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:

**Definition 0.1.** Given a symmetric kernel function \( h : X^r \to \mathbb{R} \), define the associated **U-statistic** as

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

**Definition 0.2.** For each \( c \in \{0, \ldots, r\} \), define

\[
h_c(x_1, \ldots, x_c) := \mathbb{E}[h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r)].
\]

Define \( \hat{h}_c \) to be the centered version of \( h_c \), i.e.

\[
\hat{h}_c := h_c - \mathbb{E}[h_c] = h_c - \theta,
\]

where \( \theta = \mathbb{E}[U_n] \).

**Definition 0.3.** For each \( c \in \{0, \ldots, r\} \), define

\[
\zeta_c := \text{Var}[h_c(X_1, \ldots, X_c)] = \mathbb{E}[h_c(X_1, \ldots, X_c)^2].
\]

(Note that \( \zeta_0 = 0 \).)

We also proved the two following results:

**Claim 1.** For \( A, B \subseteq [n] \) if \( |A \cap B| = c \) (i.e. sets \( A \) and \( B \) have \( c \) common elements) then

\[
\text{Cov}(h(X_A), h(X_B)) = \zeta_c
\]

**Claim 2.** As a consequence, in an asymptotic sense (i.e. for \( r \) fixed and \( n \to \infty \)), we have

\[
\text{Var}(U_n) = \frac{r^2}{n} \zeta_1 + O(n^{-2}),
\]
1 Projections

Definition 1.1. A vector space $\mathcal{H}$ is a Hilbert space if it is a complete normed vector space and we have an inner product
\[ \langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R} \]
which is linear in both arguments and $\langle u, u \rangle = ||u||^2$.

Example: $\mathbb{R}^n$ with $\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$.

Example: $L^2(P) = \{ f : \mathcal{X} \to \mathbb{R}, \int f(x)^2 dP(x) < \infty \}$ 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.

Definition 1.2. Let $S \subseteq \mathcal{H}$ be a closed linear subspace of $\mathcal{H}$ (i.e. $S$ contains 0 and all the linear combinations of elements in itself). For any $v \in \mathcal{H}$ we define the projection of $v$ onto $S$ as
\[ \pi_S(v) := \operatorname{argmin}_{s \in S} \{ \|s - v\|^2 \}. \]

Theorem 3. The projection $\pi_S(v)$ exists, is unique, and is uniquely defined by the inequality
\[ \langle v - \pi_S(v), s \rangle = 0 \]
for all $s \in S$.

Example: In $L^2(P)$, let $S$ be a collection of random variables such that $\mathbb{E}(s^2) < \infty$ for all $s \in S$. Then for $T \in L^2(P)$, the projection of $T$ onto $\operatorname{span}(S)$: $\hat{s}$, is the best $L^2$-approximation of $T$ by random variables in $S$ and we have $\mathbb{E}_P[(T - \hat{s})s] = 0$ for all $s \in S$.

Conditional Expectations

Conditional expectations considered as projections in $L^2(P)$. Let’s define $S = \operatorname{linear span}\{g(Y)\}$ for all measurable functions $g$ with $\mathbb{E}[g^2(Y)] < \infty$.

Definition 1.3. If $X \in L^2(P)$, $Y$ is a random variable, we define the conditional expectation of $X$ given $Y$: $\mathbb{E}[X | Y]$, as the projection of $X$ onto $S$, or as the prediction of $X$ (in mean square) given observation $Y$, i.e. $\mathbb{E}[X | Y]$ 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 $g \in S$.

A few consequences:

1. $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$ (take $g = 1$)
2. For any $f$, $\mathbb{E}[f(Y)X \mid Y] = f(Y)\mathbb{E}[X \mid Y]$
3. Tower property of $\mathbb{E}$: $\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y] = \mathbb{E}[X \mid Y]$
Consequence: this allows us to ignore smaller order terms in non-i.i.d. sums of random variables.

Let \( T_n \) be random variables and \( S_n \) be a sequence of subspaces of \( L^2(P_n) \). Let’s define \( \hat{S}_n = \pi S_n(T_n) \)

**Proposition 4.** Let \( \sigma^2(X) = \text{Var}(X) \), if \( \frac{\sigma^2(T_n)}{\sigma^2(S_n)} \rightarrow 1 \) as \( n \rightarrow \infty \) then

\[
\frac{T_n - \mathbb{E}[T_n]}{\sigma(T_n)} - \frac{\hat{S}_n - \mathbb{E}[\hat{S}_n]}{\sigma(S_n)} \overset{p}{\rightarrow} 0
\]

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

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

\[
= 2 - \frac{2 \, \text{Cov}(T_n, \hat{S}_n)}{\sqrt{\sigma(T_n)\sigma(\hat{S}_n)}}
\]

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

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

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

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

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

Hence,

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

\( \square \)

**Hájek Projections**

**Lemma 5** (11.10 in VDV). Let \( X_1, \ldots, X_n \) be independent. Let \( S = \left\{ \sum_{i=1}^{n} g_i(X_i) : g_i \in L^2(P) \right\} \).

If \( \mathbb{E}T^2 < \infty \), then the projection \( \hat{S} \) of \( T \) onto \( S \) is given by

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

**Proof** Note that

\[
\mathbb{E} \left[ \mathbb{E}[T \mid X_i] \mid X_j \right] = \begin{cases} \mathbb{E}[T \mid X_i] & \text{if } i = j, \\ \mathbb{E}T & \text{if } i \neq j. \end{cases}
\]
If \( \hat{S} \) is as stated in Equation 2, then
\[
\mathbb{E}[\hat{S} \mid X_j] = (n - 1)E_T + \mathbb{E}[T \mid X_j] - (n - 1)E_T = \mathbb{E}[T \mid X_j],
\]
\[
\mathbb{E}[(T - \hat{S})g_j(X_j)] = \mathbb{E}[\mathbb{E}[T - \hat{S} \mid X_j]g_j(X_j)] = 0,
\]
\[
\mathbb{E} \left( (T - \hat{S}) \sum_{j=1}^{n} g_j(X_j) \right) = 0.
\]
Thus, \( \hat{S} \) must be the projection of \( T \) onto \( S \).

\[\square\]

2 Application to U-statistics

The main idea is to use (Hájek) projections onto sets of the form :
\[
S_n = \left\{ \sum_{i=1}^{n} g_i(X_i) : g_i(X_i) \in L_2(P) \right\}.
\]
to approximate \( U_n \) by a sum of independent random variables.

**Theorem 6.** 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 \( S_n \) then
\[
\hat{U}_n = \sum_{i=1}^{n} \mathbb{E}[U_n - \theta \mid X_i] = \frac{r}{n} \sum_{i=1}^{n} h_1(X_i)
\]

**Proof** The first equality is just a direct application of Lemma 5.

Let \( \beta \subseteq [n], |\beta| = r \), then
\[
\mathbb{E}[h(X_\beta) - \theta \mid X_i] = \begin{cases} 
0 & \text{if } i \notin \beta \\
1 & \text{if } i \in \beta
\end{cases}
\]

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

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

\[\square\]
Theorem 7. Using the same notations as in the preceding theorem, we have:

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

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

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

Proof \( \sqrt{n}\hat{U}_n \xrightarrow{d} \mathcal{N}(0, r^2\zeta_1) \) 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 \( \frac{\text{Var}(U_n)}{\text{Var}(\hat{U}_n)} \to 1 \) as \( n \to \infty \).

Using, Property 4, 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. \( \square \)