Causal Inference: Influence Functions and von Mises Calculus

Hyunseung Kang

May 1, 2024

Abstract

We discuss another approach to estimating causal estimands based on the efficient influence function (EIF). A lot of this document follows the beautiful exposition in Kennedy [2022]; this paper is especially useful if you are already familiar with empirical processes. I also suggest reading Hines et al. [2022] if you want to dive into the "mechanics" of constructing estimators based on the EIF and `https://alejandroschuler.github.io/mci/introduction-to-modern-causal-inference.html` if you need a more comprehensive, but gentle introduction to this topic. This document assumes that you have taken a Ph.D. course in mathematical statistics.

1 Review of Some Concepts

We review some concepts to help us understand influence functions.

- **O_p and o_p notation**: Given a sequence of random variables X_n and a sequence of positive, fixed numbers r_n, $X_n = o_p(r_n)$ means that $X_n/r_n \to 0$ in probability and $X_n = O_p(r_n)$ is X_n/r_n is bounded in probability, i.e. $\forall \epsilon > 0$, there exist $M > 0$ and $N > 0$ where $P(|X_n/r_n| > M) < \epsilon$ for all $n > N$. Some related results include:
  - $X_n = o_p(1)$ implies that $X_n \to 0$ in probability.
  - $X_n \to X$ in distribution implies that $X_n = O_p(1)$.
  - $O_p(r_n) = r_n O_p(1)$ and $o_p(r_n) = r_n o_p(1)$
- **Taylor's Theorem**: Consider any function $f : \mathbb{R} \to \mathbb{R}$ and with at least 2 derivatives at and near the neighborhood of x_0. Then, we have

$$f(x) = f(x_0) + \underbrace{f'(x_0)(x - x_0)}_{\text{First order}} + \underbrace{\frac{1}{2} f''(x_{\text{mid}})(x - x_0)^2}_{\text{Remainder } R(f(x), f(x_0))}$$

  where x_{mid} is in between x and x_0 (i.e. in the neighhorhood of x_0). Notably, this theorem implies that when x deviates from x_0 by Δ, we have

$$f(x_0 + \Delta) - f(x_0) = f'(x_0)\Delta + \frac{1}{2} f''(x_{\text{mid}})\Delta^2$$

2 The Approach

2.1 A Functional Perspective on Statistical Estimands

Suppose we are interested in studying some low-dimensional feature of a distribution F where F is the cumulative distribution function. A bit more formally, we are interested in a functional $\psi(F) : \mathcal{F} \to \mathbb{R}$ where $\mathcal{F}$ denotes a set of cumulative distribution functions. Some examples include:

- The population mean: $\psi(F) = \mathbb{E}[O]$.
- The population variance: $\psi(F) = \text{Var}[O]$.
- Mean squared error of a fixed decision rule δ: $\psi(F) = \mathbb{E}[(O - \delta)^2]$.
- The average treatment effect: $\psi(F) = \mathbb{E}[\mathbb{E}[Y \mid A = 1, X]]$.

To study ψ, we take n i.i.d. samples $O_1, \ldots, O_n$ from a distribution $F \in \mathcal{F}$. Note that we can obtain a uniformly consistent estimate of F with the empirical cumulative distribution function, i.e. $F_n = n^{-1} \sum_{i=1}^n I(O_i \leq t)$ by the Glivenko-Cantelli Theorem; in other words, with sufficient sample size, F is reasonably close to F_n.

Given this, a natural choice to estimate $\phi(F)$ is to replace F with F_n and study the behavior of

$$\psi(F_n) - \psi(F) \tag{1}$$

as F_n gets close to F.

2.2 Derivative of $\psi(\cdot)$ and the influence function

Inspired by Taylor's theorem, a natural way to study equation (1) would be to conduct a version of Taylor expansion of (1). This exercise requires extending the notion of differentiability of ψ with respect to a distribution function F. We define this derivative in two steps:

1. First, we describe how F changes in the space of distribution functions $\mathcal{F}$. Consider a small deviation from F in the form of $F_\epsilon = (1-\epsilon)F + \epsilon\delta_o = F + \epsilon(\delta_o - F)$ where $H \in \mathcal{F}$ and $\epsilon \geq 0$. Here, δ_o is the diract delta function at some support point o, i.e. $\delta_o = I(O = o)$.
2. Second, we then measure the infinitesimal change in ψ as it moves from F_ϵ to F:

$$\text{IF}(o; \psi, F) = \lim_{\epsilon \downarrow 0} \frac{\psi(F_\epsilon) - \psi(F)}{\epsilon} = \frac{\delta}{\delta\epsilon}\psi(F_\epsilon) \mid_{\epsilon=0} \tag{2}$$

