The Annals of Statistics
2018, Vol. 46, No. 2, 642–678
https://doi.org/10.1214/17-AOS1563

GAUSSIAN AND BOOTSTRAP APPROXIMATIONS FOR HIGH-DIMENSIONAL U-STATISTICS AND THEIR APPLICATIONS[1]

BY XIAOHUI CHEN

University of Illinois at Urbana–Champaign

This paper studies the Gaussian and bootstrap approximations for the probabilities of a nondegenerate U-statistic belonging to the hyperrectangles in $\mathbb{R}^d$ when the dimension d is large. A two-step Gaussian approximation procedure that does not impose structural assumptions on the data distribution is proposed. Subject to mild moment conditions on the kernel, we establish the explicit rate of convergence uniformly in the class of all hyperrectangles in $\mathbb{R}^d$ that decays polynomially in sample size for a high-dimensional scaling limit, where the dimension can be much larger than the sample size. We also provide computable approximation methods for the quantiles of the maxima of centered U-statistics. Specifically, we provide a unified perspective for the empirical bootstrap, the randomly reweighted bootstrap and the Gaussian multiplier bootstrap with the jackknife estimator of covariance matrix as randomly reweighted quadratic forms and we establish their validity. We show that all three methods are inferentially first-order equivalent for high-dimensional U-statistics in the sense that they achieve the same uniform rate of convergence over all d-dimensional hyperrectangles. In particular, they are asymptotically valid when the dimension d can be as large as $O(e^{n^c})$ for some constant $c \in (0, 1/7)$.

The bootstrap methods are applied to statistical applications for high-dimensional non-Gaussian data including: (i) principled and data-dependent tuning parameter selection for regularized estimation of the covariance matrix and its related functionals; (ii) simultaneous inference for the covariance and rank correlation matrices. In particular, for the thresholded covariance matrix estimator with the bootstrap selected tuning parameter, we show that for a class of sub-Gaussian data, error bounds of the bootstrapped thresholded covariance matrix estimator can be much tighter than those of the minimax estimator with a universal threshold. In addition, we also show that the Gaussian-like convergence rates can be achieved for heavy-tailed data, which are less conservative than those obtained by the Bonferroni technique that ignores the dependency in the underlying data distribution.

1. Introduction. Let $X_1^n = \{X_1, \dots, X_n\}$ be a sample of independent and identically distributed (i.i.d.) random vectors in $\mathbb{R}^p$ with the distribution F. Let $h : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^d$ be a fixed and measurable function such that $h(x_1, x_2) = h(x_2, x_1)$

Received March 2016; revised February 2017.
[1]Supported in part by NSF Grant DMS-14-04891 and UIUC Research Board Award RB15004.
MSC2010 subject classifications. Primary 62E17; secondary 62F40.
Key words and phrases. U-statistics, high-dimensional inference, Gaussian approximation, bootstrap.

for all $x_1, x_2 \in \mathbb{R}^p$ and $\mathbb{E}|h_k(X_1, X_2)| < \infty$ for all $k = 1, \dots, d$. Consider the U-statistic of order two:

$$U_n = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j). \tag{1}$$

In this paper we consider the uniform approximation of the probabilities of U_n over a class of the Borel subsets in $\mathbb{R}^d$. More specifically, let $T_n = \sqrt{n}(U_n - \theta)/2$, where $\theta = \mathbb{E}[h(X_1, X_2)]$ is the parameter of interest, and $\mathcal{A}^{\mathrm{re}}$ be the class of all hyperrectangles A in $\mathbb{R}^d$ of the form

$$A = \{x \in \mathbb{R}^d : a_j \le x_j \le b_j \text{ for all } j = 1, \dots, d\}, \tag{2}$$

where $-\infty \le a_j \le b_j \le \infty$ for $j = 1, \dots, d$. Our main goal are to construct a random vector $T_n^\natural$ in $\mathbb{R}^d$ and to derive nonasymptotic bounds for

$$\rho^{\mathrm{re}}(T_n, T_n^\natural) = \sup_{A \in \mathcal{A}^{\mathrm{re}}} |\mathbb{P}(T_n \in A) - \mathbb{P}(T_n^\natural \in A)|. \tag{3}$$

When p (and therefore d) is fixed, the classical central limit theorems (CLT) for approximating T_n by a Gaussian random vector $T_n^\natural \sim N(0, \Gamma)$, where $\Gamma = \mathrm{Cov}(g(X_1))$ and $g(X_1) = \mathbb{E}[h(X_1, X_2)|X_1] - \theta$, have been extensively studied in literature [3, 28, 32–38, 60, 68]. Recently, due to the explosive data enrichment, regularized estimation and dimension reduction of high-dimensional data (i.e., d is larger or even much larger than n) have attracted a lot of research attention such as covariance matrix estimation [9, 10, 20, 30], graphical models [11, 27, 67], discriminant analysis [48], factor models [31, 44] among many others. Those problems all involve the consistent estimation of an expectation $\mathbb{E}[h(X_1, X_2)]$ of U-statistics of order two. Below are three examples.

