Lecture Notes on *Asymptotic Statistics* *

Instructor: Zongming Ma

1 Contiguity

1.1 Motivation

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$, where $\theta \in \Theta$ is an open interval on $\mathbb{R}$. We want to test

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta_1$$

If θ_0 and θ_1 are fixed, then as $n \to \infty$,

$$\mathbb{P}(\text{type I error of LRT}) \to 0 \quad \text{and} \quad \mathbb{P}(\text{type II error of LRT}) \to 0$$

where LRT stands for likelihood ratio test. So, for fixed θ_0, we can make our hypothesis testing problem harder by looking at $\theta_{1,n} \to \theta_0$ as $n \to \infty$ and see whether we still have the good property using any reasonable test.

Several questions arise naturally.

Q1: At what rate shall we let $\theta_{1,n} \to \theta_0$ such that

$$\mathbb{P}_{\theta_0}(\text{type I error of LRT}) \to \alpha \in (0,1) \quad \text{and} \quad \mathbb{P}_{\theta_{1,n}}(\text{type II error of LRT}) \to \beta \in (0, 1-\alpha)$$

Q2: Suppose you have derived this asymptotic null distribution of the likelihood ratio, can we obtain the asymptotic distribution without much effort? Not just about the significance, but also about the power of the test.

Q3: Suppose you have derived the asymptotic null distribution of your favorite test statistic T_n. Can you get its asymptotic distribution without too much effort?

- Yes, as long as you can derive the joint null distribution of the likelihood ratio (LR) and T_n, and the parametric family is smooth.

1.2 Absolute Continuity

Suppose $\mathbb{P}$ and $\mathbb{Q}$ are two probability measures on $(\Omega, \mathcal{A})$.

X is a random variable that either follows the distribution rule $\mathbb{P}$ or $\mathbb{Q}$. If you can evaluate $\mathbb{E}_\mathbb{P}[h(X)]$, when can you also evaluate $\mathbb{E}_\mathbb{Q}[h(X)]$?

The answer is yes, when $\mathbb{Q}$ is absolute continuous w.r.t. $\mathbb{P}$:

$$\mathbb{E}_\mathbb{Q}[h(X)] = \mathbb{E}_\mathbb{P}\left[h(X)\frac{d\mathbb{Q}}{d\mathbb{P}}(X)\right]$$

*This course uses *Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 2000* as textbook.

1.2.1 Definition

Definition 1.1. *Suppose $\mathbb{P}$ and $\mathbb{Q}$ are measures on measurable space $(\Omega, \mathcal{A})$.*

(i.) *$\mathbb{Q}$ is absolutely continuous w.r.t. $\mathbb{P}$ if and only if:*

$$\mathbb{P}(A) = 0 \Rightarrow \mathbb{Q}(A) = 0, \quad \forall A \in \mathcal{A}$$

This is denoted as $\mathbb{Q} \ll \mathbb{P}$.

(ii.) *$\mathbb{Q}$ is orthogonal w.r.t. $\mathbb{P}$ if and only if $\exists \Omega_{\mathbb{P}}, \Omega_{\mathbb{Q}} \subseteq \Omega$ such that:*

$$\Omega = \Omega_{\mathbb{P}} \cup \Omega_{\mathbb{Q}}, \varnothing = \Omega_{\mathbb{P}} \cap \Omega_{\mathbb{Q}}, \mathbb{P}(\Omega_{\mathbb{Q}}) = \mathbb{Q}(\Omega_{\mathbb{P}}) = 0$$

This is denoted as $\mathbb{P} \perp \mathbb{Q}$.

Suppose there is a common dominating measure μ for both $\mathbb{P}$ and $\mathbb{Q}$ so that we have densities p and q, each with support $\{p > 0\}$ and $\{q > 0\}$. Then

- $\mathbb{Q} \ll \mathbb{P}$ is approximately $\{q > 0\} \subseteq \{p > 0\}$.
- $\mathbb{P} \perp \mathbb{Q}$ is approximately $\{p > 0\} \cap \{q > 0\} = \varnothing$.

1.2.2 Lebesgue Decomposition

Theorem 1.2. *Let $\mathbb{P}$ and $\mathbb{Q}$ be two (probability) measures on $(\Omega, \mathcal{A})$. Define $\Omega_{\mathbb{P}} = \{p > 0\}$ and $\Omega_{\mathbb{Q}} = \{q > 0\}$, and two new measures, $\mathbb{Q}^a_{\mathbb{P}}$ and $\mathbb{Q}^{\perp}_{\mathbb{P}}$ on $(\Omega, \mathcal{A})$ as follows:*

$$\mathbb{Q}^a_{\mathbb{P}}(A) = \mathbb{Q}(A \cap \Omega_{\mathbb{P}}), \mathbb{Q}^{\perp}_{\mathbb{P}} = \mathbb{Q}(A \cap \Omega^c_{\mathbb{P}}), \quad \forall A \in \mathcal{A}$$

Then we have the following properties:

(i) $\mathbb{Q} = \mathbb{Q}^a_{\mathbb{P}} + \mathbb{Q}^{\perp}_{\mathbb{P}}$.

(ii) $\mathbb{Q}^a_{\mathbb{P}} \ll \mathbb{P}$ *and* $\mathbb{Q}^{\perp}_{\mathbb{P}} \perp \mathbb{P}$.

As a result, the random variable

$$\left.\frac{d\mathbb{Q}^a_{\mathbb{P}}}{d\mathbb{P}}\right|_{\Omega_{\mathbb{P}}} = \frac{q}{p}, \quad \mathbb{P}\text{-a.s.}$$

is well defined on $\Omega_{\mathbb{P}}$ and is unique $\mathbb{P}$-a.s. We will write it as $\frac{d\mathbb{Q}}{d\mathbb{P}}$ from now on as long as we are only interested in the properties of $\frac{q}{p}$ under $\mathbb{P}$. We can let it be zero on $\Omega^c_{\mathbb{P}}$.

In statistics, the Radon-Nikodym derivative is *likelihood ratio*, which is a random variable that we should study under $\mathbb{P}$, the null hypothesis.

1.2.3 Lemma

Lemma 1.3. *$\mathbb{P}$ and $\mathbb{Q}$ are probability measures on $(\Omega, \mathcal{A})$ and have densities p and q w.r.t. μ. Then for $\mathbb{Q}^a_{\mathbb{P}}$ and $\mathbb{Q}^{\perp}_{\mathbb{P}}$ defined above, we have:*

(i) $\mathbb{Q} = \mathbb{Q}^a_{\mathbb{P}} + \mathbb{Q}^{\perp}_{\mathbb{P}}$, *where* $\mathbb{Q}^a_{\mathbb{P}} \ll \mathbb{P}$ *and* $\mathbb{Q}^{\perp}_{\mathbb{P}} \perp \mathbb{P}$.

(ii) $\mathbb{Q}^a_{\mathbb{P}}(A) = \int_A \frac{q}{p} d\mathbb{P}$, $\forall A \in \mathcal{A}$.

(iii) $\mathbb{Q} \ll \mathbb{P} \Longleftrightarrow \mathbb{Q}(\{p = 0\}) = 0 \Longleftrightarrow \int_\Omega \frac{q}{p} d\mathbb{P} = 1$

Proof. (i) $\forall A \in \mathcal{A}$, $A \cap \Omega_{\mathbb{P}}, A \cap \Omega_{\mathbb{P}}^c \in \mathcal{A}$ and they are disjoint, so $\mathbb{Q}(A) = \mathbb{Q}(A \cap \Omega_{\mathbb{P}}) + \mathbb{Q}(A \cap \Omega_{\mathbb{P}}^c) = \mathbb{Q}_{\mathbb{P}}^a(A) + \mathbb{Q}_{\mathbb{P}}^{\perp}(A)$ by definition and probability axiom. As $A \in \mathcal{A}$ is arbitrary, $\mathbb{Q} = \mathbb{Q}_{\mathbb{P}}^a + \mathbb{Q}_{\mathbb{P}}^{\perp}$.

Let $B \in \mathcal{A}$ such that $\mathbb{P}(B) = 0$. As $\Omega_{\mathbb{P}}$ is the support of p, $B \cap \Omega_{\mathbb{P}} = \varnothing$, so $\mathbb{Q}_{\mathbb{P}}^a(B) = \mathbb{Q}(B \cap \Omega_{\mathbb{P}}) = \mathbb{Q}(\varnothing) = 0$. Hence, $\mathbb{Q}_{\mathbb{P}}^a \ll \mathbb{P}$. Then note that $\Omega_{\mathbb{P}} \cup \Omega_{\mathbb{P}}^c = \Omega$, $\Omega_{\mathbb{P}} \cap \Omega_{\mathbb{P}}^c = \varnothing$, $\mathbb{P}(\Omega_{\mathbb{P}}^c) = 0$, and $\mathbb{Q}_{\mathbb{P}}^{\perp}(\Omega_{\mathbb{P}}) = \mathbb{Q}(\Omega_{\mathbb{P}} \cap \Omega_{\mathbb{P}}^c) = \mathbb{Q}(\varnothing) = 0$. Hence, $\mathbb{Q}_{\mathbb{P}}^{\perp} \perp \mathbb{P}$.

(ii) Let $A \in \mathcal{A}$. By definition,

$$\mathbb{Q}_{\mathbb{P}}^a(A) = \mathbb{Q}(A \cap \Omega_{\mathbb{P}}) = \int_{A \cap \Omega_{\mathbb{P}}} q d\mu = \int_{A \cap \Omega_{\mathbb{P}}} \frac{q}{p} p d\mu = \int_{A \cap \Omega_{\mathbb{P}}} \frac{q}{p} d\mathbb{P} = \int_A \frac{q}{p} d\mathbb{P}$$

(iii) If $\mathbb{Q} \ll \mathbb{P}$, then for any $A \in \mathcal{A}$ such that $\mathbb{P}(A) = 0$, we have $\mathbb{Q}(A) = 0$. Take $A = \{p = 0\}$, we immediately get $\mathbb{Q}(\{p = 0\}) = 0$. On the contrary, if $\mathbb{Q}(\{p = 0\}) = 0$, i.e., $\mathbb{Q}(\Omega_{\mathbb{P}}^c) = 0$, then for any $A \in \mathcal{A}$ such that $\mathbb{P}(A) = 0$, by (i), $\mathbb{Q}(A) = \mathbb{Q}_{\mathbb{P}}^a(A) + \mathbb{Q}_{\mathbb{P}}^{\perp}(A) = 0 + \mathbb{Q}(A \cap \Omega_{\mathbb{P}}^c) \leq \mathbb{Q}(\Omega_{\mathbb{P}}^c) = 0$ as $\mathbb{Q}_{\mathbb{P}}^a \ll \mathbb{P}$. This shows that $\mathbb{Q} \ll \mathbb{P}$.

If $\mathbb{Q}(\Omega_{\mathbb{P}}^c) = 0$ (by above, we also have $\mathbb{Q} \ll \mathbb{P}$), then $\int_\Omega \frac{q}{p} d\mathbb{P} = \int_{\Omega_{\mathbb{P}}} \frac{d\mathbb{Q}}{d\mathbb{P}} d\mathbb{P} = 1 - \mathbb{Q}(\Omega_{\mathbb{P}}^c) = 1$. Notice that each equal sign here is an <u>if and only if</u> statement, so we also have $\int_\Omega \frac{q}{p} d\mathbb{P} = 1 \Rightarrow \mathbb{Q}(\Omega_{\mathbb{P}}^c) = 0$.

□

1.3 Contiguity

As mentioned before, when $\mathbb{Q} \ll \mathbb{P}$, for a random variable $X : \Omega \mapsto \mathbb{R}^k$, we can recover $\mathbb{Q}(X \in B)$ by $\mathbb{E}\left[\mathbb{1}_B(X) \frac{d\mathbb{Q}}{d\mathbb{P}}(X)\right]$ for $B \in \mathcal{B}(\mathbb{R}^k)$. Yet when $\mathbb{Q}$ is not absolute continuous w.r.t. $\mathbb{P}$, we cannot recover the $\mathbb{Q}$-measurable part that is not in the support of p.

Now suppose that there is a sequence of measurable spaces $\{(\Omega_n, \mathcal{A}_n)\}$, where each is equipped with two probability measures $\mathbb{P}_n$ and $\mathbb{Q}_n$. We need some tools to study the problem above asymptotically.

1.3.1 Definition

Definition 1.4. *$\mathbb{Q}_n$ is contiguous w.r.t. $\mathbb{P}_n$ if for any sequence of events $\{A_n\}$ we have $\mathbb{P}_n(A_n) \to 0 \Rightarrow \mathbb{Q}_n(A_n) \to 0$. This is denoted as $\mathbb{Q}_n \triangleleft \mathbb{P}_n$.*

If $\mathbb{P}_n \triangleleft \mathbb{Q}_n$ and $\mathbb{Q}_n \triangleleft \mathbb{P}_n$, we say that they are mutually contiguous, i.e., $\mathbb{P}_n \triangleleft\triangleright \mathbb{Q}_n$.

Remark 1.5. *At each n, let $\frac{d\mathbb{Q}_n}{d\mathbb{P}_n}$ denote $\frac{d\mathbb{Q}_n^a}{d\mathbb{P}_n}$, where $\mathbb{Q}_n^a$ be the part of $\mathbb{Q}_n$ that is absolute continuous w.r.t. $\mathbb{P}_n$, which is well defined on $\Omega_{\mathbb{P}_n}$. $\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}$ is defined in the same manner. Then $\mathbb{E}_{\mathbb{P}_n}\left[\frac{d\mathbb{Q}_n}{d\mathbb{P}_n}\right] = \mathbb{Q}_n(\Omega_{\mathbb{P}_n}) \leq 1$ and $\mathbb{E}_{\mathbb{Q}_n}\left[\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}\right] = \mathbb{P}_n(\Omega_{\mathbb{Q}_n}) \leq 1$. The two sequences are both tight. By Prokhorov's theorem, every subsequence has a weakly convergent subsequence.*

1.3.2 Le Cam's First Lemma

Lemma 1.6. *The following statements are equivalent:*

(i) $\mathbb{Q}_n \triangleleft \mathbb{P}_n$

(ii) If $\frac{d\mathbb{P}_n}{d\mathbb{Q}_n} \overset{\mathcal{D}}{\rightsquigarrow} U$ along a subsequence under $\mathbb{Q}_n$, then $\mathbb{P}(U > 0) = 1$.

(iii) If $\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \overset{\mathcal{D}}{\rightsquigarrow} V$ along a subsequence under $\mathbb{P}_n$, then $\mathbb{E}[V] = 1$.

(iv) For any sequence of random variables $T_n : (\Omega_n, \mathcal{A}_n) \mapsto \mathbb{R}^k$, if $T_n \overset{\mathbb{P}}{\to} 0$ under $\mathbb{P}_n$, then $T_n \overset{\mathbb{P}}{\to} 0$ under $\mathbb{Q}_n$.

Remark 1.7. *By Lemma 1.2.3, $\mathbb{Q} \ll \mathbb{P}$ is equivalent to $\frac{p}{q} > 0$ ($\mathbb{Q}$-a.s.) and is equivalent to $\int \frac{q}{p} d\mathbb{P} = 1$, which is also equivalent to $\mathbb{P}(T \neq 0) \Rightarrow \mathbb{Q}(T \neq 0) = 0$.*

Proof. We first show that (i) and (iv) are equivalent. When $\mathbb{Q}_n \lhd \mathbb{P}_n$, take any $\{A_n\}$ and let $T_n = \mathbb{1}_{A_n}$. If $T_n \overset{\mathbb{P}}{\to} 0$ under $\mathbb{P}_n$, then $\mathbb{P}_n(\{\omega \in \Omega : \omega \in A_n\}) = \mathbb{P}_n(A_n) \to 0$, which implies $\mathbb{Q}_n(A_n) \to 0$ by contiguity, so $\mathbb{Q}_n(T_n = 0) \to 0$, i.e., $T_n \overset{\mathbb{P}}{\to} 0$ under $\mathbb{Q}_n$. Conversely, when (iv) holds, fix $\varepsilon > 0$ and let $A_n = \{||T_n|| > \varepsilon\}$. Then (i) holds by definition.

Next, we show that (i) $\Rightarrow$ (ii). Suppose $\mathbb{P}(U = 0) > 0$, i.e., $\mathbb{P}(U > 0) < 1$. Then there exist an $\varepsilon > 0$, a subsequence $n_k \to \infty$, and a subsequence $\varepsilon_{n_k} \downarrow 0$ such that $\lim_{k\to\infty} \mathbb{Q}_{n_k}(\frac{d\mathbb{P}_{n_k}}{d\mathbb{Q}_{n_k}} < \varepsilon_{n_k}) > \varepsilon$. For each n_k, let $A_{n_k} = \{\frac{d\mathbb{P}_{n_k}}{d\mathbb{Q}_{n_k}} < \varepsilon_{n_k}\} \cap \{q_{n_k} > 0\}$. Then we have

$$\mathbb{P}_{n_k}(A_{n_k}) = \int_{\{\frac{d\mathbb{P}_{n_k}}{d\mathbb{Q}_{n_k}} < \varepsilon_{n_k}\}\cap\{q_{n_k}>0\}} \frac{d\mathbb{P}_{n_k}}{d\mathbb{Q}_{n_k}} d\mathbb{Q}_{n_k} = \int_{\{\frac{d\mathbb{P}_{n_k}}{d\mathbb{Q}_{n_k}} < \varepsilon_{n_k}\}} \frac{d\mathbb{P}_{n_k}}{d\mathbb{Q}_{n_k}} d\mathbb{Q}_{n_k} < \varepsilon_{n_k} \overset{k\to\infty}{\longrightarrow} 0$$

As $\mathbb{Q}_{n_k} \lhd \mathbb{P}_{n_k}$, $\mathbb{Q}_{n_k}(A_{n_k}) = \mathbb{Q}_{n_k}(\frac{d\mathbb{P}_{n_k}}{\mathbb{Q}_{n_k}} < \varepsilon_{n_k}) \to 0$ as $k \to \infty$, which is a contradiction.

Then, we show that (iii) $\Rightarrow$ (i). Consider any sequence of events $\{A_n\}$ such that $\mathbb{P}_n(A_n) \to 0$. We need to show that $\mathbb{Q}_n(A_n) \to 0$, or, equivalently, $\mathbb{Q}_n(A_n^c) \to 1$. In fact,

$$\mathbb{Q}_n(A_n^c) = \int_{A_n^c} d\mathbb{Q}_n \geq \int_{A_n^c} \frac{d\mathbb{Q}_n}{d\mathbb{P}_n} d\mathbb{P}_n = \int_{\Omega_n} \frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \mathbb{1}_{A_n^c} d\mathbb{P}_n$$

By (iii), $\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \overset{\mathcal{D}}{\rightsquigarrow} V \geq 0$ under $\mathbb{P}_n$. Also, $\mathbb{1}_{A_n^c} \overset{\mathcal{D}}{\rightsquigarrow} 1$ under $\mathbb{P}_n$ by definition of $\{A_n\}$. By Fatou's leamma,

$$\liminf_{n\to\infty} \mathbb{Q}_n(A_n^c) \geq \int V d\mathbb{P} = \mathbb{E}[V] = 1 \Rightarrow \mathbb{Q}_n(A_n^c) = 1 \Rightarrow \mathbb{Q}_n \lhd \mathbb{P}_n$$

Finally, we show that (ii) $\Rightarrow$ (iii). Define $\mu_n = \frac{1}{2}(\mathbb{P}_n + \mathbb{Q}_n)$ that dominates both $\mathbb{P}_n$ and $\mathbb{Q}_n$ for each n. Let $W_n = \frac{d\mathbb{P}_n}{d\mu_n} \in [0, 2]$ bounded. From this definition, $\mathbb{E}[W_n] = 1$. Since both $\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}$ and $\frac{d\mathbb{Q}_n}{d\mathbb{P}_n}$ are tight sequences, there exists a subsequence along which

$$\frac{d\mathbb{P}_n}{d\mathbb{Q}_n} \overset{\mathcal{D}}{\rightsquigarrow} U, \frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \overset{\mathcal{D}}{\rightsquigarrow} V, \frac{d\mathbb{P}_n}{d\mu_n} \overset{\mathcal{D}}{\rightsquigarrow} W$$

under $\mathbb{Q}_n, \mathbb{P}_n, \mu_n$, respectively. By DCT, $\mathbb{E}[W] = 1$. To prove (iii), we first show that $\mathbb{P}(W = 0) = 0$, then show $\mathbb{E}[V] = 1$.

