Foundations of Semiparametric Theory

MATH 470

Mila Pourali

Supervised by Dr. Mehdi Dagdoug
Co-supervised by Dr. Tim Hoheisel

Department of Mathematics and Statistics
McGill University
Summer 2024

Contents

1 Introduction

In statistics, we generally have data in the form of a collection of random vectors that have been generated with respect to some true data-generating distribution that is unknown to us. Our goal becomes to figure out what this process is in order to be able to either generate some more of this data, or make inference about its behaviour. In order to carry out this search for the truth, we turn to statistical models.

We define a *statistical model* to be a family of distributions

$$\mathcal{P} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\},$$

where θ is some parameter that characterizes the distribution. Often times, we are not interested in estimating the entire parameter space, but only a subset. For example, if we are dealing with data that we assume may be distributed according to a $\mathcal{N}(\mu, \sigma^2)$ distribution, we concern ourselves with the class of densities

$$\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\},$$

where we can define the full parameter $\theta = \begin{bmatrix}\mu & \sigma^2\end{bmatrix}^\top$ to be and element of $\mathbb{R} \times \mathbb{R}^+$. We might only be interested in estimating the mean μ, but we still have to deal with the problem of estimating σ^2 in order to fully characterize the distribution of our data. Notice that by assuming our data comes from the class of normal densities, we are, in fact, adding restrictions to our model.

Restrictions may come in many forms, and the more of them we have, the easier our estimation problem becomes because our search for the truth requires less depth. However, it also means that we are more likely to be wrong. Suppose we do wish to loosen the restrictions on our model. One way of doing this is by allowing our parameter to be infinite-dimensional as opposed to finite-dimensional. In other words, we may want to consider a parameter θ which can be partitioned into a vector of the form

$$\theta = \begin{bmatrix}\beta \\ \eta\end{bmatrix},$$

where β remains a finite-dimensional parameter that we are truly interested in estimating, but η is an infinite-dimensional parameter that may not be our main point of focus, but that we do still need to deal with in order to properly characterize our probability distribution. We call β the *parameter of interest*, and η a *nuisance parameter*. A model that is characterized as such is called a semiparametric model. Although this makes our estimation problem more difficult, it also means that if we do find solutions, they will be a lot more robust since we have fewer restrictions on our model.

The theory of semiparametric statistics begins with influence functions, which will be the subject of section 4, as well as the main focus of this report. However, in order to build the theory of influence functions and how we can use them to perform estimation in semiparametric settings, we require some tools from asymptotic statistics and functional analysis, which we will cover in sections 2 and 3. The main sources of reference for the contents of this report when it comes to semiparametric theory are Tsiatis (2006), as well as Bickel et al. (1993) and Kosorok (2008).

Note 1.1. Sometimes, it may be useful to define β as a function of θ, $\beta(\theta)$, given that $\beta(\theta)$ may not always be a natural projection.

2 Some Tools from Functional Analysis

In this section, we will recall some important results and notions from functional analysis ranging from norms to projections. It is also worth noting that we will not prove nor immediately use every result in this section, but for the sake of completeness, we include these results and some of their proofs regardless. For the most part, the results in this section are taken from Axler (2020).

Definition 2.1. A metric on a nonempty set V is a function $d : V \times V \to [0, \infty)$ such that for $f, g, h \in V$,

- $d(f, f) = 0$ for all $f \in V$,
- $d(f, g) = 0$ implies $f = g$,
- $d(f, g) = d(g, f)$,
- $d(f, h) \leq d(f, g) + d(g, h)$ (triangle inequality).

Definition 2.2. Let (V, d_V) and (W, d_W) be metric spaces. A map $T : V \to W$ is continuous at $f \in V$ if for all $\epsilon > 0$, there exists $\delta > 0$ such that

$$d_W(T(f), T(g)) < \epsilon,$$

for all $g \in V$ such that $d_V(f, g) < \delta$.

Definition 2.3. A metric space is complete if every Cauchy sequence in it converges to a limit, also contained in the space.

Definition 2.4. A norm on a real vector space V is a function $|| \cdot || : V \to [0, \infty)$ such that

- $||f|| = 0$ if and only if $f = 0$,
- $||\alpha f|| = |\alpha| \cdot ||f||$ for all $f \in V$, and $\alpha \in \mathbb{R}$,
- $||f + g|| \leq ||f|| + ||g||$ for all $f, g \in V$.

Definition 2.5. A normed vector space, complete for the induced metric, is called a Banach space.

Definition 2.6. Let $\alpha \in \mathbb{R}$. We call $T : V \to W$ a linear map if

- $T(f + g) = T(f) + T(g)$,
- $T(\alpha f) = \alpha T(f)$.

We define the operator norm on V

$$||T||_{\text{op}} = \sup\{||T(f)|| : f \in V, ||f|| \leq 1\}.$$

We say that T is bounded if $||T||_{\text{op}} < \infty$.

Theorem 2.1. If $T : V \to W$ is a linear map, then T is continuous if and only if T is bounded.

Definition 2.7. A linear map from a vector space V to its field is called a linear functional, and the normed space of all bounded linear functionals on V is called the dual space of V.

Definition 2.8. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $0 < p < \infty$. We define $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ to be the space of measurable functions $f : \Omega \to [-\infty, \infty]$ such that $||f||_p < \infty$, where $||f||_{L^p} = \left(\int_\Omega |f|^p d\mu\right)^{1/p}$ is the p-norm, and when $p = \infty$, we define the essential supremum by

$$||f||_\infty = \inf\{t > 0 : \mu(\{\omega \in \Omega : |f(x)| > t\}) = 0\}.$$

We also say that $\alpha \in \mathbb{R}$ is an essential bound for $f : \Omega \to [0, \infty]$ (measurable) if

$$\mu(f^{-1}(\alpha, \infty]) = \mu(\{\omega : f(\omega) > \alpha\}) = 0,$$

meaning that $f(\omega) \leq \alpha$ almost everywhere.

Note 2.1. Recall that we call real numbers p and q conjugate if

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Theorem 2.2 (Hölder's inequality)**.** Suppose $1 \leq p, q \leq \infty$ are conjugate indices, and $f, g : \Omega \to \mathbb{R}$ are $\mathcal{F}$-measurable. Then

$$||fg||_{L^1} \leq ||f||_{L^p} ||g||_{L^q}.$$

Proof. Consider the case where $1 < p < \infty$.
Step 1: Suppose $||f||_{L^p} = ||g||_{L^q} = 1$.
By Young's inequality (see (Axler, 2020, 196)),

$$\begin{aligned} |f(\omega)g(\omega)| &\leq \frac{|f(\omega)|^p}{p} + \frac{|g(\omega)|^q}{q}, \\ \iff \int |f(\omega)g(\omega)| d\mu &\leq \frac{1}{p}\int |f(\omega)|^p d\mu + \frac{1}{q}\int |g(\omega)|^q d\mu \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1. \end{aligned}$$

Step 2: Define $f_1 = \frac{f}{||f||_{L^p}}$ and $g_1 = \frac{g}{||g||_{L^q}}$.
Then

$$\left(\int |f_1|^p d\mu\right)^{1/p} = \left(\int \frac{|f|^p}{||f||_{L^p}^p} d\mu\right)^{1/p}$$

$$\begin{aligned}\iff \qquad ||f_1||_{L^p} &= \left[\frac{1}{\int |f|^p d\mu}\int |f|^p d\mu\right]^{1/p}\\ &= 1,\end{aligned}$$

and similarly, $||g_1||_{L^q} = 1$. This implies that

$$\begin{aligned} &\int |f_1(\omega)g_1(\omega)|d\mu \leq 1 && \text{(by step 2)}\\ \iff \qquad &\int \left|\frac{f(\omega)}{||f||_{L^p}}\frac{g(\omega)}{||g||_{L^q}}\right| d\mu \leq 1\\ \iff \qquad &\int |f(\omega)g(\omega)|d\mu \leq ||f||_{L^p}||g||_{L^q}.\end{aligned}$$

Next, we consider the case where $p = \infty$ and $q = 1$.
This time, we have

$$\begin{aligned} |f(\omega)g(\omega)| &\leq |f(\omega)||g(\omega)|\\ &\leq \operatorname{ess\,sup}|f(\omega)||g(\omega)|\\ &= ||f||_\infty |g(\omega)|,\end{aligned}$$
$$\iff ||fg||_{L^1} = \int |f(\omega)g(\omega)|d\mu \leq ||f||_\infty \int |g(\omega)|d\mu = ||f||_\infty ||g||_{L^1}.$$

□

Theorem 2.3 (Minkowski's Inequality)**.** Suppose $1 \leq p \leq \infty$, and $f, g \in \mathcal{L}^p(\Omega, \mathcal{F}, \mu)$. Then

$$||f+g||_{L^p} \leq ||f||_{L^p} + ||g||_{L^p}.$$

Proof. We may assume $\int(|f|+|g|)^p d\mu < \infty$. Suppose p and q are conjugate indices. Then, we can see that

$$\begin{aligned} (|f|+|g|)^p &= (|f|+|g|)(|f|+|g|)^{p-1}\\ &= |f|(|f|+|g|)^{p-1} + |g|(|f|+|g|)^{p-1} && \text{(distribute the first term)}\\ \iff \qquad \int(|f|+|g|)^p d\mu &= \int |f|(|f|+|g|)^{p-1}d\mu + \int |g|(|f|+|g|)^{p-1}d\mu\\ &\leq \left(\int |f|^p d\mu\right)^{1/p}\left(\int(|f|+|g|)^{q(p-1)}d\mu\right)^{1/q}\\ &\quad + \left(\int |g|^p d\mu\right)^{1/p}\left(\int(|f|+|g|)^{q(p-1)}d\mu\right)^{1/q} && \text{(Hölder's)}\\ &= \left(\int |f|^p d\mu\right)^{1/p}\left(\int(|f|+|g|)^{p}d\mu\right)^{1/q}\\ &\quad + \left(\int |g|^p d\mu\right)^{1/p}\left(\int(|f|+|g|)^{p}d\mu\right)^{1/q} && (q(p-1)=p)\end{aligned}$$

$$\iff \frac{\int(|f|+|g|)^p d\mu}{\left[\int(|f|+|g|)^p d\mu\right]^{1/q}} \leq \left(\int |f|^p d\mu\right)^{1/p} + \left(\int |g|^p d\mu\right)^{1/p}$$
$$\iff \left(\int(|f|+|g|)^p d\mu\right)^{1/p} \leq \left(\int |f|^p d\mu\right)^{1/p} + \left(\int |g|^p d\mu\right)^{1/p}.$$

□

We now have the tools to define a specific space of functions denoted by $L^p(\Omega, \mathcal{F}, \mu)$.

2.1 L^p Spaces

Let $\mathcal{A}$ be the collection of all functions $f : \Omega \to \mathbb{R}$. Note that $\mathcal{A}$ is a vector space with respect to addition and scalar multiplication. Let M be the subset of $\mathcal{A}$ that consists of all measurable functions with respect to the σ-algebra $\mathcal{F}$. We also note that M is a vector subspace of $\mathcal{A}$ since it's closed under addition and scalar multiplication, and the zero vector is the zero function.

Proposition 2.1. Define $\sim$ by $f \sim g$ if and only if $\mu(\{\omega : f(\omega) \neq g(\omega)\}) = 0$. Then $\sim$ is an equivalence relation on M.

Denote by $[M]_\mu$ the set of all equivalence classes. Let $\lambda \in \mathbb{R}$. Addition and scalar multiplication are well-defined:

$$[f] + [g] = [f + g],$$
$$\lambda[f] = [\lambda f],$$

which makes $[M]_\mu$ a vector space. With this, we define

$$L^p(\Omega, \mathcal{F}, \mu) = \{[f] : f \in \mathcal{L}^p(\Omega, \mathcal{F}, \mu) \cap M\},$$

where $1 \leq p \leq \infty$. In other words, $[f] \in L^p(\Omega, \mu)$ means that $[f]$ consists of all functions $g \in M$ such that $\mu\{\omega : f(\omega) \neq g(\omega)\} = 0$, which implies that

$$\int_\Omega |g|^p d\mu = \int_\Omega |f|^p d\mu < \infty.$$

This can be shown as follows:

Proof. Let $E = \{\omega : f(\omega) = g(\omega)\}$. Then E is measurable and $\mu(E^c) = 0$. Since E and E^c form a partition of Ω, this gives us that

$$\begin{aligned}\int_\Omega |g|^p d\mu &= \int_E |g|^p d\mu \int_{E^c} |g|^p d\mu \\ &= \int_E |f|^p d\mu \\ &= \int_E |f|^p d\mu + \int_{E^c} |f|^p d\mu\end{aligned}$$

$$= \int_\Omega |f|^p d\mu.$$

Thus, any $g \in [f]$ is such that $||f||_p = ||g||_p$. □

It is possible to show that $L^p(\Omega, \mu)$ is a vector subspace of $[M]_\mu$ and $||[f]||_p = ||f||_p$ is a norm on $L^p(\Omega, \mu)$, but we will omit the proof here.

