Semiparametric Efficiency Theory in Causal Inference

Marcelo Ortiz-Villavicencio

Emory University

April 1, 2024

Econometrics Reading Group

Roadmap

Introduction

Introduction

- Most of this presentation is based on Kennedy (2016, 2023) work.
- In this presentation I want to review important aspects of *semiparametric theory* and *empirical process* that arise in causal inference problems.
- Under *semiparametric models*, we would like to allow parts of the DGP to be *unrestricted* if they are not of particular interest (i.e., nuisance functions).
- Semiparametric Theory gives us a framework for benchmarking *efficiency* and constructing estimators in such settings.
- All these tools support the incorporation of machine learning and other data-driven methods in causal inference (The basics before DML!).

Setup

Setup

- The first step in any causal inference application is define the *causal parameter of interest*.
- This parameter (or even a function) is formulated in terms of hypothetical interventions and counterfactual data (i.e, what would have been observed under some intervention?).
- Let $Y \in \mathbb{R}$ denote the outcome of interest and $D \in \{0, 1\}$ denote a *binary treatment*.
- Let $Y(d)$ denote the *potential outcome* that would have been observed under treatment level $D = d$.
- Throughout this presentation let's assume that our causal parameter of interest is the ATE:= $\psi = \mathbb{E}[Y(1) - Y(0)]$

Examples

- ATE: $\mathbb{E}\left[Y(1) - Y(0)\right]$
- conditional ATE: $\mathbb{E}\left[Y(1) - Y(0) \mid X = x\right]$
- local ATE: $\mathbb{E}\left[Y(1) - Y(0) \mid D(1) > D(0)\right]$
- dose-response curve: $\mathbb{E}\left[Y(d)\right]$
- heterogenous response curve: $\mathbb{E}\left[Y(d) \mid X = x\right]$
- optimal treatment strategy: $\arg\max_d \mathbb{E}\left[Y^{d(X)}\right]$

Identification Assumptions

- *Identification* is nothing more than translate the causal question of interest into a statistical problem defined in terms of observed data. For ATE we typically consider the following:
  1. **Consistency:** $D = d \implies Y = Y(d)$.
  2. **Unconfoundedness:** $Y(d) \perp D \mid X, \quad d = \{0, 1\}$. This assumption could be stronger that needed for ATE. We need $\mathbb{E}[Y(d) \mid X = x] = \mathbb{E}[Y(d) \mid D = d, X = x]$.
  3. **Positivity:** $p(D = d \mid X = x) \geq \delta > 0$ whenever $p(X = x) > 0$. This means each unit has a *non-zero* probability to receive treatment level $D = d$ regardless of covariate value.
- If the 3 conditions above hold, it follows that

$$p(Y(d) = y \mid X = x) = p(Y = y \mid X = x, D = d)$$

G-computation formula

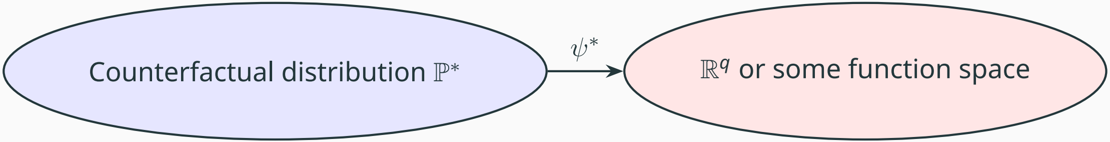
- The previous result means we can express the conditional distribution of the potential outcome $Y(d)$ given X in terms of observed data.
- Thus we can also identify the *conditional distribution* given any subset of X by simply *marginalizing*.

$$\psi = \int_{\mathscr{X}} \{\mathbb{E}(Y \mid X = x, D = 1) - \mathbb{E}(Y \mid X = x, D = 0)\} dP(X = x)$$

- This identification result is an example of the g-computation formula which was proposed by Robins (1986).