Fix any bounded continuous function f, we define bounded function $g : [0, 2] \mapsto \mathbb{R}$ as following:

$$g(x) = \begin{cases} f(\frac{x}{2-x}) \cdot (2 - x) & 0 \leq x < 2 \\ 0 & x = 2 \end{cases}$$

Then, $\mathbb{Q}$-a.s. we have $\frac{\mathbb{P}_n}{\mathbb{Q}_n} = \frac{d\mathbb{P}_n/d\mu_n}{d\mathbb{Q}_n/d\mu_n} = \frac{W_n}{2-W_n}$, so

$$\mathbb{E}_{\mathbb{Q}_n}\left[f\left(\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}\right)\right] = \int f\left(\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}\right)\frac{d\mathbb{Q}_n}{d\mu_n} d\mu_n = \int f\left(\frac{W_n}{2-W_n}\right)(2 - W_n) d\mu_n = \int g(W_n) d\mu_n$$

Hence, by DCT, $\lim_{n\to\infty} \mathbb{E}_{\mathbb{Q}_n}\left[f\left(\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}\right)\right] = \lim_{n\to\infty} \mathbb{E}_{\mu_n}[g(W_n)] = \mathbb{E}[g(W)]$, $\lim_{n\to\infty} \mathbb{E}_{\mathbb{Q}_n}\left[f\left(\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}\right)\right] = \mathbb{E}[f(U)]$.

We pick a sequence of bounded continuous functions $\{f_m\}$ such that $|f_m| \leq 1$ and $f_m \to \mathbb{1}_{\{0\}}$. By above, for each m, we have $\mathbb{E}[f_m(U)] = \mathbb{E}[g_m(W)]$. Let $m \to \infty$ and use DCT again,

$$\lim_{m\to\infty} \mathbb{E}[f_m(U)] = \mathbb{P}(U=0) = \lim_{m\to\infty} \mathbb{E}[g_m(W)] = \mathbb{E}\left[\mathbb{1}_{\{\frac{W}{2-W}=0\}}(2-W)\right] = 2\mathbb{P}(W=0)$$

By assumption of (ii), $\mathbb{P}(U=0)=0$, so $\mathbb{P}(W=0)=0 \Rightarrow \mathbb{P}(W \neq 0) = 1$.

For each bounded continuous function $\tilde{f}$, define $\tilde{g} : [0,2] \mapsto \mathbb{R}$ as:

$$\tilde{g}(x) = \begin{cases} \tilde{f}\left(\frac{2-x}{x}\right)x & \text{if } x \in (0,2] \\ 0 & \text{if } x = 0 \end{cases}$$

By similar arguments using DCT, $\mathbb{E}[\tilde{V}] = \mathbb{E}[\tilde{W}]$. Pick a sequence of bounded continuous functions:

$$\tilde{f}_m(x) = \begin{cases} \min\{x, m\} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Then, by MCT, $\mathbb{E}[\tilde{f}_m(V)] \uparrow \mathbb{E}[V]$, which is equal to

$$\mathbb{E}[\tilde{g}_m(W)] \to \mathbb{E}\left[\left(\frac{2-W}{W}\right)W\mathbb{1}_{\{w=0\}}\right] = 2\mathbb{P}(W \neq 0) - 1 = 1$$

$\square$

1.3.3 Example

Suppose $\frac{d\mathbb{P}_n}{d\mathbb{Q}_n} \overset{\mathcal{D}}{\rightsquigarrow} e^{\mathcal{N}(\mu,\sigma^2)}$ under $\mathbb{Q}_n$.

(i) Since e^z is always positive, by Le Cam's First Lemma (ii), we immediately have $\mathbb{Q}_n \lhd \mathbb{P}_n$.

(ii) When can we have $\mathbb{P}_n \lhd \mathbb{Q}_n$? We need $\mathbb{E}[e^{\mathcal{N}(\mu,\sigma^2)}] = 1$ by Le Cam's First Lemma (iii). That is, if and only if $\mu + \frac{1}{2}\sigma^2 = 0$.

1.3.4 Le Cam's Third Lemma

Theorem 1.8. *Let $\{\mathbb{P}_n\}, \{\mathbb{Q}_n\}$ be two sequences of probability measures on $\{(\Omega_n, \mathcal{A}_n)\}$ and $\{X_n\}$ is a sequences of random variables where $X_n : \Omega_n \mapsto \mathbb{R}^k$ for each n. Suppose that*

(i) $\mathbb{Q}_n \lhd \mathbb{P}_n$

(ii) $\begin{bmatrix} X_n \\ \frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \end{bmatrix} \overset{\mathcal{D}}{\rightsquigarrow} \begin{bmatrix} X \\ V \end{bmatrix}$ *under* $\mathbb{P}_n$

Then we have:

(i) $\mathcal{L}(B) = \mathbb{E}[\mathbb{1}_B(X)V], \forall B \in \mathcal{B}(\mathbb{R}^k)$ *defines a probability measure.*

(ii) $X_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}$ *under* $\mathbb{Q}_n$.

Proof. (i) $\mathcal{L}(\mathbb{R}^k) = \mathbb{E}[V] = 1$ because $\mathbb{Q}_n \lhd \mathbb{P}_n$ and Le Cam's First Lemma (iii).

(ii) Fix any lower semi-continuous function f bounded from below, then $(x, \theta) \mapsto \theta f(x)$ is lower semi-continuous and bounded from below as long as $\theta \in [0, +\infty)$ (well, θ is the likelihood ratio).

$$\liminf_{n\to\infty} \mathbb{E}_{\mathbb{Q}_n}[f(X_n)] \geq \liminf_{n\to\infty} \int f(x)\frac{d\mathbb{Q}_n}{d\mathbb{P}_n}(x)d\mathbb{P}_n(x) \geq \mathbb{E}[f(X)V] = \int f d\mathcal{L}$$

The first $\geq$ since we only integrate the absolute continuous part; the second $\geq$ by Fatou's Lemma. Then choose $f = \mathbb{1}_B$, by the Portmanteau Theorem, we have $X_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}$ under $\mathbb{Q}_n$.

$\square$

1.3.5 A Special (The Original) Case of Le Cam's Third Lemma

Corollary 1.9. $\begin{bmatrix} X_n \\ \log \frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \end{bmatrix} \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}_{k+1}\left(\begin{bmatrix} \mu_{k\times 1} \\ -\frac{1}{2}\sigma^2 \end{bmatrix}, \begin{bmatrix} \Sigma_{k\times k} & \tau_{k\times 1} \\ \tau^T & \sigma^2 \end{bmatrix} \right)$ *under* $\mathbb{P}_n$*, then* $X_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}_k(\mu + \tau, \Sigma)$ *under* $\mathbb{Q}_n$.

Proof. By continuity of $\exp(\cdot)$, $\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \overset{\mathcal{D}}{\rightsquigarrow} e^W$ under $\mathbb{P}_n$, where $W \sim \mathcal{N}(-\frac{1}{2}\sigma^2, \sigma^2)$. By Example 1.3.3., we have $\mathbb{Q}_n \lhd\rhd \mathbb{P}_n$. By the general form of Le Cam's Third Lemma, $X_n \overset{\mathcal{D}}{\rightsquigarrow} Y$ under $\mathbb{Q}_n$ for some Y. For any bounded continuous function f, $\mathbb{E}[f(Y)] = \mathbb{E}[f(X)e^W]$. Take $f(y) = e^{it^T y}, \forall t \in \mathbb{R}^k$. Then

$$M_Y(t) = \mathbb{E}[e^{it^T Y}] = \mathbb{E}[e^{it^T X} e^W] = \mathbb{E}[\exp\{ \begin{bmatrix} it^T & 1 \end{bmatrix} \begin{bmatrix} X \\ W \end{bmatrix} \}] = \mathbb{E}[e^\xi]$$

where $M_Y(t)$ is the characteristic function of random variable Y and ξ is Normal with mean $it^T\mu - \frac{1}{2}\sigma^2$ and variance $\begin{bmatrix} it^T & 1 \end{bmatrix} \begin{bmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{bmatrix} \begin{bmatrix} it \\ 1 \end{bmatrix}$. As a side note, for a random variable $V \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{E}[e^V] = \int \frac{e^{v - \frac{1}{2\sigma^2}(v-\mu)^2}}{\sqrt{2\pi}\sigma} dv = \int \frac{e^{-\frac{(v-\mu-\sigma^2)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dv \, e^{\frac{(\mu+\sigma^2)^2 - \mu^2}{2\sigma^2}} = e^{\mu + \frac{1}{2}\sigma^2}$$

Hence,

$$\begin{aligned} M_Y(t) = \mathbb{E}[e^\xi] &= \exp\left\{ it^T\mu - \frac{1}{2}\sigma^2 + \frac{1}{2} \begin{bmatrix} it^T & 1 \end{bmatrix} \begin{bmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{bmatrix} \begin{bmatrix} it \\ 1 \end{bmatrix} \right\} \\ &= \exp\left\{ it^T\mu - \frac{1}{2}\sigma^2 + \frac{1}{2}(-t^T\Sigma t + 2it^T\tau + \sigma^2) \right\} \\ &= \exp\{it^T(\mu+\tau) - \frac{1}{2}t^T\Sigma t\} \end{aligned}$$

which is the characteristic function of $\mathcal{N}_k(\mu + \tau, \Sigma)$. $\square$

2 Quadratic Mean Differentiability and Local Asymptotic Normality

2.1 Quadratic Mean differentiability

2.1.1 Definition

Definition 2.1. $\{\mathbb{P}_\theta : \theta \in \Theta\}$ *is a parametric family of probability measures (we have a parametric statistical model), where* $\Theta \subseteq \mathbb{R}^k$. *Let* $\theta_0 \in \Theta^o$ *and suppose* $\mathbb{P}_\theta$ *has pdf* p_θ *around* θ_0.

Then the model is q.m.d. at $\theta = \theta_0$ *if there exists a measurable function* $\dot{\ell}_{\theta_0}$ *such that as* $\theta \to \theta_0$,

$$\int \left(\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right)^2 d\mu = o(||\theta - \theta_0||^2)$$

This looks close to Hellinger distance, which is easy to work with in scenarios with i.i.d. data.

Remark 2.2. *When* $k = 1$, *q.m.d. is the same as*

$$\int \left(\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0) \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right)^2 d\mu = o((\theta - \theta_0)^2)$$
$$\Rightarrow \int \left(\frac{\sqrt{p_\theta} - \sqrt{p_{\theta_0}}}{\theta - \theta_0} - \frac{1}{2} \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right)^2 d\mu = o(1)$$

and $\dot{\ell}_{\theta_0}$ *is the score function at* θ_0.

2.1.2 Implication of q.m.d. for a Parametric Family

Theorem 2.3 (Likelihood Ratio under q.m.d.)**.** *Suppose* $\{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ *is q.m.d. at* $\theta_0 \in \Theta^o$ *with some* $\dot{\ell}_{\theta_0}$. *Then*

(i) $\mathbb{E}_{\theta_0}[\dot{\ell}_{\theta_0}(X)] = 0$, *and* $I_{\theta_0} = \mathbb{E}_{\theta_0}\left[\dot{\ell}_{\theta_0}(X)\dot{\ell}_{\theta_0}(X)^T\right]$ *is well-defined entrywise (both* L_1 *and* L_2*).*

(ii) $\forall h \in \mathbb{R}^k$, *as* $n \to \infty$,

$$\log \prod_{i=1}^n \frac{\mathbb{P}_{\theta_0 + \frac{k}{\sqrt{n}}}}{\mathbb{P}_{\theta_0}}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^T \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2} h^T I_{\theta_0} h + o_{\mathbb{P}_{\theta_0}}(1)$$

which means that we can approximate at a specific rate

Remark 2.4.
$$\frac{1}{\sqrt{n}} \sum_{i=1}^n h^T \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2} h^T I_{\theta_0} h \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(-\frac{1}{2} h^T I_{\theta_0} h, h^T I_{\theta_0} h)$$

i.e., $\mathcal{L}(X_1, \ldots, X_n | \theta_0 + \frac{h}{\sqrt{n}}) \lhd\rhd \mathcal{L}(X_1, \ldots, X_n | \theta_0)$.

Proof. Step 0 : Preparation.

Denote $p_n := p_{\theta_0 + \frac{h}{\sqrt{n}}}$, $p := p_{\theta_0}$, and $g = h^T \dot{\ell}_{\theta_0}$. So, the parametric family being q.m.d. implies

$$\int \left[\sqrt{n}(\sqrt{p_n} - \sqrt{p}) - \frac{1}{2} g \sqrt{p} \right]^2 d\mu = o(1) \quad \cdots (*)$$

which is equivalent to

$$\sqrt{n}(\sqrt{p_n} - \sqrt{p}) - \frac{1}{2} g\sqrt{p} \overset{L^2(\mu)}{\longrightarrow} 0 \quad \text{as } n \to \infty$$

Step 1 : We are going to prove part (i).

$(*)$ implies that $\sqrt{n}(\sqrt{p_n} - \sqrt{p}) - \frac{1}{2}g\sqrt{p} \in L^2(\mu)$. As $\sqrt{p}, \sqrt{p_n} \in L^2(\mu)$, we have $\frac{1}{2}g\sqrt{p} \in L^2(\mu)$. So,

$$+\infty > \int g^2 p d\mu = \int h^T \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T h \, p_{\theta_0} d\mu = \mathbb{E}_{\theta_0}[h^T \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T h]$$

Letting h be one-hot vectors, we conclude that the diagonal elements of $\mathbb{E}_{\theta_0}[\dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T]$ are all finite. By Cauchy-Schwartz inequality, all off-diagonal elements are also finite.

Another implication of $g\sqrt{p} \in L^2(\mu)$ is that we have $\sqrt{n}(\sqrt{p_n} - \sqrt{p}) \to \frac{1}{2}g\sqrt{p}$ in $L^2(\mu)$, which implies $\sqrt{n}(\sqrt{\frac{p_n}{p}} - 1) \to \frac{1}{2}g$ in $L^2(p)$. Since p is a pdf, $L^2(p) \subseteq L^1(p)$, so $\sqrt{n}(\sqrt{\frac{p_n}{p}} - 1) - \frac{1}{2} \to 0$ in $L^1(p)$. Since $g \in L^2(p)$, g is also in $L^1(p)$, so is $\sqrt{n}(\sqrt{\frac{p_n}{p}} - 1)$. Then it suffices to show that $\int \sqrt{n}(\sqrt{\frac{p_n}{p}} - 1) p d\mu \to 0$:

$$\begin{aligned}
\int \sqrt{n}(\sqrt{\frac{p_n}{p}} - 1) p d\mu &= \int \sqrt{n}(\sqrt{p_n p} - p) d\mu \\
&= \sqrt{n}\left[\int \sqrt{p_n p} d\mu - 1\right] \\
&= \sqrt{n} \int (\sqrt{p_n p} - \frac{1}{2}p - \frac{1}{2}p_n) d\mu \\
&= -\frac{1}{2}\sqrt{n} \int (\sqrt{p} - \sqrt{p_n})^2 d\mu
\end{aligned}$$

We know that $\sqrt{n}(\sqrt{p_n} - \sqrt{p}) \to \frac{1}{2}gp$ in $L^2(\mu)$, so $n(\sqrt{p_n} - \sqrt{p})^2$ is bounded from above. Hence, for each n, the above is bounded by $\frac{C}{\sqrt{n}}$ for some $C > 0$, which converges to 0 as $n \to \infty$.

Finally, we have

$$\left| \int \frac{1}{2} g \, p d\mu - \int \sqrt{n}(\sqrt{\frac{p_n}{p}} - 1) p d\mu \right| \leq \int \left| \frac{1}{2}gp - \sqrt{n}(\sqrt{\frac{p_n}{p}} - 1) \right| p d\mu \overset{L^1}{\longrightarrow} 0$$

so $0 = 2\int \frac{1}{2} g \, p d\mu = \mathbb{E}_{\theta_0}[g] = \mathbb{E}_{\theta_0}[h^T \dot{\ell}_{\theta_0}]$ for any $h \in \mathbb{R}^k$, which proves (i) entirely.

Step 2 : We are going to prove part (ii).

$$\begin{aligned}
\log \prod_{i=1}^n \frac{p_n}{p}(X_i) &= 2\sum_{i=1}^n \log \sqrt{\frac{p_n}{p}(X_i)} \\
&= 2\sum_{i=1}^n \log(1 + \frac{W_{n,i}}{2}) \quad W_{n,i} = 2\left(\sqrt{\frac{p_n}{p}(X_i)} - 1\right) \\
&= 2\sum_{i=1}^n \left(\frac{W_{n,i}}{2} - \frac{W_{n,i}^2}{8} + W_{n,i}^2 R(W_{n,i}) \right) \quad \text{by 2nd order Taylor expansion} \\
&= \underbrace{\sum_{i=1}^n W_{n,i}}_{\textbf{(I)}} \underbrace{-\frac{1}{4}\sum_{i=1}^n W_{n,i}^2}_{\textbf{(II)}} + \underbrace{2\sum_{i=1}^n W_{n,i}^2 R(W_{n,i})}_{\textbf{(III)}}
\end{aligned}$$

where $R(x) \to 0$ as $|x| \to 0$.

As **(I)** is a random variable, we study its mean and variance, respectively:

$$\begin{aligned}
\mathbb{E}_p[\textbf{(I)}] = n\mathbb{E}_p[W_{1,i}] &= 2n\mathbb{E}_p\left[\sqrt{\frac{p_n}{p}(X_i)} - 1\right] \quad \text{by definition and iid} \\
&= 2n\int (\sqrt{\frac{p_n}{p}} - 1)p d\mu \\
&= -\int (\sqrt{n}(\sqrt{p_n} - \sqrt{p}))^2 d\mu \quad \text{use same trick as before} \\
&= -\int (\frac{1}{2}g\sqrt{p})^2 d\mu \quad \text{since } \sqrt{n}(\sqrt{p_n} - \sqrt{p}) \xrightarrow{L^2(\mu)} \frac{1}{2}gp \\
&= -\frac{1}{4}\mathbb{E}_p[(g(X_1))^2]
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}_p\left(\sum_{i=1}^n W_{n,i} - \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i)\right) &= n\,\mathrm{Var}_p(W_{n,1} - \frac{1}{\sqrt{n}}g(X_1)) \quad \text{by i.i.d.} \\
&\le \mathbb{E}_p[(W_{n,1} - \frac{1}{\sqrt{n}}g(X_1))^2] \\
&= n\int \left(2\left(\sqrt{\frac{p_n}{p}} - 1\right) - \frac{1}{\sqrt{n}}g\right)^2 p d\mu \\
&= 4\int (\sqrt{n}(\sqrt{p_n} - \sqrt{p}) - \frac{1}{2}g\sqrt{p})^2 d\mu = o_{\mathbb{P}}(1)
\end{aligned}$$

So, $\textbf{(I)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i) - \frac{1}{4}\mathbb{E}_p[(g(X_1))^2] + o_{\mathbb{P}}(1)$.

For **(II)**, recall that $\sqrt{n}(\sqrt{\frac{p_n}{p}} - 1) \xrightarrow{L^2(p)} \frac{1}{2}g$, so $2\sqrt{n}(\sqrt{\frac{p_n}{p}} - 1) \to g$ in $L^2(p)$, which implies that $nW_{n,i}^2 = (g(X_i))^2 + A_{n,i}$, where $\mathbb{E}_p[|A_{n,i}|] \to 0$ as $n \to \infty$. By LLN, $\bar{A}_n = \frac{1}{n}\sum_{i=1}^n A_{n,i} \xrightarrow{\mathbb{P}} 0$. Thus,

$$\sum_{i=1}^n W_{n,i}^2 = \underbrace{\frac{1}{n}\sum_{i=1}^n (g(X_i))^2}_{\xrightarrow{\mathbb{P}}\mathbb{E}_p[(g(X_1))^2]} + \underbrace{\frac{1}{n}\sum_{i=1}^n A_{n,i}}_{\xrightarrow{\mathbb{P}}0} \Longrightarrow \textbf{(II)} = -\frac{1}{4}\sum_{i=1}^n W_{n,i}^2 \xrightarrow{\mathbb{P}} -\frac{1}{4}\mathbb{E}_p[(g(X_1))^2]$$

which is a constant (no need to consider the variance of **(II)**).