Theorem 2.4. $L^p(\Omega, \mu)$ is a Banach space ($1 \leq p \leq \infty$).

Definition 2.9. An inner product on a vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ such that for $f, g \in V$,

- $\langle f, f \rangle \in [0, \infty)$,
- $\langle f, f \rangle = 0$ if and only if $f = 0$,
- $\langle f + g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$,
- for $\alpha \in \mathbb{R}$, $\langle \alpha f, g \rangle = \alpha \langle f, g \rangle$,
- $\langle f, g \rangle = \langle g, f \rangle$.

V is called an inner product space.

Definition 2.10. If V is an inner product space, we can define a norm on V by

$$||f|| = \sqrt{\langle f, f \rangle}.$$

Theorem 2.5 (Cauchy-Schwarz Inequality)**.** If $f, g \in V$, then

$$|\langle f, g \rangle| \leq ||f|| \; ||g||,$$

and equality holds if and only if f and g are scalar multiples of each other.

2.2 Hilbert Spaces and Projections

Definition 2.11. An inner product space that is a Banach space with its induced norm is a Hilbert space.

The space of functions that we will concern ourselves with will be $\mathcal{H}$, the Hilbert space of q-dimensional functions of random vectors with mean zero and finite variance, where the inner product will be defined by

$$\langle h_1, h_2 \rangle = \mathbb{E}[h_1^\top h_2],$$

for all $h_1, h_2 \in \mathcal{H}$.

Definition 2.12. Suppose $U \subseteq V$ are nonempty normed vector spaces and $f \in V$. We define the distance from f to the set U to be

$$\inf\{||f - g|| : g \in U\}.$$

Definition 2.13. Suppose V is an inner product space. We call $f, g \in V$ orthogonal if $\langle f, g \rangle = 0$.

Next, we wish to define the notion of an orthogonal projection.

Theorem 2.6. Let $\mathcal{H}$ be a Hilbert space and let $\mathcal{U}$ be a closed linear subspace. For any $h \in \mathcal{H}$, there exists a unique $u_0 \in \mathcal{U}$ such that

$$||h - u_0|| \leq ||h - u||,$$

for all $u \in \mathcal{U}$, and $h - u_0$ is orthogonal to the space $\mathcal{U}$, meaning that

$$\langle h - u_0, u \rangle = 0,$$

for all $u \in \mathcal{U}$. We call u_0 the orthogonal projection of h onto the space $\mathcal{U}$.

We will not prove this version of the result, but we will introduce a different version that directly relates to random vectors in the next section.

3 Some Tools from Asymptotic Statistics

We will always implicitly consider an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, $\mathcal{F}$ is the corresponding σ-algebra, and $\mathbb{P}$ is the probability measure. For some general results, we will use λ to refer to the Lebesgue measure on $\mathbb{R}$, and $\mu(\cdot)$ to refer to any arbitrary measure. In this section, we will use $||\cdot||_2$ to refer to the Euclidean norm, defined by

$$||x||_2 = \sqrt{x_1^2 + \cdots + x_p^2}$$

for any $x \in \mathbb{R}^p$. Once again, we will not prove nor use everything, but we have included some additional material for the sake of completeness. The results in this section can be found in Ferguson (2017), Jiang et al. (2010), Lehmann (1999), and Van der Vaart (2000).

3.1 Notions of Convergence

There are many different ways in which random vectors can converge. We will define the most important ones and showcase some results related to equivalences between various modes of convergence.

Definition 3.1. Let $X \in \mathbb{R}^p$ be a random vector such that $X = (X_1, \ldots, X_p)$. We define the cumulative distribution function (CDF) of X to be

$$F_X(x) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X \leq x).$$

Definition 3.2. A sequence of random vectors X_n converges in distribution to $X \in \mathbb{R}^p$, denoted $X_n \xrightarrow[n\to\infty]{\text{d}} X$, if $F_{X_n}(x)$ converges to $F_X(x)$ as $n \to \infty$ for all $x \in \mathbb{R}^p$ for which $F_X(\cdot)$ is continuous.

Definition 3.3. A sequence of random vectors X_n converges in probability to $X \in \mathbb{R}^p$, denoted $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, if $\mathbb{P}(||X_n - X||_2 > \epsilon) \to 0$ as $n \to \infty$.

Definition 3.4. Let $r \in [1, \infty)$. A sequence of random vectors X_n converges in r^{th} mean to X, denoted $X_n \xrightarrow[n\to\infty]{\text{r}} X$ if $\mathbb{E}||X_n - X||_2^r \to 0$ as $n \to \infty$.

Definition 3.5. A sequence of random vectors X_n converges almost surely (a.s.) to $X \in \mathbb{R}^p$, denoted $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$, if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.

Note 3.1. We will use LOTP to refer to the law of total probability.

These modes of convergence have varying levels of strengths. The following result shows that some modes of convergence directly imply other modes of convergence, which may be useful when proving more difficult results.

Theorem 3.1. (a) If $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$, then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

(b) If $X_n \xrightarrow[n\to\infty]{\text{r}} X$, then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

(c) If $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, then $X_n \xrightarrow[n\to\infty]{\text{d}} X$.

Proof. (a) We start by proving two short results that will be useful to us later:

Claim: *If* $\{C_n\}$ *is an increasing sequence* ($C_1 \leq C_2 \leq \cdots \leq C_n \leq C_{n+1} \leq \cdots$), *then* $\mathbb{P}(\cup_{n=1}^{\infty} C_n) = \lim_{n\to\infty} \mathbb{P}(C_n)$.

Set $C_0 = \varnothing$. Let

$$\begin{aligned} B_1 &= C_1, \\ B_2 &= C_2 \setminus C_1, \\ B_3 &= C_3 \setminus C_2, \\ &\vdots \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P}(\cup_{n=1}^{\infty} C_n) &= \mathbb{P}(\cup_{n=1}^{\infty} B_n) \\ &= \sum_{n=1}^{\infty} \mathbb{P}(B_n) && \text{(disjoint)} \\ &= \sum_{k=1}^{\infty} \mathbb{P}(B_k) && \text{(relabel the index)} \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(B_k) \end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(C_k \setminus C_{k-1}) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} [\mathbb{P}(C_k) - \mathbb{P}(C_{k-1})] \\
&= \lim_{n\to\infty} \mathbb{P}(C_n) - \mathbb{P}(C_0) && \text{(telescoping series)} \\
&= \lim_{n\to\infty} \mathbb{P}(C_n). && (\mathbb{P}(\varnothing) = 0)
\end{aligned}
$$

Claim: $\mathbb{P}(\cap_{n=1}^{\infty} C_n^c) = \lim_{n\to\infty} \mathbb{P}(C_n^c)$.
If $\{C_n\}$ is an increasing sequence, then the sequence of its complements will be decreasing. We can see that

$$
\begin{aligned}
\mathbb{P}(\cap_{n=1}^{\infty} C_n^c) &= 1 - \mathbb{P}(\cup_{n=1}^{\infty} C_n) \\
&= 1 - \lim_{n\to\infty} \mathbb{P}(C_n) && \text{(by the previous claim)} \\
&= \lim_{n\to\infty} (1 - \mathbb{P}(C_n)) \\
&= \lim_{n\to\infty} (1 - [1 - \mathbb{P}(C_n^c)]) \\
&= \lim_{n\to\infty} \mathbb{P}(C_n^c).
\end{aligned}
$$

Now, let $\epsilon > 0$. Let $A_n = \{||X_m - X||_2 > \epsilon\}$. Let $B_n = \cup_{m\geq n} A_m$. Then we can see that

$$
\begin{aligned}
B_1 &= \{||X_1 - X||_2 > \epsilon\} \cup \{||X_2 - X||_2 > \epsilon\} \cup \cdots \\
B_2 &= \{||X_2 - X||_2 > \epsilon\} \cup \cdots \\
&\vdots
\end{aligned}
$$

Thus, $\{B_n\}_{n\geq 1}$ is a decreasing sequence $B_1 \supset B_2 \supset B_3 \supset \cdots$, which means that $B_n \searrow \cap_{n=1}^{\infty} B_n$. We can see that

$$
\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(||X_n - X||_2 > \epsilon) &= \lim_{n\to\infty} \mathbb{P}(A_n) \\
&\leq \lim_{n\to\infty} \mathbb{P}(B_n) && \text{(since } A_n \subseteq B_n) \\
&= \mathbb{P}(\cap_{n=1}^{\infty} B_n) \\
&= \mathbb{P}(\varnothing) = 0.
\end{aligned}
$$

The last equality is due to the almost sure convergence of X_n to X, which tells us that the number of times $||X_n - X||_2 > \epsilon$ occurs is finite. This means that as $n \to \infty$, we eventually do not see this event occur anymore, so the intersection must be the empty set.

(b) Suppose $X_n \xrightarrow[n\to\infty]{\text{r}} X$. We wish to show that $\lim_{n\to\infty} \mathbb{P}(||X_n - X||_2 > \epsilon) = 0$. By premise, we know that $\lim_{n\to\infty} \mathbb{E}||X_n - X||_2^r = 0$. By Markov's inequality, $\forall \epsilon > 0$,

$$
\begin{aligned}
&\mathbb{E}||X_n - X||_2^r \geq \epsilon^r \mathbb{P}(||X_n - X||_2 \geq \epsilon), \\
\iff\quad &\frac{\mathbb{E}||X_n - X||_2^r}{\epsilon^r} \geq \mathbb{P}(||X_n - X||_2 \geq \epsilon).
\end{aligned}
$$

If we take the limit of both ends, we get

$$0 = \lim_{n\to\infty} \frac{\mathbb{E}||X_n - X||_2^r}{\epsilon^r} \geq \lim_{n\to\infty} \mathbb{P}(||X_n - X||_2 \geq \epsilon) \geq 0.$$

Therefore, X_n converges to X in probability.

(c) We wish to show that $\lim_{n\to\infty} F_n(x) = F(x)$. Let $\epsilon > 0$. On one hand,

$$\begin{aligned} F_{X_n}(x) &= \mathbb{P}(X_n \leq x) \\ &= \mathbb{P}([X_n \leq x, ||X_n - X||_2 > \epsilon] \cup [X_n \leq x, ||X_n - X||_2 < \epsilon]) && \text{(LOTP)} \\ &= \mathbb{P}(X_n \leq x, ||X_n - X||_2 > \epsilon) + \mathbb{P}(X_n \leq x, ||X_n - X||_2 < \epsilon) && \text{(disjoint)} \\ &= \mathbb{P}(X_n \leq x, ||X_n - X||_2 > \epsilon) + \mathbb{P}(X_n \leq x, X_n - \epsilon < X < X_n + \epsilon) \\ &\leq \mathbb{P}(||X_n - X||_2 > \epsilon) + \mathbb{P}(X < x + \epsilon) \\ &= \mathbb{P}(||X_n - X||_2 > \epsilon) + F_X(x + \epsilon). \end{aligned}$$

Thus, we have that

$$\begin{aligned} \limsup_{n\to\infty} F_{X_n}(x) &\leq \limsup_{n\to\infty} \mathbb{P}(||X_n - X||_2 > \epsilon) + F_X(x + \epsilon) \\ &= \lim_{n\to\infty} \mathbb{P}(||X_n - X||_2 > \epsilon) + F_X(x + \epsilon) = F_X(x + \epsilon). \end{aligned}$$

On the other hand,

$$\begin{aligned} 1 - F_{X_n}(x) &= \mathbb{P}(X_n > x) \\ &= \mathbb{P}([X_n > x, ||X_n - X||_2 > \epsilon] \cup [X_n > x, ||X_n - X||_2 < \epsilon]) \\ &= \mathbb{P}(X_n > x, ||X_n - X||_2 > \epsilon) + \mathbb{P}(X_n > x, ||X_n - X||_2 < \epsilon) \\ &\leq \mathbb{P}(X > x - \epsilon) + \mathbb{P}(||X_n - X||_2 > \epsilon) \\ &= 1 - F_X(x - \epsilon) + \mathbb{P}(||X_n - X||_2 > \epsilon). \end{aligned}$$

This implies that

$$\begin{aligned} &\liminf_{n\to\infty} [1 - F_{X_n}(x)] \leq 1 - F_X(x - \epsilon) + \liminf_{n\to\infty} \mathbb{P}(||X_n - X||_2 > \epsilon), \\ &\implies \liminf_{n\to\infty} F_{X_n}(x) \geq F_X(x - \epsilon). \end{aligned}$$

Therefore,

$$F_X(x - \epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x + \epsilon).$$

By continuity of $F_X(\cdot)$, we can let $\epsilon \to 0$ to obtain

$$\begin{aligned} & F_X(x) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x), \\ \therefore \quad & \lim_{n\to\infty} F_{X_n}(x) = \liminf_{n\to\infty} F_{X_n}(x) = \limsup_{n\to\infty} F_{X_n}(x) = F_X(x). \end{aligned}$$

□

Theorem 3.2. If $c \in \mathbb{R}^p$ and $X_n \xrightarrow[n\to\infty]{\mathrm{d}} c$, then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} c$.

