Nonparametric efficiency theory and machine learning in causal inference

Edward Kennedy

Department of Statistics & Data Science
Carnegie Mellon University

Atlantic Causal Inference Conference, 21 May 2018

Roadmap

My disclaimers

1. I cannot take credit for most theory I'll talk about today
   - see: Andrews, Begun, Bickel, Pfanzagl, Pollard, Robins, Stein, 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 ψ^*

The *target parameter* ψ^* 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 $\mathbb{P}$

Picking the target parameter

Ideally target ψ^* is chosen based only on scientific concerns, but

- often it is only defined vaguely (e.g., "the effect")
- or chosen based on convenience (e.g., β in logistic regression)

I have encountered two cultures in applied statistics

1. model entire data generating process, use model to answer any/all scientific questions
2. start with specific question, and tailor analysis accordingly

I am a big fan of the 2nd approach

- one-size-fits-all model often not best for all questions
- 2nd forces you to think hard about science/goals


FINCH
21

Picking the target parameter

To pick target parameter ψ^* we can ask:

What experiment would you have conducted if there were no ethical or feasibility concerns?

For example:

- force everyone to give lab values
- give everyone treatment, then go back in time and withhold
- force all to become obese, assess outcomes after 30 years

Potential outcomes

Causal language lets us define target wrt. idealized interventions

- 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 $\overline{A}_T = (A_1, \ldots, A_T)$ to $\overline{a}_T = (a_1, \ldots, a_T)$
- Y^G denotes outcome under stochastic intervention G that assigns $A = 1$ with probability $g(x)$ depending on covariates X

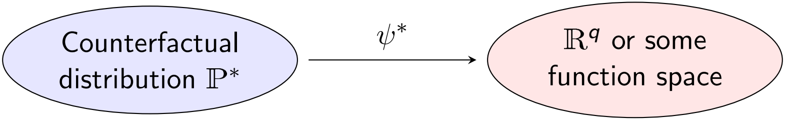
Example causal parameters

$\rightarrow$ 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:	$\mathbb{E}(Y^{\overline{a}_T} \mid V)$
$\rightarrow$ SNM:	$\mathbb{E}(Y^{\overline{a}_t,0} - Y^{\overline{a}_{t-1},0} \mid \overline{L}_t, \overline{A}_t)$

Causal target parameter

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

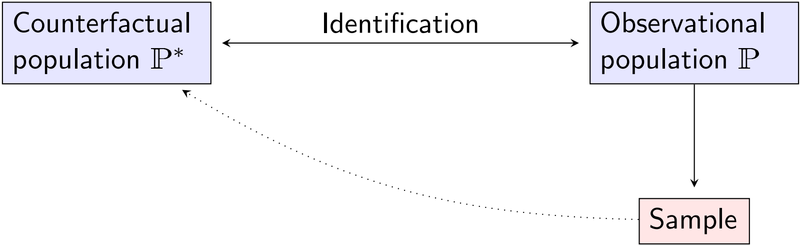
- can be a number, or function, or even more complex object



Identification

After picking ψ^*, 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 ψ



Identification example: ATE

Assume:

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

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

Then

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

Identification example 3: LATE

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

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

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

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

Then:

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

More on identification

Other identification schemes can also be used

- Pearl et al. have pursued extensive graphical criteria

Sometimes (often?) no reasonable assumptions point identify ψ

- but in some cases we can still get bounds on ψ^*
- then we can treat bounds as (often non-smooth) target
- large literature on this: see Manski, etc.

Keep 'em separated

I find it essential to keep causal & statistical issues separate

The causal stuff (what we want to estimate, what we believe about causality, confounding, etc.) only tells us **what** we should be estimating with observed data, not **how** to estimate it

I try not to mix the two, i.e., don't:

- interpret statistical models causally
- restrict observed data with parametric causal assumptions

Causal inference is over

After we identify the causal target parameter $\psi^*(\mathbb{P}^*)$ by writing it as an observed data parameter $\psi^*(\mathbb{P}^*) = \psi(\mathbb{P})$...

