The Annals of Statistics
2019, Vol. 47, No. 6, 3127–3156
https://doi.org/10.1214/18-AOS1773

RANDOMIZED INCOMPLETE U-STATISTICS IN HIGH DIMENSIONS[2]

BY XIAOHUI CHEN[1] AND KENGO KATO

University of Illinois at Urbana-Champaign and Cornell University

This paper studies inference for the mean vector of a high-dimensional U-statistic. In the era of big data, the dimension d of the U-statistic and the sample size n of the observations tend to be both large, and the computation of the U-statistic is prohibitively demanding. Data-dependent inferential procedures such as the empirical bootstrap for U-statistics is even more computationally expensive. To overcome such a computational bottleneck, incomplete U-statistics obtained by sampling fewer terms of the U-statistic are attractive alternatives. In this paper, we introduce randomized incomplete U-statistics with sparse weights whose computational cost can be made independent of the order of the U-statistic. We derive nonasymptotic Gaussian approximation error bounds for the randomized incomplete U-statistics in high dimensions, namely in cases where the dimension d is possibly much larger than the sample size n, for both nondegenerate and degenerate kernels. In addition, we propose generic bootstrap methods for the incomplete U-statistics that are computationally much less demanding than existing bootstrap methods, and establish finite sample validity of the proposed bootstrap methods. Our methods are illustrated on the application to nonparametric testing for the pairwise independence of a high-dimensional random vector under weaker assumptions than those appearing in the literature.

1. Introduction. Let $X_1, \dots, X_n$ be independent and identically distributed (i.i.d.) random variables taking values in a measurable space $(S, \mathcal{S})$ with common distribution P. Let $r \geq 2$ and $d \geq 1$ be given positive integers, and let $h = (h_1, \dots, h_d)^T : S^r \to \mathbb{R}^d$ be a fixed and jointly measurable function that is symmetric in its arguments, that is, $h(x_1, \dots, x_r) = h(x_{i_1}, \dots, x_{i_r})$ for every permutation $i_1, \dots, i_r$ of $1, \dots, r$. Suppose that $\mathbb{E}[|h_j(X_1, \dots, X_r)|] < \infty$ for all $j = 1, \dots, d$, and consider inference on the mean vector $\theta = (\theta_1, \dots, \theta_d)^T = \mathbb{E}[h(X_1, \dots, X_r)]$. To this end, a commonly used statistic is the U-statistic with kernel h, that is, the sample average of $h(X_{i_1}, \dots, X_{i_r})$ over all distinct r-tuples

Received December 2017; revised October 2018.

[1]Supported in part by NSF Grant DMS-1404891, NSF CAREER Award DMS-1752614 and UIUC Research Board Awards (RB17092, RB18099).

[2]This work is completed in part with the high-performance computing resource provided by the Illinois Campus Cluster Program at UIUC.

MSC2010 subject classifications. Primary 62E17, 62F40; secondary 62H15.

Key words and phrases. Incomplete U-statistics, randomized inference, Gaussian approximation, bootstrap, divide and conquer, Bernoulli sampling, sampling with replacement.

$(i_1, \dots, i_r)$ from $\{1, \dots, n\}$

$$U_n := U_n^{(r)}(h) := \frac{1}{|I_{n,r}|} \sum_{(i_1,\dots,i_r)\in I_{n,r}} h(X_{i_1}, \dots, X_{i_r}), \tag{1.1}$$

where $I_{n,r} = \{(i_1, \dots, i_r) : 1 \le i_1 < \cdots < i_r \le n\}$ and $|I_{n,r}| = n!/\{r!(n-r)!\}$ denotes the cardinality of $I_{n,r}$.

U-statistics are an important and general class of statistics, and applied in a wide variety of statistical problems; we refer to [27] as an excellent monograph on U-statistics. For univariate U-statistics ($d = 1$), the asymptotic distributions are derived in the seminal paper [20] for the nondegenerate case and in [34] for the degenerate case. There is also a large literature on bootstrap methods for univariate U-statistics [1, 4, 6, 14, 23, 24, 39]. A more recent interest lies in the high-dimensional case where d is much larger than n. Chen [8] develops Gaussian and bootstrap approximations for nondegenerate U-statistics of order two in high dimensions, which extends the work [11, 12] from sample averages to U-statistics; see also [17].

However, a major obstacle of inference using the complete U-statistic (1.1) is its computational intractability. Namely, the computation of the complete U-statistic (1.1) requires $O(n^r d)$ operations, and its computational cost can be prohibitively demanding even when n and d are moderately large, especially when the order of the U-statistic $r \ge 3$. For instance, the computation of a complete U-statistic with order 3 and dimension $d = 5000$ when the sample size is $n = 1000$ requires $\binom{n}{3} \times d \approx 0.8 \cdot 10^{12}$ (0.8 trillion) operations. In addition, the naive application of the empirical bootstrap for the U-statistic (1.1) requires even more operations, namely, $O(Bn^r d)$ operations, where B is the number of bootstrap repetitions.

This motivates us to study inference using *randomized incomplete U-statistics* with sparse weights instead of complete U-statistics. Specifically, we consider the Bernoulli sampling and sampling with replacement to construct random weights in Section 2. For a prespecified *computational budget parameter* $N \le |I_{n,r}|$, these sampling schemes randomly choose (on average) N indices from $I_{n,r}$, and the resulting incomplete U-statistics $U'_{n,N}$ are defined as the sample averages of $h(X_{i_1}, \dots, X_{i_r})$ taken over the subset of chosen indices $(i_1, \dots, i_r)$. Hence the computational cost of the incomplete U-statistics is reduced to $O(Nd)$, which can be much smaller than $n^r d$ as long as $N \ll n^r$ and can be made independent of the order of the U-statistic provided that N does not depend on r.

The goal of this paper is to develop computationally scalable and statistically correct inferential methods for the incomplete U-statistics with high-dimensional kernels and massive data, where d is possibly much larger than n but n can be also large. Specifically, we study distributional approximations to the randomized incomplete U-statistics in high dimensions. Our first main contribution is to derive Gaussian approximation error bounds for the incomplete U-statistics on the hyperrectangles in $\mathbb{R}^d$ for both nondegenerate and degenerate kernels. In Section 3, we

show that the derived Gaussian approximation results display an interesting computational and statistical trade-off for nondegenerate kernels (see Remark 3.1), and reveal a fundamental difference between complete and randomized incomplete U-statistics for degenerate kernels (see Remark 3.2). The mathematical insight of introducing the random weights is to create the (conditional) independence for the terms in the U-statistic sum in order to obtain a Gaussian limit. The Gaussian approximation results are, however, often not directly applicable since the covariance matrices of the approximating Gaussian distributions depend on the underlying distribution P that is unknown in practice. Our second contribution is to propose fully data-dependent bootstrap methods for incomplete U-statistics that are computationally (much) less demanding than existing bootstrap methods for U-statistics [1, 8, 9]. Specifically, we introduce generic bootstraps for incomplete U-statistics in Section 4.1. Our generic bootstrap constructions are flexible enough to cover both nondegenerate and degenerate kernels, and meanwhile they take the computational concern into account for estimating the associated (and unobserved) Hájek projection in the nondegenerate case. In particular, we propose two concrete estimation procedures for the Hájek projection: one is a deterministic construction based on the divide-and-conquer algorithm (Section 4.2), and another is a random construction based on a second randomization independent of everything else (Section 4.3). For both constructions, the overall computational complexity of the bootstrap methods can be made independent of the U-statistic order r.

As a leading example to illustrate the usefulness of the inferential methods developed in the present paper, we consider testing for the pairwise independence of a high-dimensional random vector $X = (X^{(1)}, \ldots, X^{(p)})^T$, that is, testing for the hypothesis that

$$H_0 : X^{(1)}, \ldots, X^{(p)} \quad \text{are pairwise independent.} \tag{1.2}$$

Let $X_1, \ldots, X_n$ be i.i.d. copies of X. Several dependence measures are proposed in the literature, including: Kendall's τ, Spearman's ρ, Hoeffding's D [21], Bergsma and Dassios' t^* [2] and the distance covariance [36, 40], all of which can be estimated by U-statistics. So various nonparametric tests for H_0 can be constructed based on those U-statistics. To compute the test statistics, we have to compute U-statistics with dimension $d = p(p-1)/2$, which corresponds to the number of upper triangular entries in the $p \times p$ dependence matrix and can be quite large. In addition, the orders of the U-statistics are at least 3 (except for Kendall's τ which is of order 2). So the computation of the test statistics is prohibitively demanding, not to mention the empirical bootstrap or subsampling for those U-statistics. It should be noted that there are efficient algorithms to reduce the computational costs for computing some of those U-statistics (cf. [28], Section 6.1), but such computational simplifications are case-by-case and not generically applicable, and more importantly they do not yield computationally tractable methods to approximate or estimate the sampling distributions of the U-statistics. The Gaussian and

bootstrap approximation theorems developed in the present paper can be applicable to calibrating critical values for test statistics based upon incomplete versions of those U-statistics. Detailed comparisons and discussions of nonparametric pairwise independence test statistics are presented in Section 5. In addition to pairwise independence testing, values of the dependence measures are also interesting *per se* in some applications. For instance, Spearman's ρ is related to the copula correlation if the marginal distributions are continuous ([15], Chapter 8) and our bootstrap methods can be used to construct simultaneous confidence intervals for the copula correlations uniformly over many pairs of variables.

To verify the finite sample performance of the proposed bootstrap methods for randomized incomplete U-statistics, we conduct simulation experiments in Section 5 on the leading example for nonparametric testing for the pairwise independence hypothesis in (1.2). Specifically, we consider to approximate the null distributions of the incomplete versions of the (leading term of) Spearman ρ and Bergsma–Dassios' t^*-test statistics, and examine the cases where $n = 300, 500, 1000$ and $p = 30, 50, 100$ (and hence $d = p(p-1)/2 = 435, 1225, 4950$). Statistically, we observe that the Gaussian approximation of the test statistics is quite accurate and the empirical rejection probability of the null hypothesis with the critical values calibrated by our bootstrap methods is very close to the nominal size for (almost) all setups. Computationally, we find that the (log-) running time for our bootstrap methods scales linearly with the (log-)sample size, and in addition, the slope coefficient matches very well with the computational complexity of the bootstrap methods. Therefore, the simulation results demonstrate a promising agreement between the empirical evidences and our theoretical analysis.

1.1. *Existing literature.* Incomplete U-statistics are first considered in [5], and the asymptotic distributions of incomplete U-statistics (for fixed d) are derived in [7] and [25]; see also Section 4.3 in [27] for a review on incomplete U-statistics. Closely related to the present paper is [25], which establishes the asymptotic properties of univariate incomplete U-statistics based on sampling with and without replacement and Bernoulli sampling. To the best of our knowledge, the present paper is the first paper that establishes approximation theorems for the distributions of randomized incomplete U-statistics in high dimensions. See also Remark 3.4 for more detailed comparisons with [25]. Incomplete U-statistics can be viewed as a special case of weighted U-statistics, and there is a large literature on limit theorems for weighted U-statistics; see [19, 22, 29, 32, 33, 35] and references therein. These references focus on the univariate case and do not cover the high-dimensional case. There are few references that study data-dependent inferential procedures for incomplete U-statistics that take computational considerations into account. An exception is [3], which proposes several inferential methods for univariate (generalized) incomplete U-statistics, but do not develop formal asymptotic

justifications for these methods. It is also interesting to note that incomplete U-statistics have gained renewed interests in the recent statistics and machine learning literatures [13, 30], although the focuses of these references are substantially different from ours.

From a technical point of view, this paper builds on recent development of Gaussian and bootstrap approximation theorems for averages of independent high-dimensional random vectors [11, 12] and for high-dimensional U-statistics of order two [8]. Importantly, however, developing Gaussian approximations for the randomized incomplete U-statistics in high dimensions requires a novel proof-strategy that combines iterative conditioning arguments and applications of Berry–Esseen-type bounds, and extends some of results in [8] to cover general order incomplete U-statistics. In addition, these references do not consider bootstrap methods for incomplete U-statistics that take computational considerations into account.

1.2. *Organization*. The rest of the paper is organized as follows. In Section 2, we introduce randomized incomplete U-statistics with sparse weights generated from the Bernoulli sampling and sampling with replacement. In Section 3, we derive nonasymptotic Gaussian approximation error bounds for the randomized incomplete U-statistics in high dimensions for both nondegenerate and degenerate kernels. In Section 4, we first propose generic bootstrap methods for the incomplete U-statistics and then incorporate the computational budget constraint by two concrete estimates of the Hájek projection: one deterministic estimate by the divide and conquer, and one randomized estimate by incomplete U-statistics of a lower order. Simulation examples are provided in Section 5 and in the Supplementary Material (SM) [10]. All the technical proofs are gathered in Appendix C in the SM. We conclude the paper in Section 6 with a brief discussion on some extensions.