Proof. Let $\epsilon > 0$. We begin by noting that

$$F_{X_n}(x) \to F(x) = \mathbb{P}(X \le x) = \mathbb{P}(c \le x) = \begin{cases} 1, & c \le x, \\ 0, & c > x. \end{cases}$$

Therefore,

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(||X_n - c||_2 > \epsilon) &= \lim_{n\to\infty} \mathbb{P}(x + \epsilon < X_n < c - \epsilon) \\ &= \lim_{n\to\infty} [\mathbb{P}(X_n < c - \epsilon) - \mathbb{P}(x + \epsilon < X_n)] \\ &= \lim_{n\to\infty} [F_{X_n}(c+\epsilon) - 1 + F_{X_n}(c - \epsilon)] \\ &= 1 - 1 + 0 = 0. \end{aligned}$$

□

Theorem 3.3 (Borel-Cantelli Lemma)**.** If $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$, then $\mathbb{P}(\limsup_{n\to\infty} E_n) = 0$.

Proof. By definition, $\limsup_{n\to\infty} E_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} E_k$. Let $A_n = \cup_{k=n}^{\infty} E_k$. Then we know that $\{A_n\}_{n=1}^{\infty}$ is decreasing since

$$\begin{aligned} A_1 &= E_1 \cup E_2 \cup \cdots, \\ A_2 &= E_2 \cup \cdots, \end{aligned}$$

which means that $A_1 \supset A_2 \supset A_3 \supset \cdots$. This implies that

$$\lim_{n\to\infty} A_n = \cap_{n=1}^{\infty} A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} E_k = \limsup_{n\to\infty} E_n.$$

Thus,

$$\begin{aligned} \mathbb{P}\left(\limsup_{n\to\infty} E_n\right) &= \mathbb{P}(\cap_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n) \\ &= \lim_{n\to\infty} \mathbb{P}(\cup_{k=n}^{\infty} E_k) \\ &\le \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(E_k) \\ &= 0. \end{aligned}$$

The last equality holds because we know that the sum is convergent, but we also know that as n gets larger, the sum gets smaller and smaller. This means that it must converge to zero. □

Theorem 3.4 (2$^{\text{nd}}$ Borel-Cantelli Lemma)**.** If $\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty$ and $\{E_n\}_{n=1}^{\infty}$ are independent, then $\mathbb{P}(\limsup_{n\to\infty} E_n) = 1$.

Proof. We see that

$$\begin{aligned}
\mathbb{P}(\limsup_{n\to\infty} E_n) &= \mathbb{P}(\cap_{n=1}^{\infty} \cup_{k\geq n} E_k) \\
&= 1 - \mathbb{P}(\cup_{n=1}^{\infty} \cap_{k\geq n} E_k^c) \\
&\leq 1 - \sum_{n=1}^{\infty} \mathbb{P}(\cap_{k\geq n} E_k^c) \\
&= 1 - \sum_{n=1}^{\infty} \prod_{k\geq n} \mathbb{P}(E_k^c) && \text{(indep.)} \\
&= 1 - \sum_{n=1}^{\infty} \prod_{k\geq n} [1 - \mathbb{P}(E_k^c)] \\
&= 1 - 0 && \left(\text{since } \sum_{n=1}^{\infty} \mathbb{P}(E_k) = \infty\right) \\
&= 1.
\end{aligned}$$

□

Note 3.2. We want to quickly define the concept of big "O" and little "o" since it will be useful to us later on. The notation $X_n = o_p(Y_n)$ means that

$$\frac{X_n}{Y_n} \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

and $X_n = O_p(Y_n)$ means that for all $\epsilon > 0$, there exists a constant M such that

$$\sup_n \mathbb{P}\left(\left|\left| \frac{X_n}{Y_n} \right|\right|_2 > M \right) < \epsilon.$$

3.2 Laws of Large Numbers and the Central Limit Theorem

The two arguably most important results to us (both of which may come in a variety of forms) are the law of large numbers and the central limit theorem. These are the backbone of asymptotics in statistics and will be incredibly useful to us later on when we examine asymptotic distributions.

Theorem 3.5 (Laws of Large Numbers)**.** Let $X, X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and $\mu = \mathbb{E}(X)$.

(a) Weak Law (WLLN)

$$\mathbb{E}||X||_2 < \infty \implies \bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu.$$

(b)

$$\mathbb{E}||X||_2^2 < \infty \implies \mathbb{E}||\bar{X}_n - \mu||_2^2 \to 0.$$

(c) Strong law

$$\bar{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \mu \iff \mathbb{E}||X||_2 < \infty.$$

Proof. (a) We will prove this using characteristic functions. Since μ is a constant vector, we have that $\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu \iff \bar{X}_n \xrightarrow[n\to\infty]{\text{d}} \mu$. By Lévy's continuity theorem, it suffices to show that $\varphi_{\bar{X}_n}(t) \to \varphi_\mu(t)$. Note that

$$\begin{aligned}
\varphi_{\bar{X}_n}(t) &= \mathbb{E}\left[e^{it^\top \bar{X}_n}\right] \\
&= \mathbb{E}\left[e^{i\frac{t^\top}{n}\sum_{j=1}^n X_j}\right] \\
&= \mathbb{E}\left[e^{i\frac{t^\top}{n}(X_1+\cdots+X_n)}\right] \\
&= \mathbb{E}\left[e^{i\frac{t^\top}{n}X_1}e^{i\frac{t^\top}{n}X_2}\cdots e^{i\frac{t^\top}{n}X_n}\right] \\
&= \mathbb{E}\left[e^{i\frac{t^\top}{n}X_1}\right]\mathbb{E}\left[e^{i\frac{t^\top}{n}X_2}\right]\cdots\mathbb{E}\left[e^{i\frac{t^\top}{n}X_n}\right] \quad (X_i \perp\!\!\!\perp X_j \implies f(X_i) \perp\!\!\!\perp f(X_j)) \\
&= \mathbb{E}\left[e^{i\frac{t^\top}{n}X_1}\right]^n \quad \text{(ident. dist.)} \\
&= [\varphi_{X_1}(t/n)]^n = [\varphi_X(t/n)]^n.
\end{aligned}$$

By properties of characteristic functions, φ' exists for all t and is continuous. Thus, we have that

$$\begin{aligned}
\varphi_X(t/n) &= \varphi_X(0) + \frac{\varphi_X'(0)}{1!}\frac{t}{n} + \cdots \quad \text{(Maclaurin Expansion)} \\
&= \mathbb{E}\left[e^{i(0)X}\right] + \left[\frac{\partial}{\partial(t/n)}\mathbb{E}\left[e^{iX^\top \frac{t}{n}}\right]\right]_{\frac{t}{n}=0}\frac{t}{n} + o(t/n) \\
&= 1 + \mathbb{E}\left[iX^\top e^{i\frac{t}{n}X}\right]_{\frac{t}{n}=0}\frac{t}{n} + o(t/n) \\
&= 1 + \mathbb{E}[iX^\top]\frac{t}{n} + o(t/n) \\
&= 1 + \mu^\top\frac{it}{n} + o(t/n), \\
\iff \qquad [\varphi_X(t/n)]^n &= \left[1 + \mu^\top\frac{it}{n} + o(t/n)\right]^n, \\
\iff \lim_{n\to\infty}[\varphi_X(t/n)]^n &= \lim_{n\to\infty}\left[1 + \frac{i\mu^\top t}{n} + o(t/n)\right]^n \\
&= e^{i\mu^\top t} \quad \left(\left[1+\frac{x}{n}\right]^n \overset{n\to\infty}{\to} e^x\right) \\
&= \varphi_\mu(t).
\end{aligned}$$

$$\therefore\ \varphi_{\bar{X}_n}(t) \to \varphi_\mu(t) \implies \bar{X}_n \xrightarrow[n\to\infty]{\mathrm{d}} \mu \implies \bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu.$$

(b)

$$\begin{aligned}
E||\bar{X}_n - \mu||_2^2 &= \mathbb{E}\left[(\bar{X}_n - \mu)^\top (\bar{X}_n - \mu)\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)^\top \left(\frac{1}{n}\sum_{j=1}^n X_j - \mu\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[(X_i - \mu)^\top (X_j - \mu)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[(X_i - \mathbb{E}(X_i))^\top (X_j - \mathbb{E}(X_j))\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathrm{tr}(\mathrm{Cov}(X_i, X_j)) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Cov}(X_i, X_i) && (\mathrm{Cov} = 0 \text{ when } i \neq j) \\
&= \frac{n}{n^2}\mathbb{V}(X) && (\text{ident. dist.}) \\
&= \frac{1}{n}\mathbb{V}(X) \xrightarrow[n\to\infty]{} 0. && (\mathbb{V} \text{ is finite since } \mathbb{E}||X||_2^2 < \infty)
\end{aligned}$$

(c) We will not prove the strong law here, but a detailed proof of it can be found in (Durrett, 2019, 76). It is worth noting that despite not being proven very often, the result is used quite frequently.

□

It is easy to notice that the strong and weak law, as stated above, appear to have the same set of conditions in order to hold. So, one may ask themselves: what exactly *is* the difference between the two? There is an interesting blog post about it by Tao (2008). In short, the weak law actually does hold under weaker conditions than $\mathbb{E}||X||_2 < \infty$, but it is simply easier to prove when we assume that to be the case. The strong law, however, does not hold under any weaker conditions.

Theorem 3.6 (Central Limit Theorem (CLT))**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ be random vectors such that $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \Sigma < \infty$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{\mathrm{d}} N(0, \Sigma).$$

Proof. We use Lévy's continuity theorem to prove this result. First, we see that

$$\sqrt{n}(\bar{X}_n - \mu) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)$$

$$\begin{aligned}
&= \frac{\sqrt{n}}{n}\left(\sum_{i=1}^{n} X_i - \frac{\mu}{n}\right)\\
&= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}(X_i - \mu)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu).
\end{aligned}$$

This implies that

$$\begin{aligned}
\varphi_{\sqrt{n}(\bar{X}_n-\mu)}(t) &= \mathbb{E}\left[e^{it^\top \sqrt{n}(\bar{X}_n-\mu)}\right]\\
&= \mathbb{E}\left[e^{\frac{it^\top}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)}\right]\\
&= \mathbb{E}\left[e^{\frac{it^\top}{\sqrt{n}}(X_1-\mu)}e^{\frac{it^\top}{\sqrt{n}}(X_2-\mu)}\cdots e^{\frac{it^\top}{\sqrt{n}}(X_n-\mu)}\right]\\
&= \prod_{i=1}^{n}\mathbb{E}\left[e^{\frac{it^\top}{\sqrt{n}}(X_i-\mu)}\right] && \text{(indep.)}\\
&= \mathbb{E}\left[e^{\frac{it^\top}{\sqrt{n}}(X_i-\mu)}\right]^n && \text{(ident. dist.)}\\
&= \left[\varphi_{X_1-\mu}\left(\frac{t}{\sqrt{n}}\right)\right]^n.
\end{aligned}$$

We can do a Maclaurin expansion of the characteristic function of $X_1 - \mu$ to get

$$\begin{aligned}
\varphi_{X_1-\mu}\left(\frac{t}{\sqrt{n}}\right) &= \varphi_{X_1-\mu}(0) + \varphi'(0)\frac{t}{\sqrt{n}} + \frac{\varphi''(0)}{2}\frac{t^2}{n} + \cdots\\
&= 1 + 0 - \frac{\mathbb{E}[(X_1-\mu)^2]}{2}\frac{t^2}{n} + \cdots\\
&= 1 - \frac{\Sigma t^2}{2n} + O\left(\frac{t^3}{n^{3/2}}\right),\\
\implies \lim_{n\to\infty}\left[1 - \frac{\Sigma t^2}{2n} + O\left(\frac{t^3}{n^{3/2}}\right)\right]^n &= e^{\frac{-\Sigma t^2}{2}} = \varphi_X(t),
\end{aligned}$$

where $X \sim N(0, \Sigma)$. By Lévy's continuity theorem, we have that

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{\text{d}} N(0, \Sigma).$$

□

3.3 Convergence of Transformations

We might not always get to deal with random vectors directly, but rather, with transformations of them. This is where results such as the continuous mapping theorem and Slutsky's come into play to make our lives easier.

