U-Statistic Reduction: Higher-Order Accurate Risk Control and Statistical-Computational Trade-Off

Meijia Shao
Meta Platforms
and
Dong Xia
Department of Mathematics
The Hong Kong University of Science and Technology
and
Yuan Zhang
Department of Statistics
The Ohio State University

Abstract

U-statistics play central roles in many statistical learning tools but face the haunting issue of scalability. Despite extensive research on accelerating computation by U-statistic reduction, existing results almost exclusively focused on power analysis. Little work addresses risk control accuracy, which requires distinct and much more challenging techniques. In this paper, we establish the first statistical inference procedure with provably higher-order accurate risk control for incomplete U-statistics. The sharpness of our new result enables us to reveal how risk control accuracy also trades off with speed, for the first time in literature, which complements the well-known variance-speed trade-off. Our general framework converts the challenging and case-by-case analysis for many different designs into a surprisingly principled and routine computation. We conducted comprehensive numerical studies and observed results that validate our theory's sharpness. Our method also demonstrates effectiveness on real-world data applications.

Keywords: Nonparametrics, statistical learning, Edgeworth expansion, fast computation.

1 Introduction

A *U-statistic*, denoted by U_n, is associated with an i.i.d. sample $X_1, \ldots, X_n$ drawn from a general probability space and a degree-r permutation-invariant kernel function $h(x_1, \ldots, x_r)$, s.t. $h(x_1, \ldots, x_r) = h(x_{\pi(1)}, \ldots, x_{\pi(r)})$ for any bijection $\pi : [1:r] \leftrightarrow [1:r]$. It is defined as

$$U_n := \binom{n}{r}^{-1} \sum_{1 \leqslant i_1 < \ldots < i_r \leqslant n} h(X_{i_1}, \ldots, X_{i_r}) =: \binom{n}{r}^{-1} \sum_{I_r \in \mathcal{C}_n^r} h(X_{I_r}), \tag{1}$$

where $\mathcal{C}_n^k := \{(i_1, \ldots, i_k) : 1 \leqslant i_1 < \ldots < i_k \leqslant n\}$ is the collection of all r-tuples and define the shorthand $X_{I_k} := (X_{i_1}, \ldots, X_{i_k})$ for any $k \in [1:r]$. U-statistics play central roles in many contemporary statistical learning methods, such as in the following applications:

Example 1.1 (Example 1 of [24]). *Test the symmetry of the distribution of $X \in \mathbb{R}$ by $h(x_1, x_2, x_3) := \operatorname{sign}(2x_1 - x_2 - x_3) + \operatorname{sign}(2x_2 - x_3 - x_1) + \operatorname{sign}(2x_3 - x_1 - x_2)$.*

Example 1.2 (Bergsma-Dassios sign covariance [2, 31]). *To test the independence of $X \in \mathbb{S}_X$ and $Y \in \mathbb{S}_Y$, where $\mathbb{S}_X$ and $\mathbb{S}_Y$ are Banach spaces equipped with metrics ρ_X and ρ_Y, respectively, define $h\big((x_1, y_1), \ldots (x_4, y_4)\big) := s_X(x_{i_1}, \ldots, x_{i_4}) s_Y(y_{i_1}, \ldots, y_{i_4})$, where $s_X(t_1, \ldots, t_4) := \operatorname{sign}\big\{\rho_X(t_1, t_2) + \rho_X(t_3, t_4) - \rho_X(t_1, t_3) - \rho_X(t_2, t_4)\big\}$, and define s_Y similarly.*

Example 1.3 (Treatment effect measurement [34, 44]). *Let $Y_1, \ldots, Y_n$ denote the observed treated-minus-control matched pair differences. Given integers $\underline{r}, r$ and $\bar{r}$ satisfying $1 \leqslant \underline{r} \leqslant \bar{r} \leqslant r$, consider any r observations $Y_{I_r} := (Y_{i_1}, \ldots, Y_{i_r})$. Define $h(Y_{I_r}) := \sum_{\ell=\underline{r}}^{\bar{r}} \mathbb{1}_{[Y_{I_{r,(\ell)}} > 0]}$, where $I_{r,(\ell)}$ denotes the index of the ℓ-th largest $|Y_{i_k}|$ for $k = 1, \ldots, r$.*

One primary challenge in the practical use of U-statistics is the high computational cost. Even just evaluating U_n costs $O(n^r)$ time, where r varies across applications, ranging from $r = 2$ for Maximum Mean Discrepancy (MMD) [16, 35] and energy distance [40], to $r = 4$ for dCov [40, 42] and SignCov (Example 1.2), and even up to around 20 in Example 1.3

(see Tables 3 and 4 in [44]). To mitigate this burden, researchers have developed two main approaches. The first explores shortcuts to fast-compute U_n: [34, 22, 6, 14] showed that some U-statistics can be computed in $O(n \log n)$ time. However, these shortcuts only work for scalar inputs[1], limiting their applicability to complex input data types. For instance, the Bergsma-Dassios sign covariance (Example 1.2) with manifold-valued functional trajectories as inputs [31] cannot benefit from the acceleration tricks in [20, 14]. Moreover, for non-scalar X-inputs, even evaluating a single $h(X_{I_r})$ term can sometimes be expensive. In our data analysis in Section 5.2, we consider earthquake and starlight change curves $X_i(t)$ for $t \in [0, T]$, see Figure 1. We aim to assess their within- and between- cluster dissimilarity by mean pairwise distance for different earthquake scales and star types, using a distance $h(X_i(\cdot), X_j(\cdot))$ between curves, eliminating nuisance phase discrepancy. A mature technique for aligning curves by matching their key landscape features is to compute a "warping function" [38, 39]. However, evaluating a single $h(X_i(\cdot), X_j(\cdot))$ using this method can take a few seconds on a high-performance computing (HPC) server.

This naturally motivates the second acceleration strategy: *U-statistic reduction*, that is, to average over a much smaller set of r-tuples. Let

$$\mathcal{J}_{n,\alpha} := \left(I_r^{(1)}, \ldots, I_r^{(|\mathcal{J}_{n,\alpha}|)}\right) \tag{2}$$

be a collection of elements in $\mathcal{C}_n^r$ with $|\mathcal{J}_{n,\alpha}| \asymp n^\alpha$ for some $\alpha \in (1, r)$ – we shall treat $\mathcal{J}_{n,\alpha}$ almost like a subset of $\mathcal{C}_n^r$, except that $\mathcal{J}_{n,\alpha}$ allows duplication. The *reduced U-statistic* (also known as an *incomplete U-statistic* [3, 8]) with design $\mathcal{J}_{n,\alpha}$ is defined as

$$U_J := |\mathcal{J}_{n,\alpha}|^{-1} \sum_{I_r \in \mathcal{J}_{n,\alpha}} h(X_{I_r}). \tag{3}$$

There are two kinds of prices we must pay for computation reduction. First, this reduction inflates $\mathrm{Var}(U_J)$, which further determines: (i) the confidence interval radius; and (ii) the

[1] [14] exploits the coordinate-wise order relations, but its trick cannot apply to non-vector inputs.

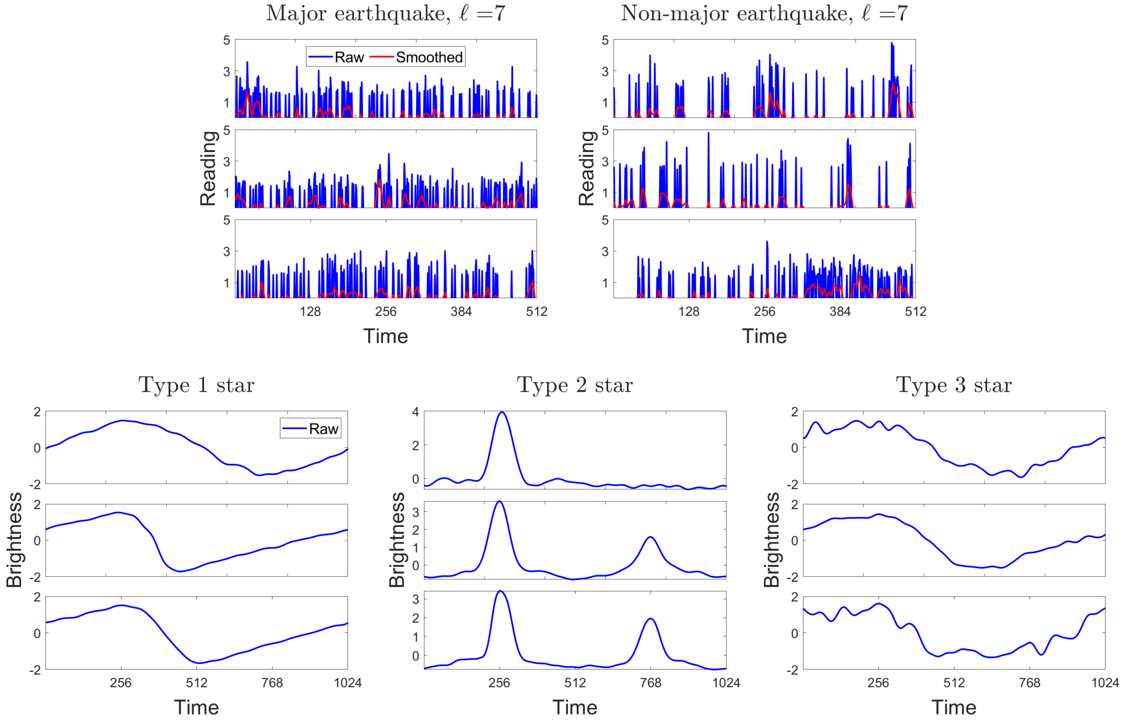

Figure 1: UCR data sets: row 1: *Earthquakes*, where blue curves show raw data and red curves show a moving average smoothing of window size $\ell = 7$; row 2: *Starlight*.

minimum separation condition $|\mu_{H_a} - \mu_{H_0}|$ for consistently[2] testing $H_0 : \mu = \mu_{H_0}$ versus $H_a : \mu = \mu_{H_a}$, where $\mu := \mathbb{E}[U_n]$. This aspect of computational-statistical trade-off is easy to quantify thus well-understood. The overwhelming majority of existing literature on U-statistic reduction regards this aspect, pioneered by [3] and followed up by many works aiming at designing $\mathcal{J}_{n,\alpha}$ smartly to minimize $\text{Var}(U_J)$ under a given computational budget $O(n^\alpha)$ [26, 27, 28, 33, 10, 24, 13].

The second kind of price for speeding-up, namely, the deterioration of *risk control accuracy* in statistical inference, is much more elusive and difficult to characterize. Here, by "risk control accuracy", we refer to: (i) $|\mathbb{P}(\text{true } \mu \in \text{CI}) - (1 - \beta)|$ for confidence intervals; and (ii) $|\mathbb{P}(\text{actual type I error rate}) - \beta|$ for hypothesis testing, where $1 - \beta$ and

[2] Test consistency: a test is called consistent if its type-I and type-II errors both converge to 0.

β are the nominal confidence and significance levels, respectively. Characterizing this accuracy requires a *higher-order accurate* approximation to the sampling distribution of the *studentized reduced U-statistic*, which most existing works fail to describe by only providing asymptotic results [4, 23, 7, 8]. Our paper is the first to uncover the computational-statistical trade-off in risk control accuracy, filling in a critical gap in the literature.

This paper makes several significant contributions. We present the first comprehensive study on risk control accuracy in statistical inference for reduced U-statistics. We establish the first higher-order accurate distribution approximation for non-degenerate reduced U-statistics under general designs, leading to Cornish-Fisher confidence intervals and tests both with higher-order accurate risk controls. Our approach requires only two natural, weak, and easy-to-verify assumptions that are satisfied by many popular designs. Notably, our method strictly complies with the $O(n^\alpha)$ computational budget in all parts and allows for easy parallel computing.

Our method's accuracy significantly improves over the best existing results. The sharpness of our error bounds enables us to reveal, for the first time, the trade-off between computation complexity (speed) and risk control accuracy of reduced U-statistics. Interestingly, we discovered that higher-order risk control accuracy can be achieved for any $\alpha > 1$; meanwhile, it may be surprising that we also find that the computation reduction from $O(n^r)$ to $O(n^2)$ is *nearly free lunch*, without deteriorating risk control error rate and only inflating $\mathrm{Var}(U_J)$ imperceptibly. For practitioners, our method provides fast and easy-to-implement solutions with tuning guidance, as well as advice on the minimum sample size requirement to achieve a target risk control accuracy goal.