1.3. *Notation*. For a hyperrectangle $R = \prod_{j=1}^{d}[a_j, b_j]$ in $\mathbb{R}^d$, a constant $c > 0$, and a vector $y = (y_1, \ldots, y_d)^T \in \mathbb{R}^d$, we use the notation $[cR + y] = \prod_{j=1}^{d}[ca_j + y_j, cb_j + y_j]$. For vectors $y = (y_1, \ldots, y_d)^T, z = (z_1, \ldots, z_d)^T \in \mathbb{R}^d$, the notation $y \le z$ means that $y_j \le z_j$ for all $j = 1, \ldots, d$. For $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a finite set J, $|J|$ denotes the cardinality of J. Let $|\cdot|_\infty$ denote the max-norm for vectors and matrices, that is, for a matrix $A = (a_{ij})$, $|A|_\infty = \max_{i,j}|a_{ij}|$. "Constants" refer to finite, positive and nonrandom numbers.

For $0 < \beta < \infty$, let ψ_β be the function on $[0, \infty)$ defined by $\psi_\beta(x) = e^{x^\beta} - 1$, and for a real-valued random variable ξ, define $\|\xi\|_{\psi_\beta} = \inf\{C > 0 : \mathbb{E}[\psi_\beta(|\xi|/C)] \le 1\}$. For $\beta \in [1, \infty)$, $\|\cdot\|_{\psi_\beta}$ is an Orlicz norm, while for $\beta \in (0, 1)$, $\|\cdot\|_{\psi_\beta}$ is not a norm but a quasi-norm, that is, there exists a constant C_β depending only on β such that $\|\xi_1 + \xi_2\|_{\psi_\beta} \le C_\beta(\|\xi_1\|_{\psi_\beta} + \|\xi_2\|_{\psi_\beta})$. (Indeed, there is a norm

equivalent to $\|\cdot\|_{\psi_\beta}$ obtained by linearizing ψ_β in a neighborhood of the origin; cf. Lemma C.2 in the SM.)

For a generic random variable Y, let $\mathbb{P}_{|Y}(\cdot)$ and $\mathbb{E}_{|Y}[\cdot]$ denote the conditional probability and expectation given Y, respectively. For a given probability space $(\mathcal{X}, \mathcal{A}, Q)$ and a measurable function f on $\mathcal{X}$, we use the notation $Qf = \int f\,dQ$ whenever the latter integral is well defined. For a jointly measurable symmetric function f on S^r and $k = 1, \dots, r$, let $P^{r-k}f$ denote the function on S^k defined by

$$P^{r-k}f(x_1, \dots, x_k) = \int \cdots \int f(x_1, \dots, x_k, x_{k+1}, \dots, x_r)\,dP(x_{k+1}) \cdots dP(x_r)$$

whenever the integral exists and is finite for every $(x_1, \dots, x_k) \in S^k$. For given $1 \le k \le \ell \le n$, we use the notation $X_k^\ell = (X_k, \dots, X_\ell)$. Throughout the paper, we assume that $n \ge 4 \vee r$ and $d \ge 3$.

2. Randomized incomplete U-statistics. In this paper, to construct sparsely weighted U-statistics, we shall use random sparse weights. For $\iota = (i_1, \dots, i_r) \in I_{n,r}$, let us write $X_\iota = (X_{i_1}, \dots, X_{i_r})$, and observe that the complete U-statistic (1.1) can be written as

$$U_n = \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} h(X_\iota).$$

Now, let $N := N_n$ be an integer such that $0 < N \le |I_{n,r}|$, and let $p_n = N/|I_{n,r}|$. Instead of taking the average over all possible ι in $I_{n,r}$, we will take the average over a subset of about N indices chosen randomly from $I_{n,r}$. In the present paper, we study Bernoulli sampling and sampling with replacement.

2.1. *Bernoulli sampling.* Generate i.i.d. $\mathsf{Ber}(p_n)$ random variables $\{Z_\iota : \iota \in I_{n,r}\}$ with success probability p_n, that is, $Z_\iota, \iota \in I_{n,r}$ are i.i.d. with $\mathbb{P}(Z_\iota = 1) = 1 - \mathbb{P}(Z_\iota = 0) = p_n$. Consider the following weighted U-statistic with random weights:

$$U'_{n,N} = \frac{1}{\widehat{N}} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota), \tag{2.1}$$

where $\widehat{N} = \sum_{\iota \in I_{n,r}} Z_\iota$ is the number of nonzero weights. We call $U'_{n,N}$ the randomized incomplete U-statistic based on the Bernoulli sampling. The variable $\widehat{N}$ follows $\mathsf{Bin}(|I_{n,r}|, p_n)$, the binomial distribution with parameters $(|I_{n,r}|, p_n)$. Hence $\mathbb{E}[\widehat{N}] = |I_{n,r}|p_n = N$ and the computation of the incomplete U-statistic (2.1) only requires $O(Nd)$ operations on average. In addition, by Bernstein's inequality (cf. Lemma 2.2.9 in [38]),

$$\mathbb{P}\big(|\widehat{N}/N - 1| > \sqrt{2t/N} + 2t/(3N)\big) \le 2e^{-t} \tag{2.2}$$

for every $t > 0$, and hence $\widehat{N}$ concentrates around its mean N. Therefore, we can view N as a *computational budget parameter* and p_n as a *sparsity design parameter* for the incomplete U-statistic.

The reader may wonder that generating $|I_{n,r}| \approx n^r$ Bernoulli random variables is computationally demanding, but there is no need to do so. In fact, we can equivalently compute the randomized incomplete U-statistic in (2.1) as follows:

1. Generate $\widehat{N} \sim \mathsf{Bin}(|I_{n,r}|, p_n)$.
2. Choose indices $\iota_1, \ldots, \iota_{\widehat{N}}$ randomly without replacement from $I_{n,r}$.
3. Compute $U'_{n,N} = \widehat{N}^{-1} \sum_{j=1}^{\widehat{N}} h(X_{\iota_j})$.

In fact, define $Z_\iota = 1$ if ι is one of $\iota_1, \ldots, \iota_{\widehat{N}}$, and $Z_\iota = 0$ otherwise; then it is not difficult to see that $\{Z_\iota : \iota \in I_{n,r}\}$ are i.i.d. $\mathsf{Ber}(p_n)$ random variables. So, we can think of the Bernoulli sampling as a sampling without replacement with a random sample size.

REMARK 2.1 (Comments on the random normalization). Interestingly, changing the normalization in (2.1) *does* affect approximating distributions to the resulting incomplete U-statistic. Namely, if we change $\widehat{N}$ to N in that is, $\breve{U}'_{n,N} = N^{-1} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota)$, then we have different approximating distributions unless $\theta = 0$. In general, changing $\widehat{N}$ to N in (2.1) results in the approximating Gaussian distributions with larger covariance matrices, and hence it is recommended to use $U'_{n,N}$ rather than $\breve{U}'_{n,N}$. See also Remark 3.3 ahead.

2.2. *Sampling with replacement.* Conditionally on $X_1^n = (X_1, \ldots, X_n)$, let $X^*_{\iota_j}$, $j = 1, \ldots, N$ be i.i.d. draws from the empirical distribution $|I_{n,r}|^{-1} \times \sum_{\iota \in I_{n,r}} \delta_{X_\iota}$ (δ_{X_ι} denotes the point mass at X_ι). Let

$$U'_{n,N} = \frac{1}{N} \sum_{j=1}^{N} h\big(X^*_{\iota_j}\big) \tag{2.3}$$

be the incomplete U-statistic obtained by sampling with replacement. We call $U'_{n,N}$ the randomized incomplete U-statistic based on sampling with replacement. Observe that $U'_{n,N}$ in (2.3) can be efficiently computed by sampling r distinct terms from $\{X_1, \ldots, X_n\}$ independently for N times. The statistic $U'_{n,N}$ can be written as a weighted U-statistic. Indeed, for each $\iota \in I_{n,r}$, let Z_ι denote the number of times that X_ι is redrawn in the sample $\{X^*_{\iota_1}, \ldots, X^*_{\iota_N}\}$. Then the vector $Z = (Z_\iota)_{\iota \in I_{n,r}}$ (ordered in an arbitrary way) follows a multinomial distribution with parameters N and probabilities $1/|I_{n,r}|, \ldots, 1/|I_{n,r}|$ independent of X_1^n, and $U'_{n,N}$ can be written as

$$U'_{n,N} = \frac{1}{N} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota). \tag{2.4}$$

Hence we can think of $U'_{n,N}$ as a statistic of $X_1, \ldots, X_n$ and $Z_\iota, \iota \in I_{n,r}$, but we will use both representations (2.3) and (2.4) interchangeably in the subsequent analysis.

REMARK 2.2. All the theoretical results presented below apply to incomplete U-statistics based on either the Bernoulli sampling or sampling with replacement. Both sampling schemes will be covered in a unified way.

3. Gaussian approximations. In this section, we will derive Gaussian approximation results for the incomplete U-statistics (2.1) and (2.3) on the hyperrectangles in $\mathbb{R}^d$. Let $\mathcal{R}$ denote the class of (closed) hyperrectangles in $\mathbb{R}^d$, that is, $\mathcal{R}$ consists sets of the form $\prod_{j=1}^d [a_j, b_j]$ where $-\infty \le a_j \le b_j \le \infty$ for $j = 1, \ldots, d$ with the convention that $[a_j, b_j] = (-\infty, b_j]$ for $a_j = -\infty$ and $[a_j, b_j] = [a_j, \infty)$ for $b_j = \infty$. For the expository purpose, we mainly focus on the nondegenerate case where $\min_{1 \le j \le d} \mathrm{Var}(\mathbb{E}[h_j(X_1, \ldots, X_r) \mid X_1])$ is bounded away from zero in the following discussion. However, our Gaussian approximation results also cover the degenerate case (cf. Theorem 3.3). The intuition behind and the proof sketch for the Gaussian approximation results are given in Section C.2 in the SM.

To state the formal Gaussian approximation results, we assume the following conditions. Let $\underline{\sigma} > 0$ and $D_n \ge 1$ be given constants, and define $g := (g_1, \ldots, g_d)^T := P^{r-1}h$. Suppose that:

(C1) $P^r |h_j|^{2+k} \le D_n^k$ for all $j = 1, \ldots, d$ and $k = 1, 2$.
(C2) $\|h_j(X_1^r)\|_{\psi_1} \le D_n$ for all $j = 1, \ldots, d$.

In addition, suppose that either one of the following conditions holds:

(C3-ND) $P(g_j - \theta_j)^2 \ge \underline{\sigma}^2$ for all $j = 1, \ldots, d$.
(C3-D) $P^r(h_j - \theta_j)^2 \ge \underline{\sigma}^2$ for all $j = 1, \ldots, d$.

Conditions (C1) and (C2) are adapted from [12] and [8]. Condition (C2) assumes the kernel h to be subexponential, which in particular covers bounded kernels. In principle, it is possible to extend our analysis under milder moment conditions on the kernel h, but this would result in more involved error bounds. For the sake of clear presentation, we mainly work with Condition (C2) and point out the differences when the kernel satisfies a polynomial moment condition in Remark 3.5. By Jensen's inequality, Conditions (C1) and (C2) imply that $P|g_j|^{2+k} \le D_n^k$ for all j and for $k = 1, 2$, and $\|g_j(X_1)\|_{\psi_1} \le D_n$ for all j. Here, we allow the exponential moment bound D_n to depend on n since the distribution P may depend on n in the high-dimensional setting. In addition, Condition (C1) implies that $P^r h_j^2 \le 1 + P|h_j|^3 \le 1 + D_n$ for all j. Condition (C3-ND) implies that the kernel h is nondegenerate. In the degenerate case, we will require Condition (C3-D) to derive Gaussian approximations.

In all what follows, we assume that

$$p_n = N/|I_{n,r}| \le 1/2$$

without further mentioning. The value $1/2$ has no special meaning; we can allow $p_n \le c$ for any constant $c \in (0,1)$, and in that case, the constants appearing in the following theorems depend in addition on c. Since we are using randomization for the purpose of computational reduction, we are mainly interested in the case where $N \ll |I_{n,r}|$, and the assumption that p_n is bounded away from 1 is immaterial.

The following theorem derives bounds on the Gaussian approximation to the randomized incomplete U-statistics on the hyperrectangles in the case where the kernel h is nondegenerate. Recall that $\alpha_n = n/N$, $p_n = N/|I_{n,r}|$, $\theta = P^r h = Pg$, $\Gamma_g = P(g-\theta)(g-\theta)^T$, and $\Gamma_h = P^r(h-\theta)(h-\theta)^T$.

