DESIGN BASED INCOMPLETE U-STATISTICS

Xiangshun Kong[1], Wei Zheng[2]

[1]*Beijing Institute of Technology and* [2]*University of Tennessee*

Generalization of Theorem 2.

The following conditions on g or F will be needed by Theorem 2 in Section 2 and Theorems 7–9 in this section.

(g.1) *Lipschitz continuous*: The function, $g : R^d \to R$, is said to be Lipschitz continuous if there exists a constant $c > 0$ such that $|g(\mathbf{a_1}) - g(\mathbf{a_2})| \leq c||\mathbf{a_1} - \mathbf{a_2}||_2$ for any $\mathbf{a_1}, \mathbf{a_2} \in R^d$. Example: First-order polynomial functions.

(g.2) *Order-p continuous*: The function, $g : R^d \to R$, is said to be order-p continuous if there exists a constant $c > 0$ and $\phi_p(\mathbf{a_1} - \mathbf{a_2}) \leq c + \max^p(||\mathbf{a_1}||_2, ||\mathbf{a_2}||_2)$ for any $\mathbf{a_1}, \mathbf{a_2} \in R^d$ such that $|g(\mathbf{a_1}) - g(\mathbf{a_2})| \leq \phi(\mathbf{a_1}, \mathbf{a_2})||\mathbf{a_1} - \mathbf{a_2}||_2$ for any $\mathbf{a_1}, \mathbf{a_2} \in R^d$. Example: All polynomial functions.

(g.3) *Uniformly bounded-variation*: For a real valued function $f : R \to R$, the total variation of f is defined as $V_R(f) = \sup_{p>0} \sup_{-\infty<c_1,\ldots,c_p<\infty} \sum_{i=1}^{p-1} |f(c_{i+1}) - f(c_i)|$. The function, $g : R^d \to R$, is said to be uniformly bounded-variation if there exists a constant $c > 0$ such that $V_R(g(\cdot, x_2, \ldots, x_d)) < c$ for any $(x_2, \ldots, x_d) \in R^{d-1}$. Example: Linear combinations of sign functions, e.g. $g(x_1, x_2) = \text{sign}(x_1x_2) +$

sign$(x_1 + x_2)$.

(F) *Light-tailed* distribution: The distribution of a random variable X is said to be light-tailed if there exists constants $c, c_1 > 0$ such that $P(|X| > x) \leq e^{-cx}$ for all $x > c_1$. Example: Normal distribution, exponential distribution, and truncated distributions.

Lemma 4. *Suppose F is light-tailed. Let $X_{\max} = \max\{|X_1|, \ldots, |X_n|\}$. Then, for arbitrary $a > 0$ with $n \to \infty$, we have*

$$EX_{\max}^a = O(\log n)^a.$$

Proof. Since the distribution is light-tailed, we have $P(|X| > x) \leq e^{-cx}$ for any $|x| > c_0$, where c and c_0 are two fixed positive numbers.

$$\begin{aligned}
E(X_{\max})^a &= \int_{x>0} ax^{a-1} P(X_{\max} > x) dx \\
&\leq \int_0^{2c^{-1}\log n} ax^{a-1} dx + \int_{2c^{-1}\log n}^{\infty} ax^{a-1} P(X_{\max} > x) dx \\
&= O(\log n)^a + \int_{2c^{-1}\log n}^{\infty} ax^{a-1} P(X_{\max} > x) dx \\
&= O(\log n)^a + \int_{2c^{-1}\log n}^{\infty} ax^{a-1} n e^{-cx} dx \;=\; O(\log n)^a + O(1). \quad \square
\end{aligned}$$

Lemma 5. *Suppose (i) g is order-p continuous, and (ii) F is light-tailed. We have*

$$E(U_{oa} - \bar{V})^2 = O\left(\frac{1}{mL}(\log n)^{2p+2}\right).$$

Proof. Let $X_{\max} = \max\{|X_1|, \ldots, |X_n|\}$. For $l \in \mathcal{Z}_L$, define $d_l = \max\{|X_{i_1} - X_{i_2}| : i_1, i_2 \in G_l\}$. Since g is order-p continuous, for $\boldsymbol{\eta} \sim \boldsymbol{\eta}'$ in $\mathcal{G}_{\boldsymbol{a}}$, $|g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'})| \leq (c_1 + X_{\max}^p) d^{1/2} d_l$, and so $|g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'})|^2 \leq (c_1 + X_{\max}^p)^2 \cdot d \cdot \sum_{j=1}^d d_{a_j}^2$.

