu}(X) \right\} \right]$$

# Relation to estimating equations/TMLE

The previous bias correction corresponds to estimating equation or one-step correction

An alternative is TMLE - this does the same bias correction, but approximately, by constructing $\hat{\mathbb{P}}^*$ such that

$$\psi(\hat{\mathbb{P}}^*) \approx \psi(\hat{\mathbb{P}}) + \mathbb{P}_n\{\varphi(\hat{\mathbb{P}})\}$$

This is asymptotically equivalent to the estimating equation or one-step correction approach

- but can give better finite-sample properties if $\psi(\hat{\mathbb{P}}^*)$ bounded

# One-step decomposition

One-step estimator has important & general decomposition:

$$\begin{aligned}\widehat{\psi} - \psi &= \psi(\widehat{\mathbb{P}}) + \mathbb{P}_n\{\widehat{\varphi}(Z)\} - \psi(\mathbb{P}) \\ &= (\mathbb{P}_n - \mathbb{P})\widehat{\varphi}(Z) + R_2(\widehat{\mathbb{P}}, \mathbb{P}) \\ &= (\mathbb{P}_n - \mathbb{P})\varphi(Z) + (\mathbb{P}_n - \mathbb{P})\{\widehat{\varphi}(Z) - \varphi(Z)\} + R_2(\widehat{\mathbb{P}}, \mathbb{P}) \\ &\equiv S^* + T_1 + T_2\end{aligned}$$

Note:

- 1st term is a sample average of a fixed function $\rightarrow$ CLT
- 2nd term is sample average of term with shrinking variance
- 3rd term is 2nd-order, can be negligible under NP conditions

If we can kill 2nd and 3rd terms, we have efficient estimator!

# The game is afoot

So now the game is finding conditions under which

- (empirical process term) $T_1 = (\mathbb{P}_n - \mathbb{P})(\widehat{\varphi} - \varphi)$
- (bias term) $T_2 = R_2(\hat{\mathbb{P}}, \mathbb{P})$

are negligible, i.e., of order $o_{\mathbb{P}}(1/\sqrt{n})$.

Then $\sqrt{n}(\hat{\psi} - \psi) = \sqrt{n}\mathbb{P}_n(\varphi) + o_{\mathbb{P}}(1) \rightsquigarrow N(0, \mathrm{var}(\varphi))$, i.e.,

1. $\widehat{\psi}$ is root-n consistent,
2. $\widehat{\psi}$ is asymptotically normal, with asymptotically valid 95% CIs given by $\widehat{\psi} \pm 1.96\sqrt{\widehat{\mathrm{var}}\{\varphi(Z; \widehat{\mathbb{P}})\}/n}$,
3. $\widehat{\psi}$ is optimal in a local asymptotic minimax sense.

# Main idea

When $\widehat{\varphi}$ and $\mathbb{P}_n$ use the same data, the bias correction in the one-step estimator uses same data twice, for two different tasks

- (1) estimating nuisance fns $\widehat{\varphi}$, (2) estimating bias $\mathbb{P}_n(\widehat{\varphi})$
- a kind of "double-dipping", which can yield overfitting

Natural way to avoid overfitting is to not fit overly complex models

- here one can assume nuisance fns lie in Donsker classes
- i.e., bracketing/entropy integrals finite (see vdV Ch 18)

# Donsker examples

Donsker classes include

- indicator functions
- VC classes
- bounded monotone functions
- Lipschitz parametric functions
- smooth functions with bounded partials
- Sobolev classes
- uniform sectional variation

# Why Donsker matters (for $T_1$)

Lemma 19.24, van der Vaart (2000): Suppose

1. $\hat{f}, f \in \mathcal{F}$ for some Donsker class $\mathcal{F}$
2. $\|\hat{f} - f\|^2 = o_{\mathbb{P}}(1)$

Then:

$$(\mathbb{P}_n - \mathbb{P})\hat{f} = (\mathbb{P}_n - \mathbb{P})f + o_{\mathbb{P}}(1/\sqrt{n})$$

This follows from the continuous mapping theorem applied to $(\mathbb{G}_n, \hat{f}) \rightsquigarrow (\mathbb{G}, f)$ with function $h(z, f') = z(f') - z(f)$.