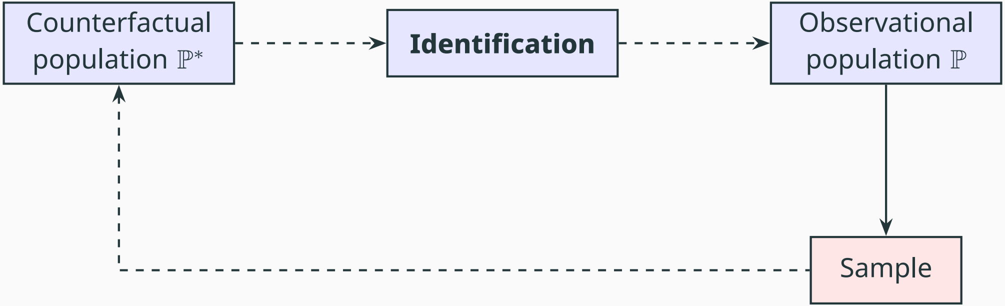
Identification

- $\psi^*(\mathbb{P}^*)$ is a map from a counterfactual distribution $\mathbb{P}^*$
- $\rightarrow$ can be a number, or function, or even more complex object



Causal Inference is over after identification

- A helpful approach is to think of the problem of causal identification and the problem of statistical estimation as separate issues.
- Causal identification only tells us what we should be estimating, not how to estimate it.
- After picking ψ^*, we need to express $\psi^*(\mathbb{P}^*) = \psi(\mathbb{P})$ for some mapping ψ and observational population distribution $\mathbb{P}$



- Now we have a pure functional estimation problem.

Semiparametric Theory

Semiparametric Models

- In this section, we give a general review of *semiparametric theory*, using as a running example the common problem of estimating an ATE.
- Standard semiparametric theory generally considers the following setting:
  - Observe iid sample $Z_1, \ldots, Z_n$ with $Z \sim \mathbb{P}$, assuming $\mathbb{P} \in \mathcal{P}$ is a unknown probability distribution on the Borel σ-field $\mathcal{B}$ for some sample space.
  - The general goal is estimation and inference for some target parameter $\psi = \psi(\mathbb{P}) \in R^p$, where $\psi = \psi(\mathbb{P})$ is a map from a probability distribution to the parameter space (assumed to be Euclidean here).
  - We want to construct a *good estimator* $\hat{\psi}$ of $\psi = \psi(\mathbb{P})$
- A *statistical model* $\mathcal{P}$ is a set of possible probability distributions, which is assumed to contain the observed data distribution $\mathbb{P}$.

Semiparametric Models

- In a parametric model, $\mathcal{P}$ is assumed to be indexed by a finite-dimensional real-valued parameter $\theta \in \mathbb{R}^q$. For example, if *Z* is a scalar RV one might assume it is normally distributed in which case the model is indexed by $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$.
- Semiparametric models are simply sets of probability distributions that cannot be indexed by only a Euclidean parameter, that is, models that are indexed by an *infinite-dimensional* parameter.
- Examples:
  - *nonparametric models* for which $\mathcal{P}$ consists of all possible probability distributions.
  - simple regression models that characterize the regression function *parametrically* but leave the residual error distribution unspecified.

Influence Functions

- Influence functions allow us to characterize a wide range of estimators and their *efficiency*.
- Let $\mathbb{P}_n = n^{-1}\sum_i \delta_{Z_i}$ denote the *empirical distribution* of the data, where δ_z is the *Dirac measure* that simply indicates whether $Z = z$.
- This means for example that empirical averages can be written as $n^{-1}\sum_i f(Z_i) = \int f(z)d\mathbb{P}_n = \mathbb{P}_n\{f(Z)\}$.

Definition

An estimator $\hat{\psi} = \hat{\psi}(\mathbb{P}_n)$ is *asymptotically linear* with influence function ϕ if the estimator can be approximated by an empirical average in the sense that

$$\hat{\psi} - \psi_0 = \mathbb{P}_n\{\phi(Z)\} + o_p(1/\sqrt{n}),$$

where ϕ has mean zero and finite variance (i.e., $\mathbb{E}\{\phi(Z)\} = 0$ and $\mathbb{E}\left\{\phi(Z)^{\otimes 2}\right\} < \infty$).

Asymptotic Normality

Theorem

By CLT, an estimator $\hat{\psi}$ with influence function ϕ is asymptotically normal with

$$\sqrt{n}\left(\hat{\psi} - \psi_0\right) \rightsquigarrow N\left(0, \mathbb{E}\left\{\phi(Z)^{\otimes 2}\right\}\right)$$

- Thus if we know the Influence functions for an estimator, we know its asymptotic distribution, and we can easily construct confidence intervals and hypothesis tests.
- Furthermore, *efficient influence function* for an asymptotically linear estimator is almost surely unique, so in a sense, the influence function contains all information about the asymptotic behavior of an estimator.