THEOREM 3.1 (Gaussian approximation under nondegeneracy). *Suppose that Conditions* (C1), (C2) *and* (C3-ND) *hold. Then there exists a constant C depending only on $\underline{\sigma}$ and r such that*

$$\begin{aligned}&\sup_{R\in\mathcal{R}} \bigl|\mathbb{P}\bigl\{\sqrt{n}\bigl(U'_{n,N}-\theta\bigr)\in R\bigr\} - \mathbb{P}(Y\in R)\bigr| \\ &\quad= \sup_{R\in\mathcal{R}} \bigl|\mathbb{P}\bigl\{\sqrt{N}\bigl(U'_{n,N}-\theta\bigr)\in R\bigr\} - \mathbb{P}\bigl(\alpha_n^{-1/2}Y\in R\bigr)\bigr| \\ &\quad\le C\biggl(\frac{D_n^2\log^7(dn)}{n\wedge N}\biggr)^{1/6}, \qquad (3.1)\end{aligned}$$

where $Y \sim N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$.

Theorem 3.1 shows that the distribution of $\sqrt{n}(U'_{n,N}-\theta)$ can be approximated by the Gaussian distribution $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$ on the hyperrectangles provided that $D_n^2\log^7(dn) \ll n\wedge N$, from which we deduce that the Gaussian approximation on the hyperrectangles holds for $U'_{n,N}$ even when $d \gg n$. Asymptotically, if, for example, D_n is bounded in n and $N \ge n$, then as $n\to\infty$,

$$\sup_{R\in\mathcal{R}} \bigl|\mathbb{P}\bigl\{\sqrt{n}\bigl(U'_{n,N}-\theta\bigr)\in R\bigr\} - \mathbb{P}(Y\in R)\bigr| \to 0$$

whenever $d = d_n$ satisfies that $\log d = o(n^{1/7})$, so that the high-dimensional CLT on the hyperrectangles holds for the incomplete U-statistics even in ultrahigh-dimensional cases where d is much larger than n. Similar comments apply to all the other results we will derive.

For complete and nondegenerate U-statistics (a special case of incomplete U-statistics with the complete design and $N = |I_{n,r}|$), it has been argued in [12] ($r = 1$) and [8] ($r = 2$) that the rate of convergence in Theorem 3.1 is nearly optimal in the regime where d grows subexponentially fast in n. On the other hand, the rate of convergence can be improved to $n^{-1/4}$ (up to logarithmic factors) if $d = O(n^{1/7})$, namely if the dimension increases at most polynomially fast with the sample size.

In the cases where $N \gg n$ (i.e., $\alpha_n \ll 1$) and $N \ll n$ (i.e., $\alpha_n \gg 1$), the approximating distribution can be simplified to $N(0, r^2\Gamma_g)$ and $N(0,\Gamma_h)$, respectively.

COROLLARY 3.2. *Suppose that Conditions* (C1), (C2) *and* (C3-ND) *hold. Then there exists a constant C depending only on $\underline{\sigma}$ and r such that*

$$\sup_{R\in\mathcal{R}}\left|\mathbb{P}\{\sqrt{n}(U'_{n,N}-\theta)\in R\}-\gamma_A(R)\right| \le C\left\{\left(\frac{nD_n\log^2 d}{N}\right)^{1/3}+\left(\frac{D_n^2\log^7(dn)}{n\wedge N}\right)^{1/6}\right\},$$

where $\gamma_A = N(0, r^2\Gamma_g)$, and

$$\sup_{R\in\mathcal{R}}\left|\mathbb{P}\{\sqrt{N}(U'_{n,N}-\theta)\in R\}-\gamma_B(R)\right| \le C\left\{\left(\frac{ND_n\log^2 d}{n}\right)^{1/3}+\left(\frac{D_n^2\log^7 d}{n\wedge N}\right)^{1/6}\right\},$$

where $\gamma_B = N(0, \Gamma_h)$.

REMARK 3.1 (Comments on the computational and statistical trade-off for the randomized incomplete U-statistics with nondegenerate kernels). Theorem 3.1 and Corollary 3.2 reveal an interesting phase transition phenomenon between the computational complexity and the statistical efficiency for the randomized incomplete U-statistics. Suppose that $n\wedge N \gg D_n^2\log^7(dn)$ and $\underline{\sigma}$ is bounded away from zero. First, if the computational budget parameter N is *superlinear* in the sample size n (i.e., $N \gg nD_n\log^2 d$), then both the incomplete U-statistic $\sqrt{n}(U'_{n,N}-\theta)$ and its complete version $\sqrt{n}(U_n-\theta)$ can be approximated by the same Gaussian distribution $\gamma_A = N(0, r^2\Gamma_g)$ (cf. [8] for $r=2$ case). Second, if N is of the same order as n, then the scaling factor of $U'_{n,N}$ remains the same as for U_n, namely, $\sqrt{n}$. However, the approximating Gaussian distribution for $\sqrt{n}(U'_{n,N}-\theta)$ has covariance matrix $r^2\Gamma_g+\alpha_n\Gamma_h$, which is larger than the the corresponding covariance matrix $r^2\Gamma_g$ for $\sqrt{n}(U_n-\theta)$ in the sense that their difference $\alpha_n\Gamma_h$ is positive semidefinite. In this case, we sacrifice the statistical efficiency for the sake of keeping the computational cost linear in n. Third, if we further reduce the computational budget parameter N to be *sublinear* in n (i.e., $N \ll n/(D_n\log^2 d)$), then the scaling factor of $U'_{n,N}$ changes from $\sqrt{n}$ to $\sqrt{N}$, and the distribution of $U'_{n,N}$ is approximated by $N(\theta, N^{-1}\Gamma_h)$ on the hyperrectangles. Hence, the decay rate of the covariance matrix of the approximating Gaussian distribution is now N^{-1}, which is slower than the n^{-1} rate for the previous two cases.

Next, we consider the case where the kernel h is degenerate, that is, $P(g_j-\theta_j)^2=0$ for all $j=1,\dots,d$. We consider the case where the kernel h is degenerate of order $k-1$ for some $k=2,\dots,r$, that is, $P^{r-k+1}h(x_1,\dots,x_{k-1})=P^r h$ for all $(x_1,\dots,x_{k-1})\in S^{k-1}$. Even in such cases, a Gaussian approximation holds for $\sqrt{N}(U'_{n,N}-\theta)$ on the hyperrectangles provided that $N \ll n^k$ up to logarithmic factors. More precisely, we obtain the following theorem.

THEOREM 3.3 (Gaussian approximation under degeneracy). *Suppose the kernel h is degenerate of order $k-1$ for some $k=2,\dots,r$. In addition, suppose that Conditions* (C1), (C2) *and* (C3-D) *hold. Then there exists a constant C depending only on $\underline{\sigma}$ and r such that*

$$\begin{aligned}\sup_{R\in\mathcal{R}}&\left|\mathbb{P}\{\sqrt{N}(U'_{n,N}-\theta)\in R\}-\gamma_B(R)\right|\\&\le C\left\{\left(\frac{ND_n^2\log^{k+3}d}{n^k}\right)^{1/4}+\left(\frac{D_n^2(\log n)\log^5(dn)}{n}\right)^{1/6}\right.\\&\qquad\left.+\left(\frac{D_n^2\log^7(dn)}{N}\right)^{1/6}\right\},\end{aligned}\tag{3.2}$$

where $\gamma_B=N(0,\Gamma_h)$.

REMARK 3.2 (Comments on the Gaussian approximation under degeneracy). In the degenerate case, for the Gaussian approximation to hold, we must have $N\ll n^k$ (more precisely, $N\ll n^k/(D_n^2\log^{k+3}d)$), which is an indispensable condition even for the $d=1$ case. To see this, consider the Bernoulli sampling case (similar arguments apply to the sampling with replacement case) and observe that $\sqrt{N}(U'_{n,N}-\theta)=(N/\widehat{N})\cdot\sqrt{N}W_n=(N/\widehat{N})(\sqrt{N}A_n+\sqrt{N(1-p_n)}B_n)$, where $A_n=U_n-\theta$ and $B_n=U'_{n,N}-U_n$. According to Theorem 12.10 in [37], $n^{k/2}A_n$ converges in distribution to a Gaussian chaos of order k. Hence, in order to approximate $\sqrt{N}(U'_{n,N}-\theta)\approx\sqrt{N}W_n$ by a Gaussian distribution, it is necessary that $\sqrt{N}A_n$ is stochastically vanishing, which leads to the condition $N\ll n^k$.

It is worth noting that Theorem 3.3 reveals a fundamental difference between complete and randomized incomplete U-statistics with the degenerate kernel. Namely, in the degenerate case, the complete U-statistic $n^{k/2}(U_n-\theta)$ is known to have a non-Gaussian limiting distribution when d is fixed, while thanks to the randomizations, our incomplete U-statistics $\sqrt{N}(U'_{n,N}-\theta)$ can be approximated by the Gaussian distribution, and in addition the Gaussian approximation can hold even when $d\gg n$. On one hand, the rate of convergence of the incomplete U-statistics is $N^{-1/2}$ and is slower than that of the complete U-statistic, namely, $n^{-k/2}$. So in that sense we are sacrificing the rate of convergence by using the incomplete U-statistics instead of the complete U-statistic, although the rate $N^{-1/2}$ can be arbitrarily close to $n^{-k/2}$ up to logarithmic factors. On the other hand, the approximating Gaussian distribution for the incomplete U-statistics is easy to estimate by using a multiplier bootstrap developed in Section 4. The multiplier bootstrap developed in Section 4 is computationally much less demanding than, for example, the empirical bootstraps for complete (degenerate) U-statistics (cf. [1, 6]), and can consistently estimate the approximating Gaussian distribution γ_B on the hyperrectangles even when $d\gg n$; see Theorem 4.1. To the best of our knowledge, there is no existing work that formally derives Gaussian chaos approximations to degenerate U-statistics in high dimensions where $d\gg n$, and in addition

such non-Gaussian approximating distributions appear to be more difficult to estimate in high dimensions. Hence, in the degenerate case, the randomizations not only reduce the computational cost but also provide more tractable alternatives to make statistical inference on θ in high dimensions.

REMARK 3.3 (Effect of deterministic normalization in the Bernoulli sampling case). In the Bernoulli sampling case, consider the deterministic normalization, that is, $\breve{U}'_{n,N} = N^{-1}\sum_{\iota\in I_{n,r}} Z_\iota h(X_\iota)$, instead of the random one, that is, $U'_{n,N} = \widehat{N}^{-1}\sum_{\iota\in I_{n,r}} Z_\iota h(X_\iota)$. Then, in the nondegenerate case, the distribution of $\sqrt{n}(\breve{U}'_{n,N} - \theta)$ can be approximated by $N(0, r^2\Gamma_g + \alpha_n P^r hh^T)$, and in the degenerate case, $\sqrt{N}(\breve{U}'_{n,N} - \theta)$ can be approximated by $N(0, P^r hh^T)$ (provided that $N \ll n^k$ for the degenerate case). To see this, observe that $\breve{U}'_{n,N} - \theta = (U_n - \theta) + N^{-1}\sum_{\iota\in I_{n,r}}(Z_\iota - p_n)h(X_\iota)$, and the distribution of $N^{-1}\sum_{\iota\in I_{n,r}}(Z_\iota - p_n)h(X_\iota)$ can be approximated by $N(0, (1-p_n)P^r hh^T)$. Since $P^r hh^T$ is larger than Γ_h unless $\theta = 0$ (in the sense that $P^r hh^T - \Gamma_h = \theta\theta^T$ is positive semidefinite), the approximating Gaussian distributions have larger covariance matrices for $\breve{U}'_{n,N}$ than those for $U'_{n,N}$, and hence it is in general recommended to use the random normalization rather than the deterministic one. A numerical comparison between these normalizations can be found in Section E of the SM.