Since U_{oa} and $\bar{V}$ always use the same $S_{oa} = \{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^m\}$, we have

$$E(U_{oa} - \bar{V})^2 = E\left(\frac{1}{m}\sum_{i=1}^m (g(\mathcal{X}_{\boldsymbol{\eta}^i}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}^i}))\right)^2.$$

For $\boldsymbol{i}_1 \neq \boldsymbol{i}_2$, $E(g(X_{\boldsymbol{\eta}^{i_1}}) - \bar{g}(X_{\boldsymbol{\eta}^{i_1}}))(g(X_{\boldsymbol{\eta}^{i_2}}) - \bar{g}(X_{\boldsymbol{\eta}^{i_2}})) = 0$.

$$\begin{aligned} E(U_{oa} - \bar{V})^2 &= m^{-2} E \sum_{i=1}^m (g(\mathcal{X}_{\boldsymbol{\eta}^i}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}^i}))^2 \\ &\leq m^{-2} E \sum_{i=1}^m (c_1 + X_{\max}^p)^2 \cdot d \cdot \sum_{j=1}^d d_{a_j^i}^2 \end{aligned}$$

Since $\sum_{l=1}^L d_l \leq 2X_{\max}$, we have $\sum_{l=1}^L d_l^2 \leq 4X_{\max}^2$. Using Lemma 4, we have

$$\begin{aligned} E(U_{oa} - \bar{V})^2 &\leq m^{-2} d E\left((c_1 + X_{\max}^p)^2 \sum_{i=1}^m \sum_{j=1}^d d_{a_j^i}^2\right) = m^{-2} d E\left((c_1 + X_{\max}^p)^2 \sum_{j=1}^d \sum_{i=1}^m d_{a_j^i}^2\right) \\ &= m^{-2} d E\left((c_1 + X_{\max}^p)^2 \sum_{j=1}^d m L^{-1} 4 X_{\max}^2\right) = O\left(\frac{1}{mL}(\log n)^{2p+2}\right). \quad \square \end{aligned}$$

Theorem 7. *Suppose (i) The kernel function g is order-p continuous, and (ii) F is light-tailed. For U_{oa} based on $OA(m, d, L, t)$, we have*

$$\mathrm{MSE}(U_{oa}) = \mathrm{MSE}(U_0) + \frac{R(t)}{m} + O\left(\frac{(\log n)^{2p+2}}{mL}\right) + O\left(\frac{1}{n^2}\right). \tag{6.14}$$

Proof. This is the direct result of (6.6), Lemma 1(*ii*), Lemmas 2 and 5. $\square$

Theorem 8. *Suppose the kernel function g has uniformly bounded variation. For U_{oa} based on $OA(m,d,L,t)$, we have*

$$\mathrm{MSE}(U_{oa}) = \mathrm{MSE}(U_0) + \frac{R(t)}{m} + O\left(\frac{1}{mL}\right) + O\left(\frac{1}{n^2}\right). \tag{6.15}$$

Proof. From (6.6), Lemma 1(ii) and Lemma 2, we only need to prove $E(U_{oa} - \bar{V})^2 = O(m^{-1}L^{-1})$. First, we introduce some notations that will be used only in the proof of this theorem. Given the order statistic of $\{X_1, \ldots, X_n\}$ denoted by $X_{(1)}, \ldots, X_{(n)}$, for $l = 1, \ldots, L$ and $(x_2, \ldots, x_d) \in R^{d-1}$, define $D(l|x_2, \ldots, x_k) = \max_{(l-1)nL^{-1} < i_1 < i_2 \leq l \cdot nL^{-1}} |g(X_{(i_1)}, x_2, \ldots, x_k) - g(X_{(i_2)}, x_2, \ldots, x_k)|$. Since g has uniformly bounded variation, g is bounded, say $|g| \leq M$.

$$\begin{aligned} E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2 | \boldsymbol{\eta} \sim \boldsymbol{\eta}'] &= L^{-d} \sum_{\boldsymbol{a} \in \mathcal{Z}_L^d} |\mathcal{G}_{\boldsymbol{a}}|^{-2} \sum_{\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}} \sum_{\boldsymbol{\eta}' \in \mathcal{G}_{\boldsymbol{a}}} (g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2 \\ &\leq 2ML^{-d} |\mathcal{G}_{\boldsymbol{a}}|^{-2} \sum_{\boldsymbol{a} \in \mathcal{Z}_L^d} \sum_{\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}} \sum_{\boldsymbol{\eta}' \in \mathcal{G}_{\boldsymbol{a}}} |g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'})|. \end{aligned}$$

Note that $g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'})$ can be written as the summation of the difference in changing each element of $\mathcal{X}_{\boldsymbol{\eta}} = (X_{\eta_1}, \ldots, X_{\eta_d})$ to $\mathcal{X}_{\boldsymbol{\eta}'} = (X_{\eta'_1}, \ldots, X_{\eta'_d})$ one by one as follows.