Theorem 3.7 (Continuous Mapping Theorem)**.** If $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$ and $f : \mathbb{R}^p \to \mathbb{R}^m$ is such that $\mathbb{P}(X \in C(f)) = 1$, where $C(f)$ is the set of continuities of f, then

$$f(X_n) \xrightarrow[n\to\infty]{\mathrm{d}} f(X).$$

Proof. By the Portmanteau theorem (see (Van der Vaart, 2000, 6)), $f(X_n) \xrightarrow[n\to\infty]{\mathrm{d}} f(X)$ if and only if

$$\mathbb{E}[g(f(X_n))] \to \mathbb{E}[g(f(X))],$$

for all bounded and continuous functions $g : \mathbb{R}^m \to \mathbb{R}$. So, we let g be arbitrary and take $h = g \circ f$. We know that h is bounded and continuous on $C(f)$, so

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)], \\ \iff &\lim_{n\to\infty} \mathbb{E}[g(f(X_n))] = \mathbb{E}[g(f(X))]. \end{aligned}$$

□

Lemma 3.1 (Asymptotic Equivalence)**.** If $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$ and $(X_n - Y_n) \xrightarrow[n\to\infty]{\mathbb{P}} 0$, then

$$Y_n \xrightarrow[n\to\infty]{\mathrm{d}} X.$$

In this case, we call X_n and Y_n asymptotically equivalent.

Proof. Let $\epsilon > 0$ and $\delta > 0$ be given. By definition, we know that

$$\lim_{n\to\infty} P(||X_n - Y_n||_2 > \epsilon) = 0.$$

For all bounded (by some arbitrary constant M) and continuous functions g, we have that

$$\begin{aligned} &||\mathbb{E}[g(Y_n)] - \mathbb{E}[g(X)]||_2 \\ &= ||\mathbb{E}[g(Y_n)] - \mathbb{E}[g(X_n)] + \mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]||_2 \\ &\leq ||\mathbb{E}[g(Y_n)] - \mathbb{E}[g(X_n)]||_2 + ||\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]||_2 && \text{(triangle ineq.)} \\ &\leq \mathbb{E}||g(Y_n) - g(X_n)||_2 + \mathbb{E}||g(X_n) - g(X)||_2 && \text{(Jensen's ineq.)} \\ \\ &= \mathbb{E}\big\{||g(Y_n) - g(X_n)||_2[\mathbb{1}(||X_n - Y_n||_2 > \epsilon) \\ &\qquad + \mathbb{1}(||X_n - Y_n||_2 \leq \epsilon)]\big\} + ||\mathbb{E}[g(X_n) - g(X)]||_2 \\ \\ &= \mathbb{E}\big\{||g(Y_n) - g(X_n)||_2\mathbb{1}(||X_n - Y_n||_2 > \epsilon)\big\} \end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}\big\{||g(Y_n) - g(X_n)||_2 \mathbb{1}(||X_n - Y_n||_2 \le \epsilon)\big\} \\
&+ ||\mathbb{E}[g(X_n) - g(X)]||_2
\end{aligned}
$$

$$\le 2M\mathbb{E}[\mathbb{1}(||X_n - Y_n||_2 > \epsilon)] + \mathbb{E}[\delta] + ||\mathbb{E}[g(X_n) - g(X)]||_2 \quad (g \text{ is unif. continuous})$$

$$
\begin{aligned}
&= 2M\mathbb{P}(||X_n - Y_n||_2 > \epsilon) + \delta + ||\mathbb{E}[g(X_n) - g(X)]||_2 \\
&\underset{n\to\infty}{\longrightarrow} 2M \cdot 0 + \delta + 0 = \delta.
\end{aligned}
$$

□

Theorem 3.8 (Slutsky's Theorem)**.** If $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$ and $Y_n \xrightarrow[n\to\infty]{\mathrm{d}} c$, where c is a constant vector, then

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow[n\to\infty]{\mathrm{d}} \begin{pmatrix} X \\ c \end{pmatrix}.$$

Proof. Since Y_n converges to a constant in distribution, it equivalently must also converge to that same constant in probability. First, we observe that

$$
\begin{aligned}
\mathbb{P}\left(\left|\left|\begin{pmatrix} X_n \\ Y_n \end{pmatrix} - \begin{pmatrix} X_n \\ c \end{pmatrix}\right|\right|_2 > \epsilon\right) &= \mathbb{P}\left(\left|\left|\begin{pmatrix} 0 \\ Y_n - c \end{pmatrix}\right|\right|_2 > \epsilon\right) \\
&= \mathbb{P}\left(||Y_n - c||_2 > \epsilon\right) \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}
$$

Second, we can also see that by the Portmanteau theorem,

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)],$$

for any bounded and continuous g. We can take $g(x) = f(x, c)$, where c is kept fixed. This means that

$$\begin{pmatrix} X_n \\ c \end{pmatrix} \xrightarrow[n\to\infty]{\mathrm{d}} \begin{pmatrix} X \\ c \end{pmatrix}.$$

By asymptotic equivalence, we get that

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow[n\to\infty]{\mathrm{d}} \begin{pmatrix} X \\ c \end{pmatrix}.$$

□

We will now present and prove a result from Kosorok (2008). This is a less general version of Theorem 2.6 that we introduced earlier. This one directly pertains to random variables with an inner product defined with respect to the expected value.

Theorem 3.9 (Projection Theorem)**.** Let $\mathcal{S}$ be a linear space of real random variables with finite second moments. Then $\hat{S}$ is the *projection* of T onto $\mathcal{S}$ if and only if

(a) $\hat{S} \in \mathcal{S}$,

(b) $\mathbb{E}[(T - \hat{S})S] = 0$ for all $S \in \mathcal{S}$.

If S_1 and S_2 are both projections of T onto $\mathcal{S}$, then $S_1 \overset{\text{a.s.}}{=} S_2$. If $\mathcal{S}$ contains constants, then

$$\mathbb{E}[T] = \mathbb{E}[\hat{S}] \qquad \text{and} \qquad \operatorname{Cov}[T - \hat{S}, S] = 0$$

for all $S \in \mathcal{S}$.

Proof. We begin by proving the first part of the statement.
($\Longleftarrow$) Suppose $\hat{S} \in \mathcal{S}$ and $\mathbb{E}[(T - \hat{S})S] = 0 \ \forall S \in \mathcal{S}$. We are essentially assuming that $T - \hat{S}$ is orthogonal to every element of $\mathcal{S}$. We can see that

$$\begin{aligned}
\mathbb{E}[(T - \hat{S})] &= \mathbb{E}[(T - \hat{S} + \hat{S} - S)^2] && (\pm\, \hat{S}) \\
&= \mathbb{E}[(T - \hat{S})^2 + 2(T - \hat{S})(\hat{S} - S) + (\hat{S} - S)^2] \\
&= \mathbb{E}[(T - \hat{S})^2] + 2\mathbb{E}[(T - \hat{S})(\hat{S} - S)] + \mathbb{E}[(\hat{S} - S)^2] \\
&= \mathbb{E}[(T - \hat{S})^2] + 2\mathbb{E}[(T - \hat{S})\hat{S} - (T - \hat{S})S)] + \mathbb{E}[(\hat{S} - S)^2] \\
&= \mathbb{E}[(T - \hat{S})^2] + 2\big\{\mathbb{E}[(T - \hat{S})\hat{S}] - \mathbb{E}[(T - \hat{S})S)]\big\} + \mathbb{E}[(\hat{S} - S)^2] \\
&= \mathbb{E}[(T - \hat{S})^2] + \mathbb{E}[(\hat{S} - S)^2] && \text{(by assumption)} \\
&\geq \mathbb{E}[(T - \hat{S})^2].
\end{aligned}$$

Thus, $\hat{S}$ is the projection of T onto the space $\mathcal{S}$.
($\Longrightarrow$) Now, assume that $\hat{S}$ is the projection of T into $\mathcal{S}$. Then $\hat{S}$ must be in $\mathcal{S}$ by definition so condition (a) is satisfied. Let $\alpha \in \mathbb{R}, \alpha \neq 0, S \in \mathcal{S}$ and consider $\hat{S} + \alpha S$. By linearity, this is also an element of $\mathcal{S}$. We can see that

$$\begin{aligned}
&& ||T - (\hat{S} + \alpha S)|| \geq ||T - \hat{S}|| \geq 0, \\
\iff && \mathbb{E}[(T - (\hat{S} + \alpha S))^2] \geq \mathbb{E}[(T - \hat{S})^2].
\end{aligned}$$

Following some algebra, we see that

$$\begin{aligned}
\mathbb{E}[(T - (\hat{S} + \alpha S))^2] - \mathbb{E}[(T - \hat{S})^2] &= \mathbb{E}[((T - \hat{S}) + \alpha S)^2] - \mathbb{E}[(T - \hat{S})^2] \\
&= \mathbb{E}[(T - \hat{S})^2 - 2(T - \hat{S})\alpha S + \alpha^2 S^2] - \mathbb{E}[(T - \hat{S})^2] \\
&= -2\alpha\mathbb{E}[(T - \hat{S})S] + \alpha^2\mathbb{E}[S^2].
\end{aligned}$$

The left-hand side is non-negative by the previous equation. On the right-hand side, we have a polynomial in α, which is non-negative if and only if its discriminant is equal to zero. This requires

$$\Big[-2\mathbb{E}[(T - \hat{S})S]\Big]^2 = 0,$$

$$\iff \qquad \mathbb{E}[(T-\hat{S})S] = 0,$$

for all $S \in \mathcal{S}$, which means that condition (b) is also satisfied.
Now, suppose S_1 and S_2 are both projections of T onto $\mathcal{S}$. Then $S_1 - S_2 \in \mathcal{S}$ by linearity. We know that

$$\begin{aligned}
\mathbb{E}[(T-S_1)(S_1-S_2)] &= 0, \\
\mathbb{E}[(T-S_2)(S_1-S_2)] &= 0.
\end{aligned}$$

If we equate the two, we get that

$$\begin{aligned}
&& \mathbb{E}[(T-S_1)(S_1-S_2)] &= \mathbb{E}[(T-S_2)(S_1-S_2)], \\
\iff && \mathbb{E}[(T-S_1)(S_1-S_2)] - \mathbb{E}[(T-S_2)(S_1-S_2)] &= 0, \\
\iff && \mathbb{E}[(T-S_1-T+S_2)(S_1-S_2)] &= 0, \\
\iff && \mathbb{E}[(-S_1+S_2)(S_1-S_2)] &= 0, \\
\iff && \mathbb{E}[(S_1-S_2)^2] &= 0.
\end{aligned}$$

Thus, $S_1 \overset{\text{a.s.}}{=} S_2$.
Next, if $c \in \mathbb{R}$ is in $\mathcal{S}$, then we should be able to pick any c such that condition (b) still holds true. If we take $c = 1$, we have that

$$\mathbb{E}[(T-\hat{S})c] = \mathbb{E}[T-\hat{S}] = 0 \iff \mathbb{E}[T] = \mathbb{E}[\hat{S}].$$

Furthermore, we see that

$$\begin{aligned}
\mathrm{Cov}[T-\hat{S}, S] &= \mathbb{E}[(T-\hat{S})S] - \mathbb{E}[T-\hat{S}]\mathbb{E}[S] \\
&= -\mathbb{E}[T-\hat{S}]\mathbb{E}[S] && \text{(by (b))} \\
&= -(\mathbb{E}[T] - \mathbb{E}[\hat{S}])\mathbb{E}[S] \\
&= 0. && \text{(by above)}
\end{aligned}$$

□

4 Influence Functions

When seeking out an estimator for any parameter, we are typically interested in finding its asymptotic distribution. This is a task that may be very simple in the case of estimators that have straightforward forms such as the sample mean, but in many cases, things become complicated because a wide variety of estimators do not have “nice” forms. In our search for estimators, we may have to limit ourselves to smaller classes of estimators whose asymptotic behaviour we are able to determine more easily. This is where influence functions come in. We will see that they are a convenient structure when it comes to finding asymptotic distributions.

