Electronic Journal of Statistics
Vol. 18 (2024) 2135–2207
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2247

On variance estimation of random forests with Infinite-order U-statistics

Tianning Xu, Ruoqing Zhu and Xiaofeng Shao

Department of Statistics, University of Illinois Urbana-Champaign
e-mail: tx8@illinois.edu; rqzhu@illinois.edu; xshao@illinois.edu

Abstract: Infinite-order U-statistics (IOUS) have been used extensively in subbagging ensemble learning algorithms such as random forests to quantify its uncertainty. While normality results of IOUS have been studied extensively, its variance estimation and theoretical properties remain mostly unexplored. Existing approaches mainly utilize the leading term dominance property in the Hoeffding decomposition. However, such a view usually leads to biased estimation when the kernel size is large relative to sample size. On the other hand, while several unbiased estimators exist in the literature, their relationships and theoretical properties, (e.g., ratio consistency), have never been studied. These limitations lead to unguaranteed asymptotic coverage of constructed confidence intervals. To bridge these gaps in the literature, we propose a new view of the Hoeffding decomposition for variance estimation that leads to an unbiased estimator. Instead of leading term dominance, our view utilizes the dominance of the peak region. Moreover, we establish the connection and equivalence of our estimator with several existing unbiased variance estimators. Theoretically, we are the first to establish the ratio consistency of such a variance estimator, which justifies the coverage rate of confidence intervals constructed from random forests. Numerically, we further propose a local smoothing procedure to improve the estimator's finite sample performance. Extensive simulation studies show that our estimators enjoy lower bias and achieve targeted coverage rates.

MSC2020 subject classifications: Primary 62G99; secondary 62G05.
Keywords and phrases: Infinite-order U-statistics, random forests, ensemble learning, variance estimation, Hoeffding decomposition, ratio consistency.

Received April 2023.

Contents

1. Introduction

Given a set of n i.i.d. observations $\mathcal{D}_n = \{X_i\}_{i=1}^n$ and an unbiased estimator, $h(X_1, \ldots, X_k)$, of the parameter of interest θ with $k \leqslant n$, the U-statistic [14] defined in the following is a minimum-variance unbiased estimator of θ:

$$U_n = \binom{n}{k}^{-1} \sum_{S_i \subset \mathcal{D}_n} h(S_i) = \binom{n}{k}^{-1} \sum_{1 \leqslant j_1 < \cdots < j_k \leqslant n} h\left(X_{j_1}, \ldots, X_{j_k}\right), \tag{1.1}$$

where each S_i is a subset of k samples from the original $\mathcal{D}_n$, k is called the kernel size and h is a symmetric kernel function. When k grows with n, U_n becomes an Infinite-Order U-statistic (IOUS) [11]. U-statistics are used extensively in problems such as non-parametric testing [17], empirical risk minimization in large-scale machine learning [21, 6], distributed computing and inference for big-data [19, 18, 5] and many others. In recent years, there has been an increasing interest in statistical inference with IOUS, particularly with their application on ensemble approaches, such as random forests [3, 13]. Random forest models for survival analysis [16], estimating heterogeneous treatment effect [28] can all benefit from such developments.

It is easy to see that large $\binom{n}{k}$ renders the computational challenge to exhaust all subsamples. Instead, random forests sample B subsamples from $\mathcal{D}_n$ to build trees and average. This leads to incomplete U-statistics [17]. Further incorporating randomness in the kernel function h, Mentch and Hooker [20] first show the asymptotic normality of random forests under the U-statistics framework when k grows at the rate of $o(\sqrt{n})$. DiCiccio and Romano [8] further relax its assumptions. Zhou, Mentch and Hooker [35] set the connection between U- and V-statistics. Peng, Coleman and Mentch [22] extend the kernel size to $k = o(n)$ under a generalized U-statistic framework. We also note that there is a large literature outside the applications of random forests. For example, for incomplete high-dimensional U-statistics, where $h \in \mathbb{R}^d$, Chen and Kato [4] and Song, Chen and Kato [27] study the asymptotic normality for fixed and growing k, respectively.

With the normality of random forest estimators established under the U-statistics [20] or other frameworks [28, 1], another line of the topic is the variance estimation. Wager, Hastie and Efron [29] propose to use jackknife and infinitesimal jackknife (IJ [9]). Mentch and Hooker [20] use Monte Carlo methods to estimate the leading term in the Hoeffding decomposition of $\mathrm{Var}(U_n)$. Recent developments include Zhou, Mentch and Hooker [35], who propose a computationally efficient approach and set the connection with the IJ estimator. Peng, Mentch and Stefanski [23] further study the bias and consistency of the IJ estimator.

However, an essential practical issue is that these estimators can display a significant amount of bias when the sample size n is small or k is large compared to n. In practice, it is common to use a fixed proportion of the total sample size [13] as the kernel size k. Variance estimators in the aforementioned literature often suffer from this bias issue because they all rely on some form of leading term dominance phenomenon. However, when k is large compared to n, such dominance is weak. Searching through the literature, several unbiased estimators have been proposed in different forms and motivated from different perspectives based on the U-statistics view. Some of them can handle a subsampling size k as large as $n/2$. Folsom [10] propose a variance estimator of complete U-statistics following a sequence of literature on sampling design [15, 34, 25]. Schucany and Bankson [24] propose to estimate all terms in the Hoeffding decomposition [14] of the variance of an order-2 complete U-statistic. However, they do not extend the estimator to a general case with $k \leqslant n/2$. Note that Folsom [10], Schucany and Bankson [24] do not consider the incomplete case; hence their estimators are computationally infeasible for large k or large n. More recently, Wang and Lindsay [31] propose partition-based, unbiased variance estimators of both complete and incomplete U-statistics motivated from the second-moment expression $E(U_n^2) - E^2(U_n)$. Wang and Wei [33] further apply this estimator to random forest variance estimation. However, there is a lack of theoretical justification for these estimators in terms of their ratio consistency, which is crucial for achieving a proper coverage rate based on the derived confidence interval. Moreover, there is a lack of understanding of their connections and differences with the estimators mentioned previously.

To address these limitations in the literature, the major contribution of our paper is three-fold. First, we re-analyze the Hoeffding decomposition and propose a peak region dominance view of the variance estimation of U-statistics to address the bias issue. This leads to a class of unbiased estimation approaches for both complete and incomplete U-statistics, called *Matched Sample Variance Estimator*, which can handle a subsampling size k as large as $n/2$. Computationally, our incomplete variance estimator is efficient and can be directly applied to random forests. Besides, we discuss two extensions of our estimators. One is a local smoothing strategy to mitigate negative variance estimation [24, 31], and the other extends our method to $k > n/2$. Secondly, we are the first to establish the connection and equivalence of the three existing estimators [10, 24, 31]. We show that our proposed estimator coincides with each under specific settings (see Section 3.5 for a detailed discussion). Thirdly, we establish the ratio consistency for our complete variance estimator under $k = o(\sqrt{n})$. To the best of our knowledge, this is the first result for such estimators, even for fixed k. This is a crucial step to achieve the nominal coverage level when we plug in the variance estimator in constructing a confidence interval. To this end, we fill a significant gap in the literature by proposing a set of interpretable conditions.

We proceed with additional notation and preliminaries of U-statistics to motivate the proposed variance estimator and establish the peak region dominance view.

2. Variance of U-statistics

Our analysis starts with a classical result of the variance of U-statistics. We first review the Hoeffding decomposition of the variance of a complete U-statistic. Then, we present the connection between the complete and incomplete versions. In particular, the variance of an order-k complete U-statistic is given by Hoeffding [14]:

$$\operatorname{Var}(U_n) = \binom{n}{k}^{-1} \sum_{d=1}^{k} \binom{k}{d} \binom{n-k}{k-d} \xi_{d,k}^2, \tag{2.1}$$

where $\xi_{d,k}^2$ is the covariance between two kernels $h(S_1)$ and $h(S_2)$ with S_1 and S_2 sharing d overlapping observations, i.e., $\xi_{d,k}^2 = \operatorname{Cov}(h(S_1), h(S_2))$, with $|S_1 \cap S_2| = d$. Here both S_1 and S_2 are size-k subsamples. Alternatively, we can represent $\xi_{d,k}^2$ as [17]

$$\xi_{d,k}^2 = \operatorname{Var}\left[\operatorname{E}\left(h(S)|X_1, ..., X_d\right)\right]. \tag{2.2}$$

This form will be utilized later.

When k grows with n, it is computationally almost infeasible to exhaust all subsamples due to large $\binom{n}{k}$. Instead, it is typical in random forests and other ensemble algorithms to build incomplete infinite-order U-statistics [17] by sampling B many S_i's, which gives

$$U_{n,B} = \frac{1}{B} \sum_{i=1}^{B} h(S_i). \tag{2.3}$$

The gap between variances of an incomplete U-statistic and its complete counterpart can be understood as

$$\begin{aligned}\operatorname{Var}(U_{n,B}) &= \operatorname{Var}\left[\operatorname{E}(U_{n,B}|\mathcal{X}_n)\right] + \operatorname{E}\left[\operatorname{Var}(U_{n,B}|\mathcal{X}_n)\right] \\ &= \operatorname{Var}(U_n) + \operatorname{E}\left[\operatorname{Var}(U_{n,B}|\mathcal{X}_n)\right],\end{aligned} \tag{2.4}$$

where $\mathcal{X}_n = (X_1, ..., X_n)$ and the additional term $\operatorname{E}[\operatorname{Var}(U_{n,B}|\mathcal{X}_n)]$ depends on the subsampling scheme. In particular, when all subsamples are drawn independently from the collection of all such subsamples [17], we have

$$\operatorname{Var}(U_{n,B}) = (1 - \frac{1}{B})\operatorname{Var}(U_n) + \frac{1}{B}\xi_{k,k}^2. \tag{2.5}$$

This suggests that we can close the gap by using a large B. Hence, we will first discuss the complete U-statistics setting and then propose the incomplete one. We also note that for applications to random forests, random kernels (trees) are involved. However, the difference can be negligible when using a large B [20].

3. Methodology

The main technical challenge for estimating the variance is when k is relatively large compared with n. Besides the aforementioned obvious computational issue in the complete version, most existing methods will also encounter a significant bias due to only estimating the leading term in the Hoeffding decomposition. By establishing a peak region dominance view, we develop a new unbiased estimator for $\mathrm{Var}(U_n)$ in both complete and incomplete forms whenever $k \leqslant \frac{n}{2}$. Its connection with existing methods will be discussed in Section 3.5. Its extension to $n/2 < k < n$ setting will be presented in Section 5.1. We demonstrate the application to random forests in Section 5, where we also introduce a locally smoothed version for better numerical performances.

3.1. Existing methods and limitations

Continuing from the decomposition of $\mathrm{Var}(U_n)$ in Equation (2.1), we define $\gamma_{d,k,n} = \binom{n}{k}^{-1}\binom{k}{d}\binom{n-k}{k-d}$ for convenience. Then $\mathrm{Var}(U_n) = \sum_{d=1}^{k} \gamma_{d,k,n}\xi_{d,k}^2$. It is easy to see that $\gamma_{d,k,n}$ corresponds to the probability mass function of a hypergeometric distribution with parameters n, k and d. A graphical demonstration of such coefficients under different k and d settings, with $n = 100$, is provided in Figure 1. Many existing methods [20, 8] rely on the asymptotic approximation of $\mathrm{Var}(U_n)$ when k is small, e.g., $k = o(n^{1/2})$. Under such settings, the first coefficient $\gamma_{1,k,n} = [1 + o(1)]\frac{k^2}{n}$ dominates all remaining ones, as we can see in Figure 1 when $k = 10$. In this case, to estimate $\mathrm{Var}(U_n)$, it suffices to estimate the leading covariance term $\xi_{1,k}^2$ if $\xi_{k,k}^2/(k\xi_{1,k}^2)$ is bounded.

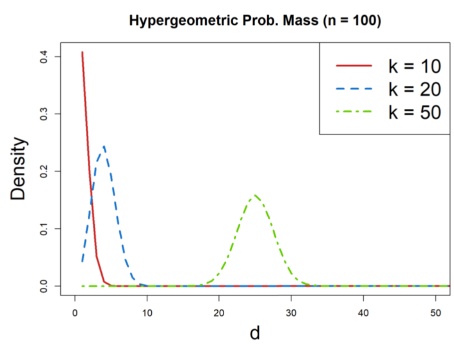

Fig 1. *Probability mass function of hypergeometric distribution with $n = 100$ for different k.*

However, as k becomes larger, the density of the hypergeometric distribution concentrates around $d = \beta^2 n$ instead of $d = 1$, where β denotes ratio k/n. Hence, the variance will be mainly determined by terms in a range of large d values, which we refer to as the peak region. In comparison, estimating just $\xi_{1,k}^2$ will introduce a significant bias even if we are able to exhaust all possible subsamples. Also note that when considering incomplete U statistics, most of the overlaping counts would fall into this region.

Another source of bias for using the leading term dominance property is the lack of samples to estimate $\xi_{1,k}^2$ realistically. Note that the definition involves approximating the Var and E operations in Equation (2.2) [20, 35]. A natural strategy is to hold one shared sample, e.g., $X_{(1)}$, and vary the remaining samples in S among existing observations $\mathcal{D}_n$ to approximate $\mathrm{E}\left[h(S)|X_1\right]$. However, this causes trouble for the variance estimator since we won't have enough samples to independently produce estimators of $\mathrm{E}\left[h(S)|X_i\right]$ with varying X_i when k becomes slightly larger. Overall, a new strategy is needed to better utilize the Hoeffding decomposition.

We also note that another theoretical strategy proposed by Wager and Athey [28], Peng, Coleman and Mentch [22] can be used for $k = o(n)$ if the U-statistic can be understood through the Hajek projection with additional regularity conditions. In this case, the variance of a U-statistic can be well approximated by the variance of a linearised version, while the infinitesimal jackknife procedure provides a valid estimator. However, it is difficult to assess whether the kernel function satisfies these assumptions. In practice, a significant bias can still occur, as seen in the simulation section.

3.2. An alternative view

At this point, estimating $\xi_{d,k}^2$'s for some d values seems inevitable. However, we may utilize the law of total variance to change the estimation procedure, which could gain a significant computational advantage. Note that for any given d,

$$\begin{aligned}\xi_{d,k}^2 =& \mathrm{Var}\left[\mathrm{E}\left(h(S)|X_1,...,X_d\right)\right]\\ =&\mathrm{Var}(h(S)) - \mathrm{E}\left[\mathrm{Var}(h(S)|X_1,...,X_d)\right]\\ :=&V^{(h)} - \tilde{\xi}_{d,k}^2,\end{aligned} \tag{3.1}$$

where we define $\tilde{\xi}_{d,k}^2 := \mathrm{E}[\mathrm{Var}(h(S)|X_1,...,X_d)]$. In this representation, $V^{(h)}$ is equivalent to $\xi_{k,k}^2$, the variance of a single kernel. It is also equivalent to $\tilde{\xi}_{0,k}^2$ since $\xi_{0,k}^2 = 0$. Incorporating these into the decomposition formula in Equation (2.1), we obtain an interesting connection:

$$\begin{aligned}\mathrm{Var}(U_n) =& \sum_{d=1}^{k} \gamma_{d,k,n}\left(V^{(h)} - \tilde{\xi}_{d,k}^2\right)\\ =& \sum_{d=0}^{k} \gamma_{d,k,n} V^{(h)} - \sum_{d=0}^{k} \gamma_{d,k,n}\tilde{\xi}_{d,k}^2\\ :=&V^{(h)} - V^{(s)},\end{aligned} \tag{3.2}$$

where we define $V^{(s)}$ as $\sum_{d=0}^{k} \gamma_{d,k,n}\tilde{\xi}_{d,k}^2$.

While this alternative view is valid for all k, the difficulty lies in finding a computationally feasible estimator, especially when we have to deal with incomplete U-statistics, instead of the complete version. In particular, when $k \leqslant n/2$, both terms can be unbiasedly estimated with a proper sampling design. In the following, we first present a straightforward formula for estimating $V^{(h)}$ and $V^{(s)}$ in a complete U-statistics version. The main result is Theorem 3.1, which shows that $V^{(s)}$ can be estimated using sample variance of all trees. Section 3.4 extends these estimators to incomplete versions.

3.3. Variance estimation for complete U-statistics

Our goal is to create estimators of $V^{(s)}$ and $V^{(h)}$ such that they can be directly computed from the trees (kernels) fitted in the random forest itself. This seems to be a challenging task given that we are estimating an infinite sum $V^{(s)}$. However, the fundamental idea we will utilize is to estimate $\mathrm{Var}[h(S)|X_1, ..., X_d]$ using pairs of trees. We proceed with the complete case when all trees are already available.

3.3.1. Joint estimation of the infinite sum $V^{(s)}$

Suppose we pair subsamples S_i and S_j among $\binom{n}{k}$ subsamples and let $d = |S_i \cap S_j| = 0, 1, ..., k$. Then for each d, there exist $N_{d,k,n} = \binom{n}{k}^2 \gamma_{d,k,n} = \binom{n}{k}\binom{k}{d}\binom{n-k}{k-d}$ pairs of subsamples S_i, S_j such that $|S_i \cap S_j| = d$. Note that for any such pair, $(h(S_i) - h(S_j))^2/2$ is an unbiased estimator of $\tilde{\xi}^2_{d,k} = \mathrm{Var}[h(S)|X_1, ..., X_d]$. We may then construct an unbiased estimator of $\tilde{\xi}^2_{d,k}$ by averaging them:

$$\hat{\tilde{\xi}}^2_{d,k} = \frac{1}{\binom{n}{k}\binom{k}{d}\binom{n-k}{k-d}} \sum_{|S_i \cap S_j| = d} \left[h(S_i) - h(S_j)\right]^2 /2. \tag{3.3}$$

This motivates us to combine all such terms in the infinite sum, which surprisingly leads to the sample variance of all kernels. The result is given in the following proposition, with its proof collected in Appendix D.

Proposition 3.1. *Given a complete U-statistic U_n, and the estimator $\hat{\tilde{\xi}}^2_{d,k}$ defined in Equation (3.3), when $k \leqslant n/2$, we have the following unbiased estimator of $V^{(s)}$:*

$$\hat{V}^{(s)} := \binom{n}{k}^{-1} \sum_i \left(h(S_i) - U_n\right)^2 = \sum_{d=0}^{k} \gamma_{d,k,n} \hat{\tilde{\xi}}^2_{d,k}. \tag{3.4}$$

Furthermore, when $k > n/2$, the first $2k - n$ terms in the summation $\sum_{d=0}^{k}$ is removed, since corresponding $\gamma_{d,k,n}$ terms are zero.

Since $\hat{V}^{(s)}$ enjoys a sample variance form, its incomplete version would also be easy to calculate. The advantage is that it can be computed without any hassle because all $h(S_i)$'s are ready to use when we calculate U_n. However, additional consideration may facilitate the estimation of $V^{(h)}$ so that both $V^{(s)}$ and $V^{(h)}$ can be done using the same set of $h(S_i)$'s.

3.3.2. Estimation of kernel variance $V^{(h)}$

Estimating $V^{(h)}$ may follow the same idea using $\left[h(S_i) - h(S_j)\right]^2/2$ if the pair S_i and S_j are disjoint. However, this is only possible when $k \leqslant n/2$, given a finite sample. In this case, following Equation (3.3), we have an unbiased estimator of $V^{(h)}$:

$$\hat{V}^{(h)} = \hat{\xi}^2_{0,k} = \frac{1}{\binom{n}{k}\binom{n-k}{k}} \sum_{|S_i \cap S_j|=0} \left[h(S_i) - h(S_j)\right]^2/2. \tag{3.5}$$

Therefore, we combine estimators $\hat{V}^{(s)}$ (3.4) and $\hat{V}^{(h)}$ (3.5) to get an unbiased estimator of $\mathrm{Var}(U_n)$:

$$\widehat{\mathrm{Var}}(U_n) = \hat{V}^{(h)} - \hat{V}^{(s)}. \tag{3.6}$$

3.4. Variance estimation for incomplete U-statistics

In random forests and other ensemble learning models, we often construct incomplete U-statistics by drawing random subsamples instead of exhausting all $\binom{n}{k}$ subsamples. This creates difficulties in calculating $\hat{V}^{(h)}$ since very few of these subsamples would be mutually exclusive ($d = 0$). Hence, a new subsampling strategy is needed to allow sufficient pairs of subsamples to estimate both $V^{(h)}$ and $V^{(s)}$.

The following "matched sample" sampling scheme is proposed to have enough disjoint samples to estimate $\hat{V}^{(h)}$. For any $2 \leqslant M \leqslant \lfloor n/k \rfloor$, we can sample a set of the matched sample group that consists M mutually exclusive subsamples $\{S_1, \dots, S_M\}$ from $\mathcal{D}_n$. This enables us to estimate $V^{(h)}$ by the sample variance of $\{h(S_1), \dots, h(S_M)\}$. Then, we repeat this procedure B times to average the estimator. To be precise, denote the subsamples in the b-th matched sample group as $S_1^{(b)}, S_2^{(b)}, ..., S_M^{(b)}$, such that $S_i^{(b)} \cap S_{i'}^{(b)} = \varnothing$ for any $i \neq i'$. Define

$$U_{n,B,M} = \frac{1}{MB} \sum_{i=1}^{M} \sum_{b=1}^{B} h(S_i^{(b)}). \tag{3.7}$$

This differs from the conventional incomplete U-statistic due to the new sampling scheme. Though $M = 2$ is enough for estimating $V^{(h)}$, we recommend using $M = \lfloor n/k \rfloor$ for a smaller variance. This is guaranteed by the following proposition.

Proposition 3.2. *For an incomplete U-statistic with $M \cdot B$ samples obtained using the matched sample sampling scheme,*

$$\mathrm{Var}(U_{n,B,M}) = \left(1 - \frac{1}{B}\right) Var(U_n) + \frac{1}{MB} V^{(h)}. \tag{3.8}$$

The proof is collected in Appendix D.1. We should note that when fixing the total number of kernels, $M \cdot B$ and let $M \geqslant 2$, the variance of $U_{n,B,M}$ is always smaller than the variance of $U_{n,B}$ given in (2.5). However, these two are identical when $M = 1$.

Based on this new sampling scheme, we can propose estimators $\hat{V}^{(h)}_{B,M}$ and $\hat{V}^{(s)}_{B,M}$ as analogs to $\hat{V}^{(h)}$ and $\hat{V}^{(s)}$, respectively. Denote the collection of kernels as $\{h(S_i^{(b)})\}_{i,b}$, for $i = 1, 2, \ldots, M$, $b = 1, 2, \ldots, B$. A sample variance within each group b, $\frac{1}{M-1}\sum_{i=1}^{M}[h(S_i^{(b)}) - \bar{h}^{(b)}]^2$, is an unbiased estimator of $V^{(h)}$. Here $\bar{h}^{(b)} = \frac{1}{M}\sum_{i=1}^{M} h(S_i^{(b)})$ is the group mean. Hence, the average over all groups becomes $\hat{V}^{(h)}_{B,M}$.

$$\hat{V}^{(h)}_{B,M} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{M-1}\sum_{i=1}^{M}\left[h(S_i^{(b)}) - \bar{h}^{(b)}\right]^2. \tag{3.9}$$

Similarly, with some algebra, we can define $\hat{V}^{(s)}_{B,M}$ as

$$\hat{V}^{(s)}_{B,M} = \frac{1}{MB-1}\sum_{b=1}^{B}\sum_{i=1}^{M}\left[h(S_i^{(b)}) - U_{n,B,M}\right]^2. \tag{3.10}$$

Note that $\hat{V}^{(h)}_{B,M}$ is still an unbiased estimator of $V^{(h)}$ while $\hat{V}^{(s)}_{B,M}$ introduces a small bias when estimating $V^{(s)}$ because these subsamples are not randomly obtained — there is an over-representation of non-overlapping pairs. The following proposition quantifies this bias.

Proposition 3.3. *For the sample variance estimator $\hat{V}^{(s)}_{B,M}$ defined on the matched sample groups subsamples with $M \cdot B \geqslant 2$, we denote $\delta_{M,B} := \frac{M-1}{MB-1}$. Then,*

$$\mathrm{E}\left(\hat{V}^{(s)}_{B,M}\right) = (1 - \delta_{M,B})V^{(s)} + \delta_{M,B}V^{(h)}. \tag{3.11}$$

This proposition leads to the following unbiased estimator of $\mathrm{Var}(U_{n,B,M})$. The proofs of both Propositions 3.3 and 3.4 are collected in Appendix D.

Proposition 3.4. *Given $M \cdot B$ subsamples from the matched sample sampling scheme, with $B \geqslant 1$ and $M \geqslant 2$, the "Matched Sample Variance Estimator" given below is an unbiased estimator of* $\mathrm{Var}(U_{n,B,M})$*:*

$$\widehat{\mathrm{Var}}(U_{n,B,M}) = \hat{V}^{(h)}_{B,M} - \frac{MB-1}{MB}\hat{V}^{(s)}_{B,M}. \tag{3.12}$$

3.5. Unifying existing unbiased estimators

To conclude this section, we discuss the relationships and differences between our view of the variance decomposition versus existing approaches. As noted in the introduction, various variance estimators appeared in the literature to correct the bias when the leading term does not dominate. Folsom [10] and Schucany and Bankson [24] primarily focus on unbiased estimators for complete U-statistics with a very small sample. In particular, Schucany and Bankson [24] propose two estimators of $\xi_{1,2}^2$ in the Hoeffding decomposition (denoted as $\hat{\zeta}_1^2$ and $\tilde{\zeta}_1^2$ in their page 418 and 422, respectively). Interestingly, the estimators of $\xi_{1,k}^2$ introduced by Mentch and Hooker [20] and Zhou, Mentch and Hooker [35] are efficient incomplete approximations of the former, $\hat{\zeta}_1^2 := \frac{1}{n-1}\sum_{i=1}[\hat{h}_1(X_i) - U_n]^2$, where $\hat{h}_1(X_i) = \binom{n-1}{k-1}^{-1}\sum_{S_j:X_i\in S_j} h(S_j)$. Meanwhile, our estimator $\hat{V}^{(h)} - \hat{\hat{\xi}}_{1,k}^2$ is equivalent to the latter, $\tilde{\zeta}_1^2$. A comprehensive derivation is provided in Appendix D.3.

Wang and Lindsay [31] propose an unbiased estimator motivated by $E(U_n^2) - E^2(U_n)$. Their complete variance estimator [31, page 1120] is,

$$N_k^{-1}\sum_{P_k} h(S_i)h(S_j) - N_0^{-1}\sum_{P_0} h(S_0)h(S_0),$$

where $P_c = \{(S_i, S_j) \text{ s.t. } |S_i \cap S_j| \leqslant c\}$, and N_c is the cardinality of P_c. Motivated by this formulation, they further propose an ANOVA form of the estimator and its corresponding incomplete version.

Although various unbiased estimators exist in the literature, they are all motivated by entirely different perspectives. The unique motivation of our estimator is its peak-region dominance phenomenon and the corresponding conditional variance view, which allows unbiased estimation. While we are not restricting the estimating of covariance terms with d values within a certain region, this is automatically done through resamplings. In the incomplete version, the overlaps are dominated by terms from the peak-region. This is in contrast with the traditional leading term dominance notation, which forces the estimator to concentrate on a single term. On the other hand, it is interesting that the connections among existing estimators have never been investigated. To complete our analysis, we further established several connections. In Appendix D.3 and D.4 we show that all existing unbiased complete estimators are essentially the same estimator, presented in different formats and settings. In particular, we show that Folsom [10]'s formula is identical to our complete version and also equivalent to Wang and Lindsay [31]'s version. We further restrict a setting with $k = 2$ for a direct comparison with the estimator proposed by Schucany and Bankson [24]. In Appendix D.4, we show the equivalence between our incomplete estimators and Wang and Lindsay [31]'s.

4. Theoretical results

To the best of our knowledge, ratio consistency of variance estimator in the context of infinite-order U-statistics has not been investigated. In this section, we attempt to fill some gaps in the literature by establishing the results of our proposed estimator, meaning that we want to show

$$\frac{\widehat{\mathrm{Var}}(U_n)}{\mathrm{E}[\widehat{\mathrm{Var}}(U_n)]} \xrightarrow{P} 1.$$

where $\xrightarrow{P}$ denotes convergence in probability. The notion of ratio consistency is important here since the variance of U-statistics would naturally converge to 0 as n grows. Hence any variance estimator that converges to 0 is consistent. However, a consistent estimator does not guarantee nominal coverage. In the following, we shall rewrite $\widehat{\mathrm{Var}}(U_n)$ as $\hat{V}_u$, and show a sufficient condition of the above

$$\mathrm{Var}(\hat{V}_u)/\mathrm{E}^2(\hat{V}_u) = \mathrm{Var}(\hat{V}_u)/\mathrm{Var}^2(U_n) \to 0, \text{ as } n \to \infty.$$

We want to note that such a result under general k settings is likely impossible without strong assumptions or knowledge of the specific form of the kernel h. The main difficulty in the proof is caused by the fourth-order term in the form of $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$ which naturally appears in the variance of $\hat{V}_u$. Untangling the dependencies of the fourth-order term under large k is a difficult task. Hence, we focus on the $k = o(\sqrt{n})$ setting in which the result is more attainable, although computationally, the estimator can still be applied whenever $k \leqslant n/2$. Even though the $k = o(\sqrt{n})$ setting is somewhat restrictive, it is still the first in the literature under the context of this paper. And further investigations may be established by extending the proposed strategy to higher orders.

Our main strategy can be summarized as follows. First, we observe that the proposed estimator $\hat{V}_u = \hat{V}^{(h)} - \hat{V}^{(s)}$ can be written an order-$2k$ U-statistic:

$$\hat{V}_u = \binom{n}{2k}^{-1} \sum_{S^{(2k)} \subseteq \mathcal{X}_n} \psi\left(S^{(2k)}\right), \tag{4.1}$$

where $S^{(2k)}$ is a size-$2k$ subsample set and $\psi\left(S^{(2k)}\right)$ is the corresponding size-$2k$ kernel, defined as

$$\psi\left(S^{(2k)}\right) := \psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right).$$

Here $\psi_{k'}\left(S^{(2k)}\right)$ for $k' = 0, 1, 2, \ldots, k$ satisfies

$$\psi_{k'}\left(S^{(2k)}\right) = N_{n,k,k'} \sum_{d=0}^{k'} \frac{1}{N_d} \sum_{\substack{S_1, S_2 \subset S^{(2k)} \\ |S_1 \cap S_2| = d}} h\left(S_1\right) h\left(S_2\right), \tag{4.2}$$

where $N_{n,k,k'} = \binom{n}{2k}\binom{n}{k}^{-1}\binom{n-k+k'}{k}^{-1}$ and $N_d = \binom{n-2k+d}{d}$ is the number of different size-$2k$ sets such that its two size-k subsets S_1 and S_2 share d overlaps. We remark that in this paper, S refers to a size-k set, and $S^{(2k)}$ refers to a size-$2k$ set.