Note that **(I)** + **(II)** is already what we want to show for part (ii). Hence, our goal now is to show that $\textbf{(III)} = 2\sum_{i=1}^n W_{n,i}^2 R(W_{n,i}) \xrightarrow{\mathbb{P}} 0$. As $\sum_{i=1}^n W_{n,i}^2$ is $\mathcal{O}_{\mathbb{P}}(1)$, it suffices to show that $\max_i |W_{n,i}| \xrightarrow{\mathbb{P}} 0$.

$$\begin{aligned}
\forall \varepsilon > 0, \quad & \mathbb{P}(\max_i |W_{n,i}| > \varepsilon) \\
&\le n\mathbb{P}(|W_{n,1}| > \varepsilon) \quad \text{by union bound and i.i.d.} \\
&= n\mathbb{P}\left(W_{n,i}^2 = \frac{1}{n}\left[(g(X_1))^2 + A_{n,1}\right] > \varepsilon^2\right) \quad \text{recall in \textbf{(II)}} \\
&\le n\mathbb{P}\left(\frac{1}{n}(g(X_1))^2 > \frac{1}{2}\varepsilon^2\right) + n\mathbb{P}\left(\frac{1}{n}|A_{n,1}| > \frac{1}{2}\varepsilon^2\right) \quad \text{by Triangular Inequality} \\
&\le \underbrace{\frac{n}{\frac{1}{2}n\varepsilon^2}}_{\text{constant}} \underbrace{\mathbb{E}_p[(g(X_1))^2 \mathbb{1}_{(g(X_1))^2 > \frac{1}{2}n\varepsilon^2}]}_{\xrightarrow{\mathbb{P}}0} + \underbrace{\frac{n}{\frac{1}{2}n\varepsilon^2}}_{\text{constant}} \underbrace{\mathbb{E}_p[|A_{n,1}|]}_{\xrightarrow{\mathbb{P}}0} \quad \text{by Markov Inequality} \\
&\Longrightarrow \max_i |W_{n,i}| \xrightarrow{\mathbb{P}} 0
\end{aligned}$$

□

2.1.3 A Sufficient Condition for q.m.d.

Lemma 2.5. *Suppose that for every $\theta \in \Theta$ (an open set of $\mathbb{R}^k$), the mapping $\theta \mapsto \sqrt{p_\theta(x)}$ is continuously differentiable in θ for μ-a.e. x. Then p_θ is continuously differentiable in θ for μ-a.e. x.*

If further, the elements of $I_\theta = \int (\frac{\dot{p}_\theta}{p_\theta})(\frac{\dot{p}_\theta}{p_\theta})^T p_\theta d\mu$ is well-defined and continuous in θ, then $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is q.m.d. at every $\theta \in \Theta$, with $\dot{\ell}_\theta = \frac{\dot{p}_\theta}{p_\theta} = (\log p_\theta)'$.

Proof. Let $s_\theta(x) = \sqrt{p_\theta(x)}$. $s_\theta(x)$ is continuously differentiable in θ for μ-a.e. x. So for the mapping $\theta \mapsto p_\theta(x)$, $\dot{p}_\theta(x) = 2s_\theta(x)\dot{s}_\theta(x)$ is continuous ($p_\theta(x)$ is continuously differentiable).

Since $s_\theta(x)$ is nonnegative and continuous, when $s_\theta(x) = 0$, $\dot{s}_\theta(x) = 0$. So we can write

$$\dot{s}_\theta(x) = \frac{1}{2} \cdot \frac{\dot{p}_\theta(x)}{s_\theta(x)} = \frac{1}{2}\frac{\dot{p}_\theta(x)}{p_\theta(x)}\sqrt{p_\theta(x)}$$

where $\frac{0}{0} := 0$. Now we have a map $\theta \mapsto I_\theta = 4\int \dot{s}_\theta(x)\dot{s}_\theta(x)^T d\mu$ that is continuous.

Consider some index $t \to 0$ and $h_t \in \mathbb{R}^k$ with $||h_t|| \le 1$. Without loss of generality, assume that $h_t \to h$ as $t \to 0$ (or take a subsequence, by Bolzano-Weierstrass Theorem). So for μ-a.e. x, we have $\frac{s_{\theta_t h_t} - s_\theta}{t} \to h^T s_\theta$ as $t \to 0$. In addition,

$$\begin{aligned}
\int \left[\frac{s_{\theta+th_t} - s_\theta}{t}\right]^2 d\mu &= \int \left(\int_0^1 h_t^T s_{\theta+uth_t}\right)^2 d\mu \quad \text{by Fundamental Theorem of Calculus} \\
&\le \int \int_0^1 \left(h_t^T s_{\theta+uth_t}\right)^2 du d\mu \quad \text{Cauchy-Schwartz} \\
&= \frac{1}{4}\int_0^1 h_t^T I_{\theta+uth_t} h_t du \quad \text{Fubini}
\end{aligned}$$

As $t \to 0$, this converges to $\frac{1}{4}h^T I_\theta h = \int \left(h^T s_\theta\right)^2 d\mu$.

This implies that

$$\begin{aligned}
&\int \left[\frac{s_{\theta+th_t} - s_\theta}{t} - h^T s_\theta\right]^2 d\mu = o(1) \\
\Leftrightarrow &\int (s_{\theta+th_t} - s_\theta - th^T s_\theta)^2 d\mu = o(t^2) \\
\Leftrightarrow &\int (s_{\theta'} - s_\theta - (\theta' - \theta)^T s_\theta)^2 d\mu = o(||\theta' - \theta||^2) \quad \text{let } \theta' = \theta + th_t
\end{aligned}$$

□

Examples:

(i) $p_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}, \forall\theta \in \mathbb{R}$. Cauchy location family.

$$\frac{d}{d\theta}\sqrt{p_\theta(x)} = \frac{1}{\sqrt{\pi}}(x-\theta)\left(1+(x-\theta)^2\right)^{-\frac{3}{2}}$$

$$I_\theta = \int_{\mathbb{R}} \frac{4}{\pi} \cdot \frac{x^2}{(1+x^2)^3}dx = \frac{1}{2}$$

so this family is q.m.d.

(ii) $p_\theta(x) = \frac{1}{\theta}\mathbb{1}_{x\in[0,\theta]}, \forall\theta \in (0, +\infty)$. If this family is q.m.d., then fix $h > 0$, we would have

$$\limsup_{n\to\infty} \int_{\mathbb{R}} \left[\frac{\sqrt{p_{\theta+\frac{h}{\sqrt{n}}}(x)} - \sqrt{p_\theta(x)}}{\frac{h}{\sqrt{n}}} \right]^2 dx < +\infty$$

which implies

$$\limsup_{n\to\infty} n \int_{\mathbb{R}} \left(\sqrt{p_{\theta+\frac{h}{\sqrt{n}}}(x)} - \sqrt{p_\theta(x)} \right)^2 dx < +\infty$$

However,

$$n \int_{\mathbb{R}} \left(\sqrt{p_{\theta+\frac{h}{\sqrt{n}}}(x)} - \sqrt{p_\theta(x)} \right)^2 dx \geq \int_\theta^{\theta+\frac{h}{\sqrt{n}}} \frac{1}{\theta + \frac{h}{\sqrt{n}}} dx = \frac{\sqrt{n}h}{\theta + \frac{h}{\sqrt{n}}}$$

This term is unbounded as $n \to \infty$, which is a contradiction, so this family is not q.m.d.

2.2 Local Asymptotic Normality

Suppose we have a single observation $X \sim \mathcal{N}_k(h, I_{\theta_0}^{-1})$ for some $h \in \mathbb{R}^k$.

The log-likelihood ratio with respect to $\mathcal{N}_k(0, I_{\theta_0}^{-1})$ is

$$\log \frac{\exp\left\{-\frac{1}{2}(x-h)^T I_{\theta_0}(x-h)\right\}}{\exp\left\{-\frac{1}{2}x^T I_{\theta_0} x\right\}} = h^T I_{\theta_0} x - \frac{1}{2} h^T I_{\theta_0} h$$

Take $\dot{\ell}_{\theta_0} = (\log p_\theta)' \Big|_{\theta=\theta_0}$. The covariance for $I_{\theta_0}X$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)$ are

$$\begin{aligned} &\mathrm{Cov}(I_{\theta_0}X) = I_{\theta_0} I_{\theta_0}^{-1} I_{\theta_0} = I_{\theta_0} \\ &\mathrm{Cov}(\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)) = \frac{1}{n} \cdot n\,\mathrm{Cov}(\dot{\ell}_{\theta_0}(X_1)) = \mathbb{E}_p[\dot{\ell}_{\theta_0}(X_1)\dot{\ell}_{\theta_0}(X_1)^T] = I_{\theta_0} \end{aligned}$$

$\{\mathcal{N}_k(\theta, I_{\theta_0}) : \theta \in \mathbb{R}^k\}$ is q.m.d., by theorem, the q.m.d. likelihood-ratio should be

$$h^T \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2} h^T I_{\theta_0} h + o_{\mathbb{P}}(1)$$

The two likelihood ratios are the same up to some negligible terms $o_{\mathbb{P}}(1)$.

Theorem 2.6 (Local Asymptotic Normality)**.** *$\{\mathbb{P}_\theta : \theta \in \Theta\}$ is q.m.d. at θ_0 for some $\dot{\ell}_{\theta_0}$. $I_{\theta_0} := \mathbb{E}[\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^T]$. Let $T_n = T_n(X_1, \ldots, X_n) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}_h$ under $\mathbb{P}^n_{\theta+\frac{1}{\sqrt{n}}h}$ for any fixed $h \in \mathbb{R}^k$, where $\mathcal{L}_h$ is some probability law. Then there exists a random variable $T = T(X, U)$, where $X \sim \mathcal{N}_k(h, I_{\theta_0}^{-1})$, $U \sim U(0,1)$, and $X \perp\!\!\!\perp U$, such that $\mathcal{L}(T(X,U)) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}_h$.*

Remark 2.7. *This means that for any testing statistic T_n that is weakly convergent is "matched" by a statistic T asymptotically. Now, we only need to study Gaussian location family. Asymptotically, the score function is a sufficient statistic for a smooth parametric family.*

Proof. Let $\mathbb{P}_{n,h} := \mathbb{P}^n_{\theta_0+\frac{1}{\sqrt{n}}h}$, $\forall h \in \mathbb{R}^k$, and $\Delta_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i)$. By q.m.d.-LR theorem,

$$\log \frac{d\mathbb{P}_{n,h}}{d\mathbb{P}_{n,0}} = h^T\Delta_n - \frac{1}{2}h^T I_{\theta_0} h + o_{\mathbb{P}}(1)$$

So, after some calculation, under $\mathbb{P}_{n,0}$, $\begin{bmatrix}\Delta_n \\ \log\frac{\mathbb{P}_{n,h}}{\mathbb{P}_{n,0}}\end{bmatrix} \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}_{k+1}\left(\begin{bmatrix}0 \\ -\frac{1}{2}h^T I_{\theta_0} h\end{bmatrix}, \begin{bmatrix} I_{\theta_0} & I_{\theta_0}h^T \\ h^T I_{\theta_0} & h^T I_{\theta_0} h\end{bmatrix}\right)$.

By Corollary 1.9, $\Delta_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}_k\left(I_{\theta_0}h, I_{\theta_0}\right)$ under $\mathbb{P}_{n,h}$ as $n \to \infty$ for any $h \in \mathbb{R}^k$. Then we consider $\begin{bmatrix}T_n \\ \Delta_n\end{bmatrix}$ where under $\mathbb{P}_{n,0}$, $T_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}_0$ and $\Delta_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}_k(0, I_{\theta_0})$. $\left\{\begin{bmatrix}T_n \\ \Delta_n\end{bmatrix}\right\}$ is then a tight sequence, which implies that there exists a subsequence $\{n_k\}$ and (S, Δ) such that $\begin{bmatrix}T_{n_k} \\ \Delta_{n_k}\end{bmatrix} \overset{\mathcal{D}}{\rightsquigarrow} \begin{bmatrix}S \\ \Delta\end{bmatrix}$ under $\mathbb{P}_{n,0}$. The marginal distributions $S \overset{\mathcal{D}}{=} \mathcal{L}_0$ and $\Delta \overset{\mathcal{D}}{=} \mathcal{N}_k(0, I_{\theta_0})$. By continuous mapping theorem,

$$\begin{bmatrix}T_{n_k} \\ \frac{d\mathbb{P}_{n_k,h}}{d\mathbb{P}_{n_k,0}}\end{bmatrix} \overset{\mathcal{D}}{\rightsquigarrow} \begin{bmatrix} S \\ \exp\left\{h^T\Delta - \frac{1}{2}h^T I_{\theta_0} h\right\}\end{bmatrix}$$

Without loss of generality, assume that S is a scalar. Fix $\Delta = \delta \in \mathbb{R}^k$. To analyze $\mathcal{L}(S|\Delta = \delta) := F_\delta$, let $F_\delta^{-1}(U) = F_\delta := \tilde{T}(\Delta, U)$, where $U \sim U(0,1)$. This implies that $\begin{bmatrix}\tilde{T}(\Delta, U) \\ \Delta\end{bmatrix} \overset{\mathcal{D}}{=} \begin{bmatrix}S \\ \Delta\end{bmatrix}$.

Suppose $X \sim \mathcal{N}_k(h, I_{\theta_0}^{-1})$, where $X \perp\!\!\!\perp U$. Then $\forall B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
\mathbb{P}_h(\tilde{T}(I_{\theta_0}X, U) \in B) &= \int_{\mathbb{R}^k} \mathbb{P}_h(\tilde{T}(I_{\theta_0}x, U) \in B)\sqrt{\frac{\det(I_{\theta_0})}{(2\pi)^n}} \exp\left\{-\frac{1}{2}(x-h)^T I_{\theta_0}(x-h)\right\} dx \\
&= \int_{\mathbb{R}^k} \mathbb{P}_h(\tilde{T}(\delta, U) \in B)\frac{1}{\sqrt{(2\pi)^n \det(I_{\theta_0})}} \exp\left\{-\frac{1}{2}\delta^T I_{\theta_0}^{-1}\delta + h^T\delta - \frac{1}{2}h^T I_{\theta_0} h\right\} d\delta \\
&\quad \text{(change of variable: } \delta = I_{\theta_0}x - h) \\
&= \mathbb{E}_0\left[\mathbb{1}_{\{\tilde{T}(\Delta,U)\in B\}} e^{h^T\Delta - \frac{1}{2}h^T I_{\theta_0} h}\right] \quad \Delta \sim \mathcal{N}_k(0, I_{\theta_0}), U \sim U(0,1),\ U \perp\!\!\!\perp X \\
&= \mathbb{E}_0\left[\mathbb{1}_{S\in B} e^{h^T\Delta - \frac{1}{2}h^T I_{\theta_0} h}\right] \\
&= \mathcal{L}_h \quad \text{by Le Cam Lemma 3}
\end{aligned}$$

so the distribution of S under $\mathbb{P}_{n,0}$ is $\mathcal{L}_h$.

When S is not a scalar-valued random variable, say $S = (S_1, S_2)$ in the simplest case. In the proof above, we should look at $S_1|\Delta = \delta$, then $S_2|\Delta = \delta, S_1 = s_1$. For a realization of $U \sim U(0,1)$, $U = 0.u_1u_2u_3\cdots$ and let $U_1 = 0.u_1u_3u_5\cdots$ and $U_2 = 0.u_2u_4u_6\cdots$. $U_1, U_2 \overset{i.i.d.}{\sim} U(0,1)$. Hence, we are able to generate entries of S successively. □

2.3 Examples

2.3.1 Asymptotic Linear Statistic

Suppose the parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is q.m.d. at θ_0 and $\mathbb{P}^n_{\theta_0} \lhd\rhd \mathbb{P}^n_{\theta_0+\frac{h}{\sqrt{n}}}$. Let $\hat{\theta}_h$ be a statistic such that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i) + o_{\mathbb{P}^n_{\theta_0}}(1)$$

Assume $\mathbb{E}_{\theta_0}[\psi(X_1)] = 0$ and $\mathrm{Var}_{\theta_0}(\psi(X_1)) = \tau^2 < \infty$. Then

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ \log \frac{d\mathbb{P}^n_{\theta_0+\frac{h}{\sqrt{n}}}}{\mathbb{P}^n_{\theta_0}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i) \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T \dot{\ell}_{\theta_0}(X_i) - \frac{1}{2}h^T I_{\theta_0} h \end{bmatrix} + o_{\mathbb{P}^n_{\theta_0}}(1) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}_2\left(\begin{bmatrix} 0 \\ -\frac{1}{2}h^T I_{\theta_0} h \end{bmatrix}, \begin{bmatrix} \tau^2 & \sigma_{12} \\ \sigma_{12} & h^T I_{\theta_0} h \end{bmatrix} \right)$$

under $\mathbb{P}^n_{\theta_0}$, where $\sigma_{12} = \mathrm{Cov}_{\theta_0}(\psi(X_1), h^T \dot{\ell}_{\theta_0}(X_1))$. By Le Cam's Third Lemma, under $\mathbb{P}^n_{\theta_0+\frac{1}{\sqrt{n}}h}$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(\sigma_{12}, \tau^2)$$

2.3.2 T-test without Normal Distribution

Let $X \sim f_\theta(x) = f(x-\theta), \forall \theta \in \mathbb{R}$, with mean 0 and variance σ^2. If f is differentiable and $\int (f')^2 \frac{1}{f} d\mu < +\infty$, then $\{f_\theta\}$ is q.m.d. By applying Lemma in the context that $H_0 : \theta = 0$, $T_n = \frac{\sqrt{n}\bar{X}_n}{S} = \sqrt{n}\frac{\bar{X}_n}{\sigma} + o_{\mathbb{P}^n_0}(1)$. So it fits the previous example with $\psi(X_i) = \frac{X_i}{\sigma}$ and $\dot{\ell}_{\theta_0}(X_i) = -\frac{f'(X_i)}{f(X_i)}$. Hence,

$$\sigma_{12} = \mathrm{Cov}_0\left(\frac{X_i}{\sigma}, -h\frac{f'(X_i)}{f(X_i)} \right) = -\frac{h}{\sigma}\int x f'(x) dx = \frac{h}{\sigma}$$

where we used integration by parts and assuming $xf(x) \to 0$ as $n \to \infty$.

So under $\frac{h}{\sqrt{n}}$, $T_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(\frac{h}{\sigma}, 1)$.

2.3.3 Sign Test

Assume further that $f(x)$ is positive and continuous at 0, with $\int_{-\infty}^0 f(x)dx = \frac{1}{2} = \int_0^\infty f(x)dx$.

Consider test statistic

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\mathbb{1}_{X_i>0} - \frac{1}{2} \right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{2}\mathrm{sign}(x)$$

Under $H_0 : \theta = 0$, $S_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0, \frac{1}{4})$, and

$$\sigma_{12} = \mathrm{Cov}_0\left(\mathbb{1}_{X_i>0}, -h\frac{f'(X_i)}{f(X_i)} \right) = -h\int_0^\infty f'(x)dx = h \cdot f(0)$$

so, under $\frac{h}{\sqrt{n}}$, $2S_n \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(2hf(0), 1)$.

Compare with T_n: when $f(0)$ is peaked, then using sign test gives better power, so it all depends on $\frac{1}{\sigma}$ and $2f(0)$.

2.3.4 NP Statistic

$$\log(\mathcal{L}_n, h) = \log \prod_{i=1}^n \frac{p_{\theta_0+\frac{h}{\sqrt{n}}}(X_i)}{p_{\theta_0}(X_i)}$$

Under $H_0 : \theta = \theta_0$, $\log(\mathcal{L}_n, h) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(-\frac{1}{2}\sigma_h^2, \sigma_h^2)$, where $\sigma_h^2 = h^T I_{\theta_0} h$.