Nonparametric Efficiency Bounds

- Our next goal is to understand how well can we possibly hope to estimate the parameter ψ over the model $\mathcal{P}$.
- A classic *benchmarking* or *lower bound* results for smooth parametric models in the so-called *Cramer-Rao Lower Bound*.

Definition (CRLB)

For smooth parametric models $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}\}$ and smooth functionals (i.e., with P_θ and $\psi(\theta)$ differentiable in θ), the variance of any unbiased estimator $\widehat{\psi}$ must satisfy

$$\mathrm{var}_\theta(\widehat{\psi}) \geq \frac{\psi'(\theta)^2}{\mathrm{var}_\theta\left\{s_\theta(Z)\right\}},$$

where $s_\theta(z) = \frac{\partial}{\partial\theta} \log p_\theta(z)$ is the score function.

- i.e., no unbiased estimator can have smaller variance than the above ratio.

- A standard way to benchmark estimation error more generally is through minimax lower bounds of the form

$$\inf_{\widehat{\psi}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\{\widehat{\psi} - \psi(P)\}^2 \right] \geq R_n$$

Intuition: the risk for estimating ψ (in this case, in terms of worstcase mean squared error), over the model $\mathcal{P}$, cannot be smaller than R_n

Theorem (Theorem 8.11, van der Vaart (2000))

Assume P_θ is differentiable in quadratic mean at θ with nonsingular Fisher information $I_\theta = \text{var}_\theta \{s_\theta(Z)\}$. If $\psi(\theta)$ is differentiable at θ, with $\psi'(\theta) = \frac{\partial}{\partial \theta}\psi(\theta)$, then for any estimator $\widehat{\psi}$ it follows that

$$\inf_{\delta > 0} \liminf_{n \to \infty} \sup_{\|\theta' - \theta\| < \delta} n \mathbb{E}_{\theta'} \left[\left\{ \widehat{\psi} - \psi(\theta') \right\}^2 \right] \geq \psi'(\theta) \, \text{var}_\theta \{s_\theta(Z)\}^{-1} \psi'(\theta)^\top$$

Intuition: the (asymptotic, worst-case) mean squared error cannot be smaller than $\psi'(\theta)^2 / n \, \text{var}_\theta \{s_\theta(Z)\}$, for any estimator $\widehat{\psi}$ in a smooth parametric model.

Paremetric Submodels

EMORY UNIVERSITY

- Can the above *Cramer-Rao bounds* be exploited to construct lower bound benchmarks in semi- or non-parametric models?
- The standard way to do so is through a *parametric submodel*.

Definition

A parametric submodel is a smooth parametric model $\mathcal{P}_\epsilon = \{P_\epsilon : \epsilon \in \mathbb{R}\}$ that satisfies (i) $\mathcal{P}_\epsilon \subseteq \mathcal{P}$, and (ii) $P_{\epsilon=0} = \mathbb{P}$.

- The high-level idea behind the use of submodels is that it is never harder to estimate a parameter over a *smaller model*, relative to a larger one in which the smaller model is *contained*.
- Therefore, any lower bound for a submodel will also be a valid lower bound for the larger model $\mathcal{P}$.
- Since any lower bound for the submodel $\mathcal{P}_\epsilon$ is also a lower bound for $\mathcal{P}$, the best and most informative is the *greatest* such lower bound.

- Recall the CRLB for submodel $\mathcal{P}_\epsilon$ is given by $\frac{\left\{ \frac{\partial}{\partial \epsilon} \psi(P_\epsilon) \big|_{\epsilon=0} \right\}^2}{\mathbb{E}_{P_\epsilon}\{s_\epsilon(Z)^2\}}$
- To find the best such lower bound, we would like to optimize the above over all P_ϵ in some submodel.

Definition (Distributional Taylor Expansion)

Suppose the functional $\psi : \mathcal{P} \mapsto \mathbb{R}$ is smooth, in the sense that it admits a kind of distributional Taylor expansion

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

for distributions $\bar{P}$ and P, often called a *von Mises expansion*, where $\varphi(z; P)$ is a meanzero, finite-variance function satisfying $\int \varphi(z; P) dP(z) = 0$ and $\int \varphi(z; P)^2 dP(z) < \infty$, and $R_2(\bar{P}, P)$ is a 2nd-order remainder term.

Intuition: describes how ψ changes locally, when moving from P to $\bar{P}$. Any φ satisfying above is an influence function for ψ.