Similar to a regular U-statistic, the variance of an order-$2k$ U-statistic $\hat{V}_u$ can be decomposed as

$$\operatorname{Var}\left(\hat{V}_u\right) = \binom{n}{2k}^{-1} \sum_{c=1}^{2k} \binom{2k}{c}\binom{n-2k}{2k-c}\sigma^2_{c,2k}, \tag{4.3}$$

where $\sigma^2_{c,2k}$ is the covariance between $\psi(S_1^{(2k)})$ and $\psi(S_2^{(2k)})$ for $|S_1^{(2k)} \cap S_2^{(2k)}| = c$ and $c = 1, 2, \dots, 2k$:

$$\sigma^2_{c,2k} := \operatorname{Cov}\left[\psi(S_1^{(2k)}), \psi(S_2^{(2k)})\right]. \tag{4.4}$$

If we follow the existing literature, it is common to impose high-level assumptions on the kernel ψ and also bound the ratio of the last term, $\sigma^2_{2k,2k}$ over the first term $\sigma^2_{1,2k}$ [8]. However, not only such assumptions are difficult to verify and can be possibly violated (see discussion in Appendix B.4), but also ψ is viewed as some form of a "black box", which does not help in analyzing the convergence of $\operatorname{Var}(\hat{V}_u)/\operatorname{Var}^2(U_n)$.

Hence, the key strategy of our approach is to avoid explicit assumptions on $\hat{V}_u$'s kernel ψ and $\sigma^2_{c,2k}$, instead only impose assumptions on a fourth-order term of h: $\operatorname{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$. This leads to the main technical challenge in this work, $\operatorname{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$ involves the 4-way overlaps among S_1, S_2, S_3, S_4, although it shares similar intuition as $\xi^2_{d,k} = \operatorname{Cov}[h(S_1), h(S_2)]$ which involves the overlaps between S_1, S_2. However, since $\operatorname{Var}(\hat{V}_u)$ is the variance of variance estimator of U-statistics, it becomes inevitable to study the fourth-order term of h instead of a second-order term.

We first establish the Double U-statistics notion of $\hat{V}_u$ in Section 4.1. The double U-statistic structure in Proposition 4.2 shows a cancellation effect (see Appendix E) inside of $\hat{V}_u$, which helps accelerate the convergence rate of $\operatorname{Var}(\hat{V}_u)$. Using this structure, we can further decompose each $\sigma^2_{c,2k}$ in the Hoeffding decomposition (4.3) into $\eta^2_{c,2k}(d_1, d_2)$ terms (see Proposition 4.3). Then, we bound all $\eta^2_{c,2k}(d_1, d_2)$'s by decomposing each term into a basic covariance term $\operatorname{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$. Hence, it suffices to impose primitive assumptions on $\operatorname{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$ to analyze the behavior of $\hat{V}_u$. We should highlight the challenge that we need to use 11 parameters to describe the 4-way overlapping among S_1, S_2, S_3, S_4. Details are left in the discussion in the assumption section (Section 4.2). We also note that it is easier to understand the difficulties and strategies related to the nature of *Double U-statistic* structure through a simplified example, the linear average kernel, presented in Appendix H. And finally, in Section 4.3, we present the ratio consistency. Section 4.4 is used to summarize a roadmap of the proof.

4.1. Double U-statistic structure

We define a notion of *Double U-statistic* to facilitate our discussion and show that $\hat{V}_u$ is a *Double U-statistic*. The advantage of this tool is to break down our variance estimator into lower-order terms, which alleviates the difficulty involved in analyzing $\sigma^2_{c,2k}$.

Definition 4.1 (Double U-statistic)**.** For an order-k U-statistic, we call it *Double U-statistic* if its kernel function h is a weighted average of U-statistics.

Essentially, a *Double U-statistic* is a "U-statistic of U-statistic". By (4.1), $\hat{V}_u = \binom{n}{2k}^{-1} \sum_{S^{(2k)} \subseteq \mathcal{X}_n} \psi\left(S^{(2k)}\right)$. $\hat{V}_u$ involves a size-2k kernel ψ. However, by Equation (4.2), the kernel ψ has a complicated form. The following proposition shows that we can further decompose ψ into linear combinations of φ_d's, which are still U-statistics.

Proposition 4.2 ($\hat{V}_u$ is a Double U-statistic)**.** *The order-*$2k$ *U-statistic* $\hat{V}_u$ *defined in Equation* (4.1) *is a* Double U-statistic. *Its kernel* $\psi\left(S^{(2k)}\right)$ *can be represented as a weighted average of U-statistics, such that*

$$\psi\left(S^{(2k)}\right) := \sum_{d=1}^{k} w_d \left[\varphi_d\left(S^{(2k)}\right) - \varphi_0\left(S^{(2k)}\right)\right]. \tag{4.5}$$

Here, for $d = 0, 1, ..., k$, φ_d *is the U-statistic with size-*$(2k-d)$ *asymmetric kernel as following*

$$\varphi_d\left(S^{(2k)}\right) = M_{d,k}^{-1} \sum_{\substack{S_1, S_2 \subset S^{(2k)} \\ |S_1 \cap S_2| = d}} h(S_1)h(S_2); \tag{4.6}$$

$M_{d,k} := \binom{2k}{d}\binom{2k-d}{d}\binom{2k-2d}{k-d}$, *which is the number of pairs* $S_1, S_2 \subset S^{(2k)}$, *s.t.* $|S_1 \cap S_2| = d$; *and* $w_d := \binom{n}{2k}\binom{n}{k}^{-2}\binom{2k}{d}\binom{2k-d}{d}\binom{2k-2d}{k-d}/\binom{n-2k+d}{d}$, $\forall d \geqslant 1$, $w_0 = \left(\binom{n}{k}^{-1} - \binom{n-k}{k}^{-1}\right)\binom{n}{k}^{-1}\binom{n}{2k}\binom{2k}{k}$. *The* w_d*'s defined above satisfy the following.* $\sum_{d=0}^{k} w_d = 0. w_d > 0$, $\forall d > 0$.

$$w_d = \mathcal{O}\left(\frac{k^{2d}}{d!\, n^d}\right), \text{ for } d = 1, 2, ..., k. \tag{4.7}$$

Particularly, for fixed d*,*

$$w_d = (1 + o(1))\, \frac{k^{2d}}{d!\, n^d}. \tag{4.8}$$

The proof is collected in Appendix C.1. We observe that given $k = o(\sqrt{n})$, w_d decays with d at a speed even faster than the geometric series. In our later analysis, we can show that the first term, $w_1[\varphi_1(S^{(2k)}) - \varphi_0 S^{(2k)})]$, can be a dominating term in $\psi(S^{(2k)})$. Moreover, with kernel $\varphi_d(S^{(2k)})$, we introduce the following decomposition of $\sigma^2_{c,2k}$.

Proposition 4.3 (Decomposition of $\sigma^2_{c,2k}$)**.** *For any size-*$2k$ *subsample sets* $S_1^{(2k)}, S_2^{(2k)}$, *s.t.* $|S_1^{(2k)} \cap S_2^{(2k)}| = c$ *and* $1 \leqslant c \leqslant 2k, 1 \leqslant d_1, d_2 \leqslant k$, *we define*

$$\eta^2_{c,2k}(d_1, d_2) := \operatorname{Cov}\big[\varphi_{d_1}\left(S_1^{(2k)}\right) - \varphi_0\left(S_1^{(2k)}\right), \varphi_{d_2}\left(S_2^{(2k)}\right) - \varphi_0\left(S_2^{(2k)}\right)\big]. \tag{4.9}$$

Then, we can represent $\sigma^2_{c,2k}$ as a weighted sum of $\eta^2_{c,2k}(d_1, d_2)$'s.

$$\sigma^2_{c,2k} = \sum_{d_1=1}^{k}\sum_{d_2=1}^{k} w_{d_1} w_{d_2} \eta^2_{c,2k}(d_1, d_2). \tag{4.10}$$

This proposition can be directly concluded by combining the alternative form of U_n's kernel ψ in Equation (4.5) and the definition of $\eta^2_{c,2k}(d_1, d_2)$. With the help of the *Double U-statistic* structure, upper bounding $\sigma^2_{c,2k}$ can be boiled down to analyzing $\eta^2_{c,2k}(d_1, d_2)$. Detailed analysis of this connection is provided in Section 4.4 and Appendix E. Note that we can further decompose $\eta^2_{c,2k}(d_1, d_2)$ (see Appendix F.5), so $\sigma^2_{c,2k}$ can be viewed as a weighted sum of $\mathrm{Cov}[h(S_2)h(S_2), h(S_3)h(S_4)]$'s.

4.2. Assumptions

Assumption 1 limits the kernel size k as a lower-order of $\sqrt{n}$, while Assumption 2 controls the growth rate of $\xi^2_{d,k}$ with d. Assumption 3, 4, and 5 are related to $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$. As previously mentioned, $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$ can be viewed as an extension of $\xi^2_{d,k} = \mathrm{Cov}[h(S_1), h(S_2)]$, the classical covariance of two kernels. While $\xi^2_{d,k}$ only depends on one parameter, i.e., $d = |S_1 \cap S_2|$, 11 parameters are needed to fully determine $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$, since it involves a 4-way overlapping structure. This can be visualized in Figure 4 in Appendix. We denote the number of parameters as "Degree of Freedom (*DoF*)" of the covariance. Essentially, Assumptions 3, 4, and 5 are about reducing this *DoF* and controlling the growth of $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$ with overlapping samples.

In Appendix B, we provide further discussion and examples of our assumptions. In Appendix G, we propose a relaxation of Assumption 3 and present the proof of the main results under the new assumptions.

Assumption 1. There exist a constant $\epsilon \in (0, 1/2)$, so that the growth rate of kernel size k regarding sample size n is bounded as $k = \mathcal{O}(n^{1/2-\epsilon})$.

Assumption 2. $\forall k \in \mathbb{N}^+, \xi^2_{1,k} > 0$ and $\xi^2_{k,k} < \infty$. There exist a universal constant $a_1 \geqslant 1$ independent of k, satisfying that

$$\sup_{d=2,3,\ldots,k} \frac{\xi^2_{d,k}}{d^{a_1}\xi^2_{1,k}} = \mathcal{O}(1).$$

Note that a smaller a_1 in Assumption 2 implies a stronger assumption. It is well known that $k\xi^2_{d,k} \leqslant d\xi^2_{k,k}$ [17], the smallest possible value of a_1 is 1, which is used in the existing literature [20, 8, 35, 22]. Hence, if we force $a_1 = 1$ and only focus on the upper bound of $\mathrm{Var}(U_n)$, the growth rate of k in Assumption 1 can be relaxed to $o(n)$. However, this trade-off between Assumptions 1 and 2 cannot be applied to ratio consistency directly.

To motivate our other assumptions, we provide a brief discussion on the 4-way overlap of $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$. As we mentioned before, the goal is to avoid direct assumptions of $\psi(S^{(2k)})$ and its covariance $\sigma^2_{c,2k}$ and study the fourth-moment term $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$. To simplify the notation, we let

$$\rho := \mathrm{Cov}\left[h(S_1)h(S_2), h(S_3)h(S_4)\right]. \tag{4.11}$$

Then ρ involves 11 different overlap schemes,

2-set: $|S_1 \cap S_2|$, $|S_1 \cap S_3|$, $|S_1 \cap S_4|, |S_2 \cap S_3|, |S_2 \cap S_4|$, $|S_3 \cap S_4|$;
3-set: $|S_1 \cap S_2 \cap S_3|$, $|S_1 \cap S_2 \cap S_4|, |S_1 \cap S_3 \cap S_4|$, $|S_2 \cap S_3 \cap S_4|$;
4-set: $|S_1 \cap S_2 \cap S_3 \cap S_4|$.

Hence, 11 parameters are needed to describe ρ. We denote the number of these parameters as the "Degrees of Freedom" (DoF) of ρ. Furthermore, there are two types of these parameters: $d_1 = |S_1 \cap S_2|$ and $d_2 = |S_3 \cap S_4|$ describes the overlapping within $S_1^{(2k)}$ and $S_2^{(2k)}$ respectively; while other 9 overlapping sets are subsets of $S_1^{(2k)} \cap S_2^{(2k)}$, so they describe the overlapping between $S_1^{(2k)}$ and $S_2^{(2k)}$. We can describe these 9 overlapping sets by a 9-dimensional vector $\underline{r}$, whose definition is collected in Appendix B.1. Hence, the 11 DoF can be denoted by tuple $(\underline{r}, d_1, d_2)$.

However, it may not be necessary to know all $\underline{r}, d_1, d_2$ values to calculate this covariance ρ. For example, in the linear average kernel (Example B.2 in Appendix B.2), ρ only depends on $\underline{r}$. This may be expected for an estimator that is approximately linear. Hence, we propose the following assumption.

Assumption 3. ρ only depends on the 9 DoF vector $\underline{r}$. Hence, without the risk of ambiguity, we define a function $\rho(\underline{r})(\cdot)$ with

$$\rho(\underline{r}) = \rho. \tag{4.12}$$

The assumption implies that the within $S_1^{(2k)}$ or $S_2^{(2k)}$ overlapping counts have no impact on ρ. This simplifies a cancellation pattern when analyzing $\eta_{c,2k}(d_1, d_2)$ (4.9). A comprehensive discussion of this assumption can be found in Appendix B. We first demonstrate that this assumption is valid for the linear average kernel, as previously mentioned. Next, we provide an example to illustrate the challenges of reducing DoF below 9 by only considering two-way overlaps, indicating that further simplification of this assumption may require specific assumptions about the kernel functions. In addition, in Appendix G, we suggest a relaxation of this assumption and provide an alternative proof of the main results based on this relaxed assumption.

Assumption 4 (Ordinal Covariance). For all size-k subsets S_1, S_2, S_3, S_4 and S_1', S_2', S_3', S_4', let ρ and ρ' denote the corresponding covariance as defined in Equation 4.11 with DOFs $\underline{r}$ and $\underline{r}'$ (defined in Appendix B.1), respectively. Then, we have:

$$\rho \geqslant \rho', \quad \text{if } r_{ij} \geqslant r'_{ij}, \forall i,j = 0,1,2.$$

Moreover, given size-k sets S', S'', and $\underline{r}$ such that $|S' \cap S''| = c$ and $|\underline{r}| = c$, we have:

$$\rho \leqslant \mathrm{Cov}[h(S')^2, h(S'')^2] =: F_c^{(k)}. \tag{4.13}$$

Assumption 4 implies that more overlapping leads to larger ρ. This is a reasonable result to expect. For every $c = |S_1^{(2k)} \cap S_2^{(2k)}|$, it also provides an upper bound of ρ, where $F_c^{(k)}$ refers to ρ with "maximum possible overlaps" given c such that $S_1 = S_2, S_3 = S_4$. The overlapping associated with $F^{(k)}$ is visualized in Figure 5 in Appendix B. It's easy to see that $\rho \geqslant 0$, an analog to $\xi_{d,k}^2 \geqslant 0$ in a regular U-statistics setting [17].

Assumption 5. For $F_c^{(k)}$ defined in Assumption 4, when $c = |S_1 \cap S_2| = 1$, we have

$$\frac{F_1^{(k)}}{\xi_{1,k}^4} = \frac{\mathrm{Cov}[h(S_1)^2, h(S_2)^2]}{\left(\mathrm{Cov}[h(S_1), h(S_2)]\right)^2} = \mathcal{O}(1) \tag{4.14}$$

In addition, there exist a universal constant $a_2 \geqslant 1$ independent of k, satisfying

$$\sup_{c=2,3,\ldots,2k} \frac{F_c^{(k)}}{c^{a_2}\, F_1^{(k)}} = \mathcal{O}(1). \tag{4.15}$$

Equation (4.14) states that a fourth-moment term cannot exceed a second-moment term $\xi_{1,k}^2$. This can be verified for the linear average kernel with basic moment conditions. Similarly to the polynomial growth rate of $\xi_{d,k}^2$ specified in Assumption 2, Equation (4.15) controls a polynomial growth rate of ρ with respect to c, as $F^{(k)}$ is an upper bound of ρ. It is worth noting that Assumption 5 can be implied by Assumption 2 for certain specific kernels (see Example B.4 in Appendix B).

4.3. Main results

We now present our main results. As a direct consequence of the following theorem, the ratio consistency property is provided in Corollary 4.6.

Theorem 4.4 (Asymptotic variance of U_n and $\hat{V}_u$)**.** *Under Assumptions 1-5, we can bound* $\mathrm{Var}(U_n)$ (2.1) *and* $\mathrm{Var}(\hat{V}_u)$ (4.3) *as*

$$\mathrm{Var}(U_n) = (1 + o(1))\, \frac{k^2 \xi_{1,k}^2}{n}, \tag{4.16}$$

$$\mathrm{Var}(\hat{V}_u) = \mathcal{O}\left(\frac{k^2 \check{\sigma}_{1,2k}^2}{n}\right), \tag{4.17}$$

where $\check{\sigma}_{1,2k}^2 \asymp \frac{k^2 \xi_{1,k}^4}{n^2}$ *is the upper bound of* $\sigma_{1,2k}^2$ *given by Proposition E.2 in Appendix E. Here, "*$f \asymp g$*" implies* $f = \mathcal{O}(g)$ *and* $g = \mathcal{O}(f)$*.*

The proof of the results is provided in Appendix C.3. The calculation of $\mathrm{Var}(U_n)$ in (4.16) and $\mathrm{Var}(\hat{V}_u)$ in (4.17) requires controlling the growth of $\xi^2_{d,k}$ and $\sigma^2_{c,2k}$. In particular, (4.16) can be derived from a general proposition (Proposition 4.5) provided below. However, the proof of (4.17) is more complex, as it relies on the *double U-statistic* structure of $\hat{V}_u$. A proof roadmap is presented in Section 4.4, and technical lemmas to upper bound $\eta^2_{c,2k}(d_1, d_2)$ (4.9) and $\sigma^2_{c,2k}$ are provided in Appendix E.

Proposition 4.5 (Leading covariance domination)**.** *For a complete U-statistic U_n with size-k kernel and $k = o(\sqrt{n})$, assume that $\xi^2_{1,k} > 0$ and there exists a non-negative constant C such that*

$$\limsup_{k\to\infty, 2\leqslant d\leqslant k} \xi^2_{d,k} / \left(d!\xi^2_{1,k}\right) = C.$$

Then,

$$\lim_{n\to\infty} \mathrm{Var}(U_n) / \left(k^2 \xi^2_{1,k}/n\right) = 1.$$

The proof of this proposition can be found in Appendix C.4. This proposition relaxes the conditions from Theorem 3.1 used by DiCiccio and Romano [8] and provides a foundation for our approach to bounding $\mathrm{Var}(\hat{V}_u)$. Specifically, our condition allows for the ratio $\xi^2_{d,k}/\xi^2_{1,k}$ to grow at a factorial rate of d, whereas the conditions in [20, 35, 8] only allow for linear growth. A comparison between our assumption on $\xi^2_{d,k}/\xi^2_{1,k}$ and existing literature is provided in Section 4.2.

Corollary 4.6 (Ratio consistency of $\hat{V}_u$)**.** *Under Assumptions 1-5,*

$$\frac{\mathrm{Var}(\hat{V}_u)}{\left[\mathrm{E}(\hat{V}_u)\right]^2} = \mathcal{O}\left(\frac{1}{n}\right),$$

which implies that $\hat{V}_u / \mathrm{E}(\hat{V}_u) \xrightarrow{P} 1$.

This result is a corollary of Theorem 4.4 and demonstrates the consistency of the variance estimator $\hat{V}_u$ in terms of ratios. The proof can be found in Appendix C.2. To the best of our knowledge, this is the first proof of the ratio consistency of an unbiased variance estimator for growing order U-statistics.

4.4. Proof roadmap

The roadmap to upper bound $\mathrm{Var}(\hat{V}_u)$ (4.17) in Theorem 4.4 is provided in Equation (4.18). The relevant technical lemmas are summarized in Appendix E.

$$\mathrm{Var}(\hat{V}_u) = \sum_{c=1}^{2k} v_c \sigma^2_{c,2k} \leqslant \sum_{c=1}^{2k} v_c \breve{\sigma}^2_{c,2k} \overset{(*)}{\asymp} \sum_{c=1}^{T_1} v_c \breve{\sigma}^2_{c,2k} \overset{(\dagger)}{\asymp} v_1 \breve{\sigma}^2_{1,2k} \overset{(\ddagger)}{\asymp} \frac{k^4}{n^3} F_1^{(k)}. \quad (4.18)$$

The quantity $v_c := \binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}$ represents the coefficients in the Hoeffding decomposition of $\mathrm{Var}(\hat{V}_u)$ (4.3); $\breve{\sigma}^2_{c,2k}$ is the upper bound of $\sigma^2_{c,2k}$ given by Propositions E.2 and E.3; and "f $\asymp$ g" means that $f = \mathcal{O}(g)$ and $g = \mathcal{O}(f)$. The inequalities in (4.18) should be interpreted as follows.

- The first inequality $\leqslant$ is a result of replacing $\sigma^2_{c,2k}$ with either its tighter bound for $c = 1, 2, ..., T_1$ (Proposition E.2) or its looser bound for $c = T_1 + 1, 2, ..., 2k$ (Proposition E.3). Each $\sigma^2_{c,2k}$ can be decomposed into $\eta^2_{c,2k}(d_1, d_2)$'s (4.10). Propositions E.2 and E.3 are based on the tighter and looser upper bounds of $\eta^2_{c,2k}(d_1, d_2)$ (Lemma E.5 and E.6).
- The first asymptotic notation $\asymp$ (denoted with $*$) is concluded from Lemma E.4. The value of $T_1 = \left\lfloor \frac{1}{\epsilon} \right\rfloor + 1$ only depends on the growth rate of k, not n, as we assume $k = o(n^{1/2-\epsilon})$ in Assumption 1.
- The second asymptotic notation $\asymp$ (denoted with $\dagger$) is a result of comparing the finite $\breve{\sigma}^2_{c,2k}$ terms for $c = 1, 2, ..., T_1$.
- The last asymptotic notation $\asymp$ (denoted with $\ddagger$) is concluded from Lemma E.2.

5. Application to random forests

Random forests can be viewed as an incomplete infinite-order U-statistic with a random kernel [20]. The purpose of this section is to present a comprehensive algorithm, as well as two extensions: one for the case when $k > n/2$ and another one that uses local smoothing to address the issue of negative estimation values.

Notation-wise, we present the algorithm in the context of regression, where we observe a vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for observations i. Hence, define $X_i = (\boldsymbol{x}_i, y_i)$, and the kernel function $h(S_i)$ can be viewed as the tree prediction on a given target point x^* with subsample S_i. The implementation of the variance estimator is straightforward using this setting and is summarized in Algorithm 1. We want to make a few comments. First, the original random forest [3] uses bootstrap samples, i.e., sampling with replacement, to build each tree. However, sampling without replacement [13] is also prevalent and achieves similar performances. Secondly, most random forest models utilize a random kernel instead of fixed ones. This is mainly due to the random feature selection [3] and random splitting point [13] when fitting each tree. Mentch and Hooker [20] show that U-statistics with random kernel converge in probability to its fixed kernel counterpart by viewing the fixed kernel version as the expectation of the random version. Under suitable conditions, given B large enough, the theoretical analysis of random U-statistic can be reasonably reduced to analyzing the non-random counterpart, allowing our method to be applied. It is possible that both our estimators of $V^{(h)}$ and $V^{(s)}$ are inflated by the influence of the randomness due to their U statistic representation. However, such inflations are likely canceled out by the difference, and our simulation results in Section 6 confirm this speculation by showing that the estimator is mostly unbiased.

Algorithm 1: Matched Sample Variance Estimator ($k \leqslant n/2$)

Input: n, k, M, B, training set $\mathcal{X}_n$, and testing sample x^*
Output: $\widehat{\text{Var}}(U_{n,B,M})$
Construct matched samples:
for $b = 1, 2, \ldots, B$ **do**
  Sequentially sample $\{S_1^{(b)}, S_2^{(b)}, ..., S_M^{(b)}\}$ from $\mathcal{X}_n$ such that $S_i^{(b)}$ s are mutually exclusive, i.e., $S_i^{(b)} \cap S_{i'}^{(b)} = \varnothing$ for $i \neq i'$.
end
Fit trees and obtain predictions:
  Fit random trees for each subsample $S_i^{(b)}$ and obtain prediction $h(S_i^{(b)})$ on the target point x^*.
Calculate the variance estimator components:
  Forest average: $U_{n,B,M} = \frac{1}{MB} \sum_{i=1}^{M} \sum_{b=1}^{B} h(S_i^{(b)})$
  Within-group average: $\bar{h}^{(b)} = \frac{1}{M} \sum_{i=1}^{M} h(S_i^{(b)})$
  Tree variance (3.9): $\hat{V}_{B,M}^{(h)} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{M-1} \sum_{i=1}^{M} (h(S_i^{(b)}) - \bar{h}^{(b)})^2$
  Tree sample variance (3.10): $\hat{V}_{B,M}^{(s)} = \frac{1}{MB-1} \sum_{i=1}^{M} \sum_{b=1}^{B} (h(S_i^{(b)}) - U_{n,B,M})^2$
The final variance estimator (3.12)
  $\widehat{\text{Var}}(U_{n,B,M}) = \hat{V}_{B,M}^{(h)} - (1 - \frac{1}{MB}) \hat{V}_{B,M}^{(s)}$

5.1. Extension to $k > n/2$

The previous estimator $\widehat{\text{Var}}(U_{n,B,M})$ (3.12) is restricted to $k \leqslant n/2$ due to the sampling scheme. However, this does not prevent the application of formulation (3.2), $\text{Var}(U_n) = V^{(h)} - V^{(s)}$. To the best of our knowledge, the existing literature does not provide further discussion under $k > n/2$ for general kernels, while some theoretical strategies such as Wang and Lindsay [32] simplify the kernel into a low-order approximation. Alternatively, the infinitesimal jackknife [28] has been shown to be almost equivalent to the leading term estimator in V-statistics by Zhou, Mentch and Hooker [35]. Here, we discuss a generalization of our formulation for $k > n/2$. Re-applying Propositions (3.2) and (3.2) with $M = 1$, we can obtain the variance of an incomplete U statistic sampled randomly with replacement:

$$\text{Var}(U_{n,B,M=1}) = V^{(h)} - \frac{B-1}{B} V^{(s)}.$$

By Proposition 3.3, $\hat{V}_{B,M=1}^{(S)}$ is still an unbiased estimator of $V^{(s)}$. However, $V^{(h)}$ has to be estimated with a different approach, since any pair of subsamples would share at least some overlapping samples. A simple strategy is to use bootstrapping. Hence, we generate another set of size-k samples, sampled with replacement, and evaluate the kernel, using their sample variance as an estimator of $V^{(h)}$. We remark that the bootstrap procedure introduced will introduce an additional computational burden.

5.2. Locally smoothed variance estimator for random forest

Even though the proposed estimator is unbiased, large variance of this estimator may still result in possible under-coverage of the corresponding confidence interval (CI). Note that due to its variation, our variance estimator might be negative, though this rarely happens in our simulations. A similar phenomenon is also noticed by Schucany and Bankson [24], and Wang and Lindsay [31]. To alleviate this issue, we propose a local smoothing estimator, namely Matched Sample Smoothing Variance Estimator (MS-s). The improvement is especially effective when the number of trees is small. This will be demonstrated in the simulation study, see, e.g., Table 1 and Figure 2.

Denote a variance estimator on a future test sample $\boldsymbol{x}^*$ as $\hat{\sigma}^2_{RF}(\boldsymbol{x}^*)$. We randomly generate N neighbor points $\boldsymbol{x}^*_1, \dots, \boldsymbol{x}^*_N$ and obtain their variance estimators $\hat{\sigma}^2_{RF}(\boldsymbol{x}^*_1), \dots, \hat{\sigma}^2_{RF}(\boldsymbol{x}^*_N)$. Then, the locally smoothed estimator is defined as the average:

$$\overline{\hat{\sigma}^2_{RF}}(\boldsymbol{x}^*) = \frac{1}{N+1}\Big[\hat{\sigma}^2_{RF}(\boldsymbol{x}^*) + \sum_{i=1}^{N} \hat{\sigma}^2_{RF}(\boldsymbol{x}^*_i)\Big]. \tag{5.1}$$

The algorithm is presented as follows.

Algorithm 2: Matched Sample Smoothing Variance Estimator ($k \leqslant n/2$)

Input: n, k, M, B, training set $\mathcal{X}_{train}$, testing sample $\boldsymbol{x}^*$ and number of neighbors N
Output: Smooth Variance estimator $\overline{\hat{\sigma}^2_{RF}}(\boldsymbol{x}^*)$

1. Find the closed distance $D_{min} = \min_{\boldsymbol{x}\in\mathcal{X}_{train}} d(\boldsymbol{x}^*, \boldsymbol{x})$;
2. Randomly generate N neighbors $\boldsymbol{x}^*_1, ..., \boldsymbol{x}^*_N$ that satisfy $\boldsymbol{x} : d(\boldsymbol{x}, \boldsymbol{x}^*) \leqslant D_{min}$ or $\boldsymbol{x} : d(\boldsymbol{x}, \boldsymbol{x}^*) = D_{min}$;
3. Obtain variance estimators $\hat{\sigma}^2_{RF}(\boldsymbol{x}^*), \hat{\sigma}^2_{RF}(\boldsymbol{x}^*_1), ..., \hat{\sigma}^2_{RF}(\boldsymbol{x}^*_N)$ by Algorithm 1 ;
4. $\overline{\hat{\sigma}^2_{RF}}(\boldsymbol{x}^*) = \frac{1}{N+1}[\hat{\sigma}^2_{RF}(\boldsymbol{x}^*) + \sum_{i=1}^{N} \hat{\sigma}^2_{RF}(\boldsymbol{x}^*_i)]$ (5.1).

In Algorithm 2, $d(\cdot, \cdot)$ can be Euclidean distance for continuous covariates and other metrics for categorical covariates. In practice, we can pre-process data before fitting random forest models, such as performing standardization and feature selection. Due to the averaging with local target samples, there is naturally a bias-variance trade-off in choosing D_{min} and neighbors. This is a rather classical topic, and there can be various ways to improve such an estimator based on the literature. Our goal here is to provide a simple illustration. In the simulation section, we consider generating 10 neighbors on an ℓ_2 ball centered at $\boldsymbol{x}^*$. The radius of the ball is set to be the Euclidean distance from $\boldsymbol{x}^*$ to the closest training sample. We found that the performance is not very sensitive to the choice of neighbor distance. Also, the computational cost of this smoothing estimator only involves new predictions, which is also minor compared to fitting random forests.

5.3. A discussion on existing normality theories of random forests

Before demonstrating the simulation results, we would like to discuss the normality theories of random forests briefly. The main concern is that there is no universal guarantee of normality for random forests, and a variance estimator may not ensure the desired coverage rate. Hence, the use of any variance estimators should be done with a reasonable understanding of the random forest itself, especially by considering the impact of its tuning parameters.

Many existing works in the literature have studied the asymptotic normality of U_n given $k = o(\sqrt{n})$ to $o(n)$ under various regularity conditions [20, 28, 8, 35, 22, 1]. Existing empirical study also shows that the normality usually holds when k is small while begins to break down for certain cases [35, Table 2]. As we will see in the following, there are both examples and counter-examples for the asymptotic normality of U_n with a large k, depending on the specific form of the kernel.

Essentially, when a kernel $h(\cdot)$ is very adaptive to local observations without much randomness, e.g., 1-nearest neighbors and the kernel size is at the same order of n, there is too much dependency across different $h(S_i)$'s. This prevents the normality of U_n. On the other hand, when the kernel size is relatively small, there is enough variation across different kernel functions to establish normality. This is the main strategy used in the literature for establishing normality. The following example demonstrates these ideas.

Example. Given covariate-response pairs: $Z_1 = (x_1, Y_1), ..., Z_n = (x_n, Y_n)$ as training samples, where x_i's are unique and deterministic numbers and Y_i's i.i.d. F such that $\mathrm{E}(Y_i) = \mu > 0, \mathrm{Var}(Y_i) = \sigma^2$, for $i = 1, 2, ..., n$. We want to predict the response for a given testing sample x^*.