The theoretical analysis in this paper differs significantly from the complete U-statistic literature [21, 30, 32] and features several innovations. Incompleteness introduces new and complicated leading terms and breaches the symmetry of remainder terms, rendering existing bounds and analysis routines in [21, 30] inapplicable. We tackle these challenges with our original analyses. A key methodological contribution of this paper is the development of a succinct and weak condition on the reduction design, formalized as Assumption 2, which was distilled from our theoretical explorations. In the proof of Lemma 3.2, a crucial supporting result for Corollary 3.2, we address the intricate dependency structures that arise in certain random sampling schemes.

Our paper goes beyond any single application or specific data structure, focusing instead on the fundamental question of risk control accuracy in U-statistic reduction. The general and comprehensive theoretical and methodological framework we present fills in a critical gap in the literature, providing a much-needed toolkit for many U-statistic-based learning methods that aim to maintain accurate risk control while scaling up.

1.1 Notation

We write $B_n = \widetilde{O}_p(b_n)$ if $\mathbb{P}(B_n \geqslant C \cdot b_n) = O(n^{-1})$ for some constant $C > 0$ and large enough n. Let $\Phi(\cdot)$ and $\phi(\cdot)$ be the CDF and PDF of $N(0,1)$, respectively. For simplicity, we regard n^α and $n^{\alpha-1}$ as integers throughout, omitting duly floor/ceiling operations. We adopt the Matlab style notation for arithmetic sequence: $[a_1 : a_2]$ denotes $(a_1, a_1+1, a_1+2, \ldots, a_2)$, whereas $[a_1 : \delta : a_2]$ denotes $(a_1, a_1+\delta, a_1+2\delta, \ldots, a_2)$.

2 Reduction of non-degenerate noiseless U-statistics

Recall that the reduced U-statistic U_J is the average of individual $h(X_{I_r}) := h(X_{i_1}, \ldots, X_{i_r})$ terms, where $I_r := \{i_1, \ldots, i_r\}$ ranges over a small subset $\mathcal{J}_{n,\alpha}$ inside $\mathcal{C}_n^r :=$ {the collection of all r-tuples in $[1 : n] := \{1, \ldots, n\}$}. Our goal is to perform accurate statistical inference for $\mu = \mathbb{E}[U_J]$ based on U_J, within a limited computational budget of $O(n^\alpha)$ for a given constant $\alpha : \alpha < r$. Following the convention [28], we call $\mathcal{J}_{n,\alpha}$ the *design* of U_J. Throughout this section, $\mathcal{J}_{n,\alpha}$ is fixed. To ease narration, we set up two sets of symbols.

- Define the *projection term* g_k's recursively: first set $g_1(X_1) := \mathbb{E}[h(X_{[1:r]})|X_1] - \mu$; then for each $k = 2, \ldots, r$ in order, define $g_k(X_{[1:k]}) := \mathbb{E}[h(X_{[1:r]})|X_{[1:k]}] - \mu - \sum_{k'=1}^{k-1} \sum_{I_{k'} \in \mathcal{C}_{[1:k]}^{k'}} g_{k'}(X_{I_{k'}})$. All g_k terms are mean-zero and mutually uncorrelated [30].

- For any size-k subset I_k of $[1 : n]$, let $a_{n,\alpha;k}(I_k) := \left|\{\widetilde{I}_r \in \mathcal{J}_{n,\alpha} : I_k \subseteq \widetilde{I}_r\}\right|$ count how many times I_k shows up in the design $\mathcal{J}_{n,\alpha}$. For example, if $r = 3$, $n = 7$ and $\mathcal{J}_{n,\alpha} = \{(1, 2, 4), (2, 5, 7), (3, 4, 6)\}$, then $a_{n,\alpha;1}(2) = 2$ and $a_{n,\alpha;2}(\{3, 4\}) = 1$.

Example 2.1. *To understand the random variation in U_J, suppose $r = 3$ and inspect just one term $h(X_{i_1}, X_{i_2}, X_{i_3})$. For example, suppose $(i_1, i_2, i_3) = (1, 2, 4)$, we have*

$$h(X_1, X_2, X_4) = \mu + g_1(X_1) + g_1(X_2) + g_1(X_4) + g_2(X_1, X_2) + g_2(X_1, X_4) + g_2(X_2, X_4) + g_3(X_1, X_2, X_4). \tag{4}$$

We call the form like (4) the "one-term Hoeffding's decomposition" *of $h(X_{I_r})$. Consequently, $h(X_1, X_2, X_4)$ contributes a count of 1 to each $a_{n,\alpha;k}(I_k)$, where $\varnothing \neq I_k \subseteq \{1, 2, 4\}$.*

In general, decomposing each $h(X_{I_r})$ in U_J as in Example 2.1, by [19], we have

$$U_J = |\mathcal{J}_{n,\alpha}|^{-1} \sum_{k=1}^{r} \sum_{I_k \in \mathcal{C}_n^k} a_{n,\alpha;k}(I_k) g_k(X_{I_k}). \tag{5}$$

Next, we address two fundamental questions regarding (5) in Sections 2.1 and 2.2.

2.1 What makes a good/bad design?

There are two main considerations that define a good design. They will both translate into our regularity assumptions and be reflected in our proposed methods.

(i) This design should comply with computation budget and be easy to implement.

(ii) Under the premise of (i), this design minimizes $\mathrm{Var}(U_J)$.

Our first regularity assumption reflects consideration (i).

Assumption 1. *The design of* $\mathcal{J}_{n,\alpha}$ *is* data-oblivious*, namely,*

$$\mathcal{J}_{n,\alpha} \perp (X_1, \ldots, X_n). \tag{6}$$

For a deterministic $\mathcal{J}_{n,\alpha}$*,* (6) *means that* $\mathcal{J}_{n,\alpha}$ *is designed without consulting the data* $X_{[1:n]}$*.*

The motivation behind Assumption 1 is two-fold, both weighing on consideration (i). First, although as pointed out by [24] that *data-aware* designs may have superior variance reduction, the step of adapting the design $\mathcal{J}_{n,\alpha}$ to the data $X_{[1:n]}$ may require expensive computation that can exceed the $O(n^\alpha)$ budget. The second motivation regards implementation feasibility. It is inspired by the study of network moments as "noisy U-statistics"[3], where X_i's are not only unobserved, but inestimable due to identifiability issues [15, 43].

Consideration (ii) has long been the focus in existing literature (but not always with much attention to consideration (i)). Clearly, the dummy construction of $\mathcal{J}_{n,\alpha}$ by repeating $[1:r]$ for $|\mathcal{J}_{n,\alpha}|$ times is useless. What makes $\mathrm{Var}(U_J)$ small then? By (5), we have

$$\mathrm{Var}(U_J) = |\mathcal{J}_{n,\alpha}|^{-2} \sum_{k=1}^{r} \Big\{ \sum_{I_k \in \mathcal{C}_n^k} a_{n,\alpha;k}^2(I_k) \Big\} \xi_k^2, \tag{7}$$

while for each $k \in [1:r]$, it always holds that $\sum_{I_k \in \mathcal{C}_n^k} a_{n,\alpha;k}(I_k) \equiv \binom{r}{k} |\mathcal{J}_{n,\alpha}|$. Therefore, minimizing $\mathrm{Var}(U_J)$ demands that for each k, all $a_{n,\alpha;k}(I_k)$'s are as similar as possible – this lets U_J maximally explore different index combinations. For instance, if $\alpha \in (1,2)$, this is requiring $a_{n,\alpha;1}(i) \equiv (r/n) \cdot |\mathcal{J}_{n,\alpha}|$ and $a_{n,\alpha;k}(I_k) \in \{0,1\}$ for all $k \in [2:r]$. In other words,

[3] Despite this paper exclusively studies conventional, noiseless U-statistics, in a closely related work, we will make use of the analysis techniques in this paper to analyze network U-statistics.

we need different $h(X_{I_r})$ terms to contribute *unique* $g_2, \ldots, g_r$ terms (while g_1 terms will unavoidably repeat as $\alpha > 1$) – we call this the "*non-overlapping property*" of the design. In alignment with these discussions, our second assumption aims at avoiding bad designs.

Assumption 2. *Set* $\alpha \in (1, r)\backslash\mathbb{Z}$. *It holds for all* $k \in [1 : r]$ *and* $I_k \in \mathcal{C}_n^k$ *that*

$$a_{n,\alpha;k}(I_k) \in \begin{cases} [C_1, C_2]n^{\alpha-k}, & \text{if } k < \alpha, \\ [0, C_2], & \text{if } k > \alpha, \end{cases} \tag{8}$$

where $C_1, C_2 : 0 < C_1 < C_2$ *are universal constants.*

In Assumption 2, we exclude integer α choices for sophisticated technical reasons – but in plain language, this would make theoretical analysis much cleaner. Practitioners who set a working $\alpha = 2$ can use our formulas for $\alpha = 2.001$, without causing noticeable error.

Last but important, the two considerations (i) and (ii) intertwine: to our best knowledge, principled and fast construction of a variance-minimizing design remains an open challenge before this paper. The variance-minimal methods in existing literature typically depend on brilliant, but case-by-case, constructions for special $(n, r, |\mathcal{J}_{n,\alpha}|)$ configurations. They provide little clue for handling general $(n, r, |\mathcal{J}_{n,\alpha}|)$ settings. In Section 3.1, we will solve this standing problem with an innovative design method.

2.2 How to develop a higher-order accurate statistical inference?

2.2.1 Non-degeneracy, variance estimation and studentization

With Assumptions 1 and 2, we can consider the design as "reasonably good" that provides a solid basis for downstream analysis. In this section, we will develop higher-order accurate statistical inference method for *any given design* that satisfies both assumptions. In other words, through this section, we fix $\mathcal{J}_{n,\alpha}$. Like in the study of complete U-statistics, we will first formulate a variance estimator and use it to studentize U_J, then formulate an accurate distribution approximation to the studentization. All these steps critically depend on the *degeneracy status* of the U-statistic.

Definition 2.1. *We call* U_J "non-degenerate", *if* $\xi_1^2 := \mathrm{Var}(g_1(X_1)) \geqslant \text{constant} > 0$.

Due to page limit, we leave the degenerate case, i.e., $\xi_1 = 0$ to future work. Next up, we face two routes for variance estimation: we could target at either the full variance $\sigma_J^2 := \mathrm{Var}(U_J)$ or just the dominating term $\sigma_{J;1}^2 := |\mathcal{J}_{n,\alpha}|^{-2}\sum_{i=1}^n a_{n,\alpha;1}^2(i)\xi_1^2$. This was not a question for complete U-statistics, where σ_J^2 and $\sigma_{J;1}^2$ differ only by $O(n^{-2})$ [30, 43]; but for an incomplete U_J, we have $|\sigma_J^2 - \sigma_{J;1}^2| \asymp n^{-\alpha}$, which cannot be directly ignored. We choose to estimate $\sigma_{J;1}^2$, because it leads to cleaner formulation and faster computation. The discrepancy between $\sigma_{J;1}^2$ and $\mathrm{Var}(U_J)$ will be accounted for by our Edgeworth correction terms, see Remark 2.4 for more details.