REMARK 3.4 (Comparisons with [25] for $d = 1$). The Gaussian approximation results established in Theorems 3.1, 3.3 and Corollary 3.2 can be considered as (partial) extensions of Theorem 1 and Corollary 1 in [25] to high dimensions. Janson [25] focuses on the univariate case ($d = 1$) and derives the asymptotic distributions of randomized incomplete U-statistics based on sampling without replacement, sampling with replacement and Bernoulli sampling ([25] considers the deterministic normalization for the Bernoulli sampling case). For the illustrative purpose, consider sampling with replacement. Suppose that $p_n \to p \in [0, 1]$ and the kernel h is degenerate of order $k-1$ for some $k = 1, \ldots, r$ (the $k = 1$ case corresponds to a nondegenerate kernel). Then Theorem 1 in [25] shows that $(n^{k/2}(U_n - \theta), N^{1/2}(U'_{n,N} - U_n)) \xrightarrow{d} (V, W)$, where V is a Gaussian chaos of order k (in particular, $V \sim N(0, r^2 P(g-\theta)^2)$ if $k = 1$) and $W \sim N(0, P^r(h-\theta)^2)$ such that V and W are independent. Hence, provided that $n^k/N \to \alpha \in [0, \infty]$, $n^{k/2}(U'_{n,N} - \theta) \xrightarrow{d} V + \alpha W$ if $\alpha < \infty$ and $\sqrt{N}(U'_{n,N} - \theta) \xrightarrow{d} W$ if $\alpha = \infty$. The present paper focuses on the cases where the approximating distributions are Gaussian (i.e., the cases where $k = 1$ and α is finite, or $k \geq 2$ and $\alpha = \infty$), but covers high-dimensional kernels and derives explicit and nonasymptotic Gaussian approximation error bounds that are not obtained in [25]. In addition, the proof strategy of our Gaussian approximation results differs substantially from that of [25]. Janson [25] shows the convergence of the joint characteristic function of $(n^{k/2}(U_n - \theta), N^{1/2}(U'_{n,N} - U_n))$ to obtain his Theorem 1, but the characteristic

function approach is not very useful to derive explicit error bounds on distributional approximations in high dimensions. Instead, our proofs iteratively use conditioning arguments combined with Berry–Esseen-type bounds.

Finally, we expect that the results of the present paper can be extended to the case where $k \geq 2$ and α is finite; in that case, the approximating distribution to $n^{k/2}(U'_{n,N} - \theta)$ will be non-Gaussian and the technical analysis will be more involved in high dimensions. We leave the analysis of this case as a future research topic.

REMARK 3.5 (Relaxation of subexponential moment Condition (C2)). It is possible to relax the subexponential moment Condition (C2) to a polynomial moment condition. Suppose that

(C2′) $(P^r|h|_\infty^q)^{1/q} \leq D_n$ for some $q \in [4, \infty)$.

Condition (C2′) covers a kernel with bounded polynomial moment of a finite degree q.

THEOREM 3.4 (Gaussian approximation under polynomial moment condition). (i) *If Conditions* (C1), (C2′) *and* (C3-ND) *hold, then there exists a constant C depending only on $\underline{\sigma}$, r and q such that*

$$\begin{aligned}&\sup_{R\in\mathcal{R}}\left|\mathbb{P}\left\{\sqrt{n}\left(U'_{n,N}-\theta\right)\in R\right\}-\mathbb{P}(Y\in R)\right|\\&\qquad\leq C\left\{\left(\frac{D_n^2\log^7(dn)}{n\wedge N}\right)^{1/6}+\left(\frac{D_n^2n^{2r/q}\log^3(dn)}{(n\wedge N)^{1-2/q}}\right)^{1/3}\right\},\end{aligned}\tag{3.3}$$

where $Y \sim N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$.

(ii) *Suppose the kernel h is degenerate of order $k-1$ for some $k = 2, \ldots, r$. If Conditions* (C1), (C2′) *and* (C3-D) *hold, then there exists a constant C depending only on $\underline{\sigma}$, r and q such that*

$$\begin{aligned}&\sup_{R\in\mathcal{R}}\left|\mathbb{P}\left\{\sqrt{N}\left(U'_{n,N}-\theta\right)\in R\right\}-\gamma_B(R)\right|\\&\qquad\leq C\left\{\left(\frac{ND_n^2\log^{k+3}d}{n^k}\right)^{1/4}+\left(\frac{D_n^2\log^5(dn)}{n}\right)^{1/6}\right.\\&\qquad\quad\left.+\left(\frac{D_n^2\log^7 d}{N}\right)^{1/6}+\left(\frac{D_n^2n^{2r/q}\log^3 d}{(n\wedge N)^{1-2/q}}\right)^{1/3}\right\},\end{aligned}\tag{3.4}$$

where $\gamma_B = N(0, \Gamma_h)$.

Comparing Theorem 3.4 with Theorems 3.1 and 3.3, we see that the same approximating Gaussian distributions under the subexponential moment condition (C2) are valid under the polynomial moment condition (C2′) as well. The rates

of convergence to the Gaussian distributions under (C2′) involve an extra Nagaev-type term similar to the sample average and complete U-statistic cases (cf. [8, 12]), and so the rates may be slower than those obtained under the subexponential moment condition (C2). In particular, the rates in (3.3) and (3.4) now depend on the order r through the term $n^{2r/q}$. Still, the leading orders in (3.3) and (3.4) coincide with those under the subexponential moment condition (C2) as long as q is sufficiently large compared with r. For example, if D_n is bounded in n, $N \geq n$, and $q \geq 4(r+1)$, then the leading order of (3.3) is $(n^{-1}\log^7(dn))^{1/6}$, which coincides with that in the subexponential case.

4. Bootstrap approximations. The Gaussian approximation results developed in the previous section are often not directly applicable in statistical applications since the covariance matrix of the approximating Gaussian distribution, $r^2\Gamma_g + \alpha_n\Gamma_h$ (or Γ_h in the degenerate case), is unknown to us. In this section, we develop data-dependent procedures to further approximate or estimate the $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$ distribution (or the $N(0,\Gamma_h)$ distribution in the degenerate case) that are computationally (much) less-demanding than existing bootstrap methods for U-statistics such as the empirical bootstrap.

4.1. *Generic bootstraps for incomplete U-statistics.* Let $\mathcal{D}_n = \{X_1, \ldots, X_n\} \cup \{Z_\iota : \iota \in I_{n,r}\}$. For the illustrative purpose, consider to estimate the $N(0, r^2\Gamma + \alpha_n\Gamma_h)$ distribution and let $Y \sim N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$. The basic idea of our approach is as follows. Since $Y \overset{d}{=} Y_A + \alpha_n^{1/2} Y_B$ where $Y_A \sim N(0, r^2\Gamma_g)$ and $Y_B \sim N(0,\Gamma_h)$ are independent, to approximate the distribution of Y, it is enough to construct data-dependent random vectors $U_{n,A}^\sharp$ and $U_{n,B}^\sharp$ such that, conditionally on $\mathcal{D}_n$, (i) $U_{n,A}^\sharp$ and $U_{n,B}^\sharp$ are independent, and (ii) the conditional distributions of $U_{n,A}^\sharp$ and $U_{n,B}^\sharp$ are computable and "close" enough to $N(0, r^2\Gamma_g)$ and $N(0,\Gamma_h)$, respectively. Then the conditional distribution of $U_n^\sharp = U_{n,A}^\sharp + \alpha_n^{1/2} U_{n,B}^\sharp$ should be close to $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$, and hence to the distribution of $\sqrt{n}(U'_{n,N} - \theta)$. Of course, if the target distribution is $N(0, r^2\Gamma_g)$ or $N(0,\Gamma_h)$, then it is enough to simulate the conditional distribution of $U_{n,A}^\sharp$ or $U_{n,B}^\sharp$ alone, respectively.

Construction of $U_{n,B}^\sharp$ is straightforward; in fact, it is enough to apply the (Gaussian) multiplier bootstrap to $\sqrt{Z_\iota}h(X_\iota)$, $\iota \in I_{n,r}$.

Construction of $U_{n,B}^\sharp$.

1. Generate i.i.d. $N(0,1)$ variables $\{\xi'_\iota : \iota \in I_{n,r}\}$ independent of the data $\mathcal{D}_n$.
2. Construct

$$U_{n,B}^\sharp = \frac{1}{\sqrt{\widehat{N}}} \sum_{\iota \in I_{n,r}} \xi'_\iota \sqrt{Z_\iota}\{h(X_\iota) - U'_{n,N}\},$$

where $\widehat{N}$ is replaced by N for the sampling with replacement case.

In the Bernoulli sampling case, $U_{n,B}^{\sharp}$ reduces to $U_{n,B}^{\sharp} = \widehat{N}^{-1/2} \sum_{j=1}^{\widehat{N}} \xi'_{\iota_j} \times \{h(X_{\iota_j}) - U'_{n,N}\}$, while in the sampling with replacement case, simulating $U_{n,B}^{\sharp}$ can be equivalently implemented by simulating $U_{n,B}^{\sharp} = N^{-1/2} \sum_{j=1}^{N} \eta_j \{h(X_{\iota_j}^*) - U'_{n,N}\}$ for $\eta_1, \ldots, \eta_N \sim N(0,1)$ i.i.d. independent of $X_{\iota_1}^*, \ldots, X_{\iota_N}^*$; in fact, the distribution of $U_{n,B}^{\sharp}$ in the latter definition (conditionally on $X_{\iota_1}^*, \ldots, X_{\iota_N}^*$) is Gaussian with mean zero and covariance matrix $N^{-1} \sum_{j=1}^{N} \{h(X_{\iota_j}^*) - U'_{n,N}\}\{h(X_{\iota_j}^*) - U'_{n,N}\}^T$, which is identical to the conditional distribution of $U_{n,B}^{\sharp}$ in the original definition. In either case, in practice, we only need to generate (on average) N multiplier variables. The following theorem establishes conditions under which the conditional distribution of $U_{n,B}^{\sharp}$ is able to consistently estimate the $N(0, \Gamma_h)$ $(= \gamma_B)$ distribution on the hyperrectangles with polynomial error rates.

THEOREM 4.1 (Validity of $U_{n,B}^{\sharp}$). *Suppose that* (C1), (C2) *and* (C3-D) *hold. If*

$$\frac{D_n^2 (\log^2 n) \log^5(dn)}{n \wedge N} \leq C_1 n^{-\zeta} \tag{4.1}$$

for some constants $0 < C_1 < \infty$ *and* $\zeta \in (0,1)$, *then there exists a constant* C *depending only on* $\underline{\sigma}, r$, *and* C_1 *such that*

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n} \left(U_{n,B}^{\sharp} \in R \right) - \gamma_B(R) \right| \leq C n^{-\zeta/6} \tag{4.2}$$

with probability at least $1 - Cn^{-1}$.

REMARK 4.1 (Bootstrap validity under the polynomial moment condition). Analogous bootstrap validity results for $U_{n,B}^{\sharp}$ in Theorem 4.1, as well as those for $U_n^{\sharp}$ and $U_{n,A}^{\sharp}$ in Theorem 4.2, 4.3 and Proposition 4.4, 4.5 ahead, can be obtained under the polynomial moment Condition (C2′). Due to the space concern, detailed results can be found in Section B of the SM.

In the degenerate case, the approximating distribution is $\gamma_B = N(0, \Gamma_h)$. So, in that case, we can approximate the distribution of $\sqrt{N}(U'_{n,N} - \theta)$ on the hyperrectangles by the conditional distribution of $U_{n,B}^{\sharp}$, which can be simulated by drawing multiplier variables many times. We call the simulation of $U_{n,B}^{\sharp}$ the *multiplier bootstrap under degeneracy* (MB-DG). On average, the computational cost of the MB-DG is $O(BNd)$ (where B denotes the number of bootstrap iterations), which can be independent of the order of the U-statistic provided that N is so. In the remainder of this section, we will focus on the nondegenerate case.

In contrast to $U_{n,B}^{\sharp}$, construction of $U_{n,A}^{\sharp}$ is more involved. We might be tempted to apply the multiplier bootstrap to the Hájek projection, $rn^{-1} \sum_{i_1=1}^{n} g(X_{i_1})$, but

the function $g = P^{r-1}h$ is unknown so the direct application of the multiplier bootstrap to the Hájek projection is infeasible. Instead, we shall construct estimates of $g(X_{i_1})$ for $i_1 \in \{1, \dots, n\}$ or a subset of $\{1, \dots, n\}$, and then apply the multiplier bootstrap to the estimated Hájek projection. Generically, construction of $U_{n,A}^\sharp$ is as follows:

Generic construction of $U_{n,A}^\sharp$.

1. Choose a subset S_1 of $\{1, \dots, n\}$ and generate i.i.d. $N(0,1)$ variables $\{\xi_{i_1} : i_1 \in S_1\}$ independent of the data $\mathcal{D}_n$ and $\{\xi_\iota' : \iota \in I_{n,r}\}$. Let $n_1 = |S_1|$.
2. For each $i_1 \in S_1$, construct an estimate $\widehat{g}^{(i_1)}$ of g based on X_1^n.
3. Construct

$$U_{n,A}^\sharp = \frac{r}{\sqrt{n_1}} \sum_{i_1 \in S_1} \xi_{i_1} \{\widehat{g}^{(i_1)}(X_{i_1}) - \breve{g}\},$$

where $\breve{g} = n_1^{-1} \sum_{i_1 \in S_1} \widehat{g}^{(i_1)}(X_{i_1})$.