EXAMPLE 1.1. The sample mean vector $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is an unbiased estimator of $\mathbb{E}X_1$ and $\bar{X}_n$ can be written as a U-statistic of form (1) with the linear kernel $h(x_1, x_2) = (x_1 + x_2)/2$ for $x_1, x_2 \in \mathbb{R}^p$ and $d = p$.

EXAMPLE 1.2. Let $d = p \times p$. The sample covariance matrix $\hat{S}_n = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^\top$ is an unbiased estimator of the covariance matrix $\Sigma = \mathrm{Cov}(X_1)$. Here, $\hat{S}_n$ is a matrix-valued U-statistic of form (1) with the quadratic kernel $h(x_1, x_2) = (x_1 - x_2)(x_1 - x_2)^\top/2$ for $x_1, x_2 \in \mathbb{R}^p$.

EXAMPLE 1.3. The covariance matrix quantifies the linear dependency in a random vector. The rank correlation is another measure for the nonlinear dependency in a random vector. Two generic vectors $y = (y_1, y_2)$ and $z = (z_1, z_2)$ in $\mathbb{R}^2$ are said to be *concordant* if $(y_1 - z_1)(y_2 - z_2) > 0$. For $m, k = 1, \dots, p$, define

$$\tau_{mk} = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \mathbf{1}\{(X_{im} - X_{jm})(X_{ik} - X_{jk}) > 0\}.$$

Then Kendall's tau rank correlation coefficient matrix $T = \{\tau_{mk}\}_{m,k=1}^{p}$ is a matrix-valued U-statistic with a bounded kernel. It is clear that τ_{mk} quantifies the monotonic dependency between (X_{1m}, X_{1k}) and (X_{2m}, X_{2k}) and it is an unbiased estimator of $\mathbb{P}((X_{1m} - X_{2m})(X_{1k} - X_{2k}) > 0)$, that is, the probability that (X_{1m}, X_{1k}) and (X_{2m}, X_{2k}) are concordant.

In this paper we are interested in the following central questions: *How does the dimension impact the asymptotic behavior of U-statistics and how can we make practical statistical inference when $d \to \infty$?* Bounds on (3) with the explicit dependence on d are particularly useful in large-scale statistical inference problems. In particular, motivation of this paper comes from the estimation and inference problems for large covariance matrix and its related functionals [10, 12, 20, 21, 51, 54, 58, 66, 67]. To establish rate of convergence for the regularized estimators or to approximate the limiting null distribution of ℓ^∞-tests in high-dimensions, a key issue is to characterize the distribution of the supremum norm $|U_n - \mathbb{E}U_n|_\infty$ that relates to the probabilities of $\mathbb{P}(T_n \in A)$ for A belonging to the family of max-hyperrectangles in $\mathbb{R}^d$ of the form $A = \{x \in \mathbb{R}^d : x_j \le a \text{ for all } j = 1, \ldots, d\}$ and $-\infty \le a \le \infty$.

Our first main contribution is to provide a Gaussian approximation scheme for the high-dimensional *nondegenerate* U-statistics. Different from the CLT- type results for the sums of independent random vectors [22, 24], which are directly approximated by the Gaussian counterparts with the matching first and second moments, approximation of the U-statistics is more subtle because of its dependency and nonlinearity structures. Here, we propose a *two-step* Gaussian approximation method in Section 2. In the first step, we approximate the U-statistics by the leading component of a linear form in the Hoeffding decomposition (a.k.a. the Hájek projection); in the second step, the linear term is further approximated by the Gaussian random vectors. To approximate the distribution of U-statistics by a linear form, a maximal moment inequality is developed to control the nonlinear and *canonical*, that is, *completely degenerate*, form of the reminder term. Then the linear projection is handled by the recent development of Gaussian approximation in high-dimensions [22, 24, 69, 70]. Explicit rate of convergence of the Gaussian approximation for high-dimensional U-statistics uniformly in the class of all hyperrectangles in $\mathbb{R}^d$ is established for unbounded kernels subject to subexponential and uniform polynomial moment conditions. Specifically, under either moment conditions, we show that the validity of the Gaussian approximation holds for a high-dimensional scaling limit, where d can be larger or even much larger than n. In our results, symmetry of U-statistics is an key ingredient in the Hoeffding decomposition. Therefore, our result can be viewed as nonlinear generalizations of the Gaussian approximation for the high-dimensional sample mean vector of i.i.d. $X_1, \ldots, X_n$.

