U-Statistic with Side Information

Ao Yuan[1], Wenqing He[2], Binhuan Wang[3], and Gengsheng Qin[3]

1. National Human Genome Center, Howard University, Washington DC, USA
2. Department of Statistics and Actuarial Science, University of Western Ontario, Canada
3. Department of Mathematics and Statistics, Georgia State University, Atlanta, USA.

Introduction

- U-statistics
- Empirical likelihood with side information
- Incorporate side information into U-statistic
- Asmptotic properties
- Examples
- Simulation studies
- Summary

U-statistic

$X_1, ..., X_n$ i.i.d. F unknown. $\mathbf{i} = (i_1, ..., i_m)$, $\mathbf{X_i} = (X_{i_1}, ..., X_{i_m})$, $D_{n,m} = \{\mathbf{i} : 1 \leq i_1 < \cdots < i_m \leq n\}$, C_n^m: combination number, $F_m(\mathbf{x}) = \prod_{j=1}^m F(x_j)$, $F_{n,m}(\mathbf{x})$: empirical distribution function of F_m based on $\{\mathbf{X_i} : \mathbf{i} \in D_{n,m}\}$, with mass $1/C_n^m$ at each point. h: m-variate symmetric kernel. U-statistic:

$$U_n = (C_n^m)^{-1} \sum_{\mathbf{i} \in D_{n,m}} h(\mathbf{X_i}) = E_{F_{n,m}} h(\mathbf{X}).$$

Goal: estimate $\theta = E_{F_m} h(\mathbf{X})$, U-statistic: the minimal variance unbiased estimator of θ.

Empirical Likelihood (EL)

Since Owen (1988), EL has gained increasing popularity: wide range of applications, simplicity to use, incorporate side information. Side infor. be incorporated into EL through a d-dimensional known function $g(x) = (g_1(x), ..., g_d(x))'$ with

$$E_F[g(X_1)] = 0.$$

Denote $w_i = F(\{X_i\})$. EL subject to the side information constraints:

$$\max_w \prod_{i=1}^n w_i \quad \text{subject to} \quad \sum_{i=1}^n w_i = 1 \quad \text{and} \quad \sum_{i=1}^n w_i g(X_i) = 0.$$

Let $t = (t_1, ..., t_d)'$: Lagrange multipliers, then

$$w_i = \frac{1}{n} \frac{1}{1 + t' g(X_i)},$$

$t = t(X_1, ..., X_n)$ determined by

$$\sum_{i=1}^{n} \frac{g(X_i)}{1 + t' g(X_i)} = 0.$$

Existence of t as solution to the above equation can be found, eg. Owen.

Empirical Weights for U-statistic

$w_{\mathbf{i}} := F_m(\{\mathbf{X}_{\mathbf{i}}\})$, $w := (w_{\mathbf{i}} : \mathbf{i} \in D_{n,m})$.
Define EL subject to side infor. constraints as

$$\max_w \prod_{\mathbf{i}\in D_{n,m}} w_{\mathbf{i}} \quad \text{subject to} \quad \sum_{\mathbf{i}\in D_{n,m}} w_{\mathbf{i}} = 1, \sum_{\mathbf{i}\in D_{n,m}} w_{\mathbf{i}} g(\mathbf{X}_{\mathbf{i}}) = 0.$$

Similarly as before, we get

$$w_{\mathbf{i}} = (C_n^m)^{-1} \frac{1}{1 + t' g(\mathbf{X}_{\mathbf{i}})} \tag{2}$$

$t = t_n(X_1, ..., X_n)$ determined by

$$\sum_{\mathbf{i} \in D_{n,m}} \frac{g(\mathbf{X}_{\mathbf{i}})}{1 + t' g(\mathbf{X}_{\mathbf{i}})} = 0. \tag{3}$$

U-statistic with Side Information

With $w_{\mathbf{i}}$'s given in (2) and (3), we define the U-statistic with side infor. given by the constraints g as

$$\tilde{U}_n = \sum_{\mathbf{i} \in D_{n,m}} w_{\mathbf{i}} h(\mathbf{X}_{\mathbf{i}}) = E_{\tilde{F}_{n,m}} h(\mathbf{X}). \tag{4}$$

Comparison: commonly used U-statistic U_n has weight $(C_n^m)^{-1}$ at each observation $h(\mathbf{X}_{\mathbf{i}})$, with side infor., the weights are $w_{\mathbf{i}}$.