# When Donsker fails...

The Donsker condition is quite a bit weaker than requiring $\eta$ to follow particular parametric models, but it is still pretty restrictive

- requires sufficient smoothness relative to dimension ($s > d/2$)
- generally fails in high-dimensional settings with $p > n$
- unclear for modern methods that are very complex/adaptive

Luckily there is a simple fix that has been around for a long time

- sample splitting!
- this not only removes all complexity conditions (only requiring consistency) but also greatly simplifies proofs
- Hasminskii & Ibragimov (1979), Bickel (1982), vdV (1998), Robins (2008), vdL (2011), Chernozhukov et al (2018), etc.

# Pfanzagl (1982)

11.4.5. Remark. Hasminskii and Ibragimov (1979, pp. 44-46, Theorems 2-4) establish the as. efficiency of improved estimator-sequences for non-parametric models. For technical reasons, they use a slightly modified version of the improvement procedure, in which the estimator $P_n(\underline{x},\cdot)$ is based on $x_1,\ldots,x_{k(n)}$, whereas the summation in $\Sigma\kappa^{\bullet}(x_\nu,P_n(\underline{x},\cdot))$ extends over $\nu = k(n)+1,\ldots,n$. By this trick they achieve that $x_\nu$ for $\nu = k(n)+1,\ldots,n$ are stochastically independent of $P_n(\underline{x},\cdot)$. Letting $k(n)$, $n \in \mathbb{N}$, tend to infinity sufficiently f a s t, one obtains estimators $P_n(\underline{x},\cdot)$ which are sufficiently close to P. If

# Sample splitting

Randomly split observations $(Z_1, ..., Z_n)$ into $K$ disjoint groups

- using random variable $S$ drawn independently of data, where $S_i \in \{1, ..., K\}$ denotes group number for unit $i$

Let $\widehat{\mathbb{P}}_{-k}$ denote nuisance estimator constructed excluding group $k$, i.e., using units $\{i : S_i \neq k\}$

