Reading:  VDV Chapter 11, 12

Outline: Asymptotics of U-Statistics

- Projections in Hilbert spaces
- Conditional expectations
- Hájek projections
- Asymptotic normality of U-statistics

Recap: Recall these definitions that we set up last lecture:
Given a symmetric kernel function $h : X^r \to \mathbb{R}$, the goal is to estimate

$$\theta := \mathbb{E}[h(X_1, \ldots, X_r)], X_i \overset{iid}{\sim} P.$$

Define the U-Statistic as

$$U_n := \frac{1}{{n \choose r}} \sum_{\beta \subseteq [n], |\beta| = r} h(X_{\beta}).$$

For each $c \in \{0, \ldots, r\}$, define

$$h_c(x_{1:c}) := \mathbb{E}[h(X_{1:r}|X_{1:c} = x_{1:c})],$$

and define

$$\zeta_c := \text{Var}[h_c(X_{1:c})] = \text{Cov}(h(X_A), h(X_B)),$$

where $|A \cap B| = c$.

$$\text{Var}(U_n) = \frac{\sigma^2}{n} \zeta_1 + O(n^{-2}),$$

1 Projections

Definition 1.1. A vector space $\mathcal{H}$ is a Hilbert space if it is a complete normed vector space with inner product $\langle \cdot, \cdot \rangle$, where the norm $||u||^2 = \langle u, u \rangle$ and

$$\langle \alpha x, y \rangle = \alpha \langle x, y \rangle = \alpha \langle y, x \rangle, \text{ all } \alpha \in \mathbb{R},$$

and

$$\langle x + y, u + v \rangle = \langle x, u \rangle + \langle y, u \rangle + \langle x, v \rangle + \langle y, v \rangle.$$
Example: $\mathbb{R}^n$ with $\langle x, y \rangle = x^Ty = \sum_{i=1}^n x_iy_i$ ♦

Example: $L^2(P) = \{ f : X \to \mathbb{R}, \int f(x)^2dP(x) < \infty \}$ with $\langle f, g \rangle = \int f(x)g(x)dP(x)$, we have $\langle f, g \rangle \leq ||f|| ||g||$ by Cauchy-Schwartz inequality. ♦

Let $\mathcal{S} \subseteq \mathcal{H}$ be a closed linear subspace of $\mathcal{H}$ (i.e. $\mathcal{S}$ contains 0 and all the linear combinations of elements in itself).

**Definition 1.2.** For any $v \in \mathcal{H}$, we define the projection of $v$ onto $\mathcal{S}$ as

$$\pi_{\mathcal{S}}(v) := \text{argmin}_{s \in \mathcal{S}} \{ ||s-v||^2 \}.$$

**Theorem 1.** The projection $\pi_{\mathcal{S}}(v)$ exists, is unique, and is unique and characterized by

$$\langle v - \pi_{\mathcal{S}}(v), s \rangle = 0$$

for all $s \in \mathcal{S}$ (orthogonality).

**Example:** In $L^2(P)$, let $\mathcal{S}$ be a collection of random variables (or functions) with $\mathbb{E}(s^2) < \infty$ for all $s \in \mathcal{S}$ and closed under linear combinations (i.e. $\forall s_1, s_2 \in \mathcal{S}$ then $\alpha_1s_1 + \alpha_2s_2 \in \mathcal{S}$). Then $\hat{s}$ is a projection of $T$ onto $\mathcal{S}$ iff

$$\mathbb{E}[(T - \hat{s})s] = 0$$

for all $s \in \mathcal{S}$. ♦

**Proposition 2** (Moreau Decomposition). For any $v \in \mathcal{H}$ and $\mathcal{S}$ is a subspace, we have

$$||v||^2 = ||\pi(v)||^2 + ||v - \pi(v)||^2.$$

**Proof of Proposition**

Since $\langle v - \pi(v), \pi(v) \rangle = 0$, then

$$||v||^2 = ||v - \pi(v) + \pi(v)||^2 = ||\pi(v)||^2 + ||v - \pi(v)||^2 + 2\langle v - \pi(v), \pi(v) \rangle = 0.$$

□

**Conditional Expectations (Projections in $L^2(P)$)**

Let’s define $\mathcal{S} = \{ \text{linear span of } g(Y) \text{ for all measurable functions } g \text{ and some random variable } Y \}$.