Note 4.1. We just want to bring attention to a slight abuse of notation. The previous section was the only section in which we used $\varphi(\cdot)$ to denote characteristic functions because that is the notation used by convention. However, in this section, we will also use $\varphi(\cdot)$, but this time, to denote influence functions. The two are not to be confused, but given the context, it is usually pretty clear which one is being used.

Let $\hat{\beta}_n$ be a measurable random function of independent, identically-distributed (i.i.d.) random vectors $Z_1, \ldots, Z_n$.

Definition 4.1. An estimator $\hat{\beta}_n$ of β is said to be *asymptotically linear* if there exists a random vector $\varphi(Z)$ such that $\mathbb{E}[\varphi(Z)] = 0$,

$$\sqrt{n}(\hat{\beta}_n - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(Z_i) + o_p(1),$$

and $\mathbb{E}[\varphi\varphi^\top]$ is finite and non-singular. We call $\varphi(Z)$ the *influence function* of the estimator $\hat{\beta}_n$, and $\varphi(Z_i)$ the i^{th} influence function of $\hat{\beta}_n$.

We can see that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(Z_i) + o_p(1) = \sqrt{n} \cdot \overline{\varphi(Z_n)} + o_p(1),$$

where $\overline{\varphi(Z_n)}$ is the notation we will use to denote $\frac{1}{n}\sum_{i=1}^{n} \varphi(Z_i)$, and more generally, we use the symbol $-$ overhead to denote any form of sample mean.

Rewriting our estimator in terms of its influence function, if it exists, is beneficial to us because we are now essentially dealing with some form of a sample mean of independent, identically-distributed random variables of which we know the limiting distribution by the central limit theorem. To illustrate this, consider the following example.

Example 4.1. Suppose $Z_1, \ldots, Z_n$ are independent and identically-distributed according to a $\mathcal{N}(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown. We recall that their maximum likelihood estimators are

$$\begin{aligned}
\hat{\mu}_n &= \frac{1}{n} \sum_{i=1}^{n} Z_i, \\
\hat{\sigma}^2_n &= \frac{1}{n} \sum_{i=1}^{n} (Z_i - \hat{\mu}_n)^2.
\end{aligned}$$

We can see that $\hat{\mu}_n$ is asymptotically linear for the parameter μ since

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_n - \mu) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right) \\
&= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}(Z_i - \mu) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \mu)
\end{aligned}$$

and

$$\mathbb{E}[Z_i - \mu] = \mathbb{E}[Z_i] - \mu = \mu - \mu = 0.$$

Thus, the i^{th} influence function of $\hat{\mu}_n$ is $\varphi(Z_i) = Z_i - \mu$. As for $\hat{\sigma}^2_n$, we have that

$$\sqrt{n}(\hat{\sigma}^2_n - \sigma^2) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \hat{\mu}_n)^2 - \sigma^2\right)$$

$$
\begin{aligned}
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu + \mu - \hat{\mu}_n)^2 - \sigma^2\right)\\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}[(Z_i - \mu)^2 + 2(Z_i - \mu)(\mu - \hat{\mu}_n) + (\mu - \hat{\mu}_n)^2] - \sigma^2\right)\\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)^2 + 2\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)(\mu - \hat{\mu}_n) + \frac{1}{n}\sum_{i=1}^{n}(\mu - \hat{\mu}_n)^2 - \sigma^2\right)\\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)^2 + 2(\mu - \hat{\mu}_n)\left(\frac{1}{n}\sum_{i=1}^{n}Z_i - \mu\right) + (\mu - \hat{\mu}_n)^2 - \sigma^2\right)\\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)^2 + 2(\mu - \hat{\mu}_n)\left(\hat{\mu}_n - \mu\right) + (\mu - \hat{\mu}_n)^2\right)\\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)^2 - 2(\mu - \hat{\mu}_n)^2 + (\mu - \hat{\mu}_n)^2 - \sigma^2\right)\\
&= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}[(Z_i - \mu)^2 - \sigma^2] - \sqrt{n}(\mu - \hat{\mu}_n)^2\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(Z_i - \mu)^2 - \sigma^2] - \sqrt{n}(\hat{\mu}_n - \mu)^2.
\end{aligned}
$$

Notice that as $n \to \infty$, $\sqrt{n}(\hat{\mu}_n - \mu)$ converges to a normal distribution by the central limit theorem, and $(\hat{\mu}_n - \mu)$ converges in probability to zero by the law of large numbers. Hence, by Slutsky's theorem, we have that $\sqrt{n}(\hat{\mu}_n - \mu)^2 = \sqrt{n}(\hat{\mu}_n - \mu) \cdot (\hat{\mu}_n - \mu) = o_p(1)$. We also note that

$$
\begin{aligned}
\mathbb{E}[(Z_i - \mu)^2 - \sigma^2] &= \mathbb{E}[(Z_i - \mu)^2] - \sigma^2\\
&= \mathbb{V}(Z_i - \mu) + \mathbb{E}[Z_i - \mu]^2 - \sigma^2\\
&= \mathbb{V}(Z_i) + (\mathbb{E}[Z_i] - \mu)^2 - \sigma^2\\
&= \sigma^2 - \sigma^2 = 0.
\end{aligned}
$$

Therefore, the i^{th} influence function of $\hat{\sigma}_n^2$ is $\varphi(Z_i) = (Z_i - \mu)^2 - \sigma^2$.

In general, we can see that asymptotically linear estimators are asymptotically normal since by the central limit theorem,

$$\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i) \xrightarrow[n\to\infty]{\text{d}} N\Big(0, \mathbb{E}[\varphi\varphi^\top]\Big),$$

and by Slutsky's theorem

$$\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i) + o_p(1) \xrightarrow[n\to\infty]{\text{d}} N\Big(0, \mathbb{E}[\varphi\varphi^\top]\Big),$$

which means that

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow[n\to\infty]{\text{d}} N\Big(0, \mathbb{E}[\varphi\varphi^\top]\Big).$$

Theorem 4.1. An asymptotically linear estimator has an almost surely unique influence function.

Proof. By contradiction. Suppose $\varphi(Z)$ and $\varphi^*(Z)$ are two different influence functions for $\hat{\beta}_n$, an asymptotically linear estimator of β. Then, we know that

$$\begin{aligned}\mathbb{E}[\varphi(Z)] &= 0,\\ \mathbb{E}[\varphi^*(Z)] &= 0.\end{aligned}$$

Notice that

$$\begin{aligned}\mathbb{V}[\varphi - \varphi^*] &= \mathbb{E}[\{(\varphi - \varphi^*) - \mathbb{E}[\varphi - \varphi^*]\}\{(\varphi - \varphi^*) - \mathbb{E}[\varphi - \varphi^*]\}^\top]\\ &= \mathbb{E}[\{(\varphi - \varphi^*) - (\mathbb{E}[\varphi] - \mathbb{E}[\varphi^*])\}\{(\varphi - \varphi^*) - (\mathbb{E}[\varphi] - \mathbb{E}[\varphi^*])\}^\top]\\ &= \mathbb{E}[(\varphi - \varphi^*)(\varphi - \varphi^*)^\top]\end{aligned}$$

Furthermore, on one hand, we have that by definition

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi(Z_i) + o_p(1),\\ \sqrt{n}(\hat{\beta}_n - \beta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi^*(Z_i) + o_p(1).\end{aligned}$$

If we equate the right-hand sides, we get

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi(Z_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi^*(Z_i) = o_p(1),\\ \iff \qquad &\frac{1}{\sqrt{n}}\sum_{i=1}^n [\varphi(Z_i) - \varphi^*(Z_i)] = o_p(1).\end{aligned}$$

On the other hand, by the central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n [\varphi(Z_i) - \varphi^*(Z_i)] \xrightarrow[n\to\infty]{\mathrm{d}} N\Big(0, \mathbb{E}[(\varphi - \varphi^*)(\varphi - \varphi^*)^\top]\Big).$$

However, since we have already established that $n^{-1/2}\sum_{i=1}^n [\varphi(Z_i) - \varphi^*(Z_i)]$ converges to zero in probability, the only way both results are true is if we have

$$\mathbb{E}[(\varphi - \varphi^*)(\varphi - \varphi^*)^\top] = 0,$$

which holds if and only if $\varphi = \varphi^*$ almost surely. □

Example 4.2 (OLS)**.** Suppose we have n observations (y_i, x_i), where $x_i \in \mathbb{R}^p$ and we assume a linear model of the form

$$y = X\beta + \epsilon$$

where $X \in \mathbb{R}^{n \times p}$, and $y, \epsilon \in \mathbb{R}^n$. Define the Ordinary Least Squares (OLS) estimator $\hat{\beta}_n$ of β as the solution to

$$\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} ||y - X\beta||_2^2.$$

We wish to find the influence function of the OLS estimator $\hat{\beta}_n$ of

$$\beta = \mathbb{E}[x_1 x_1^\top]^{-1} \mathbb{E}[x_1 y_1].$$

It can be shown that

$$\hat{\beta}_n = \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(\sum_{j=1}^n x_j y_j \right) = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{j=1}^n x_j y_j \right).$$

We can multiply by $\sqrt{n}$ on both sides to obtain

$$\sqrt{n}\hat{\beta}_n = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{j=1}^n x_j y_j \right).$$

Let $\hat{A} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top$. By the law of large numbers,

$$\hat{A} \xrightarrow[n \to \infty]{\mathbb{P}} A = \mathbb{E}[x_1 x_1^\top],$$

and since A is non-singular, the continuous mapping theorem implies

$$\hat{A}^{-1} \xrightarrow[n \to \infty]{\mathbb{P}} A^{-1}.$$

Thus, we have

$$\begin{aligned}
\sqrt{n}\hat{\beta}_n &= \frac{\sqrt{n}}{n} \sum_{j=1}^n \hat{A}^{-1} x_j y_j \\
&= \frac{\sqrt{n}}{n} \sum_{j=1}^n (\hat{A}^{-1} - A^{-1} + A^{-1}) x_j y_j \\
&= \frac{\sqrt{n}}{n} \sum_{j=1}^n A^{-1} x_j y_j + \frac{\sqrt{n}}{n} \sum_{j=1}^n (\hat{A}^{-1} - A^{-1}) x_j y_j
\end{aligned}$$

First, we want to show that the second term is $o_p(1)$. We notice that by the law of large numbers and the central limit theorem,

$$\begin{aligned}
(\hat{A}^{-1} - A^{-1}) &\xrightarrow[n \to \infty]{\mathbb{P}} 0, \\
\sqrt{n} \cdot \frac{1}{n} \sum_{j=1}^n x_j y_j &\xrightarrow[n \to \infty]{\mathrm{d}} N\Big(\mathbb{E}[x_1 y_1], \mathbb{V}[x_1 y_1] \Big).
\end{aligned}$$

Thus, by Slutsky's,

$$\frac{\sqrt{n}}{n} \sum_{j=1}^n (\hat{A}^{-1} - A^{-1}) x_j y_j = o_p(1).$$

This gives us

$$\begin{aligned}\sqrt{n}\hat{\beta}_n &= \frac{\sqrt{n}}{n}\sum_{j=1}^{n} A^{-1}x_j y_j + o_p(1),\\ \sqrt{n}(\hat{\beta}_n - \beta) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(A^{-1}x_j y_j - \beta) + o_p(1).\end{aligned}$$