Example

The average treatment effect

$$\psi(P) = \mathbb{E}_P \{\mathbb{E}_P(Y \mid X, D = 1)\}$$

satisfies *von Mises expansion* with

$$\varphi(Z; P) = \frac{1(D = 1)}{P(D = 1 \mid X)} \{Y - \mathbb{E}_P(Y \mid X, D = 1)\} + \mathbb{E}_P(Y \mid X, D = 1) - \psi(P)$$

and

$$R_2(\bar{P}, P) = \int \left\{ \frac{1}{\bar{\pi}(x)} - \frac{1}{\pi(x)} \right\} \{\mu(x) - \bar{\mu}(x)\} \pi(x) dP(x)$$

where $\pi(x) = P(D = 1 \mid X = x)$ and $\bar{\pi}(x) = \bar{P}(D = 1 \mid X = x)$, and similarly for $\mu(x) = \mathbb{E}_P(Y \mid X = x, D = 1)$.

Efficient Influence Function

- We now have enough to characterize the greatest *lower bound* for generic smooth parametric submodels.
- A common choice of submodel for nonparametric $\mathcal{P}$ is, for some mean-zero function $h : \mathcal{Z} \to \mathbb{R}$,

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

  where $\|h\|_\infty \leq M < \infty$ and $\epsilon < 1/M$ so that $p_\epsilon(z) \geq 0$. For this submodel the score function is $\frac{\partial}{\partial \epsilon} \log p_\epsilon(z)\big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon} \log\{1 + \epsilon h(z)\}\big|_{\epsilon=0} = h(z)$.

Efficient Influence Function

- For previous submodel, the score is $s_\epsilon(z) = h(z)$ and by *pathwise differentiability* we have

$$\frac{\partial}{\partial \epsilon}\psi\left(P_\epsilon\right)\bigg|_{\epsilon=0} = \int \varphi(z;\mathbb{P})h(z)d\mathbb{P}(z).$$

  Therefore over all CRLB at $\epsilon = 0$ we have

$$\sup_{P_\epsilon} \frac{\psi'\left(P_\epsilon\right)^2}{\text{var}\left\{s_\epsilon(Z)\right\}} = \sup_h \frac{\mathbb{E}\{\varphi(Z;\mathbb{P})h(Z)\}^2}{\mathbb{E}\left\{h(Z)^2\right\}} \le \mathbb{E}\left\{\varphi(Z;\mathbb{P})^2\right\} = \text{var}\{\varphi(Z)\}$$

  where the first equality follows by *pathwise differentiability* and the form of the submodel, and the inequality by *Cauchy-Schwarz*.
- Therefore $\text{var}\{\varphi(Z)\}$ is nonparametric analog of CRLB! - we call φ the efficient influence function.

Deriving Influence Functions

- There are at least 3 ways to derive IFs.
- Most general: compute 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
- Kennedy Method: use chain/product rules w/ simple IFs as building blocks:

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.,

$$\mathbb{IF}(\mathbb{E}(Y \mid X = x)) = \frac{1(X = x)}{\mathbb{P}(X = x)}\{Y - \mathbb{E}(Y \mid X = x)\}$$

Efficient Influence Function for ATE

Example

Let $\mu(x) = \mathbb{E}(Y \mid X = x, D = 1), \pi(x) = \mathbb{P}(D = 1 \mid X = x)$, and $p(x) = \mathbb{P}(X = x)$, and let $\psi = \mathbb{E}\{\mathbb{E}(Y \mid X, D = 1)\}$ denote the ATE. Then the influence function is given by

$$\begin{aligned}
\mathbb{IF}(\psi) &= \mathbb{IF}\left\{\sum_x \mu(x)p(x)\right\} = \sum_x [\mathbb{IF}\{\mu(x)\}p(x) + \mu(x)\mathbb{IF}\{p(x)\}] \\
&= \sum_x \frac{1(X = x, D = 1)}{p(1, x)}\{Y - \mu(x)\}p(x) + \mu(x)\{1(x = X) - p(x)\}\Bigg] \\
&= \frac{D}{\pi(X)}\{Y - \mu(X)\} + \mu(X) - \psi
\end{aligned}$$

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

Application: Non-parametric methods for doubly robust estimation of continuous treatments