Suppose we have two size-k ($k = \beta n$) kernels: 1) a simple (linear) average kernel: $h(S) = \frac{1}{k} \sum_{Z_j \in S} Y_j$; 2) a 1-nearest neighbor (1-NN) kernel, which predicts using the closest training sample of x^* based on the distance of x. Without loss of generality, we assume that x_i's are ordered such that x_i is the i-th nearest sample to x^*. We denote corresponding sub-bagging estimator as U_{mean} and $U_{\text{1-NN}}$ respectively. It is trivial to show that

$$U_{\text{mean}} = \frac{1}{n} \sum_{i=1}^{n} Y_n, \quad U_{\text{1-NN}} = \sum_{i=1}^{n-k+1} a_i Y_i,$$

where $a_i = \binom{n-i}{k-1} / \binom{n}{k}$ and $\sum_{i=1}^{n-k+1} a_i = 1$. Accordingly, we have $\mathrm{Var}(U_{\text{mean}}) = \frac{1}{n}\sigma^2$ and $\mathrm{Var}(U_{1-NN}) \geqslant a_1^2 \mathrm{Var}(Y_1) = \frac{k^2}{n^2}\sigma^2 = \beta^2\sigma^2$. Since U_{mean} is a sample average, we still obtain asymptotic normality after scaling by $\sqrt{n}$. However, $\beta = k/n > 0$, a_1 makes a significant proportion in the sum of all a_i's and $\mathrm{Var}(U_{1-NN})$ does not decay to 0 as n grows. Hence, asymptotic normality is not satisfied for $U_{\text{1-NN}}$.

In practice, it is difficult to know apriori what type of data dependence structure these $h(S_i)$'s may satisfy. Thus, the normality of a random forest with a large subsampling size is still an open question and requires further understanding of its kernel. In our simulation study, we observe that the confidence intervals constructed with normal quantiles work well, given that data are generated with Gaussian noise (see Section 6.1).

6. Simulation study

We present simulation studies to compare our variance estimator with existing methods [35, 28] on random forests. We consider both the smoothed and non-smoothed versions, denoted as "MS-s" and "MS", respectively. The balance estimator and its bias-corrected version proposed by Zhou, Mentch and Hooker [35] are denoted as "BM" and "BM-cor". The infinitesimal jackknife [28] is denoted as "IJ". Our simulation does not include the Internal Estimator and the External Estimator [20], since the BM method has been shown to be superior to these estimators [35]. Note that the BM estimator works for both U-statistics and V-statistics [35, Section4, paragraph 1]. However, the V-statistics version is almost equivalent to IJ [35, Theroem 3.3 and 3.4]. Hence, in our simulation, we only include the U-statistics version.

6.1. Simulation settings

We consider two regression settings:

1. MARS: $f(\boldsymbol{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.05)^2 + 10x_4 + 5x_5$
2. MLR: $f(\boldsymbol{x}) = 2x_1 + 3x_2 - 5x_3 - x_4 + 1$

The first setting, MARS (multivariate adaptive regression splines), is proposed by Friedman [12]. It has been used previously by Biau [2], Mentch and Hooker [20]. The second setting, MLR, refers to multivariate linear regression. In both settings, we generate a six-dimensional feature $\boldsymbol{x} = (x_1, \ldots, x_6)$ with independent entries uniformly from $[0, 1]$, and responses are generated by $f(\boldsymbol{x}) + \epsilon$, where $\epsilon \overset{iid}{\sim} \mathcal{N}(0, 1)$. Note that $f(\boldsymbol{x})$ only depends on a subset of 6 variables.

We use $n = 200$ as the total training sample size and pick different subsample sizes: $k = 100, 50, 25$ when $k \leqslant n/2$ and $k = 160$ when $k > n/2$. The numbers of trees are `nTrees` $= B \cdot M = 2000, 10000, 20000$. For tuning parameters, we set `mtry` as 3, which is half of the dimension, and set `nodesize` parameter to $2\lfloor \log(n) \rfloor = 8$. We repeat the simulation $N_{mc} = 1000$ times to evaluate the performance of different estimators. Our proposed methods (MS, MS-s), BM and BM-cor estimators are implemented using the `RLT` package available on GitHub. The IJ estimators are implemented using `grf` and `ranger`. Each estimation method and its corresponding ground truth (see details in the following) is generated by the same package. Note that we do not use the honest tree setting by Wager and Athey [28], since it is not essential for estimating the variance. However, it may affect the coverage rate due to the normality behavior.

The performance of the variance estimator is evaluated in terms of its bias and the coverage rate of its corresponding confidence interval. We denote the random forest estimator as $\widehat{f}(\boldsymbol{x})$ and evaluate the coverage based on the mean of the random forest estimator, $\mathrm{E}[\widehat{f}(\boldsymbol{x})]$, instead of the true model value, $f(\boldsymbol{x})$, as our focus is the variance estimation of $\widehat{f}(\boldsymbol{x})$ and the random forest itself may be a biased model. To obtain the ground truth of the variance, we generate the training dataset 10000 times and fit a random forest to each, using the mean and variance of the 10000 forest predictions as approximations of $\mathrm{E}[\widehat{f}(\boldsymbol{x})]$ and $\mathrm{Var}[\widehat{f}(\boldsymbol{x})]$. The relative bias and the confidence interval (CI) convergence are the evaluation criteria, with the relative bias defined as the ratio of the bias to the ground truth of the variance estimation. The $1-\alpha$ CI is constructed using $\hat{f} \pm Z_{\alpha/2}\sqrt{\hat{V}_u}$, where $Z_{\alpha/2}$ is the standard normal quantile.

We evaluate the variance estimation on two types of testing samples for both MARS and MLR data. The first is a central sample with $x^* = (0.5, \ldots, 0.5)$ and the second includes 50 random samples whose coordinates are independently sampled from a uniform distribution between $[0, 1]$. These testing samples are fixed for all experiments. The central sample is used to show the distribution of variance estimators over 1000 simulations, while the 50 random samples are used to evaluate the average bias and CI coverage rate. The results of the evaluation are presented in Figure 2 and Tables 1 and 2. A small difference in the ground truth generated by different packages is noted in Appendix I due to subtle differences in the packages' implementations.

6.2. Results for $k \leqslant n/2$

TABLE 1
90% CI Coverage Rate averaged on 50 testing samples. The number in the bracket is the standard deviation of coverage over 50 testing samples.

	$k = n/2$		$k = n/4$		$k = n/8$	
nTrees	2000	20000	2000	20000	2000	20000
MARS						
MS	81.5% (2.1%)	85.9% (1.6%)	82.2% (2.6%)	88.3% (1.1%)	82.5% (2.4%)	88.2% (1.4%)
MS-s	88.2% (2.9%)	89.3% (2.9%)	87.8% (2.6%)	89.9% (2.4%)	87.5% (1.8%)	89.3% (2.2%)
BM	80.8% (3.0%)	65.7% (1.8%)	91.4% (1.8%)	81.2% (1.5%)	93.8% (1.3%)	86.7% (1.1%)
BM-cor	13.9% (9.2%)	60.1% (1.6%)	70.6% (3.0%)	78.7% (1.4%)	82.9% (1.2%)	85.2% (1.1%)
IJ	95.5% (1.0%)	96.7% (1.0%)	89.8% (1.6%)	91.1% (1.0%)	92.3% (1.3%)	88.4% (1.1%)
MLR						
MS	83.3% (1.1%)	86.7% (1.1%)	84.1% (1.5%)	88.2% (0.8%)	84.1% (1.5%)	88.2% (0.8%)
MS-s	89.5% (1.7%)	89.9% (1.4%)	89.1% (1.4%)	90.2% (1.4%)	88.6% (1.4%)	90.5% (1.0%)
BM	78.6% (1.6%)	64.4% (1.8%)	90.1% (1.2%)	81.2% (1.3%)	93.3% (1.0%)	86.2% (0.9%)
BM-cor	19.0% (5.5%)	59.4% (1.8%)	72.0% (1.5%)	78.9% (1.3%)	83.3% (1.2%)	84.8% (0.9%)
IJ	95.7% (0.9%)	96.6% (0.7%)	89.6% (1.0%)	91.6% (0.7%)	91.4% (1.1%)	88.7% (1.0%)

Figure 2 presents the evaluation results for the MARS data. The subfigures show the distribution of variance estimators on the central test sample and the corresponding 90% confidence interval (CI) coverage on 50 testing samples. As previously mentioned, the bias of each estimator is compared to its population mean, eliminating the influence from different packages. The results for the MLR data are provided in Appendix I and show similar patterns. Tables 1 and 2 present the 90% CI coverage rate and relative bias of the variance estimation, respectively. The coverage for each method is calculated as the average over

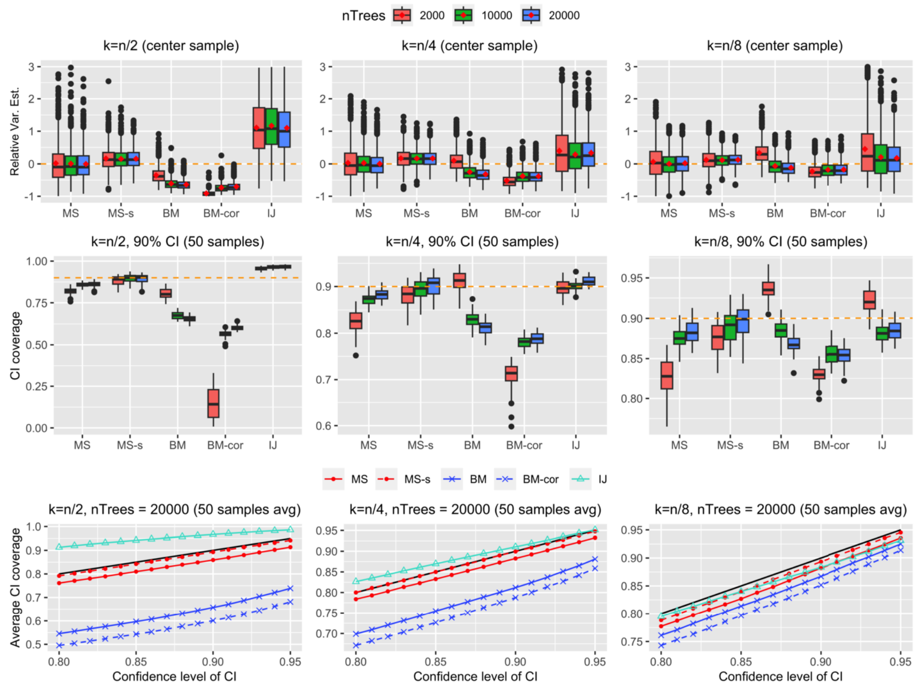

FIG 2. *A comparison of different methods on the MARS data is presented. Each column in the figure represents a different tree size: $k = n/2, n/4, n/8$ respectively. The first row displays boxplots of the relative variance estimators on a central test sample, evaluated over 1000 simulations. The range of y-coordinate is restricted within $[-1, 3]$. The mean is represented by the red diamond symbol in each boxplot. The second row displays boxplots of the 90 % confidence interval (CI) coverage for 50 testing samples. The third row displays the average coverage rate over 50 testing samples, with* $\texttt{nTrees} = 20000$. *The black reference line, $y = x$, represents the desired coverage rate.*

TABLE 2
Relative bias (standard deviation) over 50 testing samples. For each method and testing sample, the relative bias is evaluated over 1000 simulations.

	$k = n/2$		$k = n/4$		$k = n/8$	
nTrees	2000	20000	2000	20000	2000	20000
MARS						
MS	0.9% (2.6%)	0.3% (2.3%)	0.3% (2.3%)	0.7% (1.8%)	−0.5% (2.2%)	−0.4% (1.5%)
MS-s	3.8% (14.2%)	3.6% (13.9%)	3.6% (13.1%)	3.8% (13.2%)	2.1% (8.8%)	2.2% (8.8%)
BM	−31.5% (8.7%)	−64.3% (1.2%)	17.1% (12.2%)	−31.9% (2.0%)	38.4% (9.3%)	−12.3% (1.8%)
BM-cor	−103.1% (8.3%)	−71.5% (1.1%)	−54.4% (3.9%)	−39.1% (1.1%)	−25.5% (1.9%)	−18.8% (1.3%)
IJ	108.9% (15.3%)	111.1% (15.5%)	44.6% (12.8%)	25.2% (3.9%)	73.8% (19.3%)	15.9% (3.8%)
MLR						
MS	−0.3% (2.3%)	0.8% (2.1%)	−0.4% (2.5%)	−0.9% (2.1%)	−0.5% (2.3%)	0.1% (1.6%)
MS-s	5.8% (7.6%)	6.6% (7.9%)	5.2% (6.7%)	4.7% (6.7%)	3.9% (4.9%)	4.1% (4.8%)
BM	−38.9% (3.4%)	−65.6% (0.9%)	8.0% (5.7%)	−33.3% (1.4%)	31.4% (5.3%)	−14.5% (1.5%)
BM-cor	−97.3% (3.5%)	−71.4% (0.9%)	−52.5% (2.1%)	−39.4% (1.1%)	−25.1% (1.6%)	−20.2% (1.5%)
IJ	98.4% (11.5%)	101.0% (11.5%)	34.1% (4.2%)	22.9% (3.1%)	61.5% (9.4%)	11.1% (3.2%)

50 testing samples, and the standard deviation, indicated within the bracket, reflects the variation among these samples. Our simulation results show that the random forest estimators are approximately normally distributed, as the CIs constructed using the true variance achieve the desired confidence level (see Appendix I). In summary, MS and MS-s demonstrate consistently better performance compared to other methods, especially when the tree size k is large, i.e., $k = n/2$. The improved performance can be seen in terms of accurate CI coverage and reduced bias.

First, the third row of Figure 2 shows that the MS-s method achieves the best CI coverage under every k, i.e., the corresponding line is nearest to the reference line: $y = x$. The MS method performs the second best when $k = n/2$ and $n/8$. Furthermore, the CI coverages of the proposed methods are stable over different testing samples with a small standard deviation (less than 3%), as seen in Table 1. Secondly, with regards to the bias of the variance estimation, our methods show a much smaller bias than all other approaches (Figure 2, first row). More details of the relative bias are summarized in Table 2. The average bias of MS is smaller than 0.5% with a small standard deviation, mainly due to the Monte Carlo error. The MS-s method has a slightly positive average bias (0% to 6.2%), but it is still much smaller than the competing methods. The standard deviation of bias for MS-s is around 4.3% to 13.6%, which is comparable to IJ.

On the other hand, the performance of the competing methods varies. When the tree size is $k = n/2$, the BM, BM-cor, and IJ methods show a large bias, but their performance improves for smaller tree sizes. It is worth noting that these methods are theoretically designed for small k. BM and BM-cor tend to underestimate the variance in most settings, while IJ tends to overestimate. In Table 2, on the MARS data with 20000 `nTrees`, the bias of both BM and BM-cor is more than -50%, resulting in severe under-coverage (65.4%, 59.8%), while IJ leads to over-coverage. Even when the tree size is as small as $k = n/8$, these methods still display a noticeable bias. However, the proposed methods still outperform them when more trees (`nTrees` $= 20000$) are used, as shown in the last column of Table 2.

The results indicate that the choice of the number of trees has a significant effect on the performance of the estimators. This is to be expected due to the influence of the random kernels, the variation involved in incomplete U-statistics, and other theoretical aspects. As the number of trees increases, the variation of all estimators decreases, as can be seen in the first row of Figure 2. Our estimators, being mostly unbiased, benefit from larger `nTrees` values. For instance, the 90% CI coverages of the MS method on the MARS data increase from 81.5% ($k = n/2$) and 85.9% ($k = n/8$) with `nTrees` $= 2000$ to 82.5% and 88.2% respectively with `nTrees` $= 20000$. On the other hand, the performance of competing methods does not necessarily improve with an increase in `nTrees`. For example, BM shows over-coverage with `nTrees` $= 2000$ but under-coverage with `nTrees` $= 20000$ when $k = n/4$ or $n/8$. This phenomenon, known as estimation inflation, has been discussed by Zhou, Mentch and Hooker [35] and is addressed by the BM-cor method, which reduces the bias. When $k = n/8$, the gap between BM and BM-cor decreases as `nTrees` increases. However, this trend is no longer evident when k is large, as the dominating term used in their theory is no longer applicable.

We also present the computational cost for variance estimation methods in Appendix I. In short, once these tree predictions are obtained, the variance estimation is done immediately at little cost for all methods. After all predictions are done, the cost of MS is $\mathcal{O}(\texttt{nTrees})$ per testing sample. BM, BM-corr (bias-corrected BM) and IJ estimators add additional cost to this. They all involve using the number of training samples in each tree (see, e.g., Section 4.2 in [35]) and hence the total cost is at $\mathcal{O}(\texttt{nTrees} \cdot \texttt{nTrain})$. On the other hand, our MS-s estimator adds additional computational cost based on predicting additional neighboring samples for each testing sample, which will increase the cost in a different way. An additional set of analysis under a different tuning parameter ($\texttt{mtry} = 2$) is presented in Appendix I. Overall, our method is not significantly affected by this change.

Finally, we would like to emphasize the relationship between the bias of the estimator and the coverage rate of the confidence interval. Even though a random forest predictor is normally distributed and the variance estimator is unbiased, large fluctuations of the variance estimator can still lead to under-coverage. The same also applies to the IJ estimator. For example, on MARS data with $k = n/8$ and $\texttt{nTrees} = 20000$, IJ has a positive bias (11.5%), but its confidence interval is still under-coverage and even more severe than the proposed methods. Increasing the number of trees can improve this performance to some extend. An alternative strategy is to perform local averaging as implemented in the MS-s method, especially when `nTrees` is relatively small. The heights of the boxplots in the figure clearly demonstrate the variance reduction effect. As a result, the MS-s method with 2000 trees shows better coverage than the MS method with 20000 trees when $k = n/2$ (see Table 1). However, this maybe at the cost of larger bias. Hence, we still recommend using a larger number of trees whenever it is computationally feasible.

6.3. Results for $k > n/2$

As discussed in Section 5.1, when $n/2 < k < n$, we cannot jointly estimate $V^{(h)}$ and $V^{(s)}$. Additional computational cost is introduced using the bootstrap approach for estimating $V^{(h)}$. In this simulation study, we attempt to fit additional `nTrees` with bootstrapping (sampling with replacement) subsamples to estimate $V^{(h)}$ so we denote our proposed estimator and smoothing estimator as "MS(bs)" and "MS-s(bs)". We note that the `grf` package does not provide IJ estimator when $k > n/2$ so we generate the IJ estimator and corresponding ground truth by the `ranger` package.

As seen from Table 3, all methods suffer from severe bias, but our methods and IJ are comparable and better than BM and BM-cor. More specifically, our proposed method generally over-covers due to overestimating the variance. The IJ method shows good accuracy on MARS data but has more severe over-coverage than our methods on MLR. Overall, to obtain a reliable conclusion

TABLE 3

90 % CI coverage, relative bias, and standard deviation averaged on 50 testing samples. Tree size $k = 0.8n$. The calculation follows previous tables.

		90% CI Coverage		Relative Bias	
Model	Method \ nTrees	2000	20000	2000	20000
MARS	MS(bs)	94.2% (2.8%)	95.4% (2.4%)	128.4% (64.8%)	136.6% (67.2%)
	MS-s(bs)	97.7% (1.5%)	98.1% (1.3%)	132.2% (66.7%)	140.6% (69.1%)
	BM	51.4% (3.8%)	33.9% (1.7%)	−80.4% (3.1%)	−92.1% (0.5%)
	BM-cor	0.0% (0.0%)	13.5% (4.5%)	−143.0% (12.1%)	−98.3% (1.3%)
	IJ	88.0% (4.6%)	87.1% (3.7%)	−0.8% (25.2%)	−5.6% (16.3%)
MLR	MS(bs)	94.3% (1.9%)	95.2% (1.7%)	98.4% (24.7%)	103.9% (25.4%)
	MS-s(bs)	96.6% (1.3%)	97.0% (1.2%)	104.8% (24.9%)	110.3% (25.6%)
	BM	47.9% (2.3%)	32.4% (1.5%)	−83.4% (1.2%)	−92.6% (0.3%)
	BM-cor	0.0% (0.0%)	15.9% (2.4%)	−132.7% (4.3%)	−97.5% (0.5%)
	IJ	99.4% (0.3%)	99.2% (0.3%)	182.8% (21.7%)	175.8% (16.7%)

of statistical inference, we recommend avoiding using $k > n/2$. This can be a reasonable setting when n is relatively large, and $k = n/2$ can already provide an accurate model.

7. Real data illustration

We use the Seattle Airbnb Listings dataset, which was obtained from Kaggle[1]. The purpose of this analysis is to predict the price of Airbnb units in Seattle. The dataset consists of 7515 samples and nine covariates, including latitude, longitude, room type, number of bedrooms, number of bathrooms, number of accommodates, number of reviews, presence of a rating, and the rating score. Further information about the dataset, including the missing value processing, can be found in Appendix J.

Given the large sample size, we fit 40000 trees to obtain a variance estimator. The tree size is fixed as half of the sample size: $k = 3757$. We construct 12 testing samples at 3 locations: Seattle-Tacoma International Airport (SEA Airport), Seattle downtown, and Mercer Island. We further consider four bedroom/bathroom settings as $1B1B$, $2B1B$, $2B2B$, and $3B2B$. Details of the latitude and longitude of these locations and other covariates are described in Appendix J. The price predictions, along with 95% confidence intervals, are presented in Figure 3. Overall, the predictions match our intuitions. In particular, we can observe that the confidence interval of 1B1B units at SEA Airport does not overlap with those corresponding to the same unit type at the other two locations. This is possible because the accommodations around an airport usually have lower prices due to stronger competition. We also observe that 2-bathroom units at SEA Airport and downtown have higher prices than 1-bathroom units. However, the difference between 2B2B and 3B2B units at SEA Airport is insignificant.

[1]https://www.kaggle.com/shanelev/seattle-airbnb-listings

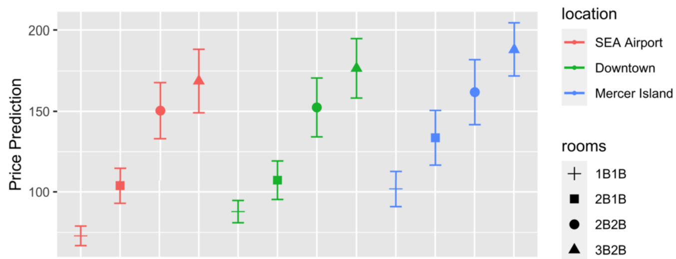

Fig 3. *Random Forest prediction on Airbnb testing data. The* 95% *confidence error bar is generated with our variance estimator, Matched Sample Variance Estimator. "2B1B" denotes the house/apartment has two bedrooms and one bathroom.*

8. Discussion

From the perspective of U-statistics, we have proposed a new framework of variance estimator for infinite-order U statistics. In contrast to estimating only the leading term, we establish a peak region dominance notion that utilize the hypergeometric density of the overlapping mechanism. This addresses the bias issue under large subsampling size k or small training size n. Additionally, new tools and strategies have been developed to study the ratio consistency behavior which is crucial for obtaining a proper coverage rate. Here, we discuss several open issues and possible extensions for future research.

First, our current methods are computationally valid for $k \leqslant n/2$. The difficulty of extending to the $k > n/2$ region is to estimate the tree variance, i.e., $V^{(h)}$. We proposed to use bootstrapped trees to extend the method to $k > n/2$. However, this could introduce additional bias and also leads to large variation, as we can see in the simulation study. We suspect Bootstrapping may be sensitive to the randomness involved in fitting trees. Since we estimate $V^{(h)}$ and $V^{(s)}$ separately, the randomness of the tree kernel could introduce different added variances, which leads to non-negligible bias. When $k > n/2$, Wang and Lindsay [32] propose an asymptotic unbiased variance estimator for the U-statistic estimator of a Kullback-Leibler risk in the k-fold cross-validation. However, this depends on a specific approximation of the kernel of Kullback-Leibler risk. The problem remains open for a general kernel.

Secondly, we developed a new double-U statistics tool to prove ratio consistency. This is the first work that analyzes the ratio consistency of a minimum-variance unbiased estimator (UMVUE) of a U-statistic's variance. The tool can be potentially applied to theoretical analyses of a general family of U-statistic problems. However, our ratio consistency result is still limited to $k = o(n^{1/2-\epsilon})$, introducing a gap between theoretical and practical versions. The limitation comes from the procedure we used to drive the Hoeffding decomposition of the variance estimator's variance. In particular, we want the leading term to dominate the variance while allowing a super-linear growth rate of each $\sigma^2_{c,2k}$ in terms of c. Hence, the extension to the $k = \beta n$ setting is still open and may require further assumptions on the overlapping structures of double-U statistics.

Thirdly, in our smoothed estimator, the choice of testing sample neighbors can be data-dependent and relies on the forest-defined distance. It is worth considering more robust smoothing methods for future work. Lastly, this paper focuses on the regression problem using random forest. This variance estimator can also be applied to the general family of subbagging estimators. Besides, we may further investigate the uncertainty quantification for variable importance, the confidence interval for classification probability, the confidence band of survival analysis, etc.

Appendix A: Notation table

TABLE 4
Summary of Notations.

Notations	Description
$\mathcal{O}$	$a = \mathcal{O}(b)$: exists $C > 0$, s.t. $a \leqslant Cb$..
Ω, $\asymp$	$a = \Omega(b) \Longleftrightarrow b = \mathcal{O}(a)$. $a \asymp b \Longleftrightarrow a = \mathcal{O}(b)$ and $a = \Omega(b)$.
U_n, h	U_n is the U-statistic with size-k kernel h.
$\hat{V}_u$, ψ	$\hat{V}_u$ denotes the estimator (4.1) of $\mathrm{Var}(U_n)$, which is a U-statistic with size-$2k$ kernel ψ.
S	S denotes the size-k subsample set associated with kernel h.
$S^{(2k)}$	$S^{(2k)}$ denotes the size-$2k$ subsample set associated with kernel ψ.
c, d_1, d_2	Given $S_1, S_2 \subset S_1^{(2k)}, S_3, S_4 \subset S_2^{(2k)}$, $c = \lvert S_1^{(2k)} \cap S_2^{(2k)} \rvert$, $d_1 = \lvert S_1 \cap S_2 \rvert$, and $d_2 = \lvert S_3 \cap S_4 \rvert$.
φ_d, w_d	See $\psi(S^{(2k)}) = \sum_{d=0}^{k} w_d \varphi_d \left(S^{(2k)}\right)$ (4.6). $\varphi_d(S^{(2k)})$ is still a U-statistic.
$\breve{w}_d$	$\breve{w}_d = \mathcal{O}(k^{2d}/(d!n^d))$ is the upper bound of w_d given by Equation (4.7).
$\xi^2_{d,k}$	$\xi^2_{d,k} = \mathrm{Cov}[h(S_1), h(S_2)]$ is first used in (2.1).
$\sigma^2_{c,2k}$	$\sigma^2_{c,2k} = \mathrm{Cov}[\psi(S_1^{(2k)}), \psi(S_2^{(2k)})]$ is first used in (4.3).
$\eta^2_{c,2k}(d_1, d_2)$	$\eta^2_{c,2k}(d_1, d_2)$ is introduced by further decomposing $\sigma^2_{c,2k}$ in (4.9).
$\breve{\sigma}^2_{c,2k}$	$\breve{\sigma}^2_{c,2k}$ is an upper bound of $\sigma^2_{c,2k}$ given by Propositions E.2 and E.3.
ρ	$\rho := \mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$ (4.11).
DoF	The number of free parameters to determine $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$.
$\underline{r}$, $\lvert\underline{r}\rvert$	$\underline{r}$ is a 9-dimensional vector defined in (B.1), describing the 4-way overlapping among S_1, S_2, S_3, S_4. $\lvert\underline{r}\rvert$ is the ℓ_1 vector norm of $\underline{r}$.
r_{i*}, r_{*j}, $\underline{r}^*$	$r_{i*} = \sum_{j=0}^{2} r_{ij}$, $r_{*j} = \sum_{i=0}^{2} r_{ij}$, and $\underline{r}^* = (r_{0*}, r_{1*}, r_{2*}, r_{*0}, r_{*1}, r_{*2})$.
$\rho(\underline{r})$	$\rho(\underline{r})$ is the 9 *DoF* representation of ρ (see Assumption 3).
$F_c^{(k)}$	$F_c^{(k)}$ (4.13) is the upper bound of ρ, given that $\lvert S_1^{(2k)} \cap S_2^{(2k)} \rvert = c$.
$\rho(\underline{r}, d_1, d_2)$	This is a notation emphasizing 11 *DoF* of ρ used in Appendix G.
$\tilde{\rho}(\underline{r})$	$\tilde{\rho}(\underline{r})$ is the 9 *DoF* benchmark of used in Assumption 6.
Influential Overlaps	The samples in $S_1^{(2k)} \cap S_2^{(2k)}$.

Appendix B: Discussion of assumptions

We present discussion and examples for Assumption 3-5, which are related to the covariance term $\rho = \text{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$ (4.11).

B.1. Definition of $\underline{r}$

We first define a 9-dimensional vector $\underline{r}$ to quantify ρ, which characterizes the overlaps between $S_1^{(2k)}, S_2^{(2k)}$.

Definition B.1 ($\underline{r}$)**.** Given size-$2k$ subsample sets $S_1^{(2k)}, S_2^{(2k)}$, and size-k subsample sets $S_1, S_2 \subset S_1^{(2k)}, S_3, S_4 \subset S_2^{(2k)}$, such that $c = |S_1^{(2k)} \cap S_2^{(2k)}|, d_1 = |S_1 \cap S_2|, d_2 = |S_3 \cap S_4|$. Denote $T_0 = S_1 \cap S_2$, $T_1 = S_1 \backslash S_2, T_2 = S_2 \backslash S_1$, $T_0' = S_3 \cap S_4, T_1' = S_3 \backslash S_4$, and $T_2' = S_3 \backslash S_4$ (see Figure 4).

Based on this, we denote the samples in $S_1^{(2k)} \cap S_2^{(2k)}$ as *Influential Overlaps* of ρ (4.11).In addition, we denote $R_{ij} := T_i \cap T_j'$, and $r_{ij} := |R_{ij}|$, for $i, j = 0, 1, 2$. Then, a 9-dimensional vector $\underline{r}$ is defined as follows:

$$\underline{r} := (r_{00}, r_{01}, r_{02}, r_{10}, r_{11}, r_{12}, r_{20}, r_{21}, r_{22})^T. \tag{B.1}$$

We define the norm of $\underline{r}$ as $|\underline{r}| = \sum_{i=0}^{2} \sum_{j=0}^{2} r_{ij}$. Note that each sample in $(S_1 \cup S_2 \cup S_3 \cup S_4) \cap (S_1^{(2k)} \cap S_2^{(2k)})$ is counted exactly once in $\underline{r}$ so $|\underline{r}| \leqslant c$.