To estimate $\sigma_{J;1}^2$, we need to estimate $\xi_1^2 := \mathrm{Var}(g_1(X_1))$ (since we know $a_{n,\alpha;1}(i)$'s). Classical variance estimators, such as jackknife ([30], Section 2) and [43], do not comply with the $O(n^\alpha)$ computation budget limit. Therefore, we propose the following estimator

$$\widetilde{\xi}_1^2 := n^{-\alpha}\sum_{i=1}^n\sum_{d=1}^{n^{\alpha-1}} h(X_{[i:d:(i+(r-1)d)]})h(X_{[i:(-d):(i-(r-1)d)]}) - \widetilde{\mu}^2, \tag{9}$$

where $\widetilde{\mu}^2 := n^{-\alpha}\sum_{i=1}^n\sum_{d=1}^{n^{\alpha-1}} h(X_{[i:d:(i+(r-1)d)]})h(X_{[(i+rd):d:(i+(2r-1)d)]})$. The formula (9) may seem intricate at first sight, but its idea is very simple. To illustrate, set $r = 3$ as in Example 2.1 and inspect the summands corresponding to $d = 1$ in the first term in (9):

$$n^{-1}\sum_{i=1}^n h(X_i, X_{i+d}, X_{i+2d})h(X_i, X_{i-d}, X_{i-2d}) = \mu^2 + \xi_1^2 + \mathfrak{R}, \tag{10}$$

where $\mathfrak{R}$ consists of several types of terms, such as $g_1^2(X_i) - \xi_i$, and $g_1(X_{i-d})\mu$, and $g_1(X_i)g_1(X_{i+2d})$, and so on, all averaged over i. Clearly, $\mathfrak{R}$ is mean-zero and concentrates. Similarly, we can understand why $\widetilde{\mu}^2$ is also an unbiased estimator for μ^2. We stress that our variance estimator (9) strictly complies with the $O(n^\alpha)$ computation budget constraint. With the variance estimator, we can studentize U_J as

$$T_J := \frac{U_J - \mu}{|\mathcal{J}_{n,\alpha}|^{-1}\big\{\sum_{i=1}^n a_{n,\alpha;1}^2(i)\big\}^{1/2} \cdot \widetilde{\xi}_1}. \tag{11}$$

2.3 Accurate distribution approximation to studentization

An accurate distribution approximation for T_J is the premise of accurate inference. For this goal, it is important to understand the stochastic variations in U_J. A natural method is to compare T_J to the *standardization* of U_J (replacing $\widetilde{\xi}_1$ in (11) by the true ξ_1) and then account for the plug-in error on the denominator. Define

$$M_\alpha := |\mathcal{J}_{n,\alpha}|^{-1}\sum_{i=1}^n a_{n,\alpha;1}^2(i) \asymp n^{\alpha-1}, \tag{12}$$

$$\mathcal{T}_1 := \frac{\sum_{i=1}^n a_{n,\alpha;1}(i) g_1(X_i)}{\big\{\sum_{i=1}^n a_{n,\alpha;1}^2(i)\big\}^{1/2}\xi_1}, \quad \mathcal{T}_2 := \frac{\sum_{k=2}^r \sum_{I_k \in \mathcal{C}_n^k} a_{n,\alpha;k}(I_k) g_k(X_{I_k})}{\big\{\sum_{i=1}^n a_{n,\alpha;1}^2(i)\big\}^{1/2}\xi_1},$$

$$\mathcal{T}_3 := \sum_{i=1}^n \frac{g_1^2(X_i) - \xi_1^2}{n\xi_1^2} + \frac{1}{nM_\alpha\xi_1^2}\sum_{i=1}^n\sum_{d=1}^{M_\alpha}\sum_{\ell=1}^{r-1} g_1(X_i)\Big\{g_2(X_i, X_{i+\ell d}) + g_2(X_i, X_{i-\ell d})\Big\}. \tag{13}$$

Let us explain these definitions for general audience. First, M_α accounts for a frequently-used non-random factor. Then $\mathcal{T}_1 + \mathcal{T}_2$ is the standardization of U_J: we separate $\mathcal{T}_1$ and $\mathcal{T}_2$ because $\mathcal{T}_1$ is a weighted i.i.d. sum and the dominating term, while $\mathcal{T}_2$ is a higher-order bias-correction to enhance risk control accuracy. Finally, $\mathcal{T}_3$ captures the plug-in error in using $\widetilde{\xi}_1$ in T_J. Formally, we have the following lemma.

Lemma 2.1. *Set $\alpha \in (1, 2)$, we have $\widetilde{\xi}_1^2 - \xi_1^2 = \xi_1^2 \cdot \mathcal{T}_3 + \widetilde{O}_p(n^{-\alpha/2}\log n)$.*

With the above notation preparation and supporting results, we can decompose T_J:

$$T_J = (\mathcal{T}_1 + \mathcal{T}_2)(1 + \mathcal{T}_3)^{-1/2} = \mathcal{T}_1 + \mathcal{T}_2 - \frac{1}{2}\mathcal{T}_1\mathcal{T}_3 + \widetilde{O}_p\big(n^{-\alpha/2}\log n\big). \tag{14}$$

So far, everything may seem familiar to readers who know the U-statistic literature. However, next we will see how U-statistic reduction leads to very different bias-correction terms in the Edgeworth expansion. Before that, we make a quick technical remark.

Remark 2.1. *Aside from Assumptions 1 and 2, another commonly required assumption in U-statistic literature is Cramér's condition [21, 30]:* $\limsup_{t\to\infty}\left|\mathbb{E}[e^{\mathbb{i}t\xi_1^{-1}\cdot g_1(X_1)}]\right| < 1$. *This condition is undesirably restrictive and violated by important applications, e.g., Example 1.1 with a discrete* X_1 *distribution. Inspired by [25] and [36], we add to* T_J *an* artificial *smoothing term* $\delta_J \sim N(0, \sigma_\delta^2 = C_{\delta_J}\log n \cdot n^{-\alpha})$ *independent of* T_J *with a large enough constant* $C_{\delta_J} > 0$. *We will show that* δ_J *waives Cramér's condition without altering the distribution approximation formula*[4].

Now we present our main results and accompanying remarks. Let $\xi_k^2 := \mathrm{Var}(g_k(X_{[1:k]}))$. Define the population Edgeworth expansion formula for T_J to be

$$G_{\mathcal{J}_{n,\alpha}}(u) := \Phi(u) + \phi(u)\left\{\frac{\Gamma_0(u)}{\sqrt{n}} + \sum_{\ell=1}^{\lfloor\frac{\alpha/2}{\alpha-1}\rfloor}\frac{\Gamma_\ell(u)}{M_\alpha{}^\ell}\right\}, \tag{15}$$

where we recall the definition of M_α from (12), and define shorthand Γ_0 and Γ_ℓ's, as follows.

$$\begin{aligned}\frac{\Gamma_0(u)}{\sqrt{n}} :=& \left(-\frac{\sum_{i=1}^n a_{n,\alpha;1}^3(i)(u^2-1)}{6\xi_1^3\{\sum_{i'=1}^n a_{n,\alpha;1}^2(i')\}^{3/2}} + \frac{r|\mathcal{J}_{n,\alpha}|u^2}{2\{\sum_{i'=1}^n a_{n,\alpha;1}^2(i')\}^{1/2}n\xi_1^3}\right)\mathbb{E}[g_1^3(X_1)]\\ &+\left(-\frac{\sum_{1\leqslant i<j\leqslant n} a_{n,\alpha;1}(i)a_{n,\alpha;1}(j)a_{n,\alpha;2}(\{i,j\})(u^2-1)}{\{\sum_{i'=1}^n a_{n,\alpha;1}^2(i')\}^{3/2}\xi_1^3} + \frac{r(r-1)|\mathcal{J}_{n,\alpha}|u^2}{\{\sum_{i'=1}^n a_{n,\alpha;1}^2(i')\}^{1/2}n\xi_1^3}\right)\\ &\mathbb{E}[g_1(X_1)g_1(X_2)g_2(X_1,X_2)],\end{aligned} \tag{16}$$

$$\frac{\Gamma_\ell(u)}{M_\alpha{}^\ell} := -\frac{H_{2\ell-1}(u)}{(2\ell)!\{\sum_{i'=1}^n a_{n,\alpha;1}^2(i')\}^\ell\xi_1^{2\ell}}\times\left\{\sum_{k=2}^r\sum_{I_k\in\mathcal{C}_n^k} a_{n,\alpha;k}^2(I_k)\xi_k\right\}^\ell, \tag{17}$$

[4] This means that the same Edgeworth expansion formula accurately approximates both $F_{T_J+\delta_J}$ without Cramér's condition and F_{T_J} assuming this condition.

where $H_k(u) := (-1)^k e^{u^2/2}\mathrm{d}^k/\mathrm{d}u^k(e^{-u^2/2})$ is the kth Hermite polynomial ([37], page 99). In (15), the first correction term (16) generalizes its familiar counterpart in literature. To see this, consider the special case of complete U-statistic, where $\mathcal{J}_{n,\alpha} = \mathcal{C}_n^r$, We have $|\mathcal{J}_{n,\alpha}| = \binom{n}{r}$ and $a_{n,\alpha;k}(I_k) \equiv \binom{r}{k}\binom{n}{r}/\binom{n}{k} = \binom{n-k}{r-k}$, thus (16) reproduces Eq. (1.6) in [21].

The second term (17), however, is unique to reduced U-statistics and was never seen in existing literature. To facilitate understanding, in Table 2, we sketch some important properties of the main terms in the decomposition (5). Here, while the first term in $\mathcal{T}_3$ is clearly $\asymp n^{-1/2}$, its second term is also $\asymp n^{-1/2}$ – to see this, simply notice that for each ℓ,

Term in T_J's decomp.	Asymp. order	Corresp. Edgeworth terms
$\mathcal{T}_1$	1	Φ and Γ_0
$\mathcal{T}_2$	$n^{-(\alpha-1)/2}$	Γ_0 and Γ_ℓ's, $\ell \geqslant 1$
$\mathcal{T}_1 \cdot \mathcal{T}_3$	$n^{-1/2}$	Γ_0

Table 1: Properties of main terms in T_J's decomposition (5)

$$\frac{1}{nM_\alpha}\sum_{i=1}^{n}\sum_{d=1}^{M_\alpha} g_1(X_i)g_2(X_i, X_{i+\ell d}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[g_1(X_{i-\ell d})g_2(X_{i+\ell d}, X_i)|X_i]$$
$$+ \frac{1}{nM_\alpha}\sum_{i=1}^{n}\sum_{d=1}^{M_\alpha}\big\{g_1(X_i)g_2(X_i, X_{i+\ell d}) - \mathbb{E}[g_1(X_i)g_2(X_i, X_{i+\ell d})|X_{i+\ell d}]\big\}, \qquad (18)$$

where the second term on the RHS of (18) is $\widetilde{O}_p(n^{-\alpha/2}\log n)$, thus ignorable[5]. From Table 2, we see that $\mathcal{T}_2$ leads to our newly-discovered Edgeworth expansion terms. It is crucial that we clarify that "$\mathcal{T}_2$ lying in the $n^{-(\alpha-1)/2}$ order" does not automatically guarantee that there will exist an $O(n^{-(\alpha-1)/2})$ term in the Edgeworth expansion. Roughly speaking, this all depends on which terms will lead in the Taylor expansion $\mathbb{E}\big[e^{\mathbb{i}t(\mathcal{T}_1+\mathcal{T}_2)}\big] = \mathbb{E}\big[e^{\mathbb{i}t\mathcal{T}_1}\big(1 + \mathbb{i}t\mathcal{T}_2 + \frac{(\mathbb{i}t)^2}{2}\mathcal{T}_2^2 + \cdots\big)\big]$, while others enter the remainder. See the proofs of Lemma S.1.3-(d) and Proposition S.1.1 in Supplementary Material for more details.

[5] Notice that although this term has similar numerator as $\mathcal{T}_2$, its denominator is much larger.