Then instead use

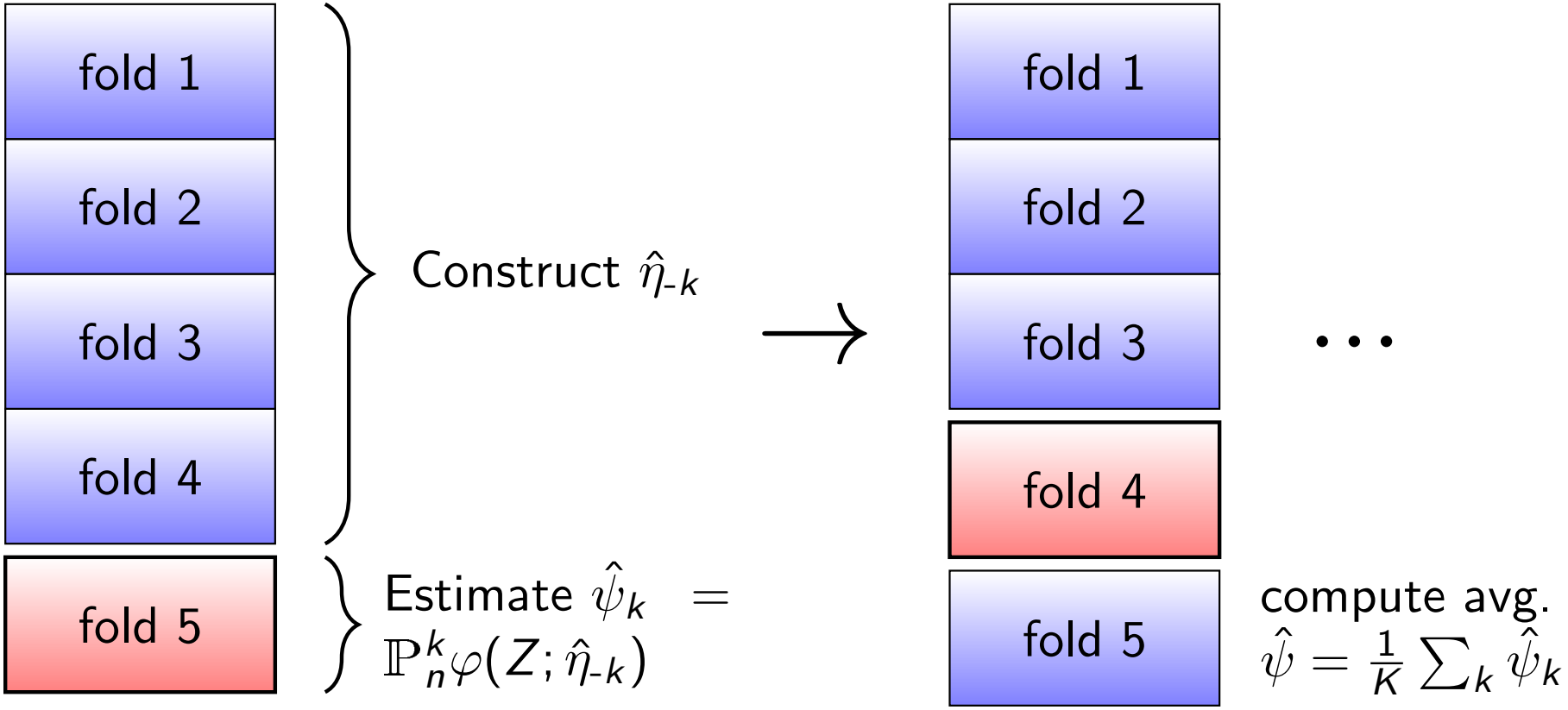
$$\hat{\psi} = \sum_{k=1}^{K} \left( \frac{N_k}{n} \right) \widehat{\psi}_k$$

where $\widehat{\psi}_k$ is the usual estimator in the $k$th fold

$$\widehat{\psi}_k = \psi(\widehat{\mathbb{P}}_{-k}) + \mathbb{P}_n^k \Big\{ \varphi(Z; \widehat{\mathbb{P}}_{-k}) \Big\}$$

where $\mathbb{P}_n^k$ denotes empirical over $k$th fold

# Sample splitting schematic



# Sample splitting analysis

For each $\widehat{\psi}_k$ we have a similar decomposition for $\widehat{\psi}_k - \psi$ as before

$$= (\mathbb{P}_n^k - \mathbb{P})\varphi(Z;\mathbb{P}) + (\mathbb{P}_n^k - \mathbb{P})\{\varphi(Z;\widehat{\mathbb{P}}_{-k}) - \varphi(Z;\mathbb{P})\} + R_2(\widehat{\mathbb{P}}_{-k},\mathbb{P})$$
$$\equiv S_k^* + T_{1k} + T_{2k}$$

by the exact same logic as before, and similarly

$$\widehat{\psi} - \psi = S^* + \sum_{k=1}^{K}\left(\frac{N_k}{n}\right)\left(T_{1k} + T_{2k}\right) \equiv S^* + T_1 + T_2$$

Now each $T_{1k}$ can be analyzed by conditioning on $\{i : S_i \neq k\}$

# Sample splitting lemma

**Lemma (see e.g., "Sharp instruments..." Kennedy et al.)**
*Let $\hat{f}$ be estimated from a sample $Z^N = (Z_{n+1}, ..., Z_N)$ and let $\mathbb{P}_n$ denote the empirical measure over $(Z_1, ..., Z_n)$, independent of $Z^N$. Then*

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\hat{f} - f\|}{\sqrt{n}}\right).$$

**Proof.**
Conditional on $Z^N$, the term $\mathbb{G}_n(\hat{f} - f)$ has mean zero and variance bounded above by $\|\hat{f} - f\|^2$. The result then follows by Markov's inequality. □

# Sample splitting

This is a beautifully simple result!