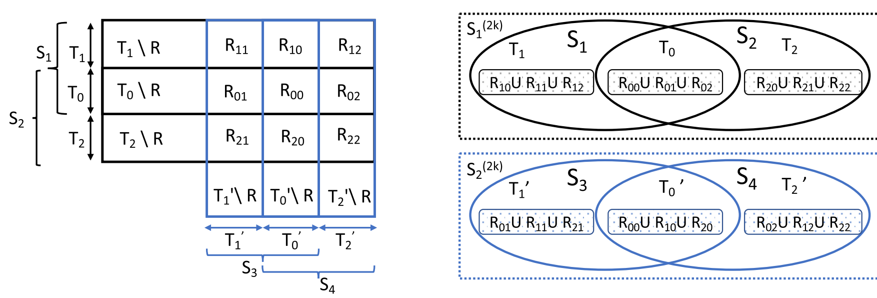

Fig 4. *Relationships among $S_1, S_2, S_3, S_4, S_1^{(2k)}$, and $S_2^{(2k)}$. Here, $R := S_1^{(2k)} \cap S_2^{(2k)}$.*

We note that the value of r_{ij} is naturally bounded by the sample size in the corresponding overlapping sets. For example, $r_{0*} = |(S_1 \cap S_2) \cap (S_3 \cup S_4)| \leqslant |(S_1 \cap S_2)| = d_1$. The 11 *DoF* of ρ can also be illustrated by the left panel of Figure 4. The figure shows 15 blocks, but given constraints $|S_1| = |S_2| = |S_3| = |S_4| = k$, we have $11 = 15 - 4$ *DoF*.

B.2. Discussion of Assumption 3

Assumption 3 reduces the *DoF* of ρ from 11 to 9 by dropping $d_1 = |S_1 \cap S_2|$ and $d_2 = S_3 \cap S_4$. The main reason being that, given the 9-dimensional vector $\underline{r}$, d_1

and d_2 only describe the overlap within $S_1^{(2k)}$ and $S_2^{(2k)}$, respectively, without adding new information for the overlap between them. This is illustrated using the following linear kernel example.

Example. Suppose h is a linear average kernel, $h(X_1,...,X_k) = \frac{1}{k}\sum_{i=1}^k X_i$, where $X_1,...,X_n$ *i.i.d.* X, where $\mathrm{E}X = 0, \mathrm{E}X^2 = \mu_2, \mathrm{E}X^3 = \mu_3, \mathrm{E}X^4 = \mu_4$. Notice that

$$\mathrm{E}(X_1X_2X_3X_4) = \begin{cases} \mu_2^2, & \text{if } X_i = X_j, X_k = X_l, X_i \neq X_k, \\ & \text{where } \{i,j,k,l\} = \{1,2,3,4\}; \\ \mu_4, & \text{if } X_1 = X_2 = X_3 = X_4; \\ 0, & \text{otherwise.} \end{cases}$$

Hence, we have

$$\mathrm{Cov}(X_1X_2, X_3X_4) = \begin{cases} cov_1 := \mu_2^2, & \text{if } X_1 = X_3 \neq X_2 = X_4; \\ & \text{or } X_1 = X_4 \neq X_2 = X_3; \\ cov_2 := \mu_4 - \mu_2^2, & \text{if } X_1 = X_2 = X_3 = X_4; \\ cov_3 := 0, & \text{otherwise.} \end{cases}$$

Hence, ρ can be represented as a weighted average in the form of $a_{1,n}cov_1 + a_{2,n}cov_2 + a_{3,n}cov_3$, where cov_3 is 0. Furthermore, by the definition of cov_1 and cov_2, $a_{1,n}$ only depends on r_{ij} for $(i,j) \neq (0,0)$ and $a_{2,n}$ only depends on $r_{0,0}$. Besides, we can also show that $F^{(k)}$ in (4.13) (see Assumption 4) is a quadratic function of c for this kernel.

We also shows that we may not be able to further reduce the *DoF*. When $\mathrm{E}(h(S)) = 0$, it is natural to consider the following fourth cumulant of ρ:

$$\begin{aligned} cum_4[h(S_1), h(S_2), h(S_3), h(S_4)] =& \rho - \mathrm{Cov}[h(S_1), h(S_3)]\mathrm{Cov}[h(S_2), h(S_4)] \\ & - \mathrm{Cov}[h(S_1), h(S_4)]\mathrm{Cov}[h(S_2), h(S_3)]. \end{aligned} \tag{B.2}$$

If $cum_4[h(S_1), h(S_2), h(S_3), h(S_4))]$ in (B.2) is a lower order term of ρ, the *DoF* can be reduced to 4, i.e., $|S_1 \cap S_3|$, $|S_1 \cap S_4|$, $|S_2 \cap S_3|$, and $|S_2 \cap S_4|$. However, the following example shows that this does not hold even for a linear average kernel. We can further verify this under a quadratic average kernel $h(S_1) = h(X_1,..,X_k) = \frac{1}{k^2}[\sum_{i=1}^k X_i]^2$, if X_i's are i.i.d. standard Gaussian.

Example. Given size-k sets S_1, S_2, S_3, S_4 s.t. $S_l = (X_1, Y_1^{(l)}, ..., Y_{k-1}^{(l)})$, $X_1, Y_j^{(l)}$ are i.i.d. $E(X_1) = 0, \mathrm{Var}(X_1) > 0$, for $j = 1,2,..,k-1$ and $l = 1,2,3,4$. By (B.2) and some direct calculations, we have $\rho = \frac{\mathrm{Var}(X_1^2)}{k^4}$, $\mathrm{Cov}[h(S_1), h(S_3)] = \frac{\mathrm{Var}(X_1)}{k^2}$. Plugging in the above equations, we have

$$\frac{cum_4[h(X_1), h(X_2), h(X_3), h(X_4)]}{2\mathrm{Cov}[h(S_1), h(S_3)]^2} = \frac{\mathrm{Var}(X_1^2) - 2\mathrm{Var}^2(X_1)}{2\mathrm{Var}^2(X_1)}. \tag{B.3}$$

As long as $\mathrm{Var}(X_1^2) - 2\mathrm{Var}^2(X_1) > 0$, which is common for non-Gaussian X_1, Equation (B.3) is larger than $o(1)$.

B.3. Discussion of Assumption 4

In Equation (4.13), $F_c^{(k)}$ is defined as an upper bound for ρ for a given c. As illustrated in Figure 5: $h(S_1)h(S_2)$ goes to $h(S')^2$ as more samples are shared between S_1 and S_2. Therefore, given that $|S_1^{(2k)} \cap S_2^{(2k)}| = c$, F_c has the most overlap among all ρ.

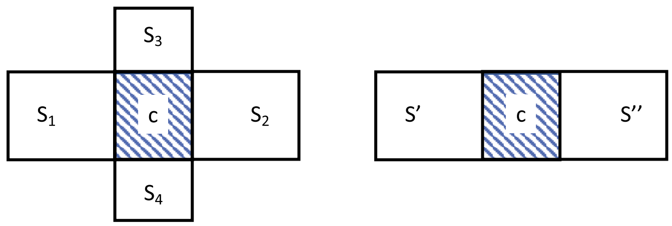

FIG 5. *An example of ordinal covariance assumption.*

We can verify this assumption on the linear average kernel again: $h(S_l) = \frac{1}{k}\sum_{i=1}^{k} X_i^{(l)}$, for $l = 1, 2, 3, 4$. In particular, considering (S_1, S_2, S_3, S_4) s.t. $|S_1 \cap S_2 \cap S_3 \cap S_4| = |(X_1^{(1)}, ..., X_c^{(1)})| = c$ and (S', S'') s.t. $|S' \cap S''| = c$, the equality in Equation (4.13) attains:

$$\rho = k^{-4}\mathrm{Var}\left[(X_1^{(1)} + ... + X_c^{(1)})^2\right] = \mathrm{Cov}\left[h(S')^2, h(S'')^2\right]. \tag{B.4}$$

It is also straightforward to verify the assumption under simple quadratic average kernel function $h(S_1) = \frac{1}{k^2}(\sum_{i=1}^{k} X_i)^2$

B.4. Discussion of Assumption 5

Assumption 2 imposes a polynomial growth rate of the second moment term $\xi_{d,k}^2$. Assumption 5 imposes a polynomial growth rate of the fourth moment term $F^{(k)}$. The following example helps illustrate the idea of Assumption 5.

Example. Suppose that there is no fourth-order cumulant term in Equation (B.2), which is valid for a linear kernel as average of i.i.d. standard Gaussian X_i's. Then, by Equation (B.4), $F_c^{(k)}$ can be simplified as $\xi_{r_{22},k}^2\xi_{r_{22},k}^2 + \xi_{r_{22},k}^2\xi_{r_{22},k}^2 = 2\xi_{r_{22},k}^4 = 2\xi_{c,k}^4$. This also implies (4.14): $F_1^{(k)}/\xi_{1,k}^4 = \mathcal{O}(1)$. We further remark that given this example, Assumption 5 can be implied by Assumption 2.

$$\frac{F_c^{(k)}}{F_1^{(k)}} = \frac{2\xi_{c,k}^4}{\xi_{1,k}^4} = 2\left(\frac{\xi_{c,k}^2}{\xi_{1,k}^2}\right)^2 = \mathcal{O}(c^{2a_1}), \text{ for } c = 1, 2, ..., k. \tag{B.5}$$

Here, given a_1 from Assumption 2, we can derive a_2 in Assumption 5 as $a_2 = 2a_1$.

We conclude the discussion by commenting on why we do not simply assume $\sigma^2_{2k,2k}/(2k\sigma^2_{1,2k}) = \mathcal{O}(1)$. In the recent literature, a similar assumption is used for $\xi^2_{d,k}$ [20, 35, 8]. By Lemma E.3 (see Appendix E), a natural upper bound for $\sigma^2_{c,2k}$ is $F_c^{(k)}$ (4.13), which can still grow at a quadratic rate of c (see (B.5)) even if $\xi^2_{d,k}$ has a linear growth rate in relation to d. This leads to a potential violation of $\sigma^2_{2k,2k}/(2k\sigma^2_{1,2k}) = \mathcal{O}(1)$.

Appendix C: Proof of results in theoretical section

C.1. Proof of Proposition 4.2 (Double U-statistic property)

Proof of Proposition 4.2.

Proof of Equation (4.5).

We first show the following equation.

$$\psi\left(S^{(2k)}\right) = \sum_{d=0}^{k} w_d \varphi_d \left(S^{(2k)}\right) \tag{C.1}$$

Wang and Lindsay [31] have demonstrated that $\hat{V}_u$ is an U-statistic with size-2k kernel (Equation (4.1)):

$$\hat{V}_u = Q(k) - Q(0) = \binom{n}{2k}^{-1} \sum_{S^{(2k)} \subseteq \mathcal{X}_n} \left[\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right)\right],$$

$$\psi_k\left(S^{(2k)}\right) = \underbrace{\binom{n}{2k}\binom{n}{k}^{-1}\binom{n}{k}^{-1}}_{A_1} \sum_{d=0}^{k} \frac{1}{N_d} \sum_{\substack{S_1, S_2 \subset S^{(2k)} \\ |S_1 \cap S_2| = d}} h\left(S_1\right) h\left(S_2\right),$$

$$\psi_0\left(S^{(2k)}\right) = \underbrace{\binom{n}{2k}\binom{n}{k}^{-1}\binom{n-k}{k}^{-1}}_{A_{1,0}} \frac{1}{N_0} \sum_{\substack{S_1, S_2 \subset S^{(2k)} \\ |S_1 \cap S_2| = 0}} h\left(S_1\right) h\left(S_2\right),$$

where $N_d = \binom{n-2k+d}{d}$.

Denote

$$A_1 := \binom{n}{2k}\binom{n}{k}^{-2}, \quad A_{1,0} := \binom{n}{2k}\binom{n}{k}^{-1}\binom{n-k}{k}^{-1}. \tag{C.2}$$

Rewrite $\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right)$ by the order of d. Notice that there is a $\sum_{d=0}^{k}$ in $\psi_k\left(S^{(2k)}\right)$ but d can only be 0 in $\psi_0\left(S^{(2k)}\right)$. Hence, there is a cancellation for $h(S_1)h(S_2)$ s.t. $d = |S_1 \cap S_2| = 0$, thus we have

$$\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right) = A_1 \sum_{d=1}^{k} \frac{1}{N_d} \sum_{\substack{S_1, S_2 \subset S^{(2k)} \\ |S_1 \cap S_2| = d}} h(S_1)\, h(S_2)$$
$$+ (A_1 - A_{1,0}) \frac{1}{N_0} \sum_{\substack{S_1, S_2 \subset S^{(2k)} \\ |S_1 \cap S_2| = 0}} h(S_1)\, h(S_2).$$

For the RHS of above equation, multiply and divide $M_{d,k}$ (4.6) inside $\sum_{d=1}^{k}$:

$$\begin{aligned}
&\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right) \\
&= \sum_{d=1}^{k} \underbrace{\left[A_1 \frac{1}{N_d} M_{d,k}\right]}_{w_d} \underbrace{\frac{1}{M_{d,k}} \sum_{S_1, S_2 \subset S^{(2k)},\, |S_1 \cap S_2| = d} h(S_1)\, h(S_2)}_{\varphi_d(S^{(2k)})} \\
&\quad + \underbrace{\left[(A_1 - A_{1,0}) \frac{1}{N_0} M_{0,k}\right]}_{w_0} \underbrace{\frac{1}{M_{0,k}} \sum_{S_1, S_2 \subset S^{(2k)},\, |S_1 \cap S_2| = 0} h(S_1)\, h(S_2)}_{\varphi_0(S^{(2k)})}.
\end{aligned}$$

We denote

$$w_0 := \frac{(A_1 - A_{1,0}) M_{0,k}}{N_0}; \quad w_d := \frac{A_1}{N_d} M_{d,k}, \text{ for } d = 1, 2, ..., k;$$
$$\varphi_d(S^{(2k)}) := \frac{1}{M_{d,k}} \sum_{\substack{S_1, S_2 \subset S^{(2k)} \\ |S_1 \cap S_2| = d}} h(S_1)\, h(S_2), \text{ for } d = 0, 1, 2, ..., k.$$

Given that $w_0 = -\sum_{d=1}^{k} w_d$, which will be proved later, we have

$$\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right) = \sum_{d=0}^{k} w_d \varphi_d(S^{(2k)}) = \sum_{d=1}^{k} w_d \left[\varphi_d\left(S^{(2k)}\right) - \varphi_0\left(S^{(2k)}\right)\right].$$

Proof of Equation Equation (4.7).

First, we show that $\sum_{d=0}^{k} w_d = 0$. As discussed above, w_d is a product of three normalization constants: $A_1 = \binom{n}{2k}\binom{n}{k}^{-2}$ and $A_{1,0} = \binom{n}{2k}\binom{n}{k}^{-1}\binom{n-k}{k}^{-1}$ are the normalization constant to rewrite $Q(k)$ and $Q(0)$ as a U-statistic; $M_{d,k} := \binom{2k}{d}\binom{2k-d}{d}\binom{2k-2d}{k-d}$ is the number of pairs $S_1, S_2 \subset S^{(2k)}$ s.t. $|S_1 \cap S_2| = d$; $N_d = \binom{n-2k+d}{d}$ is defined in Equation (4.2).

$$w_0 = \frac{(A_1 - A_{1,0}) M_{0,k}}{N_0}; \; w_d = \frac{A_1 M_{d,k}}{N_d}, \quad \text{for } d = 1, 2, ..., k.$$

Since $A_{1,0} > A_1 > 0$, $M_{d,k} > 0$, $N_d > 0$, we have $w_d > 0, \forall d \geqslant 1$ and $w_0 < 0$. Then we show $\sum_{d=0}^{k} w_d = 0$. Though this can be justified by direct calculation, we present a more intuitive proof. Recall $\hat{V}_u = Q(k) - Q(0)$. By the definition of $Q(k)$, $Q(k)$ can be represented as a weighted sum of $h(S_1)h(S_2)$, i.e., $\sum_{1\leqslant i,j\leqslant n} a_{ij} h(S_i)h(S_j)$, where $\sum_{1\leqslant i<j\leqslant n} a_{ij} = 1$. Thus, $Q(k) - Q(0)$ can be represented in a similar way:

$$Q(k) - Q(0) = \sum_{1\leqslant i<j\leqslant n} a'_{ij} h(S_i)h(S_j),$$

where $\sum_{1\leqslant i,j\leqslant n} a'_{ij} = 0$. Therefore $\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right)$, as the kernel of U-statistic $Q(k) - Q(0)$, can also be represented in the form of a weighted sum:

$$\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right) = \sum_{1\leqslant i<j\leqslant n} b_{ij} h(S_i)h(S_j), \tag{C.3}$$

where $\sum_{1\leqslant i<j\leqslant n} b_{ij} = 0$ since $\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right)$ is an unbiased estimator of $Q(k) - Q(0)$. On the other hand, for $d = 0, 1, 2, ..., k$, $\varphi_d(S^{(2k)})$ is still a U-statistic, which can be represented in the form of a weighted sum:

$$\varphi_d\left(S^{(2k)}\right) = \sum_{1\leqslant i<j\leqslant n} c_{ij}^{(d)} h(S_i)h(S_j), \tag{C.4}$$

where $\sum_{1\leqslant i<j\leqslant n} c_{ij}^{(d)} = 1$. Since $\psi_k\left(S^{(2k)}\right) - \psi_0\left(S^{(2k)}\right) = \sum_{d=0}^{k} w_d \varphi_d\left(S^{(2k)}\right)$, by comparing Equation (C.3) and (C.4), we have $\sum_{d=1}^{k} w_d \sum_{1\leqslant i<j\leqslant n} c_{ij}^{(d)} = \sum_{1\leqslant i<j\leqslant n} b_{ij}$. Since $\sum_{1\leqslant i<j\leqslant n} c_{ij}^{(d)} = 1$ and $\sum_{1\leqslant i<j\leqslant n} b_{ij} = 0$, we can take $h(S_i) = 1$ for $i = 1, 2, ..., n$ and conclude that $\sum_{d=0}^{k} w_d = 0$.

Secondly, we present the details to bound $w_d = A_1 M_{d,k}/N_d$, for $d = 1, 2, ..., k$. Plug in the expression of $A_1, M_{d,k}, N_d$, we have

$$\begin{aligned} w_d &= \left[\binom{n}{2k}\binom{n}{k}^{-2}\right]\left[\binom{2k}{d}\binom{2k-d}{d}\binom{2k-2d}{k-d}\right]\Big/\binom{n-2k+d}{d} \\ &= \left[\frac{n!}{(n-2k)!(2k)!}\frac{(n-k)!(n-k)!k!k!}{n!n!}\right] \\ &\quad \times \left[\frac{(2k)!(2k-d)!(2k-2d)!}{(2k-d)!d!(2k-2d)!d!(k-d)!(k-d)!}\right]\left[\frac{d!(n-2k)!}{(n-2k+d)!}\right]. \end{aligned}$$

After direct cancellation of the same factorials, we have

$$w_d = \underbrace{\frac{(n-k)!(n-k)!}{n!(n-2k+d)!}}_{\text{Part I}} \underbrace{\frac{k!k!}{(k-d)!(k-d)!}}_{\text{part II}} \underbrace{\frac{1}{d!}}_{\text{part III}}. \tag{C.5}$$

For Part I in Equation (C.5),

$$\frac{(n-k)!(n-k)!}{n!(n-2k+d)!} = \frac{\prod_{i=0}^{k-d-1}(n-k-i)}{\prod_{i=0}^{k-1}(n-i)} = [1 + o(1)]\frac{1}{n^d}.$$

The last equality is because for any $k = o(\sqrt{n}), d \leqslant k$, we have

$$\frac{\prod_{i=0}^{k-d-1}(n-k-i)}{\prod_{i=0}^{k-1}(n-i)} \leqslant \frac{\prod_{i=0}^{k-d-1}(n-i)}{\prod_{i=0}^{k-1}(n-i)} = \frac{1}{\prod_{i=0}^{d-1}(n-k+d-i)}$$
$$\leqslant \frac{1}{(n-k)^d} = [1+o(1)]\frac{1}{n^d}.$$

On the other hand, $\frac{\prod_{i=0}^{k-d-1}(n-k-i)}{\prod_{i=0}^{k-1}(n-i)} \geqslant \frac{1}{n^d}$. Thus, we have $= [1+o(1)]\frac{1}{n^d}$.

For Part II in Equation (C.5),

$$\frac{k!k!}{(k-d)!(k-d)!} = [k(k-1)...(k-d+1)]^2] \leqslant k^{2d}.$$

Particularly, when d is fixed, we have $\frac{k!k!}{(k-d)!(k-d)!} = k(k-1)...(k-d+1)]^2 = [1+o(1)]k^{2d}$. Combining Part I, II, III in (C.5), we have

$$w_d = \begin{cases} [1+o(1)]\left[\frac{1}{d!}(\frac{k^2}{n})^d\right] & \forall \text{ finite } d; \\ \mathcal{O}\left[\frac{1}{d!}(\frac{k^2}{n})^d\right] & \forall d = 1,2,...,k. \end{cases}$$

□

C.2. Proof of Corollary 4.6

Proof of Corollary 4.6. To show $\frac{\hat{V}_u}{E(\hat{V}_u)} \xrightarrow{P} 1$ as $n \to \infty$, it suffice to show the L_2 convergence of $\hat{V}_u/E(\hat{V}_u)$, i.e., $\mathrm{Var}(\hat{V}_u)/\left(\mathrm{E}(\hat{V}_u)\right)^2 \to 0$ as $n \to \infty$.

By plugging Equation (4.16) and (4.17) from Theorem 4.4, we have

$$\frac{\mathrm{Var}(\hat{V}_u)}{\left(\mathrm{E}(\hat{V}_u)\right)^2} = \frac{\mathcal{O}\left(\frac{k^2}{n}\check{\sigma}^2_{1,2k}\right)}{\left[(1+o(1))\,\frac{k^2}{n}\xi^2_{1,k}\right]^2} = \mathcal{O}\left(\frac{n}{k^2}\frac{\check{\sigma}^2_{1,2k}}{\xi^4_{1,k}}\right) = \mathcal{O}\left(\frac{1}{n}\right). \tag{C.6}$$

Here $\check{\sigma}^2_{1,2k} \asymp \frac{k^2}{n^2}F_1^{(k)}$ is the upper bound of $\sigma^2_{1,2k}$ given by Proposition E.2. By Assumption 5, $F_1^{(k)} = \mathcal{O}(\xi^4_{1,k})$. The last equality is by plugging in $\check{\sigma}^2_{1,2k}$. □

C.3. Proof of Theorem 4.4

We first present a technical proposition.

Proposition C.1. *For any integer c, s.t. $1 \leqslant c \leqslant k$ and $k = o(\sqrt{n})$,*

$$\binom{n}{k}^{-1}\binom{k}{c}\binom{n-k}{k-c} \leqslant \frac{1}{c!}\left(\frac{k^2}{n-k-1}\right)^c.$$

Proof of Proposition C.1. This proof is provided by DiCiccio and Romano [8]. We first write the combinatorial numbers as factorial numbers

$$\begin{aligned}\binom{n}{k}^{-1}\binom{k}{c}\binom{n-k}{k-c} &= \frac{(n-k)!k!}{n!}\frac{k!}{(k-c)!c!}\frac{(n-k)!}{(k-c)!(n-2k+c)!}\\ &= \frac{1}{c!}\left[\frac{k!k!}{(k-c)!(k-c)!}\right]\left[\frac{(n-k)!(n-k!)}{n!(n-2k+c)!}\right]\\ &\leqslant \frac{1}{c!}(k^{2c})\left[\frac{1}{n-k+1}\right]^{c}. \end{aligned} \tag{C.7}$$

□

Proof of Theorem 4.4. First, we show Equation (4.16). By Proposition 4.5 and Assumption 2, we can conclude that

$$\lim_{n\to\infty}\frac{\mathrm{Var}(U_n)}{\frac{k^2}{n}\xi_{1,k}^2} = 1.$$

Secondly, we show Equation (4.17).

$$\mathrm{Var}(\hat{V}_u) \asymp \mathrm{Var}^{(T_1)}(\hat{V}_u) = \sum_{c=1}^{T_1}\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}\sigma_{c,2k}^2 \tag{C.8}$$

$$= \sum_{c=1}^{T_1}\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}\mathcal{O}(\frac{k^2}{n^2}F_c^{(k)}) \tag{C.9}$$

$$= \sum_{c=1}^{T_1}\mathcal{O}(\frac{k^{2c+2}}{n^{c+2}}F_c^{(k)}) \tag{C.10}$$

$$=\mathcal{O}(\frac{k^4}{n^3}F_1^{(k)}). \tag{C.11}$$

Here, (C.8) is concluded by Lemma E.4. (C.9) is concluded by $\sigma_{c,2k}^2 = \mathcal{O}(\frac{k^2}{n^2}F_c^{(k)})$ (Proposition E.2). (C.10) is concluded by $\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c} = [1+o(1)]\frac{k^{2c}}{n^c}$ for $c = 1, 2, ..., T_1$ (Proposition C.1). (C.11) is concluded by the bounded growth rate of $F^{(k)}$ in Assumption 5 and finite T_1. By denoting $\frac{k^2}{n^2}F_1^{(k)}$ as $\breve{\sigma}_{1,2k}^2$, we conclude that $\mathrm{Var}(\hat{V}_u) = \mathcal{O}(\frac{k^2}{n}\breve{\sigma}_{1,2k}^2)$. □

C.4. Proof of Proposition 4.5

Proof of Proposition 4.5. For $k = o(\sqrt{n})$, we want to show $\lim_{n\to\infty}\frac{\mathrm{Var}(U_n)}{\frac{k^2}{n}\xi_{1,k}^2} = 1$. First notice that for the coefficient leading term $\binom{n}{k}^{-1}\binom{k}{1}\binom{n-k}{k-1}\xi_{1,k}^2$, we have

$$\frac{\binom{n}{k}^{-1}\binom{k}{1}\binom{n-k}{k-1}}{\frac{k^2}{n}} = \frac{(n-k)!(n-k)!}{(n-1)!(n-2k+1)!} \to 1, \text{ as } n\to\infty. \tag{C.12}$$

Therefore, it suffices to show that the rest part of $\mathrm{Var}(U_n)$ is dominated by the leading term:

$$\lim_{n\to\infty} \frac{\binom{n}{k}^{-1}\sum_{d=2}^{k}\binom{k}{d}\binom{n-k}{k-d}\xi_{d,k}^2}{\frac{k^2}{n}\xi_{1,k}^2} = 0.$$

By Proposition C.1, the numerator of the above can be bounded as

$$\binom{n}{k}^{-1}\sum_{d=2}^{k}\binom{k}{d}\binom{n-k}{k-d}\xi_{d,k}^2 \leqslant \sum_{d=2}^{k}\frac{k^{2d}}{d!(n-k+1)^d}\xi_{d,k}^2 : \sum_{d=2}^{k}(d!)^{-1}b_n^d\xi_{d,k}^2, \tag{C.13}$$

where $b_n = \frac{k^2}{n-k-1}$. Notice that $n < 2(n-k+1)$, so we have

$$\frac{\binom{n}{k}^{-1}\sum_{d=2}^{k}\binom{k}{d}\binom{n-k}{k-d}\xi_{d,k}^2}{\frac{k^2}{n}\xi_{1,k}^2} \leqslant \frac{n}{n-k+1}\sum_{d=2}^{k}\frac{1}{d!}b_n^{d-1}\frac{\xi_{d,k}^2}{\xi_{1,k}^2} \leqslant \sum_{d=2}^{k}\frac{2}{d!}b_n^{d-1}\frac{\xi_{d,k}^2}{\xi_{1,k}^2}. \tag{C.14}$$

By Assumption 2, the growth rate of $\xi_{d,k}^2$ is bounded, there exists a uniform constant C s.t. $\frac{\xi_{d,k}^2}{\xi_{1,k}^2} \leqslant Cd!$ for $d = 2, 3, ..., k$. Therefore, the RHS of Equation (C.14) is bounded as

$$\sum_{d=2}^{k}\frac{1}{d!}b_n^{d-1}\frac{\xi_{d,k}^2}{\xi_{1,k}^2} \leqslant \sum_{d=2}^{k}\frac{1}{d!}b_n^{d-1}\frac{Cd!\xi_{1,k}^2}{\xi_{1,k}^2} \leqslant C\sum_{d=2}^{k}b_n^{d-1} = Cb_n\frac{1-b_n^{k-1}}{1-b_n} \leqslant C\frac{b_n}{1-b_n}. \tag{C.15}$$

The RHS of (C.15) goes to 0 when $n \to \infty$, since $b_n = \frac{k^2}{n-k-1} \to 0$. This completes the proof. □

Appendix D: Proof of results in methodology section

D.1. Variance of incomplete U-statistics $U_{n,B,M}$

Proof of Proposition 3.2. This is an extension of the results by Wang [30, Section 4.1.1] and Wang and Lindsay [31].