In practice, we use the empirical version of (15) with estimated coefficients. Define

$$\widetilde{\mathbb{E}}[g_1^3(X_1)] := \frac{1}{n}\sum_{i=1}^{n} h(X_{[i:(i+r-1)]})h(X_{\{i,[(i+r):(i+2r-2)]\}})h(X_{\{i,[(i+2r-1):(i+3r-3)]\}}) - \widetilde{\mu}^3 \tag{19}$$

$$\widetilde{\mathbb{E}}[g_1(X_1)g_1(X_2)g_2(X_1,X_2)] := \frac{1}{n}\sum_{i=1}^{n} h(X_{[(i-r+1):i]})h(X_{[i:(i+r-1)]})h(X_{[(i+r-1):(i+2r-2)]}) - \widetilde{\mu}^3 - 2U_J \cdot \widetilde{\xi}_1^2 \tag{20}$$

$$\widetilde{\xi}_k^2 := \frac{1}{n^{\alpha}}\sum_{i=1}^{n}\sum_{d=1}^{n^{\alpha-1}} h(X_{[i:d:(i+(r-1)d)]})h(X_{[(i+(k-1)d):(-d):(i-(r-k)d)]}) - \widetilde{\mu}^2 - \sum_{k'=1}^{k-1}\binom{k}{k'}\widetilde{\xi}_{k'}^2, \tag{21}$$

for $k \in [2:r]$. These estimators all share the same idea in our development of $\widetilde{\xi}_1^2$ in (9), thus can be understood similarly. Let $\widetilde{G}_{\mathcal{J}_{n,\alpha}}(u)$ be the empirical version of $G_{\mathcal{J}_{n,\alpha}}(u)$ with coefficients estimated by (9), (19), (20) and (21). We have

Theorem 2.1. *Set $\alpha \in (1,2)$. If U_J is non-degenerate and $\mathcal{J}_{n,\alpha}$ satisfies Assumptions 1 and 2, then we have*

$$\left\|F_{T_J+\delta_J|\mathcal{J}_{n,\alpha}}(u) - G_{\mathcal{J}_{n,\alpha}}(u)\right\|_{\infty} = O\left(n^{-\alpha/2}\log^{1/2} n\right), \tag{22}$$

$$\left\|F_{T_J+\delta_J|\mathcal{J}_{n,\alpha}}(u) - \widetilde{G}_{\mathcal{J}_{n,\alpha}}(u)\right\|_{\infty} = \widetilde{O}_p(n^{-\alpha/2}\log^{1/2} n). \tag{23}$$

Remark 2.2. *Theorem 2.1 highlights an important practical guidance that for non-degenerate U-statistics, setting $\alpha > 2$ will not further merit risk control accuracy, since the error bound at $\alpha = 2$ already matches that for a complete U-statistic [21, 30]. Also, increasing α beyond 2 only brings $O(n^{-2})$ improvement to $\mathrm{Var}(U_J)$ [28]. Considering the computational cost grows exponentially in α, it is therefore not worthwhile to set $\alpha > 2$ under non-degeneracy.*

Remark 2.3. *Remark 3.1 in [8] points out that as α decreases, $\sigma_{J;1}$ becomes a poorer approximation to σ_J; when $\alpha = 1$, $|\sigma_{J;1} - \sigma_J|$ no longer vanishes as $n \to \infty$, which [41, 8] refer to as a "phase change". While [41, 8] exclusively studied $\mathrm{Var}(U_J)$ as $\alpha \to 1$, our results reveal how risk control accuracy behaves in this regime, completing the missing piece* *in the big picture. We find that the Edgeworth expansion becomes lengthier, and the risk control accuracy also depreciates. If we do not incorporate an increasing number of bias-correction terms in the Edgeworth expansion, the risk control accuracy depreciates even faster: the $n^{-\alpha/2}$ term in Theorem 2.1 will be replaced by $n^{-(\alpha-1)}$, which is the Berry-Esseen bound of the normal approximation to T_J.*

2.3.1 Higher-order accurate statistical inference

To test the hypotheses

$$H_0 : \mu = \mu_0; \quad \text{vs.} \quad H_a : \mu \neq \mu_0,$$

we use the empirical p-value, denoted by $\mathfrak{p}$ and defined as follows

$$\mathfrak{p} := 2\min\Big\{\widetilde{G}_{\mathcal{J}_{n,\alpha}}(T_J^{(\text{obs})} + \delta_J), 1 - \widetilde{G}_{\mathcal{J}_{n,\alpha}}(T_J^{(\text{obs})} + \delta_J)\Big\}, \tag{24}$$

where $T_J^{(\text{obs})} := (U_J - \mu_0)/\{|\mathcal{J}_{n,\alpha}|^{-1}\{\sum_{i=1}^n a_{n,\alpha;1}^2(i)\}^{1/2}\widetilde{\xi}_1\}$.

Corollary 2.1. *Under the conditions of Theorem 2.1, the test (24) enjoys a higher-order accurate type-I error control:* $\mathbb{P}_{H_0}\big(\mathfrak{p} < \beta\big|\mathcal{J}_{n,\alpha}\big) = \beta + O(n^{-\alpha/2}\log^{1/2} n)$.

Next, we invert the Edgeworth expansion to formulate the Cornish-Fisher confidence interval (CF-CI) with higher-order accurate confidence level control. Before presenting our method, for readers who are not familiar with this topic, we give a quick review of how the CF-CI was derived in the classical setting. Constructing a CI requires quantiles of the distribution of the pivot, but the Edgeworth expansion G is not guaranteed to be a valid CDF, as its value may exceed the range $[0, 1]$, thus cannot be naively inverted. The Edgeworth expansions for an i.i.d. sample mean and a complete U-statistic both take the form $G(u) = \Phi(u) + n^{-1/2}\phi(u)\Gamma_0(u)$, at $O(n^{-1})$ accuracy. Given the significance level $\beta \in (0, 1/2)$, we need to find a u that well approximates the lower-β quantile of the distribution approximated by G, that is, the u such that $G(u) = \beta + O(n^{-1})$. This can be achieved by the *Cornish-Fisher expansion* [17, 18], which takes the form $u = G^{-1}(z_\beta) := z_\beta - n^{-1/2}\Psi_0(z_\beta)$, where $z_\beta := \Phi^{-1}(\beta)$. To determine $\Psi_0(z_\beta)$, we expand $G(z_\beta - n^{-1/2}\Psi_0(z_\beta))$ and set all $n^{-1/2}$ terms to sum to zero. This gives $\Psi_0(u) = \Gamma_0(z_\beta)$. Therefore, $G^{-1}(z_\beta) = \phi(z_\beta) - n^{-1/2}\Gamma_0(z_\beta)$.

In contrast, the Cornish-Fisher expansion in our setting is much complicated by the $\Gamma_\ell \asymp n^{-(\alpha-1)\ell}$ terms in the Edgeworth expansion. Our C-F expansion reads:

$$G^{-1}_{\mathcal{J}_{n,\alpha}}(z_\beta) =: z_\beta - \frac{\Gamma_0(z_\beta)}{\sqrt{n}} + \sum_{\ell=1}^{\lfloor\frac{\alpha/2}{\alpha-1}\rfloor} \frac{\Psi_\ell(z_\beta)}{{M_\alpha}^{\ell}}. \tag{25}$$

Technically speaking, when plugging $u = G^{-1}_{\mathcal{J}_{n,\alpha}}(z_\beta)$ into (15), the term Ψ_k will release expansion terms at the orders of ${M_\alpha}^{-k}, {M_\alpha}^{-(k+1)}, \ldots, {M_\alpha}^{-\lfloor(\alpha/2)/(\alpha-1)\rfloor}$. Therefore, we formulate Ψ_k's recursively. We describe step 1 ($\ell = 1$):

(i) Only keep Γ_0 and Γ_1 on the RHS of (15), temporarily ignoring other Γ_ℓ's. Do the same for $G^{-1}_{\mathcal{J}_{n,\alpha}}$ (only keep Γ_0 and Ψ_1).

(ii) Plug $u = G^{-1}_{\mathcal{J}_{n,\alpha}}(z_\beta)$ into (15).

(iii) Set the sum of ${M_\alpha}^{-1}$ terms to zero. This would solve Ψ_1.

To solve Ψ_2, add Γ_2 and Ψ_2 back into consideration in (i) and set the sum of ${M_\alpha}^{-2}$ terms to zero in (iii). Repeat this procedure until all Ψ_k's are solved.

Now, we formalize the above method. Readers who do not wish to read involved math may jump to Theorem 2.2. To start, set $\Psi_1(z_\beta) := -\Gamma_1(z_\beta)$. Then for each $k = 2, \ldots, \lfloor(\alpha/2)/(\alpha-1)\rfloor$ in order, recursively compute $\Psi_k(z_\beta)$ by

$$\Psi_k(z_\beta) \cdot \phi(z_\beta) = -\sum_{\ell'=2}^{k} \left\{ \sum_{\substack{j_1,\ldots,j_{\ell'}: \\ 1\leqslant\{j_1,\ldots,j_{\ell'}\}\leqslant k-\ell'+1 \\ j_1+\cdots+j_{\ell'}=k}} \Psi_{j_1}(z_\beta)\cdots\Psi_{j_{\ell'}}(z_\beta) \cdot \frac{\phi^{(\ell'-1)}(z_\beta)}{(\ell')!} \right\}$$

$$- \sum_{\substack{k_1,k_2: k_1+k_2=k \\ k_1=0,\ldots,k-1}} \left[\left\{ \phi(z_\beta) \cdot \mathbb{1}_{[k_1=0]} + \sum_{\ell'=1}^{k_1} \sum_{\substack{j_1,\ldots,j_{\ell'}: \\ 1\leqslant\{j_1,\ldots,j_{\ell'}\}\leqslant k_1-\ell'+1 \\ j_1+\cdots+j_{\ell'}=k_1}} \Psi_{j_1}(z_\beta)\cdots\Psi_{j_{\ell'}}(z_\beta) \cdot \frac{\phi^{(\ell')}(z_\beta)}{(\ell')!} \right\} \right.$$

$$\times\left\{\Gamma_{k_2}(z_\beta)+\sum_{\ell'=1}^{k_2-1}\sum_{\ell''=1}^{k_2-\ell'}\left\{\sum_{\substack{j_1,\dots,j_{\ell''}:\\ 1\leqslant\{j_1,\dots,j_{\ell''}\}\leqslant k_2-\ell'-\ell''+1\\ j_1+\cdots+j_{\ell''}=k_2-\ell'}}\Psi_{j_1}(z_\beta)\cdots\Psi_{j_{\ell''}}(z_\beta)\cdot\frac{\Gamma_{\ell'}^{(\ell'')}(z_\beta)}{(\ell'')!}\right\}\right\}\Bigg]. \tag{26}$$

To provide readers a more concrete view of the result, let us calculate the first three Ψ_k's.

Range of α	k[6]	Formula for computing Ψ_k[7]
$[4/3, 2]$	1	$\Psi_1 = -\Gamma_1$
$[6/5, 4/3)$	2	$-\Psi_2 = (\Gamma_1'\Psi_1 + \Gamma_2) + (\Psi_1^2/2 + \Psi_1\Gamma_1)\phi'/\phi$ [8]
$[8/7, 6/5)$	3	$-\Psi_3 = (\Gamma_3 + \Psi_2\Gamma_1' + \Psi_1^2\Gamma_1''/2 + \Psi_1\Gamma_2')$ $+(\Psi_1\Psi_2{+}\Psi_1\Gamma_2 + \Psi_1^2\Gamma_1' + \Psi_2\Gamma_1)\phi'/\phi + (\Psi_1^3/6 + \Psi_1^2\Gamma_1/2)\phi''/\phi$

Table 2: Examples of C-F expansion formulas

From (26) and Table 2, we see that all C-F expansion terms are functions of Γ_ℓ's. Thus, replacing Γ_ℓ's by $\widetilde{\Gamma}_\ell$'s, we obtain the empirical C-F expansion, denoted by $\widetilde{G}^{-1}_{\mathcal{J}_{n,\alpha}}(\cdot)$.

Theorem 2.2. *Under the conditions of Theorem 2.1, for any given $\beta \in (0,1)$, the population and empirical Cornish-Fisher expansions respectively satisfy*

$$F_{T_J+\delta_J|\mathcal{J}_{n,\alpha}}\big(G^{-1}_{\mathcal{J}_{n,\alpha}}(z_\beta)\big) = \beta + O\big(n^{-\alpha/2}\log^{1/2} n\big), \tag{27}$$

$$\big\|\widetilde{G}^{-1}_{\mathcal{J}_{n,\alpha}}(u) - G^{-1}_{\mathcal{J}_{n,\alpha}}(u)\big\|_\infty = O\big(n^{-\alpha/2}\log^{1/2} n\big). \tag{28}$$

Corollary 2.2. *Under the conditions of Theorem 2.1, the Cornish-Fisher confidence interval $\mathcal{I}_\beta$ defined by*

$$\mathcal{I}_\beta := \Bigg(U_J - (\widetilde{G}^{-1}_{\mathcal{J}_{n,\alpha}}(z_{1-\beta/2}) - \delta_J)\cdot|\mathcal{J}_{n,\alpha}|^{-1}\Big\{\sum_{i=1}^n a^2_{n,\alpha;1}(i)\Big\}^{1/2}\cdot\widetilde{\xi}_1,$$
$$U_J - (\widetilde{G}^{-1}_{\mathcal{J}_{n,\alpha}}(z_{\beta/2}) - \delta_J)\cdot|\mathcal{J}_{n,\alpha}|^{-1}\Big\{\sum_{i=1}^n a^2_{n,\alpha;1}(i)\Big\}^{1/2}\cdot\widetilde{\xi}_1\Bigg)$$

[6] This is the maximum k such that Ψ_k appears in the C-F expansion. It equals $\lfloor(\alpha/1)/(\alpha-1)\rfloor$.