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

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

- int. sq. density: $\psi = \int p(x)^2 \ dx$
- entropy: $\psi = -\int p(x) \log p(x) \ dx$
- support size: $\psi = \sum_x \mathbb{1}\{p(x) > 0\}$
- 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 sometimes 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
- $\varphi(\cdot)$ is some function, e.g., $\varphi(Z) = Z$ for sample mean

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.,

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

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

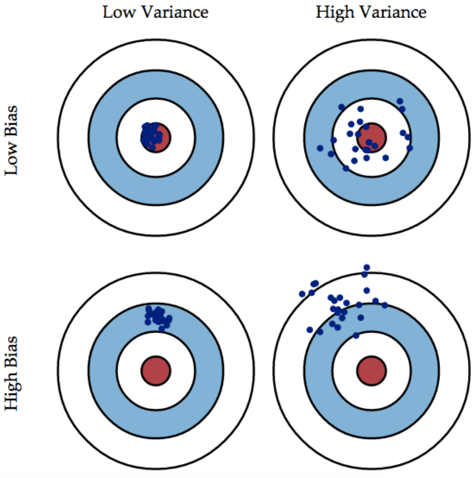
$$\mathbb{P}(|X_n| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty$$

We use L_2 and sup norms to measure 'distance' between functions

$$\|\hat{f} - f\| = \sqrt{\int \{\hat{f}(x) - f(x)\}^2 \, d\mathbb{P}(x)} = \sqrt{\mathbb{P}\{(\hat{f} - f)^2\}}$$

$$\|\hat{f} - f\|_\infty = \sup_x |\hat{f}(x) - f(x)|$$

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

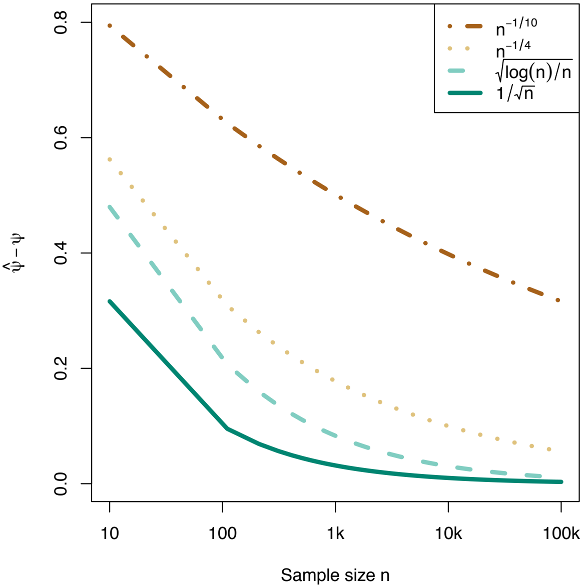
$$r_n(\hat{\psi} - \psi) = O_{\mathbb{P}}(1) \quad \Longleftrightarrow \quad \hat{\psi} - \psi = O_{\mathbb{P}}(1/r_n)$$

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

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

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

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


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

Aside: Asymptotic normality

An estimator is root-n consistent and asymptotically normal if

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

$\rightarrow$ and then we say $\hat{\psi}$ has influence function ϕ

This is typically as well as we can possibly do

- if 2+ estimators are $\sqrt{n}$-CAN then can choose among them based on variances (only in proper semiparametric models)

In some cases $\sqrt{n}$-CAN is not attainable

- in nonparametric models at most one estimator is $\sqrt{n}$-CAN

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 $\mu(X) = \mathbb{E}(Y \mid X)$ with β partial derivatives, best possible convergence rate for any $\hat{\mu}$ is