If this limit exists, this is called the influence curve of ψ at the point F. More loosely stated, $\text{IF}(o; \psi, F)$ is the derivative of the function ψ at the point F. Note that the derivative $\frac{\delta}{\delta\epsilon}$ is the "usual" derivative from calculus.[1]

Some examples of this derivative are included below:

- Population mean: We have $F_\epsilon = (1-\epsilon)F + \epsilon\delta_o$, $\psi(F_\epsilon) = (1-\epsilon)\mathbb{E}[O] + \epsilon o$, and $\psi(F) = \mathbb{E}[O]$. Then

$$\text{IF}(o; \psi, F) = \lim_{\epsilon \to 0} \frac{(1-\epsilon)\mathbb{E}[O] + \epsilon o - \mathbb{E}[O]}{\epsilon} = \lim_{\epsilon \to 0} \frac{\epsilon(o - \mathbb{E}[O])}{\epsilon} = o - \mathbb{E}[O]$$

- Population variance: $\psi(\mathbb{P}) = \text{Var}[O]$ and $\psi(F_\epsilon) = (1-\epsilon)\text{Var}[O] + \epsilon(o - \mathbb{E}[O])^2$. Then,

$$\text{IF}(o; \psi, F) = \lim_{\epsilon \to 0} \frac{(1-\epsilon)\text{Var}[O] + \epsilon(o - \mathbb{E}[O])^2 - \text{Var}[O]}{\epsilon} = (o - \mathbb{E}[O])^2 - \text{Var}[O]$$

- Z estimator: Suppose $\mathbb{E}[g(O, \theta^*)] = 0$ for some θ^* and we are interested in estimating $\theta^* = \psi(F)$. Then, Example 20.4 of van der vaart shows that

$$\text{IF}(o; \psi, F) = -\mathbb{E}\left[\nabla_\theta g(O, \theta)|_{\theta=\theta^*}\right]^{-1} g(o, \theta^*).$$

- Average treatment effect. $\psi(F) = \mathbb{E}[\mathbb{E}[Y \mid A = 1, X]]$. Let $\mu_1 = \mathbb{E}[Y \mid A = 1, X]$ and $\pi(X) = P(A = 1 \mid X)$. For pedagogy, we'll assume the density of the distribution $P(Y, A = 1, X)$ exists and is denoted by $f(y, 1, x)$. We also denote the density of F_ϵ as f_ϵ. Finally, we assume that we can exchange derivatives and integrals; see below.

  From the definition of conditional distributions, we arrive at

$$\psi(F_\epsilon) = \int\int y \frac{f_\epsilon(y, 1, x) f_\epsilon(x)}{f_\epsilon(1, x)} dy dx$$

  Taking the derivative w.r.t. ϵ and exchanging derivatives with integrals give us

$$\begin{aligned}
\frac{\delta}{\delta\epsilon}\psi(F_\epsilon) &= \int\int \frac{\delta}{\delta\epsilon} \frac{f_\epsilon(y, 1, x) f_\epsilon(x)}{f_\epsilon(1, x)} dy dx \\
&= \int\int \left(\frac{y(I(y, 1, x) - f(y, 1, x)) f_\epsilon(x)}{f_\epsilon(1, x)} + \frac{y f_\epsilon(y, 1, x)(I(x) - f(x))}{f_\epsilon(1, x)} - \frac{y f_\epsilon(y, 1, x) f_\epsilon(x)(I(1, x) - f(1, x))}{f_\epsilon^2(1, x)} \right) dy dx
\end{aligned}$$

[1] If it's helpful, think of ϵ as a parameter θ of a distribution F.

Evaluating the derivative at $\epsilon = 0$ gives us

$$
\begin{aligned}
&\frac{\delta}{\delta\epsilon}\psi(F_\epsilon)\mid_{\epsilon=0} \\
=&\int\int\left(\frac{y(I(y,1,x)-f(y,1,x))f(x)}{f(1,x)}+\frac{yf(y,1,x)(I(x)-f(x))}{f(1,x)}-\frac{yf(y,1,x)f(x)(I(1,x)-f(1,x))}{f^2(1,x)}\right)dydx \\
=&\int\int y\frac{f(y,1,x)f(x)}{f(1,x)}\left[\left(\frac{I(y,1,x)}{f(y,1,x)}-1\right)+\left(\frac{I(x)}{f(x)}-1\right)-\left(\frac{I(1,x)}{f(1,x)}-1\right)\right]dydx \\
=&\int\int y\frac{f(y,1,x)f(x)}{f(1,x)}\left[\frac{I(y,1,x)}{f(y,1,x)}+\frac{I(x)}{f(x)}-\frac{I(1,x)}{f(1,x)}-1\right]dydx \\
=&\frac{I(A=1)}{\pi(x)}\left(y-\mu_1(x)\right)+\mu_1(x)-\mathbb{E}[\mu_1(X)]
\end{aligned}
$$

2.3 Important Properties of Influence Functions

There are two key properties of the influence functions.