As long as $\widehat{\varphi}$ is consistent for $\varphi$ in $L_2$ norm, and $K$ is finite:

- we have $T_1 = o_{\mathbb{P}}(1)$
- don't need any complexity conditions whatsoever - free to use any modern ML methods we like
- also somewhat more transparent than Donsker approach

# Recap

We have seen two nonparametric ways to control empirical process terms of the form

$$(\mathbb{P}_n - \mathbb{P})\{\widehat{\varphi}(Z) - \varphi(Z)\}$$

which allows us to kill the first term $T_1$ in the decomposition of our influence function-based bias-corrected estimator

- restrict the complexity of $\varphi(\mathbb{P})$ via Donsker conditions
- sample splitting to avoid any restrictions (only consistency)

We are one-step closer to an efficient estimator

- now need to study the second-order remainder

# Second-order remainders

The 2nd-order remainder terms $R_2(\hat{\mathbb{P}}, \mathbb{P})$ typically need to be studied on a case-by-case basis

This term is really what makes the bias-corrected estimator special

- for general plug-ins, the $T_1$ and CLT-type terms also appear
- but the remainder is generally non-negligible, resulting in a slower rate of convergence

The amazing thing about the bias-corrected approach is that $R_2$ can be negligible even in complex nonparametric models

# Zero remainder examples

Recall earlier we showed that for $\psi = \mathbb{E}(Z)$ we have

$$\psi(\mathbb{Q}) - \psi(\mathbb{P}) = -\int \left\{ z - \psi(\mathbb{Q}) \right\} d\mathbb{P}(z)$$

so that $R_2 = 0$ exactly.

Of course this holds for any known $f(Z)$, e.g., $f = \frac{A}{\pi}(Y - g) + g$

- this shows that IPW-style IFs are in fact IFs when $\pi$ is known
- each satisfies the von Mises expansion with $R_2 = 0$

# Integrated density squared

Recall the IF for $\psi = \int p(z)^2 \, dz$ is

$$\varphi(z; p) = 2\{p(z) - \psi\}$$

We showed that general plug-in $\mathbb{P}_n(\hat{p})$ will not be $\sqrt{n}$-consistent

- kernel plug-in can, *with undersmoothing & strong smoothness*

Consider instead IF-based bias-corrected estimator

$$\hat{\psi} = 2\mathbb{P}_n(\hat{p}) - \int \hat{p}^2$$

- (suppose we've constructed $\hat{p}$ from a separate sample)

# Integrated density squared

The IF-based estimator satisfies

$$\hat{\psi} - \psi = 2(\mathbb{P}_n - \mathbb{P})(\hat{p} - p) + 2(\mathbb{P}_n - \mathbb{P})p + R_2(\hat{p}, p)$$

where

$$R_2(\hat{p}, p) = -\int (\hat{p} - p)^2$$

The optimal rate for estimating smooth $p$ in $L_2$ norm is $n^{\frac{-1}{2+d/\beta}}$

- therefore need $\frac{2}{2+d/\beta} > 1/2$, i.e., $\beta > d/2$ rather than $\beta > d$

$\rightarrow$ we need half smoothness of $p$, & no undersmoothing required!

# ATE

Recall the IF for $\psi = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1)\} \equiv \mathbb{E}\{\mu(X)\}$ is

$$\varphi(Z; \mathbb{P}) = \frac{A}{\pi(X)}\Big\{Y - \mu(X)\Big\} + \mu(X) - \psi$$

Plug-in: $\mathbb{P}_n(\hat{\mu}) - \psi = (\mathbb{P}_n - \mathbb{P})(\hat{\mu} - \mu) + (\mathbb{P}_n - \mathbb{P})\mu + \mathbb{P}(\hat{\mu} - \mu)$

- generally dominated by last term, which is not $O_{\mathbb{P}}(1/\sqrt{n})$

Consider instead IF-based bias-corrected estimator

$$\hat{\psi} = \mathbb{P}_n\left[\frac{A}{\hat{\pi}(X)}\Big\{Y - \hat{\mu}(X)\Big\} + \hat{\mu}(X)\right]$$