Step 1 chooses a subset S_1 to reduce the computational cost of the resulting bootstrap. Construction of estimates $\widehat{g}^{(i_1)}, i_1 \in S_1$ can be flexible. For instance, the estimates $\widehat{g}^{(i_1)}, i_1 \in S_1$ may depend on another randomization independent of everything else. In Sections 4.2 and 4.3, we will consider deterministic and random constructions of $\widehat{g}^{(i_1)}, i_1 \in S_1$, respectively.

It is worth noting that the jackknife multiplier bootstrap (JMB) developed in [8] (for the $r = 2$ case) and [9] (for the general r case) is a special case of $U_{n,A}^\sharp$ where $S_1 = \{1, \dots, n\}$ and $\widehat{g}^{(i_1)}(X_{i_1})$ is realized by its jackknife estimate, that is, by the U-statistic with kernel $(x_2, \dots, x_r) \mapsto h(X_{i_1}, x_2, \dots, x_r)$ for the sample without the i_1-th observation. Nevertheless, the bottleneck is that the computation of the jackknife estimates of $g(X_{i_1}), i_1 = 1, \dots, n$ requires $O(n^r d)$ operations, and hence implementing the JMB can be computationally demanding.

Now, consider $U_n^\sharp = U_{n,A}^\sharp + \alpha_n^{1/2} U_{n,B}^\sharp$. We call the simulation of $U_n^\sharp$ the *multiplier bootstrap under nondegeneracy* (MB-NDG). The following theorem establishes conditions under which the conditional distribution of $U_n^\sharp$ is able to consistently estimate the $N(0, r^2\Gamma_g + \alpha_n \Gamma_h)$ distribution on the hyperrectangles with polynomial error rates. Define

$$\widehat{\Delta}_{A,1} := \max_{1 \le j \le d} \frac{1}{n_1} \sum_{i_1 \in S_1} \{\widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1})\}^2,$$

which quantifies the errors of the estimates $\widehat{g}^{(i_1)}, i_1 \in S_1$. In addition, let $\overline{\sigma}_g := \max_{1 \le j \le d} \sqrt{P(g_j - \theta_j)^2}$.

THEOREM 4.2 (Generic bootstrap validity under nondegeneracy). *Let* $U_n^\sharp = U_{n,A}^\sharp + \alpha_n^{1/2} U_{n,B}^\sharp$. *Suppose that Conditions* (C1), (C2) *and* (C3-ND) *hold. In addi-*

tion, suppose that

$$\frac{D_n^2(\log^2 n)\log^5(dn)}{n_1 \wedge N} \le C_1 n^{-\zeta_1} \quad \text{and}$$
$$\mathbb{P}\big(\overline{\sigma}_g^2 \widehat{\Delta}_{A,1} \log^4 d > C_1 n^{-\zeta_2}\big) \le C_1 n^{-1} \tag{4.3}$$

for some constants $0 < C_1 < \infty$ and $\zeta_1, \zeta_2 \in (0,1)$. Then there exists a constant C depending only on $\underline{\sigma}, r$ and C_1 such that

$$\sup_{R \in \mathcal{R}} \big|\mathbb{P}_{|\mathcal{D}_n}\big(U_n^\sharp \in R\big) - \mathbb{P}(Y \in R)\big| \le C n^{-(\zeta_1 \wedge \zeta_2)/6} \tag{4.4}$$

with probability at least $1 - Cn^{-1}$, where $Y \sim N(0, r^2\Gamma_g + \alpha_n \Gamma_h)$. If the estimates $g^{(i_1)}, i_1 \in S_1$ depend on an additional randomization independent of $\mathcal{D}_n$, $\{\xi_{i_1} : i_1 \in S_1\}$, and $\{\xi'_\iota : \iota \in I_{n,r}\}$, then the result (4.4), with $\mathcal{D}_n$ replaced by the augmentation of $\mathcal{D}_n$ with variables used in the additional randomization, holds with probability at least $1 - Cn^{-1}$.

The second part of Condition (4.3) is a high-level condition on the estimation accuracy of $\widehat{g}^{(i_1)}, i_1 \in S_1$. In Sections 4.2 and 4.3, we will verify the second part of Condition (4.3) for deterministic and random constructions of $\widehat{g}^{(i_1)}, i_1 \in S_1$. The bootstrap distribution is taken with respect to the multiplier variables $\{\xi_{i_1} : i_1 \in S_1\}$ and $\{\xi'_\iota : \iota \in I_{n,r}\}$, and so if the estimation step for g depends on an additional randomization, then the variables used in the additional randomization have to be generated outside the bootstrap iterations.

When the approximating distribution can be simplified to $\gamma_A = N(0, r^2\Gamma_g)$, then it suffices to estimate $N(0, r^2\Gamma_g)$ by the conditional distribution of $U_{n,A}^\sharp$.

COROLLARY 4.3 (Validity of $U_{n,A}^\sharp$). *Suppose that all the conditions in Theorem 4.2 hold. Then there exists a constant C depending only on $\underline{\sigma}, r$ and C_1 such that*

$$\sup_{R \in \mathcal{R}} \big|\mathbb{P}_{|\mathcal{D}_n}\big(U_{n,A}^\sharp \in R\big) - \gamma_A(R)\big| \le C n^{-(\zeta_1 \wedge \zeta_2)/6} \tag{4.5}$$

with probability at least $1 - Cn^{-1}$. If the estimates $g^{(i_1)}, i_1 \in S_1$ depend on an additional randomization independent of $\mathcal{D}_n$, $\{\xi_{i_1} : i_1 \in S_1\}$, and $\{\xi'_\iota : \iota \in I_{n,r}\}$, then the result (4.5), with $\mathcal{D}_n$ replaced by the augmentation of $\mathcal{D}_n$ with variables used in the additional randomization, holds with probability at least $1 - Cn^{-1}$.

REMARK 4.2 (Comments on the partial bootstrap simplification under nondegeneracy). When the approximating distribution of $\sqrt{N}(U'_{n,N} - \theta)$ can be simplified to $\gamma_B = N(0, \Gamma_B)$, it is also possible to use the partial bootstrap $U_{n,B}^\sharp$ to estimate $N(0, \Gamma_B)$. In this case, we must take N to be sublinear in n (i.e., $N \ll$

$n/(D_n \log^2 d)$) to ensure the Gaussian approximation validity (cf. Remark 3.1). However, we do not recommend this simplification because the decay rate of the covariance matrix of the approximating Gaussian distribution $N(\theta, N^{-1}\Gamma_B)$ to $U'_{n,N}$ is N^{-1}, which is slower than the n^{-1} rate for the linear and superlinear cases. In particular, this implies a power loss in the testing problems if the critical values are calibrated by $U^{\sharp}_{n,B}$.

The rest of this section is devoted to concrete constructions of the estimates $\widehat{g}^{(i_1)}$, $i_1 \in S_1$.

4.2. *Divide-and-conquer estimation.* We first propose a deterministic construction of $\widehat{g}^{(i_1)}$, $i_1 \in S_1$ via the divide-and-conquer (DC) algorithm (cf. [41]).

1. For each $i_1 \in S_1$, choose K *disjoint* subsets $S^{(i_1)}_{2,k}$, $k = 1, \dots, K$ with common size $L \geq r-1$ from $\{1, \dots, n\} \setminus \{i_1\}$.
2. For each $i_1 \in S_1$, estimate g by computing U-statistics with kernel $(x_2, \dots, x_r) \mapsto h(x, x_2, \dots, x_r)$ applied to the subsamples $\{X_i : i \in S^{(i_1)}_{2,k}\}$, $k = 1, \dots, K$, and taking the average of those U-statistics of order $r-1$, that is,

$$\widehat{g}^{(i_1)}(x) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_{L,r-1}|} \sum_{\substack{i_2, \dots, i_r \in S^{(i_1)}_{2,k} \\ i_2 < \cdots < i_r}} h(x, X_{i_2}, \dots, X_{i_r}).$$

The DC algorithm can be viewed as an estimation procedure for g via incomplete U-statistics of order $r-1$ with a *block diagonal* sampling scheme (up to a permutation on the indices). We call the simulation of $U^{\sharp}_n$ with the DC algorithm the *MB-NDG-DC*. In Section 4.3, we will propose a different estimation procedure for g via randomized incomplete U-statistics of order $r-1$ based on an additional Bernoulli sampling. As a practical guidance to implement the DC algorithm, we suggest to choose $S_1 = \{1, \dots, n\}$, $L = r-1$ and $K = \lfloor (n-1)/L \rfloor$ consecutive blocks, which are the parameter values used in our simulation examples in Section 5. In this case, the DC algorithm turns out to be calculating Hoeffding's averages of the U-statistics of order $r-1$, which requires $O(nd)$ operations for each i_1. In contrast, the JMB constructs $\widehat{g}^{(i_1)}$ by complete U-statistics of order $r-1$, which requires $O(n^{r-1}d)$ operations for each i_1. Since the estimation step for g can be done outside the bootstrap iterations, the overall computational cost of the MB-NDG-DC is $O((BN + n_1 KL + Bn_1)d) = O(n^2 d + B(N+n)d)$ (where B denotes the number of bootstrap iterations), which is independent of the order of the U-statistic. In addition, if we choose to only simulate $U^{\sharp}_{n,A}$, then the computational cost is $O(n^2 d + Bnd)$, since the $O(BNd)$ computations come from simulating $U^{\sharp}_{n,B}$. We can certainly make the computational cost even smaller by taking n_1 and K smaller than n. For instance, if we choose n_1 and K in such a way

that $n_1 K = O(n)$ and $L = r - 1$, then the overall computational cost is reduced to $O(nd + B(N+n)d) = O(B(N+n)d)$ (or $O(Bnd)$ if we only simulate $U^{\sharp}_{n,A}$). In general, choosing smaller n_1 and K would sacrifice the statistical accuracy of the resulting bootstrap, but if $O(n^2 d)$ computations are difficult to implement, then choosing smaller n_1 and K would be a reasonable option.

Our MB-NDG-DC substantially differs from the the Bag of Little Bootstraps (BLB) proposed in [26], which is another generically scalable bootstrap method for large data sets based on the DC algorithm. Due to the space concern, we defer the comparison of our MB-NDG-DC with the BLB in Section A.1 of the SM.

The following proposition provides conditions for the validity of the multiplier bootstrap equipped with the DC estimation (MB-NDG-DC).

PROPOSITION 4.4 (Validity of bootstrap with DC estimation). *Suppose that Conditions* (C1), (C2) *and* (C3-ND) *hold. In addition, suppose that*

$$\frac{D_n^2(\log^2 n)\log^5(dn)}{n_1 \wedge N} \vee \left\{ \frac{\overline{\sigma}_g^2 D_n^2 \log^7 d}{KL}\left(1 + \frac{\log^2 d}{K^{1-1/\nu}}\right)\right\} \le C_1 n^{-\zeta} \tag{4.6}$$

for some constants $0 < C_1 < \infty$, $\zeta \in (0,1)$ *and* $\nu \in (1/\zeta, \infty)$. *Then there exists a constant* C *depending only on* $\underline{\sigma}$, r, ν *and* C_1 *such that each of the results* (4.4) *and* (4.5) *with* $(\zeta_1, \zeta_2) = (\zeta, \zeta - 1/\nu)$ *holds with probability at least* $1 - Cn^{-1}$.

For instance, consider to take $N = n$, $S_1 = \{1, \dots, n\}$, $L = r - 1$, and $K = \lfloor (n-1)/L \rfloor$, and suppose that $D_n^2 \log^7(dn) \le n^{1-\zeta}$ for some $\zeta \in (0,1)$. Then, by Theorem 3.1 and Proposition 4.4, for arbitrarily large $\nu \in (1/\zeta, \infty)$, there exists a constant C depending only on $\overline{\sigma}_g$, $\underline{\sigma}$, r and ν such that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}\left(\sqrt{n}(U'_{n,N} - \theta) \in R\right) - \mathbb{P}_{|\mathcal{D}_n}\left(U^{\sharp}_n \in R\right) \right| \le C n^{-(\zeta - 1/\nu)/6} \tag{4.7}$$

with probability at least $1 - Cn^{-1}$. Hence, the conditional distribution of the MB-NDG-DC approaches uniformly on the hyperrectangles in $\mathbb{R}^d$ to the distribution of the randomized incomplete U-statistic at a polynomial rate in the sample size.