Comparing $\mathrm{Var}(U_{n,B,M}) = (1-\frac{1}{B})Var(U_n)+\frac{1}{MB}V^{(h)}$ (3.8) with $\mathrm{Var}(U_{n,B}) = \mathrm{Var}(U_n) + \mathrm{E}[\mathrm{Var}(U_{n,B}|\mathcal{X}_n)]$ (2.4) , it suffices to show that

$$\mathrm{E}\left[\mathrm{Var}(U_{n,B}|\mathcal{X}_n)\right] = \frac{1}{MB}V^{(h)} - \frac{1}{B}\mathrm{Var}(U_n).$$

Here we adopt an alternative view of a complete U-statistic U_n with $k \leqslant n/2$ by Wang and Lindsay [31]. Follow our notation of "matched group", we can always take $M = \lfloor n/k \rfloor$ mutually disjoint subsamples $S_1, ..., S_M$ from $(X_1, ..., X_n)$, such that $|S_i \cap S_j| = 0$ for $1 \leqslant i < j \leqslant M$. Wang and Lindsay [31] take integer $M = n/k$ while we allow $2 \leqslant M \leqslant \lfloor n/k \rfloor$. Recall such $(S_1^{(b)}, ..., S_M^{(b)})$ as a "matched group", where b is the index of group. Let $\mathcal{G}_{n,k,M}$ be the collection of all such matched groups constructed from n samples, i.e.,

$$\mathcal{G}_{n,k,M} = \left\{(S_1^{(b)}, \ldots, S_M^{(b)}) : \cup_j S_j^{(b)} \subset \mathcal{X}_n, \text{ and } S_i^{(b)} \cap S_j^{(b)} = \varnothing, \forall 1 \leqslant i, j \leqslant M\right\}. \tag{D.1}$$

Then, an alternative representation of U_n is

$$U_n = \frac{1}{M|\mathcal{G}_{n,k,M}|} \sum_{b=1}^{|\mathcal{G}_{n,k,M}|} \sum_{i=1}^{M} S_i^{(b)}. \tag{D.2}$$

This form seems redundant because there are some replicate subsample among all $S_i^{(b)}$'s. However, for incomplete U-statistic $U_{n,B,M}$, each $(S_1^{(b)}, ..., S_M^{(b)})$ can be viewed as a sample from $\mathcal{G}_{n,k,M}$. Hence, Wang [30] show that $B \cdot \mathrm{Var}(U_{n,B}|\mathcal{X}_n) = \frac{1}{|\mathcal{G}_{n,k,M}|} \sum_{b=1}^{|\mathcal{G}_{n,k,M}|} (\bar{h}^{(b)} - U_n)^2$, where $\bar{h}^{(b)} = \frac{1}{M} \sum_{i=1}^{M} h(S_i^{(b)})$, $S_i^{(b)}$'s are all subsamples associated with the complete U-statistic U_n on $\mathcal{X}$. However, Wang [30] and Wang and Lindsay [31] do not provide a simple expression in the form of $V^{(h)}$ and $\mathrm{Var}(U_n)$. We further simplify $B\mathrm{Var}(U_{n,B}|\mathcal{X}_n)$ as follows,

$$\begin{aligned}
B \cdot \mathrm{E}\left[\mathrm{Var}(U_{n,B}|\mathcal{X}_n)\right] &= \mathrm{E}\left(\frac{1}{|\mathcal{G}_{n,k,M}|} \sum_{b=1}^{|\mathcal{G}_{n,k,M}|} (\bar{h}^{(b)} - U_n)^2\right) \\
&= \mathrm{E}\left[\frac{1}{|\mathcal{G}_{n,k,M}|} \sum_{b=1}^{|\mathcal{G}_{n,k,M}|} \Big((\bar{h}^{(b)} - \mathrm{E}(U_n)) - (U_n - \mathrm{E}(U_n))\Big)^2\right] \\
&= \frac{1}{|\mathcal{G}_{n,k,M}|}\left[\sum_{b=1}^{|\mathcal{G}_{n,k,M}|} \mathrm{E}\left(\Big(\bar{h}^{(b)} - \mathrm{E}(U_n)\Big)^2\right)\right] \\
&\quad + \frac{1}{|\mathcal{G}_{n,k,M}|}\left[\sum_{b=1}^{|\mathcal{G}_{n,k,M}|} \mathrm{E}\left((U_n - \mathrm{E}(U_n))^2\right)\right] \\
&\quad - 2\mathrm{E}\left(\frac{1}{|\mathcal{G}_{n,k,M}|} \sum_{b=1}^{|\mathcal{G}_{n,k,M}|} (\bar{h}^{(b)} - \mathrm{E}(U_n))(U_n - \mathrm{E}(U_n)\right) \\
&= \mathrm{Var}\left(\bar{h}^{(1)}\right) - \mathrm{Var}(U_n) \\
&= \frac{1}{M} V^{(h)} - \mathrm{Var}(U_n).
\end{aligned}$$

In the above equations, the first equality is the conclusion by Wang [30]; the next-to-last equality holds since $\frac{1}{|\mathcal{G}_{n,k,M}|} \sum_{b=1}^{|\mathcal{G}_{n,k,M}|} \bar{h}^{(b)} = U_n$; the last equality holds since $h(S_1^{(1)}), ..., h(S_M^{(1)})$ are independent. $\square$

D.2. Unbiasedness of variance estimators

Proof of Proposition 3.1. First, we restrict the discussion given $k \leqslant n/2$. We first show that $\hat{V}^{(s)} = \sum_{d=0}^{k} \gamma_{d,k,n} \hat{\tilde{\xi}}^2_{d,k}$. By the discussion in Section 3.3.1, we have $N_{d,k,n} = \binom{n}{k}^2 \gamma_{d,k,n}$. For a complete U-statistic with $k \leqslant n/2$, $N_{d,k,n} = \binom{n}{k}\binom{n-k}{k-d}\binom{k}{d}$ and we denote $\mathbb{N} = \binom{n}{k} > 0$. Then,

$$
\begin{aligned}
\sum_{d=0}^{k} \gamma_{d,k,n} \hat{\tilde{\xi}}^2_{d,k} =& \mathbb{N}^{-2} \left[2 \sum_{1 \leqslant i < j \leqslant \mathbb{N}} \sum_{d=0}^{k-1} \mathbb{1}\{|S_i \cap S_j| = d\} [h(S_i) - h(S_j)]^2/2 \right] \\
=& \frac{\mathbb{N}(\mathbb{N}-1)}{\mathbb{N}^2} \left[\binom{\mathbb{N}}{2}^{-1} \sum_{1 \leqslant i < j \leqslant \mathbb{N}} [h(S_i) - h(S_j)]^2/2 \right] \\
=& \frac{\mathbb{N}-1}{\mathbb{N}} \left[\frac{1}{\mathbb{N}-1} \sum_{i=1}^{\mathbb{N}} [h(S_i) - U_n]^2 \right] \\
=& \frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} [h(S_i) - U_n]^2 = \hat{V}^{(s)}.
\end{aligned}
$$

Here, the second equality holds by plugging in the definition of $N_{d,k,n}$ and interchanging the finite summation $\sum_{d \in \mathcal{D}}$ with $\sum_{1 \leqslant i < j \leqslant B}$. The third equality omits the cases with $i = j$, where $h(S_i) - h(S_j) = 0$. The second to last equality holds because the sample variance is essentially an order-2 U-statistic, with kernel $(h(S_i) - h(S_j))^2/2$.

Then, as we demonstrated in Section 3.3.1, $\mathrm{E}(\hat{\tilde{\xi}}^2_{d,k}) = \tilde{\xi}^2_{d,k}$ for $d = 0, 1, ..., k$. Hence we conclude that

$$
\mathrm{E}\left(\hat{V}^{(s)}\right) = \sum_{d=0}^{k} \gamma_{d,k,n} \mathrm{E}\left(\hat{\tilde{\xi}}^2_{d,k}\right) = \sum_{d=0}^{k} \gamma_{d,k,n} \tilde{\xi}^2_{d,k} = V^{(s)}.
$$

Secondly, we extend the previous argument to the setting $n/2 < k < n$. We denote $\mathcal{D} = \{d \in N^* | 0 \leqslant d \leqslant k, \gamma_{d,k,n} > 0\}$. We can define $\hat{V}^{(s)} = \sum_{d \in \mathcal{D}} \gamma_{d,k,n} \hat{\tilde{\xi}}^2_{d,k}$. We want to show that

$$
\hat{V}^{(s)} = \frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} [h(S_i) - U_n]^2, \tag{D.3}
$$

$$
\mathrm{E}(\hat{V}^{(s)}) = V^{(s)}. \tag{D.4}
$$

Similar to previous proof

$$
\begin{aligned}
\hat{V}^{(S)'} =& \mathbb{N}^{-2} \left[2 \sum_{1 \leqslant i < j \leqslant \mathbb{N}} \sum_{d \in \mathcal{D} \setminus \{k\}} \mathbb{1}\{|S_i \cap S_j| = d\} [h(S_i) - h(S_j)]^2/2 \right] \\
=& \frac{\mathbb{N}(\mathbb{N}-1))}{\mathbb{N}^2} \left[\binom{\mathbb{N}}{2}^{-1} \sum_{1 \leqslant i < j \leqslant \mathbb{N}} [h(S_i) - h(S_j)]^2/2 \right]
\end{aligned}
$$

$$=\frac{\mathbb{N}-1}{\mathbb{N}}\left[\frac{1}{\mathbb{N}-1}\sum_{i=1}^{\mathbb{N}}[h(S_i)-U_n]^2\right]=\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}[h(S_i)-U_n]^2$$

Since each $\hat{\tilde{\xi}}^2_{d,k}$ is still an unbiased estimator of $\tilde{\xi}^2_{d,k}$, similarly, we have $\mathrm{E}(\hat{V}^{(S)'}) = V^{(s)}$. Remark that the summation in Equation (D.3) is over $d \in \mathcal{D}$ instead of $d = 0, 1, 2, ..., n$. This is because $\gamma_{d,k,n}$ is 0 for small d, given $k > n/2$. In other words, when $k > n/2$, several terms of $\gamma_{d,k,n}\xi^2_{d,k}$ in the Hoeffding decomposition (2.1) is already 0. □

Proof of Proposition 3.3. Since a sample variance is an order-2 U-statistics and we denote $\hat{v}_{(i,j,i',j')} = [h(S_i^{(j)}) - h(S_{i'}^{(j')})]^2/2$:

$$\begin{aligned}\hat{V}^{(s)}_{B,M} =&[(BM-1)BM]^{-1}\sum_{i=1}^{M}\sum_{j=1}^{B}\sum_{(i',j')\neq(i,j)}\left[h(S_i^{(j)})-h(S_{i'}^{(j')})\right]^2/2\\ =&[(BM-1)BM]^{-1}\sum_{i=1}^{M}\sum_{j=1}^{B}\left(\sum_{(i',j')\in\mathcal{A}(i,j)}+\sum_{(i',j')\in\mathcal{B}(i,j)}\right)\hat{v}_{(i,j,i',j')},\end{aligned}$$

where $\mathcal{A}(i,j) = \{(i',j')|i' \neq i, j' = j, 1 \leqslant i' \leqslant M, 1 \leqslant j' \leqslant B\}$; $\mathcal{B}(i,j) = \{(i',j')|j' \neq j, 1 \leqslant i' \leqslant M, 1 \leqslant j' \leqslant B\}$. We note that $|\mathcal{A}(i,j)| = M-1$ and $|\mathcal{B}(i,j)| = [(B-1)(M-1)]$ for any (i,j).

Fixing (i,j), for any $(i',j') \in \mathcal{A}(i,j)$, $S_{i'}^{(j')}$ and $S_i^{(j)}$ are the same j but not identical. Hence, $\hat{v}_{(i,j,i',j')}$ is an unbiased estimator of $\hat{V}^{(h)}$. Furthermore, the sample variance within group j is also a U-statistic, which can be alternatively represented as an order-2 U-statistic: $[M(M-1)]^{-1}\sum_{i=1}^{M}\sum_{i'\neq i}\hat{v}_{(i,j,i',j')}$. Thus, let $\delta_{M,B} = \frac{M-1}{MB-1}$, by summation over all j and the symmetry, we have

$$\begin{aligned}&[(BM-1)BM]^{-1}\sum_{i=1}^{M}\sum_{j=1}^{B}\sum_{(i',j')\in\mathcal{A}(i,j)}\mathrm{E}(\hat{v}_{(i,j,i',j')})\\ &=[(BM-1)BM]^{-1}\sum_{i=1}^{M}\sum_{j=1}^{B}|\mathcal{A}(i,j)|V^{(h)}\\ &=\delta_{M,B}V^{(h)}.\end{aligned}\tag{D.5}$$

Fixing (i,j), for any $(i',j') \in \mathcal{B}(i,j)$, $S_i^{(j)}$ and $S_{i'}^{(j')}$ are in different matched group. Since each matched group are sampled independently, $S_i^{(j)}$ and $S_{i'}^{(j')}$ are independently sampled from $\mathcal{X}_n$. By the theory of finite population sampling [7], for $(i',j') \in \mathcal{B}(i,j)$,

$$\mathrm{E}(\hat{v}_{(i,j,i',j')}) = \mathrm{E}[\mathrm{E}(\hat{v}_{(i,j,i',j')}|\mathcal{X}_n)] = \mathrm{E}\left[\binom{n}{k}^{-1}\sum_{S_i\in\mathcal{X}_n}(h(S_i)-U_n)^2\right] = V^{(s)}$$

Thus, the normalized summation over all such $\hat{v}_{(i,j,i',j')}$ satisfies that

$$\begin{aligned}
&[[(BM-1)BM]^{-1}\sum_{i=1}^{M}\sum_{j=1}^{B}\sum_{(i',j')\in\mathcal{B}(i,j)}E(\hat{v}_{(i,j,i',j')})\\
&=[(BM-1)BM]^{-1}\sum_{i=1}^{M}\sum_{j=1}^{B}|\mathcal{B}(i,j)|V^{(s)}\\
&=(1-\delta_{M,B})V^{(s)}. \qquad \text{(D.6)}
\end{aligned}$$

Combining Equations (D.5) and (D.6), we conclude that

$$\mathrm{E}\left(\hat{V}_{B,M}^{(s)}\right)=(1-\delta_{M,B})V^{(s)}+\delta_{M,B}V^{(h)}. \qquad \square$$

Proof of Proposition 3.4. On one hand, by Proposition 3.3

$$\begin{aligned}
\mathrm{E}\left(\widehat{\mathrm{Var}}(U_{n,B,M})\right)=&\mathrm{E}(\hat{V}_{B,M}^{(h)})-\frac{MB-1}{MB}\mathrm{E}(\hat{V}_{B,M}^{(s)})\\
=&V^{(h)}-\frac{MB-1}{MB}\left[(1-\frac{M-1}{MB-1})V^{(s)}+\frac{M-1}{MB-1}V^{(h)}\right]\\
=&\frac{MB-M+1}{MB}V^{(h)}-\frac{B-1}{B}V^{(s)}.
\end{aligned}$$

On the other hand, by $\mathrm{Var}(U_n)=V^{(h)}-V^{(s)}$ (3.2) and Proposition and 3.2,

$$\mathrm{Var}(U_{n,B,M})=\frac{B-1}{B}Var(U_n)+\frac{1}{MB}V^{(h)}=\frac{MB-M+1}{MB}V^{(h)}-\frac{B-1}{B}V^{(s)}.$$

Hence, we conclude the unbiasedness of our incomplete variance estimator:

$$\mathrm{E}\left(\widehat{\mathrm{Var}}(U_{n,B,M})\right)=\mathrm{Var}(U_{n,B,M}). \qquad \square$$

D.3. Equivalence of complete variance estimators

We denote the complete variance estimator by us, Schucany and Bankson [24], and Wang and Lindsay [31] as $\hat{V}_u$ (3.6), $\hat{V}_u^{(S\&B)}$, and $\hat{V}_u^{(W\&L)}$ respectively.

First, our complete U-statistic variance estimator is identical to the estimator in page 79 of Folsom [10]'s work.

Secondly, we restrict $k=2$ and show that $\hat{V}_u=\hat{V}_u^{(S\&B)}$. Schucany and Bankson [24] estimate two terms, $\xi_{1,2}^2$ and $\xi_{2,2}^2$ in the Hoeffding decomposition as $\tilde{\zeta}_1^2$ and $\tilde{\zeta}_{2,2}^2$ respectively as follows. We adapt their notation to simplify $h(X_i,X_j)$ as h_{ij}.

$$\begin{aligned}
\tilde{\zeta}_{1,2}^2&=\binom{n}{3}^{-1}\sum_{i<j<l}h_0^*(X_i,X_j,X_l)-\binom{n}{4}^{-1}\sum_{i<j<l<m}h_1^*(X_i,X_j,X_l,X_m),\\
\tilde{\zeta}_{2,2}^2&=\binom{n}{4}^{-1}\sum_{i<j<l<m}g^*(X_i,X_j,X_l,X_m),
\end{aligned}$$

where

$$h_0^*(X_i,X_j,X_l)=\frac{1}{3}\left[h_{ij}h_{il}+h_{ij}h_{jl}+h_{il}h_{jl}\right]$$
$$h_1^*(X_i,X_j,X_l,X_m)=\frac{1}{3}\left[h_{ij}h_{lm}+h_{il}h_{jm}+h_{im}h_{jl}\right],$$
$$g^*(X_i,X_j,X_l,X_m)=\frac{1}{6}\left[(h_{ij}-h_{lm})^2+(h_{il}-h_{jm})^2+(h_{im}-h_{jl})^2\right].$$

Then, by estimating corresponding terms in the Hoeffding decomposition (2.1),

$$\hat{V}_u^{(S\&B)}=\binom{n}{2}^{-1}\binom{2}{1}\binom{n-1}{1}\tilde{\zeta}_{1,2}^2+\binom{n}{2}^{-1}\tilde{\zeta}_{2,2}^2.$$

By our proposed decomposition, $\mathrm{Var}(U_n)=V^{(h)}-V^{(s)}$ and Proposition 3.1, our estimation approach is equivalent to estimate $\xi_{1,2}^2$ and $\xi_{2,2}^2$ by $\hat{V}^{(h)}-\hat{\hat{\xi}}_{1,2}^2$ and $\hat{V}^{(h)}$ respectively. When $k=2$, $\hat{V}^{(h)}=(\binom{n}{2}\binom{n-2}{2})^{-1}\sum_{|S_i\cap S_j|=0}(h(S_i)-h(S_j))^2$ $/2$ and $\hat{\hat{\xi}}_{1,2}^2=(2\binom{n}{2}\binom{n-2}{1})^{-1}\sum_{|S_i\cap S_j|=1}(h(S_i)-h(S_j))^2/2$. Hence, to show that $\hat{V}_u=\hat{V}_u^{(S\&B)}$, it suffices to show that $\hat{V}^{(h)}-\hat{\hat{\xi}}_{1,2}^2=\tilde{\zeta}_1^2$ and $\hat{V}^{(h)}=\tilde{\zeta}_2^2$ respectively. For the first equality, we can simplify these terms as follows.

$$\begin{aligned}
\tilde{\zeta}_{1,2}^2 &= \left(3\binom{n}{3}\right)^{-1}\sum_{i=1}^{n}\sum_{j:j>i.}\sum_{l:l\neq i,j.}h_{ij}h_{jl}\\
&\quad-\left(\binom{n}{2}\binom{n-2}{2}\right)^{-1}\sum_{i=1}^{n}\sum_{j:j>i.}\sum_{l:l\neq i,j.}\sum_{m:m\neq i,j;m>l.}h_{ij}h_{lm},\\
\hat{V}^{(h)} &= \binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j:j>i.}h_{ij}^2\\
&\quad-\left(\binom{n}{2}\binom{n-2}{2}\right)^{-1}\sum_{i=1}^{n}\sum_{j:j>i}\sum_{l:l\neq i,j.}\sum_{m:m\neq i,j;m>l.}h_{ij}h_{lm},\\
\hat{\hat{\xi}}_{1,2}^2 &= \binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j:j>i.}h_{ij}^2-\left(3\binom{n}{3}\right)^{-1}\sum_{i=1}^{n}\sum_{j:j>i.}\sum_{l:l\neq i,j.}h_{ij}h_{jl}.
\end{aligned}$$

Therefore, $\tilde{\zeta}_{1,2}^2=\hat{V}^{(h)}-\hat{\hat{\xi}}_{1,2}^2$. For the latter equality, $\tilde{\zeta}_2^2$'s kernel g^* is composed by sample variance between two kernels with disjoint subsamples, such as h_{ij} and h_{lm}, $\tilde{\zeta}_{2,2}^2$ is a redundant version of our $\hat{V}^{(h)}$, which implies that $\hat{V}^{(h)}=\tilde{\zeta}_{2,2}^2$. This, concludes the equivalence.

We remark that Schucany and Bankson [24] consider an alternative estimator of $\xi_{1,2}^2$, denoted as $\hat{\zeta}_1^2$ [26]. However, that one shows connection to the work of Mentch and Hooker [20] and Zhou, Mentch and Hooker [35] (see Section 3.5) but is not the focus of this appendix.

Thirdly, note that Wang and Lindsay [31]'s estimator involves the definition of their partitioning scheme and we use the notation of our matching group and assume $M = n/k$ to present their estimator (see $\mathcal{G}_{n,k,M}$ (D.1) and in Appendix D.1). Here we present Wang and Lindsay [31]'s estimator in their ANOVA form, which is the alternative to their second-moment view. This alternative form uses the within and between-variances of the groups [see 31, page 1122]. However, the form is still different from ours. To simplify the notation, we denote $\mathbb{B} = |\mathcal{G}_{n,k,M}|$ and $\bar{h}_{(b)} = \frac{1}{M}\sum_{i=1}^{M} h(S_i^{(b)})$, for $b = 1, 2, ..., \mathbb{B}$. Under this notation, the alternative form (with $\mathcal{G}_{n,k,M}$) of U_n (D.2) is $\frac{1}{M\mathbb{B}}\sum_{i=1}^{M}\sum_{b=1}^{\mathbb{B}} h(S_i^{(b)})$. Then our estimator, $\hat{V}_u$, and Wang and Lindsay [31]'s estimator, $\hat{V}_u^{(W\&L)}$, can be represented as follows.

$$\hat{V}_u = \hat{V}^{(h)} - \hat{V}^{(s)},$$
$$\hat{V}_u^{(W\&L)} = \sigma_{WP}^2/M - \sigma_{BP}^2,$$

where $\hat{V}^{(s)} = \binom{n}{k}^{-1}\sum_{i=1}^{\binom{n}{k}}(h(S_i) - U_n)^2$ (3.4); $\hat{V}^{(h)} = \binom{n}{k}\binom{n-k}{d}\sum_{|S_i\cap S_j|=0}[h(S_i) - h(S_j)]^2/2$ (3.5); $\sigma_{WP}^2 = \frac{1}{\mathbb{B}}\sum_{b=1}^{\mathbb{B}}\frac{1}{M-1}\sum_{i=1}^{M}(h(S_i^{(b)}) - \bar{h}_{(b)})^2$; $\sigma_{BP}^2 = \frac{1}{\mathbb{B}}(\bar{h}_{(b)} - U_n)^2$.

Proposition D.1. *Our complete variance estimator $\hat{V}_u$ is equivalent to the estimator $\hat{V}_u^{(W\&L)}$ proposed by Wang and Lindsay [31].*

Proof. To simplify the notation, we denote

$$A_1 := \frac{1}{M\mathbb{B}}\sum_{b=1}^{\mathbb{B}}\sum_{i=1}^{M}(h(S_i^{(b)}))^2,\ A_2 := U_n^2,\ A_3 := \frac{1}{\mathbb{B}}\sum_{b=1}^{\mathbb{B}}(\bar{h}_{(b)})^2. \tag{D.7}$$

To show the equivalence between $\hat{V}_u$ and $\hat{V}_u^{(W\&L)}$, we will show that they are the same linear combination of A_1, A_2, A_3. First, it is trivial to verify that σ_{WP}^2 and σ_{BP}^2 are linear combinations of A_1, A_2, A_3:

$$\sigma_{WP}^2 = \frac{M}{M-1}(A_1 - A_3),\ \text{and}\ \sigma_{BP}^2 = A_3 - A_2. \tag{D.8}$$

Secondly, we show that $\hat{V}^{(h)} = \sigma_{WP}^2$ by showing that $\hat{V}^{(h)}$ also equals to $\frac{M}{M-1}(A_1 - A_3)$. Considering the summation $\frac{1}{M-1}\sum_{i=1}^{M}(h(S_i^{(b)}) - \bar{h}_{(b)})^2$ in σ_{WP}^2, it can be represented as

$$\binom{M}{2}^{-1}\sum_{i=1}^{M}\sum_{j\neq i}(h(S_i^{(b)}) - h(S_j^{(b)}))^2/2.$$

Since $\mathcal{G}_{n,k,M}$ is a set of all permutation of disjoint $(S_1^{(b)}, ..., S_M^{(b)})$, we have

$$\begin{aligned}\sigma_{WP}^2 &= \frac{1}{\mathbb{B}}\sum_{b=1}^{\mathbb{B}}\binom{M}{2}^{-1}\sum_{i=1}^{M}\sum_{j\neq i}(h(S_i^{(b)}) - h(S_j^{(b)}))^2/2 \\ &= \binom{n}{k}^{-1}\binom{n-k}{k}^{-1}\sum_{|S_i\cap S_j|=0}[h(S_i) - h(S_j)]^2/2 = \hat{V}^{(h)}.\end{aligned}$$

Thirdly, we show that $\hat{V}^{(s)}$ is also a linear combination of A_1, A_2, A_3. We start with $\hat{V}^{(s)} = \binom{n}{k}^{-1} \sum_{i=1}^{\binom{n}{k}} h(S_i)^2 - U_n^2$. Due to the definition of $\mathcal{G}_{n,k,M}$, the collection $\{S_i^{(b)}\}_{i,b}$ are basically replications of $\{S_i\}_{i=1}^{\binom{n}{k}}$. Hence, $\binom{n}{k}^{-1} \sum_{i=1}^{\binom{n}{k}} h(S_i)^2 = A_1$, which implies that

$$\hat{V}^{(s)} = A_1 - A_2. \tag{D.9}$$

Therefore, we can represent both $\hat{V}_u$ and $\hat{V}_u^{(W\&L)}$ with A_1, A_2, A_3 by (D.8) and (D.9) as follows:

$$\begin{aligned}
\hat{V}_u = \hat{V}^{(h)} - \hat{V}^{(s)} &= \frac{M}{M-1}(A_1 - A_3) - (A_1 - A_2) \\
&= \frac{1}{M-1}A_1 + A_2 - \frac{M}{M-1}A_3, \\
\hat{V}_u^{(W\&L)} = \sigma_{WP}^2/M - \sigma_{BP}^2 &= \frac{1}{M-1}(A_1 - A_3) - (A_3 - A_2) \\
&= \frac{1}{M-1}A_1 + A_2 - \frac{M}{M-1}A_3.
\end{aligned}$$

This conclude that $\hat{V}_u = \hat{V}_u^{(W\&L)}$. Note that $\hat{V}^{(h)} - \hat{V}^{(s)} = \sigma_{WP}^2/M - \sigma_{BP}^2$, however, $\hat{V}^{(h)} \neq \sigma_{WP}^2/M$ and $\hat{V}^{(s)} \neq \sigma_{BP}^2$. Our and Wang and Lindsay [31]'s estimators are proposed under different perspectives. □

D.4. Equivalence of incomplete variance estimators

Only Wang and Lindsay [31] and our paper propose variance estimator for in complete U-statistics. Similar to the analysis in Appendix D.3, we will show that our incomplete Variance estimator (3.12) is equivalent to the counterpart in Wang and Lindsay [31, page 1124]. Given B matching groups and M subsamples in each group, we denote the above estimators as $\hat{V}_u^{(inc)}$ and $\hat{V}_u^{(inc,W\&L)}$ respectively:

$$\hat{V}_u^{(inc)} := \hat{V}_{B,M}^{(h)} - \frac{MB-1}{MB}\hat{V}_{B,M}^{(s)}, \ \hat{V}_u^{(inc,W\&L)} := \tilde{\sigma}_{WP}^2/M - \tilde{\sigma}_{BP}^2, \tag{D.10}$$

where $\tilde{\sigma}_{WP}^2 := \frac{1}{(M-1)B} \sum_{b=1}^{B} \sum_{i=1}^{M} (h(S_i^{(b)}) - \tilde{h}_{(b)})^2$, $\tilde{\sigma}_{BP}^2 := \frac{1}{B} \sum_{b=1}^{B} (\tilde{h}_{(b)} - U_n)^2$, and $\tilde{h}_{(b)} = \frac{1}{M} \sum_{i=1}^{M} h(S_i^{(b)})$.

As analogues to A_1, A_2, A_3 (D.7) in Appendix D.3, we denote

$$\tilde{A}_1 := \frac{1}{MB} \sum_{b=1}^{B} \sum_{i=1}^{M} (h(S_i^{(b)}))^2, \ \tilde{A}_2 := U_{n,B,M}^2, \ \tilde{A}_3 := \frac{1}{B} \sum_{b=1}^{B} (\bar{h}_{(b)})^2.$$

Similarly, it is trivial to verify that $\hat{V}_{B,M}^{(h)}$ (3.9), $\hat{V}_{B,M}^{(s)}$ (3.10), $\tilde{\sigma}_{WP}^2$ and $\tilde{\sigma}_{BP}^2$ can be represented as linear combinations of $\tilde{A}_1, \tilde{A}_2$ and $\tilde{A}_3$ as follows:

$$\hat{V}_{B,M}^{(h)} = \tilde{\sigma}_{WP}^2 = \frac{M}{M-1}(\tilde{A}_1 - \tilde{A}_3), \ \hat{V}_{B,M}^{(s)} = \frac{MB}{MB-1}(\tilde{A}_1 - \tilde{A}_2), \ \sigma_{BP}^2 = \tilde{A}_3 - \tilde{A}_2.$$

By plugging the above into equation (D.10), we have

$$\hat{V}_u^{(inc)} = \hat{V}_{B,M}^{(h)} - \frac{MB-1}{MB}\hat{V}_{B,M}^{(s)} = \frac{1}{M-1}\tilde{A}_1 + \tilde{A}_2 - \frac{M}{M-1}\tilde{A}_3$$
$$\hat{V}_u^{(inc,W\&L)} = \tilde{\sigma}_{WP}^2/M - \tilde{\sigma}_{BP}^2 = \frac{1}{M-1}\tilde{A}_1 + \tilde{A}_2 - \frac{M}{M-1}\tilde{A}_3.$$

Hence, we conclude that $\hat{V}_u^{(inc)} = \hat{V}_u^{(inc,W\&L)}$.

Appendix E: Technical propositions and lemmas

In this section, we present the technical propositions and lemmas. The proofs of these results are collected in Appendix F.

Proposition E.1. *The value of $\psi\left(S^{(2k)}\right)$ does not depend on $\mathrm{E}\left[h(X_1,..,X_k)\right]$. Therefore, WLOG, we can assume the kernel is zero-mean, i.e., $\mathrm{E}\left[h(X_1,..,X_k)\right] = 0$*

The proof of this proposition is collected in Appendix F.1.

E.1. Results of $\sigma_{c,2k}^2$

First, we present Propositions E.2 and E.3. The former provides a precise bound of $\sigma_{c,2k}^2$ for some fixed c while the latter provides rough bound for $1 \leqslant c \leqslant 2k$.

Proposition E.2 (Bound $\sigma_{c,2k}^2$ for finite c)**.** *Fix $T_1 = \left\lfloor \frac{1}{\epsilon} \right\rfloor + 1$. Under Assumptions 1-5, for any c that $1 \leqslant c \leqslant T_1$,*

$$\sigma_{c,2k}^2 = \mathcal{O}\left(\frac{k^2}{n^2}F_c^{(k)}\right).$$

Based on the upper bound, we define $\check{\sigma}_{c,2k}^2 := \frac{Ck^2}{n^2}F_c^{(k)}$, where C is a generic positive constant. Equivalently, we write it as $\check{\sigma}_{c,2k}^2 \asymp \frac{k^2 F_c^{(k)}}{n^2}$.

Proposition E.3 (Bound $\sigma_{c,2k}^2$ for any c)**.** *Under Assumptions 1-5, for any $1 \leqslant c \leqslant 2k$, we have*

$$\sigma_{c,2k}^2 = \mathcal{O}(F_c^{(k)}).$$

The proof of the above propositions is collected in Appendix F.2 and Appendix F.3 respectively. Note that the upper bound in Proposition E.2 actually works for any fixed and finite c but it suffices to restrict $c \leqslant T_1$ to show our main results. These two propositions depend on the further decomposition of $\sigma_{c,2k}^2$ into weighted sum of $\eta_{c,2k}^2(d_1,d_2)$'s, which is later discussed in Appendix E.2. In particularly, we can show that $\eta_{c,2k}^2(1,1)$ dominates $\sigma_{c,2k}^2$ for $c = 1, 2, ..., T_1$. As a corollary of the above results, we can show the following lemma.

Lemma E.4 (Truncated Variance Lemma I). *Under Assumptions 1-5, there exists a constant $T_1 = \left\lfloor \frac{1}{\epsilon} \right\rfloor + 1$, such that*

$$\lim_{n\to\infty} \frac{\mathrm{Var}\left(\hat{V}_u\right) - \mathrm{Var}^{(T_1)}\left(\hat{V}_u\right)}{\widetilde{\mathrm{Var}}^{(T_1)}\left(\hat{V}_u\right)} = 0, \tag{E.1}$$

where

$$\mathrm{Var}^{(T_1)}\left(\hat{V}_u\right) := \binom{n}{2k}^{-1} \sum_{c=1}^{T_1} \binom{2k}{c}\binom{n-2k}{2k-c} \sigma^2_{c,2k}; \tag{E.2}$$

$$\widetilde{\mathrm{Var}}^{(T_1)}\left(\hat{V}_u\right) := \binom{n}{2k}^{-1} \sum_{c=1}^{T_1} \binom{2k}{c}\binom{n-2k}{2k-c} \breve{\sigma}^2_{c,2k}. \tag{E.3}$$

Here, we denote $\breve{\sigma}^2_{c,2k}$ as the upper bound of $\sigma^2_{c,2k}$ given by Proposition E.2.