- (suppose we've constructed $\hat{\eta}$ from a separate sample)

# ATE

The IF-based estimator satisfies

$$\hat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})(\hat{\varphi} - \varphi) + (\mathbb{P}_n - \mathbb{P})\varphi + R_2(\hat{\eta}, \eta)$$

where

$$R_2(\hat{\eta}, \eta) = \mathbb{P}\left\{\frac{1}{\hat{\pi}}(\pi - \hat{\pi})(\mu - \hat{\mu})\right\} \lesssim \|\hat{\pi} - \pi\|\|\hat{\mu} - \mu\|$$

Thus e.g., $n^{-1/4}$ rates suffice! can be attained with sufficient smoothness/sparsity

Optimal rate for smooth $\pi$ (resp., $\mu$) is $n^{\frac{-1}{2+d/\alpha}}$ (resp., $n^{\frac{-1}{2+d/\beta}}$ )

- therefore need $\frac{1}{2+d/\alpha} + \frac{1}{2+d/\beta} > 1/2$, i.e., $\sqrt{\alpha\beta} > d/2$

# Pfanzagl (1982)

To prove (11.4.3), assume that

$$\kappa(P_n(\underline{x},\cdot)) = \kappa(P) + \int \kappa^{\bullet}(\xi,P)P_n(\underline{x},d\xi) + o_P(n^{-1/2}) . \tag{11.4.4}$$

This holds, for instance, if $\kappa$ is strongly differentiable in the sense that

$$\kappa(Q) = \kappa(P) + \int \kappa^{\bullet}(\xi,P)Q(d\xi) + O(\delta(Q,P)^2)$$

uniformly for $Q \in \mathfrak{P}$, and if $P_n$ fulfills $\delta(P_n(\underline{x},\cdot),P) = o_P(n^{-1/4})$. No-

# LATE

The IF for $\psi = \frac{\mathbb{E}\{\mathbb{E}(Y|X,R=1)-\mathbb{E}(Y|X,R=0)\}}{\mathbb{E}\{\mathbb{E}(A|X,R=1)-\mathbb{E}(A|X,R=0)\}}$ is

$$\Big\{\varphi_Y(Z;\mathbb{P}) - \psi\ \varphi_A(Z;\mathbb{P})\Big\}\Big/\mathbb{E}\{\varphi_A(Z;\mathbb{P})\}$$

for $\varphi_T = \frac{(2R-1)\{T-\mathbb{E}(T|X,R)\}}{\pi_R} + \mathbb{E}(T \mid X, R=1) - \mathbb{E}(T \mid X, R=0)$

The remainder term is bounded above by

$$R_2(\hat{\mathbb{P}},\mathbb{P}) \lesssim \|\hat{\pi}-\pi\| \left(\max_r \|\hat{\lambda}_r - \lambda_r\| + \max_r \|\hat{\mu}_r - \mu_r\|\right)$$

where $\lambda_R = \mathbb{E}(A \mid X, R)$ and $\mu_R = \mathbb{E}(Y \mid X, R)$

# Software

I have an R package `npcausal` that implements influence function-based estimators with sample splitting

Details can be found at:

https://www.ehkennedy.com/code.html

https://github.com/ehkennedy/npcausal

# Example npcausal code

Loading `npcausal` in R is easy:

```
> install.packages("devtools")
>
> library(devtools)
> install_github("ehkennedy/npcausal")
> library(npcausal)
```