$$\begin{aligned} & |g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'})| \\ = \ & |g(X_{\eta_1}, X_{\eta_2}, \cdots) - g(X_{\eta'_1}, X_{\eta_2}, \cdots)| + |g(X_{\eta'_1}, X_{\eta_2}, X_{\eta_3}, \cdots) - g(X_{\eta'_1}, X_{\eta'_2}, X_{\eta_3}, \cdots)| \\ + \ & \cdots + |g(X_{\eta'_1}, X_{\eta'_2}, X_{\eta'_3}, \cdots, X_{\eta'_{d-1}}, X_{\eta_d}) - g(X_{\eta'_1}, X_{\eta'_2}, X_{\eta'_3}, \cdots, X_{\eta'_{d-1}}, X_{\eta'_d})| \\ \leq \ & D(a_1 | X_{\eta_2}, \ldots, X_{\eta_d}) + D(a_2 | X_{\eta'_1}, X_{\eta_3}, \ldots, X_{\eta_d}) + \cdots + D(a_d | X_{\eta'_1}, X_{\eta'_3}, \ldots, X_{\eta'_{d-1}}) \end{aligned}$$

For orthogonal arrays, we can separate $\sum_{\boldsymbol{a} \in \mathcal{Z}_L^d} \sum_{\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}} \sum_{\boldsymbol{\eta}' \in \mathcal{G}_{\boldsymbol{a}}} D(a_1 | X_{\eta_2}, \ldots, X_{\eta_d})$

into $|\mathcal{Z}_L^d||\mathcal{G}_{\boldsymbol{a}}|^2/L$ groups such that each group contains L elements whose summation is control by the total variation $c > 0$. So we have

$$\sum_{\boldsymbol{a}\in\mathcal{Z}_L^d}\sum_{\boldsymbol{\eta}\in\mathcal{G}_{\boldsymbol{a}}}\sum_{\boldsymbol{\eta}'\in\mathcal{G}_{\boldsymbol{a}}} D(a_1|X_{\eta_2},\ldots,X_{\eta_d}) \le cL^d|\mathcal{G}_{\boldsymbol{a}}|^2/L.$$

Similarly analyzing the $D(a_2|X_{\eta_1'},X_{\eta_3},\ldots,X_{\eta_d})$, $\ldots$, $D(a_d|X_{\eta_1'},X_{\eta_3'},\ldots,X_{\eta_{d-1}'}$, we have $E[(g(\mathcal{X}_{\boldsymbol{\eta}})-g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta}\sim\boldsymbol{\eta}'] = O(L^{-1})$ and so $E(U_{oa}-\bar{V})^2 = O(m^{-1}L^{-1})$. Theorem 8 is the direct result of (6.6), Lemma 1(ii), Lemma 2. $\square$

Theorem 9. *Suppose* (i) *The kernel function* g *is a linear combination of some order-*p *continuous functions and some uniformly bounded-variation functions, and* (ii) F *is light-tailed. Then (6.14) still holds with* $L^2 \le n(\log n)^{-1}$.

Proof. This is the direct result of Theorems 7 and 8.

Choosing L and t.

From Eq(2.13) of Theorem 3 in the manuscript and the relation $m = \lambda L^t$, we know that the trade-off between L and t depends on the variance of each component in the Heoffding's decomposition, i.e., δ_j^2, $j = 1,\ldots,d$. We shall give these variances a estimator $\hat{\delta}_j^2$. Using Eq(2.13) with $R(t)$ and $E\gamma^2(X_1,\ldots,X_d)$ being estimated as a function of $\hat{\delta}_j^2$, we should choose the combination of L and t which minimizes

$$\phi(L,t) = \frac{\hat{R}(t)}{m} + \frac{d}{12mL^2}\hat{E}\gamma^2(X_1,\ldots,X_d),$$

where $\hat{R}(t)$ and $\hat{E}\gamma^2(X_1,\ldots,X_d)$ are functions of $\hat{\delta}_j^2$'s.

Now we provide two methods for generating $\hat{\delta}_j^2$. (1) When the Heoffding's decomposition is easy to calculate, one can write down the analytical expression and give a direct estimation of δ_j^2's. (2) We can use a bootstrap approach for $\hat{\delta}_j^2$'s. With a small sample size $n' \ll n$, it is easy to bootstrap MSE(U_0) (the complete U-statistic). For details of the bootstrap approach, we may refer to Marie Huskova and Paul Janssen (1993a,b). Now, let us review the formula of MSE(U_0):

$$\mathrm{MSE}(U_0) = \binom{n}{d}^{-1} \sum_{j=1}^{d} \binom{d}{j}\binom{n-d}{d-j}\sigma_j^2 \;=\; \sum_{j=1}^{d} \binom{d}{j}^2 \binom{n}{j}^{-1} \delta_j^2.$$

Usually, with at most d different $n'(> d)$, we can generate linear equations of δ_j^2 based on the d different $\widehat{\mathrm{MSE}}(U_0)$ based on the bootstrap approach. And the solution of these linear equations can be used as the estimation of $\hat{\delta}_j^2$'s.