The second contribution of this paper is to provide *computable* methods for approximating the probabilities $\mathbb{P}(T_n \in A)$ uniformly for $A \in \mathcal{A}^{\mathrm{re}}$. This allows us to compute the quantiles of the maxima $|U_n - \mathbb{E}U_n|_\infty$. Since the covariance matrix Γ of the Hájek projection of the centered U-statistics depends on the underlying data distribution F which is unknown in many real applications, a practically feasible alternative is to use data-dependent approaches such as the bootstrap to approximate $\mathbb{P}(T_n \in A)$, where the insight is to implicitly construct a consistent estimator of Γ under the supremum norm. In Section 3, we provide a unified perspective for the empirical bootstrap (EB), the randomly reweighted bootstrap, and the Gaussian multiplier bootstrap with the jackknife estimator of covariance matrix as randomly reweighted quadratic forms and we establish their validity. Specifically, we show that all three methods are inferentially first-order equivalent for high-dimensional U-statistics in the sense that they achieve the same uniform rate of convergence over $\mathcal{A}^{\mathrm{re}}$. In particular, they are asymptotically valid when the dimension d can be as large as $O(e^{n^c})$ for some constant $c \in (0, 1/7)$. One important feature of the Gaussian and bootstrap approximations is that no structural assumption on the distribution F is made and the strong dependency in F is allowed, which in fact helps the Gaussian and bootstrap approximations.

In Section 4, we apply the proposed bootstrap method to a number of important high-dimensional problems, including the data-dependent tuning parameter selection in the thresholded covariance matrix estimator and the simultaneous inference of the covariance and Kendall's tau rank correlation matrices. Two additional applications for the estimation problems of the sparse precision matrix and the sparse linear functionals of the precision matrix are given in the Supplementary Material (SM, [19]). In those problems, we show that the Gaussian-like convergence rates can be achieved for non-Gaussian data with heavy-tails, which are less conservative than those obtained by the Bonferroni technique that ignores the dependency in the underlying data distribution. For the sparse covariance matrix estimation problem, we also show that the thresholded estimator with the tuning parameter selected by the bootstrap procedure adapts the the dependency and moment in the underlying data distribution and, therefore, the bounds can be much tighter than those of the minimax estimator with a universal threshold that ignores the dependency in F [10, 13, 20].

To establish the Gaussian approximation result and the validity of the bootstrap methods, a key step is to bound the the expected supremum norm of the second-order canonical term in the Hoeffding decomposition of the U-statistics and establish its nonasymptotic maximal moment inequalities. An alternative simple data splitting approach by reducing the U-statistics to sums of i.i.d. random vectors can give the exact rate for bounding the moments in the nondegenerate case [29, 42, 50, 62]. Nonetheless, the reduction to the i.i.d. summands in terms of data splitting does not exploit the complete degeneracy structure of the canonical term and it does not lead to the convergence result in the Gaussian approximation for the nondegenerate U-statistics; see Section 5.1 for details. In addition, unlike the Hoeffding decomposition approach, the data splitting approximation is not asymptotically tight in distribution and, therefore, it is less useful in making inference of the high-dimensional U-statistics.

Relation to the existing literature. For univariate U-statistics, the empirical bootstrap was studied in [2, 7] and the randomly reweighted bootstrap of the form (15) was proposed in [39, 41], where a different class of random weights w_i was considered satisfying $w_i = \xi_i/(n^{-1}\sum_{i=1}^n \xi_i)$ such that ξ_i are i.i.d. nonnegative random variables and $\mathbb{E}\xi_i^2 < \infty$. Weights of such form contain the Bayesian bootstrap as a special case [47, 59]. The randomly reweighted bootstrap with i.i.d. mean-zero weights was considered for the nondegenerate case in [65] and for the degenerate case in [26]. More general exchangeably weighted bootstraps can be found in [40, 49, 57]. However, none of those results in literature can be used to establish the bootstrap validity for high-dimensional U-statistics when $d \gg n$. The Gaussian and bootstrap approximations for the maxima of sums of high-dimensional independent random vectors were considered in [22, 24]. For an i.i.d. sample, this corresponds to a U-statistic with the kernel $h(x_1, x_2) = (x_1 + x_2)/2$ for $x_1, x_2 \in \mathbb{R}^d$. Thus, our results are nonlinear generalizations of those in [22, 24] when $X_1, \dots, X_n$ are i.i.d.