Big Picture

- Kennedy et al (2017) shows how we can apply the previous concepts for more complex causal parameter like the ones with continuous treatments.
- This paper develops a novel *kernel smoothing* approach with mild smoothness assumptions on the *effect curve* allowing for *doubly robust covariate adjustment*.
- Derives asymptotic properties and provides a data-driven procedure for *bandwidth selection*.
- Illustrates its perks via simulations and a study of the effect of nurse staffing on hospital readmission penalties.
- Empirical Application: Study whether *nurse staffing* (the treatment, measured in nurse hours per patient day) affected hospitals' risk of excess readmission penalty (in the context of the Hospital re-admissions reduction program (2012)).
- $\Rightarrow$ Hospitals differ in many important ways that could be related to both nurse staffing and excess re-admissions like size, location, teaching status, etc. To make fair comparisons, we must adjust for hospital characteristics!

Background

- We are interested in *continuous treatments* such as dose, duration, or frequency that arise often in observational studies.
- Such treatments lead to effects that are described by *dose-response curves* rather than scalars as in binary treatments.
- There 2 methodological challenges in this setting:
  1. How to discover underlying structure of dose-response curves *without imposing* a prior *shape restrictions*.
  2. How to adjust properly for *high dimensional confounders*.

Literature

- One of the approaches for estimating continuous treatment effects is based on *regression modeling*.
  - Needs correct specification of the outcome model
  - Does not incorporate available information about the treatment mechanism
  - Sensitive to the curse of dimensionality.
- Another one is *semiparametric doubly robust*
  - Rely on parametric models for the dose-response estimation.
- Recent work extended semiparametric doubly robust methods to non-parametric and high dimensional settings
- This paper: New approach for *causal dose-response* that is *DR* without requiring parametric assumptions and can incorporate general ML methods.

Setup

- i.i.d sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ where $\mathbf{Z} = (\mathbf{L}, A, Y)$ has support $\mathcal{Z} = (\mathcal{L} \times \mathcal{A} \times \mathcal{Y})$
- $\mathbf{L}$ denotes a vector of covariates, A a continuous treatment and Y outcome of interest
- Let Y^a potential outcome under *treatment level* a
- Denote the distribution of $\mathbf{Z}$ by P, with density $p(\mathbf{z}) = p(y \mid \mathbf{l}, a)p(a \mid \mathbf{I})p(\mathbf{l})$
- Denote mean outcome given covariates and treatment as $\mu(\mathbf{l}, a) = \mathbb{E}(Y \mid \mathbf{L} = \mathbf{l}, A = a)$
- Let conditional treatment density given covariates $\pi(a \mid \mathbf{l}) = \partial P(A \leqslant a \mid \mathbf{L} = \mathbf{l})/\partial a$
- Let marginal treatment density $\varpi(a) = \partial P(A \leqslant a)/\partial a$.

Identification

- Our goal is to estimate the *effect curve* $\theta(a) = \mathbb{E}[Y^a]$.

Assumption (Consistency)

$A = a$ *implies* $Y = Y^a$. *No interference and no different versions of the treatment*

Assumption (Positivity)

$\pi(a \mid \mathbf{l}) \geqslant \pi_{\min} > 0$ *for all* $\mathbf{l} \in \mathcal{L}$. *Every subject has some chance of receiving treatment level a, regardless of covariates.*

Assumption (Ignorability)

$\mathbb{E}\left(Y^a \mid \mathbf{L}, A\right) = \mathbb{E}\left(Y^a \mid \mathbf{L}\right)$. *Treatment assignment is unrelated to potential outcomes within strata of covariates.*

Previous assumptions are satisfied in RCTs, but in observational studies may be violated and generally untestable.

Definition (Identification)

Under assumptions 1-3, the effect curve $\theta(a)$ can be identified with observed data as

$$\theta(a) = \mathbb{E}\{\mu(\mathbf{L}, a)\} = \int_{\mathcal{L}} \mu(\mathbf{l}, a) \mathrm{d}P(\mathbf{l})$$

Doubly robust mapping

EMORY UNIVERSITY

- This paper derive *doubly robust estimators* for $\theta(a)$ without relying on parametrics models.
- Our goal is to *find a function* $\xi(\mathbf{Z}; \pi, \mu)$ of the observed data $\mathbf{Z}$ and nuisance functions (π, μ) such that

$$\mathbb{E}\{\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu}) \mid A = a\} = \theta(a)$$

  if either $\bar{\pi} = \pi$ or $\bar{\mu} = \mu$ (not necessarily both).
- Given such a mapping, *off-the-shelf non-parametric regression* and *machine learning* methods could be used to estimate $\theta(a)$ by regressing $\xi(\mathbf{Z}; \hat{\pi}, \hat{\mu})$ on treatment A, based on estimates $\hat{\pi}$ and $\hat{\mu}$.