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

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 θ

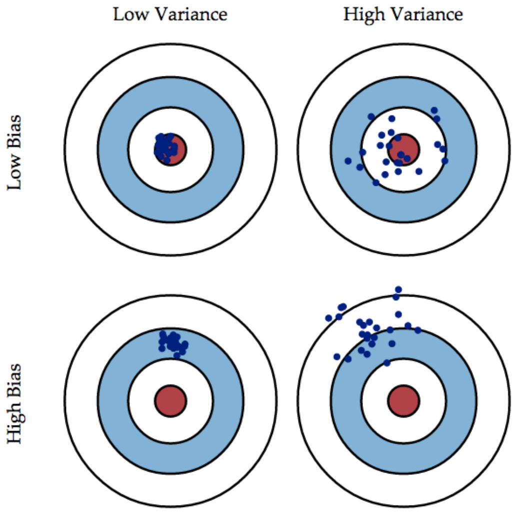
- then, for $\hat{\theta}$ the MLE, classical theory says $\psi(\mathbb{P}_{\hat{\theta}})$ is asymptotically efficient (minimax optimal) under smoothness

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 θ? 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

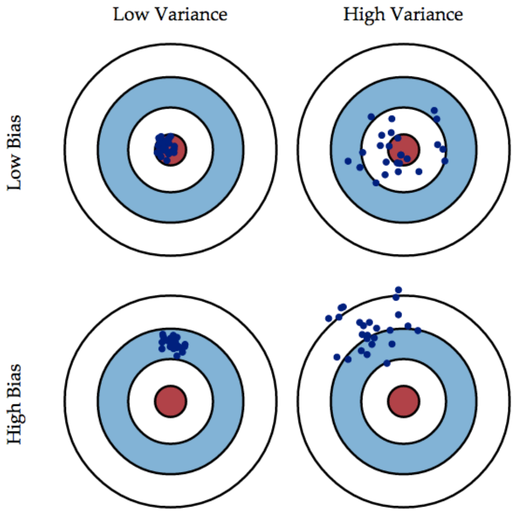
$$\hat{\psi} = \mathbb{P}_n\{\hat{p}(X)\}.$$

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 $O_{\mathbb{P}}(n^{-\beta/(2\beta+d)})$ for Hölder(β)


Low Variance
High Variance
Low Bias
High Bias

Ex: integrated density squared

However if p is Hölder(β) 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*

Example 1: nonparametric model

The simplest example of a semiparametric model is...

- a nonparametric model

$\rightarrow \mathcal{P}$ consists of all possible probability measures

Thus semiparametrics *includes nonparametrics as a special case*

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

Example 2: GEE

Suppose $Z = (X, Y)$ for covariates X and outcome Y, and

$$Y = \mu(X; \psi) + \epsilon$$

for $\psi \in \mathbb{R}^q$ and density of ϵ unrestricted beyond $\mathbb{E}(\epsilon \mid X) = 0$.

- i.e., we only assume $\mathbb{E}(Y \mid X) = \mu(X; \psi)$

This is just: GEE / m-estimation / restricted moment model

Example 3: Cox model

Suppose $Z = (X, T)$ for covariates X and survival time T, and

$$\frac{\lambda(t \mid X = x)}{\lambda(t \mid X = 0)} = \exp(x^{\mathrm{T}}\psi)$$

for $\lambda(t \mid x)$ the conditional hazard of T given $X = x$.

This is the Cox proportional hazards model

Semiparametric causal models

In causal problems, often know/put structure on treatment process

- e.g., in randomized trial, treatment process is known, whereas outcome might be complex process outside human control
- in observational settings, treatment may not be known exactly but still well-understood

Thus may have $Z = (X, A, Y) \sim \mathbb{P}$ with

$$p(z; \eta, \alpha) = p(y \mid x, a)\ p(a \mid x; \alpha)\ p(x)$$

with $\eta = \{p(y \mid x, a), p(x)\}$ **infinite-dimensional** & unrestricted but $\alpha \in \mathbb{R}^q$.