The current paper supersedes and improves the preliminary work [18] (available as an arXiv preprint) by the author. In [18], a Gaussian multiplier bootstrap was proposed by estimating the individual Hájek projection terms using the idea of decoupling on an independent dataset. The bootstrap validity therein is established under the Kolmogorov distance, which is a subset of $\mathcal{A}^{\text{re}}$ corresponding to max-hyperrectangles in $\mathbb{R}^d$. In addition, the rate of convergence in [18] is suboptimal while the rate derived in this paper is nearly optimal; see Remark 3 for detailed comparisons.

Notation and definitions. For a vector x, we use $|x|_1 = \sum_j |x_j|$, $|x| := |x|_2 = (\sum_j x_j^2)^{1/2}$, and $|x|_\infty = \max_j |x_j|$ to denote its entry-wise ℓ^1, ℓ^2, and ℓ^∞ norms, respectively. For a matrix M, we use $|M|_F = (\sum_{i,j} M_{ij}^2)^{1/2}$ and $\|M\|_2 = \max_{|a|=1} |Ma|$ to denote its Frobenius and spectral norms, respectively. We shall use $C, C_1, C_2, \dots$ to denote positive constants that do not depend on n and d and whose values may change from place to place. Denote $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$, $a \asymp b$ if $C_1 a \le b \le C_2 b$ for some constants $C_1, C_2 > 0$. For a random variable X, we write $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ for $q > 0$. For $r = 1, \dots, n$, we shall write $x_1^r = (x_1, \dots, x_r)$ and $\mathbb{E}h = \mathbb{E}h(X_1^r)$ for the random variables $X_1, \dots, X_r$ taking values in a measurable space $(S, \mathcal{S})$ and a measurable function $h : S^r \to \mathbb{R}^d$. For two vectors $x, y \in \mathbb{R}^d$, we use $x \le y$ (or $x > y$) to mean that $x_j \le y_j$ (or $x_j > y_j$) for all $j = 1, \dots, d$. We use $\mathcal{L}(X)$ to denote the law or distribution of the random variable X. For $\alpha > 0$, let $\psi_\alpha(x) = \exp(x^\alpha) - 1$ be a function defined on $[0, \infty)$ and L_{ψ_α} be the collection of all real-valued random variables ξ such that $\mathbb{E}[\psi_\alpha(|\xi|/C)] < \infty$ for some $C > 0$. For $\xi \in L_{\psi_\alpha}$, we define $\|\xi\|_{\psi_\alpha} = \inf\{C > 0 : \mathbb{E}[\psi_\alpha(|\xi|/C)] \le 1\}$. Then, for $\alpha \in [1, \infty)$, $\|\cdot\|_{\psi_\alpha}$ is an Orlicz norm and $(L_{\psi_\alpha}, \|\cdot\|_{\psi_\alpha})$ is a Banach space [45]. For $\alpha \in (0, 1)$, $\|\cdot\|_{\psi_\alpha}$ is a quasi-norm, that is, there exists a constant $C(\alpha) > 0$ such that $\|\xi_1 + \xi_2\|_{\psi_\alpha} \le C(\alpha)(\|\xi_1\|_{\psi_\alpha} + \|\xi_2\|_{\psi_\alpha})$

holds for all $\xi_1, \xi_2 \in L_{\psi_\alpha}$ [1]. We denote the Kolmogorov distance between two real-valued random variables X and Y as $\rho(X, Y) = \sup_{t\in\mathbb{R}} |\mathbb{P}(X \le t) - \mathbb{P}(Y \le t)|$. Throughout the paper, we assume that $n \ge 4$ and $d \ge 3$.

2. Gaussian approximation. In this section we study the approximation for $\mathbb{P}(T_n \in A)$ where $T_n = \sqrt{n}(U_n - \theta)/2$ and $A \in \mathcal{A}^{\mathrm{re}}$. We shall derive a Gaussian approximation result (GAR) for nondegenerate U-statistics, which is the stepping stone to study various bootstrap procedures in Section 3. Let X' and X be two independent random vectors with the distribution F that are also independent of X_1^n. In Sections 2 and 3, since we consider centered U-statistics T_n, we assume without loss of generality that $\theta = 0$. Define $g(X) = \mathbb{E}[h(X, X')|X]$ and $f(X, X') = h(X, X') - g(X) - g(X')$.

DEFINITION 2.1. The kernel $h : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^d$ is said to be: (i) nondegenerate if $\mathrm{Var}(g_m(X)) > 0$ for all $m = 1, \dots, d$; (ii) degenerate of order one, that is, completely degenerate or F-canonical, if $\mathbb{P}(g(X) = 0) = 1$ or equivalently $\mathbb{E}[h(x_1, X')] = \mathbb{E}[h(X, x_2)] = \mathbb{E}[h(X, X')] = 0$ for all $x_1, x_2 \in \mathbb{R}^p$. The corresponding U-statistic in (1) is nondegenerate if h is nondegenerate.