Under $H_1 : \theta = \theta_0 + \frac{h}{\sqrt{n}}$, $\log(\mathcal{L}_n, h) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(\frac{1}{2}\sigma_h^2, \sigma_h^2)$.

If we are at level α, we reject when $\dfrac{\log(\mathcal{L}_n, h) + \frac{1}{2}\sigma_h^2}{\sigma_h} > z_{1-\alpha}$.

The power is then equal to

$$\mathbb{P}\left(\frac{\mathcal{N}(\frac{1}{2}\sigma_h^2, \sigma_h^2) + \frac{1}{2}\sigma_h^2}{\sigma_h} > z_{1-\alpha}\right) = \mathbb{P}(\mathcal{N}(0,1) > z_{1-\alpha} - \sigma_h)$$

3 Asymptotic Efficiency of Estimators

3.1 Motivation

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}_\theta$. We want to estimate some function $\psi(\theta)$ of θ (ψ is known).

Suppose $\sqrt{n}(T_n - \psi(\theta)) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}_\theta$ under $\mathbb{P}_\theta^n$. Then T_n is "good" if $\mathcal{L}_\theta$ is concentrated, such as $\int |x|^2 d\mathcal{L}_\theta(x)$, $\int |x| d\mathcal{L}_\theta(x)$, or $\int \frac{|x|^2}{2} \wedge a(|x| - \frac{1}{2}a)_+ d\mathcal{L}_\theta(x)$ being small.

Counterexample: Hodge's Estimator

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$. Let $T_n = \bar{X}_n$, $S_n = \begin{cases} T_n & \text{if } |T_n| \geq n^{-\frac{1}{4}} \\ 0 & \text{otherwise} \end{cases}$

For any fixed θ, $\sqrt{n}(T_n - \theta) \overset{\mathcal{D}}{=} \mathcal{N}(0,1)$ (this is exact, not even using CLT).

- When $\theta = 0$, $\mathbb{P}(S_n = 0) = \mathbb{P}(|T_n| < n^{-\frac{1}{4}}) \overset{n\to\infty}{\to} 1$ because $T_n \overset{\mathbb{P}}{\to} 0$ exponentially fast. So $\sqrt{n} S_n \overset{\mathbb{P}}{\to} 0$, which implies $\sqrt{n}(S_n - 0) \overset{\mathcal{D}}{\rightsquigarrow} 0$. Hence, the MSE for S_n goes to 0, better than that of T_n, which is always equal to 1.
- When $\theta \neq 0$, as $n \to \infty$, $|T_n|$ would be concentrated around an $\frac{1}{\sqrt{n}}$ neighborhood of θ, which moves away from $n^{-\frac{1}{4}}$. Hence, $\mathbb{P}(|T_n| \geq n^{-\frac{1}{4}}) \to 1$ as $n \to \infty$, i.e., $\mathbb{P}(S_n - T_n \neq 0) \to 0$.

  We use quadratic loss function when comparing how large the improvement is from T_n to S_n:

$$
\begin{aligned}
\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta[(\sqrt{n}(T_n - \theta))^2] &= 1 \qquad \text{this holds for all } \theta \\
\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta[(\sqrt{n}(S_n - \theta)^2] &\geq \mathbb{E}_{n^{-\frac{1}{4}}}[(\sqrt{n}(S_n - n^{-\frac{1}{4}}))^2] \\
&\geq \mathbb{E}_{n^{-\frac{1}{4}}}[(\sqrt{n}(S_n - n^{-\frac{1}{4}}))^2 \mathbb{1}_{S_n=0}] \\
&= \sqrt{n}\mathbb{E}_{n^{-\frac{1}{4}}}[\mathbb{1}_{S_n=0}] \\
&= \frac{\sqrt{n}}{2}(1 + o(1))
\end{aligned}
$$

  As $n \to \infty$, the latter could be arbitrary large. As another way to see this, we choose $h \in \mathbb{R}$, then

$$
\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta\left[(\sqrt{n}(S_n - \theta)^2\right] \geq \mathbb{E}_{\frac{h}{\sqrt{n}}}\left[\left(\sqrt{n}\left(S_n - \frac{h}{\sqrt{n}}\right)\right)^2 \mathbb{1}_{S_n=0}\right] = h^2 \mathbb{P}_{\frac{h}{\sqrt{n}}}(|T_n| < n^{-\frac{1}{4}}) \to h^2
$$

  We can choose h arbitrarily large. So, S_n beats T_n only when $\theta = 0$.

3.2 The Special Role of Normal Mean Estimation

Suppose $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is q.m.d. at some $\theta_0 \in \Theta^o$ and

$$
\sqrt{n}\left(T_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}_{\theta_0,h} \qquad \cdots (*)
$$

under $\mathbb{P}^n_{\theta_0 + \frac{h}{\sqrt{n}}}$ for all $h \in \mathbb{R}^k$, where $\psi : \mathbb{R}^k \mapsto \mathbb{R}^m$, with derivative matrix $\dot{\psi}_{\theta_0} \in \mathbb{R}^{m\times k}$ at θ_0.

From $(*)$ we have

$$\sqrt{n}(T_n - \psi(\theta_0)) = \underbrace{\sqrt{n}\left(T_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right)}_{\overset{\mathcal{D}}{\rightsquigarrow}\mathcal{L}_{\theta_0,h}} + \underbrace{\sqrt{n}\left(\psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - \psi(\theta_0)\right)}_{\overset{\mathbb{P}}{\to}\dot{\psi}_{\theta_0}h}$$
$$\overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}_{\theta_0,h} * \delta_{\dot{\psi}_{\theta_0}h}$$

for all $h \in \mathbb{R}^k$ under $\mathbb{P}^n_{\theta_0+\frac{h}{\sqrt{n}}}$. δ is delta measure.

By Local Asymptotic Normality Theorem, there exists $T(X,U)$ such that $T(X,U) \overset{\mathcal{D}}{=} \mathcal{L}_{\theta_0,h} * \delta_{\dot{\psi}_{\theta_0}h}$, where $X \sim \mathcal{N}_k(h, I_{\theta_0}^{-1})$ and $U \sim U(0,1)$, $X \perp\!\!\!\perp U$.

The implication of this is that if we understand what is optimal for Gaussian location family, then we understand the smooth parametric family.

3.3 Equivariant-in-Law Estimator in Normal Location Family

Let $X \sim \mathcal{N}_k(h, \Sigma = I_{\theta_0}^{-1})$, $U \sim U(0,1)$, and $X \perp\!\!\!\perp U$. We want to estimate Ah (A is known) with $T(X,U)$.

Definition 3.1. *(i) The estimator $T(X,U)$ is equivariant if $T(x+h',u) - Ah' = T(x,u), \forall h' \in \mathbb{R}^k$.*

(ii) The estimator $T(X,U)$ is equivariant-in-law if $T(X+h',U) - Ah' \overset{\mathcal{D}}{=} T(X,U), \forall h' \in \mathbb{R}^k$. This is weaker than equivariant.

Proposition 3.2. *The null distribution $h = 0$ of any randomized equivariant-in-law estimator of Ah can be decomposed as $\mathcal{L} = \mathcal{N}(0, A\Sigma A^T) * \mathcal{M}$, i.e., it has the distribution as $AX + Y$, where $X \sim \mathcal{N}(0,\Sigma)$, $Y \sim \mathcal{M}$, and $X \perp\!\!\!\perp Y$.*

Proof. Assume $A = I_k$, the $k \times k$ identity matrix.

Consider a sequence of priors π_m on h, e.g., $\pi_m := \mathcal{N}_k(0, m\Sigma)$.

Let $H_m \sim \pi_m$. Then $X|H_m = h \sim \mathcal{N}(h,\Sigma)$, so $\begin{bmatrix} X \\ H_m \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma & m\Sigma \\ m\Sigma & m\Sigma \end{bmatrix}\right) \quad \cdots (**)$

Let F_m be the marginal distribution of X when $H_m \sim \pi_m$ and $\mathcal{L}$ be the law of $T(X,U) - h$ when $X \sim \mathcal{N}(h,\Sigma)$. Then $T(X,U) - H_m \overset{\mathcal{D}}{=} \mathcal{L}$ for every m when $(X, H_m)^T$ is jointly distributed according to $(**)$. For any realization h of H_m, the law $\mathcal{L}$ is the same, so there is nothing to marginalize.

Fix any continuity point θ of $\mathcal{L}$, the cdf of $\mathcal{L}$ is then $\mathcal{L}(\theta) = \mathbb{P}(T(X,U) - H_m \le \theta)$. By $(**)$, as $n \to \infty$,

$$H_m|X = x \sim \mathcal{N}_k\left(\frac{m}{m+1}x, \frac{m}{m+1}\Sigma\right) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}_k(x,\Sigma)$$

So, given $X = x, U = u$,

$$\begin{aligned} &\mathbb{P}(T(X,U) - H_m \le \theta | X = x, U = u) \\ =&\mathbb{P}(H_m \ge T(X,U) - \theta | X = x, U = u) \\ \to& 1 - \Phi_\Sigma(T(x,u) - \theta - x) \\ =&\Phi_\Sigma(\theta - T(x,u) + x) \end{aligned}$$

where $\Phi_\Sigma(\cdot)$ is the cdf of $\mathcal{N}_k(0,\Sigma)$. With this,

$$\begin{aligned}\mathcal{L}(\theta) &= \int_0^1 \int_{\mathbb{R}^k} \mathbb{P}(T(X,U) - H_m \le \theta | X = x, U = u) dF_m(x) du \\ &= \int_0^1 \int_{\mathbb{R}^k} \Phi_\Sigma(\theta - T(x,u) + x) dF_m(x) du + o_m(1) \\ &= \int_{\mathbb{R}^k} \Phi_\Sigma(\theta - t) dG_m(t) + o_m(1)\end{aligned}$$

where $G_m(t) := \int_0^1 F_m(T(X < U) - X \le t) du$. G_m is the law of $T(X,U) - X$, with $X \sim F_m, U \sim U(0,1)$, and $X \perp\!\!\!\perp U$. Thus, $T(X,U) - X = \underbrace{T(X,U) - H_m}_{\mathcal{L}} + \underbrace{H_m - X}_{\mathcal{N}(0,\Sigma)}$, regardless of m.

So both $\{T(X,U) - H_m\}$ and $\{H_m, X\}$ are tight sequences, which implies that $\{G_m\}$ is also a tight sequence. $\square$

3.4 Hájek's Convolution Theorem

Definition 3.3 (Regular Estimator Sequence). *T_n is regular at θ_0 for estimator a parameter $\psi(\theta)$ if $\forall h \in \mathbb{R}^k$,*

$$\sqrt{n}\left(T_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}_{\theta_0}$$

under $\mathbb{P}_{\theta + \frac{h}{\sqrt{n}}}$.

The limiting distribution does not change locally. $\mathcal{L}_{\theta_0}$ can be any law.

For example, for a location family, the variance is the same, so $\mathcal{L}_{\theta_0}$ is the same.

Theorem 3.4. *Suppose $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is q.m.d. at $\theta \in \Theta$ with nondegenerate I_{θ_0}. $\psi : \mathbb{R}^k \mapsto \mathbb{R}^m$ is differentiable at θ_0 with derivative matrix $\dot\psi_{\theta_0} \in \mathbb{R}^{m \times k}$.*

If T_n is regular for estimating $\psi(\theta)$ at $\theta = \theta_0$, then

$$\mathcal{L}_{\theta_0} = \mathcal{N}_m(0, \dot\psi_{\theta_0} I_{\theta_0}^{-1} \dot\psi_{\theta_0}^T) * \mathcal{M}_{\theta_0}$$

for some $\mathcal{M}_{\theta_0}$.

In particular, if $\mathcal{L}_{\theta_0}$ has covariance matrix Σ_{θ_0}, then we have $\Sigma_{\theta_0} - \dot\psi_{\theta_0} I_{\theta_0}^{-1} \dot\psi_{\theta_0}^T \succeq \mathcal{O}_{m\times m}$.

Proof. T_n is regular, so $\sqrt{n}\left(T_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{L}_{\theta_0}$ under $\mathbb{P}_{\theta + \frac{h}{\sqrt{n}}}$. Then,

$$\sqrt{n}(T_n - \psi(\theta_0)) = \underbrace{\sqrt{n}\left(T_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right)}_{\overset{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}}{\longrightarrow} \mathcal{L}_{\theta_0}} + \underbrace{\sqrt{n}\left(\psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - \psi(\theta_0)\right)}_{\to \dot\psi_{\theta_0} h} \overset{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}}{\longrightarrow} \mathcal{L}_{\theta_0} * \delta_{\dot\psi_{\theta_0} h}$$

By Theorem 2.6 (score is asymptotically sufficient), there exists $T(X,U)$ such that $\mathcal{L}(T(X,U)) \sim \mathcal{L}_{\theta_0} * \delta_{\dot\psi_{\theta_0} h}$ when $X \sim \mathcal{N}_k(h, I_{\theta_0}^{-1})$.

If we fix $X \sim \mathcal{N}_k(h, I_{\theta_0}^{-1})$, $T(X+h, U) \sim \mathcal{L}_{\theta_0} * \delta_{\dot\psi_{\theta_0} h}$. So, $T(X+h,U) - \dot\psi_{\theta_0} h \sim \mathcal{L}_{\theta_0}$. This shows that $T(X+h,U) - \dot\psi_{\theta_0} h$ is equivariant-in-law for estimating $\dot\psi_{\theta_0} h$.

Hence, $\mathcal{L}_{\theta_0} \sim \mathcal{N}_m(0, \dot\psi_{\theta_0} I_{\theta_0}^{-1} \dot\psi_{\theta_0}^T) * \mathcal{M}$ for some law $\mathcal{M}$ by previous proposition. $\square$

Corollary 3.5. *The statement "If T_n is regular, then $\sqrt{n}(T_n - \psi(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\psi}_{\theta_0} I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i) + o_{\mathbb{P}_{\theta_0}}(1)$" if and only if T_n is the best regular estimator sequence.*

4 Projection and U-Statistics

Consider

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \overset{i.i.d.}{\sim} \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}\right), \quad \theta \in [-1, 1]$$

we can estimate θ by sample covariance

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

which may be a little bit hard to calculate. However, with U-Statistics, it can be easier to calculate, as they are symmetric functions of the sample:

$$\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} (X_i - X_j)(Y_i - Y_j)$$

Projections make U-Statistics into additive functions of the sample.

4.1 Projection

4.1.1 Definition

Definition 4.1. *T and $\{S : S \in \mathcal{S}\}$ are random variables with finite second moment. A random variable $\hat{S}$ is a projection of T onto $\mathcal{S}$ (or L_2-projection) if $\hat{S} \in \operatorname{argmin}_{S \in \mathcal{S}} \mathbb{E}[(T - S)^2]$.*

4.1.2 Theorem (Orthogonality)

Theorem 4.2. *Let $\mathcal{S}$ be a linear space, i.e., if $S_1, S_2 \in \mathcal{S}$, then $\alpha_1 S_1 + \alpha_2 S_2 \in \mathcal{S}, \forall \alpha_1, \alpha_2 \in \mathbb{R}$, and all elements of $\mathcal{S}$ have finite second moment.*

(i) $\hat{S}$ is a projection of T onto $\mathcal{S}$ if and only if $\hat{S} \in \mathcal{S}$ and $\mathbb{E}[(T - \hat{S})S] = 0, \forall S \in \mathcal{S}$.

(ii) $\hat{S}$ is unique (any two projections are equal almost surely).

(iii) If $1 \in \mathcal{S}$, then $\mathbb{E}[T] = \mathbb{E}[\hat{S}]$ and $\mathbb{E}[(T - \hat{S})S] = 0, \forall S \in \mathcal{S}$.

Proof. We first prove (i). ($\Leftarrow$) Fix $\hat{S} \in \mathcal{S}$ such that $\mathbb{E}[(T - \hat{S})S] = 0, \forall S \in \mathcal{S}$. Now take an arbitrary $S' \in \mathcal{S}$. We then have

$$\mathbb{E}[(T - S')^2] = \mathbb{E}[(T - \hat{S} + \hat{S} - S')^2] = \mathbb{E}[(T - \hat{S})^2] + \mathbb{E}[(\hat{S} - S')^2] + 2\mathbb{E}[(T - \hat{S})(\hat{S} - S')]$$

Since $\mathcal{S}$ is linear, $\hat{S} - S' \in \mathcal{S}$, so $2\mathbb{E}[(T - \hat{S})(\hat{S} - S')] = 0$. Then, $\mathbb{E}[(T - S')^2] \ge \mathbb{E}[(T - \hat{S})^2]$, so $\hat{S} \in \operatorname{argmin}_{S \in \mathcal{S}} E[(T - S)^2]$. ($\Rightarrow$) Let $\alpha \in \mathbb{R}$. Take any $S \in \mathcal{S}$.

$$\mathbb{E}[(T - \hat{S} - \alpha S)^2] = \mathbb{E}[(T - \hat{S})^2] - 2\alpha\mathbb{E}[(T - \hat{S})S] + \alpha^2\mathbb{E}[S^2]$$

This quantity should be no smaller than $\mathbb{E}[(T - \hat{S})^2$, by the definition of $\hat{S}$, which implies that

$$-2\alpha\mathbb{E}[(T - \hat{S})S] + \alpha^2\mathbb{E}[S^2] \ge 0, \quad \forall \alpha \in \mathbb{R}, \forall S \in \mathcal{S} \quad \Rightarrow \mathbb{E}[(T - \hat{S})S] = 0, \quad \forall S \in \mathcal{S}$$

since the left-hand side can be viewed as a parabola in α and we want to make it always nonnegative.

To prove (ii), note that in the proof of (i) ($\Leftarrow$), $\mathbb{E}[(T - S')^2] = \mathbb{E}[(T - \hat{S})^2]$ holds if and only if $\mathbb{E}[(S - \hat{S})^2] = 0$, which implies that $S = \hat{S}$ almost surely. For (iii), as $1 \in \mathcal{S}$, by (i), we can let $S = 1$ and conclude that $\mathbb{E}[(T - \hat{S}) \cdot 1] = 0$. Hence, $\mathbb{E}[T] = \mathbb{E}[\hat{S}]$. Then, $\forall S \in \mathcal{S}$,

$$\operatorname{Cov}(T - \hat{S}, S) = \mathbb{E}[(T - \hat{S})S] - \mathbb{E}[(T - \hat{S})]\mathbb{E}[S] = 0 - 0 \cdot \mathbb{E}[S] = 0$$

$\square$

4.1.3 Theorem (Co-Convergence)

Theorem 4.3. *Let $\mathcal{S}_n$ be a sequence of linear spaces of random variables with finite second moment that contain 1. Let T_n be a sequence of random variables with projection $\hat{S}_n \in \mathcal{S}_n$ for all n.*

If $\frac{\mathrm{Var}(T_n)}{\mathrm{Var}(\hat{S}_n)} \to 1$, then $\frac{T_n - \mathbb{E}[T_n]}{\mathrm{SD}(T_n)} - \frac{\hat{S}_n - \mathbb{E}[\hat{S}_n]}{\mathrm{SD}(\hat{S}_n)} \xrightarrow{\mathbb{P}} 0$.

Proof. Let $\Delta_n = \frac{T_n - \mathbb{E}[T_n]}{\mathrm{SD}(T_n)} - \frac{\hat{S}_n - \mathbb{E}[\hat{S}_n]}{\mathrm{SD}(\hat{S}_n)}$. We aim to show $\mathbb{E}[\Delta_n^2] \to 0$, which implies $\Delta_n \xrightarrow{\mathbb{P}} 0$.

$$
\begin{aligned}
\mathbb{E}[\Delta_n^2] &= \frac{\mathbb{E}[(T_n - \mathbb{E}[T_n])^2]}{\mathrm{Var}(T_n)} + \frac{\mathbb{E}[(\hat{S}_n - \mathbb{E}[\hat{S}_n])^2]}{\mathrm{Var}(\hat{S}_n)} - 2\mathbb{E}\left[\frac{T_n - \mathbb{E}[T_n]}{\mathrm{SD}(T_n)} \frac{\hat{S}_n - \mathbb{E}[\hat{S}_n]}{\mathrm{SD}(\hat{S}_n)}\right] \\
&= 1 + 1 - 2\frac{\mathrm{Cov}(T_n, \hat{S}_n)}{\mathrm{SD}(T_n)\,\mathrm{SD}(\hat{S}_n)} \\
&= 2 - 2\frac{\mathrm{Cov}(T_n - \hat{S}_n, \hat{S}_n) + \mathrm{Cov}(\hat{S}_n, \hat{S}_n)}{\mathrm{SD}(T_n)\,\mathrm{SD}(\hat{S}_n)} \\
&= 2 - 2\frac{\mathrm{SD}(\hat{S}_n)}{\mathrm{SD}(T_n)} \to 0
\end{aligned}
$$

□

4.1.4 Hájek Projection

Definition 4.4. *Let $X_1, \ldots, X_n$ be independent random variables. Define*

$$
\mathcal{S}_n = \left\{ \sum_{i=1}^{n} g_i(X_i) : \mathbb{E}[(g_i(X_i))^2] < +\infty, \forall i \in [n] \right\}
$$

where g_i's are arbitrary measurable functions.

