Incomplete/Reduced U-Statistics

Wei Zheng
Department of Business Analytics and Statistics
University of Tennessee
(Joint with Xiangshun Kong)

Design of Experiments: New Challenges
Marseille, France
May 02 2018

Major difference?

A closer look

Reduction in size

Definition of U-statistics

- Suppose $X_1, X_2, ..., X_n \overset{iid}{\sim} F \in \mathcal{F}$, where $\mathcal{F}$ could be any set of distribution defined on $\mathbb{R}$.
- We are interested in estimating $\theta = \theta(F) = \mathbb{E}g(X_1, ..., X_k)$, where $g : \mathbb{R}^k \to \mathbb{R}$ is a symmetric kernel function of order k.
- Let $S_0 = \{\boldsymbol{i} = (i_1, i_2, ..., i_k) : 1 \leq i_1 < i_2 < ... < i_k \leq n\}$ be the collection of all size k subset of $\{1, 2, ..., n\}$.
- The regular U-statistics is defined as

$$\begin{aligned} U_0 &= \binom{n}{k}^{-1} \sum_{\boldsymbol{i} \in S_0} g_{\boldsymbol{i}} \\ g_{\boldsymbol{i}} &= g(X_{i_1}, X_{i_2}, ..., X_{i_k}) \end{aligned}$$

- It is well known that U_0 has the smallest variance among all unbiased estimators of θ for any given F.
- One big issue: Its computational complexity is $O(n^k)$.

Incomplete U-statistics

- Imagine the data size as $n = 1000$ and the order of kernel as $k = 3$, the total number of terms to be averaged is already $\binom{1000}{3} \approx 166$ millions. (roughly 6 mins on a daily desktop)
- Adding a zero to n: $\binom{10000}{3} \approx 166$ billion. (100 hours)
- Due to the computational burden even for moderate size of data, we need to approximate U_0 by an incomplete U-statistics

$$U = m^{-1} \sum_{\boldsymbol{i} \in S} g_{\boldsymbol{i}}$$

where S is a sample of elements from S_0 with or without replacement and $|S| = m << \binom{n}{k}$.

Random sampling: case 1

$m/\binom{n}{k}$	efficiency (%)
0.2	$\simeq 100$
0.1	99.54
0.04	95.80
0.02	80.77
0.01	53.22
0.006	23.01

Table 1: The performance of incomplete U-statistics when S is a random sample of the elements in S_0 with replacement at the setting of $n = 1000$, $k = 2$, and $g(X_1, X_2) = (X_1 - X_2)^2/2$, where $X_i \sim N(0, 1)$.

Random sampling: case 2

$m/\binom{n}{k}$	time	efficiency (%)
1×10^{-3}	0.35 sec	$\simeq 100$
5×10^{-4}	0.03 sec	99.88
4×10^{-4}	0.023 sec	95.81
3×10^{-4}	0.017 sec	94.13
2×10^{-4}	0.012 sec	85.00
1×10^{-4}	3.3 μsec	81.82
6×10^{-5}	1.9 μsec	26.52

Table 2: The performance of incomplete U-statistics when S is a random sample of the elements in S_0 with replacement at the setting of $n = 1000$, $k = 3$, and $g(X_1, X_2, X_3) = \frac{1}{3}(\text{sign}(2X_1 - X_2 - X_3) + \text{sign}(2X_2 - X_1 - X_3) + \text{sign}(2X_3 - X_1 - X_2))$. The computation of the complete U-statistics takes around 6 minutes.

Insights on random sampling

- The random sample is drawn from the combination pool S_0 instead of the original data $\{1, 2, .., n\}$.
- Blom (1976): The variance of the incomplete U-statistic based on the random sampling scheme is

$$V(U_{RND}) = \frac{\sigma_g^2}{m} + (1 - \frac{1}{m})V(U_0)$$

  where $\sigma_g^2 = V(g(X_1, ..., X_k)) > V(U_0)$.
- For non-degenerated case, $V(U_0) \asymp 1/n$, so softly speaking
  - If $m/n \to 0$, we have $V(U_{RND}) \approx \sigma_g^2/m$.
  - If $m/n \to \alpha \in (0, \infty)$, we have $V(U_{RND}) \approx V(U_0) + \alpha^{-1}\sigma_g^2/n$.
  - If $m/n \to \infty$, $V(U_{RND}) \approx V(U_0)$.
- The takeaway: Instead of computing the complete U-statistics at the computational cost of $O(n^k)$, the random incomplete U-statistic with $m \succ n$ shall achieve the same variance asymptotically.