4.3. *Random sampling estimation.* Next, we propose a random construction of $\widehat{g}^{(i_1)}$, $i_1 \in S_1$ based on an additional Bernoulli sampling. For each $i_1 = 1, \dots, n$, let $I_{n-1,r-1}(i_1) = \{(i_2, \dots, i_r) : 1 \le i_2 < \cdots < i_r \le n, i_j \ne i_1 \ \forall j \ne 1\}$. In addition, define $\sigma_{i_1} : \{1, \dots, n-1\} \to \{1, \dots, n\} \setminus \{i_1\}$ as follows: if $\{1, \dots, n\} \setminus \{i_1\} = \{j_1, \dots, j_{n-1}\}$ with $j_1 < \cdots < j_{n-1}$, then $\sigma_{i_1}(\ell) = j_\ell$ for $\ell = 1, \dots, n-1$. For notational convenience, for $\iota' = (i_2, \dots, i_r) \in I_{n-1,r-1}$, we write $\sigma_{i_1}(\iota') = (\sigma_{i_1}(i_2), \dots, \sigma_{i_1}(i_r)) \in I_{n-1,r-1}(i_1)$.

Now, consider the following randomized procedure to construct $\widehat{g}^{(i_1)}$, $i_1 \in S_1$:

1. Let $0 < M = M_n \le |I_{n-1,r-1}|$ be a positive integer, and generate i.i.d. $\mathsf{Ber}(\vartheta_n)$ random variables $\{Z'_{\iota'} : \iota' = (i_2, \dots, i_r) \in I_{n-1,r-1}\}$ independent of $\mathcal{D}_n$, $\{\xi_{i_1} : i_1 \in S_1\}$, and $\{\xi'_\iota : \iota \in I_{n,r}\}$, where $\vartheta_n = M/|I_{n-1,r-1}|$.

2. For each $i_1 \in S_1$, construct $\widehat{g}^{(i_1)}(x) = M^{-1} \sum_{\iota' \in I_{n-1,r-1}} Z'_{\iota'} h(x, X_{\sigma_{i_1}(\iota')})$.

The resulting bootstrap method is called the *multiplier bootstrap under nondegeneracy with random sampling* (MB-NDG-RS). Equivalently, the above procedure can be implemented as follows:

1. Generate $\widehat{M} \sim \mathsf{Bin}(|I_{n-1,r-1}|, \vartheta_n)$.
2. Sample $\iota'_1, \ldots, \iota'_{\widehat{M}}$ randomly without replacement from $I_{n-1,r-1}$.
3. Construct $\widehat{g}^{(i_1)}(x) = M^{-1} \sum_{j=1}^{\widehat{M}} h(x, X_{\sigma_{i_1}(\iota'_j)})$ for each $i_1 \in S_1$.

So, on average, the computational cost to construct $\widehat{g}^{(i_1)}, i_1 \in S_1$ by the random sampling estimation is $O(n_1 M d)$, and the overall computational cost of the MB-NDG-RS is $O(n_1 M d + B(N + n_1)d)$ (or $O(n_1 M d + B n_1 d)$ if we only simulate $U^{\sharp}_{n,A}$). As a practical guidance to implement the random sampling estimation, we suggest to choose $S_1 = \{1, \ldots, n\}$ and M proportional to $n - 1$, which are the parameter values used in our simulation examples in Section 5. Then the overall computational cost of the MB-NDG-RS is $O(n^2 d + B(N + n)d)$ (or $O(n^2 d + Bnd)$ if we only simulate $U^{\sharp}_{n,A}$), which is independent of the order of the U-statistic. In addition, the computational cost can be made even smaller, for example, can be reduced to $O(B(N + n)d)$ by choosing n_1 and M in such a way that $n_1 M = O(n)$ (or $O(Bnd)$ if we only simulate $U^{\sharp}_{n,A}$), which would be a reasonable option if $O(n^2 d)$ computations are difficult to implement.

PROPOSITION 4.5 (Validity of bootstrap with Bernoulli sampling estimation). *Suppose that Conditions* (C1), (C2) *and* (C3-ND) *hold. In addition, suppose that*

$$\frac{D_n^2(\log^2 n)\log^5(dn)}{n_1 \wedge N} \vee \frac{\overline{\sigma}_g^2 D_n^2 \log^7(dn)}{n \wedge M} \le C_1 n^{-\zeta} \tag{4.8}$$

for some constants $0 < C_1 < \infty$ *and* $\zeta \in (0, 1)$. *Then, for arbitrarily large* $\nu \in (1/\zeta, \infty)$, *there exists a constant* C *depending only on* $\underline{\sigma}, r, \nu$ *and* C_1 *such that each of the results* (4.4) *and* (4.5), *with* $\mathcal{D}_n$ *replaced by* $\mathcal{D}'_n = \mathcal{D}_n \cup \{Z'_{\iota'} : \iota' \in I_{n-1,r-1}\}$ *and with* $(\zeta_1, \zeta_2) = (\zeta, \zeta - 1/\nu)$, *holds with probability at least* $1 - Cn^{-1}$.

For instance, consider to take $N = n$, $S_1 = \{1, \ldots, n\}$, and M proportional to $n - 1$, and suppose that $D_n^2 \log^7(dn) \le n^{1-\zeta}$ for some $\zeta \in (0, 1)$. Then, by Theorem 3.1 and Proposition 4.5, for arbitrarily large $\nu \in (1/\zeta, \infty)$, there exists a constant C depending only on $\overline{\sigma}_g, \underline{\sigma}, r$, and ν such that the result (4.7) holds with probability at least $1 - Cn^{-1}$.

REMARK 4.3 (Alternative options for random sampling estimation). In construction of $\widehat{g}^{(i_1)}$, instead of normalization by M, we may use normalization by

$\widehat{M}$, namely, $\widehat{M}^{-1}\sum_{j=1}^{\widehat{M}} h(x, X_{\sigma_{i_1}(\iota'_j)})$ for $\widehat{g}^{(i_1)}(x)$. In view of the concentration inequality for $\widehat{M}$ (cf. equation (2.2)), it is not difficult to see that the same conclusion of Proposition 4.5 holds for $\widehat{g}^{(i_1)}(x) = \widehat{M}^{-1}\sum_{j=1}^{\widehat{M}} h(x, X_{\sigma_{i_1}(\iota'_j)})$.

Next, alternatively to the Bernoulli sampling, we may use sampling with replacement to construct $\widehat{g}^{(i_1)}$, which can be implemented as follows: (1) sample $\iota'_1, \ldots, \iota'_M$ randomly with replacement from $I_{n-1,r-1}$ (independently of everything else); and (2) construct $\widehat{g}^{(i_1)}(x) = M^{-1}\sum_{j=1}^{M} h(x, X_{\sigma_{i_1}(\iota'_j)})$ for $i_1 \in S_1$. For each $i_1 \in S_1$, conditionally on X_1^n, $X_{\sigma_{i_1}(\iota'_j)}$, $j = 1, \ldots, M$ are i.i.d. draws from the empirical distribution $|I_{n-1,r-1}|^{-1}\sum_{\iota' \in I_{n-1,r-1}(i_1)} \delta_{X_{\iota'}}$. Mimicking the proof of Proposition 4.5, it is not difficult to see that the conclusion of the proposition holds for the estimation of g via sampling with replacement under the condition (4.8) (here, $Z'_{\iota'}$ is the number of times that ι' is redrawn in the sample $\{\iota'_1, \ldots, \iota'_M\}$, for which $\widehat{g}^{(i_1)}(x)$ can be expressed as $\widehat{g}^{(i_1)}(x) = M^{-1}\sum_{\iota' \in I_{n-1,r-1}} Z'_{\iota'} h(x, X_{\sigma_{i_1}(\iota')})$).

5. Numerical examples. In this section, we provide some numerical examples to verify the validity of our Gaussian approximation results and the proposed bootstrap algorithms (i.e., MB-DG, MB-NDG-DC, MB-NDG-RS) for approximating the distributions of incomplete U-statistics. In particular, we examine the statistical accuracy and computational running time of the Gaussian approximation and bootstrap algorithms in the leading example of testing for the pairwise independence of a high-dimensional vector.

5.1. *Test statistics.* In this section, we discuss several nonparametric statistics in the literature for the testing problem of the pairwise independence.

EXAMPLE 5.1 (Spearman's ρ). Let Π_r be the collection of all possible permutations on $\{1, \ldots, r\}$. Hoeffding [20] shows that Spearman's rank correlation coefficient matrix ρ can be written as

$$\rho = \frac{n-2}{n+1}\widehat{\rho} + \frac{3}{n+1}\tau,$$

where $\widehat{\rho} = U_n^{(3)}(h^S)$ is the $p \times p$ matrix-valued U-statistic associated with the kernel

$$\begin{aligned} h^S(X_1, X_2, X_3) &= \big(h^S_{j,k}(X_1, X_2, X_3)\big)_{1 \le j,k \le p} \\ &= \frac{1}{2}\sum_{\pi \in \Pi_3} \operatorname{sign}\big\{(X_{\pi(1)} - X_{\pi(2)})(X_{\pi(1)} - X_{\pi(3)})^T\big\}, \end{aligned}$$

and $\tau = (\tau_{j,k})_{1 \le j,k \le p} = U_n^{(2)}(h^K)$ is the $p \times p$ Kendall τ matrix with the kernel

$$h^K(X_1, X_2) = \operatorname{sign}\big\{(X_1 - X_2)(X_1 - X_2)^T\big\}.$$

Here, for a matrix $A = (a_{j,k})_{1\le j,k\le p}$, $\operatorname{sign}\{A\}$ is the matrix of the same size as A whose (j,k)-th element is $\operatorname{sign}(a_{j,k}) = \mathbf{1}(a_{j,k} > 0) - \mathbf{1}(a_{j,k} < 0)$. It is seen that the leading term in Spearman's ρ is $\widehat{\rho}$, and so it is reasonable to reject the null hypothesis (1.2) if $\max_{1\le j<k\le p} |\widehat{\rho}_{j,k}|$ is large. Precisely speaking, this test is testing for a weaker hypothesis that

$$H_0' : \mathbb{E}\big[\operatorname{sign}\big(X_1^{(j)} - X_2^{(j)}\big)\operatorname{sign}\big(X_1^{(k)} - X_3^{(k)}\big)\big] = 0 \qquad \text{for all } 1 \le j < k \le p.$$

EXAMPLE 5.2 (Bergsma and Dassios' t^*). [2] propose a U-statistic $t^* = (t^*_{j,k})_{1\le j,k\le p} = U_n^{(4)}(h^{\mathrm{BD}})$ of order 4 with the kernel

$$h^{\mathrm{BD}}(X_1,\dots,X_4) = \frac{1}{24}\sum_{\pi\in\Pi_4} \phi(X_{\pi(1)},\dots,X_{\pi(4)})\phi(X_{\pi(1)},\dots,X_{\pi(4)})^T,$$

where $\phi(X_1,\dots,X_4) = (\phi_j(X_1,\dots,X_4))_{j=1}^p$ and

$$\begin{aligned}&\phi_j(X_1,\dots,X_4)\\ &\quad= \mathbf{1}\big(X_1^{(j)} \vee X_3^{(j)} < X_2^{(j)} \wedge X_4^{(j)}\big) + \mathbf{1}\big(X_1^{(j)} \wedge X_3^{(j)} > X_2^{(j)} \vee X_4^{(j)}\big)\\ &\qquad- \mathbf{1}\big(X_1^{(j)} \vee X_2^{(j)} < X_3^{(j)} \wedge X_4^{(j)}\big) - \mathbf{1}\big(X_1^{(j)} \wedge X_2^{(j)} > X_3^{(j)} \vee X_4^{(j)}\big).\end{aligned}$$

Under the assumption that $(X^{(j)}, X^{(k)})$ has a bivariate distribution that is discrete or (absolutely) continuous, or a mixture of both, [2] show that $\mathbb{E}[t^*_{j,k}] = 0$ if and only if $X^{(j)}$ and $X^{(k)}$ are independent, and so it is reasonable to reject the null hypothesis (1.2) if $\max_{1\le j<k\le p} |t^*_{j,k}|$ is large (or $\max_{1\le j<k\le p} t^*_{j,k}$ is large, since in general $\mathbb{E}[t^*_{j,k}] \ge 0$).

EXAMPLE 5.3 (Hoeffding's D). Hoeffding [20] proposes a U-statistic $D = (D_{j,k})_{1\le j,k\le p} = U_n^{(5)}(h^D)$ of order 5 with the kernel

$$h^D(X_1,\dots,X_5) = \frac{1}{120}\sum_{\pi\in\Pi_5} \phi(X_{\pi(1)},\dots,X_{\pi(5)})\phi(X_{\pi(1)},\dots,X_{\pi(5)})^T,$$

where $\phi(X_1,\dots,X_5) = (\phi_j(X_1,\dots,X_5))_{j=1}^p$ and $\phi_j(X_1,\dots,X_5) = [\mathbf{1}(X_1^{(j)} \ge X_2^{(j)}) - \mathbf{1}(X_1^{(j)} \ge X_3^{(j)})][\mathbf{1}(X_1^{(j)} \ge X_4^{(j)}) - \mathbf{1}(X_1^{(j)} \ge X_5^{(j)})]/4$. Under the assumption that the joint distribution of $(X^{(j)}, X^{(k)})$ has continuous joint and marginal densities, [21] shows that $\mathbb{E}[D_{j,k}] = 0$ if and only if $X^{(j)}$ and $X^{(k)}$ are independent, and so it is reasonable to reject the null hypothesis (1.2) if $\max_{1\le j<k\le p} |D_{j,k}|$ is large (or $\max_{1\le j<k\le p} D_{j,k}$ is large, since in general $\mathbb{E}[D_{j,k}] \ge 0$). It is worth noting that Bergsma and Dassios' t^* is an improvement on Hoeffding's D since the former can characterize the pairwise independence under weaker assumptions on the distribution of X than the latter.