Asymptotic Properties of $\tilde{U}_n$

- Notations

As in Hoeffding (1948), for kernel $h(\cdot)$ with $E_{F_m}(h(\mathbf{X})) < \infty$, let $h_c(x_1, ..., x_c) = Eh(x_1, ..., x_c, X_{c+1}, ..., X_m)$, $h_c^o = h_c - \theta$ be its centered version $(c = 1, ..., m)$, $\tilde{h}_1(X_1) = h_1^o(x_1)$, $\tilde{h}_2(x_1, x_2) = h_2^o(x_1, x_2) - \tilde{h}_1(x_1) - \tilde{h}_1(x_2)$, $\tilde{h}_3(x_1, x_2, x_3) = h_3^o(x_1, x_2, x_3) - \sum_{i=1}^3 \tilde{h}_1(x_i) - \sum_{1 \le i < j \le 3} \tilde{h}_2(x_i, x_j)$,

...

$$\tilde{h}_c(x_1, ..., x_c) = h^o(x_1, ..., x_c) - \sum_{i=1}^{c} \tilde{h}_1(x_i)$$

$$- \sum_{1 \leq i < j \leq c} \tilde{h}_2(x_i, x_j) - \cdots - \sum_{1 \leq i_1 < \cdots < i_{c-1} \leq c} \tilde{h}_{c-1}(x_{i_1}, ..., x_{i_{c-1}})$$

$$= \int \cdots \int h_c(y_1, ..., y_c) \prod_{s=1}^{c} d(\delta_{x_s}(y_s) - F(y_s)), \quad (c = 1, ..., m),$$

(Korolyuk and Borovskich, 1994). $\tilde{h}_c$: *canonical forms* of h. $\tilde{U}_n$ is of *rank* k $(1 \leq k \leq m)$ if $\tilde{h}_1 = \cdots = \tilde{h}_{k-1} = 0$ and $\tilde{h}_k \neq 0$. When $k > 1$ we have $\theta = 0$, and U_n (or h) called *degenerate*.

Similarly, for g, define

$$g_c(x_1, ..., c_c) = E_{F_m} g(x_1, ..., x_c, X_{c+1}, ...X_m), \quad (c = 1, ..., m)$$

and canonical forms for g,

$$\tilde{g}_c(x_1, ..., x_c) = \int \cdots \int g_c(y_1, ..., y_c) \prod_{s=1}^{c} d(\delta_{x_s}(y_s) - F(y_s)).$$

Likewise, let q_c be the canonical forms of $g(\cdot)h(\cdot)$ $(c = 1, ..., m)$. Let $r_o = \min\{rank(g_1), ..., rank(g_d)\}$, $r = rank(h)$, $r_1 = \min\{rank(g_1 h), ..., rank(g_d h)\}$, and $\tilde{F}_{nm}$ be the empirical distribution with mass $w_{\mathbf{i}}$ at the observation $\mathbf{x_i}$.

- Regularity Conditions

(C1). $\Omega := E[g(\mathbf{X})g'(\mathbf{X})]$ is positive definite.
(C2). $E||g(\mathbf{X})||^{\alpha} < \infty$ for some $\alpha > 0$ to be specified.
(C3). $E_{F_m}|h(\mathbf{X})| < \infty$.
(C4). $E_{F_m}h^2(\mathbf{X}) < \infty$.
(C5) $E_{F_m}[||g(\mathbf{X})h(\mathbf{X})|| + ||g(\mathbf{X})||^2|h(\mathbf{X})|] < \infty$.

Note: (C2) with $\alpha \geq 4$ and (C4) implies (C5).