- The influence function has mean zero when the expectation is evaluated at F, i.e.

$$\mathbb{E}_F[\mathrm{IF}(O;\psi,F)] = 0$$

  I add subscript F in the expectation to emphasize that the expectation is evaluated with respect to F. Because of this, $\mathrm{Var}_F[\mathrm{IF}(O;\psi,F)] = \mathbb{E}_F[\mathrm{IF}^2(O;\psi,F)]$.
- If the tangent space (see below) is the entire Hilbert space of mean-zero, finite variance functions and the influence function of ψ at the point F exists (see Theorem 4.4 of Tsiatis [2006]), this is the only influence function (see Theorem 4.3 of Tsiatis [2006]). Roughly stated, if you find an influence function for ψ, this is going to be the efficient influence function.

2.4 von Mises Expansion and the One-Step Estimator

Once we have a notion of a derivative for $\psi(\cdot)$, we can use an analogy of Taylor's theorem on $\psi(\cdot)$. Specifically, the von Mises expansion states that for two distributions $F', F \in \mathcal{P}$, the difference $\psi(F') - \psi(F)$ can be written as

$$
\begin{aligned}
\psi(F') - \psi(F) &= \int \mathrm{IF}(o;\psi,F')(dF' - dF)o + R(F',F) \\
&= -\mathbb{E}_F[\mathrm{IF}(O;\psi,F')] + R(F',F).
\end{aligned}
$$

The term $\mathbb{E}_F[\mathrm{IF}(O;\psi,F')]$ represents the bias from plugging in F' instead of F into the influence function. Note that this term may still not go away if we replace F' with the empirical cumulative distribution function F_n.

Then, a natural way to correct this plug-in bias is to add $\mathbb{E}_F[\mathrm{IF}(O;\psi,F')]$ to $\psi(F')$, i.e. $\psi(F')+\mathbb{E}_F[\mathrm{IF}(O;\psi,F')]$. More formally, if we obtain $O_i \overset{\text{iid}}{\sim} F$, consider an estimate of F, say $\hat{F}$, and the bias-corrected estimator of $\psi(F)$:

$$\hat{\psi} = \psi(\hat{F}) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}(O_i;\psi,\hat{F}) \tag{3}$$

This is known as the one-step estimator. The asymptotic analysis of this estimator proceeds by looking at the three terms (A), (B), and (C) described below:

$$\begin{aligned}
\hat{\psi} - \psi(F) &= \psi(\hat{F}) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}(O_i;\psi,\hat{F}) - \psi(F) \\
&= \underbrace{\psi(\hat{F}) + \mathbb{E}_F[\mathrm{IF}(O;\psi,\hat{F})] - \psi(F)}_{R(\hat{F},F)} + \mathbb{E}_F[\mathrm{IF}(O;\psi,\hat{F})] - \frac{1}{n}\sum_{i=1}^{n}\mathrm{IF}(O_i;\psi,\hat{F}) \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathrm{IF}(O_i;\psi,F) - \mathbb{E}_F[\mathrm{IF}(O_i;\psi,F)]}_{\text{(A)}} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big[\mathrm{IF}(O_i;\psi,\hat{F}) - \mathrm{IF}(O_i;\psi,F)\Big] - \mathbb{E}_F\Big[\mathrm{IF}(O_i;\psi,\hat{F}) - \mathrm{IF}(O_i;\psi,F)\Big]}_{\text{(B)}} \\
&\quad + \underbrace{R(\hat{F},F)}_{\text{(C)}}
\end{aligned}$$

The term (A) is a mean-zero random variable and should behave like $O_p(1/\sqrt{n})$. The term (B) is an empirical process term, which requires either Donsker conditions on $\mathrm{IF}(\psi,\hat{F}) - \mathrm{IF}(\psi,F)$ or sample splitting, to ensure that it behaves like $o_p(1/\sqrt{n})$. In particular, if $\hat{F}$ is constructed from an independent sample, say $\hat{F}^{\perp}$, Lemma 1 of Kennedy [2022] showed that

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\mathrm{IF}(O_i;\psi,\hat{F}^{\perp}) - \mathrm{IF}(O_i;\psi,F)\Big] - \mathbb{E}_F\Big[\mathrm{IF}(O_i;\psi,\hat{F}^{\perp}) - \mathrm{IF}(O_i;\psi,F)\Big] = O_p\left(\frac{\|\mathrm{IF}(O_i;\psi,\hat{F}^{\perp}) - \mathrm{IF}(O_i;\psi,F)\|_2}{\sqrt{n}}\right)$$

In other words, we only need $\|\mathrm{IF}(O_i;\psi,\hat{F}^{\perp}) - \mathrm{IF}(O_i;\psi,F)\|_2 = o_p(1)$ in order for the second term to behave like $o_p(1/\sqrt{n})$. The term (C) requires a case-by-case analysis in order to ensure $o_p(1/\sqrt{n})$ and for some problems, it can be annoying to deal with. Combined, the one-step estimator $\hat{\psi}$'s asymptotic variance is determined by the first term.