Next, we see that

$$\begin{aligned}\mathbb{E}[A^{-1}x_j y_j - \beta] &= A^{-1}\mathbb{E}[x_j y_j] - \beta\\ &= \mathbb{E}[x_1 x_1^\top]^{-1}\mathbb{E}[x_1 y_1] - \mathbb{E}[x_1 x_1^\top]^{-1}\mathbb{E}[x_1 y_1]\\ &= 0.\end{aligned}$$

Hence, the j^{th} influence function of $\hat{\beta}_n$ is $\varphi(x_j, y_j) = A^{-1}x_j y_j - \beta$. We now also know the asymptotic distribution of $\hat{\beta}_n$ because

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow[n\to\infty]{\text{d}} N\Big(0, \mathbb{V}(A^{-1}x_1 y_1)\Big) = N\Big(0, A^{-1}\mathbb{V}(x_1 y_1)(A^{-1})^\top\Big).$$

4.1 Superefficiency, Regularity, and Hodges Estimator

Example 4.3 (Hodges')**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. The maximum likelihood estimator of θ is $\bar{X}_n$ and its Cramér–Rao lower bound is 1. Consider Hodges' estimator $\hat{\theta}_n$ defined by

$$\hat{\theta}_n = \begin{cases}\bar{X}_n, & |\bar{X}_n| > n^{-1/4}\\ 0, & |\bar{X}_n| \leq n^{-1/4}.\end{cases}$$

We wish to find the limiting distribution of $\hat{\theta}_n$. Let $Z \sim N(0,1)$ and $\Phi(\cdot)$ denote the cumulative distribution function of a standard normal random variable. Note that $Z + \sqrt{n}\theta \sim N(\sqrt{n}\theta, 1)$, which means that $n^{-1/2}(Z + \sqrt{n}\theta) \sim N(\theta, n^{-1})$ (this is the same distribution as $\bar{X}_n$). On one hand, we have that

$$\begin{aligned}\mathbb{P}(\hat{\theta}_n = 0) &= \mathbb{P}(|\bar{X}_n| \leq n^{-1/4})\\ &= \mathbb{P}\left(\left|\frac{1}{\sqrt{n}}(Z + \sqrt{n}\theta)\right| \leq n^{-1/4}\right)\\ &= \mathbb{P}\left(\frac{1}{\sqrt{n}}|Z + \sqrt{n}\theta| \leq n^{-1/4}\right)\\ &= \mathbb{P}\left(|Z + \sqrt{n}\theta| \leq n^{1/4}\right)\\ &= \mathbb{P}(Z + \sqrt{n}\theta \leq n^{1/4}, -Z - \sqrt{n}\theta \leq n^{1/4})\\ &= \mathbb{P}(-n^{1/4} - \sqrt{n}\theta \leq Z \leq n^{1/4} - \sqrt{n}\theta)\\ &= \Phi(n^{1/4} - \sqrt{n}\theta) - \Phi(-n^{1/4} - \sqrt{n}\theta).\end{aligned}$$

If $\theta \neq 0$, we get that, as $n \to \infty$,

$$\Phi(n^{1/4} - \sqrt{n}\theta) - \Phi(-n^{1/4} - \sqrt{n}\theta) \to \Phi(-\infty) - \Phi(-\infty) \to 0,$$

which means that

$$\mathbb{P}(\hat{\theta}_n = \bar{X}_n) = \mathbb{P}(|\bar{X}_n| > n^{-1/4}) \to 1.$$

Now, on the other hand, if $\theta = 0$, as $n \to \infty$, we get

$$\Phi(n^{1/4}) - \Phi(-n^{1/4}) \to \Phi(+\infty) - \Phi(-\infty) \to 1 - 0 = 1,$$

which means that

$$P(\hat{\theta}_n = 0) \to 1.$$

We observe that when $\theta \neq 0$,

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta) &= \sqrt{n}(\hat{\theta}_n - \bar{X}_n + \bar{X}_n - \theta) \\
&= \sqrt{n}(\hat{\theta}_n - \bar{X}_n) + \sqrt{n}(\bar{X}_n - \theta) \\
&\underset{n\to\infty}{\longrightarrow} 0 + N(0,1) = N(0,1), \qquad \text{(since } \hat{\theta}_n = \bar{X}_n \text{ with probability 1)}
\end{aligned}$$

and when $\theta = 0$, we have that

$$\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}\hat{\theta}_n \underset{n\to\infty}{\longrightarrow} 0. \qquad \text{(with probability 1)}$$

Thus, the asymptotic distribution of Hodges' estimator is

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \begin{cases} N(0,1), & \theta \neq 0 \\ 0, & \theta = 0. \end{cases}$$

This means that when $\theta = 0$, the asymptotic variance of the Hodges' estimator is 0, and when that is not the case, its asymptotic variance is 1, which is the same as the Cramér–Rao lower bound of $\bar{X}_n$. This is called *superefficiency.* Hodges' estimator is extremely efficient at zero. This might seem like a very good property to have, but we actually run into some big problems when we try to estimate at values that are very close to zero. We can take a look at the following plot where we simulated the mean squared error (MSE) of the Hodges' estimator for different values of θ and three different sample sizes (5, 50, and 500):

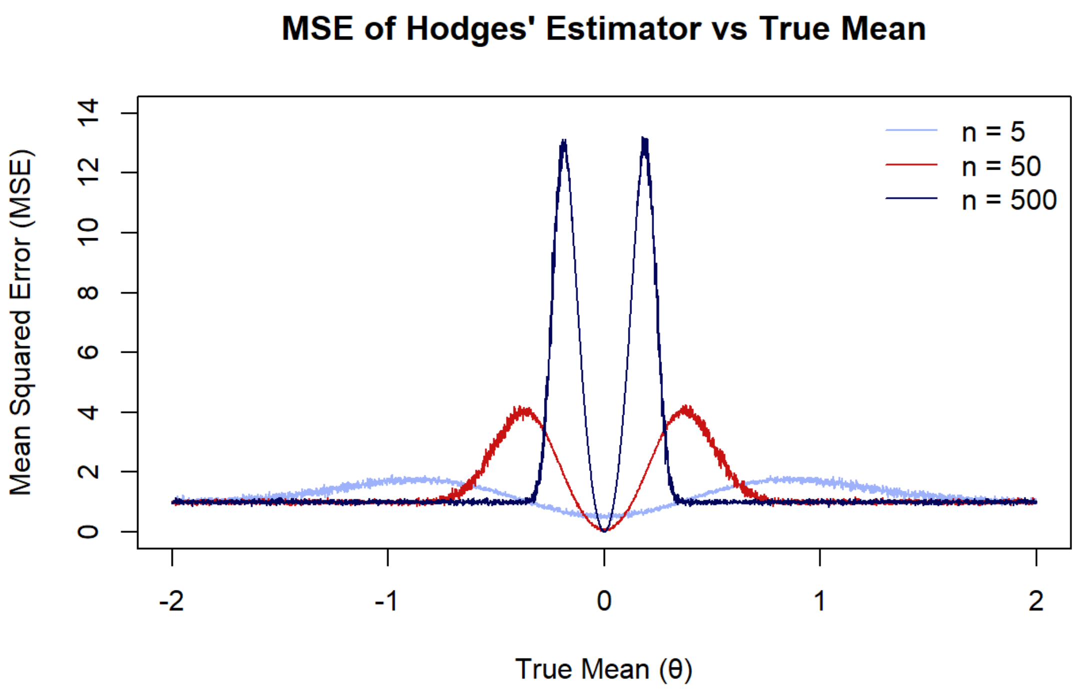

Figure 1: Hodges' estimator is not regular at zero.

We observe erratic behaviour when θ is shifted by very small values around zero. We want to avoid such bad local behaviour, so we turn to a class of estimators we call regular estimators.

Consider a local data-generating process (LDPG) where

$$Z_{1n}, \dots, Z_{nn} \overset{\text{iid}}{\sim} p(z, \theta_n),$$

such that $\sqrt{n}(\theta_n - \theta) \to \tau$, where τ is some constant vector. In other words, we are generating sequences of random vectors that take on the shape of a triangular array:

$$\begin{array}{llllll} Z_{11} & & & & \sim & p(z, \theta_1), \\ Z_{12} & Z_{22} & & & \sim & p(z, \theta_2), \\ \vdots & & \ddots & & & \vdots \\ Z_{1n} & \cdots & & Z_{nn} & \sim & p(z, \theta_n). \end{array}$$

Definition 4.2. An estimator $\hat{\theta}_n$ of θ is called *regular* if its limiting distribution is unaffected by small perturbations. In other words, if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\text{d}} d_\theta,$$

then for $\theta_n = \theta + \frac{h}{\sqrt{n}}$ and any arbitrary $h \in \mathbb{R}^p$,

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow[n\to\infty]{\text{d}} d_\theta.$$

Example 4.4 (Hodges', continued)**.** We can now show that Hodges' estimator is not regular by considering a slight perturbation of the form $\theta_n = \theta + \frac{h}{\sqrt{n}}$. In fact, we only want to consider the case where $\theta = 0$ because regularity is a property related to a point and not the entire function, meaning that Hodges' estimator is not regular at the point $\theta = 0$. Consider the perturbed parameter

$$\theta_n = \theta + \frac{h}{\sqrt{n}}.$$

We see that

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_n) &= \sqrt{n}\left(\hat{\theta}_n - \left(\theta + \frac{h}{\sqrt{n}}\right)\right)\\ &= \sqrt{n}\left(\hat{\theta}_n - \frac{h}{\sqrt{n}}\right)\\ &= \sqrt{n}\hat{\theta}_n - h \xrightarrow[n\to\infty]{} -h,\end{aligned}$$

which is not zero, so Hodges' estimator is not regular at $\theta = 0$.

All in all, we will want to restrict ourselves to the class of regular asymptotically linear estimators (RAL). Next up, we will take a look at another example of a large class of estimators called m-estimators, for which we can find influence functions.

4.2 m-Estimators

Definition 4.3. Let $m(Z,\theta)$ be a p-dimensional function such that

$$\begin{aligned}\mathbb{E}_\theta[m(Z,\theta)] &= 0_{p\times 1},\\ \mathbb{E}_\theta[m(Z,\theta)^\top m(Z,\theta)] &< \infty,\end{aligned}$$

and $\mathbb{E}_\theta[m(Z,\theta)m(Z,\theta)^\top]$ is positive definite for all $\theta \in \Theta$. An m-estimator is any solution to

$$\sum_{i=1}^n m(Z_i, \hat{\theta}_n) = 0.$$

To show that the m-estimator $\hat{\theta}_n$ is consistent, it is enough to assume that $\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial \theta^\top}\right]$ is non-singular and that

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial m(Z_i,\theta)}{\partial \theta^\top} \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}_{\theta_0}\left[\frac{\partial m(Z,\theta)}{\partial \theta^\top}\right]$$

uniformly in θ in a neighbourhood of the truth, θ_0. We will not prove that result here, but a complete proof of it can be found in Van der Vaart (2000). We will, however, need to use this result to find the influence function of $\hat{\theta}_n$. By the Mean Value Theorem (MVT), there exists θ_n^* between $\hat{\theta}_n$ and θ_0 such that

$$\sum_{i=1}^n m(Z_i, \hat{\theta}_n) = \sum_{i=1}^n m(Z_i, \theta_0) + \left\{\sum_{i=1}^n \frac{\partial m(Z_i, \theta_n^*)}{\partial \theta^\top}\right\}(\hat{\theta}_n - \theta_0),$$

where $\sum_{i=1}^{n} \frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^\top} \in \mathbb{R}^{p\times p}$. We know that

$$0 = \sum_{i=1}^{n} m(Z_i, \hat\theta_n),$$

so we get

$$\begin{aligned}
0 &= \sum_{i=1}^{n} m(Z_i,\theta_0) + \left\{\sum_{i=1}^{n} \frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^\top}\right\}(\hat\theta_n - \theta_0),\\
\iff \qquad (\hat\theta_n - \theta_0) &= -\left\{\sum_{i=1}^{n} \frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^\top}\right\}^{-1}\sum_{i=1}^{n} m(Z_i,\theta_0),\\
\iff \sqrt{n}(\hat\theta_n - \theta_0) &= -\sqrt{n}\left\{\sum_{i=1}^{n} \frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^\top}\right\}^{-1}\sum_{i=1}^{n} m(Z_i,\theta_0),\\
\iff \sqrt{n}(\hat\theta_n - \theta_0) &= -\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^\top}\right\}^{-1} n^{-1/2}\sum_{i=1}^{n} m(Z_i,\theta_0) \quad \left(\text{notice that } \sqrt{n} = \frac{n}{\sqrt{n}}\right).
\end{aligned}$$