Literature Review

- Incomplete U-statistic: Blom (1976).
- Reduced U-statistic: Brown and Kildea (1978).
- Constructions: Lee (1982), Rempala and Wesolowski (2003), Rempala and Srivastav (2004).
- Statistical properties: Lee (1979), Janson (1984).
- Multi-sample and machine learning: Clemencon et al (2016), Colin (2016).
- High dimensional case: Chen (2017), Chen and Kato (2017).

Some basics of the regular U-statistics

- For $1 \le c \le k$, let $g_c(x_1, ..., x_c) = \mathbb{E}g(x_1, ..., x_c, X_{c+1}, ..., X_k)$.
- Define the projections $h^{(1)}(x_1) = g_1(x_1) - \theta$ and

$$h_c(x_1, ..., x_c) = g_c(x_1, ..., x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} h_j(x_{i_1}, ..., x_{i_j}) - \theta$$

- Hoeffding decomposition (1948):

$$U_0 = \theta + \sum_{c=1}^{k} \binom{k}{c} H_c,$$

where H_c is a U-statistics defined on the kernel function h_c.
- By the projection property, we have

$$V(U_0) = \sum_{c=1}^{k} \binom{k}{c}^2 \binom{n}{c}^{-1} \delta_c^2 \quad (1)$$

where $\delta_c^2 = V(h_c(X_1, ..., X_c))$.

Variance of an incomplete U-stat

- Lee (1982): the variance of $U = m^{-1} \sum_{\boldsymbol{i} \in S} g_{\boldsymbol{i}}$ is

$$
\begin{aligned}
V(U) &= \sum_{c=1}^{k} \eta_c \delta_c^2 \\
\eta_c &= m^{-2} \sum_{(n,c)} \lambda(i_1, ..., i_c)^2
\end{aligned}
$$

  where $\lambda(i_1, ..., i_c) = \#\{\boldsymbol{i} \in S : \{i_1, ..., i_c\} \subset \boldsymbol{i}\}$ is the number of k-tuples (blocks) in S containing the c-tuple $(i_1, ..., i_c)$.
- Since $\sum_{(n,c)} \lambda(i_1, ..., i_c) = m\binom{k}{c}$, the quantity η_c is minimized when $\lambda(i_1, ..., i_c)$ differs from each other by 1 or 0 all c-tuples.
- For given m, an equal replicate design minimizes η_1
- For given m, **A BIBD minimizes** η_1 **and** η_2**.**
- PBIBD with two associate classes ($\lambda_1 = 1$ and $\lambda_0 = 0$).

BIBD

- When $k = 3$, a BIBD minimizes the variance of incomplete U-statistics among all designs with the same m.
- When $k \geq 4$, a BIBD with $\lambda = 1$ minimizes the variance of incomplete U-statistics among all designs with the same m.
- We will see later that the above statements of optimality is only conditionally true.
- Raghavarao (1971): For each integer t, there exist a BIBD for $n = 6t + 3$ (data size), $m = (3t + 1)(2t + 1)$.
- However, m is forced to be at the scale of $m \asymp n^2$.
- Recall random sampling only require $m \succ n$.
- In Example 2 with $n = 1000$, we have $m = 166,167$ for the BIBD. This makes the ratio $m/\binom{1000}{3} = 0.001$, where **the random sampling reaches the efficiency of nearly 100%.**
- Similar observations for PBIBD.