Lemma. *Assume (C1) and (C2) for* $\alpha > 2m/r_o$*, we have (i)*

$$w_{\mathbf{i}} \stackrel{a.s.}{=} \frac{1}{C_n^m}\Bigg(1 - g'(\mathbf{X_i})\Omega^{-1}\frac{1}{C_n^m}\sum_{\mathbf{j}\in D_{n,m}} g(\mathbf{X_j})$$

$$+g(\mathbf{X_i})O(\rho_n n^{-1/2}(\log\log n)^{1/2})$$

$$+[g(\mathbf{X_i}) + ||g(\mathbf{X_i})||^2]O(\rho_n^2)\Bigg),$$

where,

$$\rho_n = \begin{cases} O(n^{-1/2}(\log\log n)^{1/2}), & r_o = 1; \\ o(n^{-r_o/2}\log n), & 1 < r_o \le m. \end{cases}$$

(ii)

$$w_{\mathbf{i}} = \frac{1}{C_n^m}\Bigg(1 - g'(\mathbf{X_i})\Omega^{-1}\frac{1}{C_n^m}\sum_{\mathbf{j}\in D_{n,m}} g(\mathbf{X_j})$$

$$+g(\mathbf{X_i})O_p(n^{-(r_o+1)/2}) + [g(\mathbf{X_i}) + ||g(\mathbf{X_i})||^2]O_p(n^{-r_o})\Bigg).$$

The $O_p(\cdot)$ terms above are uniformly for all the $\mathbf{x_i}$'s and $\mathbf{i}$'s.

- Strong consistency of $\tilde{U}_n$

Theorem 1. *(i). Assume the conditions in the Lemma and (C3) and (C5), if $r = 1$, then*

$$n^q(\tilde{U}_n - \theta) \to 0, \quad a.s. \quad \text{for all } q < 1/2.$$

(ii) Assume conditions in the Lemma and (C4) and (C5), if $r > 1$, then

$$a_n \tilde{U}_n \to 0, \ \ (a.s.), \ \ a_n = \begin{cases} n^q \ \text{ for all } \ q < 1/2, & r_1 = r_o = 1; \\ n^{\min\{r/2,1\}}/\log n, & r_1 > r_o = 1; \\ n^{\min\{r_o,r\}/2}/\log n, & 1 = r_1 < r_o; \\ n^{\min\{r,r_o+r_1,2r_o\}/2}/\log n, & r_o, r_1 > 1. \end{cases}$$

(iii) *Assume (C4) and conditions of Lemma (i), if $r = 1$, then*

$$\limsup_n \left(2\sigma^2 \frac{\log\log n}{n}\right)^{-1/2} |\tilde{U}_n - \theta| = 1, \quad (a.s.)$$

- Asymptotic distribution of $\tilde{U}_n$

$W(A)$: Gaussian random measure, $J_r(h)$: Wiener-It$\hat{o}$ integral of order r (Koroljuk and Borovskich, 1994).

Theorem 2. *(i) Assume (C4) and conditions of the Lemma, if* $r = 1$*,*

$$\sqrt{n}(\tilde{U}_n - \theta) \xrightarrow{D} N(0, \sigma^2),$$

$$\sigma^2 = \begin{cases} m^2(\eta_1^2 - 2A'\Omega^{-1}A_1 + A'\Omega^{-1}\Omega_1\Omega^{-1}A), & r_o = 1; \\ m^2\eta_1^2, & r_o > 1; \end{cases},$$

where $\eta_1^2 = E_F\tilde{h}_1^2(X_1)$*,* $\Omega_1 = E_F(\tilde{g}_1(X_1)\tilde{g}_1'(X_1))$*,*
$A = E_{F_m}[g(\mathbf{X})h(\mathbf{X})]$ *and* $A_1 = E_F[\tilde{g}_1(X_1)\tilde{h}_1(X_1)]$*.*