[7] Since all functions are evaluated at z_β, we omit all "(z_β)" notions, e.g., we only write "Ψ_1" for "$\Psi_1(z_\beta)$".

[8] The formula for Ψ_2 uses the Ψ_1 computed in the "$k = 1$" case. The same goes for the formula for Ψ_3.

enjoys a higher-order accurate control of the actual coverage probability around $1-\beta$*:* $\mathbb{P}\big(\mu \in \mathcal{I}_\beta | \mathcal{J}_{n,\alpha}\big) = 1-\beta + O(n^{-\alpha/2}\log^{1/2} n)$.

2.3.2 Two remarks

First, as mentioned in Section 1, reducing the U-statistic inflates $\mathrm{Var}(U_J)$. However, we studentize U_J by $\widetilde{\sigma}_{J;1}$, which only captures the leading term in $\mathrm{Var}(U_J)$, whose order does *not* vary with α. Readers naturally wonder where the variance inflation is reflected in our statistical inference procedure. Here, we use our CI formula as an example to clarify.

Remark 2.4. *The radius of our Cornish-Fisher CI is* $O\big(n^{-1/2}+n^{-(\alpha-1/2)}\big)$[9]*. Studentizing* U_J *with* $\widetilde{\sigma}_J$ *will also yield a CI radius of* $\{O(n^{-1}+n^{-\alpha})\}^{1/2} = O\big(n^{-1/2}+n^{-(\alpha-1/2)}\big)$*. In other words, using* $\widetilde{\sigma}_J$ *or* $\widetilde{\sigma}_{J;1}$ *to studentize* U_J *lead to different pivots as intermediate steps, but eventually, their eventually produced CI lengths are on the same order.*

Our second remark regards test power. In fact, any test based on an asymptotically $N(0,1)$ pivot (including our method) is asymptotically power-optimal (see how Theorem 3.5 of [1] establishes asymptotic power-optimality). We reiterate that power-optimality and risk control accuracy are *distinct* goals. As pointed out in [36], achieving either goal alone is not difficult, however, achieving both is usually rather challenging. To our best knowledge, our work is the first to achieve both goals for inference based on reduced U-statistics.

[9] To see this, notice that $\Gamma_0(-u) = \Gamma_0(u)$, while $\Gamma_\ell(-u) = -\Gamma_\ell(u)$ for all $\ell \geqslant 1$. Also notice that $(\widetilde{G}^{-1}_{\mathcal{J}_{n,\alpha}}(z_{1-\beta/2}) - \delta_J) \asymp 1 + n^{-(\alpha-1)}$ and $|\mathcal{J}_{n,\alpha}|^{-1}\big\{\sum_{i=1}^n a^2_{n,\alpha;1}(i)\big\}^{1/2} \asymp n^{-1/2}$.

3 Our method: application to specific designs

In this subsection, we apply our general results in Section 2 to analyzing several designs. First, we propose and analyze a novel variance-optimal deterministic reduction scheme in Section 3.1. Then in Section 3.2, we present the first provably higher-order accurate inference for a few randomized designs [28, 8].

3.1 A novel variance-optimal deterministic design

As discussed in Section 2.1, existing works typically focused on minimizing the variance for special configurations. In this section, we present a novel method to *principally* construct variance-minimizing $\mathcal{J}_{n,\alpha}$ for general (α, r). To start, recall an important simplification that we proposed in Remark 2.2 that we only need to consider $\alpha \in (1,2)$. The key to minimize $\mathrm{Var}(U_J)$ is that the design $\mathcal{J}_{n,\alpha}$ needs to satisfy the following properties.

(D1) All $a_{n,\alpha;1}(i)$'s are equal;

(D2) For all $k \geqslant 2$ and $I_k \in \mathcal{C}_n^k$, all $a_{n,\alpha;k}(I_k)$'s are 0 or 1; or equivalently, any two member sets of $\mathcal{J}_{n,\alpha}$ may not overlap (intersect) by more than 1 index.

Now we describe our design. We set $\mathcal{J}_{n,\alpha}$ to be the union of a few $\mathcal{J}_{n,\alpha}^{(d)}$ sets, defined as

$$\mathcal{J}_{n,\alpha}^{(d)} := \left\{ \left(i + (2^{1-1}-1)d, i + (2^{2-1}-1)d, \cdots, i + (2^{r-1}-1)d \right) : i = 1, \ldots, n \right\}, \qquad (29)$$

where we circulate indexes outside the range $[1:n]$. For instance, when $r = 3$ as in Example 2.1, we have $\mathcal{J}_{n,\alpha}^{(1)} = \{(1,2,4),(2,3,5),\ldots,(n,1,3)\}$. Clearly, any individual $\mathcal{J}_{n,\alpha}^{(d)}$ satisfies both (D1) and (D2). But when we union a few $\mathcal{J}_{n,\alpha}^{(d)}$ sets, we need to watch out for the compliance with (D2). For example, $(1,2,4)$ from $\mathcal{J}_{n,\alpha}^{(1)}$ and $(2,4,8)$ from $\mathcal{J}_{n,\alpha}^{(2)}$ overlap by 2 indexes, violating (D2). We meticulously select the set of d values to avoid such multiple overlap. Our choice is:

$$\mathcal{J}_{n,\alpha} := \bigcup_{d=b_1\cdot n^{\alpha-1}}^{b_2\cdot n^{\alpha-1}} \mathcal{J}^{(d)}_{n,\alpha}, \tag{30}$$

where b_1, b_2 are chosen according to the following lemma.

Lemma 3.1. *Suppose $n \gg r$. Set $\alpha \in (1,2)$ and $b_1/b_2 \in \big((2^{r-1}-1)/2^{r-1}, 1\big)$. Our design $\mathcal{J}_{n,\alpha}$ specified by (29) and (30) satisfies $a_{n,\alpha;1}(i) = n^{\alpha-1}$ and $a_{n,\alpha;k}(I_k) \in \{0,1\}$, for all $i \in [1:r]$ and $I_k \in \mathcal{C}_n^k$, $k \in [2:r]$. Thus it satisfies (D1) and (D2) and minimizes* $\mathrm{Var}(U_J)$.

Lemma 3.1 ensures that this $\mathcal{J}_{n,\alpha}$ satisfies Assumption 2. Therefore, Theorem 2.1 and Corollaries 2.1 and 2.2 apply. This $\mathcal{J}_{n,\alpha}$ also greatly simplifies the Edgeworth formulas.

Corollary 3.1. *Under our design $\mathcal{J}_{n,\alpha}$ as described by (30) and Lemma 3.1, we have*

$$\Gamma_0(u) = \frac{2u^2+1}{6\xi_1^3}\mathbb{E}[g_1^3(X_1)] + \frac{(r-1)(u^2+1)}{2\xi_1^3}\mathbb{E}[g_1(X_1)g_1(X_2)g_2(X_1,X_2)], \tag{31}$$

$$\Gamma_\ell(u) = -\left\{\frac{\sum_{k=2}^r \xi_k^2\binom{r}{k}}{(b_2-b_1)r^2\xi_1^2}\right\}^\ell \cdot \frac{H_{2\ell-1}(u)}{(2\ell)!} = -\left\{\frac{\sigma_h^2 - r\xi_1^2}{(b_2-b_1)r^2\xi_1^2}\right\}^\ell \cdot \frac{H_{2\ell-1}(u)}{(2\ell)!}, \tag{32}$$

for $\ell = 1, \ldots, \lfloor \alpha/\{2(\alpha-1)\}\rfloor$, where $\sigma_h^2 := \mathrm{Var}(h(X_{[1:r]}))$ in (32).

We can estimate σ_h^2 by

$$\widetilde{\sigma}_h^2 := \frac{1}{nM_\alpha}\sum_{i=1}^{n}\sum_{d=1}^{M_\alpha} h^2(X_{[i:d:(i+(r-1)d)]}) - \widetilde{\mu}^2, \tag{33}$$

where in contrast to (9), we should multiply two *identical* $h(X_{I_r})$ terms in term 1 in (33). Now the empirical Edgeworth expansion formula $\widetilde{G}_{\mathcal{J}_{n,\alpha}}(u)$ for hypothesis testing can be computed by combining (15), (19), (20) and (31)–(33). Then with (25) and (26), we can compute the Cornish-Fisher confidence interval. We skip repetitive formula presentation.

Interestingly, our method not only serves as an acceleration tool itself but also enhances the performance of other acceleration tools. One example is the *divide-and-conquer acceleration* through parallel computing [7]. They utilize K parallel computing servers that return summary statistics to a main server for aggregation. But in [7], each server still computes a *complete* U-statistic, leaving significant space for further acceleration. Here, we present Algorithm 1 that couples our method with the divide-and-conquer idea in [7]. In fact, this algorithm can be viewed as a parallelized version of our own method.

Algorithm 1 Our method + Chen-Peng Reduction

```
Input: Data: $X_1, \dots, X_n$; kernel function $h(x_1, \dots, x_r)$; $\alpha$; number of servers $K$; $(b_1, b_2)$.
Output: Coefficients of the empirical Edgeworth expansion $\widehat{G}_J(u)$.
Part I: data splitting
for $k = 1 : K$ do
    Pass: $h, n, b_1, b_2, X_{[(k-1)n/K+1-(r-k)n^{\alpha-1}):(kn/K+\max\{(r-1)n^{\alpha-1},(2^{r-1}-1)n^{\alpha-1}\}]}$ to server $k$.
end for $k$
Part II: local computation
for $k = 1 : K$ do  (On the $k$th local server, compute the following quantities.)
    • Compute and return:
        $U_{J;k} := \frac{1}{n^\alpha/K} \sum_{I_r \in \mathcal{J}_{n,\alpha;k}} h(X_{I_r})$    (34)
      with $\mathcal{J}_{n,\alpha;k} := \bigcup_{d=b_1 n^{\alpha-1}}^{b_2 n^{\alpha-1}} \mathcal{J}_{n,\alpha;k}^{(d)}$, where $\mathcal{J}_{n,\alpha;k}^{(d)}$ is defined similarly to $\mathcal{J}_{n,\alpha}^{(d)}$ in (29),
      except that $i$ ranges in $[\{(k-1)n/K+1\} : (kn/K)]$ instead of $[1 : n]$.
    • Compute and return:
        $\widehat{\mathcal{E}}_{g_1;3}^{(d)} := \frac{1}{n/K} \sum_{i \in [((k-1)n/K+1):(kn/K)]} h(X_{[(i-r+1):i]}) h(X_{[i:(i+r-1)]}) h(X_{[(i+r-1):(i+2r-2)]})$,
        $\widehat{\mathcal{E}}_{g_1 g_1 g_2}^{(k)} := \frac{1}{n/K} \sum_{i \in [((k-1)n/K+1):(kn/K)]} h(X_{[(i-r+1):i]}) h(X_{[i:(i+r-1)]}) h(X_{[(i+r-1):(i+2r-2)]})$.
    • For each $\ell \in [0 : r]$, compute and return:
        $\widehat{\eta}_{\ell;k} := \frac{1}{n^\alpha/K} \sum_{i \in [((k-1)n/K+1):(kn/K)]} \sum_{d=1}^{n^{\alpha-1}} h(X_{[i:d:(i+(r-1)d)]}) h(X_{[(i+(k-1)d):(-d):(i-(r-k)d)]})$.
```

end for k

Part III: result aggregation

On the central server, compute and output:

$$\begin{aligned}
U_J &:= \frac{1}{K}\sum_{k=1}^{K} U_{J;k},\\
\widehat{\mu}^2 &:= \frac{1}{K}\sum_{k=1}^{K} \widehat{\eta}_{0;k},\\
\widehat{\xi}_1^2 &:= \frac{1}{K}\sum_{k=1}^{K} \widehat{\eta}_{1;k} - \widehat{\mu}^2,\\
\widehat{\xi}_\ell^2 &:= \frac{1}{K}\sum_{k=1}^{K} \widehat{\eta}_{\ell;k} - \widehat{\mu}^2 - \sum_{\ell'=1}^{\ell}\binom{\ell}{\ell'}\widehat{\xi}_{\ell'}^2,\\
\widehat{\mathbb{E}}[g_1^3(X_1)] &:= \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathcal{E}}_{g_1;3}^{(d)} - \widehat{\mu}^3,\\
\widehat{\mathbb{E}}[g_1(X_1)g_1(X_2)g_2(X_1,X_2)] &:= \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathcal{E}}_{g_1 g_1 g_2}^{(d)} - \widehat{\mu}^3 - 2U_J\cdot\widehat{\xi}_1^3.
\end{aligned}$$

Finally, plug these estimated quantities into Corollary 3.1 for statistical inference.