Throughout this paper, we only consider the nondegenerate U-statistics and we assume that:

(M.1) There exists a constant $\underline{b} > 0$ such that $\mathbb{E}[g_m^2(X)] \ge \underline{b}$ for all $m = 1, \dots, d$.

The Hoeffding decomposition of T_n is given by $T_n = L_n + R_n$, where

$$L_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i) \quad \text{and} \quad R_n = \frac{1}{2\sqrt{n}(n-1)} \sum_{1\le i\ne j\le n} f(X_i, X_j).$$

Since f is F-canonical, we expect that L_n is the leading term (a.k.a. the Hájek projection) of T_n. Therefore, we can reasonably expect that T_n is an approximately linear statistic such that $\mathcal{L}(T_n) \approx \mathcal{L}(L_n)$, where the latter can be further approximated by its Gaussian analogue [22, 24]. This motivates the following two-step Gaussian approximation procedure. Let $\Gamma = \mathrm{Cov}(g(X)) = \mathbb{E}(g(X)g(X)^\top)$ be the $d \times d$ covariance matrix of $g(X)$ and $Y \sim N(0, \Gamma)$ be a d-dimensional Gaussian random vector. The main result of this section is to establish nonasymptotic error bounds for $\rho^{\mathrm{re}}(T_n, Y)$ under different moment conditions on h. Let $q > 0$ and $B_n \ge 1$ be a sequence of real numbers possibly tending to infinity. In particular, we shall consider the following assumptions:

(M.2) $\mathbb{E}[|h_m(X, X')|^{2+\ell}] \le B_n^\ell$ for $\ell = 1, 2$ and for all $m = 1, \dots, d$.
(E.1) $\|h_m(X, X')\|_{\psi_1} \le B_n$ for all $m = 1, \dots, d$.
(E.2) $\mathbb{E}[\max_{1\le m\le d}(|h_m(X, X')|/B_n)^q] \le 1$.

In the high-dimensional context, the dimension d grows with the sample size n and the distribution function F may also depend on n. Therefore, B_n is allowed to increase with n. In particular, under (M.1) and (M.2), B_n can be interpreted as a uniform bound on the standardized absolute moments of $g_m(X)$ for $m = 1, \dots, d$. For instance, the kurtosis parameter κ_m of $g_m(X)$ obeys $\kappa_m = [\mathbb{E}g_m^4(X)]/[\mathbb{E}g_m^2(X)]^2 - 3 \le B_n^2/\underline{b}^2 - 3$. Define

$$\varpi_{1,n} = \left(\frac{B_n^2 \log^7(nd)}{n}\right)^{1/6} \quad \text{and} \quad \varpi_{2,n} = \left(\frac{B_n^2 \log^3(nd)}{n^{1-2/q}}\right)^{1/3}. \tag{4}$$

THEOREM 2.1 (*Main result I*: Gaussian approximation for high-dimensional U-statistics for hyperrectangles). *Assume that* (M.1) *and* (M.2) *hold. Suppose that* $\log d \le \bar{b}n$ *for some constant* $\bar{b} > 0$:

(i) *If* (E.1) *holds*, *then there exists a constant* $C := C(\underline{b}, \bar{b}) > 0$ *such that*

$$\rho^{\mathrm{re}}(T_n, Y) \le C\varpi_{1,n}. \tag{5}$$

(ii) *If* (E.2) *holds with* $q \ge 4$, *then there exists a constant* $C := C(\underline{b}, \bar{b}, q) > 0$ *such that*

$$\rho^{\mathrm{re}}(T_n, Y) \le C\{\varpi_{1,n} + \varpi_{2,n}\}. \tag{6}$$

The following corollary is an immediate consequence of Theorem 2.1.

COROLLARY 2.2. *Assume that* (M.1) *and* (M.2) *hold. Let* $K \in (0, 1)$ *and* $\bar{b} > 0$.

(i) *If* (E.1) *holds and* $B_n^2 \log^7(dn) \le \bar{b}n^{1-K}$, *then there exists a constant* $C := C(\underline{b}, \bar{b}) > 0$ *such that*

$$\rho^{\mathrm{re}}(T_n, Y) \le Cn^{-K/6}. \tag{7}$$

In particular,

$$\rho(\bar{T}_n, \bar{Y}) \le Cn^{-K/6}, \tag{8}$$

where $\bar{T}_n = \max_{1\le m\le d} T_{nm}$ *and* $\bar{Y} = \max_{1\le m\le d} Y_m$.