(ii) *Assume (C4), conditions of Lemma (ii) and* $r > 1$*, then*

$$n^{b/2}\tilde{U}_n \xrightarrow{D} Z, \qquad \text{where}$$

$$\begin{cases} b = 1, & Z = mJ_1(A'\Omega^{-1}\tilde{g}_1), & r_o = r_1 = 1; \\ b = 2, & Z = O_P(1), & 1 = r_o < r_1; \\ b = r, & Z = C_m^r J_r(\tilde{h}_r - A'\Omega^{-1}\tilde{g}_r), & 1 = r_1 < r_o = r; \\ b = r_o, & Z = -C_m^{r_o} J_{r_o}(A'\Omega^{-1}\tilde{g}_{r_o}), & 1 = r_1 < r_o < r; \\ b = r, & Z = C_m^r J_r(\tilde{h}_r), & 1 = r_1 < r < r_o; \\ b = r_o, & Z = O_P(1), & 1 < r_o \leq \min\{r_1, r/2\}; \\ b = r, & Z = C_m^r J_r(\tilde{h}_r) - C_m^{r_1} C_m^{r_0} J_{r_1}(\tilde{q}_{r_1})\Omega^{-1} J_{r_o}(\tilde{g}_{r_o}), & 1 < r_1, r_o, r = r_o + r_1; \\ b = r, & Z = C_m^r J_r(\tilde{h}_r), & 1 < r_1, r_o, r < r_o + r_1; \\ b = r_o + r_1, & Z = -C_m^{r_1} C_m^{r_0} J_{r_1}(\tilde{q}_{r_1})\Omega^{-1} J_{r_o}(\tilde{g}_{r_o}), & 1 < r_1, r_o, r > r_o + r_1, \end{cases}$$

From Theorem 2 we see that the most interesting case is $r = r_o = r_1 = 1$, in which $\sqrt{n}(\tilde{U}_n - \theta)$ is asymptotic non-degenerate normal, with asymptotic variance being smaller than that of $\sqrt{n}(U_n - \theta)$. σ^2 is the same as that of U_n either when $r_1 > 1, A = 0$, or when $r_o > 1$, $A_1 = 0$ and $\Omega_1 = 0$. Thus, for the side information to be of practical meaning, we need $r = r_o = r_1 = 1$.

- An optimality property of $\tilde{U}_n$

$f(\cdot|\theta)$: density of X given θ, $\theta_n = \theta + n^{-1/2}b$ for some $b \in C$. An estimator $T_n = T_n(X_1, ..., X_n)$ is *regular*, if under $f(\cdot|\theta_n)$, $W_n := \sqrt{n}(T_n - \theta_n) \xrightarrow{D} W$ for some W, independent of $\{\theta_n\}$. Let $Z \bigoplus U$: convolution of Z and U, $I(\theta)$: Fisher infor at θ, and $Z \sim N(0, I^{-1}(\theta))$. Convolution Theorem (H$\acute{a}$jek, 1970): for any regular T_n with weak limit W, there is a U such that

$$W = Z \bigoplus U.$$

The optimal weak limit: a normal random variable with mean zero and variance $I^{-1}(\theta)$.

Now let $\mathbb{I}(\theta|g)$: infor. bound for estimating θ given side infor. in g.

Theorem 3. *Assume* $r = r_o = 1$, *(C4) and conditions in the Lemma , we have*

$$(i) \qquad \mathbb{I}(\theta|g) = \eta_1^2 - A_1' \Omega_1^{-1} A_1.$$

Thus, if we set $g(\mathbf{x}) = (g(x_1) + \cdots + g(x_m))/m$, *then* $rank(g) = 1$, $A = mA_1$, $\Omega = m\Omega_1$, $\sigma^2 = m^2 \mathbb{I}(\theta|g)$ *and* $\tilde{U}_n$ *is efficient.*
(ii) Assume further that $f(\cdot|\theta)$ *has second order continuous partial derivative with respect to* θ, *then for any regular estimator* T_n *with weak limit* W *of* $W_n := \sqrt{n}(T_n - \theta)$, W *can be decomposed as, for some* U,

$$W = Z \oplus U, \quad \textit{with} \quad Z \sim N(0, \mathbb{I}(\theta|g)).$$

U-statistic with side information of the form $\tilde{U}_n$ is regular, thus is optimal in the sense of convolution under the conditions of Theorem 3. Without side infor, asymptotic variance of $\sqrt{n}(U_n - \theta)$ is η_1^2; with side infor, asymptotic variance of $\sqrt{n}(\tilde{U}_n - \theta)$ is $\eta_1^2 - A_1'\Omega_1^{-1}A_1$, with a reduction of $A_1'\Omega_1^{-1}A_1$.

$\mathbb{I}(\theta|g)$: length of projection of $\tilde{h}_1(X)$ onto $[\tilde{g}_1(X)^{\perp}]$, the linear span of the orthogonal complements of $\tilde{g}_1(X)$. Increasing the components in g (and thus in $\tilde{g}_1$) shrinks the space $[\tilde{g}_1(X)^{\perp}]$, and shortens the length of the projection or increases the efficiency of $\tilde{U}_n$, or increasing the number of information constraints reduces the asymptotic variance of the U-statistic.