# Example npcausal code: ATE

```
> n <- 1000; x <- matrix(rnorm(n*5),nrow=n)
> a <- sample(3,n,replace=TRUE); y <- rnorm(n)
>
> ate.res <- ate(y,a,x)
  |=================================================================| 100%
     parameter         est         se         ci.ll      ci.ul  pval
1      E{Y(1)} -0.02114171 0.05695659 -1.327766e-01 0.09049322 0.710
2      E{Y(2)} -0.09023984 0.05613025 -2.002551e-01 0.01977546 0.108
3      E{Y(3)}  0.11000161 0.05612383 -1.102083e-06 0.22000432 0.050
4 E{Y(2)-Y(1)} -0.06909813 0.07990418 -2.257103e-01 0.08751405 0.387
5 E{Y(3)-Y(1)}  0.13114332 0.07999265 -2.564228e-02 0.28792891 0.101
6 E{Y(3)-Y(2)}  0.20024145 0.07987263  4.369109e-02 0.35679180 0.012
```

# Example npcausal code: ATT

```
> n <- 1000; x <- matrix(rnorm(n*5),nrow=n)
> a <- rbinom(n,1,.3); y <- rnorm(n)
>
> att.res <- att(y,a,x)
  |=================================================================| 100%
      parameter        est         se       ci.ll     ci.ul  pval
1      E(Y|A=1) 0.07668552 0.05336217 -0.02790434 0.1812754 0.151
2   E{Y(0)|A=1} 0.01934400 0.01895868 -0.01781501 0.0565030 0.308
3 E{Y-Y(0)|A=1} 0.05734152 0.05660685 -0.05360791 0.1682910 0.311
```

# Example npcausal code: LATE

```
> n <- 100; x <- matrix(rnorm(n*5),nrow=n)
> z <- rbinom(n,1,0.5); a <- rbinom(n,1,0.6*z+0.2)
> y <- rnorm(n)
>
> ivlate.res <- ivlate(y,a,z,x)
  |=================================================================| 100%
  parameter         est         se         ci.ll     ci.ul  pval
1      LATE -0.07832395 0.39111888 -8.449169e-01 0.6882690 0.841
2  Strength  0.57774804 0.08328482  4.145098e-01 0.7409863    NA
3 Sharpness  0.04745317 0.17225263  2.841594e-05 0.9886793    NA
```

# Software

There are other functions as well, which implement procedures from some of my papers:

- "Sharp instruments for classifying compliers and generalizing causal effects"
- "Nonparametric methods for doubly robust estimation of continuous treatment effects"
- "Nonparametric causal effects based on incremental propensity score interventions"

# Some open problems

NP functional estimation may seem like an open-and-shut case

$\rightarrow$ not at all - tons of important open problems!

1. What about new functionals?
2. What if $\sqrt{n}$ rates are not attainable?
3. What if $\psi$ is not pathwise differentiable?

$\rightarrow$ *Lots of exciting work for us to do!*

# References

Readable intros:

- Newey (1990): Semiparametric efficiency bounds
- Hahn (1998): On the role of the propensity score in efficient...
- van der Vaart (2002): Lecture notes on semiparametric statistics
- Tsiatis (2006): Semiparametric theory & missing data
- Kennedy (2022): Semiparametric DR T DML

Good references:

- BKRW (1993): Efficient & adaptive estimation for SP models
- Robins, Rotnitzky, Zhao (1995): Analysis of SP regression...
- van der Laan & Robins (2003): Unified methods for censored...

Useful examples *(note this is a very small sample)*:

- van der Laan (2006): Statist. inference for variable importance
- Tchetgen & Shpitser (2012): Semiparametric theory for...
- Kandasamy et al. (2014): Influence functions for ML...
- Ogburn et al. (2015): DR estimation of the LATE curve
- Farrell (2015): Robust inference on ATEs with possibly more...
- Chernozhukov et al. (2017): Double ML
- my papers?

In addition to above authors, also check out papers by:

- Carone, Diaz, Luedtke, Newey, Tan, Vansteelandt,

Fun to read and historically important:

- Stein (1956): Efficient nonparametric testing & estimation
- Bickel & Ritov (1988): Estimating integrated squared density...
- Pfanzagl (1992): Contributions to a general asymptotic...

If you're feeling courageous:

- Robins et al. (2008): Higher order influence functions...
- Carone et al. (2014): Higher order TMLE
- Robins et al. (2017): Minimax estimation of a functional...

**Thank you!**