**Definition 1.3.** Define conditional expectation as the projection of $X$ onto $\mathcal{S}$. That is how well we can approximate $X$ as the function of $Y$.

\[ \mathbb{E}[X|Y] := \text{Projections of } X \text{ onto } \mathcal{S} = \text{Best } "\text{predictor}" \text{ of } X \text{ onto } \mathcal{S}. \]

$\mathbb{E}[X|Y]$ is the unique (up to measure 0 sets) function of $Y$ such that

$$\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y)] = 0$$

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

\[ \square \]
A few consequences:
1. (Tower Property) $E[X] = E[E[X | Y]]$ (take $g = 1$)
2. For any measurable $f$, $E[f(Y)X | Y] = f(Y)E[X | Y]

Sketch of Proof
For 2,

$$E[f(Y)X - f(Y)E[X | Y])g(Y)] = E[(X - E[X | Y])f(Y)g(Y)] = 0$$

for all measurable $g$.

Consequence: This allows us to ignore smaller order staff!

Let $T_n$ be random variables and $S_n$ be a sequence of subspaces of $L^2(P)$. Let’s define

$$\hat{S}_n = \pi_{S_n}(T_n) = E[T_n | S_n].$$

Proposition 3. Let $\sigma^2(X) = \text{Var}(X)$, if $\frac{\sigma^2(T_n)}{\sigma^2(S_n)} \to 1$ as $n \to \infty$ then

$$\frac{T_n - E[T_n]}{\sigma(T_n)} \xrightarrow{p} 0$$

Proof Let $A_n = \frac{T_n - E[T_n]}{\sigma(T_n)} - \frac{\hat{S}_n - E[\hat{S}_n]}{\sigma(S_n)}$. Note that $E[A_n] = 0$. Thus, if we can show that $\text{Var}(A_n) \to 0$, we are done.

$$\text{Var}(A_n) = \text{Var} \left( \frac{T_n - E[T_n]}{\sigma(T_n)} \right) + \text{Var} \left( \frac{\hat{S}_n - E[\hat{S}_n]}{\sigma(S_n)} \right) - \frac{2 \text{Cov}(T_n, \hat{S}_n)}{\sigma(T_n)\sigma(S_n)}$$

Now using the fact that $T_n - \hat{S}_n$ is orthogonal to $\hat{S}_n$ we have:

$$\text{Cov}(T_n, \hat{S}_n) = E[T_n \hat{S}_n] - E[T_n]E[\hat{S}_n]$$

$$= E[(T_n - \hat{S}_n)\hat{S}_n] - E[E[T_n | S_n]]E[\hat{S}_n]$$

$$= E[(T_n - E[T_n | S_n])\hat{S}_n] + E[\hat{S}_n^2] - E[\hat{S}_n]^2$$

$$= E[\hat{S}_n^2] - E[\hat{S}_n]^2$$

$$= \text{Var}(\hat{S}_n).$$

Hence,

$$\text{Var}(A_n) = 2 \left( 1 - \frac{\sigma(\hat{S}_n)}{\sigma(T_n)} \right) \to 0$$

Which also gives us $A_n \to 0$ in $L_2(P)$.
Hájek Projections

**Lemma 4** (11.10 in VDV). Let $X_1, \ldots, X_n$ be independent. Let $\mathcal{S} = \left\{ \sum_{i=1}^{n} g_i(X_i) : g_i \in L_2(P) \right\}$. If $\mathbb{E}[T^2] < \infty$, let $\hat{S} = \pi_{\mathcal{S}}(T)$, then

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

**Proof**  Note that, by independence of $X_i$s,

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

If $\hat{S}$ is as stated in Equation 2, we prove that $T - \hat{S}$ is orthogonal to $\mathcal{S}$. We have:

$$\mathbb{E}[\hat{S} | X_j] = (n-1)\mathbb{E}T + \mathbb{E}[T | X_j] - (n-1)\mathbb{E}T$$

Thus

$$\mathbb{E}[(T - \hat{S})g_j(X_j)] = \mathbb{E}[(\mathbb{E}[T - \hat{S} | X_j]g_j(X_j)]$$

$$= 0,$$

$$\mathbb{E} \left[(T - \hat{S}) \sum_{j=1}^{n} g_j(X_j) \right] = 0.$$  

Thus, $T - \hat{S}$ must be orthogonal to $\mathcal{S}$, so $\hat{S}$ is the projection of $T$.  

\[ \square \]

2 Application to U-statistics

The main idea is to use (Hájek) projections onto sets of the form :

$$\mathcal{S}_n = \left\{ \sum_{i=1}^{n} g_i(X_i) : g_i(X_i) \in L_2(P) \right\},$$

to approximate $U_n$ by a sum of independent random variables.