By our previous assumptions, we know that

$$\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^\top}\right\}^{-1} \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1}.$$

This means that

$$-\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^\top}\right\}^{-1} n^{-1/2}\sum_{i=1}^{n} m(Z_i,\theta_0) + \mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1} n^{-1/2}\sum_{i=1}^{n} m(Z_i,\theta_0) = o_p(1),$$

which implies that

$$\begin{aligned}
-\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(Z_i,\theta_n^*)}{\partial\theta^\top}\right\}^{-1} n^{-1/2}\sum_{i=1}^{n} m(Z_i,\theta_0) &= -\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1} n^{-1/2}\sum_{i=1}^{n} m(Z_i,\theta_0) + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}\left\{-\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1} m(Z_i,\theta_0)\right\} + o_p(1).
\end{aligned}$$

And since

$$\mathbb{E}\left[-\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1} m(Z_i,\theta_0)\right] = -\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1}\mathbb{E}\left[m(Z_i,\theta_0)\right] = 0,$$

the i^{th} influence function of $\hat\theta_n$ is

$$\varphi(Z_i) = -\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1} m(Z_i,\theta_0).$$

Note that

$$\mathbb{V}[m(Z_i,\theta_0)] = E[m(Z_i,\theta_0)m(Z_i,\theta_0)^\top] - E[m(Z_i,\theta_0)]\mathbb{E}[m(Z_i,\theta_0)^\top]$$

$$= E[m(Z_i,\theta_0)m(Z_i,\theta_0)^\top]. \qquad \text{(by definition)}$$

Based on the properties of influence functions, we know that

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{\mathrm{d}} N &\left(0, \mathbb{V}\left\{-\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1} m(Z_i,\theta_0)\right\}\right) \\
&= N\left(0, \mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1} \mathbb{V}\left\{m(Z_i,\theta_0)\right\} \left(\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1}\right)^\top\right) \\
&= N\left(0, \mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1} \mathbb{E}[m(Z_i,\theta_0)m(Z_i,\theta_0)^\top] \left(\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]^{-1}\right)^\top\right).
\end{aligned}$$

This result cannot be used directly: unknown expected values must be estimated. Given our previous assumptions, we can consistently estimate $\mathbb{E}\left[\frac{\partial m(Z,\theta_0)}{\partial\theta^\top}\right]$ by

$$\hat{E} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(Z_i,\hat{\theta}_n)}{\partial\theta^\top},$$

and by the law of large numbers, we can consistently estimate $\mathbb{E}[m(Z_i,\theta_0)m(Z_i,\theta_0)^\top]$ by

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\hat{\theta}_n)m(Z_i,\hat{\theta}_n)^\top.$$

This gives us the so-called “sandwich” variance estimator, $\hat{E}^{-1}\hat{V}[\hat{E}^{-1}]^\top$, for the asymptotic variance of $\hat{\theta}_n$.

Now, suppose that the parameter space can be partitioned into

$$\theta_{p\times 1} = \begin{bmatrix}\beta\\ \eta\end{bmatrix} \qquad \text{and} \qquad \hat{\theta}_n = \begin{bmatrix}\hat{\beta}_n\\ \hat{\eta}_n\end{bmatrix},$$

where β and η are respectively, q-dimensional and r-dimensional. In this case, the influence function of $\hat{\beta}_n$ is made-up of the first q elements of the p-dimensional influence function of $\hat{\theta}_n$.

4.3 Contiguity

We now want to work our way towards proving a very important result about influence functions. In order to do that, we first need to define the concept of contiguity.

Definition 4.4. Let $\{\mathbb{P}_{1n}\}_{n\geq 1}, \{\mathbb{P}_{0n}\}_{n\geq 1}$ be sequences of probability measures with respective densities $\{p_{1n}\}_{n\geq 1}$ and $\{p_{0n}\}_{n\geq 1}$ with respect to a common dominating σ-finite measure μ. We say that the sequence $\mathbb{P}_{1n}$ is contiguous to $\mathbb{P}_{0n}$ if, for any sequence of events $\{A_n\}$, as $n \to \infty$,

$$\mathbb{P}_{0n}(A_n) \to 0 \implies \mathbb{P}_{1n}(A_n) \to 0.$$

The concept of contiguity is similar to the idea of dominating measures, but in the asymptotic sense. In some ways, we could use the parallel idea that "$\mathbb{P}_{0n}$ asymptotically dominates $\mathbb{P}_{1n}$".

Lemma 4.1 (LeCam)**.** Let $\{V_n\}_{n\geq 1}$ be a sequence of random vectors. If

$$\log\left\{\frac{p_{1n}(V_n)}{p_{0n}(V_n)}\right\} \overset{\mathrm{d}(\mathbb{P}_{0n})}{\rightarrow} N\left(-\frac{\sigma^2}{2}, \sigma^2\right),$$

then $\mathbb{P}_{1n}$ is contiguous to $\mathbb{P}_{0n}$.

We omit the proof of the previous result, and refer the reader to (Van der Vaart, 2000, 88). We will use it to prove the following lemma:

Lemma 4.2. The LDGP is contiguous to the true DGP.

Proof. Let $V_n = (Z_{1n}, \ldots, Z_{nn})$, where the Z_{ij}'s are i.i.d. random vectors. The joint distributions of the true data-generating process and the LDGP are given by, respectively,

$$p_{0n}(v_n) = \prod_{i=1}^{n} p(z_{in}, \theta_0) \qquad \text{and} \qquad p_{1n}(v_n) = \prod_{i=1}^{n} p(z_{in}, \theta_n),$$

where we assume that $\sqrt{n}(\theta_n - \theta_0) \rightarrow \tau$, some p-dimensional constant vector. Let

$$\begin{aligned}\log(L_n(V_n)) &= \log\left(\frac{p_{1n}(V_n)}{p_{0n}(V_n)}\right)\\ &= \log\left(\prod_{i=1}^{n}\frac{p(Z_{in}, \theta_n)}{p(Z_{in}, \theta_0)}\right)\\ &= \sum_{i=1}^{n}\left[\log p(Z_{in}, \theta_n) - \log p(Z_{in}, \theta_0)\right].\end{aligned}$$

We also define the score function and the second derivative of the log-likelihood function

$$\begin{aligned}S_\theta(Z, \theta_0)_{p\times 1} &= \frac{\partial}{\partial\theta}\log p(Z, \theta_0),\\ S_{\theta\theta}(Z, \theta_0)_{p\times p} &= \frac{\partial^2}{\partial\theta\partial\theta^\top}\log p(Z, \theta_0).\end{aligned}$$

We can use a Taylor series expansion of the first term around θ_0 to obtain

$$\begin{aligned}&\sum_{i=1}^{n}\left[\log p(Z_{in}, \theta_n) - \log p(Z_{in}, \theta_0)\right]\\ &= \sum_{i=1}^{n}\Bigg[\log p(Z_{in}, \theta_0) + (\theta_n - \theta_0)^\top\frac{\partial}{\partial\theta}\log p(Z_{in}, \theta_0) \qquad (\theta_n^* \text{ is some value between } \theta_n \text{ and } \theta_0)\end{aligned}$$

$$
\begin{aligned}
&\qquad + \frac{(\theta_n - \theta_0)^\top}{2} \frac{\partial^2}{\partial\theta\partial\theta^\top} \log p(Z_{in}, \theta_n^*)(\theta_n - \theta_0) - \log p(Z_{in}, \theta_0)\Bigg] \\
&= \sum_{i=1}^n \Bigg[(\theta_n - \theta_0)^\top \frac{\partial}{\partial\theta} \log p(Z_{in}, \theta_0) + \frac{(\theta_n - \theta_0)^\top}{2} \frac{\partial^2}{\partial\theta\partial\theta^\top} \log p(Z_{in}, \theta_n^*)(\theta_n - \theta_0)\Bigg] \\
&= \sum_{i=1}^n \Bigg[(\theta_n - \theta_0)^\top S_\theta(Z_{in}, \theta_0) + \frac{(\theta_n - \theta_0)^\top}{2} S_{\theta\theta}(Z_{in}, \theta_n^*)(\theta_n - \theta_0)\Bigg] \\
&= (\theta_n - \theta_0)^\top \sum_{i=1}^n S_\theta(Z_{in}, \theta_0) + \frac{(\theta_n - \theta_0)^\top}{2} \sum_{i=1}^n S_{\theta\theta}(Z_{in}, \theta_n^*)(\theta_n - \theta_0) \\
&= \sqrt{n}(\theta_n - \theta_0)^\top \left[\sqrt{n}\, \frac{1}{n} \sum_{i=1}^n S_\theta(Z_{in}, \theta_0)\right] \\
&\qquad + \frac{\sqrt{n}(\theta_n - \theta_0)^\top}{2} \left[\frac{1}{n} \sum_{i=1}^n S_{\theta\theta}(Z_{in}, \theta_n^*)\right] \sqrt{n}(\theta_n - \theta_0), \qquad \left(\text{multiply by } \frac{\sqrt{n}}{\sqrt{n}}\right)
\end{aligned}
$$

Note that the term containing θ_n^* is the Lagrange remainder of our expansion. We can now deal with each part of the expression above separately.

First, we notice that under $\mathbb{P}_{0n}$, the $S_\theta(Z_{in}, \theta_0)$ are i.i.d. random vectors with mean zero and variance equal to the Fisher information matrix,

$$\mathcal{I}(\theta_0) = \mathbb{E}\left[(S_\theta(Z_{in}, \theta_0))^2\right] = -\mathbb{E}[S_{\theta\theta}(Z_{in}, \theta_0)].$$

Since we are taking some sort of sample mean, we can use the CLT to infer that

$$\sqrt{n}\, \frac{1}{n} \sum_{i=1}^n S_\theta(Z_{in}, \theta_0) \xrightarrow[n\to\infty]{\mathrm{d}} N(0, \mathcal{I}(\theta_0)).$$

Second, we recall our initial assumption that $\sqrt{n}(\theta_n - \theta_0) \to \tau$.

Third, since $\theta_n^* \xrightarrow[n\to\infty]{\mathbb{P}} \theta_0$, and we assume S_θ and $S_{\theta\theta}$ are continuous in a neighbourhood of θ_0, we can use the continuous mapping theorem to infer that

$$\frac{1}{n} \sum_{i=1}^n S_{\theta\theta}(Z_{in}, \theta_n^*) \xrightarrow[n\to\infty]{\mathbb{P}} \frac{1}{n} \sum_{i=1}^n S_{\theta\theta}(Z_{in}, \theta_0).$$

By Slutsky's,

$$\frac{1}{n} \sum_{i=1}^n \Big[S_{\theta\theta}(Z_{in}, \theta_n^*) - S_{\theta\theta}(Z_{in}, \theta_0)\Big] \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

By the WLLN,

$$\frac{1}{n} \sum_{i=1}^n S_{\theta\theta}(Z_{in}, \theta_0) \xrightarrow[n\to\infty]{\mathbb{P}} -\mathcal{I}(\theta_0),$$

since by definition $-\mathbb{E}[S_{\theta\theta}(Z_{in}, \theta_0)] = \mathcal{I}(\theta_0)$, which also means that

$$\frac{1}{n} \sum_{i=1}^n S_{\theta\theta}(Z_{in}, \theta_n^*) \xrightarrow[n\to\infty]{\mathbb{P}} -\mathcal{I}(\theta_0).$$

We can now go back to our original expression to conclude that by Slutsky's,

$$\sqrt{n}(\theta_n-\theta_0)^\top\left[\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}S_\theta(Z_{in},\theta_0)\right]+\frac{\sqrt{n}(\theta_n-\theta_0)^\top}{2}\left[\frac{1}{n}\sum_{i=1}^{n}S_{\theta\theta}(Z_{in},\theta_n^*)\right]\sqrt{n}(\theta_n-\theta_0)$$

$$\begin{aligned}\xrightarrow{d_{\mathbb{P}_{0n}}}\tau^\top N\big(0,\mathcal{I}(\theta_0)\big)-\frac{\tau^\top\mathcal{I}(\theta_0)\tau}{2}&=N\big(0,\tau^\top\mathcal{I}(\theta_0)\tau\big)-\frac{\tau^\top\mathcal{I}(\theta_0)\tau}{2}\\&=N\left(-\frac{\tau^\top\mathcal{I}(\theta_0)\tau}{2},\tau^\top\mathcal{I}(\theta_0)\tau\right).\end{aligned}$$

By LeCam's lemma, $\mathbb{P}_{1n}$ is contiguous to $\mathbb{P}_{0n}$. □

Now, the next result is an important theorem that will allow us to, later on, take advantage of the geometry of influence functions.