Here, h^S is nondegenerate, while h^{BD} and h^D are degenerate of order 1 under H_0. The above testing problem is motivated from recent papers by [28] and [18], which study testing for the null hypothesis

$$H_0'' : X^{(1)}, \ldots, X^{(p)} \qquad \text{are mutually independent,}$$

and develop tests based on functions of the U-statistics appearing in Examples 5.1–5.3. Note that H_0'' is a stronger hypothesis than H_0. Specifically, [28] consider tests statistics such as, for example, $S_{\widehat{\rho}} = \sum_{1 \le j < k \le p} \widehat{\rho}_{j,k}^2 - 3\mu_{\widehat{\rho}}$ with $\mu_{\widehat{\rho}} = \mathbb{E}[\widehat{\rho}_{1,2}^2]$ under H_0'' and show that $nS_{\widehat{\rho}}/(9p\zeta_1^{\widehat{\rho}}) \xrightarrow{d} N(0,1)$ under H_0'' as $(n,p) \to \infty$ where $\zeta_1^{\widehat{\rho}} = \mathrm{Var}(\mathbb{E}[h_{1,2}^S(X_1, X_2, X_3) \mid X_1])$. On the other hand, [18] consider test statistics such as, for example, $L_n = \max_{1 \le j < k \le p} |\widehat{\rho}_{j,k}|$ and show that $L_n^2 / \mathrm{Var}(\widehat{\rho}_{1,2}) - 4\log p + \log\log p$ converges in distribution to a Gumbel distribution as $n \to \infty$ and $p = p_n \to \infty$ under H_0'' provided that $\log p = o(n^{1/3})$ (precisely speaking, [18] rule out degenerate kernels). Importantly, compared with the tests developed in [28] and [18] based on analytical critical values, our bootstrap-based tests can directly detect the pairwise dependence for some pair of coordinates (or $\mathbb{E}[\mathrm{sign}(X_1^{(j)} - X_2^{(j)})\,\mathrm{sign}(X_1^{(k)} - X_3^{(k)})] \neq 0$ for some $1 \le j < k \le p$ for Spearman's ρ) rather than the nonmutual independence and also work for noncontinuous random vectors (see, e.g., [16] for interesting examples of pairwise independent but jointly dependent random variables; in particular, their examples include continuous random variables). In contrast, the derivations of the asymptotic null distributions in [28] and [18] critically depend on the mutual independence between the coordinates of X. In addition, they both assume that X is continuously distributed so that there are no ties in $X_1^{(j)}, \ldots, X_n^{(j)}$ for each coordinate j, thereby ruling out discrete components. It is worth noting that the U-statistics appearing Examples 5.1–5.3 are rank-based, and so if X is continuous and H_0'' is true, then those U-statistics are pivotal, that is, they have known (but difficult-to-compute) distributions, which is also a critical factor in their analysis; however, that is not the case under the weaker hypothesis of pairwise independence and without the continuity assumption on X [31].

In our simulation studies, we consider two test statistics: Spearman's ρ and Bergsma–Dassios' t^*. Under H_0 in (1.2), the leading term $\widehat{\rho}$ of Spearman's ρ is nondegenerate while Bergsma–Dassios' t^* is degenerate of order 1, both having zero mean. Slightly abusing notation, we will use $\widehat{\rho}$ as Spearman's ρ statistic throughout this section. We consider tests of the forms

$$\max_{1 \le j < k \le p} |\widehat{\rho}'_{j,k}| > c \quad \Rightarrow \quad \text{reject } H_0 \quad \text{and} \quad \max_{1 \le j < k \le p} |t^{*'}_{j,k}| > c \quad \Rightarrow \quad \text{reject } H_0,$$

where $\widehat{\rho}'_{j,k}$ and $t^{*'}_{j,k}$ are incomplete versions of $\widehat{\rho}_{j,k}$ and $t^*_{j,k}$, respectively, and their critical values are calibrated by the bootstrap methods. In particular, for any nominal size $\alpha \in (0,1)$, the value of $c := c(\alpha)$ can be chosen as the $(1-\alpha)$-th quantile

of an appropriate bootstrap conditional distribution given $\mathcal{D}_n$. For Spearman's ρ, we use $U_n^\sharp$ for MB-NDG-DC and MB-NDG-RS. For Bergsma–Dassios' t^*, we use $U_{n,B}^\sharp$ for MB-DG. In addition, we also test the performance of the *partial* versions of MB-NDG-DC and MB-NDG-RS (i.e., $U_{n,A}^\sharp$; cf. Corollary 4.3) for Spearman's ρ statistic when its distribution can be approximated by $\gamma_A = N(0, r^2\Gamma_g)$ (cf. Corollary 3.2).

5.2. *Simulation setup*. We simulate i.i.d. data from the noncentral t-distribution with 3 degrees of freedom and noncentrality parameter 2. This data generating process implies H_0. We consider $n = 300, 500, 1000$ and $p = 30, 50, 100$ (so the number of the free parameters is $d = p(p-1)/2 = 435, 1225, 4950$). For each setup (n, p), we fix the bootstrap sample size $B = 200$ and report the empirical rejection probabilities of the bootstrap tests averaged over 2000 simulations. For Spearman's ρ, we apply the MB-NDG-DC and MB-NDG-RS (full version $U_n^\sharp$) and set the computational budget parameter value $N = 2n$. In addition, we implement the MB-NDG-DC with the parameter values suggested in Section 4.2 (i.e., $S_1 = \{1, \dots, n\}$, $L = r - 1$, and $K = \lfloor (n-1)/L \rfloor$), and the MB-NDG-RS with the parameter values suggested in Section 4.3 (i.e., $S_1 = \{1, \dots, n\}$ and $M = 2(n-1)$). For Bergsma–Dassios' t^*, we apply the MB-DG $U_{n,B}^\sharp$ with $N = n^{4/3}$. Moreover, we also apply the partial versions of MB-NDG-DC and MB-NDG-RS $U_{n,A}^\sharp$ with $N = 4n^{3/2}$. These computational budget parameter values are chosen to minimize the rate in the error bounds of the corresponding Gaussian and bootstrap approximations. We only report the simulation results for the randomized incomplete U-statistic with the Bernoulli sampling since the simulation results for the sampling with replacement case are qualitatively similar.

5.3. *Simulation results*. We first examine the statistical accuracy of the bootstrap tests in terms of size for $U_n^\sharp$ for the incomplete versions of Spearman's ρ and $U_{n,B}^\sharp$ for Bergsma–Dassios' t^*. For each nominal size $\alpha \in (0, 1)$, we denote by $\widehat{R}(\alpha)$ the empirical rejection probability of the null hypothesis, where the critical values are calibrated by our bootstrap methods. The uniform errors-in-size on $\alpha \in [0.01, 0.10]$ of our bootstrap tests are summarized in Table 1. We observe that the bootstrap approximations become more accurate as n increases, and they work quite well for small values of α, which are relevant in the testing application. Due to the space concern, we defer the empirical size graph $\{(\alpha, \widehat{R}(\alpha)) : \alpha \in (0, 1)\}$ of the bootstrap tests for MB-NDG-DC (Spearman's ρ), MB-NDG-RS (Spearman's ρ) and MB-DG (Bergsma–Dassios' t^*) to Appendix D in the SM. In addition, we also report the simulation results of the partial bootstrap $U_{n,A}^\sharp$ for Spearman's ρ in Appendix D in the SM.

We also report the empirical performance of the Gaussian approximation for the test statistics. The P-P plots for Spearman's ρ (i.e., $\sqrt{n}U'_{n,N}$ versus $N(0, r^2\Gamma_g +$

TABLE 1
Uniform error-in-size $\sup_{\alpha\in[0.01.0.10]} |\widehat{R}(\alpha) - \alpha|$ *of the bootstrap tests, where* α *is the nominal size*

Setup	Spearman's ρ (MB-NDG-DC)	Spearman's ρ (MB-NDG-RS)	Bergsma-Dassios' t^* (MB-DG)
$p = 30, n = 300$	0.0080	0.0110	0.0280
$p = 30, n = 500$	0.0065	0.0130	0.0225
$p = 30, n = 1000$	0.0060	0.0055	0.0095
$p = 50, n = 300$	0.0250	0.0135	0.0385
$p = 50, n = 500$	0.0105	0.0035	0.0260
$p = 50, n = 1000$	0.0145	0.0095	0.0235
$p = 100, n = 300$	0.0180	0.0125	0.0660
$p = 100, n = 500$	0.0135	0.0100	0.0290
$p = 100, n = 1000$	0.0075	0.0020	0.0170

$\alpha_n\Gamma_h)$) and Bergsma–Dassios' t^* (i.e., $\sqrt{N}U'_{n,N}$ versus $N(0, \Gamma_h)$) are shown in Figures 1 and 2, respectively. Similarly as the bootstrap approximations, Gaussian approximations become more accurate as n increases.

Next, we report the computer running time of the bootstrap tests. Figure 3 displays the computer running time versus the sample size, both on the log-scale. It is observed that the (log-)running time for the bootstrap methods scales linearly with the (log-)sample size. We further fit a linear model of the (log-)running time against the (log-)sample size (with the intercept term) for each p. For Spearman's ρ, the slope coefficient for $p = (30, 50, 100)$ is $(1.820, 1.863, 1.819)$ in the case MB-NDG-DC, and $(1.987, 1.874, 1.918)$ in the case MB-NDG-RS. In both cases, the slope coefficients are close to the theoretic value 2. Recall that the computational complexity for MB-NDG-DC and MB-NDG-RS is the same as $O((n + B)nd)$ for the suggested parameter values. For n larger than B, the computational cost is approximately quadratic in n for each p. For Bergsma–Dassios' t^*, the slope coefficient for $p = (30, 50, 100)$ is $(1.314, 1.318, 1.316)$, which matches very well to the exponent $4/3$ of the computational budget parameter value $N = n^{4/3}$. In addition, the running time lines are in parallel with each other. This also makes sense because the computational costs of all the bootstrap methods are linear in d (and thus quadratic in p) and the increase of p only affects the intercept on the log-scale.

6. Discussions. In this paper, we have derived the Gaussian and bootstrap approximation results for incomplete U-statistics with random and sparse weights in high dimensions. Specifically, we have considered two sampling schemes: Bernoulli sampling and sampling with replacement, both subject to a computational budget parameter to construct the random weights. On one hand, the sparsity in the design makes the computation of the incomplete U-statistics tractable. On the other hand, the randomness of the weights opens the possibility for us to obtain