Permanent design

- Introduced by Rempala and Wesolowski (2003).
- Suppose n is divisible by k and denote $t = n/k$.
- Randomly split $\{1, 2..., n\}$ into k disjoint sets M_1,...,M_k, each of size t.
- Form t^k distinct k-tuples by selecting one element from each of M_1,...,M_k. Let S be the set of all k-tuples such formed.
- Define the incomplete U-statistic by

$$U = t^{-k} \sum_{\boldsymbol{i} \in S} g_{\boldsymbol{i}}$$

- We have $m = O(n^k)$ for this algorithm, so not attractive considering BIBD already yields the efficiency close to 1.

Rectangular design

- Introduced by Rempala and Srivastav (2004).
- Arrange the data by a $k \times t$ array

$$\begin{gathered} X_{1,1}, \ldots, X_{1,t} \\ X_{2,1}, \ldots, X_{2,t} \\ \cdots \\ X_{k,1}, \ldots, X_{k,t} \end{gathered}$$

- Definition of Rectangular scheme
  - S consists of k-tuples with one element from each row: $\{X_{1,i_1}, \ldots, X_{k,i_k}\}$.
  - It contains all 2-subsets of the form $\{X_{i,j}, X_{k,l}\}$ where $i \neq k$ and $j \neq l$.
- It is essentially a subset of permanent design with 2-dimensional projection property.
- **Like BIBD, it also enforces $m \asymp n^2$**

Rectangular design vs BIBD

Method	time	efficiency (%)	m
Rectangular design	0.32 sec	$\simeq 100$	117306*
BIBD	0.44 sec	$\simeq 100$	166167*
Random sampling	0.35 sec	$\simeq 100$	166167

Table 3: Comparison of the three methods in Example 2: $n = 1000$, $k = 3$, and $g(X_1, X_2, X_3) = \frac{1}{3}(\text{sign}(2X_1 - X_2 - X_3) + \text{sign}(2X_2 - X_1 - X_3) + \text{sign}(2X_3 - X_1 - X_2))$. *For rectangular design and BIBD, m is fixed for given n and k, and is forced to have the scale of $m \asymp n^2$.

Stratified random sampling

Given a partition $S_0 = \cup_{j=1}^{J} S_j$, we approximate the U-statistics as follows

- Independently draw a random sample T_j from S_j (with replacement), $1 \leq j \leq J$, so that $\sum_{j=1}^{J} |T_j| = m$.
- Approximate the U-statistics by

$$U_{str} = \sum_{j=1}^{J} w_j U_j$$

where $w_j = |S_j|/|S_0|$ and $U_j = |T_j|^{-1} \sum_{\boldsymbol{i} \in T_j} g_{\boldsymbol{i}}$.

Theorem 1

Under the proportional sampling scheme, $|T_j| \propto |S_j|$, we have

$$\mathrm{Var}(U_{str}) \leq \mathrm{Var}(U_{RND}).$$

A simple but important observation

Consider the data

i	1	2	3	4	5	6
X_i	10	11	12	10	11	12

- For a kernel function g of order 3, obviously we have $g(X_1, X_2, X_3) = g(X_4, X_5, X_6)$.
- To our best knowledge, there has been no method incorporating this information.
- To utilize this information, we shall divide the original data into homogeneous groups.
- In multi-dimensional case it is called clustering.

OA based stratification: Definition

- For simplicity, suppose there is a positive integer, say L, which divides n.
- Arrange the data in ascending order $X_{(1)}, \ldots, X_{(n)}$ and divide them into L groups, $G_1, \ldots, G_L$ each of equal size.
- Let $A = (a_{jk})$ be an $OA(J, k, L, t)$, and a_{jk} be the element in the jth run and kth factor of A.
- For $1 \leq j \leq J$, draw a random sample of size m/J from $G_{a_{i1}} \times G_{a_{j2}} \times \cdots \times G_{a_{jk}}$, and calculate U_j as the average of g evaluated across the drawn sample.
- Approximate the U-statistics by

$$U_{oa} = \frac{1}{J} \sum_{j=1}^{J} U_j$$

- When $t = k$, we have $J = L^k$.

