U-Statistics

Joël Terschuur

June 14, 2022

Functionals

- Let $\mathcal{F}$ be a set of distribution functions, consider the functional

$$\theta = \theta(F), \quad F \in \mathcal{F}.$$

- Halmos (1946) asks:
    - **Q1:** Does there exist unbiased $\hat{\theta}$ for θ for all $F \in \mathcal{F}$?.
    - **Q2:** For which sets $\mathcal{F}$ and functionals θ is the answer to **Q1** affirmative?
    - **Q3:** If such an estimator exists, what is it? If several exist, which is the best?

Functionals

- **Q1:** Does there exist unbiased $\hat{\theta}$ for θ for all $F \in \mathcal{F}$?.

- **Answer:** A functional θ defined in $\mathcal{F}$ admits an unbiased estimator iff there is a function h of m variables such that

$$\theta(F) = \int \dots \int h(x_1, \dots, x_m) F(dx_1) \dots F(dx_m),$$

for all $F \in \mathcal{F}$. WLOG h can be taken symmetric.

- i.e. if $\theta(F) = \mathbb{E}_F[h(X_1, \dots, X_m)]$, $X_1, \dots, X_m$ i.i.d. distributed according to F.

- **Q2** is also answered. h is called the kernel of the functional.

V-statistics (V from Von Mises)

- Given i.i.d data $X_1, ..., X_n$ from F (take as given from now on)

- We can look at the plug in estimator (V-statistic)

$$\theta(\hat{F}_n) = \frac{1}{n^m} \sum_{i_1}^{n} ... \sum_{i_m}^{n} h(X_{i1}, ..., X_{im})$$

- Common notation
  - $m = 1$

$$V_n h \equiv \mathbb{P}_n h = n^{-1} \sum_{i=1}^{n} h(X_i)$$

  - $m = 2$

$$V_n h \equiv (\mathbb{P}_n \times \mathbb{P}_n) h = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(X_i, X_j)$$

V-statistics Bias

- Let $m = 2$, assume wlog that h is symmetric

$$\theta(\hat{F}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, X_j) = \frac{2}{n^2}\sum_{i<j} h(X_i, X_j) + \frac{1}{n^2}\sum_{i=1}^{n} h(X_i, X_i)$$

- First term sums over terms not in the diagonal (twice the sum over a triangle by symmetry). Second term sums over the diagonal.

$$\mathbb{E}[\theta(\hat{F}_n)] = \frac{n-1}{n}\theta(F) + \frac{1}{n}\mathbb{E}[h(X_i, X_i)].$$

- Bias goes away as $n \to \infty$.

U-statistics (U for unbiased)

- $\hat{\theta}(X_1, ..., X_n) = h(X_1, ..., X_m)$ is an example of an unbiased estimator. Inefficient since it does not use all the sample.

- A more efficient estimator is a symmetric function of all n observations

$$\hat{\theta}(X_1, ..., X_n) \equiv U_n h = \binom{n}{2}^{-1} \sum_{1 \leq i_1 < ... < i_m \leq n} h(X_{i_1}, ..., X_{i_m})$$

- $U_n h$ is a U-statistic, termed by Hoeffding (1948).

- **Answer to Q3:** $U_n h$ is the only symmetric estimator which is unbiased for all F for which $\theta(F)$ exists, and it can be shown to have smaller variance than any other such unbiased estimator.

U-statistics: Notation

- We use the following notation

$$U_n h = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} h(X_{i_1}, \ldots, X_{i_m})$$

- **e.g.** $m = 1$

$$U_n h = n^{-1} \sum_{i=1}^{n} h(X_i)$$

- **e.g.** $m = 2$

$$U_n h = \binom{n}{2}^{-1} \sum_{i<j} h(X_i, X_j)$$

U-statistics vs V-statistics

- It can be shown that

$$\sqrt{n}(V_n h - \theta) = \frac{n-1}{n}\sqrt{n}(U_n h - \theta) + \frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}[h(X_i, X_i) - \theta]$$

- $V_n h$ and $U_n h$ are asymptotically equivalent.

- U-statistics are unbiased while V-statistics are only asymptotically unbiased.

U-statistics: Examples

- **Sample mean:**

$$\theta(F) = \mathbb{E}[X_1], \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = U_n h,$$

where $h(x) = x$.

- **Sample variance:**

$$\theta(F) = \mathbb{E}\left[\frac{(X_1 - X_2)^2}{2}\right], \quad \hat{\theta} = \binom{n}{2}^{-1} \sum_{i<j} \frac{(X_i - X_j)^2}{2} = U_n h,$$

where $h(x_1, x_2) = (1/2)(x_1 - x_2)^2$.

U-statistics: Examples

- **Gini Mean Difference (GMD):**

$$\theta(F) = \mathbb{E}[|X_1 - X_2|], \quad \hat{\theta} = \binom{n}{2}^{-1} \sum_{i<j} |X_i - X_j| = U_n h,$$

where $h(x_1, x_2) = |x_1 - x_2|$.

- **Gini Coefficient:** (ratio of U-statistics)

$$\theta(F) = \frac{\mathbb{E}[|X_1 - X_2|]}{\mathbb{E}[X_1 + X_2]}, \quad \hat{\theta} = \frac{\sum_{i<j} |X_i - X_j|}{\sum_{i<j}(X_i + X_j)} = \frac{U_n h_1}{U_n h_2},$$

where $h_1(x_1, x_2) = |x_1 - x_2|$ and $h_2(x_1, x_2) = x_1 + x_2$.