**Theorem 5.** Let $h$ be a symmetric kernel (function) of order $r$ and let $\mathbb{E}[h^2] < \infty$, $U_n$ be the associated U-statistic, $\theta = \mathbb{E}[U_n] = \mathbb{E}[h(X_1, \ldots, X_n)]$. If $\hat{U}_n$ is the projection of $U_n - \theta$ onto $\mathcal{S}_n$ then

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

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

4
Proof The first equality is just a direct application of Lemma 4, noting that $\mathbb{E}[U_n - \theta] = 0$. We now show the second equality. Let $\beta \subseteq [n], |\beta| = r$, then

$$
\mathbb{E}[h(X_\beta) - \theta | X_i] = \begin{cases} 
0 & i \notin \beta \\
h_1(X_i) & i \in \beta 
\end{cases}
$$

Then

$$
\mathbb{E}[U_n - \theta | X_i] = \left(\frac{n}{r}\right)^{-1} \sum_{|\beta|=r} \mathbb{E}[h(X_\beta) - \theta | X_i = x] 
= \left(\frac{n}{r}\right)^{-1} \sum_{|\beta|=r, i \in \beta} h_1(X_i) 
= \left(\frac{n}{r}\right)^{-1} \left(\frac{n-1}{r-1}\right) h_1(X_i) = \frac{r}{n} h_1(X_i)
$$

It follows that

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

\[ \square \]

Theorem 6. Using the same notations as in the preceding theorem, we have:

1. \( \sqrt{n}(U_n - \theta - \hat{U}_n) \xrightarrow{p} 0 \)

2. \( \sqrt{n}\hat{U}_n \xrightarrow{d} N(0, r^2 \zeta_1) \)

3. \( \sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, r^2 \zeta_1) \)

Proof \( \sqrt{n}\hat{U}_n \xrightarrow{d} N(0, r^2 \zeta_1) \) is by direct application of the CLT. Then, since

$$
\text{Var}(U_n) = \frac{r^2}{n} \zeta_1 + O(n^{-2})
$$

$$
\text{Var}(\hat{U}_n) = \frac{r^2}{n} \zeta_1
$$

we have \( \frac{\text{Var}(U_n)}{\text{Var}(\hat{U}_n)} \rightarrow 1 \) as \( n \rightarrow \infty \).

Using, Property 3, we get that \( \sqrt{n}(U_n - \theta) - \sqrt{n}\hat{U}_n \xrightarrow{p} 0 \)

By application of Slutsky’s theorem we can conclude the desired results.

\[ \square \]
**Example 1** (Signed Rank Test): This example shows how the U-statistics can be useful because it requires minimal modelling assumptions. Consider \( \theta = \mathbb{P}[X_1 + X_2 > 0] \), with \( U_n = \binom{n}{2}^{-1} \sum_{i<j} 1 \{ X_i + X_j > 0 \} \). Let
\[
H_0 : \{ \text{Distribution } P \text{ of } X_i \text{ is symmetric about } 0 \text{ and has continuous CDF} \}
\equiv \{ F(x) = \mathbb{P}[X \leq x] = 1 - F(-x) \forall x \in \mathbb{R} \}
\]
Note that, given \( X_i \),
\[
h_1(X_i) = \mathbb{E}[1 \{ X_i + X_j > 0 \} | X_i]
= \mathbb{P}[X_j > -X_i | X_i]
= 1 - F(-X_i)
\]
As a result, we have
\[
\hat{U}_n = \sum_{i=1}^{n} \mathbb{E}[U_n - \theta | X_i]
= -\frac{2}{n} \sum_{i=1}^{n} (F(-X_i) - \mathbb{E}[F(-X_i)])
\]
Under \( H_0 \), we have \( F(x) = 1 - F(x) \) and \( \theta = \frac{1}{2} \). Because we assumed that \( F(x) \) is continuous, \( F(X_i) \sim \text{Unif}[0, 1] \). Thus we have
\[
\hat{U}_n \overset{d}{=} \frac{2}{n} \sum_{i=1}^{n} (Y_i - \frac{1}{2})
\]
where \( Y_i \overset{iid}{\sim} \text{Unif}[0, 1] \). Because the variance of a uniform random variable is \( \frac{1}{12} \), the central limit theorem gives us \( \sqrt{n}\hat{U}_n \overset{d}{\rightarrow} N(0, \frac{1}{3}) \). We can then test using quantiles of the normal distribution. ♦