(ii) *If* (E.2) *holds with* $q = 4$ *and* $B_n^4 \log^7(dn) \le \bar{b}n^{1-K}$, *then there exists a constant* $C := C(\underline{b}, \bar{b}) > 0$ *such that* (7) *and* (8) *hold.*

Theorem 2.1 and Corollary 2.2 are nonasymptotic, showing that the validity of the Gaussian approximation for centered nondegenerate U-statistics holds even if d can be much larger than n and no structural assumption on F is required. In particular, Theorem 2.1 applies to kernels with the subexponential distribution such that $\|h_m\|_q \le Cq$ for all $q \ge 1$, in which case $B_n = O(1)$ and the dimension d is allowed to have a subexponential growth rate in the sample size n, that is, $d = O(\exp(n^{(1-K)/7}))$. Condition (E.1) also covers bounded kernels $\|h\|_\infty \le B_n$, where B_n may increase with n.

REMARK 1 (Comments on the near-optimality of the convergence rate in Theorem 2.1). The rate of convergence $n^{-K/6}$ obtained in (7) is slower than the Berry–Esseen rate $n^{-1/2}$ when d is fixed. Similar observations have been made in the existing literature [4, 56] for the normalized sample mean vectors of i.i.d. mean-zero random vectors $X_i \in \mathbb{R}^d$, which corresponds to a U-statistic with the linear kernel $h(x_1, x_2) = (x_1 + x_2)/2$. Assuming $\mathrm{Cov}(X_i) = \mathrm{Id}_d$, [56] showed that $\sqrt{n}\bar{X}_n$ has the asymptotic normality if $d = o(\sqrt{n})$ and [4] showed that

$$\sup_{A \in \mathcal{A}} \left|\mathbb{P}(\sqrt{n}\bar{X}_n \in A) - \mathbb{P}(Y \in A)\right| \le C d^{1/4} \mathbb{E}|X_1|^3 / n^{1/2},$$

where $\mathcal{A}$ is the class of all convex subsets in $\mathbb{R}^d$, $Y \sim N(0, \mathrm{Id}_d)$, and $C > 0$ is an absolute constant. In either case, the dependence of the CLT rate on the dimension d is polynomial ($d/n^{1/2}$ and $d^{7/4}/n^{1/2}$, resp.). On the contrary, our Theorem 2.1 allows d can be larger than n in order to obtain the CLT type results in much higher dimensions. Since the rate $O(n^{-1/6})$ is minimax optimal in infinite-dimensional Banach spaces for the linear kernel case [6, 24], we argue that the rates derived in Theorem 2.1 for U-statistics seem un-improvable in n in the following sense. Let $\{X_{ij}\}_{i=1,\dots,n;\, j=1,\dots,d}$ be an array of i.i.d. mean-zero random variables with the distribution F such that $\mathbb{E}X_{ij}^2 = 1$ and $\|X_{ij}\|_{\psi_1} \le c$ for all $i = 1, \dots, n$ and $j = 1, \dots, d$. Consider the linear kernel. Let $Y \sim N(0, \mathrm{Id}_d)$ and $\bar{Y} = \max_{1 \le j \le d} Y_j$. Denote $\Phi(\cdot)$ and $\phi(\cdot)$ as the c.d.f. and p.d.f. of the standard normal distribution, respectively. By the moderate deviation principle for sums of subexponential random variables (cf. [16], equation (1.1), or [55], Chapter 8, equation (2.41)), there exist constants $C_0, C_1 > 0$ depending only on c such that

$$\frac{\mathbb{P}(T_{nj} > x)}{1 - \Phi(x)} = 1 + \frac{\eta_1(1 + x^3)}{n^{1/2}}, \qquad j = 1, \dots, d$$

for $0 \le x \le C_0 n^{1/6}$ and $|\eta_1| \le C_1$. Then, for all such x in the power zone of normal convergence, we have

$$\mathbb{P}(\bar{T}_n \le x) - \mathbb{P}(\bar{Y} \le x) = \mathbb{P}(\bar{Y} \le x)\{[1 + \eta_2(1 - \Phi(x))/\Phi(x)]^d - 1\},$$

where $\eta_2 = -\eta_1(1 + x^3)n^{-1/2}$. Take a distribution F such that $\eta_1 < 0$. By the inequality $(1 + x)^d \ge 1 + dx$ for $x \ge 0$,

$$\mathbb{P}(\bar{T}_n \le x) - \mathbb{P}(\bar{Y} \le x) \ge |\eta_1|(1 + x^3)n^{-1/2}\mathbb{P}(\bar{Y} \le x)d[1 - \Phi(x)].$$