The reason we study this class of random variables is that it has the form of a sum, and we can use central limit theorem to derive its weak convergence.

The projection of an arbitrary random variable $T = T(X_1, \ldots, X_n)$ onto $\mathcal{S}_n$ is its Hájek projection.

4.1.5 Theorem

Theorem 4.5. *Let T be a random variable such that $\mathbb{E}[T^2] < +\infty$. Then T has its Hájek projection as*

$$
\hat{S} = \sum_{i=1}^{n} \mathbb{E}[T|X_i] - (n-1)\mathbb{E}[T]
$$

Before proving it, here is a recap on conditional expectation, which is also a kind of projection. For a random variable X, its expectation $\mathbb{E}[X]$ minimizes $\mathbb{E}[(X-c)^2]$ over all $c \in \mathbb{R}$. Similarly, suppose we have two random variables X and Y, and we want to minimize $\mathbb{E}[(X - g(Y))^2]$ over all measurable functions g. We get the conditional expectation $\mathbb{E}[X|Y]$, which is a random variable, and $\mathbb{E}[X|Y] = g_0(Y)$ for some measurable function g_0.

Proof. By the Theoreom 4.1.2 (i), we need to show that $\hat{S} \in \mathcal{S}_n$ and $\mathbb{E}[(T - \hat{S})S] = 0, \forall S \in \mathcal{S}_n$.

Since for each i, the conditional expectation $\mathbb{E}[T|X_i]$ is a measurable function of X_i and we can absorb $-(n-1)\mathbb{E}[T]$ into measurable functions, we immediately conclude that $\hat{S} \in \mathcal{S}_n$. Then $\forall i, j \in [n]$,

$$\mathbb{E}[\mathbb{E}[T|X_i]|X_j] = \begin{cases} \mathbb{E}[\mathbb{E}[T|X_i]] = \mathbb{E}[T] & \text{if } j \neq i \\ \mathbb{E}[T|X_i] & \text{if } j = i \end{cases}$$

by independence. So $\forall j \in [n], \forall g$ measurable,

$$\begin{aligned} \mathbb{E}[\hat{S}|X_j] &= \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}[T|X_i] - (n-1)\mathbb{E}[T] \middle| X_j\right] \\ &= \sum_{i \neq j} \mathbb{E}[\mathbb{E}[T|X_i]|X_j] + \mathbb{E}[T|X_j] - (n-1)\mathbb{E}[T] \\ &= \mathbb{E}[T|X_j] \\ \mathbb{E}[(T - \hat{S})g(X_j)] &= \mathbb{E}\left[\mathbb{E}[(T - \hat{S})g(X_j)|X_j]\right] \quad \text{tower of expectation} \\ &= \mathbb{E}\left[\mathbb{E}[T - \hat{S}|X_j]g(X_j)\right] \quad \text{since } g(X_j) \text{ is measurable w.r.t. } \sigma(X_j) \\ &= 0 \quad \text{by above} \end{aligned}$$

Take any $S = \sum_{i=1}^n g_i(X_i) \in \mathcal{S}_n$, we have $\mathbb{E}[(T - \hat{S})S] = \sum_{i=1}^n \mathbb{E}[(T - \hat{S})g_i(X_i)] = 0$. □

4.2 One-sample U-Statistics

4.2.1 Definition

Definition 4.6. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$. h is a function with r arguments that is permutation symmetric, i.e.,*

$$h(x_1, \ldots, x_r) = h(x_{\pi(1)}, \ldots, x_{\pi(r)})$$

for all $(x_1, \ldots, x_r) \in Dom(h)$ and for all π that is a permutation of $\{1, \ldots, n\}$.

A U-Statistics with kernel h is then

$$U = \frac{1}{\binom{n}{r}} \sum_{1 \leq \beta_1 < \cdots < \beta_r \leq n} h(X_{\beta_1}, \ldots, X_{\beta_r})$$

Lemma 4.7. $U = \mathbb{E}[h(X_1, \ldots, X_r)|X_{(1)}, \ldots, X_{(n)}]$.

Proof. HW2. □

4.2.2 Asymptotic Normality of U-Statistics

Theorem 4.8. *Suppose $\mathbb{E}_\theta[U] = \theta = \mathbb{E}_\theta[h(X_1, \ldots, X_r)]$ and $\mathbb{E}_\theta[(h(X_1, \ldots, X_r))^2] < +\infty$. Then,*

(i) The Hájek projection of $U - \theta$ is

$$\hat{U} = \sum_{i=1}^n \mathbb{E}[U - \theta|X_i] = \frac{r}{n} \sum_{i=1}^n h_1(X_i)$$

where $h_1(x) = \mathbb{E}_\theta[h(x, X_2, \ldots, X_r)] - \theta$.

(ii) $\sqrt{n}(U-\theta-\hat{U}) \xrightarrow{\mathbb{P}} 0$ *and* $\sqrt{n}(U-\theta) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0, r^2\zeta_1)$*, where*

$$\zeta_1 = \mathrm{Cov}_\theta(h(X_1, X_2, \ldots, X_r), h(X_1, X_2', \ldots, X_r')), \quad X_1, X_2, \ldots, X_r, X_2', \ldots, X_r' \overset{i.i.d.}{\sim} F$$

Proof. (i) The first equality holds since $\mathbb{E}_\theta[U] = \theta$ and property of conditional expectation that appeared in the proof of Theorem 4.1.5. For the second equality, note that for a given r-tuple $(\beta_1, \ldots, \beta_r)$, $\forall i \in [n]$, by the permutation symmetry of h,

$$\mathbb{E}[h(X_{\beta_1}, \ldots, X_{\beta_r}) - \theta | X_i] = \begin{cases} h_1(X_i) & \text{if } i \in \{\beta_1, \ldots, \beta_r\} \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}[U-\theta|X_i] &= \sum_{i=1}^n \mathbb{E}\left[\frac{1}{\binom{n}{r}} \sum_{1\le\beta_1<\cdots<\beta_r\le n} h(X_{\beta_1}, \ldots, X_{\beta_r}) - \theta \middle| X_i\right] \\ &= \frac{1}{\binom{n}{r}} \sum_{i=1}^n \mathbb{E}\left[\sum_{1\le\beta_1<\cdots<\beta_r\le n} (h(X_{\beta_1}, \ldots, X_{\beta_r}) - \theta) \middle| X_i\right] \\ &= \frac{1}{\binom{n}{r}} \sum_{i=1}^n \sum_{1\le\beta_1<\cdots<\beta_r\le n} h_1(X_i) \mathbb{1}_{i\in\{\beta_1,\ldots,\beta_r\}} \end{aligned}$$

For each $i \in [n]$, the number of tuples is $\binom{n-1}{r-1}$, so the above is equal to

$$\frac{1}{\binom{n}{r}} \sum_{i=1}^n \binom{n-1}{r-1} h_1(X_i) = \frac{r}{n} \sum_{i=1}^n h_1(X_i)$$

(ii) We first show $\sqrt{n}(U-\theta) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0, r^2\zeta_1)$, which is equivalent to $\frac{U-\theta}{\sqrt{\frac{r^2\zeta_1}{n}}} \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0,1)$.

$$\begin{aligned} \mathrm{Var}(\hat{U}) &= \frac{r^2}{n^2} \sum_{i=1}^n \mathrm{Var}[(h_1(X_i))^2] \\ &= \frac{r^2}{n} \mathbb{E}[(h_1(X_1))^2] \quad \text{by i.i.d.} \\ &= \frac{r^2}{n} \mathbb{E}\left[(h(X_1, X_2, \ldots, X_r) - \theta)(h(X_1, X_2', \ldots, X_r'))\right] \\ &= \frac{r^2}{n} \mathrm{Cov}(h(X_1, X_2, \ldots, X_r), h(X_1, X_2', \ldots, X_r')) = \frac{r^2}{n}\zeta_1 \end{aligned}$$

U is the sum of $\binom{n}{r}$ independent random variables $h(X_{\beta_1}, \ldots, X_{\beta_r})$ with r fixed, by central limit theorem, $\frac{U-\theta}{\sqrt{\mathrm{Var}(U)}} \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0,1)$. Hence, if we can show $\frac{\mathrm{Var}(U)}{\mathrm{Var}(\hat{U})} \to 1$, then by Slutsky's lemma,

$$\frac{U-\theta}{\sqrt{\frac{r^2\zeta_1}{n}}} = \sqrt{\frac{\mathrm{Var}(U)}{\mathrm{Var}(\hat{U})}} \cdot \frac{U-\theta}{\sqrt{\mathrm{Var}(U)}} \overset{\mathcal{D}}{\rightsquigarrow} 1 \cdot \mathcal{N}(0,1) = \mathcal{N}(0,1)$$

We now calculate $\mathrm{Var}(U)$:

$$
\begin{aligned}
\mathrm{Var}(U) &= \mathrm{Var}\left(\frac{1}{\binom{n}{r}} \sum_{1\le\beta_1<\cdots<\beta_r\le n} h(X_{\beta_1},\ldots,X_{\beta_r})\right) \\
&= \frac{1}{\binom{n}{r}^2}\sum_{\beta}\sum_{\beta'} \mathrm{Cov}(h(X_{\beta_1},\ldots,X_{\beta_r}), h(X_{\beta'_1},\ldots,X_{\beta'_r})) \\
&= \frac{1}{\binom{n}{r}^2}\sum_{c=0}^{r}\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\underbrace{\mathrm{Cov}(h(X_1,\ldots,X_c,X_{c+1},\ldots,X_r), h(X_1,\ldots,X_c,X'_{c+1},\ldots,X'_r))}_{:=\zeta_c} \\
&= \sum_{i=1}^{r}\frac{r!}{c!(r-c)!}\frac{(n-r)!}{(r-c)!(n-2r+c)!}\frac{r!(n-r)!}{n!}\zeta_c
\end{aligned}
$$

Comment on summation: we have $\binom{n}{r}$ many r-tuples. For each r-tuple β, we choose the other r-tuple β'. If β and β' are disjoint, then the summand is zero. If they overlap with c elements ($\binom{r}{c}$ many of such), we need to choose the other $r-c$ elements in β' from the remaining $n-r$ elements. Also, ζ_c's are consistent with the ζ_1 in the statement of the theorem, and they are bounded.

The first term of the summand is then

$$
\frac{r!^2(n-r)!^2}{(r-1)!^2(n-2r+1)!n!}\zeta_1 = r^2\frac{\overbrace{(n-r-(r-2))\cdots(n-r)}^{r-1\text{ terms}}}{\underbrace{n(n-1)\cdots(n-r+1)}_{n\text{ terms}}}\zeta_1 = (1+o(1))\frac{r^2}{n}\zeta_1
$$

Similarly, for the second term,

$$
\frac{r!^2(n-r)!^2}{2(r-2)!^2(n-2r+2)!n!}\zeta_2 = \frac{r^2(r-1)^2\zeta_2}{2}\frac{(n-r-(r-3))\cdots(n-r)}{n(n-1)\cdots(n-r+1)} = \mathcal{O}\left(\frac{1}{n^2}\right)
$$

Later terms are also in the form of $\mathcal{O}\left(\frac{1}{n^c}\right)$, so $\mathrm{Var}(U) = (1+o(1))\,\mathrm{Var}(\hat{U})) + \mathcal{O}\left(\frac{1}{n^2}\right)$, which implies that $\frac{\mathrm{Var}(U)}{\mathrm{Var}(\hat{U})} \to 1$.

With this, by the Co-Convergence Theorem, we have $\frac{U-\theta}{\sqrt{\mathrm{Var}(U)}} - \frac{\hat{U}}{\sqrt{\mathrm{Var}(\hat{U})}} \xrightarrow{\mathbb{P}} 0$.

$$
\begin{aligned}
&\frac{U-\theta}{\sqrt{\mathrm{Var}(U)}} - \frac{\hat{U}}{\sqrt{\mathrm{Var}(\hat{U})}} \\
=&\frac{\sqrt{\frac{r^2\zeta_1}{n}}(U-\theta) - \sqrt{(1+o(1))\frac{r^2\zeta_1}{n} + \mathcal{O}(\frac{1}{n^2})}\hat{U}}{\sqrt{(1+o(1))\frac{r^4\zeta_1^2}{n^2} + \mathcal{O}(\frac{1}{n^3})}} \\
=&\frac{\sqrt{n}(U-\theta-\hat{U})}{\sqrt{1+o(1)+\mathcal{O}(\frac{1}{n})}} + \underbrace{\frac{\sqrt{\frac{r^2\zeta_1}{n}} - \sqrt{(1+o(1))\frac{r^2\zeta_1}{n} + \mathcal{O}(\frac{1}{n^2})}}{\sqrt{(1+o(1))\frac{r^4\zeta_1^2}{n^2} + \mathcal{O}(\frac{1}{n^3})}}}_{\to 0\ p.w.}\hat{U} \xrightarrow{\mathbb{P}} 0 \\
\Rightarrow&\sqrt{n}(U-\theta-\hat{U}) \xrightarrow{\mathbb{P}} 0
\end{aligned}
$$

□

4.2.3 Example (Kendall's τ)

$\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d. pairs.

Let $\tau := \frac{4}{n(n-1)} \sum_{1\le i<j\le n} \mathbb{1}_{\{(X_i-X_j)(Y_i-Y_j)\}} - 1$.

so $\tau + 1$ is a U-statistic with kernel

$$h\left(\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix}\right) = 2\mathbb{1}_{\{(X_1-X_2)(Y_1-Y_2)\}}$$

Let $\theta = \mathbb{E}[h] = \mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0)$.

By the theorem, we have $\sqrt{n}(\tau - \mathbb{E}[\tau]) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0, 4\zeta_1)$, where $\zeta_1 = \mathrm{Cov}\left(h\left(\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix}\right), h\left(\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \begin{bmatrix} X_3 \\ Y_3 \end{bmatrix}\right)\right)$.

Consider the null case, where $X_i \perp\!\!\!\perp Y_i$ within each pair. Then

$$\begin{aligned}
\mathbb{E}\left[h\left(\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix}\right)\right] &= 2\mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0) \\
&= 2[\mathbb{P}(X_1 > X_2)\mathbb{P}(Y_1 > Y_2) + \mathbb{P}(X_1 < X_2)\mathbb{P}(Y_1 < Y_2)] \\
&= 2 \times (\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}) = 1 \\
\mathbb{E}[\tau] &= \frac{4}{n(n-1)}\binom{n}{2}\mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0) - 1 \\
&= \frac{4}{n(n-1)}\frac{n(n-1)}{2}\frac{1}{2} - 1 = 0 \\
\zeta_1 &= \mathrm{Cov}\left(h\left(\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix}\right), h\left(\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \begin{bmatrix} X_3 \\ Y_3 \end{bmatrix}\right)\right) \\
&= 4\mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0 \text{ and } (X_1 - X_3)(Y_1 - Y_3) > 0) - 1 \\
&= 8[\mathbb{P}(X_1 > X_2, Y_1 > Y_2 \text{ and } X_1 > X_3, Y_1 > Y_3) + \\
&\quad\ \mathbb{P}(X_1 > X_2, Y_1 > Y_2 \text{ and } X_1 < X_3, Y_1 < Y_3)] - 1 \\
&= 8(\mathbb{P}(X_1 > X_2 \text{ and } X_1 > X_3)^2 + \mathbb{P}(X_3 > X_1 > X_2)^2) - 1 \quad \text{since } X_i \perp\!\!\!\perp Y_i
\end{aligned}$$

Since this is only about ordering, we can think of them as i.i.d. $U(0, 1)$.

$$\begin{aligned}
\mathbb{P}(X_1 > X_2 \text{ and } X_1 > X_3) &= \int_0^1 \mathbb{P}(X_3 < U)\mathbb{P}(X_2 < u)du = \int_0^1 u^2 du = \frac{1}{3} \\
\mathbb{P}(X_3 > X_1 > X_2) &= \int_0^1 \mathbb{P}(X_3 > u)\mathbb{P}(X_2 < u)du = \int_0^1 u(1-u)du = \frac{1}{6}
\end{aligned}$$

so $\zeta_1 = 8(\frac{1}{9} + \frac{1}{36}) - 1 = \frac{1}{9}$, and $\sqrt{n}\tau \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0, \frac{4}{9})$.

4.3 Two-sample U-Statistics

4.3.1 Definition

Definition 4.9 (Two-sample U-Statistics)**.** $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ *independent of* $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} G$*. Let* $h(x_1, \ldots, x_r, y_1, \ldots, y_s)$ *be seperately permutation invariant with respect to the first* r *arguments and the last* s *arguments. Then*

$$U = \frac{1}{\binom{n}{r}\binom{m}{s}} \sum_{1\le\alpha_1<\ldots<\alpha_r\le n} \sum_{1\le\beta_1<\ldots<\beta_s\le m} h(X_{\alpha_1}, \ldots, X_{\alpha_r}, Y_{\beta_1}, \ldots, Y_{\beta_s})$$

is a two-sample U-Statistics with kernel h.

4.3.2 Hájek Projection

Definition 4.10. *Let $\theta = \mathbb{E}[h(X_1, \ldots, X_r, Y_1, \ldots, Y_s)]$. The Hájek projection of $U - \theta$ is then*

$$\hat{U} = \frac{r}{n} \sum_{i=1}^{n} h_{1,0}(X_i) + \frac{s}{m} \sum_{j=1}^{m} h_{0,1}(Y_j)$$

where

$$h_{1,0}(x) = \mathbb{E}[h(x, X_2, \ldots, X_r, Y_1, \ldots, Y_s)]$$
$$h_{0,1}(y) = \mathbb{E}[h(X_1, \ldots, X_r, y, Y_2, \ldots, Y_s)]$$

4.3.3 Theorem

Theorem 4.11. *Suppose that $N = n + m$, $\frac{n}{N} \to \lambda \in (0, 1)$ and $\mathbb{E}[(h(X_1, \ldots, X_r, Y_1, \ldots, Y_s))^2] < \infty$. Then $\sqrt{N}(U - \theta - \hat{U}) \xrightarrow{\mathbb{P}} 0$ and*

$$\sqrt{N}(U - \theta - \hat{U}) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}\left(0, \frac{r^2 \zeta_{1,0}}{\lambda} + \frac{s^2 \zeta_{0,1}}{1 - \lambda}\right)$$

where

$$\zeta_{1,0} = \mathrm{Cov}_\theta(h(X_1, X_2, \ldots, X_r, Y_1, Y_2, \ldots, Y_s), h(X_1, X_2', \ldots, X_r', Y_1', Y_2', \ldots, Y_s'))$$
$$\zeta_{0,1} = \mathrm{Cov}_\theta(h(X_1, X_2, \ldots, X_r, Y_1, Y_2, \ldots, Y_s), h(X_1', X_2', \ldots, X_r', Y_1, Y_2', \ldots, Y_s'))$$

and the X's and Y's are all independent.

5 Asymptotic Relative Efficiency of Tests

5.1 Asymptotic Power Function

We have a hypothesis testing problem $H_0 : \theta \in \Theta_0$ v.s. $H_1 : \theta \in \Theta_1$. The test is that we reject H_0 when the test statistic $T_n \in K_n$, the rejection region.

Power function is defined as $\pi_n(\theta) := \mathbb{P}_\theta(T_n \in K_n)$, where $\theta \in \Theta = \Theta_0 \cup \Theta_1$.