Corollary 1

For any distribution of X and kernel function g, we have

$$\mathrm{Var}(U_{oa}) \leq \mathrm{Var}(U_{RND})$$

$$\begin{aligned}\mathrm{Var}(U_{RND}) &= \mathrm{Var}(U_0) + m^{-1}(\sigma_g^2 - V(U_0)).\\ \sigma_g^2 &= \sum_{c=1}^{k} \binom{k}{c} \delta_c^2\end{aligned}$$

Theorem 2

Suppose g is Lipschitz continuous and X is bounded, with t being the strength of OA, we have

$$\mathrm{Var}(U_{oa}) = \mathrm{Var}(U_0) + m^{-1} \sum_{k \geq c > t} \binom{k}{c} \delta_c^2 + O(m^{-1}L^{-2}),$$

- For given m, with the constraint of $L^t = J \asymp m$, we have the trade-off in selecting the values of L and t.
- The optimal choice of the strength t and hence L depend on the comparison between $\delta_c, c > t$, and the change occurred to L^{-2} when we change t.

Corollary 2

Suppose g is Lipschitz continuous and X is bounded, the U-statistic based on the OA of strength k has the following property

$$\mathrm{Var}(U_{oa}) = \mathrm{Var}(U_0) + O(m^{-1}L^{-2}),$$

- Given strength k, under the constraint of $L^k = J \leq m$, the optimal choice for L will be $m^{1/k}$.

OA-stratification vs random sampling

- $\mathrm{Eff}(U) = \mathrm{Var}(U_0)/\mathrm{Var}(U)$.
- Recall both rectangular design and BIBD enforces $m \asymp n^2$. We have argued that the random sampling performs equivalently well in this case. We shall use random sampling as the benchmark to evaluate the OA method.
- When $m \succ n$, both OA and random sampling methods are asymptotically efficient.
- With $t = k$, $\mathrm{Eff}(U_{oa})$ converges to 1 faster than $\mathrm{Eff}(U_{RND})$

$$\frac{1 - \mathrm{Eff}(U_{oa})}{1 - \mathrm{Eff}(U_{RND})} = O(L^{-2}) \to 0, \text{ as } m, L \to \infty$$

- When $m \asymp n$ or $m \prec n$, U_{RND} is no longer efficient, but U_{oa} could be efficient under some circumstances.

Simulation: Test of symmetry

- For testing the symmetry of the distribution of X, we use the U-statistics with the following kernel function of order 3

$$
\begin{aligned}
g(X_1, X_2, X_3) &= \operatorname{sign}(2X_1 - X_2 - X_3) + \operatorname{sign}(2X_2 - X_1 - X_3) \\
&\quad + \operatorname{sign}(2X_3 - X_1 - X_2)
\end{aligned}
$$

- The data is *iid* generated from $\sim N(0, 1)$.
- We will compare three different incomplete U-statistics: random sampling and OA with strengths of $t = 2$ and $t = 3$.
- The simulations will be carried out for three different cases: $m \succ n$, $m \asymp n$ and $m \prec n$.

Large m

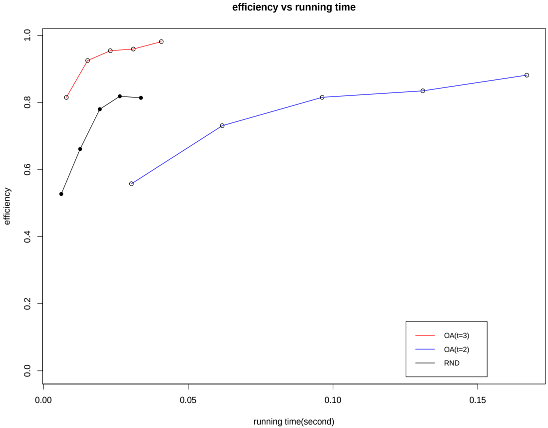

Figure 1: $n = 1024$, $m/4096 = 1, 2, 3, 4, 5$. The number of levels for OA are $L_2 = 64$ and $L_3 = 16$.