3 Example with the ATE Estimator

Let $\hat{\mu}_1(X) = \hat{\mathbb{E}}[Y \mid A=1, X]$ and $\hat{\pi}(X) = \hat{\mathbb{E}}[A \mid X]$. Throughout the exercise, we assume $0 < \pi(X_i)$. Suppose we consider the one-step estimator for $\psi(F) = \mathbb{E}[\mathbb{E}[Y \mid A=1, \mathbf{X}]]$ based on its influence function above, i.e.

$$\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=1)}{\hat{\pi}(X_i)}(Y_i - \hat{\mu}_1(X_i)) + \hat{\mu}_1(X_i)$$

The term (A) behaves like a Normal random variable:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}(O_i;\psi,F) - \mathbb{E}_F[\mathrm{IF}(O_i;\psi,F)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{A_i(Y_i - \mu_1(X_i))}{\pi(X_i)} + \mu_1(X_i) - \mathbb{E}[\mu_1(X_i)]\right) \to N(0,\sigma^2)$$

where σ^2 is the variance of the influence function $IF(O_i;\psi,F)$ evaluated at the true value F.

For the term (B), if we obtained an estimate of $\hat{\mu}_1(X_i)$ and $\hat{\pi}(X_i)$ from an independent sample, we only need to study the behavior of the term

$$\begin{aligned}
&\mathrm{IF}(O_i;\psi,\hat{F}^{\perp}) - \mathrm{IF}(O_i;\psi,F) \\
&= \left(\frac{A_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} + \hat{\mu}_1(X_i)\right) - \left(\frac{A_i(Y_i - \mu_1(X_i))}{\pi(X_i)} + \mu_1(X_i)\right) \\
&= \left(1 - \frac{A_i}{\pi(X_i)}\right)(\hat{\mu}_1(X_i) - \mu_1(X_i)) + \frac{A_i(Y_i - \hat{\mu}_1(X_i))(\pi(X_i) - \hat{\pi}(X_i))}{\hat{\pi}(X_i)\pi(X_i)}
\end{aligned}$$

As long as (a) both the estimated propensity score is bounded strictly away from 0 (b) the second moment of $Y - \hat{\mu}_1(X_i)$ is finite, and (c) the outcome regression estimator and the propensity score estimator are both consistent (i.e. $\|\hat{\mu}_1(X_i) - \mu_1(X_i)\|_2 = o_p(1)$ and $\|\hat{\pi}(X_i) - \pi(X_i)\|_2 = o_p(1)$), we have $\|\mathrm{IF}(O_i;\psi,\hat{F}^{\perp}) - \mathrm{IF}(O_i;\psi,F)\|_2 = o_p(1)$.

We remark that we can replace (c) with a condition where only one of the estimators are consistent. In this case, $\mathrm{IF}(\psi, F)$ is replaced by $\mathrm{IF}(\psi, F_{\mathrm{mis}})$ where F_{mis} denotes a model where either the propensity score or the outcome regression is mis-specified.[2]

For the term (C), its explicit form can be derived from the definition of the remainder term in the von Mises expansion:

$$\begin{aligned} R(\hat{F}, F) &= \psi(\hat{F}) - \psi(F) + \mathbb{E}_F[\mathrm{IF}(O; \psi, \hat{F}] \\ &= \mathbb{E}_F\left[\left(\frac{1}{\hat{\pi}(X_i)} - \frac{1}{\pi(X_i)}\right)(\mu_1(X_i) - \hat{\mu}_1(X_i))\,\pi(X_i)\right]. \end{aligned}$$

As long as (a) the estimated propensity score is bounded strictly away from 0, we have

$$|R(\hat{F}, F)| \leq C\|\pi(X_i) - \hat{\pi}(X_i)\|_2 \cdot \|\mu_1(X_i) - \hat{\mu}_1(X_i)\|_2$$

and $C > 0$ is some constant. Thus, as long as the product of these two estimates are of order $o_p(1/\sqrt{n})$, we get the desired rate. We remark that this is where the term "doubly robust rates" arises.

References

O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

A. A. Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

[2]If we do this, the term (A) still behaves like a mean-zero Normal random variable, albeit with a different variance.