We compare our method coupled with [7] to the vanilla [7] in Table 3. For clarity, we unified all split sizes, set $K \asymp n^{\tau'}$ as in [7] and aligned the orders of the second leading terms in the variance formulas of both approaches, by setting $\alpha = 2 - \tau'$. Table 3 shows that our method speeds up [7] by a factor of $n^{r+1-\alpha}$, without noticeable relative variance inflation and achieving a higher risk control accuracy.

Table 3: Our method enhances [7]'s method. Set $\alpha \in (1,2)$. Recall $r \geqslant 2$.

	Vanilla [7]	Our method + [7]
Time cost on each server	$O\big(n^{(r-1)(\alpha-1)+1}\big)$	$O\big(n^{(\alpha-2)(\alpha-1)+1}\big)$
Variance of aggregated U-stat.	$r^2\xi_1^2/n + O(n^{-\alpha})$	$r^2\xi_1^2/n + O(n^{-\alpha})$
CDF approximation error	$o(n^{-1/2})$[10]	$O(n^{-\alpha/2})$
Risk control accuracy	$o_p(1)$[11]	$\widetilde{O}_p(n^{-\alpha/2}\log^{1/2} n)$

3.2 Analysis of randomized incomplete U-statistics

Our general framework in Section 2 is a powerful tool for analyzing randomized designs. Here, we showcase its application to some popular designs (and close variants) in literature:

(J1) Sample n^α size-r subsets from $\mathcal{C}_n^r$ at random, with replacement.

(J2) Similar to (J1), but sample without replacement[12].

(J3) For $i = 1, \dots, n$, sample $n^{\alpha-1}$ size-r subsets from $\mathcal{C}_n^r$ containing i, with replacement.

(J4) Similar to (J3), but for each i, sample without replacement[13].

These sampling schemes are very natural, and there are many more similar randomized designs in existing literature [3, 8]. However, no available theory and methods yet exist to provide higher-order accurate risk control for inference under these schemes. Conventional analysis [8] typically starts with re-expressing U_J as follows.

$$U_J - \mu := \underbrace{(U_n - \mu)}_{\text{(Part I)}} + \underbrace{|\mathcal{J}_{n,\alpha}|^{-1} \sum_{I_r \in \mathcal{J}_{n,\alpha}} \left\{h(X_{I_r}) - U_n\right\}}_{\text{(Part II)}} =: (U_n - \mu) + V_J, \tag{35}$$

where part I is a rescaled complete U-statistic (see definition in Eq. (1)) and part II captures the randomness in $\mathcal{J}_{n,\alpha}$. One can normal-approximate both parts and eventually U_J, via careful conditioning and convolution, see page 9–20 in [9]. While (35) is useful for analyzing degenerate U-statistics, it is not a sharp tool in the non-degenerate case, where the two parts, dependent on each other, both noticeably impact the Edgeworth formula.

[10] This further requires $K = O(n^{\tau'})$ for $\tau' \in (0, 1/4)$, see Theorem 3.3-(i) in [7].

[11] [7] *standardizes* U_J, therefore, their inference is *not* higher-order accurate, that is unless it further employs a "bias-correction" that consults and eventually reproduces our method. See [18], Section 3.10.2.

[12] In theory, sampling $\mathcal{J}_{n,\alpha} : |\mathcal{J}_{n,\alpha}| = O(n^\alpha)$ without replacement could be done within $O(n^\alpha)$ budget, in terms of both time and memory, via a lexicographic indexing of $\mathcal{C}_n^r$.

[13] But subsets from different i-strata can still coincide

In sharp contrast, our analysis takes a very different route: the key is to apply our general framework in Section 2 to analyze U_J directly, without going through (35). As a premise, we first verify that these randomized designs indeed satisfies Assumption 2 with high probability. (Assumption 1 is easily verified.)

Lemma 3.2. *Let $\mathcal{J}_{n,\alpha}$ be constructed by one of (J1)–(J4). For any given constant $C_0 > 0$, there exist constants $C_1, C_2 : C_2 > C_1 > 0$ depending on C_0 and the design $\mathcal{J}_{n,\alpha}$, such that Assumption 2 with these C_1 and C_2 holds with probability at least $1 - n^{-C_0}$.*

All four designs (J1)–(J4) have clean analytical Edgeworth formulas, which can be handily found by taking another layer of expectation $\mathbb{E}_J[\cdot]$ over the randomness of $\mathcal{J}_{n,\alpha}$.

Corollary 3.2. *Under the setting $\alpha \in (1, 2)$, we have the following results.*

- *For randomized designs (J1) and (J2), we have*

$$\mathbb{E}_J[\Gamma_0(u)] := \frac{2u^2+1}{6\xi_1^3}\mathbb{E}[g_1^3(X_1)] + \frac{(r-1)(u^2+1)}{2\xi_1^3}\mathbb{E}[g_1(X_1)g_1(X_2)g_2(X_1,X_2)], \quad (36)$$

$$\mathbb{E}_J[\Gamma_\ell(u)] := -\frac{H_{2\ell-1}(u)}{(2\ell)!}\left\{\frac{\sum_{k=2}^r \binom{r}{k}\xi_k^2}{r^2\xi_1^2}\right\}^\ell, \quad \textit{for } \ell \geqslant 1. \quad (37)$$

- *For randomized designs (J3) and (J4), we have*

$$\begin{aligned}\mathbb{E}_J[\Gamma_0(u)] :=& \frac{2u^2+1}{6\xi_1^3}\mathbb{E}[g_1^3(X_1)] \\ &+ \frac{(r-1)\{(r^3+2r^2-2)u^2+r^3-2r^2+2\}}{2r^3\xi_1^3}\mathbb{E}[g_1(X_1)g_1(X_2)g_2(X_1,X_2)], \end{aligned} \quad (38)$$

$$\mathbb{E}_J[\Gamma_\ell(u)] \textit{ is the same as the } \mathbb{E}_J[\Gamma_\ell(u)] \textit{ under (J1) and (J2).} \quad (39)$$

Then set

$$G_J(u) := \Phi(u) + \phi(u)\left\{\frac{\mathbb{E}_J[\Gamma_0(u)]}{\sqrt{n}} + \sum_{\ell=1}^{\lfloor\frac{\alpha/2}{\alpha-1}\rfloor}\frac{\mathbb{E}_J[\Gamma_\ell(u)]}{\widetilde{M}_\alpha^{\;\ell}}\right\}, \quad (40)$$

where

$$\widetilde{M}_\alpha := n^{\alpha-1}\cdot\big\{1+1/(rn^{\alpha-1})\big\}\Big/\begin{cases}1+\dfrac{n^{\alpha-2}\xi_2^2\cdot r(r-1)}{\sum_{k=2}^r\binom{r}{k}\xi_k^2}, & \textit{under (J1) and (J2)},\\[2ex] 1+\dfrac{n^{\alpha-2}\xi_2^2\cdot r^2(r-1)}{2\sum_{k=2}^r\binom{r}{k}\xi_k^2}, & \textit{under (J3) and (J4)}.\end{cases} \tag{41}$$

We have

$$\big\|F_{T_J+\delta_J}(u)-G_J(u)\big\|_\infty = O(n^{-\alpha/2}\log n). \tag{42}$$

We can naturally define the empirical version $\widetilde{G}_J(u)$ with coefficient estimated by (9), (19), (20) and (33) and use it for downstream analysis, accompanied by theoretical guarantees exactly similar to Corollaries 2.1 and 2.2. We skip the repetitive detailed descriptions.

We conclude this section by instantiating the general formula for the Cornish-Fisher confidence interval, using the formula under (J1). Define $\sigma^2_{h,(-1)} := \frac{\sum_{k=2}^r\binom{r}{k}\xi_k^2}{r^2\xi_1^2}$. We have

Range of α	k	$\Psi_k(u)$
$[4/3,2]$	1	$\frac{1}{2}u\sigma^2_{h,(-1)}$
$[6/5,4/3)$	2	$\frac{1}{24}\big\{(u^3-3u)\sigma^2_{h,(-1)}+3u\sigma^4_{h,(-1)}\big\}$
$[8/7,6/5)$	3	$\frac{u}{720(u^2-1)}\Big\{(u^6-11u^4+25u^2-15)\sigma^2_{h,(-1)}$ $+45(u^2-1)^2\sigma^4_{h,(-1)}-(15u^2-45)\sigma^6_{h,(-1)}\Big\}$

Table 4: First three Ψ_k's under (J1).

4 Simulations

We assess the accuracy of the CDF approximation for noiseless non-degenerate U-statistics. The goal is to accurately approximate $F_{T_J+\delta_J}$, where we set a small variance with $C_\delta = 0.008$ for δ_J. We generate synthetic data with $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim}$ PDF: $(x+1)/2$, $x\in[-1,1]$, and use the kernel function $h(x_1, x_2, x_3) := \sin(x_1 + x_2 + x_3)$. We experiment with our proposed deterministic design from Section 3.1 and the random design (J1) from Section 3.2. We compare our method to the following benchmarks: 1. $N(0,1)$; 2. resample bootstrap (bootstrap iteration $B = 200$ [29]); and 3. subsample bootstrap (subsample size: $n^{1/2}$). To emulate the true sampling distribution of $T_J + \delta_J$, we use a Monte-Carlo approximation with $n_{\mathrm{MC}} := 10^6$ samples[14]. The performance measure is:

$$\sup_{u\in[-2,2];u\in\mathbb{Z}/10} \left|\widehat{F}_{T_J+\delta_J}(u) - F_{T_J+\delta_J}(u)\right|. \tag{43}$$

We vary $n \in \{10, 20, 40, 80\}$ and set $\alpha = 1.5$ (results for $\alpha = 1.7$ are provided in Supplementary Material). For each (n, α) setting, we repeated the experiment 30 times and recorded the mean and standard deviation of the distribution approximation errors (43).

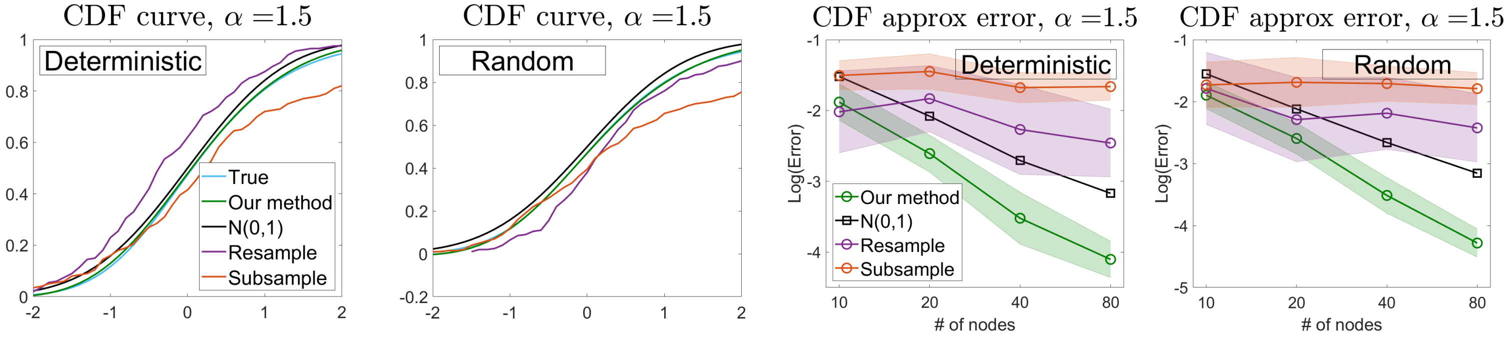

Figure 2: CDF approximation accuracy: plots 1–2: true CDF $= F_{T_J+\delta_J}(u)$, $n = 80$; plots 3–4: log-transformed CDF approximation error.