- Uniform SLLN and CLT of $\tilde{U}_n$-processes

Let $\tilde{P}_{n,m}$, $P_{n,m}$, P_m and P be the (random) probability measures induced by $\tilde{F}_{n,m}$, $F_{n,m}$, F_m and F respectively. For a function h, denote $\tilde{P}_{n,m}h = \sum_{\mathbf{i}\in D_{n,m}} w_{\mathbf{i}} h(\mathbf{X}_{\mathbf{i}})$, $P_m h = E_{P_m} h(\mathbf{X})$, $\tilde{\mathbb{G}}_{n,m}h = \sqrt{n}(\tilde{P}_{n,m}h - P_m h)$ and $\mathbb{G}_{n,m}h = \sqrt{n}(P_{n,m}h - P_m h)$. For fixed h and g, we have shown that, under suitable conditions,

$$\tilde{P}_{n,m}h \to P_m h = P\tilde{h}_1 \ \ (a.s.) \quad \text{and} \quad \tilde{\mathbb{G}}_{n,m}h \xrightarrow{D} N(0,\sigma^2)$$

with $\sigma^2 = \sigma^2(h) = P\tilde{h}_1^2 - P(\tilde{g}_1'\tilde{h}_1)\Omega_1^{-1}P(\tilde{g}_1\tilde{h}_1)$.

In contrast, $\mathbb{G}_{n,m}h \xrightarrow{D} N(0, \eta_1^2)$ with $\eta_1^2 = P\tilde{h}_1^2$. So incorporating the side information g reduces the asymptotic variance by the amount $P(\tilde{g}_1'\tilde{h}_1)\Omega_1^{-1}P(\tilde{g}_1\tilde{h})$.

It is of interest to have a uniformly version of the above SLLN and CLT over a class of functions $\mathcal{H}$.

Theorem 4. *(i) Under the conditions of Theorem 1(i), and some further conditions, we have*

$$\sup_{h \in \mathcal{H}} |\tilde{P}_{n,m} h - P_m h| = 0, \quad (a.s.^*).$$

(ii) Under the conditions of Theorem 3(ii), and further conditions, then

$$\tilde{\mathbb{G}}_{n,m} \stackrel{D}{\Rightarrow} \mathbb{G} \ \text{ in } \ L^\infty(\mathcal{H}),$$

where $\mathbb{G}$ *is a Gaussian process indexed by* $\mathcal{H}$*, with* $E_P(\mathbb{G}h) = 0$ *and* $Cov_P(\mathbb{G}h, \mathbb{G}q) = P(\tilde{h}_1 \tilde{q}_1) - P(\tilde{g}_1' \tilde{h}_1)\Omega_1^{-1} P(\tilde{g}_1 \tilde{q}_1)$ *for all* $h, q \in \mathcal{H}$*.*

- Empirical Likelihood Ratio for U-stat. with Side Infor.

Let $G(\mathbf{x}|\theta) = (g'(\mathbf{x}), h(\mathbf{x}) - \theta)'$, then $E_{F_m} G(\mathbf{X}|\theta) = 0$. We define the empirical log likelihood ratio of θ with presence of side infor by

$$R_G(\theta) = L_n(\theta)/(C_n^m)^{-C_n^m} = \prod_{\mathbf{i}\in D_{n,m}} (C_n^m w_{\mathbf{i}}),$$

where

$$L_n(\theta) = \max_{\sum_{\mathbf{i}\in D_{n,m}} w_{\mathbf{i}}=1, \sum_{\mathbf{i}\in D_{n,m}} w_{\mathbf{i}} G(\mathbf{X}_{\mathbf{i}}|\theta)=0} \prod_{\mathbf{i}\in D_{n,m}} w_{\mathbf{i}}$$

and denote

$$l(\theta) = -\log R_G(\theta) = \sum_{\mathbf{i}\in D_{n,m}} \log[1 + t' G(\mathbf{X_i}|\theta)].$$