For the second method, we now use the setup in Example 1 for illustration. For convenience, we set $n = 10^4$ and $m = 10^6$. The two choices of the combination of L and t is $(L = 100, t = 3)$ and $(L = 1000, t = 2)$. We use bootstrap method to estimate the variance of the complete U-statistic with $n' = 4, 5, 6$. The subsample size n' is so small that the computational burden of the bootstrapped complete U-statistic, i.e., $\binom{n'}{3}$ is negligible. Simulation reveals that $\hat{\delta}_1 = 0.0557$, $\hat{\delta}_2 = 0.00217$ and $\hat{\delta}_3 = 1.06257$. Simple analysis reveals that $t = 3$ shall work better than $t = 2$, which is verified by the simulation result. Actually, with $m = 10^6$, the efficiency of U_{oa} is 100.0% when $t = 3$ and 97.88% when $t = 2$.

Examples for multi-sample and multi-dimensional cases. Consider the multi-sample case. Suppose $d_1 = d_2 = 2$, $n_1 = n_2 = 9$ and the two samples are

$$X_6^{(1)} \le X_8^{(1)} \le X_2^{(1)} \le X_4^{(1)} \le X_7^{(1)} \le X_5^{(1)} \le X_3^{(1)} \le X_9^{(1)} \le X_1^{(1)}.$$

$$X_2^{(2)} \le X_7^{(2)} \le X_3^{(2)} \le X_6^{(2)} \le X_1^{(2)} \le X_4^{(2)} \le X_5^{(2)} \le X_9^{(2)} \le X_8^{(2)}.$$

Then we have $L = 3$ groups listed as $G_1^{(1)} = \{6, 8, 2\}, G_2^{(1)} = \{4, 7, 5\}, G_3^{(1)} = \{3, 9, 1\}$ and $G_1^{(2)} = \{2, 7, 3\}, G_2^{(2)} = \{6, 1, 4\}, G_3^{(2)} = \{5, 9, 8\}$. An example of $OA(m = 9, d = 4, L = 3, t = 2)$ in step 1 is given as follows in transpose.

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 1 & 2 & 3 & 1 & 2 & 3 \\ 1 & 2 & 3 & 2 & 3 & 1 & 3 & 1 & 2 \\ 1 & 2 & 3 & 3 & 1 & 2 & 2 & 3 & 1 \end{pmatrix}.$$

Then we could possibly have the $\mathcal{X}_{\eta^i}$, $i = 1, \ldots, 9$, used in the construction of 9-run multi-sample construction as follows.

$$\{\mathcal{X}_{\eta^1}, \ldots, \mathcal{X}_{\eta^9}\} = \left\{ \begin{array}{ccccccccc} X_8^{(1)} & X_2^{(1)} & X_6^{(1)} & X_4^{(1)} & X_4^{(1)} & X_5^{(1)} & X_9^{(1)} & X_1^{(1)} & X_9^{(1)} \\ X_6^{(1)} & X_7^{(1)} & X_3^{(1)} & X_8^{(1)} & X_7^{(1)} & X_1^{(1)} & X_6^{(1)} & X_7^{(1)} & X_3^{(1)} \\ X_7^{(2)} & X_1^{(2)} & X_5^{(2)} & X_4^{(2)} & X_8^{(2)} & X_3^{(2)} & X_9^{(2)} & X_2^{(2)} & X_6^{(2)} \\ X_3^{(2)} & X_6^{(2)} & X_9^{(2)} & X_5^{(2)} & X_3^{(2)} & X_1^{(2)} & X_6^{(2)} & X_8^{(2)} & X_3^{(2)} \end{array} \right\}.$$

Consider the multi-dimensional case. Suppose $X_1 = (1.0, 3.2)$, $X_2 = (0.9, 1.0)$, $X_3 = (0.9, 3.1)$, $X_4 = (0.8, 2.1)$, $X_5 = (0.7, 2.2)$, $X_6 = (0.9, 1.2)$, $X_7 = (0.9, 1.9)$, $X_8 = (0.8, 1.1)$, $X_9 = (0.9, 2.8)$. Simple clustering methods reveal $G_1 = \{6, 8, 2\}, G_2 = \{4, 7, 5\}, G_3 = \{3, 9, 1\}$. The choosing of η^i, $i = 1, \ldots, 9$, might be the same as (2.9).