Theorem 4.2. Let the parameter of interest $\beta(\theta)$ be a q-dimensional function of the p-dimensional parameter θ $(q < p)$ such that the matrix of partial derivatives

$$\Gamma(\theta)_{q\times p}=\frac{\partial\beta(\theta)}{\partial\theta^\top},$$

exists, has rank q, and is continuous in θ in a neighbourhood of θ_0. Let $\hat{\beta}_n$ be an asymptotically linear estimator with influence function $\varphi(Z)$ such that $\mathbb{E}_\theta[\varphi^\top\varphi]$ exists and is continuous in θ in a neighbourhood of θ_0.
If $\hat{\beta}_n$ is regular, then

$$\mathbb{E}\big[\varphi(Z)S_\theta^\top(Z,\theta_0)\big]=\Gamma(\theta_0).$$

Proof. Consider a true DGP $\{p_{0n}\}_{n\geq1}$ where

$$p_{0n}(v_n)=\prod_{i=1}^{n}p(z_{in},\theta_0),$$

and a LDGP $\{p_{1n}\}_{n\geq1}$ where

$$p_{1n}(v_n)=\prod_{i=1}^{n}p(z_{in},\theta_n),$$

where we assume that $\sqrt{n}(\theta_n-\theta_0)\to\tau$. Since $\hat{\beta}_n$ is asymptotically linear, we can rewrite it in terms of its influence function

$$\sqrt{n}(\hat{\beta}_n-\beta(\theta_0))=n^{-1/2}\sum_{i=1}^{n}\varphi(Z_{in})+o_{\mathbb{P}_{0n}}(1).$$

We have previously proven that $\mathbb{P}_{1n}$ is contiguous to $\mathbb{P}_{0n}$, so a term that is $o_{\mathbb{P}_{0n}}(1)$ is also $o_{\mathbb{P}_{1n}}(1)$. Thus, we have that

$$\sqrt{n}(\hat{\beta}_n-\beta(\theta_0))=n^{-1/2}\sum_{i=1}^{n}\varphi(Z_{in})+o_{\mathbb{P}_{1n}}(1).$$

If we add and subtract $\sqrt{n}\beta(\theta_n)$ on the left-hand side, and $\mathbb{E}_{\theta_n}[\varphi(Z)]$ on the right-hand side (RHS), we get

$$
\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \beta(\theta_n)) - \sqrt{n}\beta(\theta_0) + \sqrt{n}\beta(\theta_n) &= n^{-1/2}\sum_{i=1}^{n}\Big[\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)] + \mathbb{E}_{\theta_n}[\varphi(Z)]\Big] + o_{\mathbb{P}_{1n}}(1), \\
\underbrace{\sqrt{n}(\hat{\beta}_n - \beta(\theta_n))}_{\text{(LHS)}} &= n^{-1/2}\sum_{i=1}^{n}\Big[\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]\Big] + n^{-1/2}\ n\ \mathbb{E}_{\theta_n}[\varphi(Z)] \\
&\quad + \sqrt{n}\beta(\theta_0) - \sqrt{n}\beta(\theta_n) + o_{\mathbb{P}_{1n}}(1) \\
&= \underbrace{n^{-1/2}\sum_{i=1}^{n}\Big[\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]\Big]}_{\text{(a)}} + \overbrace{\sqrt{n}\mathbb{E}_{\theta_n}[\varphi(Z)]}^{\text{(b)}} \\
&\quad - \underbrace{\sqrt{n}(\beta(\theta_n) - \beta(\theta_0))}_{\text{(c)}} + o_{\mathbb{P}_{1n}}(1).
\end{aligned}
$$

We can take advantage of the regularity of $\hat{\beta}_n$ to infer that under $\mathbb{P}_{1n}$,

$$
\sqrt{n}(\hat{\beta}_n - \beta(\theta_n)) \xrightarrow[n\to\infty]{\text{d}} N\big(0, \mathbb{E}[\varphi\varphi^\top]\big)
$$

by asymptotic properties of influence functions. Now, by the equality above, the RHS should also converge to the same distribution. We can deal with each term $(\cdot)$ separately.

(a) As functions of i.i.d. random vectors, $\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]$ are also i.i.d. random vectors with

$$
\mathbb{E}_{\theta_n}\big[\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]\big] = \mathbb{E}_{\theta_n}[\varphi(Z_{in})] - \mathbb{E}_{\theta_n}[\varphi(Z)] = 0,
$$

since Z is simply one of the Z_{in}'s, and

$$
\begin{aligned}
\mathbb{V}_{\theta_n}\big[\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]\big] &= \mathbb{V}_{\theta_n}\big[\varphi(Z_{in})\big] \\
&= \mathbb{E}_{\theta_n}[\varphi\varphi^\top] - \mathbb{E}_{\theta_n}[\varphi]\mathbb{E}_{\theta_n}[\varphi^\top] \\
&\qquad\qquad \downarrow \\
&\quad\ \mathbb{E}_{\theta_0}[\varphi\varphi^\top] - \mathbb{E}_{\theta_0}[\varphi]\mathbb{E}_{\theta_0}[\varphi^\top] = \mathbb{E}_{\theta_0}[\varphi\varphi^\top].
\end{aligned}
$$

By the CLT, under $\mathbb{P}_{1n}$,

$$
\sqrt{n}\ \frac{1}{n}\sum_{i=1}^{n}\big[\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]\big] \xrightarrow[n\to\infty]{\text{d}} N\big(0, \mathbb{E}_{\theta_0}[\varphi\varphi^\top]\big).
$$

(b) We can expand this term using the definition of expected value as follows:

$$
\begin{aligned}
\sqrt{n}\mathbb{E}_{\theta_n}[\varphi(Z)] &= \sqrt{n}\int \varphi(z)p(z,\theta_n)d\lambda(z) \qquad\qquad \text{(apply the MVT on } p(z,\theta_n)) \\
&= \sqrt{n}\int \varphi(z)p(z,\theta_0)d\lambda(z) + \sqrt{n}\int \varphi(z)\left[\frac{\partial p(z,\theta_n^*)}{\partial\theta}\right]^\top (\theta_n - \theta_0)d\lambda(z)
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{n}\mathbb{E}_{\theta_0}[\varphi(Z)] + \sqrt{n}\int \varphi(z)\left[\frac{\partial p(z,\theta_n^*)}{\partial\theta}\right]^\top (\theta_n - \theta_0)d\lambda(z) \\
&= \sqrt{n}\int \varphi(z)\left[\frac{\partial p(z,\theta_n^*)}{\partial\theta}\right]^\top (\theta_n - \theta_0)d\lambda(z) \qquad \left(\text{multiply by } \frac{p(z,\theta_n^*)}{p(z,\theta_n^*)}\right) \\
&= \sqrt{n}\int \varphi(z)\left[\frac{\partial p(z,\theta_n^*)}{\partial\theta}\cdot\frac{1}{p(z,\theta_n^*)}\right]^\top p(z,\theta_n^*)(\theta_n - \theta_0)d\lambda(z) \\
&= \sqrt{n}\int \varphi(z)\left[\frac{\partial}{\partial\theta}\log p(z,\theta_n^*)\right]^\top p(z,\theta_n^*)(\theta_n - \theta_0)d\lambda(z) \\
&= \mathbb{E}_{\theta_n^*}[\varphi(Z)S_\theta^\top(Z,\theta_0)]\sqrt{n}(\theta_n - \theta_0) \\
&\underset{n\to\infty}{\xrightarrow{\text{d}}} \mathbb{E}_{\theta_0}[\varphi(Z)S_\theta^\top(Z,\theta_0)]\tau.
\end{aligned}
$$

(c) We can do a Taylor series expansion of $\beta(\theta_n)$ around θ_0 to obtain

$$
\begin{aligned}
\beta(\theta_n) &= \beta(\theta_0) + \frac{\partial\beta(\theta_0)}{\partial\theta^\top}(\theta_n - \theta_0) + o(||\theta_n - \theta_0||^2) \\
&= \beta(\theta_0) + \frac{\partial\beta(\theta_0)}{\partial\theta^\top}(\theta_n - \theta_0) + o(n^{-1}) \\
&= \beta(\theta_0) + \Gamma(\theta_0)(\theta_n - \theta_0) + o(n^{-1}). \qquad \text{(by definition)}
\end{aligned}
$$

We can rearrange the terms to obtain

$$
\sqrt{n}(\beta(\theta_n) - \beta(\theta_0)) = \Gamma(\theta_0)\sqrt{n}(\theta_n - \theta_0) + o(n^{-1}) \underset{n\to\infty}{\longrightarrow} \Gamma(\theta_0)\tau. \quad \text{(by assumption)}
$$

All of this gives us that

$$
\begin{aligned}
\text{(LHS)} &\to N\big(0, \mathbb{E}_{\theta_0}[\varphi\varphi^\top]\big), \\
\text{(a)+(b)} - \text{(c)} &\to N\big(0, \mathbb{E}_{\theta_0}[\varphi\varphi^\top]\big) + \mathbb{E}_{\theta_0}[\varphi(Z)S_\theta^\top(Z,\theta_0)]\tau - \Gamma(\theta_0)\tau.
\end{aligned}
$$

This holds only if

$$
\begin{aligned}
\mathbb{E}_{\theta_0}[\varphi(Z)S_\theta^\top(Z,\theta_0)]\tau - \Gamma(\theta_0)\tau &= 0, \\
\mathbb{E}_{\theta_0}[\varphi(Z)S_\theta^\top(Z,\theta_0)]\tau &= \Gamma(\theta_0)\tau, \\
\mathbb{E}_{\theta_0}[\varphi(Z)S_\theta^\top(Z,\theta_0)] &= \Gamma(\theta_0).
\end{aligned}
$$

□

There exists a special case of this theorem which is the case where we have the projection mapping $\beta(\theta) = \beta$. In which case, we clearly have that $\partial\beta/\partial\beta^\top = I$, the identity matrix, and $\partial\beta/\partial\eta = 0$, the zero matrix. Thus, we obtain the following result.

Corollary 4.1. Suppose Theorem 4.2 holds. Suppose further that our parameter θ can be partitioned into

$$
\theta = \begin{bmatrix}\beta \\ \eta\end{bmatrix} \in \mathbb{R}^p,
$$

meaning that $\beta(\theta) = \beta$ and $\eta(\theta) = \eta$. Then

i) $\mathbb{E}\big[\varphi(Z)S_\beta^\top(Z,\theta_0)\big] = I_{q\times q}$,

ii) $\mathbb{E}\big[\varphi(Z)S_\eta^\top(Z,\theta_0)\big] = 0_{q\times r}$.

5 Final Remarks

The main takeaway from this corollary is that influence functions are orthogonal to the space of functions spanned by the score vector of the nuisance parameter η, which we can define as the *nuisance tangent space*. We do not expand further on this, but there may potentially be a way of using this property to our advantage. In this regard, another area to investigate is a complete extension of influence functions and their geometry to semiparametric models and explore how we may use these results with specific estimators.

Due to time constraints, we do not have time to cover the full depth of the geometry of influence functions. However, one question that arises immediately is: does there exist a converse to the previous result? That is, suppose we can pull out some function φ from our space $\mathcal{H}$ such that conditions i) and ii) of the above corollary are satisfied. Is it possible to find an estimator such that φ will be its influence function? This remains to be determined.

References

Axler, S. (2020). *Measure, integration & real analysis.* Springer Nature.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer.

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Ferguson, T. S. (2017). *A course in large sample theory.* Routledge.

Jiang, J. et al. (2010). *Large sample techniques for statistics.* Springer.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*, volume 61. Springer.

Lehmann, E. L. (1999). *Elements of large-sample theory.* Springer.

Tao, T. (2008). The strong law of large numbers. Accessed: 2024-08-08.

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

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.