5.1.1 Asymptotic Level-α Test

Definition 5.1. *A test is asymptotic level-α test if* $\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} \pi_n(\theta) \le \alpha$.

This means that you have control of Type-I error.

5.1.2 Consistency

Definition 5.2. *Given a sequence of tests that is asymptotically level-α for some pre-specified $\alpha \in (0, 1)$, it is consistent if* $\lim_{n\to\infty} \pi_n(\theta) = 1$ *for all $\theta \in \Theta_1$.*

Being consistent is a rather weak property. There are many consistent tests for the same problem.

5.1.3 Local Alternative Sequence

Pick a sequence $\{\theta_n\} \subseteq \Theta_1$ such that $d(\theta_n, \Theta_0) := \psi_n \to 0$ as $n \to \infty$. Our interest is in finding the fastest vanishing rate for ψ_n such that $\lim_{n\to\infty} \pi_n(\theta_n) > \alpha$.

If $\Theta_0 = \{\theta_0\}$, and the parametric family of distributions is q.m.d., then $\psi_n \asymp \frac{1}{\sqrt{n}}$. So, to compare different sequences of tests, it reduces to compare how "soon" a desired power can be achieved along a sequence of local alternative at a given level α.

5.1.4 Slope of Tests with Asymptotic Normal Test Statistics

$H_0 : \theta = 0$ v.s. $H_1 : \theta = \frac{h}{\sqrt{n}}$

Let $\pi(h) = \lim_{n\to\infty} \pi_n(\frac{h}{\sqrt{n}})$, assuming that the limit exists.

Theorem 5.3. *Let μ, σ be functions of θ such that:*

for any h, define $\theta_n = \frac{h}{\sqrt{n}}$, under $\mathbb{P}^n_{\theta_n}$, we have $\dfrac{\sqrt{n}(T_n - \mu(\theta_n))}{\sigma(\theta_n)} \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0, 1)$

If μ is differentiable at 0 and σ is continuous at 0, then the asymptotic level-α tests that reject H_0 for large values of T_n satisfies

$$\pi(h) = 1 - \Phi\big(z_\alpha - h\frac{\mu'(0)}{\sigma(0)}\big)$$

where $\frac{\mu'(0)}{\sigma(0)}$ is called the slope of the test.

Proof. Under H_0, we have $\frac{\sqrt{n}(T_n-\mu(0))}{\sigma(0)} \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0,1)$. So an asymptotic level-α test is equivalent to rejecting H_0 when $\sqrt{n}(T_n - \mu(0)) > \sigma(0)z_\alpha + o_{\mathbb{P}}(1)$. Hence,

$$\begin{aligned}
\pi_n\left(\frac{h}{\sqrt{n}}\right) &= \mathbb{P}_{\frac{h}{\sqrt{n}}}(\sqrt{n}(T_n - \mu(0)) > \sigma(0)z_\alpha + o_{\mathbb{P}}(1)) \\
&= \mathbb{P}_{\frac{h}{\sqrt{n}}}\left(\frac{\sqrt{n}(T_n - \mu(\frac{h}{\sqrt{n})}))}{\sigma(\frac{h}{\sqrt{n}})} > \frac{\sigma(0)}{\sigma(\frac{h}{\sqrt{n}})}z_\alpha + \frac{\mu(0)-\mu(\frac{h}{\sqrt{n}})}{\sigma(\frac{h}{\sqrt{n}})/\sqrt{n}} + o_{\mathbb{P}}(1)\right) \\
&\longrightarrow 1 - \Phi\big(z_\alpha - h\frac{\mu'(0)}{\sigma(0)}\big)
\end{aligned}$$

□

The larger slope, the better test we have.

5.1.5 Example

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x-\theta)$, where f is symmetric about 0.

Test #1: $T_n = \frac{\bar{X}}{S}$, where S is sample standard deviation.

$\sqrt{n}(T_n - \frac{h}{\sqrt{n}}/\sigma) \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0,1)$. $\mu'(0) = \frac{1}{\sigma}$ and $\sigma(0) = 1$, which implies that slope is equal to $\frac{1}{\sigma}$, where $\sigma^2 = \int x^2 f(x)dx$.

Test #2: $S_n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i>0\}}$

Then, $\mu_\theta = \mathbb{P}_\theta(X_1 > 0) = \mathbb{P}_0(X_1 > -\theta) = \mathbb{P}(X_1 < \theta) = F(\theta)$ by symmetry. $\sigma^2(\theta) = F(\theta)(1 - F(\theta))$, which implies that $\mu'(0) = f(0), \sigma(0) = \frac{1}{2}$. So slope is equal to $2f(0)$.

We provide the following two concrete examples:

(a) $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$.

Slope of #1 is 1 and slope of #2 is $2\frac{1}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}$.

(b) $f(x) = \frac{1}{2}e^{-|x|}$.

Slope of #1 is $\frac{1}{\sqrt{2}}$ and slope of #2 is 1.

5.2 Pitman Asymptotic Relative Efficiency

5.2.1 Setup

$H_0 : \theta = 0$ v.s. $H_1 : \theta = \theta_\nu$, where ν is the "time" index, and $\theta_\nu \to 0$ as $\nu \to \infty$.

We require our tests to attain asymptotic level α and asymptotic power $\gamma \in (\alpha, 1)$, where $\gamma = \lim_{\nu\to\infty} \pi_{n_\nu}(\theta_\nu)$

We see two sequences of tests, the better test achieves γ sooner than the worse one.

There are three moving parts: the sequence of alternative hypotheses, the level α, and power γ. They characterize a test with asymptotic behavior.

Now we fix α and γ. As a result, we just have the sequence of alternative hypotheses moving (which sends θ_ν to the null hypothesis), then we compare how large $n_{\nu,1}$ and $n_{\nu,2}$ are required for two sequences of tests to achieve α and γ, respectively.

5.2.2 Definition

Definition 5.4 (Pitman Asymptotic Relative Efficiency). *Suppose at least $n_{\nu,1}$ and $n_{\nu,2}$ observations are needed for two sequences of tests to achieve asymptotic level-α and asymptotic power γ along the same sequence of θ_ν. Then if exists, $\lim_{\nu\to\infty} \frac{n_{\nu,2}}{n_{\nu,1}}$ is the Pitman asymptotic relative efficiency of test sequence 1 over test sequence 2. If this limit is greater than 1, then sequence 1 is better.*

5.2.3 Theorem

Theorem 5.5 (Piteman ARE & Slopes of Tests). *Consider statistical models $\{\mathbb{P}_{n,\theta} : \theta \geq 0\}$ such that*

$$||\mathbb{P}_{n,\theta} - \mathbb{P}_{n,0}|| = \int |p_{n,\theta} - p_{n,0}| d\mu \to 0$$

as $\theta \to 0$ at each fixed n (so that the alternative hypothesis H_1 can go to null).

Let $T_{n,i}, i \in \{1, 2\}$ be two sequences of test statistics such that

$$(*) \qquad \frac{T_{n,i} - \mu_i(\theta_n)}{\sigma_i(\theta_n)} \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0, 1) \quad \text{under } \mathbb{P}_{n,\theta_n}$$

for every sequence of $\theta_n \to 0$ as $n \to \infty$, where $\mu_i'(0) > 0$ and $\sigma_i(0) > 0$. Then,

(i) the Pitman ARE of tests that reject H_0 for large values of $T_{n,i} (i \in \{1, 2\})$ is equal to

$$\left[\frac{\mu_1'(0)/\sigma_1(0)}{\mu_2'(0)/\sigma_2(0)}\right]^2$$

which is independent of $\alpha > 0$ and $\gamma \in (\alpha, 1)$ as well as θ_n.

(ii) If power of $T_{n,i}$-based tests are monotone in θ at each fixed n, then it suffices to have $(*)$ for $\theta_n = \frac{h}{\sqrt{n}}$ for all fixed $h > 0$.

Proof. Step 1:

Fix α, γ, and $\theta_\nu \to 0$. Let $n_{1,\nu}$ and $n_{2,\nu}$ be our sample sizes for level-α tests to achieve power γ at θ_ν. Now we show that for any $i \in \{1, 2\}$, $n_{\nu,i} \to \infty$ as $\nu \to \infty$.

Fix n, by assumption, we know $||\mathbb{P}_{n,\theta_\nu} - \mathbb{P}_{n,0}|| \to 0$ as $\nu \to \infty$.

The sum of type I and type II error probability of a test that has rejection region K_n is

$$\begin{aligned}
&\int_{K_n} d\mathbb{P}_{n,0} + \int_{K_n^c} d\mathbb{P}_{n,\theta_\nu} \\
=&1 + \int_{K_n} (p_{n,0} - p_{n,\theta_\nu}) d\mu \\
\geq&1 + \int_{\{p_{n,0} < p_{n,\theta_\nu}\}} (p_{n,0} - p_{n,\theta_\nu}) d\mu \\
=&1 - \frac{1}{2}||\mathbb{P}_{n,\theta_\nu} - \mathbb{P}_{n,0}|| \to 1 \quad \text{as } \nu \to \infty
\end{aligned}$$

But type I and type II error should be $\alpha + 1 - \gamma < \alpha + (1-\alpha) = 1$, so we need $n_{\nu,i} = +\infty$ in order to avoid contradiction. To get the last line above, we can let

$$A = \int_{\{p_{n,0} < p_{n,\theta_\nu}\}} (p_{n,0} - p_{n,\theta_\nu}) d\mu, B = \int_{\{p_{n,0} \geq p_{n,\theta_\nu}\}} (p_{n,0} - p_{n,\theta_\nu}) d\mu$$

and solve from $A + B = 0$ and $B - A = ||\mathbb{P}_{n,\theta_\nu} - \mathbb{P}_{n,0}||$.

Step 2:

Under hypothesis $\mathbb{P}_{n_{\nu,i},0}$, as $\nu \to \infty$,

$$\sqrt{n_{\nu,i}} \frac{T_{n_{\nu,i}} - \mu_i(0)}{\sigma_i(0)} \overset{\mathcal{D}}{\rightsquigarrow} \mathcal{N}(0,1)$$

and we reject H_0 if

$$\sqrt{n_{\nu,i}}(T_{n_{\nu,i}} - \mu_i(0)) > \sigma_i(0) z_\alpha + o_{\mathbb{P}_{n_{\nu,i},0}}(1)$$

so the power for each i is

$$\pi_{n_{\nu,i}}(\theta_\nu) = 1 - \Phi\left(z_\alpha - \sqrt{n_{\nu,i}}\theta_\nu \frac{\mu_i'(0)}{\sigma_i(0)}\right) + o(1) \qquad \cdots (*)$$

which converges to γ by assumption. Hence, we have

$$z_\alpha - \sqrt{n_{\nu,i}}\theta_\nu \frac{\mu_i'(0)}{\sigma_i(0)} \to z_\gamma$$

so $\sqrt{n_{\nu,i}}\theta_\nu \frac{\mu_i'(0)}{\sigma_i(0)} \to z_\alpha - z_\gamma$ as $\nu \to \infty$. Thus,

$$\frac{n_{\nu,2}}{n_{\nu,1}} \to \left(\frac{\mu_1'(0)/\sigma_1(0)}{\mu_2'(0)/\sigma_2(0)}\right)^2$$

Step 3:

Power functions are monotonously increasing in θ. At each fixed sample size n, $(*)$ holds for $\theta_n = \frac{h}{\sqrt{n}}$, $\forall h$.

$$\pi_{n,i}(\theta_n) \to \alpha \text{ if } \sqrt{n}\theta_n \to 0$$
$$\pi_{n,i}(\theta_n) \to 1 \text{ if } \sqrt{n}\theta_n \to \infty$$

Hence, everything interesting should happen at rate $\frac{1}{\sqrt{n}}$. □

5.3 Bahadur Asymptotic Relative Efficiency

In general, we have three parameters to vary: the sequence of alternative values θ, significance level α, and power γ. Now we fix θ and γ, and let $\alpha \downarrow 0$ as $n \uparrow \infty$.

5.3.1 Definition

Definition 5.6 (Bahadur Asymptotic Relative Efficiency)**.** *Bahadur Asymptotic Relative Efficiency is defined as*

$$\lim_{\nu \to \infty} \frac{n_2(\alpha_\nu, \gamma, \theta)}{n_1(\alpha_\nu, \gamma, \theta)}$$

where $\alpha \downarrow 0$ as $\nu \uparrow \infty$, and γ and θ are fixed. The testing setup is simple v.s. simple.

The reason we want to study this is that we could have made tail behavior bad while maintaining other certain desired behavior.

5.3.2 Preliminaries

For each n, a test statistic T_n is used for the hypothesis testing $H_0 : \theta = 0$ v.s. $H_1 : \theta = \theta_\nu > 0$.

Suppose that we reject H_0 for some large values of T_n and tat

(i) $-\frac{2}{n} \log \mathbb{P}_0(T_n \geq t) \to e(t)$ for all $t \geq 0$ for some function $e(\cdot)$. This function e is about tail behavior.

(ii) $T_n \to \mu(\theta)$ under $\mathbb{P}_\theta$.

5.3.3 Theorem

Theorem 5.7. *Suppose you have two sequences of statistical models $\{\mathbb{P}_{n,0}\}$ and $\{\mathbb{P}_{n,\theta}\}$, with two sequences of test statistics $\{T_{n,i}\}$ $(i \in \{1, 2\})$ that satisfy conditions (i) and (ii) above with functions $e_i(\cdot), \mu_i(\cdot)$ $(i \in \{1, 2\})$, e_i being continuous at $\mu_i(\theta)$.*

Then the Bahadur Asymptotic Relative Efficiency of tests that reject larger values of $T_{n,i}$'s is equal to

$$\frac{e_1(\mu_1(\theta))}{e_2(\mu_2(\theta))}$$

for every $\alpha_\nu \downarrow 0$ and every $1 > \gamma > \sup_n \mathbb{P}_\theta(p_{n,0} = 0)$.

Discussing the set $\{p_{n,0} = 0\}$ is trivial because you can always reject the null hypothesis on it.

Proof. Step 1:

To show that $n_{\nu,i} \uparrow \infty$ as $\nu \uparrow \infty$, suppose that n_ν stays finite. Then there exists an $n \in \mathbb{N}$ fixed and a sequence of tests with levels going down to 0 and powers at least γ.

Index this sequence by $m = m_\nu$, each with rejection region K_m. So $\alpha_m = \mathbb{P}_0(K_m) \to 0$ as $m \to \infty$.

$$\begin{aligned}
\gamma = \mathbb{P}_{n,\theta}(K_m) &= \mathbb{P}_{n,\theta}(K_m \cap \{p_{n,0} = 0\}) + \mathbb{P}_{n,\theta}(K_m \cap \{p_{n,0} \neq 0\}) \\
&\leq \mathbb{P}_{n,\theta}(\{p_{n,0}0\}) + o_m(1) \\
&\leq \sup_n \mathbb{P}_{n,\theta}(\{p_{n,0} = 0\})
\end{aligned}$$

which is a contradiction. The reason why the second term is $o_m(1)$ is the following:

$$\begin{aligned}
\mathbb{P}_{n,\theta}(K_m \cap \{p_{n,0} \neq 0\}) &= \int_{K_m \cap \{p_{n,0} \neq 0\}} p_{n,\theta} d\mu \\
&= \int_{K_m \cap \{p_{n,0} \neq 0\}} \frac{p_{n,\theta}}{p_{n,0}} p_{n,0} d\mu \\
&= \mathbb{E}_{n,0}\left[\mathbb{1}_{K_m} \mathbb{1}_{\{p_{n,0} \neq 0\}} \frac{p_{n,\theta}}{p_{n,0}}\right]
\end{aligned}$$

As the significance level approaches 0 as m approaches infinity, under $\mathbb{P}_{n,0}$, $\mathbb{E}_{n,0}[\mathbb{1}_{K_m}] \xrightarrow{\mathbb{P}} 0$. So this term is of $o_m(1)$.

Step 2:

Since we reject when the test statistics are large, there exists $c_{n,i}$ for each $n \in \mathbb{N}$ and $i \in \{1, 2\}$ such that the rejection region is $\{T_{n,i} \geq c_{n,i}\}$. Define the corresponding p-values as

$$L_{n,i} = \mathbb{P}_{n,0}(T_{n,i} \geq t_{n,i})$$

where $t_{n,i}$'s are the observed values of those testing statistics. Hence, we reject if $L_{n,i} \leq \alpha_\nu$.

By the definition of $n_{\nu,i}$, we have

$$\mathbb{P}_{n,\theta}(-\frac{2}{n}\log L_{n,i} \geq -\frac{2}{n}\log \alpha_\nu) \begin{cases} \geq \gamma & \text{if } n \geq n_{\nu,i} \\ < \gamma & \text{if } n < n_{\nu,i} \end{cases}$$

Meanwhile,

$$\begin{aligned} -\frac{2}{n}\log L_{n,i} &= -\frac{2}{n}\log \mathbb{P}_{n,0}(T_{n,i} \geq t_{n,i}) \quad \text{by definition} \\ &= -\frac{2}{n}\log \mathbb{P}_{n,0}(T_{n,i} \geq \mu_i(\theta) + o_{\mathbb{P}_{n,\theta}}(1)) \quad \text{Preliminary (ii)} \\ &\xrightarrow{\mathbb{P}} e_i(\mu_i(\theta)) \quad \text{ under } \mathbb{P}_{n,\theta} \text{ by Preliminary (i) and continuity of } e(\cdot) \end{aligned}$$

So, since $\gamma > 0$,

$$\lim_{\nu\to\infty} -\frac{2}{n_{\nu,i}}\log \alpha_\nu \leq e_i(\mu_i(\theta))$$

and since $\gamma < 1$,

$$\lim_{\nu\to\infty} -\frac{2}{n_{\nu,i}-1}\log \alpha_\nu \geq e_i(\mu_i(\theta))$$

These imply that $n_{\nu,i} = (1 + o(1))\left[-2\frac{\log \alpha_\nu}{e_i(\mu_i(\theta))}\right]$ $\square$

5.3.4 Cramér–Chernoff Theorem

Let $Y_1, \ldots, Y_n$ be i.i.d. copies of Y, and let $K(u) = \log \mathbb{E}(e^{uY})$ be its cumulant generating function. For $t > \mathbb{E}Y$, the Cramér–Chernoff theorem gives

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Y_i \geq t\right) = \inf_{u>0}\{K(u) - tu\}.$$

Equivalently, if $K'(u_t) = t$, then the Bahadur exponent is

$$e(t) = 2\{tu_t - K(u_t)\}.$$

5.3.5 Likelihood Ratio Test

Consider $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. The likelihood-ratio test rejects for large values of

$$T_n = \frac{1}{n}\sum_{i=1}^n \log \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}.$$

Under H_0, for $Y = \log(p_{\theta_1}(X)/p_{\theta_0}(X))$,

$$K(u) = \log \int \left(\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}\right)^u p_{\theta_0}(x)\, d\mu(x).$$

Furthermore,

$$K'(1) = \mathbb{E}_{\theta_1} \log \frac{p_{\theta_1}(X)}{p_{\theta_0}(X)} = \mu(\theta_1) = D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_0}), \qquad K(1) = 0.$$

Thus $u = 1$ optimizes the exponent at $t = \mu(\theta_1)$, and

$$e(\mu(\theta_1)) = 2\mu(\theta_1) = 2D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_0}).$$

5.4 Rescaling the Power Function

5.4.1 Definition

For $H_0 : \theta = 0$ versus $H_1 : \theta = h/\sqrt{n}$, define

$$\Pi(h) = \lim_{n\to\infty} \Pi_n\left(\frac{h}{\sqrt{n}}\right),$$

whenever this limit exists.

5.4.2 Lemma

For probability measures P, Q with densities p, q with respect to μ, define

$$\|P - Q\| = \int |p - q|\, d\mu \in [0, 2].$$

Lemma 5.8 (Testing error and L_1 distance)**.** *For every test ϕ,*

$$\Pi(\theta) - \Pi(\theta_0) \leq \frac{1}{2}\|P_\theta - P_{\theta_0}\|.$$

Equality is attained pointwise by $\phi = \mathbb{1}_{\{p_\theta \geq p_{\theta_0}\}}$.