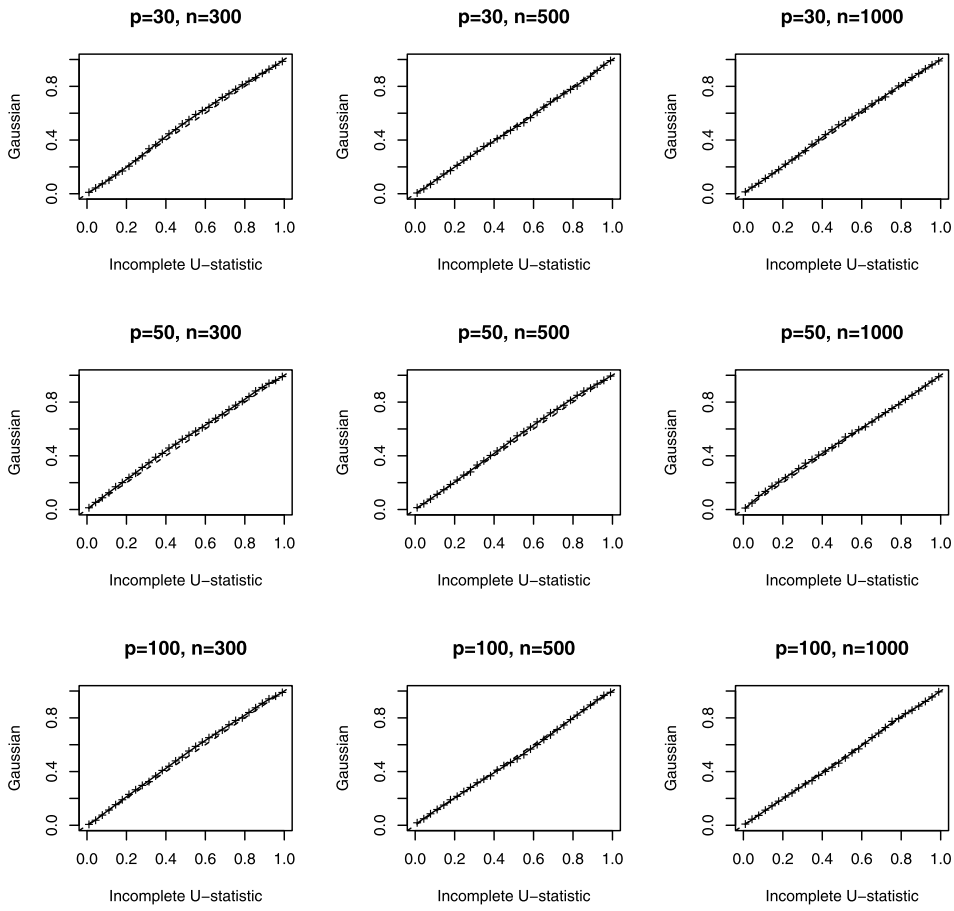

FIG. 1. *P-P plots for the Gaussian approximation* $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$ *of* $\sqrt{n}U'_{n,N}$ *for Spearman's* ρ *test statistic with the Bernoulli sampling.*

unified central limit theorem (CLT) type behaviors for both nondegenerate and degenerate kernels, thus revealing the fundamental difference between complete and randomized incomplete U-statistics. Building upon the Gaussian approximation results, we have developed novel bootstrap methods for incomplete U-statistics that take computational considerations into account, and established finite sample error bounds for the proposed bootstrap methods. Additional discussions on two extensions (extensions to normalized U-statistics and incomplete U-statistics with increasing orders) can be found in Section A of the SM.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

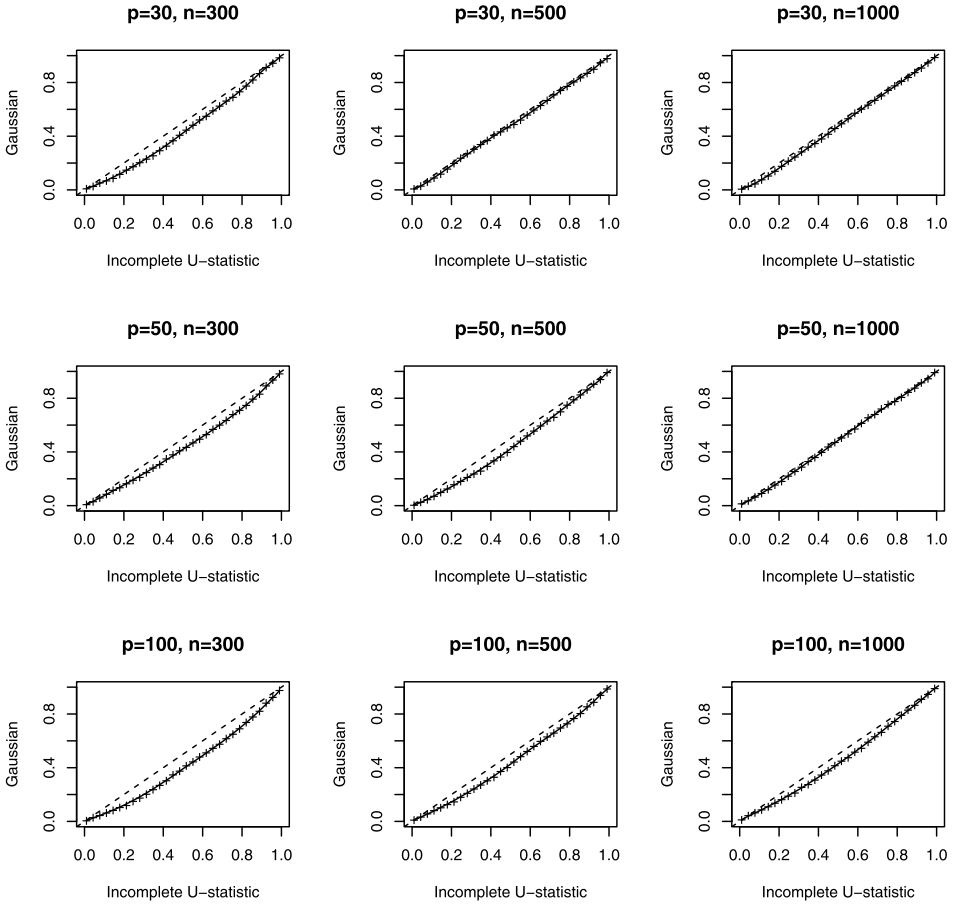

FIG. 2. *P-P plots for the Gaussian approximation* $N(0, \Gamma_h)$ *of* $\sqrt{N}U'_{n,N}$ *for Bergsma–Dassios' t* test statistic with the Bernoulli sampling.*

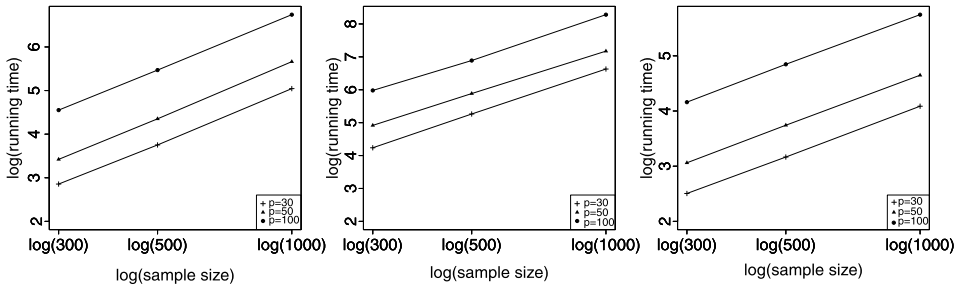

FIG. 3. *Computer running time of the bootstrap versus the sample size on the log-scale. Left*: *bootstrap* $U_n^{\natural\natural}$ *for Spearman's* ρ *with the divide-and-conquer estimation* (*MB-NDG-DC*). *Middle*: *bootstrap* $U_n^{\natural\natural}$ *for Spearman's* ρ *with the random sampling estimation* (*MB-NDG-RS*). *Right*: *bootstrap* $U_{n,B}^{\natural\natural}$ *for Bergsma–Dassios' t** (*MB-DG*).

SUPPLEMENTARY MATERIAL

Supplement to "Randomized incomplete U-statistics in high dimensions". (DOI: 10.1214/18-AOS1773SUPP; .pdf). The Supplementary Material contains the proofs and additional discussions, simulation results and applications of the main paper.

REFERENCES

[1] ARCONES, M. A. and GINÉ, E. (1992). On the bootstrap of U and V statistics. *Ann. Statist.* **20** 655–674. MR1165586

[2] BERGSMA, W. and DASSIOS, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli* **20** 1006–1028. MR3178526

[3] BERTAIL, P. and TRESSOU, J. (2006). Incomplete generalized U-statistics for food risk assessment. *Biometrics* **62** 66–74, 315. MR2226558

[4] BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103

[5] BLOM, G. (1976). Some properties of incomplete U-statistics. *Biometrika* **63** 573–580. MR0474582

[6] BRETAGNOLLE, J. (1983). Lois limites du bootstrap de certaines fonctionnelles. *Ann. Inst. H. Poincaré Sect. B (N.S.)* **19** 281–296. MR0725561

[7] BROWN, B. M. and KILDEA, D. G. (1978). Reduced U-statistics and the Hodges–Lehmann estimator. *Ann. Statist.* **6** 828–835. MR0491556

[8] CHEN, X. (2018). Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Ann. Statist.* **46** 642–678. MR3782380

[9] CHEN, X. and KATO, K. (2017). Jackknife multiplier bootstrap: Finite sample approximations to the U-process supremum with applications. Available at arXiv:1708.02705.

[10] CHEN, X. and KATO, K. (2019). Supplement to "Randomized incomplete U-statistics in high dimensions." DOI:10.1214/18-AOS1773SUPP.

[11] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

[12] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. MR3693963

[13] CLÉMENÇON, S., COLIN, I. and BELLET, A. (2016). Scaling-up empirical risk minimization: Optimization of incomplete $\mathcal{U}$-statistics. *J. Mach. Learn. Res.* **17** Paper No. 76, 36. MR3517099

[14] DEHLING, H. and MIKOSCH, T. (1994). Random quadratic forms and the bootstrap for U-statistics. *J. Multivariate Anal.* **51** 392–413. MR1321305

[15] EMBRECHTS, P., LINDSKOG, F. and MCNEIL, A. (2003). Modelling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance* (S. T. Rachev, ed.) 8. North-Holland, Amsterdam.

[16] GEISSER, S. and MANTEL, N. (1962). Pairwise independence of jointly dependent variables. *Ann. Math. Statist.* **33** 290–291. MR0137188

[17] GU, Q., CAO, Y., NING, Y. and LIU, H. (2015). Local and global inference for high dimensional nonparanormal graphical models. Available at arXiv:1502.02347.

[18] HAN, F., CHEN, S. and LIU, H. (2017). Distribution-free tests of independence in high dimensions. *Biometrika* **104** 813–828. MR3737306

[19] HAN, F. and QIAN, T. (2016). Asymptotics for asymmetric weighted U-statistics: Central limit theorem and bootstrap under data heterogeneity. Preprint.

[20] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics* **19** 293–325. MR0026294

[21] HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Statistics* **19** 546–557. MR0029139

[22] HSING, T. and WU, W. B. (2004). On weighted U-statistics for stationary processes. *Ann. Probab.* **32** 1600–1631. MR2060311

[23] HUŠKOVÁ, M. and JANSSEN, P. (1993). Consistency of the generalized bootstrap for degenerate U-statistics. *Ann. Statist.* **21** 1811–1823. MR1245770

[24] HUŠKOVÁ, M. and JANSSEN, P. (1993). Generalized bootstrap for studentized U-statistics: A rank statistic approach. *Statist. Probab. Lett.* **16** 225–233. MR1208512

[25] JANSON, S. (1984). The asymptotic distributions of incomplete U-statistics. *Z. Wahrsch. Verw. Gebiete* **66** 495–505. MR0753810

[26] KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. MR3248677

[27] LEE, A. J. (1990). *U-Statistics. Theory and Practice. Statistics: Textbooks and Monographs* **110**. Dekker, New York. MR1075417

[28] LEUNG, D. and DRTON, M. (2018). Testing independence in high dimensions with sums of rank correlations. *Ann. Statist.* **46** 280–307. MR3766953

[29] MAJOR, P. (1994). Asymptotic distributions for weighted U-statistics. *Ann. Probab.* **22** 1514–1535. MR1303652

[30] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** Paper No. 26, 41. MR3491120

[31] NANDY, P., WEIHS, L. and DRTON, M. (2016). Large-sample theory for the Bergsma–Dassios sign covariance. *Electron. J. Stat.* **10** 2287–2311. MR3541972

[32] O'NEIL, K. A. and REDNER, R. A. (1993). Asymptotic distributions of weighted U-statistics of degree 2. *Ann. Probab.* **21** 1159–1169. MR1217584

[33] RIFI, M. and UTZET, F. (2000). On the asymptotic behavior of weighted U-statistics. *J. Theoret. Probab.* **13** 141–167. MR1744988

[34] RUBIN, H. and VITALE, R. A. (1980). Asymptotic distribution of symmetric statistics. *Ann. Statist.* **8** 165–170. MR0557561

[35] SHAPIRO, C. P. and HUBERT, L. (1979). Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *Ann. Statist.* **7** 788–794. MR0532242

[36] SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

[37] VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[38] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics. Springer Series in Statistics*. Springer, New York. MR1385671

[39] WANG, Q. and JING, B.-Y. (2004). Weighted bootstrap for U-statistics. *J. Multivariate Anal.* **91** 177–198. MR2087842

[40] YAO, S., ZHANG, X. and SHAO, X. (2018). Testing mutual independence in high dimension via distance covariance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 455–480. MR3798874

[41] ZHANG, Y., DUCHI, J. and WAINWRIGHT, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **16** 3299–3340. MR3450540

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
725 S. WRIGHT STREET
CHAMPAIGN, ILLINOIS 61874
USA
E-MAIL: xhchen@illinois.edu

DEPARTMENT OF STATISTICAL SCIENCE
CORNELL UNIVERSITY
1194 COMSTOCK HALL
ITHACA, NEW YORK 14853
USA
E-MAIL: kk976@cornell.edu