Figure 2 shows the true and estimated CDF curves for $T_J + \delta_J$. Our method's estimated CDF almost overlaps the true curve; whereas all other methods exhibit much more noticeable estimation errors. It also shows the log-transformed CDF approximation errors of all methods under different (n, α) configurations. Our method shows clear advantage in accuracy across all settings, and we are the only method that exhibits an empirical

[14] We need to set n_{MC} to be much larger than $(1/e^{-5})^2 \approx 2.2 \times 10^4$, in view of DKW inequality.

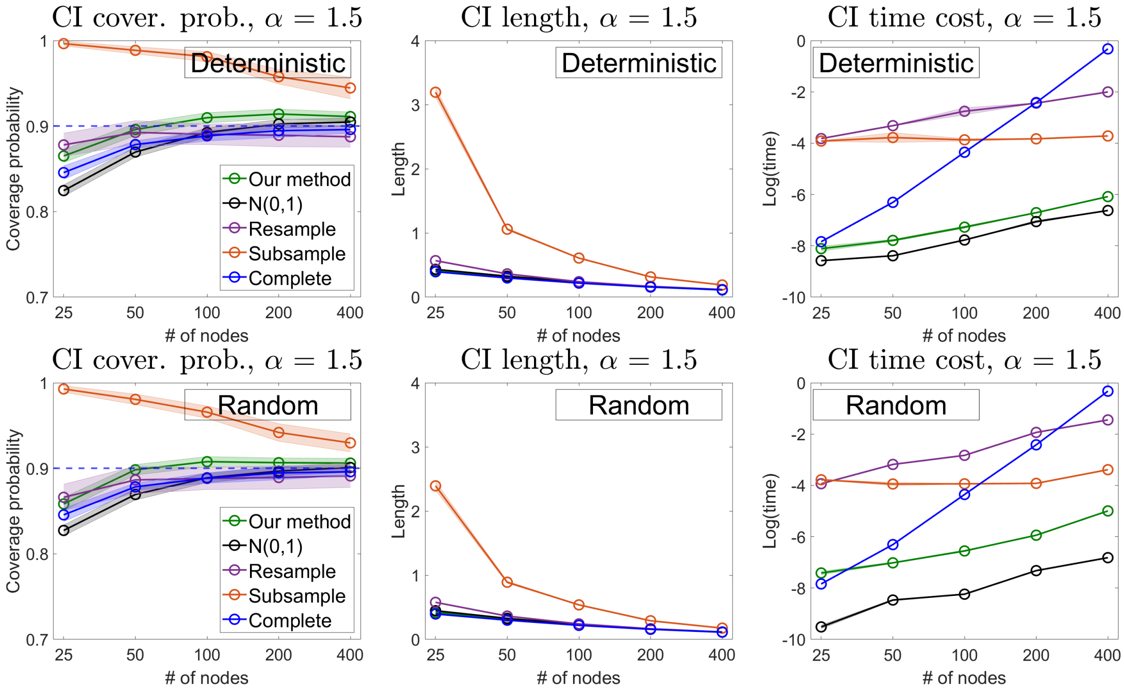

Figure 3: CI-related performance measures: column 1: CI coverage probability, dashed blue line = 90%; column 2: CI length; column 3: log-transformed time cost (log-second).

error rate faster than $n^{-1/2}$. All these results well-align with our theory's prediction and demonstrate the higher-order accuracy of our method.

Next, we compare our Cornish-Fisher confidence interval to that produced by the benchmark methods in Simulation 1, plus the C-F CI constructed based on the complete U-statistic. Performance measurements include: coverage probability, CI length and computation time. We fix the confidence level at $1 - \beta = 90\%$ and focus on the two-sided CI for simplicity. The simulation set up is mostly inherited from Simulation 1, except that now we no longer need a large n_{MC} and can test for larger n's: $n \in \{25, 50, 100, 200, 400\}$. In each experiment, which will produces one empirical CI coverage probability, we generate 3000 CI's for our method, $N(0, 1)$ and complete U-statistics; and 500 CI's for resampling and subsampling bootstraps since they are slower. Then we repeat the experiment 100 times for all methods except the complete U-statistic method (repeated 20 times) to evaluate the variance of the coverage probability of each method.

Figure 3 shows the result for deterministic and random designs. Our method shows clear advantage in accuracy of controlling the empirical coverage probability around the nominal level level of 90%, significantly improving over normal approximation, especially for small n's. As n grows large, our method's speed advantage over bootstrap methods becomes clearer. Compared to inference based on complete U-statistic, our method effectively reduces computational complexity, reflected by its much flatter log-time curve, without noticeable loss in risk control accuracy. All methods except subsampling bootstrap produce similar CI lengths. This echoes our earlier remarks that the CI length reflects a different aspect of U-statistic reduction (inference power, Section 2.3.2); and different approaches may perform similarly in this aspect, if they are all asymptotically normal approximations.

5 Data examples

5.1 Data example 1: Stock market data

The S&P 500 historical data [11] records the daily prices of 412 stocks from 11 sectors. Following [5], we computed the *monthly logarithmic return rates* of each stock from 1-Mar-2000 to 29-Aug-2022, yielding $n = 138$ observations. Our goal is to assess the pairwise dependency between sectors via independence tests. Denote the log-return sequence of stock i from sector X by $S_i^X = (S_{i,1}^X, ..., S_{i,n}^X)$; similarly define S_t^Y. We measure dependency between sectors X and Y by dCov, rewritten as a complete U-statistic (Lemma 1 of [42]):

$$\mathrm{dCov}^2(X,Y) := \binom{n}{4}^{-1} \sum_{i<j<q<r} h(Z_i, Z_j, Z_q, Z_r), \tag{44}$$

where $h(Z_i, Z_j, Z_q, Z_r) := \sum_{s,t,u,v}^{i,j,q,r}(a_{st}b_{uv} + a_{st}b_{st} - a_{st}b_{su} - a_{st}b_{tv})/24$ [15], $a_{ij} = \|S_i^X - S_j^X\|_2$, and $b_{ij} = \|S_i^Y - S_j^Y\|_2$. Set $\alpha = 1.5$. We test $H_0 : \mathbb{E}[\mathrm{dCov}^2(X, Y)] = 0$ between each sector pair, versus a two-sided alternative. As a reference, on the diagonal, we randomly split the stocks in each sector into two sets and tested their dependency. Figure 4 shows that our method well-aligns with the test decisions that would have been made using the complete U-statistic, but our method computes much faster (see Table 5). On the diagonal, the sectors that exhibit strongest inner dependency include `CD, E, F, I` and `IT`. This is understandable since they tend to be more sensitive to global economic fluctuations. In contrast, members of `CmS, CnS` and `U` sectors focus more on local markets, so their within-sector price fluctuations are less synchronized. This understanding also applies to cross-sector relations, such as the tight connection between the pairs (`CD`, `I`) and (`I`, `IT`), whereas `U` is comparatively less dependent on other sectors except `E`.

5.2 Data example 2: UCR time series data (Earthquakes, Starlight)

In the second example, we analyze two UCR time series data sets [12]: *Earthquakes* and *Starlight*. The earthquakes data consist of $n = 461$ earthquake curves, each of length $T = 512$. These curves are classified into $K = 2$ clusters: $n_0 = 368$ *non-major* and $n_1 = 93$ *major* earthquakes. Following the approach of [5] and [45], we treat each earthquake curve as a point in a Hilbert space and aim at comparing the population distributions of the curves of different types using Maximum Mean Discrepancy (MMD). We measure the distance between two earthquake curves by comparing their SRVF transforms [38], which synchronize their phases in the presence of amplitude discrepancy. However, computing the SRVF for each curve pair is slow [39]. To accelerate and also to tame the violent

[15] The summation notation "$\sum_{s,t,u,v}^{i,j,q,r}$" means summing (s, t, u, v) over all permutations of (i, j, q, r).

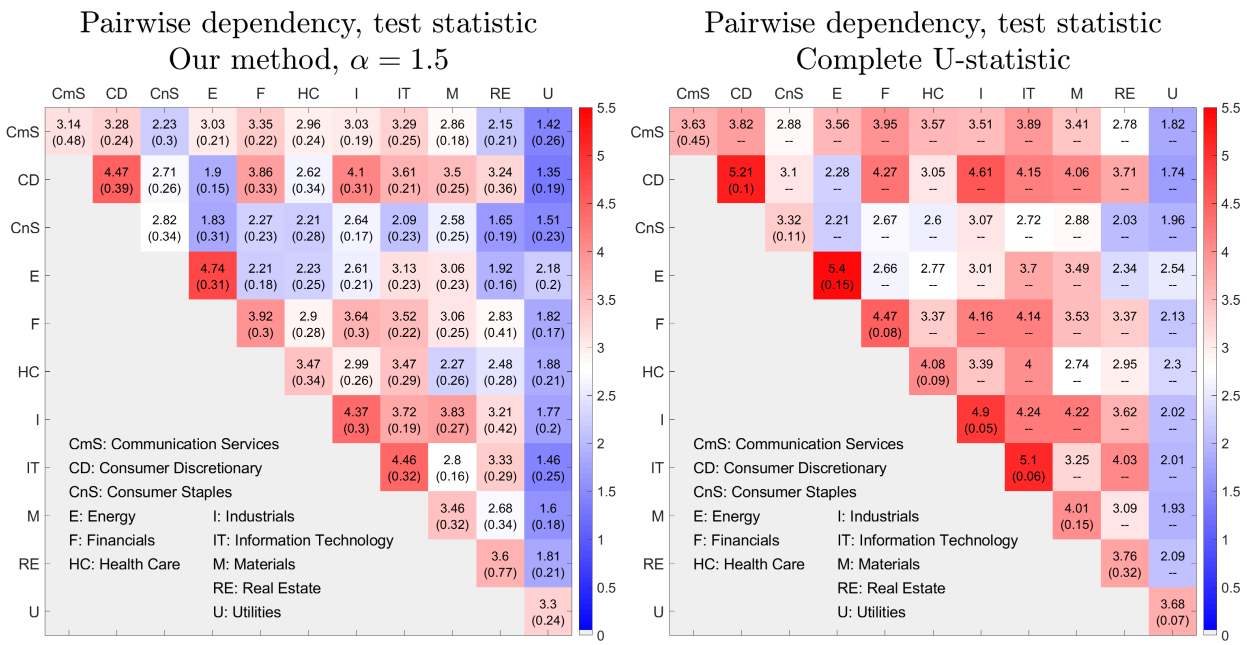

Figure 4: Pairwise dependency test: heatmaps of test statistics. High values (red): high detected dependency. Each cell reports mean(std.) of test statistics over 30 repeated experiments, except the off-diagonal of complete U-statistic method (no repetition needed).

fluctuation in the raw data, we pre-processed each curve $\{x_t\}_{t=1}^{512}$ by a moving average (window size ℓ) with down-sampling: $\left\{\widetilde{x}_t := \text{Mean}\big(x_{[\{t-(\ell-1)/2\}:\{t+(\ell-1)/2\}]}\big)\right\}_{t\in\{4k+1, k\in[0:127]\}}$. Due to page limit, we only present results for $\ell = 7$, leaving results for more window sizes to Supplementary Material.