Nonparametric causal models

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

- then a nonparametric model makes most sense

This is the perspective I 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

More semiparametric causal models

Semiparametric models can also arise via parametric assumptions about non-Euclidean functionals

- e.g., functionals like $\gamma(v) = \mathbb{E}(Y^1 - Y^0 \mid V = v)$ for $V \subseteq X$

When V is high-dimensional or has continuous components, might make sense to use parametric models $\gamma(v; \psi)$ for $\psi \in \mathbb{R}^q$

- can either assume $\gamma(v) = \gamma(v; \psi)$
- or project $\gamma(v)$ onto $\gamma(v; \psi)$ agnostically/nonparametrically

$\rightarrow$ either way, semiparametric models naturally arise

Classical lower bounds

First line of business: lower bds/benchmarks for estimation error

- first consider classical parametric setup

Recall for "smooth" $\{\mathbb{P}_\theta : \theta \in \mathbb{R}^d\}$ & *any* unbiased est. T of $\psi(\theta)$

$$\text{var}(T) \geq \psi'(\theta)^2/\mathbb{E}(s_\theta^2)$$

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

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

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

Parametric submodels

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

Let parametric submodel $\mathcal{P}_\epsilon = \{\mathbb{P}_\epsilon : \epsilon \in \mathbb{R}\} \subset \mathcal{P}$ 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,

$$p_\epsilon(z) = p(z)\{1 + \epsilon h(z)\}$$

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

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

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

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

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

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

Pathwise differentiability

What can we say about the derivative in the numerator?

Suppose ψ is smooth enough to admit a von Mises-type expansion

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

for $\phi(z; \mathbb{P})$ with $\mathbb{P}(\phi) = 0$, $\mathbb{P}(\phi^2) < \infty$, R_2 a 2nd-order remainder

- this is just a distributional analog of a Taylor expansion

This implies, under regularity conditions, pathwise differentiability:

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

Pathwise differentiability is important

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

$$\psi(\mathbb{Q}) - \psi(\mathbb{P}) = \int \phi(\mathbb{Q}) \; d(\mathbb{Q} - \mathbb{P}) + R_2(\mathbb{Q}, \mathbb{P})$$

for mean-zero/finite-variance ϕ, is really key

- will see later that this suggests how to bias-correct a plug-in

Typically say any ϕ satisfying above is an influence function for ψ

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

Efficient influence function

If ψ is pathwise differentiable, then the greatest lower bound is

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

inequality by Cauchy-Schwarz, equality if ϕ is valid submodel score

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

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

- we call ϕ the efficient influence function

This is hugely important - implies no estimator can beat

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

in an asymptotic minimax sense!

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

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

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

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

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 ϕ

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

There are several ways I know of to derive IFs, the most general being to compute pathwise derivative $\psi'(\mathbb{P}_\epsilon)$ and solve for ϕ

- often easier to pretend data are discrete and compute Gateaux derivative in direction of point mass contamination
- or use chain/product rules with 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 $\phi(z; \mathbb{P}) = z - \psi$. Also von Mises holds with $R_2 = 0$.

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

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

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

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

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

Influence function for integrated square density

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

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

so EIF is $\phi = 2(p - \psi)$. 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 ψ at $\mathbb{P}(z')$ in direction of point mass $(\delta_z - \mathbb{P}(z'))$

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

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 $h = \delta_z$ ... $\implies \int \phi s_\epsilon \, d\mathbb{P} = \phi(z)$

Influence function for ATE

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

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

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...

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 $\psi'(\mathbb{P}_\epsilon) = \text{cov}(\phi, s_\epsilon)$ 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)

How do we know EIF is unique?

Let $\mathcal{T}$ be tangent space, ϕ any IF, and $\Pi(\cdot \mid \cdot)$ projection operator. Then $\Pi(\phi \mid \mathcal{T})$ is the unique EIF. Proof:

0. Can write any IF as $\phi' = \phi + h$ for ϕ any IF and $h \in \mathcal{T}^\perp$ since

   $$\text{cov}(h, s_\epsilon) = \text{cov}(\phi + h, s_\epsilon) - \text{cov}(\phi, s_\epsilon) = \psi'(\mathbb{P}_\epsilon) - \psi'(\mathbb{P}_\epsilon) = 0$$

1. $\varphi = \Pi(\phi + h \mid \mathcal{T})$ doesn't depend on h (is unique) since

   $$\Pi(\phi + h \mid \mathcal{T}) = \Pi(\phi \mid \mathcal{T}) + \Pi(h \mid \mathcal{T}) = \Pi(\phi \mid \mathcal{T})$$

2. $\varphi = \Pi(\phi + h \mid \mathcal{T}) = \Pi(\phi \mid \mathcal{T})$ is an IF since

   $$\text{cov}(\varphi, s_\epsilon) = \text{cov}(\phi, s_\epsilon) + \text{cov}(\varphi - \phi, s_\epsilon) = \psi'(\mathbb{P}_\epsilon) + 0$$

What happens w/ATE when PS is known?