Proof.

$$\begin{aligned}\Pi(\theta) - \Pi(\theta_0) &= \int \phi(p_\theta - p_{\theta_0})\, d\mu \\ &\leq \int_{\{p_\theta \geq p_{\theta_0}\}} (p_\theta - p_{\theta_0})\, d\mu \\ &= \frac{1}{2}\int |p_\theta - p_{\theta_0}|\, d\mu.\end{aligned}$$

□

In the classical setting

$$H_0 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta_0}, \qquad H_1 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta_n},$$

there are three regimes:

(i) $\|P^n_{\theta_0} - P^n_{\theta_n}\| \to 2$: easy (power $\to 1$ and level $\to 0$);

(ii) $\|P^n_{\theta_0} - P^n_{\theta_n}\| \to 0$: hard;

(iii) the distance stays in a compact subset of $(0, 2)$: intermediate.

5.4.3 Hellinger Distance

The Hellinger distance, convenient for independent samples, is

$$H(P,Q) = \left(\int (\sqrt{p} - \sqrt{q})^2 \, d\mu\right)^{1/2}.$$

Writing $A(P,Q) = \int \sqrt{pq}\, d\mu$ for the Hellinger affinity,

$$H^2(P,Q) = 2 - 2A(P,Q).$$

For product measures,

$$A(P^n, Q^n) = [A(P,Q)]^n,$$

and hence

$$H^2(P^n, Q^n) = 2 - 2[A(P,Q)]^n = 2 - 2\left[1 - \frac{1}{2}H^2(P,Q)\right]^n.$$

5.4.4 Lemma

Lemma 5.9 (Connection with L_1 distance)**.**

$$H^2(P,Q) \le \|P - Q\| \le \min\{2H(P,Q),\, 2 - A^2(P,Q)\}.$$

Thus

$$H^2(P^n, Q^n) \le \|P^n - Q^n\| \le 2H(P^n, Q^n).$$

Consequently:

(i) if $nH^2(P,Q) \to 0$, then $\|P^n - Q^n\| \to 0$;

(ii) if $nH^2(P,Q) \to \infty$, then $\|P^n - Q^n\| \to 2$;

(iii) if $nH^2(P,Q) \to c \in (0,\infty)$, then

$$H^2(P^n, Q^n) \to 2 - 2e^{-c/2} \in (0,2).$$

Proof. The first inequality follows from

$$\int (\sqrt{p} - \sqrt{q})^2 \, d\mu \le \int |p - q| \, d\mu.$$

Since $pq \le (p+q)\min\{p,q\}$, Cauchy–Schwarz gives

$$A^2(P,Q) \le \left(\int (p+q)\, d\mu\right)\left(\int \min\{p,q\}\, d\mu\right) = 2\left(1 - \frac{1}{2}\|P - Q\|\right).$$

Finally,

$$\begin{aligned}\|P - Q\| &= \int |\sqrt{p} - \sqrt{q}|(\sqrt{p} + \sqrt{q})\, d\mu \\ &\le H(P,Q)\left(\int (\sqrt{p} + \sqrt{q})^2 \, d\mu\right)^{1/2} \le 2H(P,Q).\end{aligned}$$

□

5.4.5 Example (parametric)

Consider
$$H_0 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0,1), \qquad H_1 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta_n, 1).$$
For $P = N(0,1)$ and $Q = N(\theta_n, 1)$,

$$A(P,Q) = e^{-\theta_n^2/8}, \qquad nH^2(P,Q) = 2n(1 - e^{-\theta_n^2/8}) = \frac{n\theta_n^2}{4}(1 + o(1)) \asymp n\theta_n^2.$$

Thus the nontrivial testing rate is $\theta_n \asymp n^{-1/2}$.

5.4.6 Example (nonparametric)

Suppose
$$Y_i = f(i/n) + \sigma Z_i, \qquad i = 1, \ldots, n,$$
where $Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and f is L-Lipschitz: $|f(x) - f(y)| \leq L|x - y|$. Consider

$$H_0 : f(1/2) = 0 \qquad \text{versus} \qquad H_1 : f(1/2) \geq \theta_n.$$

The least favorable alternative is a triangular bump centered at $1/2$, of height θ_n and width of order θ_n/L. Hence roughly $n\theta_n$ observations have a mean difference of order θ_n, so

$$A\left(\bigotimes_{i=1}^n P_i, \bigotimes_{i=1}^n Q_i\right) \asymp e^{-Cn\theta_n^3}, \qquad H^2\left(\bigotimes_{i=1}^n P_i, \bigotimes_{i=1}^n Q_i\right) \asymp 1 - e^{-Cn\theta_n^3}.$$

The intermediate regime requires $n\theta_n^3 \asymp 1$, giving
$$\theta_n \asymp n^{-1/3}.$$

5.5 Testing in a q.m.d. Family

5.5.1 Abstract Theorem

Let $\mathcal{E}_n = \{P_{n,h} : h \in \mathcal{H}\}$ converge to the limit experiment $\mathcal{E} = \{P_h : h \in \mathcal{H}\}$. If power functions Π_n of tests in $\mathcal{E}_n$ converge pointwise, $\Pi_n(h) \to \Pi(h)$ for every h, then Π is a power function in the limit experiment. This follows from LAN.

5.5.2 Concrete q.m.d. Theorem

Theorem 5.10. *Suppose $\{P_\theta : \theta \in \Theta\}$ is q.m.d. at $\theta_0 \in \mathring{\Theta}$, with nonsingular Fisher information I_{θ_0}. Let $\psi : \Theta \to \mathbb{R}$ be differentiable at θ_0, with nonzero gradient $\dot\psi_{\theta_0}$ and $\psi(\theta_0) = 0$. For the power function Π_n of any sequence of level-α tests of*

$$H_0 : \psi(\theta) \leq 0 \qquad \textit{versus} \qquad H_1 : \psi(\theta) > 0,$$

and every h such that $\dot\psi_{\theta_0}^T h > 0$,

$$\limsup_{n\to\infty} \Pi_n\left(\theta_0 + \frac{h}{\sqrt{n}}\right) \leq 1 - \Phi\left(z_\alpha - \frac{\dot\psi_{\theta_0}^T h}{\sqrt{\dot\psi_{\theta_0}^T I_{\theta_0}^{-1} \dot\psi_{\theta_0}}}\right).$$

Tests that reject for large values of

$$T_n = \frac{\dot{\psi}_{\theta_0}^T I_{\theta_0}^{-1} \Delta_{n,\theta_0}}{\sqrt{\dot{\psi}_{\theta_0}^T I_{\theta_0}^{-1} \dot{\psi}_{\theta_0}}} + o_{P_{\theta_0}}(1), \qquad \Delta_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i),$$

attain the bound; these are score tests.

6 Empirical Processes

6.1 Stochastic Convergence in Metric Spaces

6.1.1 Motivation

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$, where F is continuous, and let

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}$$

be the empirical distribution function. For every fixed x, the strong law of large numbers gives $\widehat{F}_n(x) \to F(x)$ almost surely. We are often interested in uniform statements such as

$$\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \to 0 \quad \text{a.s.}$$

More generally, for a class $\mathcal{G}$ of functionals on distribution functions, we study

$$\sup_{g \in \mathcal{G}} |g(\widehat{F}_n) - g(F)|.$$

The relevant measurability and convergence questions depend on the class $\mathcal{G}$.

At a fixed x, the central limit theorem yields

$$\sqrt{n}\{\widehat{F}_n(x) - F(x)\} \overset{\mathcal{D}}{\rightsquigarrow} N(0, F(x)\{1 - F(x)\}).$$

To formulate a corresponding result uniformly in x, one must regard the empirical process as a random element in a function space.

Brownian motion illustrates the same issue. If $B_n(t) = n^{-1/2} W(nt)$, then its finite-dimensional distributions agree with those of Brownian motion, whose increments satisfy

$$B(t + \Delta t) - B(t) \sim N(0, \Delta t), \qquad B(t) - B(s) \perp\!\!\!\perp B(v) - B(u)$$

for disjoint time intervals. Convergence of finitely many coordinates alone is not enough; we need convergence of random functions.

6.1.2 Metric and Normed Spaces

Definition 6.1. *A metric on a set D is a map $d : D \times D \to [0, \infty)$ satisfying symmetry, the triangle inequality, and $d(x, y) = 0$ if and only if $x = y$. If the last condition is omitted, d is a semimetric.*

The open ball of radius r centered at x is $B(x, r) = \{y : d(x, y) < r\}$. A set is open if it is a union of open balls and closed if its complement is open. We write $x_n \to x$ when $d(x_n, x) \to 0$. The closure $\overline{S}$ consists of all limits of sequences from S; S is dense in D if $\overline{S} = D$.

A metric space is *separable* if it contains a countable dense subset. It is *compact* if it is closed and every sequence has a convergent subsequence, and *totally bounded* if, for every $\varepsilon > 0$, it can be covered by finitely many ε-balls.

Definition 6.2. *A norm on a vector space D is a map $\| \cdot \| : D \to [0, \infty)$ such that*

$$\|x + y\| \le \|x\| + \|y\|, \qquad \|\alpha x\| = |\alpha| \|x\|, \qquad \|x\| = 0 \Longleftrightarrow x = 0.$$

6.1.3 Borel σ-field and Borel Measurable Maps

The Borel σ-field $\mathcal{B}(D)$ of a metric space (D, d) is generated by its open sets. A map $f : D_1 \to D_2$ is Borel measurable if $f^{-1}(A) \in \mathcal{B}(D_1)$ for every $A \in \mathcal{B}(D_2)$. A random element $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (D, d)$ is measurable if $X^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{B}(D)$.

Important examples include:

(i) For an arbitrary index set T,

$$\ell^\infty(T) = \{z : T \to \mathbb{R} : \sup_{t \in T} |z(t)| < \infty\}, \qquad \|z\|_T = \sup_{t \in T} |z(t)|.$$

For infinite T this space is generally nonseparable; in particular, the full space $\ell^\infty(T)$ is separable only when T is finite.

(ii) For $T = [a, b]$,

$$C[a, b] \subset D[a, b] \subset \ell^\infty[a, b],$$

where $D[a, b]$ is the Skorohod space of càdlàg functions. The uniform norm is naturally used on $C[a, b]$.

(iii) If (T, ρ) is totally bounded, write $UC(T, \rho)$ for the uniformly continuous functions $z : T \to \mathbb{R}$. Then $UC(T, \rho) \subset \ell^\infty(T)$.

(iv) On the extended real line $[-\infty, \infty]$, one may use

$$d(x, y) = |\Phi(x) - \Phi(y)|,$$

where Φ is the standard normal c.d.f.; this makes the space compact.

6.1.4 Outer Probability and Outer Expectation

A map into a nonseparable function space need not be measurable. For an arbitrary function $Z : \Omega \to \overline{\mathbb{R}}$, define

$$\mathbb{E}^* Z = \inf\{\mathbb{E}U : U \text{ is measurable},\ U \geq Z,\ \mathbb{E}U \text{ exists}\},$$

and

$$\mathbb{E}_* Z = \sup\{\mathbb{E}U : U \text{ is measurable},\ U \leq Z,\ \mathbb{E}U \text{ exists}\}.$$

Outer probability is $\mathbb{P}^*(A) = \mathbb{E}^* \mathbb{1}_A$.

Let $X_n : \Omega_n \to D$ be arbitrary maps and let $X : \Omega \to D$ be Borel measurable. We say

$$X_n \overset{\mathcal{D}}{\rightsquigarrow} X$$

if $\mathbb{E}^* f(X_n) \to \mathbb{E} f(X)$ for every bounded continuous $f : D \to \mathbb{R}$. We write $X_n \overset{\mathbb{P}}{\to} X$ if

$$\mathbb{P}^*\{d(X_n, X) > \varepsilon\} \to 0 \qquad (\varepsilon > 0),$$

and $X_n \to X$ outer almost surely if there are measurable Δ_n with $d(X_n, X) \leq \Delta_n$ and $\Delta_n \to 0$ almost surely.

6.1.5 Tightness and Asymptotic Measurability

A Borel random element X is tight if, for every $\varepsilon > 0$, there is a compact $K \subset D$ such that $\mathbb{P}(X \notin K) < \varepsilon$. A sequence $\{X_n\}$ is asymptotically tight if, for all $\varepsilon, \delta > 0$, some compact $K \subset D$ satisfies

$$\limsup_{n\to\infty} \mathbb{P}^*(X_n \notin K^\delta) < \varepsilon, \qquad K^\delta = \{y : d(y, K) < \delta\}.$$

It is asymptotically measurable if

$$\mathbb{E}^* f(X_n) - \mathbb{E}_* f(X_n) \to 0$$

for every bounded continuous $f : D \to \mathbb{R}$.

Theorem 6.3 (Prohorov)**.** *Let X_n be arbitrary maps into a metric space.*

(i) If $X_n \overset{\mathcal{D}}{\rightsquigarrow} X$ for a tight Borel random element X, then $\{X_n\}$ is asymptotically tight and asymptotically measurable.

(ii) If $\{X_n\}$ is asymptotically tight and asymptotically measurable, then every subsequence has a further subsequence converging weakly to a tight Borel random element.

The Portmanteau theorem and the continuous mapping theorem continue to hold with outer probabilities and expectations in this framework.

6.1.6 Stochastic Processes and Weak Convergence

A stochastic process $X = \{X_t : t \in T\}$ is a collection of random variables on one probability space. For fixed ω, the map $t \mapsto X_t(\omega)$ is a sample path. When all paths are bounded, X may be viewed as a map into $\ell^\infty(T)$.

Theorem 6.4 (Weak convergence of stochastic processes)**.** *Let $X_n : \Omega_n \to \ell^\infty(T)$. Then X_n converges weakly to a tight random element X if and only if:*

(i) for every finite $t_1, \dots, t_k \in T$,

$$(X_n(t_1), \dots, X_n(t_k)) \overset{\mathcal{D}}{\rightsquigarrow} (X(t_1), \dots, X(t_k));$$

(ii) for every $\varepsilon, \eta > 0$, there is a finite partition $T = T_1 \cup \cdots \cup T_k$ such that

$$\limsup_{n\to\infty} \mathbb{P}^* \left(\max_i \sup_{s,t \in T_i} |X_n(s) - X_n(t)| > \varepsilon \right) \le \eta.$$

6.2 Symmetrization and Chaining

6.2.1 Preliminaries

Our goal is to bound the maximal deviation of a sum of independent stochastic processes from its mean. Let

$$S_n(\omega, t) = \sum_{i=1}^n f_i(\omega_i, t), \qquad M_n(t) = \mathbb{E} S_n(t), \qquad \Delta_n = \sup_{t\in T} |S_n(t) - M_n(t)|.$$

For a convex increasing function Ψ on $\mathbb{R}_+$, introduce the Rademacher process

$$L_n(b,\omega) = \sup_{t\in T}\left|\sum_{i=1}^n b_i f_i(\omega_i,t)\right|,$$

where $b_1,\dots,b_n$ are i.i.d. Rademacher variables.

For fixed n, this is governed by the subset

$$\mathcal{F}_\omega = \{(f_1(\omega_1,t),\dots,f_n(\omega_n,t)) : t\in T\}\subset\mathbb{R}^n.$$

6.2.2 Symmetrization

Theorem 6.5 (Symmetrization inequality)**.** *For every convex increasing* $\Psi:\mathbb{R}_+\to\mathbb{R}_+$*,*

$$\mathbb{E}^*\Psi(\Delta_n)\le\mathbb{E}^*\Psi(2L_n).$$

Proof. Let ω' be an independent copy of ω. Jensen's inequality gives

$$\mathbb{E}^*\Psi\left(\sup_t\left|\sum_i\{f_i(\omega_i,t)-\mathbb{E}f_i(\omega_i,t)\}\right|\right)\le\mathbb{E}^*\Psi\left(\sup_t\left|\sum_i\{f_i(\omega_i,t)-f_i(\omega_i',t)\}\right|\right).$$

The differences are symmetric, so Rademacher signs may be inserted. The triangle inequality then bounds the last display by $\mathbb{E}^*\Psi(2L_n)$. □

6.2.3 Orlicz Norms

Let Ψ be convex and increasing on $\mathbb{R}_+$, with $0\le\Psi(0)<1$. The Orlicz norm of a random variable Z is

$$\|Z\|_\Psi=\inf\{c>0:\mathbb{E}^*\Psi(|Z|/c)\le 1\}.$$

Write $\mathcal{L}^\Psi=\{Z:\|Z\|_\Psi<\infty\}$, identifying random variables that are equal almost everywhere.

Lemma 6.6. *Let* $\Psi(x)=e^{x^2}-1$*. For fixed* $b,f\in\mathbb{R}^n$*,*

$$\|\langle b,f\rangle\|_\Psi\le 2\|f\|_2.$$

Proof. Let $g_i\overset{\text{i.i.d.}}{\sim}N(0,1)$ and $r=\mathbb{E}|N(0,1)|=\sqrt{2/\pi}$. Jensen's inequality and the Gaussian moment generating function show that

$$\mathbb{E}_b\exp\left\{\left(\frac{\sum_i b_i f_i}{cr}\right)^2\right\}\le\frac{1}{\sqrt{1-2\|f\|_2^2/(c^2r^2)}}.$$

Choosing $c=2\|f\|_2$ proves the claim. □

Lemma 6.7. *For arbitrary random variables* $Z_1,\dots,Z_m$ *and* $\Psi(x)=e^{x^2}-1$*,*

$$\left\|\max_{1\le i\le m}|Z_i|\right\|_\Psi\le\sqrt{2+\log m}\,\max_{1\le i\le m}\|Z_i\|_\Psi.$$

This follows by the union bound and the tail-integral representation of expectation, choosing the normalization constant $\sqrt{2+\log m}$.

6.2.4 Chaining

For $T_0 \subset T$, define the packing number $D(\varepsilon, T_0, d)$ as the largest cardinality of an ε-separated subset of T_0, and the covering number $N(\varepsilon, T_0, d)$ as the least number of closed ε-balls covering T_0. They satisfy

$$N(\varepsilon, T_0, d) \leq D(\varepsilon, T_0, d) \leq N(\varepsilon/2, T_0, d).$$

Lemma 6.8 (Chaining bound)**.** *Let (T, d) be a bounded metric space and $t_0 \in T$. Suppose the process $Z = \{Z(t) : t \in T\}$ has continuous sample paths and*

$$\|Z(s) - Z(t)\|_\Psi \leq d(s, t).$$

With $\delta = \sup_{t \in T} d(t, t_0)$,

$$\left\| \sup_{t \in T} |Z(t)| \right\|_\Psi \leq \|Z(t_0)\|_\Psi + \sum_{i=0}^{\infty} \frac{\delta}{2^i} \sqrt{2 + \log D(\delta/2^{i+1}, T, d)}.$$

Proof. Let $\delta_i = \delta/2^i$ and choose nested finite nets T_i with $|T_i| \leq D(\delta_i, T, d)$. For each $t \in T_{i+1}$ select $\pi_i(t) \in T_i$ with $d(t, \pi_i(t)) \leq \delta_i$. Iterating

$$\max_{t \in T_{i+1}} |Z(t)| \leq \max_{t \in T_i} |Z(t)| + \max_{t \in T_{i+1}} |Z(t) - Z(\pi_i(t))|$$

and applying the preceding maximal Orlicz inequality yields the result. □

6.3 Packing Numbers of Sets in $\mathbb{R}^n$

6.3.1 Affine Subspaces and Coordinate Projections

Lemma 6.9. *If $\mathcal{F}$ lies in a V-dimensional affine subspace of $\mathbb{R}^n$ and has diameter at most R, then*

$$D(\varepsilon, \mathcal{F}) \leq (2R/\varepsilon)^V.$$

For $I = (i_1, \ldots, i_k)$, the map

$$(x_1, \ldots, x_n) \mapsto (x_{i_1}, \ldots, x_{i_k})$$

is a k-dimensional coordinate projection.

Definition 6.10. *For a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$, its pseudo-dimension $\mathrm{Pdim}(\mathcal{F})$ is the largest V for which there exist $x_1, \ldots, x_V$ and $y_1, \ldots, y_V$ such that every binary vector $b \in \{0,1\}^V$ is realized by some $f \in \mathcal{F}$ through*

$$f(x_i) > y_i \iff b_i = 1, \qquad i = 1, \ldots, V.$$

Lemma 6.11. *If, for every $m > V$ and every $x = (x_1, \ldots, x_m)$, the set*

$$\mathcal{F}_x = \{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}$$

lies in an affine subspace of dimension at most V, then $\mathrm{Pdim}(\mathcal{F}) \leq V$.