Let $\Lambda = E_{F_m}(G(\mathbf{x}|\theta)G'(\mathbf{X}|\theta)) = \begin{pmatrix} \Omega & A \\ A' & \eta^2 \end{pmatrix}$, $\eta^2 = Var(h(\mathbf{X}))$;

and $\Lambda_1 = Cov(\tilde{G}_1)$, $\tilde{G}_1$ the first canonical form (vector) of G. Without side infor, $G(\cdot|\theta)$ reduces to $h(\cdot) - \theta$, and t is a scalar determined by $\sum_{\mathbf{i}\in D_{n,m}} (h(\mathbf{X_i}) - \theta)/[1 + t(h(\mathbf{X_i}) - \theta)] = 0$. The corresponding log-likelihood ratio is

$$l_h(\theta) = \sum_{\mathbf{i}\in D_{n,m}} \log[1 + t(h(\mathbf{X_i}) - \theta)].$$

Theorem 5. *(i) Under conditions of Theorem 2(i) or Theorem 3(i) and assume* Λ *to be positive definite, then*

$$\frac{2n}{m^2 C_n^m} l(\theta) \xrightarrow{D} Z'_{d+1} \Lambda_1^{1/2} \Lambda^{-1} \Lambda_1^{1/2} Z_{d+1}, \quad Z_{d+1} \sim N(0, I_{d+1}).$$

(ii) Assume (C4), then

$$\frac{2n\eta^2}{m^2 C_n^m \eta_1^2} l_h(\theta) \xrightarrow{D} \chi_1^2.$$

When $m = 1$, $\Lambda_1^{1/2} = \Lambda^{1/2}$ and the above result for U-statistic automatically reduces to that for the common EL ratio, and the right hand side in Theorem 5(i) is χ^2_{d+1}.

Corollary. *If $E_{F_m} g(\mathbf{X}) = \delta \neq 0$, then*
(i) Under conditions of Theorem 1(i),

$$\tilde{U}_n - \theta \to A'\Omega^{-1}\delta.$$

(ii) Under conditions of Theorem 2(i),

$$\sqrt{n}(\tilde{U}_n - \theta - A'\Omega^{-1}\delta) \approx N(0, \sigma^2).$$

(iii) If $E_{F_m} G(\mathbf{X}) = \delta \neq 0$, then under conditions of Theorem 5(i),

$$-\frac{2n}{C_n^m} R_G(\theta) \approx Z'_{d+1} \Lambda_1^{1/2} \Lambda^{-1} \Lambda_1^{1/2} Z_{d+1}, \quad Z_{d+1} \sim N(\sqrt{n}\Lambda_1^{-1/2}\delta, I_{d+1}),$$

when $\Lambda = \Lambda_1$, $Z'_{d+1}\Lambda_1^{1/2}\Lambda^{-1}\Lambda_1^{1/2} Z_{d+1} = \chi^2_{d+1}(n\delta'\Lambda^{-1}\delta)$, the chi-squared distribution of degree $d+1$ with noncentrality parameter $n\delta'\Lambda^{-1}\delta$.

Examples

- Example 1

$\theta(F) = \int (x-\mu)^2 dF(x)$ be the variance, μ the mean. Let $\mu_k, k \geq 2$ be the k-th moment of F. For the kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$, we have $\tilde{h}_1(x_1) = [(x_1-\mu)^2 - \theta]/2$, $\eta^2 = E(h^2) - \theta^2 = (\mu_4 + \theta^2)/2$, $\eta_1^2 = E(\tilde{h}_1^2) = (\mu_4 - \theta^2)/4$. Without side infor, the asymptotic variance of U_n based on kernel $h(x_1, x_2)$ is $\sigma_0^2 = 4\eta_1^2 = \mu_4 - \theta^2$, the same as that for the sample variance estimator $\theta_n := \sum_{i=1}^n (X_i - \overline{X})^2$.

If we know that F has median at 0: $F(0) = 1/2$, we take $g(x_1, x_2) = [I(x_1 \leq 0) + I(x_2 \leq 0)]/2 - 1/2$. Then $\tilde{g}_1(x_1) = [I(x_1 \leq 0) - 1/2]/2$, $A_1 = E(\tilde{g}_1 \tilde{h}_1) = [\int_{-\infty}^{0} (x - \mu)^2 dF(x) - \theta/2]/4$, and $\Omega_1 = E(\tilde{g}_1^2) = 1/16$. So by Theorem 3(i), the asymptotic variance of $\tilde{U}_n$ is now $\sigma^2 = \sigma_0^2 - A_1^2 \Omega_1^{-1} = 4\eta_1^2 - [\int_{-\infty}^{0} (x - \mu)^2 dF(x) - \sigma^2/2]^2$, a deduction of $[\int_{-\infty}^{0} (x - \mu)^2 dF(x) - \sigma^2/2]^2$ from σ_0^2.