If π known, influence functions take the form

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

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

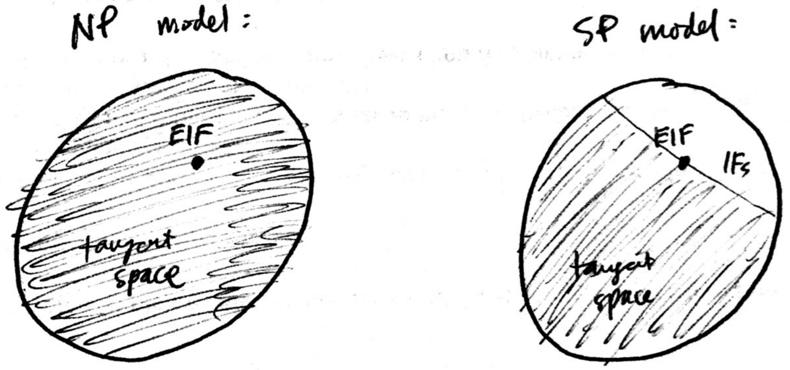
Why are these IFs not IFs in nonparametric model? Scores differ:

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

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

$\implies$ for $g \neq \mu$, we have $\text{cov}(\phi_g, s_\epsilon) \neq 0$ for s_ϵ the full NP score

- π unknown $\implies$ ϕ_g is a valid score but not a valid IF
- π known $\implies$ ϕ_g is a valid IF but not a valid score



Aside: semiparametrics vs. projections

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

- I try to 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.,

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

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 $\mathbb{P}(\hat{f}) = \mathbb{E}(\hat{f} \mid Z_1, \ldots, Z_n)$ 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 ψ satisfying von Mises/Taylor expansion

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

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

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

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

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

Bias correction

The previous formulation suggests simple bias-correction procedure

$$\hat{\psi} - \psi \equiv \left[\psi(\hat{\mathbb{P}}) + \mathbb{P}_n\{\phi(\hat{\mathbb{P}})\}\right] - \psi = (\mathbb{P}_n - \mathbb{P})\phi(\hat{\mathbb{P}}) + R_2(\hat{\mathbb{P}}, \mathbb{P})$$
$$= (\mathbb{P}_n - \mathbb{P})\{\phi(\hat{\mathbb{P}}) - \phi(\mathbb{P})\} + (\mathbb{P}_n - \mathbb{P})\phi(\mathbb{P}) + R_2(\hat{\mathbb{P}}, \mathbb{P})$$

Note:

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

If we can kill 1st and 3rd terms, we have efficient estimator!

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

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

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

The game is afoot

So now the game is finding conditions under which terms

- $R_1 = \mathbb{G}_n\{\phi(\hat{\mathbb{P}}) - \phi(\mathbb{P})\}$
- $R_2 = \sqrt{n} R_2(\hat{\mathbb{P}}, \mathbb{P})$

are negligible, i.e., of order $o_{\mathbb{P}}(1)$, where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$.

Then $\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N(0, \text{var}(\phi))$ and we have optimality

R_1 will be negligible if either

1. $\phi(\mathbb{P})$ is not too complex (empirical processes)
2. we separate $\hat{\mathbb{P}}$ & $\mathbb{P}_n$ to prevent overfitting (sample splitting)

We will discuss these kinds of conditions first, then move to R_2

Main idea

To control how close $\mathbb{G}_n\phi(Z;\hat{\eta})$ is to limiting version $\mathbb{G}_n\phi(Z;\eta)$ one approach is to restrict the complexity of $\hat{\eta}$ and η

- a nonparametric way to do this is with **Donsker classes**

We'll show $\mathbb{G}_n\phi(Z;\hat{\eta}) = \mathbb{G}_n\phi(Z;\eta) + o_{\mathbb{P}}(1)$ if