Let x^* be the median of $\bar{Y}$; that is, $\mathbb{P}(\bar{Y} \le x^*) = 1/2$. Then $x^* \asymp \sqrt{2\log d}$. In fact, by [25], Corollary 3.1, we have $x^* \le \sqrt{2\log d}$ for $d \ge 31$. Thus, if $x^* \le C_0 n^{1/6}$, then using $[1 - \Phi(x)]/[x^{-1}\phi(x)] \to 1$ as $x \to \infty$ we have

$$\rho(\bar{T}_n, \bar{Y}) \ge C_2 n^{-1/2} x^{*2} d \exp(-x^{*2}/2) \ge C_2 n^{-1/2} x^{*2}.$$

Hence, there exist constants C and C' depending only on F such that if $(\log d)^3 \le C'n$, then $\rho(\bar{T}_n, \bar{Y}) \ge Cn^{-1/2}\log d$. In particular, taking $(\log d)^3 \asymp n$, we have

$\rho(\bar{T}_n, \bar{Y}) \geq Cn^{-1/6}$. Therefore, in view of the upper bound in (5) and the lower bound in [6, 24], we conjecture that the optimal rate for $\rho(\bar{T}_n, \bar{Y})$ in the high-dimensional setting is $O((n^{-1}B_n^2 \log^a(nd))^{1/6})$ for some $a > 0$, based on which the rate of convergence in (5) is also nearly optimal in d. However, a rigorous lower bound for $\rho(\bar{T}_n, \bar{Y})$ is still an open question. By the moderate deviations for self-normalized sums [61] and the argument above, we expect that similar comments apply for X_{ij} with weaker polynomial moment conditions.

Theorem 2.1 and Corollary 2.2 can be viewed as nonlinear generalization of the results in [22, 24], which considered the Gaussian approximation for $\max_{1\leq j\leq d} \sqrt{n}\bar{X}_{nj}$. Therefore, for U-statistics with a nonlinear kernel h (possibly unbounded and discontinuous), the effect of higher-order terms than the Hájek projection to a linear subspace in the Hoeffding decomposition vanishes in the Gaussian approximation. For multivariate symmetric statistics of order two, to the best of our knowledge, the Gaussian approximation result (5), (6), (7) and (8) with the explicit convergence rate is new. When d is fixed, the rate of convergence and the Edgeworth expansion of such statistics can be found in [5, 8, 33]. In those papers, assuming the Cramér condition on $g(X_1)$ and suitable moment conditions on $h(X_1, X_2)$, the Edgeworth expansion of U-statistics was established for the univariate case ($d = 1$) with remainder $o(n^{-1/2})$ or $O(n^{-1})$ [5, 8] and the multivariate case ($d > 1$ fixed) with remainder $o(n^{-1/2})$ [33]. In the latter work [33], it is unclear that how the constant in the error bound depends on the dimensionality parameter d.

Theorem 2.1 and Corollary 2.2 allow us to approximate the probabilities of T_n belonging to the hyperrectangles in $\mathbb{R}^d$ by those probabilities of Y, with the knowledge of Γ. Such results are useful for approximating the quantiles of $\bar{T}_n$ by those of $\bar{Y}$. In practice, the covariance matrix Γ and the Hájek projection terms $g(X_i), i = 1, \ldots, n$, depend on the underlying data distribution F, which is unknown. Thus, quantiles of $\bar{Y}$ need to be estimated in real applications. However, we shall see in Section 3 that Theorem 2.1 can still be used to derive valid and computable (i.e., fully data-dependent) methods to approximate the quantiles of $\bar{T}_n$.

3. Bootstrap approximations. In this section we consider *computable* approximations of the probabilities $\mathbb{P}(T_n \in A)$ for $A \in \mathcal{A}^{\mathrm{re}}$. Before proceeding to the rigorous results, we shall explain our general strategy. The validity of the bootstrap procedures is established by a series of approximations:

$$\mathcal{L}(T_n) \approx_{(1)} \mathcal{L}(Y) \approx_{(2)} \mathcal{L}\big(Z^X | X_1^n\big) \approx_{(3)} \mathcal{L}\big(T_n^\natural | X_1^n\big), \tag{9}$$

where Z^X is a conditionally mean-zero Gaussian random vector in $\mathbb{R}^d$ given the observed sample X_1^n. The choice of Z^X and $T_n^\natural$ depends on the specific bootstrap method such that the conditional covariance matrix of Z^X given X_1^n is a consistent estimator of Γ under the supremum norm. Step (1) follows from the GAR and CLT in Section 2. Step (2) relies on a (conditional) Gaussian comparison principle and the tail probability inequalities of maximal U-statistics to bound the probability of the events on which the Gaussian comparison can be applied. Those tail probability inequalities are developed in the SM (Section E), which are of independent interest and may be used for other high-dimensional problems. Step (3) is a conditional version of Step (1) given X_1^n.