The proof of Lemma E.4 is collected in Appendix F.4. This implies that to bound $\mathrm{Var}(\hat{V}_u)$, it suffices to bound the weighted average of first T_1 terms of $\sigma^2_{c,2k}$, instead of all $2k$ terms. Note that we use $\widetilde{\mathrm{Var}}$ instead of Var in the denominator of (E.1). Here $T_1 = \left\lfloor \frac{1}{\epsilon} \right\rfloor + 1$ does not grow with n. It only relies on ϵ, which quantifies the growth rate of k with respect to n (see Assumption 1,). For example, if $k = n^{1/3}$, i.e., $\epsilon = 1/6$, then we can choose $T_1 = 7$. Hence, to show $\sigma^2_{1,2k}$ dominates in $\mathrm{Var}(\hat{V}_u)$ when $n \to \infty$, it suffices to show that $\sigma^2_{1,2k}$ dominates in T_1-truncated $\mathrm{Var}^{(T_1)}(\hat{V}_u)$ when $n \to \infty$.

E.2. Results of $\eta^2_{c,2k}(d_1, d_2)$

Given the decomposition $\sigma^2_{c,2k} = \sum_{d_1=1}^{k}\sum_{d_2=1}^{k} w_{d_1} w_{d_2} \eta^2_{c,2k}(d_1, d_2)$ (see (4.10) in Proposition 4.3). To bound $\sigma^2_{c,2k}$, we should study $\eta^2_{c,2k}(d_1, d_2)$ (4.10). The results of $\eta^2_{c,2k}(d_1, d_2)$ are presented in this section.

Lemma E.5. *Under Assumptions 1-5, for $1 \leqslant c \leqslant T_1, 1 \leqslant d_1, d_2 \leqslant T_2$,*

$$\eta^2_{c,2k}(d_1, d_2) = \mathcal{O}\left(\frac{1}{k^2} F_c^{(k)}\right).$$

Lemma E.6. *Under Assumptions 1-5, for $c = 1, 2, ..., 2k,\ d_1, d_2 = 1, 2, ..., k$,*

$$\eta^2_{c,2k}(d_1, d_2) = \mathcal{O}(F_c^{(k)}).$$

Similar to the "two-type" upper bounds of $\sigma^2_{c,2k}$, Lemma E.5 provides a precise bound of $\eta^2_{c,2k}(d_1, d_2)$ for bounded c and d_1, d_2 while Lemma E.6 provides a rough bound of $\eta^2_{c,2k}(d_1, d_2)$ for all c, d_1, d_2. Again, the result of Lemma E.5 actually holds for any fixed and finite c, d_1, d_2. The proof of the above lemmas are collected in Appendix F.5 and Appendix F.6 respectively. The proof demonstrates the cancellation pattern by matching ρ (4.11).

With the above bounds on $\eta^2_{c,2k}(d_1,d_2)$, we introduce the following truncated $\sigma^2_{c,2k}$ and show Lemma E.8, which implies that to bound $\sigma^2_{c,2k}$, it suffices to bound the first finite $\eta^2_{c,2k}(d_1,d_2)$ terms in its decomposition (4.10).

Definition E.7 (Truncated $\sigma^2_{c,2k}$)**.** Let $T_2 = \lfloor \frac{1}{\epsilon} \rfloor + 1$. We define $\psi^{(T_2)}$, a T_2-truncated ψ as

$$\psi^{(T_2)}(S_1^{(2k)}) := \sum_{d=1}^{T_2} w_c \left(\varphi_d \left(S^{(2k)}\right) - \varphi_0 \left(S^{(2k)}\right)\right). \tag{E.4}$$

Hence, given two size-$2k$ subsamples $S_1^{(2k)}$ and $S_2^{(2k)}$ that $|S_1^{(2k)} \cap S_2^{(2k)}| = c$, a T_2-truncated of $\sigma^2_{c,2k}$ are defined as:

$$\sigma^2_{c,2k,(T_2)} := \mathrm{Cov}(\psi^{(T_2)}(S_1^{(2k)}), \psi^{(T_2)}(S_2^{(2k)})) = \sum_{d_1=1}^{T_2}\sum_{d_2=1}^{T_2} w_{d_1} w_{d_2} \eta^2_{c,2k}(d_1,d_2). \tag{E.5}$$

Lemma E.8 (Truncated Variance Lemma II)**.** *Under Assumptions 1-5, there exists a constant $T_2 = \lfloor \frac{1}{\epsilon} \rfloor + 1$, such that for any $c \leqslant T_1$,*

$$\lim_{k\to\infty} \frac{\sigma^2_{c,2k} - \sigma^2_{c,2k,(T_2)}}{\breve{\sigma}^2_{c,2k}} = 0,$$

where $\breve{\sigma}^2_{c,2k,(T_2)}$ is the upper bound of $\sigma^2_{c,2k,(T_2)}$ given in Proposition E.2.

The proof is collected in Appendix F.7. Similar to the idea of Lemma E.4, by Lemma E.8 the upper bound of $\sigma^2_{c,2k}$ (4.10) only involves the sum of T_2^2 terms, i.e., $\sigma^2_{c,2k,(T_2)}$ rather than k^2 terms. Here T_2 is again finite and does not grow with k. Though T_1 and T_2 take the same value, we note that T_1 is the truncation constant for $\mathrm{Var}(\hat{V}_u)$ in (E.2) while T_2 is the truncation constant for $\sigma^2_{c,2k}$ in (E.5).

Appendix F: Proof of technical propositions and lemmas

F.1. Proof of Proposition E.1

Proof of Proposition E.1. Suppose $\mathrm{E}(h(S_1)) = \mu$ and we rewrite $h(S_1) = h_{(0)}(S_1) + \mu$, where $\mathrm{E}(h_{(0)}(S_1)) = 0$. Then $\varphi_d(S^{(2k)})$ defined in (4.6) can be written as

$$\begin{aligned}\varphi_d\left(S^{(2k)}\right) = \frac{1}{M_{d,k}} &\sum_{S_1,S_2 \subset S^{(2k)}, |S_1\cap S_2| = d} h_{(0)}(S_1)h_{(0)}(S_2) \\ &+ [h_{(0)}(S_1) + h_{(0)}(S_2)]\mu + \mu^2.\end{aligned} \tag{F.1}$$

Plug Equation (F.1) into Equations (4.1) and (4.5). By the fact that $\sum_{d=0}^k w_d = 0$ and the symmetry of U-statistic, the terms of μ and μ^2 are cancelled. Consequently, $\hat{V}_u$ does not depend on μ, so W.L.O G., we assume that $\mu = 0$ $\square$

F.2. Proof of Proposition E.2

Proof of Proposition E.2. This proof relies on the technical lemmas in Appendix E.2. First, by Lemma E.8, to upper bound $\sigma^2_{c,2k}$, it suffices to upper bound the following $\sigma^2_{c,2k,(T_2)}$ (E.5)

$$\sigma^2_{c,2k,(T_2)} = \sum_{d_1=1}^{T_2}\sum_{d_2=1}^{T_2} w_{d_1}w_{d_2}\eta^2_{c,2k}(d_1,d_2),$$

where $w_d = [1+o(1)]\left[\frac{1}{d!}(\frac{k^2}{n})^d\right], \forall d \leqslant T_2$.

By Lemma E.5: fixing any c s.t. $1 \leqslant c \leqslant T_2$, $\eta^2_{c,2k}(d_1,d_2) = \mathcal{O}(\frac{F_c^{(k)}}{k^2}), \forall d_1, d_2 \leqslant T_2$. Besides, since d_1 and d_2 are bounded by a constant T_2. w_1^2 dominates the summation $\sum_{d_1=1}^{T_2}\sum_{d_2=1}^{T_2} w_{d_1}w_{d_2}$. We have

$$\begin{aligned}\sigma^2_{c,2k,(T_2)} &= \sum_{d_1=1}^{T_2}\sum_{d_2=1}^{T_2} w_{d_1}w_{d_2}\eta^2_{c,2k}(d_1,d_2) = \mathcal{O}(\frac{F_c^{(k)}}{k^2})\sum_{d_1=1}^{T_2}\sum_{d_2=1}^{T_2} w_{d_1}w_{d_2}\\ &= \mathcal{O}(\frac{F_c^{(k)}}{k^2})[1+o(1)]w_1^2 = \mathcal{O}(\frac{k^2}{n^2})F_c^{(k)}.\end{aligned}$$

Here, the last equality is derived by plugging in $w_1 = [1+o(1)]\frac{k^2}{n}$. □

F.3. Proof of Proposition E.3

Proof of Proposition E.3. This lemma again relies on the upper bound of $\eta^2_{c,2k}(d_1,d_2)$ in Appendix E.2. By Proposition 4.3, we can decompose $\sigma^2_{c,2k}$ as

$$\sigma^2_{c,2k} = \sum_{d_1=1}^{k}\sum_{d_2=1}^{k} w_{d_1}w_{d_2}\eta^2_{c,2k}(d_1,d_2),$$

First, we investigate the coefficient of $\eta^2_{c,2k}(d_1,d_2)$. By Proposition 4.2, we have $\sum_{d_1=1}^{k} w_{d_1} < 1$. Thus, we attain

$$\sum_{d_1=1}^{k}\sum_{d_2=1}^{k} w_{d_1}w_{d_2} = (\sum_{d_1=1}^{k} w_{d_1})(\sum_{d_2=1}^{k} w_{d_2}) < 1.$$

By Lemma E.6, we have $\eta^2_{c,2k}(d_1,d_2) = \mathcal{O}(F_c)$ for $c = 1,2,3,...,2k$. Hence, combining the bounds on w_d and $\eta^2_{c,2k}(d_1,d_2)$, we conclude that $\sigma^2_{c,2k} = \mathcal{O}(F_c^{(k)})$. We remark that the summation $\sum_{d_1=1}^{k} w_{d1}$ can attain a lower order of 1, which may imply a tighter bound of $\sigma^2_{c,2k}$. □

F.4. Proof of Lemma E.4

Proof. Let $T_1 = \lfloor \frac{1}{\epsilon} \rfloor + 1$. Recall Equation (4.3)

$$\mathrm{Var}\left(\hat{V}_u\right) = \binom{n}{2k}^{-1} \sum_{c=1}^{2k} \binom{2k}{c} \binom{n-2k}{2k-c} \sigma_{c,2k}^2.$$

We first present the intuition of this lemma. $\mathrm{Var}(\hat{V}_u)$ is a weighted sum of $\sigma_{c,2k}^2$, where the coefficient of $\sigma_{c,2k}^2$ decays with c at a rate even faster than a geometric rate. If the growth rate of $\sigma_{c,2k}^2$ is not too fast, then the tail terms can be negligible. This involves both the precise upper bound of $\sigma_{1,2k^2}$ (Proposition E.2) and the rough upper bound of $\sigma_{c,2k}^2$ for $c \geqslant T_1 + 1$ (Proposition E.3).

First, $\binom{n}{2k}^{-1}\binom{2k}{1}\binom{n-2k}{2k-1}\breve{\sigma}_{1,2k}^2 \leqslant \widetilde{\mathrm{Var}}^{(T_1)}\left(\hat{V}_u\right)$ since the former is the first tern in the latter and all the other terms are positive. Therefore, it suffices to show

$$\frac{\mathrm{Var}\left(\hat{V}_u\right) - \mathrm{Var}^{(T_1)}\left(\hat{V}_u\right)}{\binom{n}{2k}^{-1}\binom{2k}{1}\binom{n-2k}{2k-1}\breve{\sigma}_{1,2k}^2} = \frac{\sum_{c=T_1+1}^{2k}\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}\sigma_{c,2k}^2}{\binom{n}{2k}^{-1}\binom{2k}{1}\binom{n-2k}{2k-1}\breve{\sigma}_{1,2k}^2} \to 0. \tag{F.2}$$

We bound the numerator and denominator in Equation (F.2) separately. For the denominator, by the analysis of Equation (C.12), we have

$$\binom{n}{2k}^{-1}\binom{2k}{1}\binom{n-2k}{2k-1}\breve{\sigma}_{1,2k}^2 = \frac{[1+o(1)]4k^2}{n}\breve{\sigma}_{1,2k}^2. \tag{F.3}$$

For the numerator, by Proposition E.3 and assumption 5, $\sigma_{c,2k}^2 = \mathcal{O}(F_c^{(k)}) = o(c^{a_2}F_1^{(k)})$. Therefore, it suffices to show that

$$\begin{aligned}&\frac{\sum_{c=T_1+1}^{2k}\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}o(c^{a_2}F_1^{(k)})}{\frac{4k^2}{n}\breve{\sigma}_{1,2k}^2}\\ &= \frac{\mathcal{O}\left(\{\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}o(c^{a_2}F_1^{(k)})\}_{c=T_1+1}\right)}{\frac{4k^2}{n}\breve{\sigma}_{1,2k}^2} \xrightarrow{n\to\infty} 0.\end{aligned} \tag{F.4}$$

The equality in (F.4) is given by Proposition F.1. The followed $\xrightarrow{n\to\infty} 0$ in (F.4) is given by Proposition F.2. This completes the proof. ☐

Proposition F.1. *Under Assumptions 1-5,*

$$\frac{\sum_{c=T_1+2}^{2k}\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}c^{a_2}F_1^{(k)}}{\binom{n}{2k}^{-1}\binom{2k}{T_1+1}\binom{n-2k}{2k-T_1-1}(T_1+1)^{a_2}F_1^{(k)}} \to 0, \textit{ as } n\to\infty \tag{F.5}$$

.

Proof. The proof of Proposition F.1 is similar to the proof of Proposition 4.5. The idea is that the sum of tail coefficients is a geometric sum and thus dominates the growth rate of moments.

First we consider the coefficient $\frac{\sum_{c=T_1+2}^{2k}\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}}{\binom{n}{2k}^{-1}\binom{2k}{T_1+1}\binom{n-2k}{2k-T_1-1}}$. By Proposition C.1 and our analysis in Equation (C.13) and (C.14), let $b_n = \frac{4k^2}{n-2k+1}$ which is the common ratio in the geometric sequence.

$$\frac{\sum_{c=T_1+2}^{2k}\binom{n}{2k}^{-1}\binom{2k}{c}\binom{n-2k}{2k-c}}{\binom{n}{2k}^{-1}\binom{2k}{T_1+1}\binom{n-2k}{2k-T_1-1}} \leqslant \sum_{c=T_1+2}^{2k}[1+o(1)]\frac{(T_1+1)!}{c!}b_n^{c-(T_1+1)}. \tag{F.6}$$

Second, combining c^{a_1} with (F.6), it's again the problem of geometric series with common ratio $b_n = o(1)$. We have

$$\begin{aligned}\text{LHS of (F.5)} &\leqslant \sum_{c=T_1+2}^{2k}\mathcal{O}\left(\frac{c^{a_2}}{c!}\right)b_n^{c-(T_1+1)}\\ &\leqslant \sum_{c=T_1+2}^{2k}\mathcal{O}(1)b_n^{c-(T_1+1)}\\ &\leqslant \sum_{c=1}^{\infty}\mathcal{O}(1)b_n^c = \mathcal{O}(1)b_n \to 0,\end{aligned} \tag{F.7}$$

where the last equality is concluded by the sum of geometric series. □

Proposition F.2. *Under Assumptions 1-5,*

$$\frac{\binom{n}{2k}^{-1}\binom{2k}{T_1+1}\binom{n-2k}{2k-T_1-1}C'(T_1+1)^{a_2}F_1^{(k)}}{\frac{k^2}{n}\breve{\sigma}_{1,2k}^2} = o(\frac{1}{k^2}) \to 0. \tag{F.8}$$

In Equation (F.8), the upper bound of $c = T_1 + 1$ term of the numerator is a lower order term compared to the denominator.

Proof. It suffices bound two separate parts in Equation (F.8),

$$\frac{\binom{n}{2k}^{-1}\binom{2k}{T_1+1}\binom{n-2k}{2k-T_1-1}}{\frac{k^2}{n}} = o(\frac{1}{n^2}), \tag{F.9}$$

$$\frac{C'(T_1+1)^{a_2}F_1^{(k)}}{\breve{\sigma}_{1,2k}^2} = \mathcal{O}(\frac{n^2}{k^2}). \tag{F.10}$$

Then, combining Equation (F.9) and (F.10), we have

$$\text{LHS of (F.8)} \leqslant o\left(\frac{1}{n^2}\right)\mathcal{O}(\frac{n^2}{k^2}) = o\left(\frac{1}{k^2}\right) \to 0.$$

We first show Equation (F.9), i.e., bound the ratio of coefficient. Similar to the analysis for Equation (C.14), by Proposition C.1, we have

$$\begin{aligned}\binom{n}{2k}^{-1}\binom{2k}{T_1+1}\binom{n-2k}{2k-T_1-1} &\leqslant \frac{(2k)^{2(T_1+1)}}{(T_1+1)!}\frac{1}{(n-2k+1)^{T_1+1}}\\ &\leqslant \frac{1}{(T_1+1)!}b_n^{T_1+1},\end{aligned}$$

where $b_n = \frac{4k^2}{n-2k+1}$. Therefore the ratio of coefficient,

$$\frac{\binom{n}{2k}^{-1}\binom{2k}{T_1+1}\binom{n-2k}{2k-T_1-1}}{\frac{4k^2}{n}} \leqslant [1+o(1)]\frac{1}{(T_1+1)!}b_n^{T_1}.$$

It remains to show $b_n^{T_1} = o(\frac{1}{n^2})$. By $T_1 = \lfloor\frac{1}{\epsilon}\rfloor + 1$ and $k = \mathcal{O}(n^{1/2-\epsilon})$, we have

$$b_n^{T_1} \leqslant (\frac{4k^2}{n})^{\lfloor 1/\epsilon\rfloor+1} = \mathcal{O}((n^{-2\epsilon})^{\lfloor 1/\epsilon\rfloor+1}) = o(\frac{1}{n^2}). \tag{F.11}$$

Second, we show (F.10), i.e., bound the ratio of moments. By Lemma E.8 and E.2, $\sigma^2_{1,2k} = \mathcal{O}(k^2/n^2F_1^{(k)})$ and $\breve{\sigma}^2_{1,2k} \asymp k^2/n^2F_1^{(k)}$. We have

$$\frac{C'(T_1+1)^{a_2}F_1^{(k)}}{\breve{\sigma}^2_{1,2k}} = \mathcal{O}\left(n^2/k^2(T_1+1)^{a_2}\right) = \mathcal{O}(\frac{n^2}{k^2}). \tag{F.12}$$

□

F.5. Proof of Lemma E.5

Proof of Lemma E.5. First, we present a sketch of this proof. Given $S_1^{(2k)}$ and $S_2^{(2k)}$, our strategy tracks the distribution of *Influential Overlaps*, i.e., the samples in $S_1^{(2k)} \cap S_2^{(2k)}$. We will decompose $\eta^2_{c,2k}(d_1,d_2)$ as a finite weighted sum:

$$\eta^2_{c,2k}(d_1,d_2) = \sum_i a_i b_i, \tag{F.13}$$

where i is the summation index to be specified later. Based on this form, we will show $a_i = \mathcal{O}(\frac{1}{k^2})$ and $b_i = \mathcal{O}(F_c^{(k)})$ for each i. Since (F.13) is a finite sum, we can conclude $\eta^2_{c,2k}(d_1,d_2) = \mathcal{O}\left(\frac{1}{k^2}F_c^{(k)}\right)$. Details of (F.13) will be presented later. We remark that it is straightforward to upper bound $\eta^2_{1,2k}(d_1,d_2)$ by enumerating all the possible 4-way overlapping cases of S_1, S_2, S_3, S_4 given c, d_1, d_2 for small c. However, the growth of c from 1 to $2,3,4,...,T_2$ makes "enumerating" impossible.

We start the proof by reviewing the definition of $\eta^2_{c,2k}(d_1,d_2)$ (4.9):

$$\eta^2_{c,2k}(d_1,d_2) = \mathrm{Cov}\left[\varphi_{d_1}\left(S_1^{(2k)}\right) - \varphi_0\left(S_1^{(2k)}\right), \varphi_{d_2}\left(S_2^{(2k)}\right) - \varphi_0\left(S_2^{(2k)}\right)\right],$$

where $\varphi_d\left(S^{(2k)}\right) = [\binom{2k}{d}\binom{2k-d}{d}\binom{2k-2d}{k-d}]^{-1}\sum_{S_1,S_2\subset S^{(2k)},|S_1\cap S_2|=d} h(S_1)h(S_2)$ (4.6). The following proof is organized in two parts. First, we propose an alternative representation of the covariance $\mathrm{Cov}[\varphi_{d_1}(S_1^{(2k)}),\varphi_{d_2}(S_2^{(2k)})]$, which helps discover the cancellation pattern of $\eta^2_{c,2k}(d_1,d_2)$ (4.9). Secondly, we derive Equation (F.13) and specify a_i's and b_i's.

First, we notice that $\varphi_d\left(S^{(2k)}\right)$ (4.6) is a weighted average of the product of two kernels $h(S_1)h(S_2)$:

$$\varphi_d\left(S^{(2k)}\right) = \left[\binom{2k}{d}\binom{2k-d}{d}\binom{2k-2d}{k-d}\right]^{-1}\sum_{S_1,S_2\subset S^{(2k)},|S_1\cap S_2|=d} h(S_1)h(S_2).$$

Denote $\sum_{P_{12}}$ as summation over all pairs of S_1,S_2, s.t. $S_1,S_2\subset S^{(2k)}$, $|S_1\cap S_2| = d_2$. Similarly, we can also denote $\sum_{P_{34}}$. Then, we can represent the covariance $\mathrm{Cov}[\varphi_{d_1}(S_1^{(2k)}),\varphi_{d_2}(S_2^{(2k)})]$ as

$$\begin{aligned}\mathrm{Cov}\left[\varphi_{d_1}(S_1^{(2k)}),\varphi_{d_2}(S_2^{(2k)})\right] &= \mathrm{Cov}\left[M_{d_1}^{-1}\sum_{P_{12}} h(S_1)h(S_2), M_{d_2}^{-1}\sum_{P_{34}} h(S_3)h(S_4)\right] &\text{(F.14)}\\ &= M_{d_1}^{-1}M_{d_2}^{-1}\sum_{P_{12}}\sum_{P_{34}}\rho &\text{(F.15)}\\ &= \sum_{\text{feasible }\underline{r}} p_{(\underline{r},d_1,d_2,c)}\rho. &\text{(F.16)}\end{aligned}$$

In the above equations,

$$p_{(\underline{r},d_1,d_2,c)} := M_{d_1}^{-1}M_{d_2}^{-1}\sum_{(S_1,S_2,S_3,S_4)} \mathbb{1}\{\text{overlapping structure of } S_1,S_2,S_3,S_4 \text{ satisfies } (\underline{r},d_1,d_2)\};$$

$M_d = \binom{2k}{d}\binom{2k-d}{d}\binom{2k-2d}{k-d}$ is the number of pairs of sets in the summation. The equality in (F.16) holds by combining the ρ terms with the same $\underline{r}$ (see definition of $\underline{r}$ in Appendix B.1). Since it is difficult to figure out the exact value of $p_{(\underline{r},d_1,d_2,c)}$, we further propose the following proposition to show an alternative representation of $\mathrm{Cov}[\varphi_{d_1}(S_1^{(2k)}),\varphi_{d_2}(S_2^{(2k)})]$.

Lemma F.3. *Denote $r_{i*} = \sum_{j=0}^2 r_{ij}$ and $r_{*j} = \sum_{i=0}^2 r_{ij}$, for $i,j = 0,1,2$ and vector $\underline{r}^* := (r_{0*},r_{1*},r_{2*},r_{*0},r_{*1},r_{*2})$, we have*

$$\mathrm{Cov}\left[\varphi_{d_1}\left(S_1^{(2k)}\right),\varphi_{d_2}\left(S_2^{(2k)}\right)\right] = \sum_{\textit{feasible }\underline{r}^*} p_{(r_{0*},r_{1*},r_{2*},d_1,c)}p_{(r_{*0},r_{*1},r_{*2},d_2,c)}g(\underline{r}^*,d_1,d_2). \tag{F.17}$$

Here $p_{(r_{0*},r_{1*},r_{2*},d_1,c)}$ and $p_{(r_{*0},r_{*1},r_{*2},d_2,c)}$ are non-negative and satisfy the following. For non-negative integers x_0,x_1,x_2 that $x_0+x_1+x_2\leqslant c$,

$$p_{(x_0,x_1,x_2,d,c)} := \binom{d}{x_0}\binom{k-d}{x_1}\binom{k-d}{x_2}\binom{d}{c-x_0-x_1-x_2}\binom{2k}{c}^{-1} = \mathcal{O}(k^{x_1+x_2-c}). \tag{F.18}$$

$g(\underline{r}^*, d_1, d_2)$ *is the following weighted average of* ρ, *where the weight is some constant* $p_{(\underline{r},\underline{r}^*)}$ *satisfying that* $\sum_{\underline{r}^*} p_{(\underline{r},\underline{r}^*)} = 1$.

$$g(\underline{r}^*, d_1, d_2) := \sum_{\underline{r}} p_{(\underline{r},\underline{r}^*)}\rho. \tag{F.19}$$

The proof of Lemma F.3 is deferred to the end of Appendix F.5. Under Assumption 3, ρ does not depend on d_1, d_2. Hence, $g(\underline{r}^*, d_1, d_2)$ also does not depend on d_1, d_2. We further denote $G(\underline{r}^*) = g(\underline{r}^*, d_1, d_2)$. Then,

$$\mathrm{Cov}\left[\varphi_{d_1}\left(S_1^{(2k)}\right), \varphi_{d_2}\left(S_2^{(2k)}\right)\right] = \sum_{\text{feasible } \underline{r}^*} p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} G(\underline{r}^*). \tag{F.20}$$

We remark that $p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)}$ and $p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)}$ can be viewed as some probability mass function with parameters c, d_1, d_2 (see the proof of Lemma F.3). As a corollary of (F.18), when $c - x_1 - x_2 > d$, $p_{(x_0, x_1, x_2, d, c)} = 0$. In particular, when $d = 0$ and $x_1 + x_2 \leqslant c - 1$, $p_{(x_0, x_1, x_2, d, c)}$ is always 0.

Notice that since $S_1^{(2k)}$ is independent of $S_2^{(2k)}$, $\sum_{\text{feasible } \underline{r}^*}$ can be written as two sequential sums: $\sum_{(r_{0*}, r_{1*}, r_{2*})} \sum_{(r_{*0}, r_{*1}, r_{*2})}$. Therefore, by plugging the expression of $\mathrm{Cov}[\varphi_{d_1}(S_1^{(2k)}), \varphi_{d_2}(S_2^{(2k)})]$ (F.20) into $\eta^2_{c,2k}(d_1, d_2)$ (4.9), we have

$$\begin{aligned} \eta^2_{c,2k}(d_1, d_2) &= \mathrm{Cov}\left[\varphi_{d_1}\left(S_1^{(2k)}\right) - \varphi_0\left(S_1^{(2k)}\right), \varphi_{d_2}\left(S_2^{(2k)}\right) - \varphi_0\left(S_2^{(2k)}\right)\right] \\ &= \sum_{\text{feasible } \underline{r}^*} \underbrace{\left[p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} - p_{(r_{0*}, r_{1*}, r_{2*}, 0, c)}\right]\left[p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} - p_{(r_{*0}, r_{*1}, r_{*2}, 0, c)}\right]}_{a_i} \\ &\quad \times \underbrace{G(\underline{r}^*)}_{b_i} \end{aligned} \tag{F.21}$$

We have two observations on the above Equation (F.21). First, this $\eta^2_{c,2k}(d_1, d_2) = \sum_i a_i b_i$ is a finite summation because $c \leqslant T_1$. Hence, to show $\eta^2_{c,2k} = \mathcal{O}(F_c^{(k)}/k^2)$, it suffices to bound every term $a_i \cdot b_i$. Secondly, by Lemma F.3, $b_i = G(\underline{r}^*)$ is a weighted average of ρ where the non-negative weights $\sum_{\underline{r}^*} p_{(\underline{r},\underline{r}^*)} = 1$. Hence, each b_i is naturally bounded by the upper bound of ρ. We conclude that $b_i = \mathcal{O}(F_c^{(k)})$. Therefore, it remains to show that $a_i = \mathcal{O}(k^{-2})$ for every i. This is provided by the following lemma.

Lemma F.4. *Fixing integer* $d, c \geqslant 0$, *for any tuple of non-negative integers* (x_0, x_1, x_2) *s.t.* $\sum_{i=0}^2 x_i \leqslant c$,

$$p_{(x_0, x_1, x_2, d, c)} - p_{(x_0, x_1, x_2, 0, c)} = \mathcal{O}(\frac{1}{k}). \tag{F.22}$$

The proof is collected later in Appendix F.5. This completes the proof of Lemma E.5. We remark that though there exists $p_{(x_0, x_1, x_2, d, c)} \asymp 1$ for some (x_0, x_1, x_2), $p_{(x_0, x_1, x_2, d, c)} - p_{(x_0, x_1, x_2, 0, c)}$ is always at the order of $\mathcal{O}(\frac{1}{k})$. $\square$

Remark F.5. This proof has proceeded under Assumption 3. It can be adapted to a weaker assumption: Assumption 6.The according proof using this new assumption will be present in Appendix G, where we cannot exactly cancel two ρ with the same $\underline{r}$ but different d.

We present the proof of two important technical facts in the above proof: Lemma F.3 and Lemma F.4.

Proof of Lemma F.3. First, we derive Equation (F.17) from Equation (F.16):

$$\mathrm{Cov}\left[\varphi_{d_1}(S_1^{(2k)}), \varphi_{d_2}(S_2^{(2k)})\right] = \sum_{\text{feasible } \underline{r}} p_{(\underline{r},d_1,d_2,c)}\rho.$$

Given that $c = |S_1^{(2k)}, S_2^{(2k)}|$, $d_1 = |S_1 \cap S_2|$ and $d_2 = |S_3 \cap S_4|$, suppose we randomly sample a feasible S_1, S_2, S_3, S_4 from all possible cases, we can use a 9-dimension random variable $\mathbf{R}$ to denote the 4-way overlapping structure of S_1, S_2, S_3, S_4. Hence, the the coefficient $p_{(\underline{r},d_1,d_2,c)}$ in (F.16) is $P(\mathbf{R} = \underline{r}|d_1, d_2, c)$. Then, denote a 6-dimension random variable $\mathbf{R}^* = (\mathbf{R}_{0*}, \mathbf{R}_{1*}, \mathbf{R}_{2*}, \mathbf{R}_{*0}, \mathbf{R}_{*1}, \mathbf{R}_{*2})$, taking all possible values of $\underline{r}^*$ given d_1, d_2, c. By Bayesian rule,

$$P(\mathbf{R} = \underline{r}|d_1, d_2, c) = P(\mathbf{R} = \underline{r}|\mathbf{R}^* = \underline{r}^*, d_1, d_2, c)P(\mathbf{R}^* = \underline{r}^*|d_1, d_2, c). \quad \text{(F.23)}$$

Since $S_1, S_2 \subset S_1^{(2k)}$, $S_3, S_4 \subset S_2^{(2k)}$ and $S_1^{(2k)}$ is independent from $S_2^{(2k)}$, $(\mathbf{R}_{0*}, \mathbf{R}_{1*}, \mathbf{R}_{2*})$ are independent from $(\mathbf{R}_{*0}, \mathbf{R}_{*1}, \mathbf{R}_{*2})$). Hence, we can further decompose $P(\mathbf{R}^* = \underline{r}^*|d_1, d_2, c)$ as $P(\mathbf{R}_{0*} = r_{0*}, \mathbf{R}_{1*} = r_{1*}, \mathbf{R}_{2*} = r_{2*})|d_1, c) \cdot P(\mathbf{R}_{*0} = r_{*0}, \mathbf{R}_{*1} = r_{*1}, \mathbf{R}_{*2} = r_{*2})|d_2, c)$. To simplify the notations, we denote

$$\begin{aligned}
&p_{(\underline{r},d_1,d_2,c)} := P(\mathbf{R}^* = \underline{r}^*|d_1, d_2, c),\\
&p_{(\underline{r},\,\underline{r}^*)} := P(\mathbf{R} = \underline{r}|\mathbf{R}^* = \underline{r}^*, d_1, d_2, c),\\
&p_{(r_{0*},r_{1*},r_{2*},d_1,c)} := P(\mathbf{R}_{0*} = r_{0*}, \mathbf{R}_{1*} = r_{1*}, \mathbf{R}_{2*} = r_{2*})|d_1, c),\\
&p_{(r_{*0},r_{*1},r_{*2},d_2,c)} := P(\mathbf{R}_{*0} = r_{*0}, \mathbf{R}_{*1} = r_{*1}, \mathbf{R}_{*2} = r_{*2})|d_2, c).
\end{aligned}$$

Given $\mathbf{R}^*$, the distribution of $\mathbf{R}$ does not depend on d_1, d_2, c so we omit the subscript d_1, d_2, c in $p_{(\underline{r},\,\underline{r}^*)}$. We also remark that $\sum_{\underline{r}^*} p_{(\underline{r},\,\underline{r}^*)} = 1$ since $\mathbf{R}|\mathbf{R}^*$ can be viewed as a random variable. Based on these notations and (F.23), we can rewrite Equation (F.16) as

$$\begin{aligned}
&\sum_{\text{feasible } \underline{r}^*}\ \sum_{\text{feasible } \underline{r}} p_{(\underline{r},\,\underline{r}^*)} p_{(r_{0*},r_{1*},r_{2*},d_1,c)} p_{(r_{*0},r_{*1},r_{*2},d_2,c)}\rho\\
&= \sum_{\text{feasible } \underline{r}^*} p_{(r_{0*},r_{1*},r_{2*},d_1,c)} p_{(r_{*0},r_{*1},r_{*2},d_2,c)} \underbrace{\sum_{\text{feasible } \underline{r}} p_{(\underline{r},\,\underline{r}^*)}\rho]}_{\text{denote as } g(\underline{r}^*, d_1, d_2)} .
\end{aligned}$$

This justifies both Equations (F.17) and (F.19).