1. $$\|\phi(;\hat{\eta}) - \phi(;\eta)\|^2 = \int \Big\{\phi(z;\hat{\eta}) - \phi(z;\eta)\Big\}^2 \, d\mathbb{P}(z) = o_{\mathbb{P}}(1)$$

(i.e., if $\hat{\eta}$ is consistent in L_2 norm) and if

2. $\{\phi(;\eta) : \eta \in H\}$ is a Donsker class

Preliminaries

Let $\mathcal{F}$ denote a class of functions $f : \mathcal{Z} \to \mathbb{R}$. Then

$$\{\mathbb{G}_n f : f \in \mathcal{F}\}$$

is called the **empirical process** indexed by $\mathcal{F}$.

As a collection of rvs indexed by a set, this is a stochastic process

- this views $\mathbb{G}_n f$ as a rv, for any f, mapping $\mathcal{Z}^n$ to $\mathbb{R}$

Can be helpful to view $\mathbb{G}_n f$ as, for any sample, map from $\mathcal{F}$ to $\mathbb{R}$

- here the empirical process is a random fn from $\mathcal{Z}^n$ to $\ell^\infty(\mathcal{F})$, which is the space of bdd fns $h : \mathcal{F} \to \mathbb{R}$ with $||h||_\mathcal{F} < \infty$

Weak convergence & Donsker

We've mentioned processes for *fixed* n. Now consider

$$\{\mathbb{G}_n f : f \in \mathcal{F}\}_{n \geq 1}$$

This sequence of random functions **converges in distribution** to $\mathbb{G}$ in the space $\ell^\infty(\mathcal{F})$, i.e., $\mathbb{G}_n \rightsquigarrow \mathbb{G}$, if

$$\mathbb{E}^* h(\mathbb{G}_n) \to \mathbb{E} h(\mathbb{G}) \ , \quad \text{for all cts. } h : \ell^\infty(\mathcal{F}) \to \mathbb{R}$$

$\to$ gives a notion of convergence for random functions

A class $\mathcal{F}$ is **Donsker** if the sequence $\{\mathbb{G}_n f : f \in \mathcal{F}\}_{n \geq 1}$ converges in distribution to some tight limit $\mathbb{G}$

- *tight* $= \mathbb{P}(\mathbb{G} \in S) > 1 - \epsilon$ for all $\epsilon > 0$ and some compact S

Why Donsker matters (for R_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:

$$\mathbb{G}_n \hat{f} = \mathbb{G}_n f + o_{\mathbb{P}}(1)$$

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

Examples

We've seen the utility of Donsker, but not what classes it covers

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

Preservation

When using $\phi(Z;\eta)$, it is more natural to put Donsker conditions on η rather than the transformed class $\{\phi(;\eta) : \eta \in H\}$

Many transformations preserve the Donsker property

1. subsets
2. unions
3. closures
4. convex combinations (useful for Super Learner / stacking)
5. Lipschitz: minimums, maximums, sums, products, ratios

Bracketing and covering numbers

How does one show a class $\mathcal{F}$ is Donsker?

An ϵ-bracket is the set of f bracketed by $[l, u]$ (i.e., $l \leq f \leq u$) with

$$||u - l|| < \epsilon$$

The bracketing number of $\mathcal{F}$ is the smallest number of ϵ-brackets needed to cover $\mathcal{F}$, and is denoted $N_B(\epsilon, \mathcal{F})$

A class $\mathcal{F}$ is **Donsker** if: $\displaystyle\int_0^1 \sqrt{\log N_B(\epsilon, \mathcal{F})}\, d\epsilon < \infty$

- $N_B(\epsilon, \mathcal{F})$ increases as $\epsilon \to 0$. Donsker: can't increase too fast
- similar results are available for covering numbers

When Donsker fails...

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

- 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
- Bickel (1982), vdV (1998), Robins (2008), vdL (2011), etc.