U-statistics: Hájek Projection

- Suppose $m = 2$. Want to "linearize" the U-statistic

$$U_n h - \theta,$$

- The closest sample mean statistic (i.e. projection on space of iid sums) is

$$\begin{aligned}\Pi_1(U_n h - \theta) &= \frac{1}{n}\sum_{i=1}^{n}[\mathbb{E}[U_n h(X_1, X_2)|X_i] - \theta] \\ &= \frac{2}{n}\sum_{i=1}^{n}[\mathbb{E}[h(X_i, X_2)|X_i] - \theta] \\ &\equiv \frac{2}{n}\sum_{i=1}^{n}[h_1(X_i) - \theta].\end{aligned}$$

- $h_1(x) \equiv \mathbb{E}[h(X_1, X_2)|X_1 = x] = \mathbb{E}[h(x, X)]$.

U-statistics: Hájek Projection

- One can show

$$U_n h = \frac{2}{n} \sum_{i=1}^{n} h_1(X_i) + o_p(n^{-1/2}).$$

- Hence, we have linearized the U-statistic and we can apply common theory for sum of i.i.d. variables.

- **Variance:** $h(x_1, x_2) = (1/2)(x_1 - x_2)^2$ and $h_1(x) = (1/2)[(x - \mu)^2 - \sigma^2]$ and

$$\frac{2}{n} \sum_{i=1}^{n} [h_1(X_i) - \sigma^2] = \frac{1}{n} \sum_{i=1} [(X_i - \mu)^2 - \sigma^2].$$

U-statistics: Hoeffding Decomposition

- Take a symmetric kernel of order m

$$\begin{aligned} h_k(x_1, \dots, x_k) &= \mathbb{E}[h(X_1, \dots, X_m) | X_1 = x_1, \dots, X_k = x_k] \\ &= \mathbb{E}[h(x_1, \dots, x_k, X_{k+1}, \dots, X_m)], \end{aligned}$$

and centered versions $\tilde{h}_k = h_k - \theta$.

- Define

$$\begin{aligned} g_1(X_1) &\equiv \tilde{h}_1(X_1), \\ g_2(X_1, X_2) &\equiv \tilde{h}_2(X_1, X_2) - g_1(X_1) - g_1(X_2), \\ g_3(X_1, X_2, X_3) &\equiv \tilde{h}_3(X_1, X_2, X_3) - \sum_{j=1}^{3} g_1(X_j) \\ &\quad - g_2(X_1, X_2) - g_2(X_1, X_3) - g_2(X_2, X_3), \dots \end{aligned}$$

U-statistics: Hoeffding Decomposition

- Define
$$H_n^{(c)} = U_n g_c$$
e.g. $c = 1 : n^{-1} \sum_{i=1}^{n} g_1(X_i)$; $c = 2 : \binom{n}{2}^{-1} \sum_{i<j} g_2(X_i, X_j)$.

- **Hoeffding Decomposition:**
$$U_n h - \theta = \sum_{c=1}^{m} \binom{m}{c} U_n g_c$$

- **Mean:** $h(x) = x$: $U_n h = n^{-1} \sum_{i=1}^{n} (X_i - \theta)$

- **GMD:** $h(x_1, x_2) = |x_1 - x_2|$ (BLACKBOARD)

U-statistics: Hoeffding decomposition Geometry

- Let $\mathcal{L}_2^j$ be the space of j-th order U-statistics with square integrable kernel
- Let $\mathcal{M}_j = \mathcal{L}_2^j \cap \mathcal{L}_2^{j-1\perp}$, then the space of all U-statistics: $\mathcal{L}_2 = \bigoplus_{i=1}^m \mathcal{M}_i$
- The H-decomposition is the projection of $U_n h$ onto $\bigoplus_{i=1}^m \mathcal{M}_i$.

Slides/HdecompG.png

Figure 1: U-statistics spaces

Influence Function (IF)

- $(X_1, ..., X_n)$ iid F_0. An asymptotically linear estimator satisfies

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(X_i) + o_p(1).$$

- φ is an IF, it has mean zero and satisfies

$$\frac{d}{d\tau}\theta(F_\tau)\bigg|_{\tau=0} = \int \varphi(x)H(dx),$$

  where $F_\tau = F_0 + \tau(H - F_0)$ and $H \in \mathcal{H}$ is an alternative distribution

- If $\mathcal{H}$ is large enough φ is unique

Examples

- **Mean:** $\theta_0(F_0) = \mathbb{E}[X_i]$

$$\frac{d}{d\tau}\int xF_\tau(dx)|_{\tau=0} = \frac{d}{d\tau}\left(\int xF_0(dx) + \tau\int x(H-F)(dx)\right)|_{\tau=0}$$
$$= \int (x-\theta_0)H(dx).$$

- Hence, $\varphi(x) = x - \theta_0$ and trivially if $\hat{\theta}$ is the sample mean

$$\sqrt{n}(\hat{\theta}-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \theta_0)$$

Examples

- **Variance:** $\theta_0 = \mathbb{E}[(1/2)(X_i - X_j)^2]$

$$\frac{d}{d\tau}\int\int \frac{(x_i - x_j)^2}{2} F_\tau(dx_i)F_\tau(dx_j)|_{\tau=0}$$
$$= \int\int \frac{(x_i - x_j)^2}{2}(F_0(dx_i)H(dx_j) + H(dx_i)F_0(dx_j) - 2F_0(dx_i)F_0(dx_j$$
$$= \int [(x - \mathbb{E}[X_i])^2 - \theta_0] H(dx).$$

- Hence, $\varphi(x) = (x - \mathbb{E}[X_i])^2 - \theta_0$ and

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{n}\sum_{i=1}^{n}[(X_i - \mathbb{E}[X_i])^2 - \theta_0] + o_p(1).$$

- IF of a U-statistic is the first term of H-projection (i.e. Hájek projection)!

References

Paul R Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43, 1946.

W Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.