Doubly robust mapping

EMORY UNIVERSITY

- This mapping is related to the *efficient influence function* for a particular parameter.
- If $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu}) \mid A = a\} = \theta(a)$ then it follows that $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu})\} = \psi$ for

$$\psi = \int_{\mathcal{A}}\int_{\mathcal{L}} \underbrace{\mu(\mathbf{l},a)}_{\text{expected outcome given cov. + treat.}} \underbrace{\varpi(a)}_{\text{marginal treat. density}} \mathrm{d}P(\mathbf{l})\mathrm{d}a.$$

- The *efficient influence function* $\phi(\mathbf{Z};\pi,\mu)$ will be doubly robust such that $\mathbb{E}\{\phi(\mathbf{Z};\pi,\mu)\} = \mathbb{E}\{\xi(\mathbf{Z};\pi,\mu) - \psi\} = 0$, so $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu})\} = \psi$ if either $\bar{\pi} = \pi$ or $\bar{\mu} = \mu$.

Efficient Influence Function

- The parameter ψ represents the *average outcome under an intervention* that *randomly* assigns treatment based on the density ϖ.
- The efficient influence function for ψ has not been given before under a *non-parametric model* (i.e., suppose that the marginal density ϖ is unknown).

Theorem (Efficient Influence Function under Non-parametric model)

Under a non-parametric model, the efficient influence function for ψ is

$$\xi(\mathbf{Z};\pi,\mu)-\psi+\int_{\mathcal{A}}\left\{\mu(\mathbf{L},a)-\int_{\mathcal{L}}\mu(\mathbf{l},a)\mathrm{d}P(\mathbf{l})\right\}\varpi(a)\mathrm{d}a$$

where $\xi(\mathbf{Z};\pi,\mu)=\frac{Y-\mu(\mathbf{L},A)}{\pi(A|\mathbf{L})}\int_{\mathcal{L}}\pi(A\mid\mathbf{l})\mathrm{d}P(\mathbf{l})+\int_{\mathcal{L}}\mu(\mathbf{l},A)\mathrm{d}P(\mathbf{l})$

Approach Proposed

- Our goal is to derive a doubly robust mapping $\xi(\mathbf{Z}; \pi, \mu)$ for which $\mathbb{E}\{\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu}) \mid A = a\} = \theta(a)$, as long as either $\bar{\pi} = \pi$ or $\bar{\mu} = \mu$, in a *two-step procedure*:
  1. Estimate nuisance functions (π, μ) and obtain predicted values.
  2. Construct pseudo-outcome $\hat{\xi}(\mathbf{Z}; \hat{\pi}, \hat{\mu})$ and regress on treatment variable A.

Estimator using Local Linear Regression

- For step 2, one can propose an estimator that uses kernel smoothing such as *Local Linear Kernel Regression*.
- Let $\hat{\theta}_h(a) = \mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\beta}_h(a)$, where $\mathbf{g}_{ha}(t) = (1, (t-a)/h)^{\mathrm{T}}$ and

$$\hat{\beta}_h(a) = \underset{\beta \in \mathbb{R}^2}{\arg\min} \mathbb{P}_n \left[K_{ha}(A) \left\{ \hat{\xi}(\mathbf{Z}; \hat{\pi}, \hat{\mu}) - \mathbf{g}_{ha}(A)^{\mathrm{T}}\beta \right\}^2 \right]$$

for $K_{ha}(t) = h^{-1}K\{(t-a)/h\}$ with K a *standard kernel function* (e.g. a symmetric probability density) and h a *scalar bandwidth* parameter.

Thanks!

marcelo.ortiz@emory.edu
marcelortiz.com
@marcelortizv

References

- Kennedy, E. H. (2015). *Semiparametric Theory and Empirical Processes in Causal Inference*. arXiv. https://arxiv.org/abs/1510.04740
- Kennedy, E. H., Ma, Z., McHugh, M. D., Small, D. S. (2017). *Nonparametric methods for doubly robust estimation of continuous treatment effects*. Journal of the Royal Statistical Society: Series B, 79(4), 1229-1245. https://doi.org/10.1111/rssb.12212 (arXiv:1507.00747)