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

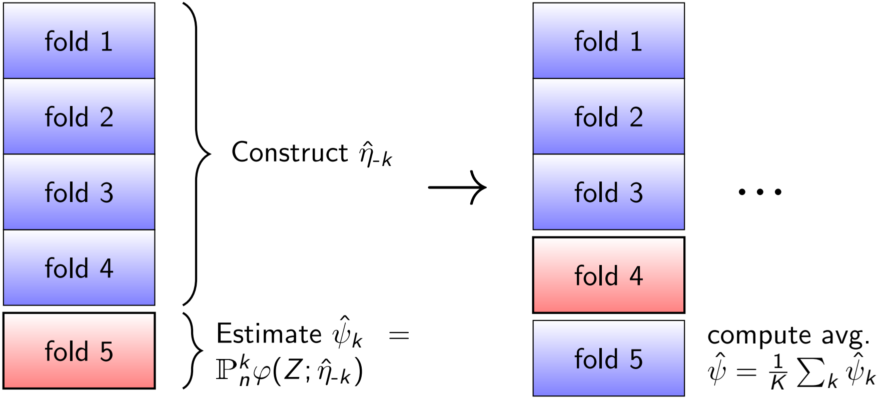
Let $\hat{\eta}_{\text{-}k}$ denote nuisance estimator constructed excluding group k, i.e., using units $\{i : S_i \neq k\}$

For simplicity consider case where $\hat{\psi} = \psi(\hat{\mathbb{P}}) + \mathbb{P}_n \phi(\hat{\mathbb{P}}) = \mathbb{P}_n \varphi(\hat{\eta})$. Then instead use

$$\hat{\psi} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{P}_n^k \varphi(Z; \hat{\eta}_{\text{-}k}) = \mathbb{P}_n \varphi(Z; \hat{\eta}_{\text{-}S})$$

where $\mathbb{P}_n^k$ denotes sub-empirical measure over units $\{i : S_i = k\}$

Sample splitting schematic



Sample splitting analysis

With sample splitting the relevant decomposition is given by

$$\sqrt{n}(\hat{\psi}-\psi) = \frac{1}{K}\sum_{k=1}^{K}\Big[\mathbb{G}_n^k\{\phi(\hat{\eta}_{\text{-}k}) - \phi(\eta)\} + \mathbb{P}\{\phi(\hat{\eta}_{\text{-}k}) - \phi(\eta)\}\Big] + \mathbb{G}_n\phi(\eta)$$

with $R_1 = K^{-1}\sum_k R_{1k}$, for $R_{1k} = \mathbb{G}_n^k\{\phi(\hat{\eta}_{\text{-}k}) - \phi(\eta)\}$

Now R_{1k} can be analyzed easily 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

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

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 $\phi(\hat{\eta})$ is consistent for $\phi(\eta)$ in L_2 norm, and K is finite:

- we have $R_1 = o_{\mathbb{P}}(1)$
- 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

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P})\{\phi(\hat{\mathbb{P}}) - \phi(\mathbb{P})\} = \mathbb{G}_n\{\phi(\hat{\mathbb{P}}) - \phi(\mathbb{P})\}$$

which allows us to kill the first term R_1 in the decomposition of our influence function-based bias-corrected estimator

- restrict the complexity of $\phi(\mathbb{P})$ 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 R_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 π 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

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

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 $n^{\frac{-\beta}{2\beta+d}}$

- therefore need $\frac{2\beta}{2\beta+d} > 1/2$, 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

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

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

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

where

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

Optimal rate for smooth π (resp., μ) is $n^{\frac{-\alpha}{2\alpha+d}}$ (resp., $n^{\frac{-\beta}{2\beta+d}}$)

- therefore need $\frac{\alpha}{2\alpha+d} + \frac{\beta}{2\beta+d} > 1/2$, i.e., $\frac{\alpha+\beta}{2} > d/2$
- note L_2 norm rates are available under other conditions as well

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

$$\Big\{\phi_Y(Z;\mathbb{P}) - \psi\ \phi_A(Z;\mathbb{P})\Big\}\Big/\mathbb{E}\{\phi_A(Z;\mathbb{P})\}$$

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

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 ψ 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 (2017): Semiparametric theory

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!