Secondly, we show Equation (F.18). Since $p_{(r_{0*},r_{1*},r_{2*},d_1,c)}$ and $p_{(r_{*0},r_{*1},r_{*2},d_2,c)}$ can be analyzed in the same way, our discussion focuses on $p_{(r_{0*},r_{1*},r_{2*},d_1,c)}$, which is boiled down to the distribution of $(\mathbf{R}_{0*}, \mathbf{R}_{1*}, \mathbf{R}_{2*})$. Given c and d_1, c *Influential Overlaps* can fall into 4 different "boxes" in $S_1^{(2k)}$: $S_1 \cap S_2$, $S_1 \backslash S_2$, $S_2 \backslash S_2$, and $S_1^{(2k)} \backslash (S_1 \cup S_2)$, with "box size" as d_1, $k - d_1$, $k - d_1$, d_1 respectively. The number of samples in each "box" follows a hypergeometric distribution. This is illustrated by the following table.

TABLE 5
The distribution of Influential Overlaps *in* $S_1^{(2k)}$.

Index	0	1	2	3
"box"	$S_1 \cap S_2$	$S_1 \backslash S_2$	$S_2 \backslash S_1$	$S_1^{(2k)} \backslash (S_1 \cup S_2)$
"box size"	d_1	$k - d_1$	$k - d_1$	d_1
# of *Influential Overlaps*	r_{0*}	r_{1*}	r_{2*}	$c - \lvert \underline{r} \rvert$

Hence, the probability mass function of $(\mathbf{R}_{0*}, \mathbf{R}_{1*}, \mathbf{R}_{2*})$: $P(\mathbf{R}_{0*} = r_{0*}, \mathbf{R}_{1*} = r_{1*}, \mathbf{R}_{2*} = r_{2*} | d_1, c)$ is

$$p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} = \binom{d_1}{r_{0*}} \binom{k - d_1}{r_{1*}} \binom{k - d_1}{r_{2*}} \binom{d_1}{c - r_{0*} - r_{1*} - r_{2*}} \binom{2k}{c}^{-1}. \tag{F.24}$$

It remains to show that $p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} = \mathcal{O}(k^{r_{1*} + r_{2*} - c})$ for any fixed c, d_1. In the following, to simplify the notation, we denote $x_i = r_{i*}$ for $i = 0, 1, 2$ and $x_3 = c - r_{0*} - r_{1*} - r_{1*}$. Then Equation (F.18) can be written as

$$\frac{c!}{x_0! x_1! x_2! x_3!} \frac{d_1!}{(d_1 - x_0)!} \frac{(k - d_1)!}{(k - d_1 - x_1)!} \frac{(k - d_1)!}{(k - d_1 - x_2)!} \frac{d_1!}{(d_1 - x_3)!} \Big/ \frac{(2k)!}{(2k - c)!}. \tag{F.25}$$

Before the formal justification, we remark that (F.25) looks similar to the probability mass function of a multinomial distribution: $\frac{c!}{x_0! x_1! x_2! x_3!} \left(\frac{d_1}{2k}\right)^{x_0} \left(\frac{k - d_1}{2k}\right)^{x_1} \left(\frac{k - d_1}{2k}\right)^{x_2} \left(\frac{d_1}{2k}\right)^{x_3}$, which is obviously $\mathcal{O}(k^{x_1 + x_2 - c})$.

We decompose Equation (F.25) as a production of three parts, denoting Part I $:= \frac{c!}{x_0! x_1! x_2! x_3!}$, Part II $:= \frac{d_1!}{(d_1 - x_0)!} \frac{(k - d_1)!}{(k - d_1 - x_1)!} \frac{(k - d_1)!}{(k - d_1 - x_2)!} \frac{d_1!}{(d_1 - x_3)!}$, Part III $:= \frac{(2k)!}{(2k - c)!}$. Since x_0, x_1, x_2, x_3, c are finite, Part I can be viewed as a constant in the asymptotic analysis. For Part II, again, $\frac{d_1!}{(d_1 - x_0)!} \frac{d_1!}{(d_1 - x_3)!}$ does not depend on k and thus can be treated as a constant. For the rest part:

$$\begin{aligned} \frac{(k - d_1)!}{(k - d_1 - x_1)!} \frac{(k - d_1)!}{(k - d_1 - x_2)!} &= \left[\prod_{i=0}^{x_1 - 1} (k - d_1 - i)\right] \left[\prod_{i=0}^{x_2 - 1} (k - d_1 - i)\right] \\ &\leqslant (k - d_1)^{x_1 + x_2}. \end{aligned}$$

For Part III,

$$\frac{(2k)!}{(2k - c)!} = \prod_{i=0}^{c-1} (2k - i) \geqslant k^c.$$

Combining Part I, II, III, we have

$$p_{(x_0,x_1,x_2,d_1,c)} = \mathcal{O}\left[(k-d_1)^{x_1+x_2}/k^c\right] = \mathcal{O}(k^{-c+x_1+x_2}).$$

This completes the proof. □

Proof of Lemma F.4. To show $p_{(x_0,x_1,x_2,d,c)} - p_{(x_0,x_1,x_2,0,c)} = \mathcal{O}(\frac{1}{k})$, we study two cases separately, where case I is $x_1+x_2 \leqslant c-1$ and case II is $x_1+x_2 = c$. This is motivated by the conclusion of Proposition F.3, $p_{(x_0,x_1,x_2,d,c)} = \mathcal{O}(k^{x_1+x_2-c})$.

We first study case I. For any finite c, d, since $x_1 + x_2 \leqslant c-1$, by (F.18), $p_{(x_0,x_1,x_2,d,c)} = \mathcal{O}(1/k)$. In particular, $p_{(x_0,x_1,x_2,0,c)} = 0$. Therefore,

$$p_{(x_0,x_1,x_2,d,c)} - p_{(x_0,x_1,x_2,0,c)} = \mathcal{O}(\frac{1}{k}) - 0 = \mathcal{O}(\frac{1}{k}).$$

Secondly, we study case II. For any finite c, d, since $x_1+x_2 = c$, $p_{(x_0,x_1,x_2,d,c)} \asymp 1$ and $p_{(x_0,x_1,x_2,0,c)} \asymp 1$. Hence, we can not conclude the order of $p_{(x_0,x_1,x_2,d,c)} - p_{(x_0,x_1,x_2,0,c)}$ directly from the order of each term. We need to study $p_{(x_0,x_1,x_2,d,c)} = \binom{d}{x_0}\binom{k-d}{x_1}\binom{k-d}{x_2}\binom{d_1}{c-x_0-x_1-x_2}\binom{2k}{c}^{-1}$ a bit more carefully. It is equivalent to showing that

$$\begin{aligned}\left[p_{(x_0,x_1,x_2,d,c)} - p_{(x_0,x_1,x_2,0,c)}\right]/p_{(x_0,x_1,x_2,d,c)} &= p_{(x_0,x_1,x_2,d,c)}/p_{(x_0,x_1,x_2,d,c)} - 1\\ &= \mathcal{O}(\frac{1}{k}).\end{aligned}$$

To prove the above, we denote $q(d) := p_{(x_0,x_1,x_2,d,c)}/p_{(x_0,x_1,x_2,0,c)}$. It suffices to show that

$$q(d) = p_{(x_0,x_1,x_2,d,c)}/p_{(x_0,x_1,x_2,0,c)} = 1 + \mathcal{O}(\frac{1}{k}).$$

Since $x_0+x_1+x_2+x_3 = c$ and $x_1+x_2 = c$, we have $x_0 = x_3 = 0$ in $p_{(x_0,x_1,x_2,d,c)}$. Therefore,

$$\begin{aligned}q(d) &= \frac{\binom{d}{0}\binom{k-d}{x_1}\binom{k-d}{x_2}\binom{d}{0}}{\binom{2k}{c}} \Big/ \frac{\binom{0}{0}\binom{k}{x_1}\binom{k}{x_2}\binom{0}{0}}{\binom{2k}{c}}\\ &= \left[\frac{(k-d)!}{(k-d-x_1)!x_1!}\frac{(k-d)!}{(k-d-x_2)!x_2!}\right] \Big/ \left[\frac{k!}{(k-x_1)!x_1!} \Big/ \frac{k!}{(k-x_2)!x_2!}\right].\end{aligned} \tag{F.26}$$

By direct cancellations of factorials, the above equation can be simplified as

$$q(d) = \frac{\prod_{i=0}^{d-1}(k-x_1-i)\prod_{i=0}^{d-1}(k-x_2-i)}{\prod_{i=0}^{d-1}(k-i)\prod_{i=0}^{d-1}(k-i)} = \left[\prod_{i=0}^{d-1}\frac{k-x_1-i}{k-i}\right]\left[\prod_{i=0}^{d-1}\frac{k-x_2-i}{k-i}\right]. \tag{F.27}$$

To upper bound these two products in Equation (F.27), we consider a general argument. For any integer $b \geqslant a \geqslant x \geqslant 0$, we have

$$\frac{a-x}{b-x} \leqslant \ldots \leqslant \frac{a-1}{b-1} \leqslant \frac{a}{b}.$$

Therefore,

$$\left(\frac{a-x}{b-x}\right)^x \leqslant \frac{a(a-1)...(a-x+1)}{b(b-1)...(b-x+1)} \leqslant \left(\frac{a}{b}\right)^x$$

Hence, let $a = k - x_1, b = k, x = d - 1$, we can bound $\prod_{i=0}^{d-1} \frac{k-x_1-i}{k-i}$ in Equation (F.27) as

$$(\frac{k-d+1-x_1}{k-d+1})^d \leqslant \prod_{i=0}^{d-1} \frac{k-x_1-i}{k-i} \leqslant (\frac{k-x_1}{k})^d.$$

Similarly, we can bound $\prod_{i=0}^{d-1} \frac{k-x_2-i}{k-i}$ in Equation (F.27) as $(\frac{k-d+1-x_2}{k-d+1})^d \leqslant \prod_{i=0}^{d-1} \frac{k-x_2-i}{k-i} \leqslant (\frac{k-x_2}{k})^d$. Therefore, the Equation (F.27) can be upper and lower bounded as

$$\left(\frac{k-d+1-x_1}{k-d+1}\right)^d \left(\frac{k-d+1-x_2}{k-d+1}\right)^d \leqslant q(d) \leqslant \left(\frac{k-x_1}{k}\right)^d \left(\frac{k-x_2}{k}\right)^d. \tag{F.28}$$

We will show both LHS and RHS of Equation (F.28) is $1 + \mathcal{O}(\frac{1}{k})$. First, consider the terms in the RHS of Equation (F.28). Recall that d, x_1 are finite compared to k, by binomial theorem

$$\left(\frac{k-x_1}{k}\right)^d = \left(1 - \frac{x_1}{k}\right)^d = \sum_{i=0}^{d} \binom{d}{i} \left(-\frac{x_1}{k}\right)^i = 1 + \mathcal{O}(\frac{1}{k}).$$

Similarly, for the other term in the RHS of Equation (F.28), we achieve

$$\left(\frac{k-x_2}{k}\right)^d = 1 + \mathcal{O}(\frac{1}{k}).$$

Similarly, for the two terms in the LHS of Equation (F.28), we have

$$\begin{aligned}
\left(\frac{k-x_1-d+1}{k-d+1}\right)^d &= 1 + \mathcal{O}(\frac{1}{k-d+1}) = 1 + \mathcal{O}(\frac{1}{k}), \\
\left(\frac{k-x_2-d+1}{k-d+1}\right)^d &= 1 + \mathcal{O}(\frac{1}{k-d+1}) = 1 + \mathcal{O}(\frac{1}{k}).
\end{aligned}$$

Putting the above analysis together for Equation (F.28), we get

$$\begin{aligned}
\left[1 + \mathcal{O}(\frac{1}{k})\right]\left[1 + \mathcal{O}(\frac{1}{k})\right] &\leqslant q(d) \leqslant \left[1 + \mathcal{O}(\frac{1}{k})\right]\left[1 + \mathcal{O}(\frac{1}{k})\right] \\
\implies \left[1 + \mathcal{O}(\frac{1}{k})\right] &\leqslant q(d) \leqslant \left[1 + \mathcal{O}(\frac{1}{k})\right].
\end{aligned}$$

This completes the proof. $\square$

F.6. Proof of Lemma E.6

Proof of Lemma E.6. By 4.9, given $S_1^{(2k)}, S_2^{(2k)} s.t. |S_1^{(2k)} \cap S_2^{(2k)}| = c$,

$$\begin{aligned}\eta^2_{c,2k}(d_1,d_2) &= \mathrm{Cov}\left[\varphi_{d_1}\left(S_1^{(2k)}\right) - \varphi_0\left(S_1^{(2k)}\right), \varphi_{d_2}\left(S_2^{(2k)}\right) - \varphi_0\left(S_2^{(2k)}\right)\right] \\ &\leqslant \mathrm{Cov}\left[\varphi_{d_1}\left(S_1^{(2k)}\right), \varphi_{d_2}\left(S_2^{(2k)}\right)\right] + \mathrm{Cov}\left[\varphi_0\left(S_1^{(2k)}\right), \varphi_0\left(S_2^{(2k)}\right)\right],\end{aligned}$$

where the last inequality is by the non-negativity of ρ. By the definition of φ_d in Equation (4.6), the RHS of above equation is upper bounded by

$$2 \max_{S_1,S_2 \subset S_1^{(2k)}, s.t. |S_1\cap S_2| \leqslant d_1, S_3,S_4\subset S_2^{(2k)}, s.t. |S_3 \cap S_4| \leqslant d_2} \rho = \mathcal{O}(F^{(k)}). \qquad \square$$

F.7. Proof of Lemma E.8

Proof of Lemma E.8. We apply the strategies we used in the proof of Lemma E.4. The truncation parameter is $T_2 = \left\lfloor \frac{1}{\epsilon} \right\rfloor + 1$. Recall in Lemma E.4, $\mathrm{Var}(\hat{V}_u) = \sum_{c=1}^{2k} \binom{n}{2k}^{-1} \binom{2k}{c} \binom{n-2k}{2k-c} \sigma^2_{c,2k}$, where

$$\sigma^2_{c,2k} = \sum_{d_1=1}^{k} \sum_{d_1=1}^{k} w_{d_1} w_{d_2} \eta^2_{c,2k}(d_1,d_2), \text{ for } c = 1,2,...,T_1.$$

We decompose $\sigma^2_{c,2k}$ into three parts:

$$\begin{aligned}\sigma^2_{c,2k} = &\underbrace{\sum_{d_1=1}^{T_2} \sum_{d_2=1}^{T_2} w_{d_1} w_{d_2} \eta^2_{c,2k}(d_1,d_2)}_{A} + 2 \underbrace{\sum_{d_1=1}^{T_2} \sum_{d_2=T_2+1}^{k} w_{d_1} w_{d_2} \eta^2_{c,2k}(d_1,d_2)}_{B} \\ &+ \underbrace{\sum_{d_1=T_2+1}^{k} \sum_{d_2=T_2+1}^{k} w_{d_1} w_{d_2} \eta^2_{c,2k}(d_1,d_2)}_{C}. \qquad \text{(F.29)}\end{aligned}$$

Similarly, denote

$$\breve{A} := \breve{\sigma}^2_{c,2k,(T_2)} = \sum_{d_1=1}^{T_2} \sum_{d_2=1}^{T_2} w_{d_1} w_{d_2} \breve{\eta}^2_{c,2k}(d_1,d_2), \qquad \text{(F.30)}$$

where $\breve{\eta}_{c,2k}(d_1,d_2)$ is the upper bound given in Lemma E.5. To prove this lemma, it suffices to show

$$\lim_{k\to\infty} \frac{2B + C}{\breve{A}} = 0. \qquad \text{(F.31)}$$

It remains to bound $\breve{A}, B, C$ as

$$\breve{A} \asymp \frac{w_1^2 F_c}{k^2}, B = \mathcal{O}\left(w_1 \breve{w}_{T_2+1} F_c\right), C = \mathcal{O}\left(\breve{w}^2_{T_2+1} F_c\right), \qquad \text{(F.32)}$$

where $\breve{w}_d = \mathcal{O}(\frac{k^{2d}}{d!n^d})$ is the rough upper bound of w_d in (4.7). We need to quantify two parts, the coefficients $w_{d_1}w_{d_2}$ and the covariance $\eta^2_{c,2k}(d_1,d_2)$. Fix one $c \leqslant T_1$ and first quantify $\eta^2_{c,2k}(d_1,d_2)$. By Lemma E.5 and E.6, we have

$$\eta^2_{c,2k}(d_1,d_2) = \mathcal{O}(\frac{1}{k^2}F_c), \text{ for } c \leqslant T_1, d_1, d_2 \leqslant T_2, \tag{F.33}$$

$$\eta^2_{c,2k}(d_1,d_2) = \mathcal{O}(F_c), \text{ for } c \leqslant 2k, d_1, d_2 \leqslant k. \tag{F.34}$$

By Proposition E.2, we have $A = \sigma^2_{c,2k,(T_2)} = \mathcal{O}(\frac{k^2}{n^2}F_c)$. Since $\check{A}$ is the upper bound of A given in Proposition E.2, $\check{A} = \mathcal{O}(\frac{k^2}{n^2}F_c)$. For B, C, we upper bound $\eta^2_{c,2k}(d_1,d_2)$ by $\mathcal{O}(F_c)$ in Equation (F.34). Hence, we can reduce the analysis for both coefficients and covariance to the analysis on only coefficients w_d, for

$$B = \mathcal{O}(F_c)\left[\sum_{d_1=1}^{T_2} w_{d_1}\right]\left[\sum_{d_2=T_2+1}^{k} w_{d_2}\right]$$
$$C = \mathcal{O}(F_c)\left[\sum_{d_1=T_2+1}^{k} w_{d_1}\right]\left[\sum_{d_2=T_2+1}^{k} w_{d_2}\right].$$

To be more specific, it remains to show that

$$\sum_{d=1}^{T_2} w_d = \mathcal{O}(w_1), \qquad \sum_{d=T_2+1}^{k} w_d = \mathcal{O}(\breve{w}_{T_2+1}).$$

where $\breve{w}_d = \mathcal{O}(\frac{k^{2d}}{d!n^d})$ is the rough upper bound of w_d in (4.7).

For $\sum_{d=1}^{T_2} w_d$, by Equation (4.7), each $w_d = [1+o(1)]\frac{k^{2d}}{d!n^d} \leqslant [1+o(1)]\frac{k^{2d}}{n^d}$. The common ratio of geometric decay is $k^2/n = o(1)$. Therefore, the first term w_1 dominates $\sum_{d=1}^{T_2} w_d$. For $\sum_{d=T_2+1}^{k} w_d$, by Equation (4.7), we have each $w_d = \mathcal{O}(\frac{k^{2d}}{d!n^d}) = \mathcal{O}(\frac{k^{2d}}{n^d})$. Similarly, by geometric decay with common ratio k^2/n,

$$\sum_{d=T_2+1}^{k} w_d \leqslant \sum_{d=T_2+1}^{k} \mathcal{O}(\frac{k^{2d}}{n^d}) = \mathcal{O}\left((\frac{k^2}{n})^{T_2+1}\right).$$

Hence we define $\breve{w}_{T_2+1} \asymp \frac{k^2}{n})^{T_2+1}$. Then $\sum_{d=T_2+1}^{k} w_d = \mathcal{O}(\breve{w}_{T_2+1})$.

We have proved the bounds in Equation (F.32). Then plug Equation (F.32) into the LHS of Equation (F.31), we can conclude that

$$\frac{2B+C}{\check{A}} = \mathcal{O}\left(\frac{k^2(w_1\breve{w}_{T_2+1} + \breve{w}^2_{T_2+1})}{w_1^2}\right). \tag{F.35}$$

For (F.35), plugging in $T_2 = \lfloor 1/\epsilon \rfloor + 1$ and the the upper bound of $w_d = [1 + o(1)]\frac{k^{2d}}{d!n^d}$ and $\breve{w}_d = \mathcal{O}(\frac{k^{2d}}{d!n^d})$ from Equation (4.7) and (4.7), we conclude

$$\frac{2B+C}{\check{A}} = \mathcal{O}\left(k^2 n^{-2\epsilon(\lfloor 1/\epsilon \rfloor+1)}\right) = \mathcal{O}\left(k^2 n^{-2\epsilon(\lfloor 1/\epsilon \rfloor+1)}\right) = o(k^2 n^{-2}) = o(1).$$

This completes the proof. ∎

Appendix G: Relaxation of Assumption 3 and according proof

Assumption 3 assumes that $\rho(\underline{r}, d_1, d_2)$ only depends on $\underline{r}$, and therefore it has 9 *DoF*. This is valid in Example B.2 but still too restrictive in practice. In this section, we first present Assumption 6 as a relaxation of Assumption 3, Then, we show that the technical lemmas can be proved under our relaxed assumption. We denote ρ as $\rho(\underline{r}, d_1, d_2)$ in this section, since ρ depends on all 11 *DoF* rather than only 9 *DoF*.

G.1. Assumption 6

Assumption 6 (Relaxation of Assumption 3). Given $\underline{r}$, for any finite d_1, d_2, we have $d_1 \geqslant r_{0*} = \sum_{j=0}^{2} r_{0j}$, $d_2 \geqslant r_{*0} = \sum_{i=0}^{2} r_{i0}$ by their definition. There exist constant B and $B(\underline{r})$ such that $B(\underline{r}) \leqslant B < \infty$,

$$\rho(\underline{r}, d_1, d_2) = \left[1 + B(\underline{r})\frac{d_1 - r_{0*} + d_2 - r_{*0}}{k} + \mathcal{O}(\frac{1}{k^2})\right]\tilde{\rho}(\underline{r}), \tag{G.1}$$

$$\rho(\underline{r}, d_1, d_2) \leqslant B\tilde{\rho}(\underline{r}), \tag{G.2}$$

where

$$\tilde{\rho}(\underline{r}) := \rho(\underline{r}, r_{0*}, r_{*0}). \tag{G.3}$$

In Equation (G.1), we refer to the benchmark $\tilde{\rho}(\underline{r})$ as *main effect*, capturing the contribution from the overlap between $S_1^{(2k)}$ and $S_2^{(2k)}$; while we refer to the rest part, $\mathcal{O}(\frac{1}{k})\tilde{\rho}(\underline{r})$, as *additional effect*, capturing the contribution from the overlap within $S_1^{(2k)}$ and the overlap within $S_2^{(2k)}$. Equation (G.2) bounds *additional effect* with respect to *main effect*. Note that in Example B.2, there is only the *main effect*, i.e., $B(\underline{r}) \equiv 0$ and thus Assumption 6 degenerates to Assumption 3.

We interpret *main effect* and *additional effect* as follows. First, $S_1 \cup S_2$ can be decomposed into two sets: $\mathcal{A} = (S_1 \cup S_2) \cap (S_1^{(2k)} \cap S_2^{(2k)})$, i.e., the set of *Influential Overlaps* and $\mathcal{B} = (S_1 \cup S_2) \backslash (S_1^{(2k)} \cap S_2^{(2k)})$. Similar analysis holds for the $\mathcal{O}(\frac{d_2}{k})$ term. When we fix $\underline{r}$ and increase d_1, $\mathcal{A}$ does not change while the structure of $\mathcal{B}$ changes with d_1, causing $\rho(\underline{r}, d_1, d_2)$ to deviate from $\tilde{\rho}(\underline{r})$. Second, we assume that this deviation has a lower order impact compared to the main effect, i.e., the order of $\mathcal{O}(\frac{d_1}{k})$.

G.2. Proof under Assumption 6

In our previous proof, only two fundamental lemmas directly rely on Assumption 3: Lemma E.5 (the precise bound for $\eta^2_{c,2k}(d_1, d_2)$) and Lemma E.6 (the rough bound for $\eta^2_{c,2k}(d_1, d_2)$). Based on these two lemmas, we can derive the upper bounds of $\sigma^2_{c,2k}$ and hence upper bound $\mathrm{Var}(\hat{V}_u)$ (see the proof roadmap in Appendix 4.4). Therefore, when Assumption 3 is relaxed to Assumption 6, it suffices to show the results in Lemma E.5 and E.6.

First, it is trivial to validate a relaxed Lemma E.6 under Assumption 6. Assumption 6 does not change the upper bound of $\rho(\underline{r}, d_1, d_2)$, which is $F_c^{(k)} = \mathrm{Cov}[h(S')^2, h(S'')^2]$ s.t. $|S' \cap S''| = c$ (4.13). The proof of Lemma E.6 in Appendix F.6 only requires the upper bound $F_c^{(k)}$, thus it still works.

Second, we need to adapt the proof of Lemma E.5 in Appendix F.5. $\rho(\underline{r}, d_1, d_2)$ can no longer be represented as $\rho(\underline{r})$. Thus, $\rho(\underline{r}, d_1, d_2) - \rho(\underline{r}, d_1', d_2')$ is not necessarily 0 for $(d_1, d_2) \neq (d_1', d_2')$.

Proof of Lemma E.5 under Assumptions 1, 2, 4 - 6. We adopt the proof in Appendix F.5 from the beginning to Lemma F.3. We note that Lemma F.3 does not rely on Assumption 3. Hence, we still have Equation (F.17):

$$\begin{aligned}
&\mathrm{Cov}\left[\varphi_{d_1}\left(S_1^{(2k)}\right), \varphi_{d_2}\left(S_2^{(2k)}\right)\right] \\
&= \sum_{\text{feasible } \underline{r}^*} p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} g(\underline{r}^*, d_1, d_2),
\end{aligned}$$

where $g(\underline{r}^*, d_1, d_2)$ is given in Equation (F.19).eq:def:g Since $\rho(\underline{r}, d_1, d_2)$ in $g(\underline{r}^*, d_1, d_2)$ (F.19) depends d_1, d_2, we cannot further simplify $g(\underline{r}^*, d_1, d_2)$ as $G(\underline{r}^*)$. Further, we denote

$$\begin{aligned}
\overline{pg}(\underline{r}^*, d_1, d_2, c) &:= p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} g(\underline{r}^*, d_1, d_2, c); \\
\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) &:= \overline{pg}(\underline{r}^*, d_1, d_2, c) - \overline{pg}(\underline{r}^*, 0, d_2, c) \\
&\quad - \overline{pg}(\underline{r}^*, d_1, 0, c) + \overline{pg}(\underline{r}^*, 0, 0, c).
\end{aligned}$$

Then $\eta_{c,2k}^2(d_1, d_2)$ can be represented as

$$\eta_{c,2k}^2(d_1, d_2) = \sum_{(r_{0*}, r_{1*}, r_{2*})} \sum_{(r_{*0}, r_{*1}, r_{*2})} \overline{\Delta pg}(\underline{r}^*, d_1, d_2, c). \tag{G.4}$$

We apply the strategy used in the proof of Lemma F.4, partitioning the summation as follows:

$$\begin{aligned}
\eta_{c,2k}^2(d_1, d_2) = &\left(\sum_{r_{1*}+r_{2*}=c} + \sum_{r_{1*}+r_{2*} \leqslant c-1}\right) \\
&\times \left(\sum_{r_{*1}+r_{*2}=c} + \sum_{r_{*1}+r_{*2} \leqslant c-1}\right) \overline{\Delta pg}(\underline{r}^*, d_1, d_2, c).
\end{aligned} \tag{G.5}$$

There are 4 cases. Case A: $r_{1*} + r_{2*} = c$ and $r_{*1} + r_{*2} = c$; case B: $r_{1*} + r_{2*} = c$ and $r_{*1} + r_{*2} \leqslant c - 1$; case C: $r_{1*} + r_{2*} \leqslant c - 1$ and $r_{*1} + r_{*2} = c$; case D: $r_{1*} + r_{2*} \leqslant c - 1$ and $r_{*1} + r_{*2} \leqslant c - 1$. Since c is finite, r_{i*}'s and r_{*j}'s are also finite. Thus, (G.5) is a finite summation. To show $\eta_{c,2k}^2(d_1, d_2) = \mathcal{O}(\frac{1}{k^2} F_c^{(k)})$, it suffices to show that the summand $\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) = \mathcal{O}(\frac{1}{k^2} F_c^{(k)})$ in all 4 cases.

First, we study case A. For the $g(\underline{r}^*, d_1, d_2)$ defined in Equation (F.19), by approximation of ρ in Assumption 6: $\rho(\underline{r}, d_1, d_2) = \left[1 + B(\underline{r})\frac{d_1' + d_2'}{k} + \mathcal{O}(\frac{1}{k^2})\right]\tilde{\rho}(\underline{r})$, we have

$$g(\underline{r}^*, d_1, d_2) = [1 + \frac{d_1 + d_2}{k}B(\underline{r}) + \mathcal{O}(\frac{1}{k^2})]g(\underline{r}^*, 0, 0),$$

Therefore, $\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c)$ can be simplified as

$$\begin{aligned}
&\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) \\
&\quad = g(\underline{r}^*, 0, 0)\big\{(p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} - p_{(r_{0*}, r_{1*}, r_{2*}, 0, c)}) \\
&\quad\times (p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} - p_{(r_{*0}, r_{*1}, r_{*2}, 0, c)}) \\
&\quad + \frac{d_1 B(\underline{r})}{k} p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)}(p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} - p_{(r_{*0}, r_{*1}, r_{*2}, 0, c)}) \\
&\quad + \frac{d_2 B(\underline{r})}{k} p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)}(p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} - p_{(r_{0*}, r_{1*}, r_{2*}, d_2, c)}) + \mathcal{O}(\frac{1}{k^2})\big\}.
\end{aligned}$$

By Lemma F.4, $p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} - p_{(r_{0*}, r_{1*}, r_{2*}, 0, c)} = \mathcal{O}(1/k)$, $p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} - p_{(r_{*0}, r_{*1}, r_{*2}, 0, c)} = \mathcal{O}(1/k)$. Besides, $g(\underline{r}^*, 0, 0, c) \leqslant F_c^{(k)}$. Thus, $\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) = \mathcal{O}(\frac{1}{k^2}F_c^{(k)})$.