Lemma 6.12 (Growth bound)**.** *If $\mathrm{Pdim}(\mathcal{F}) = V$, then for any $k > V$, any $x_1, \ldots, x_k$ and thresholds $y_1, \ldots, y_k$, at most*

$$\sum_{j=0}^{V} \binom{k}{j}$$

binary patterns of the form $\mathbb{1}_{\{f(x_i) > y_i\}}$ can be realized.

6.3.2 ℓ_1 Packing Bounds

For $x \in \mathbb{R}^n$, write $\|x\|_1 = \sum_i |x_i|$. A box is $A = \prod_{i=1}^n [\alpha_i, \beta_i]$ and has ℓ_1-diameter $\sum_i (\beta_i - \alpha_i)$.

Lemma 6.13 (Random coordinate projection)**.** *Let $\mathcal{F}$ lie in an ℓ_1-box of diameter 1 and contain m points whose pairwise ℓ_1 distances are at least ε. If*

$$k = \lceil 2\varepsilon^{-1} \log m \rceil,$$

then some k-dimensional coordinate projection of $\mathcal{F}$ occupies at least m orthants.

Proof. After rescaling, write the box as $\prod_i [0, p_i]$ with $\sum_i p_i = 1$, partition $[0, 1]$ into intervals of lengths p_i, and sample k independent uniform points. Each sample produces a coordinate and a threshold. Two ε-separated vectors fail to be separated by all k thresholds with probability at most $(1 - \varepsilon)^k$. A union bound over the $\binom{m}{2}$ pairs is strictly less than one for the displayed choice of k, so a separating projection exists. □

Theorem 6.14. *If $\mathcal{F}$ has pseudo-dimension V and $F_x = (F(x_1), \ldots, F(x_n))$ is an envelope vector for $\mathcal{F}_x = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$, then constants A, W, depending only on V, satisfy*

$$D_1(\varepsilon \|F_x\|_1, \mathcal{F}_x) \le A(1/\varepsilon)^W, \qquad 0 < \varepsilon \le 1.$$

Indeed, normalize $\mathcal{F}_x$ by $2\|F_x\|_1$. If m points are ε-separated, the projection lemma and the growth bound give $m \le \sum_{j=0}^V \binom{k}{j} \le (1 + V)^k$ with $k \asymp \varepsilon^{-1} \log m$, which rearranges to the stated polynomial bound.

The same conclusion holds after coordinatewise rescaling: for every $\alpha \in \mathbb{R}^n_+$,

$$D_1(\varepsilon \|\alpha \odot F_x\|_1, \alpha \odot \mathcal{F}_x) \le A(1/\varepsilon)^W.$$

6.3.3 Consequences for ℓ_2 Packing Numbers

For any bounded $\mathcal{F} \subset \mathbb{R}^n$ with envelope vector F,

$$D_2(\varepsilon, \mathcal{F}) \le D_1\left(\frac{\varepsilon^2}{2}, F \odot \mathcal{F}\right) \le D_2\left(\frac{\varepsilon^2}{2\|F\|_2}, \mathcal{F}\right),$$

because

$$\|f - g\|_2^2 \le 2\|F \odot f - F \odot g\|_1 \le 2\|F\|_2 \|f - g\|_2.$$

Consequently, under the assumptions above, there exist constants A_2, W_2 depending only on V such that

$$D_2(\varepsilon \|\alpha \odot F_x\|_2, \alpha \odot \mathcal{F}_x) \le A_2(1/\varepsilon)^{W_2}.$$

If two function classes $\mathcal{F}$ and $\mathcal{G}$ each have pseudo-dimension at most V, then the classes $\mathcal{F} \cup \mathcal{G}$, $\mathcal{F} \vee \mathcal{G}$ and $\mathcal{F} \wedge \mathcal{G}$ have pseudo-dimension bounded by a constant multiple of V. Corresponding packing bounds are stable under products and coordinatewise rescaling.

Finally, if $\mathcal{F} \subset B_2(1)$ and

$$\mathrm{co}(\mathcal{F}) = \left\{ \sum_{i=1}^m \lambda_i f_i : \lambda_i \ge 0,\ \sum_i \lambda_i = 1,\ f_i \in \mathcal{F} \right\},$$

then polynomial entropy of $\mathcal{F}$ implies a stretched-exponential entropy bound for its convex hull: for suitable constants $C, \tau > 0$,

$$D_2(\varepsilon, \mathrm{co}(\mathcal{F})) \le \exp\{(C/\varepsilon)^\tau\}, \qquad 0 < \varepsilon \le 1.$$

One proves this by choosing a maximal net of $\mathcal{F}$, replacing every finite convex combination by the corresponding combination of net points, and then covering the resulting simplex in coefficient space.

More explicitly, let $\mathcal{F}_\varepsilon$ be an ε^β-net of $\mathcal{F}$, where $\beta = 2/(2+W)$. Partition $\mathcal{F}_\varepsilon$ into cells $E_1, \ldots, E_m$ of ℓ_2-diameter at most ε^β, and write a convex combination as

$$\sum_{f \in \mathcal{F}_\varepsilon} \theta_f f = \sum_{i=1}^m \lambda_i e_i, \qquad \lambda_i = \sum_{f \in E_i} \theta_f, \qquad e_i = \sum_{f \in E_i} \frac{\theta_f}{\lambda_i} f.$$

For a fixed weight vector λ, approximate each e_i by the average of n_i independent draws from its representing distribution, where $n_i \asymp \lambda_i^2 \varepsilon^{-2+2\beta}$. Since

$$\mathbb{E} \left\| \lambda_i e_i - \lambda_i \overline{Y}_i \right\|_2^2 \le \frac{\lambda_i^2 \operatorname{diam}^2(E_i)}{n_i},$$

the total squared approximation error is of order ε^2. Counting the possible discretized weight vectors and sample averages gives, for every $\tau > 2W/(2+W)$,

$$\log D_2(\varepsilon, \mathrm{co}(\mathcal{F})) \lesssim \varepsilon^{-\tau}.$$

6.4 Uniform Laws of Large Numbers

Let $f_1(\omega, t), \ldots, f_n(\omega, t)$ be independent processes, with separable envelopes $F_1, \ldots, F_n$, and put

$$S_n(t) = \sum_{i=1}^n f_i(t), \qquad M_n(t) = \mathbb{E} S_n(t).$$

Theorem 6.15 (Uniform law of large numbers)**.** *Suppose that, for every $\varepsilon > 0$,*

(i) there exists $K < \infty$ such that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}\big[F_i \mathbb{1}_{\{F_i > K\}}\big] < \varepsilon \quad \textit{for every } n;$$

(ii) with

$$\mathcal{F}_{n,\omega} = \{(f_1(\omega, t), \ldots, f_n(\omega, t)) : t \in T\}, \qquad F_{n,\omega} = (F_1(\omega), \ldots, F_n(\omega)),$$

one has

$$\log D_1(\varepsilon \|F_{n,\omega}\|_1, \mathcal{F}_{n,\omega}) = o_{\mathbb{P}}(n).$$

Then

$$\frac{1}{n} \sup_{t \in T} |S_n(t) - M_n(t)| \xrightarrow{\mathbb{P}} 0.$$

Proof. Fix $\varepsilon > 0$, choose K as in (i), and truncate $f_i^* = f_i \mathbb{1}_{\{F_i \le K\}}$. The contribution of $f_i - f_i^*$ is bounded in outer expectation by

$$\frac{2}{n} \sum_{i=1}^n \mathbb{E}[F_i \mathbb{1}_{\{F_i > K\}}] < 2\varepsilon.$$

By symmetrization, the truncated part is bounded by a Rademacher process. For fixed ω, choose an $\varepsilon\|F_{n,\omega}\|_1$-net $D_{n,\omega}$ of $\mathcal{F}_{n,\omega}$. Then

$$\frac{1}{n}\mathbb{E}_b \sup_{f\in\mathcal{F}_{n,\omega}} |\langle b, f\rangle| \le \frac{\varepsilon}{n}\|F_{n,\omega}\|_1 + \frac{1}{n}\mathbb{E}_b \max_{f\in D_{n,\omega}} |\langle b, f\rangle|.$$

The maximal Orlicz inequality gives

$$\frac{1}{n}\mathbb{E}_b \max_{f\in D_{n,\omega}} |\langle b, f\rangle| \lesssim \frac{1}{n} \max_{f\in D_{n,\omega}} \|f\|_2\sqrt{2+\log|D_{n,\omega}|}.$$

Since $\|f\|_2 \le K\sqrt{n}$ after truncation and $\log|D_{n,\omega}| = o_{\mathbb{P}}(n)$, this term converges to zero. Letting $\varepsilon \downarrow 0$ proves the assertion. □

6.4.1 Consistency of M-estimators

Let Θ be a metric space and let

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i).$$

An estimator satisfying

$$M_n(\widehat{\theta}_n) \ge \sup_{\theta\in\Theta} M_n(\theta) - o_{\mathbb{P}}(1)$$

is called an approximate M-estimator.

Theorem 6.16 (Consistency of M-estimators)**.** *Let M_n be random functions and M a deterministic function on Θ. Suppose*

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$$

and, for every $\varepsilon > 0$,

$$\sup_{\{\theta: d(\theta,\theta_0)\ge\varepsilon\}} M(\theta) < M(\theta_0).$$

Then every approximate maximizer $\widehat{\theta}_n$ satisfies $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Proof. Uniform convergence and approximate maximization imply

$$M(\theta_0) - M(\widehat{\theta}_n) \le 2\sup_{\theta} |M_n(\theta) - M(\theta)| + o_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0.$$

The separation condition then forces $\mathbb{P}\{d(\widehat{\theta}_n, \theta_0) \ge \varepsilon\} \to 0$ for every $\varepsilon > 0$. □

6.4.2 Example: the empirical distribution function

Take

$$f_i(\omega, t) = \mathbb{1}_{\{X_i(\omega)\le t\}}, \qquad t \in \mathbb{R}.$$

The envelope is $F_i = 1$. For a fixed sample, the vectors

$$\left(\mathbb{1}_{\{X_1\le t\}}, \dots, \mathbb{1}_{\{X_n\le t\}}\right), \qquad t \in \mathbb{R},$$

form a monotone chain with at most $n+1$ distinct elements. Thus the empirical packing entropy is $O(\log n) = o(n)$, and the preceding theorem gives the Glivenko–Cantelli conclusion

$$\sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - F(t)| \xrightarrow{\mathbb{P}} 0.$$

Equivalently, one may partition the real line into intervals of small F-probability and use a Chernoff bound plus a union bound to construct an explicit finite cover.

6.4.3 Example: least-squares estimators

Let $\Theta = [-K_0, K_0]$ and

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \sum_{i=1}^{n} (X_i - \theta)^2.$$

For $f_i(\theta) = (X_i - \theta)^2$, an envelope is

$$F_i = 2(X_i^2 + K_0^2).$$

For fixed data,

$$\{(X_1^2 - 2X_1\theta + \theta^2, \dots, X_n^2 - 2X_n\theta + \theta^2) : \theta \in \Theta\}$$

lies in the span of $(X_i^2)_{i=1}^n$, $(X_i)_{i=1}^n$, and $(1, \dots, 1)$, hence in a three-dimensional affine space. Its packing entropy is therefore polynomial, so the uniform law applies whenever the envelope is uniformly integrable.

6.5 Maximal Inequalities

For

$$\Delta_n = \sup_{t \in T} |S_n(t) - M_n(t)|, \qquad \delta_n(\omega) = \sup_{f \in \mathcal{F}_{n,\omega}} \|f\|_2,$$

define the entropy integral

$$J_n(\omega) = 18 \int_0^{\delta_n(\omega)} \sqrt{\log D_2(x, \mathcal{F}_{n,\omega})}\, dx.$$

Theorem 6.17 (Maximal inequalities). *(i) If $J_n(\omega) \le K_n$ for every ω, then*

$$\mathbb{P}^*(\Delta_n \ge t) \le 5 \exp\left(-\frac{t^2}{4K_n^2}\right).$$

(ii) If $\|J_n\|_\Psi < \infty$ for $\Psi(x) = e^{x^2} - 1$, then

$$\mathbb{P}^*(\Delta_n \ge t) \le 2\sqrt{5} \exp\left(-\frac{t}{\|J_n\|_\Psi}\right).$$

Suppose there is a deterministic entropy bound λ_n such that

$$D_2(x\|F_n(\omega)\|_2, \mathcal{F}_{n,\omega}) \le \lambda_n(x), \qquad 0 < x \le 1,$$

and set

$$\Lambda_n(t) = \int_0^t \sqrt{\log \lambda_n(x)}\, dx < \infty.$$

Then

$$J_n(\omega) \le 18\|F_n(\omega)\|_2 \Lambda_n\left(\frac{\delta_n(\omega)}{\|F_n(\omega)\|_2}\right) \le 18\Lambda_n(1)\|F_n(\omega)\|_2.$$

Thus uniformly bounded envelopes give uniform sub-Gaussian behavior after $n^{-1/2}$ scaling. If instead $\sup_i \mathbb{E}e^{\eta F_i^2} < \infty$ for some $\eta > 0$, then $\|J_n\|_\Psi \lesssim \sqrt{n}$, which yields uniform sub-exponential tails for $\Delta_n/\sqrt{n}$.

6.6 Functional Central Limit Theorem

6.6.1 Triangular arrays and manageability

Consider a triangular array

$$\{f_{n,i}(\omega,t) : t \in T,\ i = 1,\dots,k_n\}, \qquad n = 1,2,\dots,$$

whose entries are independent within each row. Let $F_{n,i}$ be envelopes. The array is called *manageable* with respect to F_n if there is a deterministic capacity bound λ satisfying

$$\int_0^1 \sqrt{\log \lambda(x)}\, dx < \infty$$

and, for every nonnegative rescaling vector α,

$$D_2(x\|\alpha \odot F_n(\omega)\|_2, \alpha \odot \mathcal{F}_{n,\omega}) \le \lambda(x), \qquad 0 < x \le 1.$$

Define the centered process

$$X_n(t) = \sum_{i=1}^{k_n} \{f_{n,i}(t) - \mathbb{E} f_{n,i}(t)\}$$

and the intrinsic semimetric

$$\rho_n(s,t) = \left(\sum_{i=1}^{k_n} \mathbb{E}\{f_{n,i}(t) - f_{n,i}(s)\}^2 \right)^{1/2}.$$

Assume $\rho(s,t) = \lim_n \rho_n(s,t)$ exists.

Theorem 6.18 (Functional CLT). *Suppose:*

(i) the array is manageable with respect to F_n;

(ii) the covariance limit

$$H(s,t) = \lim_{n\to\infty} \mathbb{E}\{X_n(s)X_n(t)\}$$

exists for every $s, t \in T$;

(iii) $\limsup_n \sum_i \mathbb{E} F_{n,i}^2 < \infty$;

(iv) for every $\varepsilon > 0$,

$$\sum_i \mathbb{E}[F_{n,i}^2 \mathbb{1}_{\{F_{n,i} > \varepsilon\}}] \to 0;$$

(v) (T,ρ) is well defined and $\rho_n(s_n,t_n) \to 0$ whenever $\rho(s_n,t_n) \to 0$.

Then (T,ρ) is totally bounded, the finite-dimensional distributions of X_n converge to those of a centered Gaussian process G with covariance H, and

$$X_n \overset{\mathcal{D}}{\rightsquigarrow} G \qquad \text{in } \ell^\infty(T).$$

6.7 Applications to M-estimators

6.7.1 A uniform differentiability lemma

Lemma 6.19. *Let $\Theta \subset \mathbb{R}^d$ be open and let $\theta \mapsto m_\theta(x)$ be differentiable at θ_0 for almost every x, with derivative $\dot{m}_{\theta_0}(x)$. Suppose, in a neighborhood of θ_0,*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le \dot{m}(x)\|\theta_1 - \theta_2\|, \qquad \mathbb{E}\dot{m}^2(X) < \infty.$$

For every $r_n \to \infty$ and every tight sequence $\widetilde{h}_n$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[r_n\{m_{\theta_0 + \widetilde{h}_n/r_n}(X_i) - m_{\theta_0}(X_i)\} - \widetilde{h}_n^T \dot{m}_{\theta_0}(X_i) - \mathbb{E}(\cdots) \right] \xrightarrow{\mathbb{P}} 0.$$

The proof applies the functional CLT to h in a fixed bounded ball. A valid envelope is

$$F_{n,i} = \frac{R}{\sqrt{n}}\{\dot{m}(X_i) + \|\dot{m}_{\theta_0}(X_i)\|\},$$

and the induced class has polynomial packing entropy because its parameter set is finite dimensional.

6.7.2 Rate of convergence

Corollary 6.20. *Assume the preceding local Lipschitz condition. Suppose $\theta \mapsto Pm_\theta$ admits a second-order expansion at $\theta_0 = \operatorname{argmax}_\theta Pm_\theta$, with nonsingular negative Hessian V_{θ_0}, and suppose*

$$\mathbb{P}_n m_{\widehat{\theta}_n} \ge \mathbb{P}_n m_{\theta_0} - o_{\mathbb{P}}(n^{-1}), \qquad \widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0.$$

Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = O_{\mathbb{P}}(1).$$

6.7.3 Asymptotic normality

Theorem 6.21 (Asymptotic normality of M-estimators)**.** *Let $\Theta \subset \mathbb{R}^d$ be open. Assume:*

(i) $\theta \mapsto m_\theta(x)$ *is differentiable at θ_0 for almost every x, with derivative $\dot{m}_{\theta_0}(x)$;*

(ii)

$$Pm_\theta = Pm_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + o(\|\theta - \theta_0\|^2),$$

where V_{θ_0} is nonsingular and negative definite;

(iii) $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$ *and*

$$\mathbb{P}_n m_{\widehat{\theta}_n} \ge \sup_{\theta \in \Theta} \mathbb{P}_n m_\theta - o_{\mathbb{P}}(n^{-1}).$$

Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{m}_{\theta_0}(X_i) + o_{\mathbb{P}}(1),$$

and consequently

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \overset{\mathcal{D}}{\rightsquigarrow} N\left(0, V_{\theta_0}^{-1} P[\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T] V_{\theta_0}^{-1}\right).$$

Proof. Write $\widehat{h}_n = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ and

$$\widetilde{h}_n = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\dot{m}_{\theta_0}(X_i) - P\dot{m}_{\theta_0}\}.$$

The uniform differentiability lemma and the quadratic expansion give, uniformly for bounded h,

$$n\mathbb{P}_n(m_{\theta_0+h/\sqrt{n}} - m_{\theta_0}) = \frac{1}{2}h^T V_{\theta_0} h + h^T\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\dot{m}_{\theta_0}(X_i) - P\dot{m}_{\theta_0}\} + o_{\mathbb{P}}(1).$$

Completing the square yields

$$\frac{1}{2}h^T V_{\theta_0} h - h^T V_{\theta_0}\widetilde{h}_n = \frac{1}{2}(h - \widetilde{h}_n)^T V_{\theta_0}(h - \widetilde{h}_n) - \frac{1}{2}\widetilde{h}_n^T V_{\theta_0}\widetilde{h}_n.$$

Approximate maximality and negative definiteness of V_{θ_0} imply $\widehat{h}_n - \widetilde{h}_n \xrightarrow{\mathbb{P}} 0$. The multivariate central limit theorem then gives the stated limit. ☐

6.7.4 Maximum likelihood as a special case

Suppose $\{P_\theta : \theta \in \Theta\}$ is q.m.d. at an interior point θ_0, and there is a measurable ℓ with $P_{\theta_0}\ell^2 < \infty$ such that, locally,

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \le \ell(x)\|\theta_1 - \theta_2\|.$$

If the Fisher information I_{θ_0} is nonsingular and the MLE is consistent, the preceding theorem applied to $m_\theta = \log p_\theta$ gives

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = I_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}_{\theta_0}(X_i) + o_{\mathbb{P}}(1) \overset{\mathcal{D}}{\rightsquigarrow} N(0, I_{\theta_0}^{-1}).$$