Moderate m

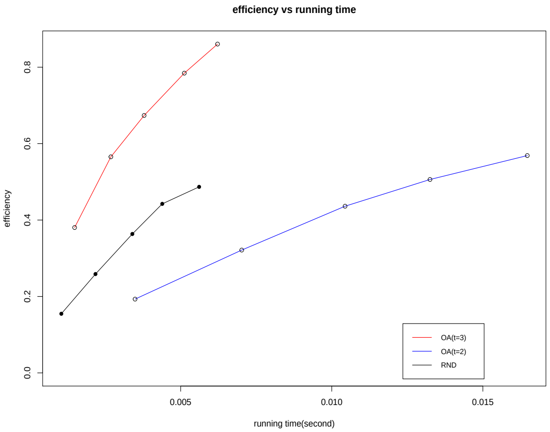

Figure 2: $n = 1215$, $m/729 = 1, 2, 3, 4, 5$. The number of levels for OA are $L_2 = 27$ and $L_3 = 9$.

Small m

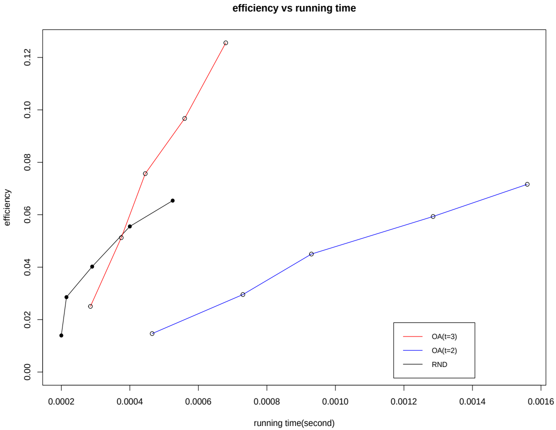

Figure 3: $n = 1000$, $m/64 = 1, 2, 3, 4, 5$. The number of levels for OA are $L_2 = 8$ and $L_3 = 4$. Ranking process takes substantial time for OA methods.

Choice of OA strength

- Even though the kernel function is non-degenerate, OA(t=2) is still dominated by OA(t=3). Why?
- Because the variance of the high order projection term does not decay fast: $\delta_1^2 = 0.0028$, $\delta_2^2 = 0.00724$, $\delta_3^2 = 0.081$.
- From previous results, we have

$$\begin{aligned} V(U_{oa(t=3)}) &= n^{-1}0.00252 + O(m^{-1}L_3^{-2}) \\ V(U_{oa(t=2)}) &= n^{-1}0.00252 + m^{-1}0.081 + O(m^{-1}L_2^{-2}) \\ V(U_{RND}) &= n^{-1}0.00252 + m^{-1}0.1111 \end{aligned}$$

  where $L_2^2 = L_3^3 = J$.
- OA(t=2) allows us to divide the original data into finer grids, but it loses the uniformity in 3-dimensional space.

When OA(t=2) is better

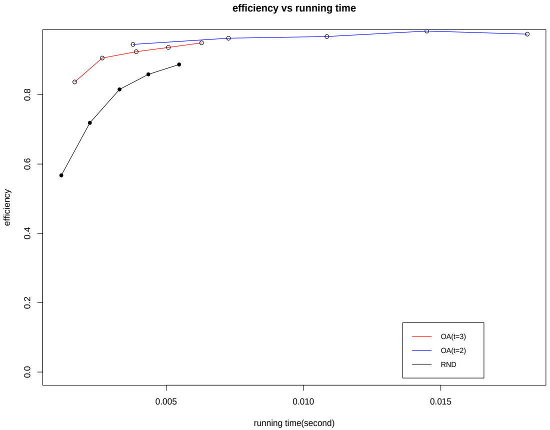

Figure 4: $n = 1215$, $m/729 = 1, 2, 3, 4, 5$;
$g(X_1, X_2, X_3) = \frac{1}{6}((X_1 - X_2)^2 + (X_2 - X_3)^2 + (X_3 - X_1)^2)$;
$\delta_1^2 = 2/9$, $\delta_2^2 = 2/9$, $\delta_3^2 = 0$.

Divide and conquer does not work well

- Proposed by Lin and Xi (2010).
- Randomly split the data into K parts and calculate U-statistics on each part and take the aggregate average.
- The computational complexity is $O(K(n/K)^k)$.