Secondly, we study case B. Recall that $p_{(r_{*0}, r_{*1}, r_{*2}, 0, c)} = 0$ by (F.18). Therefore, $\overline{pg}(\underline{r}^*, d_1, 0, c) = \overline{pg}(\underline{r}^*, 0, 0, c) = 0$. Hence, $\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c)$ can be simplified as

$$\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) = \overline{pg}(\underline{r}^*, d_1, d_2, c) - \overline{pg}(\underline{r}^*, 0, d_2, c). \tag{G.6}$$

By Assumption 6, we can approximate $g(\underline{r}^*, d_1, d_2)$ as

$$g(\underline{r}^*, d_1, d_2) = [1 + \frac{d_1}{k}B(\underline{r}) + \mathcal{O}(\frac{1}{k^2})]g(\underline{r}^*, 0, d_2).$$

Then, plug the above approximation into Equation (G.6):

$$\begin{aligned}
\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) = g(\underline{r}^*, 0, d_2)\big[&p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} \\
&\times (p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} - p_{(r_{0*}, r_{1*}, r_{2*}, 0, c)}) \\
&+ \frac{d_1 B(\underline{r})}{k} p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} + \mathcal{O}(\frac{1}{k^2})\big].
\end{aligned}$$

Similarly, we have $p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} = \mathcal{O}(1/k)$ and $p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} - p_0(\underline{x}) = \mathcal{O}(1/k)$. Given $g(\underline{r}^*, 0, d_2, c) \leqslant F_c^{(k)}$, we conclude that $\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) = \mathcal{O}(\frac{1}{k^2}F_c^{(k)})$.

Thirdly, by a similarly analysis in case B, we can bound $\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) = \mathcal{O}(\frac{1}{k^2}F_c^{(k)})$ in case C.

Finally, we study case D. Since $r_{1*} + r_{2*} \leqslant c - 1$ and $r_{*1} + r_{*2} \leqslant c - 1$, $p_{(r_{0*}, r_{1*}, r_{2*}, 0, c)} = p_{(r_{*0}, r_{*1}, r_{*2}, 0, c)} = 0$. Therefore,

$$\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c)d_1 d_2 = p_{(r_{0*}, r_{1*}, r_{2*}, d_1, c)} p_{(r_{*0}, r_{*1}, r_{*2}, d_2, c)} g(\underline{r}^*, d_1, d_2).$$

Since $p_{(r_{0*},r_{1*},r_{2*},d_1,c)} = \mathcal{O}(1/k)$, $p_{(r_{*0},r_{*1},r_{*2},d_2,c)} = \mathcal{O}(1/k)$, and $g(\underline{r}^*, d_1, d_2) \leqslant F_c^{(k)}$. Therefore, we can bound $\overline{\Delta pg}(\underline{r}^*, d_1, d_2, c) = \mathcal{O}(\frac{1}{k^2} F_c^{(k)})$ in case D.

This completes the proof. ☐

Appendix H: Low order kernel h illustration

This section provides an over-simplified analysis of the behavior of $E(\hat{V}_u)$ and $\mathrm{Var}(\hat{V}_u)$ assuming that h is a linear kernel. Then, we discuss the difficulty in generalizing the analysis to a general kernel h, which motivates us to propose the assumptions for $\mathrm{Cov}[h(S_1)h(S_2), h(S_3)h(S_4)]$ in Section 4.2.

Proposition H.1. *$X_1, X_2, ..., X_n$ i.i.d., s.t. $\mathrm{E}(X_1) = 0, \mathrm{Var}(X_1) = \gamma > 0$. Suppose kernel function $h(X_1, ..., X_k) = \frac{1}{k}\sum_{i=1}^{k} X_i$ Then, we have the ratio consistency of the estimator,*

$$\frac{\mathrm{Var}(\hat{V}_u)}{\left(\mathrm{E}(\hat{V}_u)\right)} = \mathcal{O}(\frac{1}{n}).$$

This can be concluded by showing that $\mathrm{E}(\hat{V}_u) = \Omega(\frac{1}{n})$ and $\mathrm{Var}(\hat{V}_u) = \mathcal{O}(\frac{1}{n^3})$. We skip a complete proof of this proposition but remark on three key steps to upper bound $\mathrm{Var}(\hat{V}_u)$: 1) the decomposition of $\hat{V}_u$ by a double U-statist structure (Proposition 4.2); 2) deriving an explicit form of $\varphi_d\left(S^{(2k)}\right) - \varphi_0\left(S^{(2k)}\right)$ in the above decomposition as

$$\frac{d}{k^2}\left(\frac{1}{2k}\sum_{X_i \in S^{(2k)}} X_i^2 - \frac{2}{2k(2k-1)}\sum_{X_i, X_j \in S^{(2k)}, j>i} X_i X_j\right);$$

and 3) showing that the leading term in $\mathrm{Var}(\hat{V}_u)$ dominates $\mathrm{Var}(\hat{V}_u)$.

We note that a similar analysis may be performed for an intrinsic low-order kernel, where $h(X_1, ..., X_k) = \binom{k}{l}^{-1}\sum_{i_1<...<i_l} g^{(l)}(X_{i_1}, ..., X_{i_l})$ for a fixed l and g is an order l kernel. Because we are still able to derive an explicit form of $\varphi_d(S^{(2k)}) - \varphi_0(S^{(2k)})$.

However, for a general kernel h without a low-order structure, difficulties arise in the above analysis. First, $\varphi_d(S^{(2k)}) - \varphi_0(S^{(2k)})$ no longer has a simple expression. Our remedy is to quantify the implicit cancellation in covariance $\eta^2_{c,2k}(d_1, d_2) := Cov[\varphi_{d_1}(S_1^{(2k)}) - \varphi_0(S_1^{(2k)}), \varphi_{d_2}(S_2^{(2k)}) - \varphi_0(S_1^{(2k)})]$ (see Equation (4.9)). The further decomposition of $\eta^2_{c,2k}(d_1, d_2)$ involves the following covariance (4.11)

$$\rho := Cov[h(S_1)h(S_2), h(S_3)h(S_4)].$$

Thus, assumptions are made about this term. In particular, Assumption 3 reduces its degree of freedom from 11 to 9. A relaxation of Assumption 3 is presented in Appendix B. Second, we may not be able to show that $\sigma^2_{1,2k}$ (or its upper bound) dominates $\mathrm{Var}(\hat{V}_u)$ (4.3) as $\sigma^2_{2k,2k}/\sigma^2_{1,2k}$ does not have an explicit form. Therefore, we adopt a new strategy that bounds all $\sigma^2_{c,2k}$ with a tighter upper bound for $c = 1, 2, .., T_1$ (Proposition E.2) and a looser upper bound for $c = T_1 + 1, ..., 2k$ (Proposition E.3), where T_1 is a fixed value dependent on the growth rate of k. A series of technical lemmas are collected in Appendix E.

Appendix I: Additional simulation results

This section we present additional simulation results.

I.1. Computational cost

We performed additional simulation study to address the concern of computational cost. It should be noted first that the major cost of variance calculation in MS is not the part of our proposed estimator, but it is rather the standard prediction mechanism in which we send an observation down each tree. Once these tree predictions are obtained, the variance estimation is done immediately at little cost. On the other hand, BM, BM-corr (bias-corrected BM) and IJ estimators add burden to this. They all involve using the number of training samples in each tree (see, e.g., equations $m_i = \sum_{b=1}^{B_n} \omega_{i,b} h_b$ and $\hat{\zeta}^{\mathrm{BM}}_{1,k_n} = \frac{1}{n-1} \sum_{i=1}^{n} (m_i - \bar{m})^2$ in Section 4.2 [35]) and hence the total cost is at $\mathcal{O}(\texttt{nTrees} \cdot \texttt{nTrain})$. Furthermore, our MS-s estimator adds additional computational cost based on predicting additional neighboring samples, but this is not affected by the number of training samples.

To rigorously compare the computational cost of MS and BM estimators, we implemented BM and BM-corr under the same C++ parallel computing framework of our `RLT` package, which was used to implement our MS and MS-s methods. We also use the `grf` package to fit random forests and perform IJ estimation.

We perform experiments with these settings: subsample size $k = n/2$ setting and number of trees ($\texttt{nTrees} = 2000$). All experiments are performed on an 8-core AMD Ryzen 7 4800H CPU with 16 GB ram. Each experiment is repeated 1000 times and we summarize the average cost of each experiment with average's standard error (in millisecond) in Table 6. The other experimental settings are the same as the experiments presented in Section 6.2, including using 200 training samples and 55 testing samples. We enable the use of multiple cores for model fitting and prediction procedures by setting the `ncores` parameter in `RLT` and the `num.thread` parameter in `grf` as 8.

Columns in Table 6 can be interpreted as follows. The variance estimation cost for each testing samples is collected in the last column: "Additional cost from var est". For MS, MS-s, and IJ methods, this column is calculated as a difference between prediction with and without variance estimation since variance estimation is integrated in their model prediction functions. As references, we also present the cost of 1) fitting random forest on 200 training samples (column

TABLE 6

Computational cost in milliseconds. The cost of time is reported in the form of "average over 1000 simulations (standard error of the average)". IJ is implemented in `grf` *and all other methods are implemented with* `RLT` *and Rcpp.*

<table>
<tr><th rowspan="2">Method</th><th>Train</th><th colspan="3">Post-train (pred. and var. est. per testing samples)</th></tr>
<tr><th>Model Fit</th><th>Pred. cost</th><th>Total cost</th><th>Additional cost
var est. = Total – Pred.</th></tr>
<tr><td>MS</td><td rowspan="2">39.8 (0.2)</td><td rowspan="2">0.141 (<0.001)</td><td>0.151 (<0.001)</td><td>0.010 (<0.001)</td></tr>
<tr><td>MS-s</td><td>1.830 (0.010)</td><td>1.679 (0.010)</td></tr>
<tr><td>BM</td><td rowspan="2">37.7 (0.4)</td><td rowspan="2">0.143 (<0.001)</td><td>0.204 (0.002)</td><td>0.060 (0.010)</td></tr>
<tr><td>BM-cor</td><td>1.634 (0.005)</td><td>1.491 (0.005)</td></tr>
<tr><td>IJ</td><td>65.9 (1.1)</td><td>0.547 (0.001)</td><td>0.596 (0.001)</td><td>0.050 (<0.001)</td></tr>
</table>

"model fit"), 2) predicting on each testing samples (column "Pred. cost"), and 3) total cost of prediction and variance estimation (column "Total cost") on each testing sample.

When comparing the additional variance estimation cost (last column) of all methods, MS is faster than BM and IJ, which BM and IJ show similar cost. However, the costs of all 3 methods are much lighter than the cost of model fitting. The little cost of MS confirms our previous statement that they can be immediately obtained after standard predictions, i.e., $\mathcal{O}(\texttt{nTrees})$.

For each testing sample, after obtaining the predictions of `nTrees` trees, the variance estimation cost of MS estimator is $\mathcal{O}(\texttt{nTrees})$ (see line 7-13 of Algorithm 1 in Section 6). While for BM method, their estimator of the component σ_1^2 of each testing sample involves using the number of training samples in each tree (see equations $m_i = \sum_{b=1}^{B_n} \omega_{i,b} h_b$ and $\hat{\zeta}_{1,k_n}^{\text{BM}} = \frac{1}{n-1} \sum_{i=1}^{n} (m_i - \bar{m})^2$ in Section 4.2 [35]) and hence the total cost is at least $\mathcal{O}(\texttt{nTrees} \cdot \texttt{nTrain})$, where `nTrain` is the number of training samples. Similarly, IJ's cost of variance estimation for each sample is also $\mathcal{O}(\texttt{nTrees} \cdot \texttt{nTrain})$ (see Equation (5) in Wager, Hastie and Efron [29]). BM-cor estimator is a little more computationally intensive while we notice there exists an approximate version of this estimator [35, Appendix E.], whose cost is between BM and BM-cor. In short, given our matched-group samples, MS estimation is efficient because it does not need to track how many times each training sample appears in a specific tree.

I.2. Additional experimental results for `mtry` = *2*

We perform the same experiments in Section 6.2 with `mtry` = 2, i.e., p/3, as to show that variance estimation is not very sensitive to the choice of tunning parameters. Experimental results are collected in Table 7 and 8. While we do observe a slight decrease of coverage across all models, possibly due to the worse performance and possibly larger variance of random forest itself, the coverage results and relative bias are not changed significantly compared with Table 1 and Table 2 in Section 6.2.

TABLE 7
mtry=2. 90% CI Coverage Rate averaged on 50 testing samples. The number in the bracket is the standard deviation of coverage over 50 testing samples.

	$k = n/2$		$k = n/4$		$k = n/8$	
nTrees	2000	20000	2000	20000	2000	20000
MARS						
MS	78.0% (3.2%)	86.5% (1.4%)	79.3% (3.6%)	88.2% (1.0%)	79.1% (3.1%)	88.0% (1.4%)
MS-s	87.0% (2.8%)	89.6% (2.6%)	86.7% (2.5%)	90.1% (1.8%)	86.2% (2.4%)	89.4% (1.8%)
BM	85.0% (3.6%)	66.4% (1.9%)	93.4% (1.8%)	81.6% (1.4%)	94.7% (1.1%)	86.9% (1.2%)
BM-cor	3.6% (4.2%)	57.8% (2.1%)	3.6% (4.2%)	57.8% (2.1%)	82.1% (1.3%)	85.0% (1.1%)
IJ	82.1% (1.3%)	85.0% (1.1%)	90.7% (1.5%)	90.4% (0.8%)	93.5% (1.5%)	88.1% (0.9%)
MLR						
MS	81.2% (1.8%)	87.1% (1.1%)	81.9% (2.4%)	88.3% (0.9%)	81.9% (2.4%)	88.3% (0.9%)
MS-s	89.2% (1.5%)	90.4% (1.3%)	88.7% (1.3%)	90.3% (1.4%)	87.7% (1.6%)	90.6% (1.1%)
BM	82.7% (1.9%)	64.9% (1.7%)	91.8% (1.2%)	81.5% (1.2%)	94.1% (1.1%)	86.4% (0.9%)
BM-cor	6.0% (3.5%)	57.9% (2.1%)	6.0% (3.5%)	57.9% (2.1%)	82.7% (1.2%)	84.8% (1.0%)
IJ	93.9% (0.9%)	95.7% (0.7%)	90.2% (1.0%)	90.8% (1.0%)	92.5% (1.1%)	88.3% (1.2%)

TABLE 8
mtry = 2. Relative bias (standard deviation) over 50 testing samples. For each method and testing sample, the relative bias is evaluated over 1000 simulations.

	$k = n/2$		$k = n/4$		$k = n/8$	
nTrees	2000	20000	2000	20000	2000	20000
MARS						
MS	0.7% (2.3%)	0.2% (2.3%)	1.4% (2.5%)	0.8% (1.8%)	−1.1% (2.5%)	−0.4% (1.5%)
MS-s	4.6% (13.8%)	4.8% (13.5%)	4.8% (10.6%)	4.6% (10.8%)	2.3% (6.3%)	2.8% (6.6%)
BM	−15.9% (15.0%)	−63.9% (1.6%)	33.4% (16.9%)	−31.2% (2.3%)	49.9% (11.5%)	−11.9% (1.9%)
BM-cor	−120.6% (14.2%)	−74.4% (1.7%)	−60.6% (5.4%)	−40.7% (1.2%)	−27.7% (2.2%)	−19.9% (1.4%)
IJ	85.5% (13.8%)	85.0% (13.6%)	52.7% (15.7%)	23.9% (3.7%)	91.8% (22.4%)	18.2% (5.2%)
MLR						
MS	0.0% (2.7%)	0.8% (2.4%)	−0.7% (2.5%)	−1.1% (1.9%)	−0.8% (2.2%)	0.0% (1.5%)
MS-s	8.7% (8.5%)	9.3% (8.8%)	6.8% (6.6%)	6.3% (6.8%)	4.0% (4.1%)	4.9% (3.9%)
BM	−28.0% (5.7%)	−65.3% (0.9%)	19.5% (8.0%)	−33.0% (1.4%)	39.7% (7.1%)	−14.3% (1.3%)
BM-cor	−109.7% (6.0%)	−73.5% (1.1%)	−57.5% (2.8%)	−40.8% (1.0%)	−27.0% (1.7%)	−21.1% (1.4%)
IJ	76.8% (8.8%)	78.9% (10.5%)	39.8% (7.1%)	20.0% (2.8%)	73.9% (14.1%)	12.5% (2.8%)

I.3. Ground truth in the simulation

We simulate the ground truth for our experiments in main text: the expectation of forest predictions: $\mathrm{E}(f(\boldsymbol{x}^*))$ and the variance of forest predictions: $\mathrm{Var}(f(\boldsymbol{x}^*))$, by 10000 simulations. (see Section 6.1)

Since variance estimators are produced by different packages, we use the corresponding package to generate their ground truth. IJ estimator is performed by `grf` package when $k \leqslant n/2$ and `ranger` package when $k > n/2$. Both "RLT (MS)" and "RLT" use `RLT` package. However, When $k \leqslant n/2$, they apply different way to sample incomplete U-statistics: matched-group sampling for MS and MS-s, and independent subsampling (see the description above Equation 2.5) for BM and BM-cor respectively. When $k > n/2$, the sampling schemes are the same. The result of the central testing sample (see Section 6) is presented in Table 9 and Table 10. There is a small difference between different packages though similar tunning parameters are used to train random forests.

In addition, we present the "oracle" CI coverage rate in Table 11, which matches $1-\alpha$. To construct these CIs, we still use the random forest prediction over 1000 simulations but replace the estimated variance with the "true variance", $\mathrm{Var}(f(\boldsymbol{x}^*))$. This result also shows the normality of the random forest predictor.

Table 9

Ground Truth of $\mathrm{E}(f(\boldsymbol{x}^*))$ *evaluated on central testing sample by 10000 simulations. Reported in the form of "mean (standard deviation of mean)" of* $\mathrm{E}(f(\boldsymbol{x}^*))$. *Standard deviation* (< 0.01) *and* (< 0.001) *are displayed as* (0.01) *and* (0.001) *respectively.*

		MARS			MLR		
k	nTrees	RLT (MS)	RLT	grf/ranger	RLT (MS)	RLT	grf/ranger
$n/2$	2000	17.81 (0.01)	17.81 (0.01)	18.20 (0.01)	0.497 (0.001)	0.505 (0.001)	0.499 (0.001)
$n/2$	20000	17.81 (0.01)	17.81 (0.01)	18.20 (0.01)	0.497 (0.001)	0.505 (0.001)	0.499 (0.001)
$n/4$	2000	17.42 (0.01)	17.81 (0.01)	18.01 (0.01)	0.498 (0.001)	0.500 (0.000)	0.463 (0.001)
$n/4$	20000	17.42 (0.01)	17.42 (0.01)	18.00 (0.01)	0.498 (0.001)	0.501 (0.001)	0.464 (0.001)
$n/8$	2000	17.40 (0.01)	17.40 (0.01)	18.19 (0.01)	0.499 (0.001)	0.500 (0.001)	0.416 (0.001)
$n/8$	20000	17.40 (0.01)	17.40 (0.01)	18.19 (0.01)	0.499 (0.001)	0.500 (0.001)	0.420 (0.001)
$4n/5$	2000	18.21 (0.01)		18.19 (0.01)	0.499 (0.005)		0.498 (0.005)
$4n/5$	20000	18.21 (0.01)		18.19 (0.01)	0.498 (0.005)		0.498 (0.005)

Table 10

Ground Truth of $\mathrm{Var}(f(\boldsymbol{x}^*))$ *evaluated on the central testing sample by 10000 simulations.*

		MARS			MLR		
k	nTrees	RLT (MS)	RLT	grf/ranger	RLT (MS)	RLT	grf/ranger
$n/2$	2000	0.847	0.840	0.802	0.135	0.131	0.133
$n/2$	20000	0.844	0.836	0.795	0.135	0.131	0.133
$n/4$	2000	0.519	0.518	0.527	0.077	0.077	0.076
$n/4$	20000	0.513	0.512	0.509	0.077	0.076	0.077
$n/8$	2000	0.345	0.345	0.382	0.044	0.043	0.043
$n/8$	20000	0.339	0.339	0.368	0.043	0.042	0.042
$4n/5$	2000	1.334		1.348	0.214		0.213
$4n/5$	20000	1.331		1.341	0.213		0.212

Table 11

90% *CI Coverage Rate averaged on 50 testing samples, where the true variance is used in constructing the CI. The number in the bracket is the standard deviation of coverage over 50 testing samples.*

			MARS		MLR	
Tree size		nTrees	RLT	grf/ranger	RLT	grf/ranger
$k \leqslant n/2$	$k = n/2$	2000	90.12% (0.93%)	90.00% (0.97%)	89.97% (0.86%)	90.04% (0.99%)
$k \leqslant n/2$	$k = n/2$	20000	90.10% (0.96%)	89.95% (0.97%)	89.97% (0.88%)	89.96% (1.00%)
$k \leqslant n/2$	$k = n/4$	2000	89.87% (0.76%)	89.69% (0.84%)	90.07% (1.03%)	90.06% (1.25%)
$k \leqslant n/2$	$k = n/4$	20000	89.83% (0.78%)	89.63% (0.82%)	90.14% (1.04%)	89.98% (1.21%)
$k \leqslant n/2$	$k = n/8$	2000	89.53% (0.78%)	89.35% (0.85%)	90.22% (1.13%)	89.91% (1.17%)
$k \leqslant n/2$	$k = n/8$	20000	89.38% (0.89%)	89.28% (0.85%)	90.20% (1.12%)	89.78% (1.23%)
$k > n/2$	$k = \frac{4}{5}n$	2000	90.05% (0.94%)	90.02% (0.97%)	89.86% (1.05%)	89.94% (0.98%)
$k > n/2$	$k = \frac{4}{5}n$	20000	90.05% (1.01%)	90.00% (0.96%)	89.88% (0.98%)	89.86% (0.97%)

I.4. Figures of MLR model

Figure 6 shows the performance of different methods on the MLR model. This is a counterpart of Figure 2 in Section 6.

Appendix J: Additional information and results for the real data

Table 12 describes the covariates of Airbnb data in Section 7. We use the samples with the price falling in the interval $(0, 500]$ dollars. The missing values (NA) in the rating score and bathroom number are replaced. The "having rating" covariate is created based on the "review number".

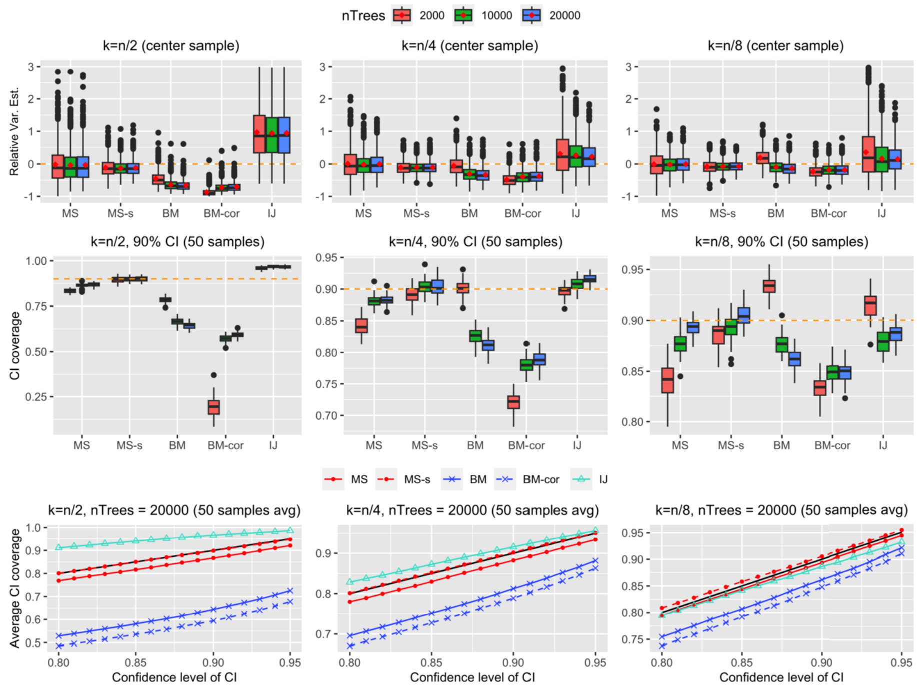

FIG 6. *A comparison of different methods on MLR data. Each column of figure panel corresponds to one tree size: $k = n/2, n/4, n/8$. The first row: boxplots of relative variance estimators of the central test sample over 1000 simulations. The diamond symbol in the boxplot indicates the mean. The range of y-coordinate is restricted within $[-1, 3]$. The second row: boxplots of 90% CI coverage for 50 testing samples. For each method, three side-by-side boxplots represent `nTrees` as 2000, 10000, 20000. The third row: the coverage rate averaged over 50 testing samples with 20000 `nTrees` and the confidence level (x-axis) from 80% to 95%. The black reference line $y = x$ indicates the desired coverage rate.*

TABLE 12
Covariates information of Airbnb data.

Covariate Name	Description
latitude	Latitude of the Airbnb unit.
longitude	Longitude of the Airbnb unit.
room type	Three types (with # of samples): Entire home/apt (5547), Private room (1839) and Shared room (129).
bedroom number	Number of bedrooms in this unit.
bathroom number	Number of bathrooms in this unit. NA values are replaced by 0.
accommodates	Maximum accommodates of this unit.
reviews number	The number of reviews of this unit.
having a rating	It is 1 if the number of reviews is greater than 0; and is 0 otherwise.
rating score	The average rating score. NA is replaced by the average score.

To train the random forest model, we set `mtry` (number of variables randomly sampled as candidates at each split) as 3, and set `nodesize` parameter as 36. Here we also present the details of testing samples. The latitude and longitude of SEA Airport, Seattle downtown, and Mercer Island are (47.4502, −122.3088), (47.6050, −122.3344), and (47.5707, −122.2221) respectively. The “room type” “accommodates” and “having a rating” are fixed as “Entire home/apt”, the double of “bedroom numbers”, and 1 respectively. We use averages in the training data as the values of “reviews number” and “rating score”.

Funding

Ruoqing Zhu was supported by NSF grant 2210657.

References

[1] Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics* **47** 1148–1178. MR3909963
[2] Biau, G. (2012). Analysis of a Random Forests Model. *The Journal of Machine Learning Research* **13** 1063–1095. MR2930634
[3] Breiman, L. (2001). Random Forests. *Machine Learning* **45** 5–32. MR3874153
[4] Chen, X. and Kato, K. (2019). Randomized incomplete U-statistics in high dimensions. *The Annals of Statistics* **47** 3127–3156. MR4025737
[5] Chen, S. X. and Peng, L. (2021). Distributed Statistical Inference for Massive Data. *The Annals of Statistics* **49** 2851–2869. MR4338895
[6] Clémençon, S., Colin, I. and Bellet, A. (2016). Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. *The Journal of Machine Learning Research* **17** 2682–2717. MR3517099
[7] Cochran, W. G. (2007). *Sampling Techniques*, 3 ed. John Wiley & Sons. MR0054199
[8] DiCiccio, C. and Romano, J. (2022). CLT for U-Statistics with Growing Dimension. *Statistica Sinica* **32** 1–22. MR4359635
[9] Efron, B. (2014). Estimation and Accuracy After Model Selection. *Journal of the American Statistical Association* **109** 991–1007. MR3265671
[10] Folsom, R. E. (1984). Probability Sample U-statistics: Theory and Applications for Complex Sample Designs, PhD thesis, The University of North Carolina at Chapel Hill.
[11] Frees, E. W. (1989). Infinite Order U-statistics. *Scandinavian Journal of Statistics* 29–45. MR1003967
[12] Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics* 1–67. MR1091842
[13] Geurts, P., Ernst, D. and Wehenkel, L. (2006). Extremely Randomized Trees. *Machine Learning* **63** 3–42.
[14] Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *Ann. Math. Statist.* **19** 293–325. `https://doi.org/10.1214/aoms/1177730196`. MR0026294
[15] Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association* **47** 663–685. MR0053460

[16] Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics* **2** 841 – 860. https://doi.org/10.1214/08-AOAS169. MR2516796
[17] Lee, A. J. (1990). *U-statistics: Theory and Practice*. CRC Press. MR1075417
[18] Li, X., Li, R., Xia, Z. and Xu, C. (2020). Distributed Feature Screening via Componentwise Debiasing. *Journal of Machine Learning Research* **21**. MR4071207
[19] Lin, N. and Xi, R. (2010). Fast Surrogates of U-statistics. *Computational Statistics & Data Analysis* **54** 16–24. MR2558454
[20] Mentch, L. and Hooker, G. (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research* **17** 841–881. MR3491120
[21] Papa, G., Clémençon, S. and Bellet, A. (2015). SGD Algorithms Based on Incomplete U-statistics: Large-Scale Minimization of Empirical Risk. *Advances in Neural Information Processing Systems* **28**.
[22] Peng, W., Coleman, T. and Mentch, L. (2022). Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics* **16** 232–292. MR4359361
[23] Peng, W., Mentch, L. and Stefanski, L. (2021). Bias, Consistency, and Alternative Perspectives of the Infinitesimal Jackknife. *arXiv preprint arXiv:2106.05918*.
[24] Schucany, W. R. and Bankson, D. M. (1989). Small sample variance Estimators for U-statistics. *Australian Journal of Statistics* **31** 417–426.
[25] Sen, A. R. (1953). On the Estimate of the Variance in Sampling With Varying Probabilities. *Journal of the Indian Society of Agricultural Statistics* **5** 127. MR0068179
[26] Sen, P. K. (1960). On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin* **10** 1–18. MR0119286
[27] Song, Y., Chen, X. and Kato, K. (2019). Approximating high-dimensional infinite-order U-statistics: Statistical and computational guarantees. *Electronic Journal of Statistics* **13** 4794–4848. MR4038726
[28] Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* **113** 1228-1242. https://doi.org/10.1080/01621459.2017.1319839. MR3862353
[29] Wager, S., Hastie, T. and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research* **15** 1625–1651. MR3225243
[30] Wang, Q. (2012). Investigation of Topics in U-statistics and Their Applications in Risk Estimation and Cross-validation, PhD thesis, Penn State University. MR3295407
[31] Wang, Q. and Lindsay, B. (2014). Variance Estimation of a General U-statistic with Application to Cross-validation. *Statistica Sinica* 1117–1141. MR3241280
[32] Wang, Q. and Lindsay, B. (2017). Pseudo-kernel method in U-statistic

variance estimation with large kernel size. *Statistica Sinica* 1155–1174. MR3699698

[33] WANG, Q. and WEI, Y. (2022). Quantifying uncertainty of subsampling-based ensemble methods under a U-statistic framework. *Journal of Statistical Computation and Simulation* 1–21. MR4503374

[34] YATES, F. and GRUNDY, P. M. (1953). Selection Without Replacement from Within Strata with Probability Proportional to Size. *Journal of the Royal Statistical Society: Series B (Methodological)* **15** 253–261.

[35] ZHOU, Z., MENTCH, L. and HOOKER, G. (2021). V-statistics and Variance Estimation. *Journal of Machine Learning Research* **22** 1–48. MR4353066