- Example 2

Wilcoxon one-sample statistic $\theta(F) = P_F(x_1 + x_2 \leq 0)$, kernel for corresponding U-statistic: $h(x_1, x_2) = I(x_1 + x_2 \leq 0)$. Then $\tilde{h}_1(x_1) = F(-x_1) - \theta$, $\eta_1^2 = E_F(\tilde{h}_1(x_1)) = \int F^2(-x)dF(x) - \theta^2$. Without side infor, asymptotic variance of U_n based on $h(x_1, x_2)$ is $\sigma_0^2 = 4\eta_1^2$.

If we know the distribution is symmetric about $a > 0$: $F(x - a) = 1 - F(a - x)$ for all x. Take $g(x_1, x_2) = [I(x_1 \le 0) + I(x_1 \le 2a) + I(x_2 \le 0) + I(x_2 \le 2a)]/2 - 1$, then $\tilde{g}_1(x_1) = [I(x_1 \le 0) + I(x_1 \le 2a)]/2 - 1/2$, $\Omega_1 = F(-a)/2$, $A_1 = [\int_{-\infty}^{a} F(-x)dF(x) + \int_{-\infty}^{-a} F(-x)dF(x)]/2 - \int F(-x)dF(x)/2$, and the deduction of asymptotic variance is $A_1^2 \Omega^{-1}$.

- Example 3

Gini difference: $\theta(F) = E_F|x_1 - x_2|$. corresponding kernel for U-stat.: $h(x_1, x_2) = |x_1 - x_2|$. Then
$\tilde{h}_1(x_1) = \int_{x_1}^{\infty} x dF(x) - \int_{-\infty}^{x_1} x dF(x) - \theta$,
$\eta_1^2 = \int \left(\int_{x_1}^{\infty} x dF(x) - \int_{-\infty}^{x_1} \right)^2 dF(x_1) - \theta^2$. Without side infor, asymptotic variance of U_n based on kernel $h(x_1, x_2)$ is $\sigma_0^2 = 4\eta_1^2$.

If we know the distribution mean μ, and take $g(x_1, x_2) = (x_1 + x_2)/2 - \mu$, then $\tilde{g}_1(x_1) = (x_1 - \mu)/2$, $\Omega_1 = \int (x - \mu)^2 dF(x)$, $A_1 = \{\int x_1[\int_{x_1}^{\infty} xdF(x) - \int_{-\infty}^{x_1} xdF(x)]dF(x_1) - \theta\}/2$, and the deduction of asymptotic variance is $A_1^2 \Omega^{-1}$.

Simulation Studies

Consider Examples 1 and 2 above.

- Example 1

Table 1: asymp variance estimation of U-stat. $X \sim \exp(1) - \ln(2)$

Method	n=50	n=100	n=150	n=200
Without side infor	8.5239	7.8569	7.3839	7.1557
With side infor	8.4572	7.5524	7.2673	7.0791
Variance reduction	0.0667	0.3045	0.1165	0.0766

- Example 2

Table 2: asymp variance estimation of U-stat. $X \sim \mathcal{N}(1, 4)$

Method	n=50	n=100	n=150	n=200
Without side infor	0.2413	0.2208	0.2199	0.2203
With side infor	0.0548	0.0526	0.0527	0.0572
Variance reduction	0.1865	0.1682	0.1673	0.1631

From Tables 1 and 2 we see reductions of the variance of estimating θ. Sometimes the reduction is significant, like in Example 2, which means the proposed method gives more accurate estimation.

Summary

- U-stat side infor., via EL approach;
- some asymp behavior
- smaller asymp. variance.
- efficiency
- confi. intervals using such U-stat. via EL ratio.

- References

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, **19**, 293-325.

Koroljuk, V.S. and Borovskich, Yu.V. (1994). *Theory of U-Statistics*, Kluwer Academic Publishers, The Netherlands.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, **75**, 237-249