Method	efficiency (%)	m	time
OA	98.81	20480	0.044 sec
Divide and conquer	76.09	35840	0.22 sec
	92.80	158720	1.26 sec
Random sampling	$\simeq 100$	166167	0.35 sec

Table 4: $n = 10^3$, $k = 3$, $g(X_1, X_2, X_3) = \frac{1}{3}(\text{sign}(2X_1 - X_2 - X_3)$ $+\text{sign}(2X_2 - X_1 - X_3)$ $+\text{sign}(2X_3 - X_1 - X_2))$.

Simulation: Wilcoxon Signed Rank Test

- The Wilcoxon Signed Rank Test is a nonparametric method to test the equality of the means of two matched up samples.
- The testing statistic can be represented as a summation of two simple U-statistics

$$W_n^+ = \sum_{1\le i\le n} I(Z_i > 0) + \sum_{1\le i<j\le n} I(Z_i + Z_j > 0)$$

- We will compare two different incomplete U-statistics: random sampling and OA with $k = t = 2$.
- The simulations will be carried out for three different cases: $m \succ n$, $m \asymp n$ and $m \prec n$.

Large m

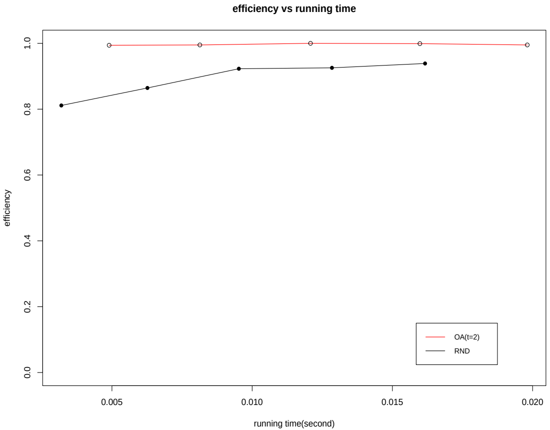

Figure 5: $n = 1000$, $m/2500 = 1, 2, 3, 4, 5$.

Moderate m

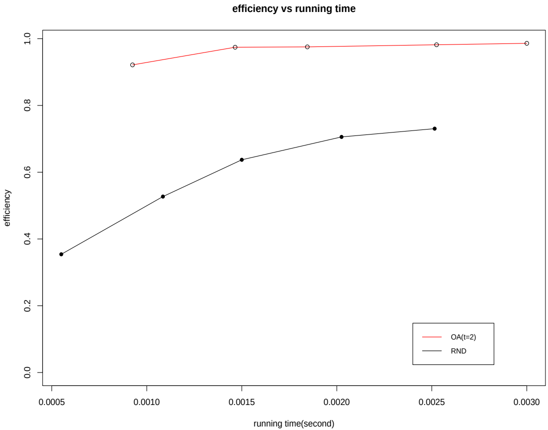

Figure 6: $n = 1000,\ m/400 = 1, 2, 3, 4.$

Small m

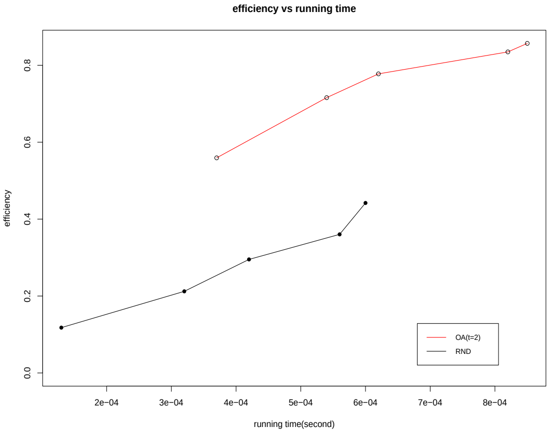

Figure 7: $n = 1000,\ m/100 = 1, 2, 3, 4, 5.$

- Notice in Figure 6, the OA based U-statistic is still highly efficient even when $m \leq n/2$.
- Recall $V(U_{oa}) = V(U_0) + O(m^{-1}L^{-2})$.
- Under the constraint of $L^2 \leq m$, we could choose $L \asymp \sqrt{m}$ so that we have $V(U_{oa}) = V(U_0) + O(m^{-2})$.
- This means the OA based U-statistic shall be asymptotically efficient as long as $m \succ \sqrt{n}$.