We applied our method with $\alpha = 1.5$ to estimate the average pairwise distance (using SRVF) within each cluster to assess its internal cohesion. For the between-cluster comparison, we sub-sampled the larger group (non-major earthquakes) and rewrote the MMD a one-sample U-statistic following Equation (6) in [35] with the RBF kernel $k(x, y) := \exp(-\text{SRVF}(x, y)^2/5000)$. The we applied our method with $\alpha = 1.5$ to reduce this MMD U-statistic. Figure 5 shows the results, in which, we used the complete two-sample MMD U-statistic value in lieu of the unknown population mean discrepancy. Our Cornish-Fisher confidence intervals with randomized design (J1) demonstrate good coverage in both inference tasks for within- and between-cluster distances, respectively.

Next, we apply this analysis method to the much larger *Starlight* data set that contains $K = 3$ types of stars, with cluster sizes $n_1 = 1329$, $n_2 = 2580$ and $n_3 = 5327$. Here, each curve is a length 1024 sequence, which we down-sampled to length 128 without smoothing, because the starlight curves are much smoother than that in the earthquake data. Even with the down-sampling, evaluating a complete U-statistic for comparing any two star types remains computationally infeasible, due to the large sample sizes. Our method with $\alpha = 1.5$ allows users to implement a reduced version of Equation (6) in [35] with the RBF kernel $k(x, y) := \exp(-\text{SRVF}(x, y)^2/100)$. Due to page limit, in Figure 5, we only present the result for the comparison between type 1 and type 2 stars, relegating the rest to Supplemental Material. We observed that the MMD CI's produced by the starlight data are much narrower than the counterpart from the earthquakes data, possibly due to the much larger sample size. Also, for the between-cluster comparison, some MMD CI's of the earthquakes data contain 0 (will not reject H_0), while all CI's for the starlight data clearly support a two-sided alternative. This is echoed by the much smaller within-cluster distance and the clearer between-cluster differences in the starlight data.

Table 5: Time cost: our method ($\alpha = 1.5$) vs. complete U-statistic

Time cost	Stock Market ($r = 4$)	Earthquakes ($r = 2$)		
(Unit = sec.)	All	Major	Non-major	Maj. vs. Non-Maj.
Our method	3.47	303.94	2471.70	1223.50
Complete U	8099.73	708.99	11199.92	17912.91

Time cost	Starlight ($r = 2$)					
(Unit = sec.)	Type 1	Type 2	Type 3	1 vs. 2	1 vs. 3	2 vs. 3
Our method	4512.95	12773.76	41282.26	19140.13	19149.33	50413.75
Complete U	48227.72	158233.7	(Time out)	(Time out)	(Time out)	(Time out)

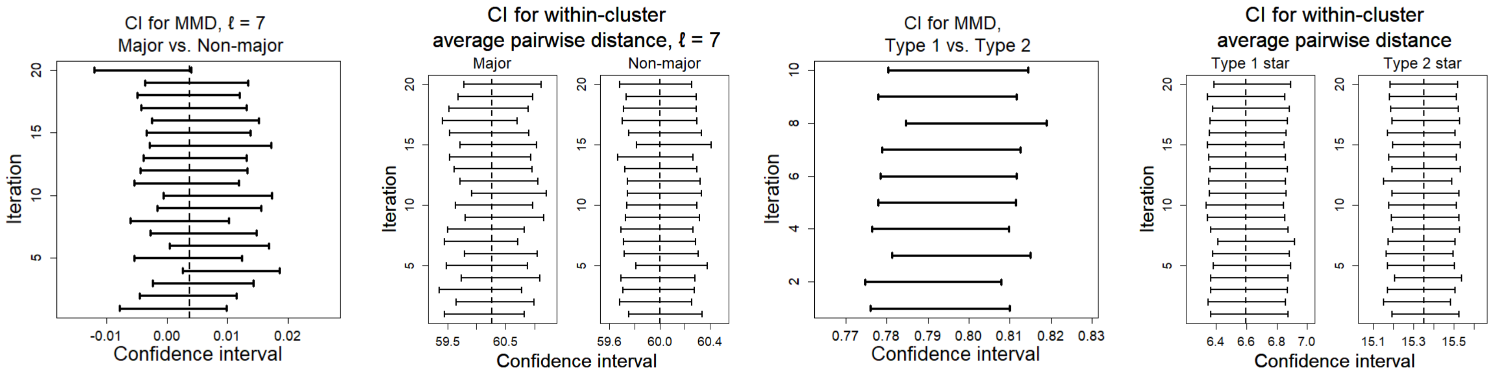

Figure 5: Results of data example 2. Plots 1–2: *Earthquakes*; plots 3–4: *Starlight*. Plots 1 & 3: 90% CI of based on reduced between-cluster MMD; column 4: 90% CI of within-cluster average pairwise distance (using SRVF [38]). Dashed line: complete U-statistic (evaluations of complete U-statistics timed out (> 48 hours) in most settings for *Starlight*).

6 Discussion

Our study throughout this paper exclusively focuses on *data-oblivious* reduction schemes. Recently, [24] proposed a *data-aware* reduction scheme, based on their key observation that $X_{[1:r]} \approx Y_{[1:r]}$ implies $h(X_{[1:r]}) \approx h(Y_{[1:r]})$, thus by clustering X_i's, one can effectively reduce the U-statistic's computation. While their method shows very attractive performance, finite-sample higher-order analysis for their method poses an interesting open challenge. There is also a computational price for *being data-aware*. For example, in the setting considered by [31], the clustering of all X_i's in some Banach space requires computing at least $O(n^2)$ many potentially expensive (like in our second data example) pairwise distances.

Acknowledgment

We thank the Editor, the AE and the Reviewer for very helpful comments that led to significant improvement of this paper. We thank Davide Giraudo for his reply to our question on Math-Stackexchange[16]; and Peter Craigmile, Sebastian Kurtek and Wen Zhou for helpful discussion; and Steven MacEachern and Ji Zhu for warm encouragements. Meijia Shao and Yuan Zhang were supported by NSF grant DMS-2311109. Dong Xia was supported by Hong Kong RGC Grant GRF 16303320.

Conflict of interest. The authors declare no conflict of interest.

References

[1] Banerjee, D. and Z. Ma (2017). Optimal hypothesis testing for stochastic block models with growing degrees. *arXiv preprint arXiv:1705.05305*.

[2] Bergsma, W. and A. Dassios (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli 20*(2), 1006–1028.

[3] Blom, G. (1976). Some properties of incomplete U-statistics. *Biometrika 63*(3), 573–580.

[4] Brown, B. and D. Kildea (1978). Reduced U-statistics and the Hodges-Lehmann estimator. *The Annals of Statistics*, 828–835.

[5] Chakraborty, S. and X. Zhang (2021). A new framework for distance and kernel-based metrics in high dimensions. *Electronic Journal of Statistics 15*(2), 5455–5522.

[6] Chaudhuri, A. and W. Hu (2019). A fast algorithm for computing distance correlation. *Computational Statistics & Data Analysis 135*, 15–24.

[7] Chen, S. X. and L. Peng (2021). Distributed statistical inference for massive data. *The Annals of Statistics 49*(5), 2851–2869.

[8] Chen, X. and K. Kato (2019a). Randomized incomplete U-statistics in high dimensions. *The Annals of Statistics 47*(6), 3127–3156.

[9] Chen, X. and K. Kato (2019b). Supplementary Material to "Randomized incomplete U-statistics in high dimensions". *The Annals of Statistics*.

[10] Clémençon, S., I. Colin, and A. Bellet (2016). Scaling-up empirical risk minimization: optimization of incomplete U-statistics. *The Journal of Machine Learning Research 17*(1), 2682–2717.

[11] (Datahub). S&P 500 companies with financial information. https://datahub.io/core/s-and-p-500-companies#data-cli. Accessed 18-Nov-2022.

[16] https://math.stackexchange.com/questions/4387849/

[12] Dau, H. A., E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, and Hexagon-ML (2018, October). The UCR time series classification archive. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

[13] Dürre, A. and D. Paindaveine (2021). On the consistency of incomplete U-statistics under infinite second-order moments. *arXiv preprint arXiv:2112.14666*.

[14] Even-Zohar, C. and C. Leng (2021). Counting small permutation patterns. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2288–2302. SIAM.

[15] Gao, C., Y. Lu, and H. H. Zhou (2015). Rate-optimal graphon estimation. *The Annals of Statistics 43*(6), 2624–2652.

[16] Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *The Journal of Machine Learning Research 13*(1), 723–773.

[17] Hall, P. (1983). Inverting an Edgeworth expansion. *The Annals of Statistics 11*(2), 569–576.

[18] Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

[19] Han, F. and T. Qian (2018). On inference validity of weighted U-statistics under data heterogeneity. *Electronic Journal of Statistics 12*(2), 2637–2708.

[20] Heller, Y. and R. Heller (2016). Computing the bergsma dassios sign-covariance. *arXiv preprint arXiv:1605.08732*.

[21] Helmers, R. (1991). On the Edgeworth expansion and the bootstrap approximation for a studentized U-statistic. *The Annals of Statistics 19*(1), 470–484.

[22] Huo, X. and G. J. Székely (2016). Fast computing for distance covariance. *Technometrics 58*(4), 435–447.

[23] Janson, S. (1984). The asymptotic distributions of incomplete U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 66*(4), 495–505.

[24] Kong, X. and W. Zheng (2021). Design based incomplete U-statistics. *Statistica Sinica 31*(3), 1593–1618.

[25] Lahiri, S. N. (1993). Bootstrapping the studentized sample mean of lattice variables. *Journal of Multivariate Analysis 45*(2), 247–256.

[26] Lee, A. J. (1979). On the asymptotic distribution of certain incomplete U-statistics. Technical report, North Carolina State University. Dept. of Statistics.

[27] Lee, A. J. (1982). On incomplete U-statistics having minimum variance. *Australian Journal of Statistics 24*(3), 275–282.

[28] Lee, A. J. (2019). *U-Statistics: Theory and Practice*. Routledge.

[29] Levin, K. and E. Levina (2019). Bootstrapping networks with latent space structure. *arXiv preprint arXiv:1907.10821*.

[30] Maesono, Y. (1997). Edgeworth expansions of a studentized U-statistic and a jackknife estimator of variance. *Journal of Statistical Planning and Inference 61*(1), 61–84.

[31] Moon, H. and K. Chen (2022). Interpoint-ranking sign covariance for the test of independence. *Biometrika 109*(1), 165–179.

[32] Putter, H. and W. R. van Zwet (1998). Empirical edgeworth expansions for symmetric statistics. *The Annals of Statistics 26*(4), 1540–1569.

[33] Rempala, G. and J. Wesolowski (2003). Incomplete U-statistics of permanent design. *Journal of Nonparametric Statistics 15*(2), 221–236.

[34] Rosenbaum, P. R. (2011). A new U-statistic with superior design sensitivity in matched observational studies. *Biometrics 67*(3), 1017–1027.

[35] Schrab, A., I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton (2021). MMD aggregated two-sample test. *arXiv preprint arXiv:2110.15073*.

[36] Shao, M., D. Xia, Y. Zhang, Q. Wu, and S. Chen (2022). Higher-order accurate two-sample network inference and network hashing. *arXiv preprint arXiv:2208.07573*.

[37] Slater, L. J. (1960). *Confluent Hypergeometric Functions*. Cambridge University Press.

[38] Srivastava, A., W. Wu, S. Kurtek, E. Klassen, and J. S. Marron (2011). Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*.

[39] Strait, J., O. Chkrebtii, and S. Kurtek (2019). Automatic detection and uncertainty quantification of landmarks on elastic curves. *Journal of the American Statistical Association 114*(527), 1002–1017.

[40] Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

[41] Weber, N. (1981). Incomplete degenerate U-statistics. *Scandinavian Journal of Statistics*, 120–123.

[42] Yao, S., X. Zhang, and X. Shao (2018). Testing mutual independence in high dimension via distance covariance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(3), 455–480.

[43] Zhang, Y. and D. Xia (2022). Edgeworth expansions for network moments. *The Annals of Statistics 50*(2), 726–753.

[44] Zhao, Q. (2019). On sensitivity value of pair-matched observational studies. *Journal of the American Statistical Association 114*(526), 713–722.

[45] Zhu, C. and X. Shao (2021). Interpoint distance based two sample tests in high dimension. *Bernoulli 27*(2), 1189–1211.