Degenerated case

- Recall the variances for U_{RND} and U_{oa}.

$$V(U_{RND}) = \text{Var}(U_0) + m^{-1}(\sigma_g^2 - V(U_0))$$

$$V(U_{oa}) = V(U_0) + O(m^{-1}L^{-2})$$

- Suppose g is degenerate of order d, we have $V(U_0) = O(n^{d+1})$.
- To be asymptotically efficient: RND requires $m \succ n^{d+1}$.
- For OA method, we choose $L \asymp m^{1/k}$, which result in $V(U_{oa}) = V(U_0) + O(m^{-(1+2/k)})$
- OA based U-stat will be asymptotically efficient if $m \succ n^{\frac{d+1}{1+2/k}}$.
- When $k = 2$, we have $m_{OA} \asymp \sqrt{m_{RND}}$
- Actually, for large enough m, we still have

$$\frac{1 - \text{Eff}(U_{oa})}{1 - \text{Eff}(U_{RND})} = O(L^{-2}) \to 0$$

$k = 3$ and $d = 2$

Method	efficiency (%)	m
OA	0.8	166000
Rectangular design	0.06	117306*
BIBD	0.1	166167*
Random sampling	0.0026	166167

Table 5: $n = 1000$; $g(X_1, X_2, X_3) = X_1X_2X_3$; $X_i \sim N(0,1)$; $\delta_1^2 = \delta_2^2 = 0$ and $\delta_3^2 = 1$. *For rectangular design and BIBD, m is fixed for given n and k, and is forced to have the scale of $m \asymp n^2$.

- We have claimed BIBD to be optimal. Why is it inferior to the new method (OA) now?
- OA stratification is playing a different game.

$k = 2$ and $d = 1$

Method	efficiency (%)	m
OA	82.17	250000
	26.84	40000
	4.167	10000
random sampling	0.506	250000
	0.076	40000
	0.021	10000

Table 6: $n = 10^4$; $g(X_1, X_2) = X_1 X_2$; $X_i \sim N(0, 1)$; $\delta_1^2 = 0$ and $\delta_2^2 = 1$.

Discussions

- It is computational rewarding to replace the original U-statistic by incomplete U-statistic.
- The random sampling method perform so well ($m \succ n$ for 100% efficiency) that there was not much gain by using existing design methods or the so called "divide and conquer".
- We proposed a simple idea of grouping which made significant improvement against random sampling. e.g. $m_{OA} = \sqrt{m_{RND}}$ for $k = 2$.
- Some possible extensions.
  - Multi-dimensional input
  - Multi-dimensional output
  - Multi-sample case
  - Hodges-Lehmann estimator
  - Machine learning applications

Key references

- Blom (1976). Some Properties of Incomplete U-Statistics. Biometrika.
- Brown and Kildea (1978). Reduced U-Statistics and the Hodges-Lehmann Estimator. The Annals of Statistics.
- Lee (1982). On incomplete U-statistics having minimum variance. Austral. J. Statist.
- Rempala and Srivastav (2004). Minimum variance rectangular designs for U-statistics. Journal of Statistical Planning and Inference.

Welcome to U.S. series of DAE 2019

- The purpose of the Design and Analysis of Experiments (DAE) conference series is to provide support and encouragement to junior researchers in the field of design and analysis of experiments, and to stimulate interest in topics of practical relevance to science and industry.
- The meetings also attract top notch senior researchers, primarily from North America and **Europe**, and emphasize interaction between junior and senior researchers.
- It has been held at Columbus, OH (2000), Vancouver, BC (2002), Chicago, IL (2003), Santa Fe, NM (2005), Memphis, TN (2007), Columbia, MO (2009), Athens, GA (2012), Cary, NC (2015), and Los Angels, CA (2017). DAE 2019 is the tenth event to be held in Knoxville, TN.

University of Tennessee

Smoky Mountain