3.1. *Empirical bootstrap*. Let $X_1^*, \dots, X_n^*$ be a bootstrap sample independently drawn from the empirical distribution $\hat{F}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ_x is the Dirac point mass at x. Define

$$U_n^* = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i^*, X_j^*). \tag{10}$$

Then the conditional distribution of $T_n^* = \sqrt{n}(U_n^* - \mathbb{E}[U_n^* | X_1^n])/2$ given X_1^n is used to approximate the distribution of T_n. Here, $T_n^\natural = T_n^*$ in (9). Note that $\mathbb{E}[U_n^* | X_1^n] = V_n$, where $V_n = n^{-2} \sum_{i,j=1}^n h(X_i, X_j)$ is a V-statistic. Let

$$\xi_i \overset{\text{i.i.d.}}{\sim} \text{multinomial}(1; 1/n, \dots, 1/n).$$

Denote $\boldsymbol{\xi}_{n \times n} = (\xi_1, \dots, \xi_n)$ and $\mathbf{X}_{p \times n} := X_1^n = (X_1, \dots, X_n)$. Then we can write $\mathbf{X}^* = (X_1^*, \dots, X_n^*) = \mathbf{X}\boldsymbol{\xi}$. The key observation is that conditional on $\mathbf{X}$, U_n^* is a U-statistic of $\xi_1, \dots, \xi_n$ since

$$U_n^* = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(\mathbf{X}\xi_i, \mathbf{X}\xi_j).$$

Therefore, we can perform the conditional Hoeffding decomposition as follows. Let

$$\begin{aligned} g^X(\xi_1) &= \mathbb{E}\big[h(\mathbf{X}\xi_1, \mathbf{X}\xi_2) | \xi_1, X_1^n\big] - V_n \\ &= \frac{1}{n} \sum_{j=1}^n h(\mathbf{X}\xi_1, X_j) - \frac{1}{n^2} \sum_{i,j=1}^n h(X_i, X_j). \end{aligned}$$

Then $\mathbb{E}[g^X(\xi_1) | X_1^n] = 0$ and

$$\hat{\Gamma}_n := \operatorname{Cov}\big(g^X(\xi_1) | X_1^n\big) = \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n h(X_i, X_j) h(X_i, X_k)^\top - V_n V_n^\top. \tag{11}$$

For the special case $d = 1$, by the strong law of large numbers for U-statistics ([60], Theorem A, page 190) we have with probability one

$$\lim_{n \to \infty} \operatorname{Var}\big(g^X(\xi_1) | X_1^n\big) = \operatorname{Var}\big(g(X_1)\big) = \mathbb{E}\big\{\mathbb{E}\big[h(X_1, X_2) | X_1\big]\big\}^2 - \big\{\mathbb{E}\big[h(X_1, X_2)\big]\big\}^2.$$

Therefore, we expect that T_n^* is a reasonable approximation of T_n and our goal to is bound the random quantity

$$\rho^B(T_n, T_n^*) = \sup_{A \in \mathcal{A}^{\mathrm{re}}} \left| \mathbb{P}(T_n \in A) - \mathbb{P}(T_n^* \in A \mid X_1^n) \right|.$$

In addition to (M.2), (E.1) and (E.2), we shall also assume that:

(M.2′) $\mathbb{E}[|h_m(X, X)|^{2+\ell}] \le B_n^\ell$ for $\ell = 1, 2$ and for all $m = 1, \dots, d$.
(E.1′) $\|h_m(X, X)\|_{\psi_1} \le B_n$ for all $m = 1, \dots, d$.
(E.2′) $\mathbb{E}[\max_{1 \le m \le d}(|h_m(X, X)|/B_n)^q] \le 1$.

(M.2′), (E.1′) and (E.2′) are the von Mises conditions on the empirical bootstrap of U-statistics [7], which require that the diagonal entries of the kernel h obey the same moment conditions as the off-diagonal ones (M.2), (E.1) and (E.2), respectively. Without (M.2′), (E.1′) and (E.2′), the empirical bootstrap (Theorem 3.1) can fail and a counterexample was given in [7]; see also [46], Chapter 6.5. For $\gamma \in (0, e^{-1})$, define

$$\varpi_{1,n}^B(\gamma) = \left(\frac{B_n^2 \log^5(nd) \log^2(1/\gamma)}{n} \right)^{1/6} \quad \text{and} \tag{12}$$
$$\varpi_{2,n}^B(\gamma) = \left(\frac{B_n^2 \log^3(nd)}{\gamma^{2/q} n^{1-2/q}} \right)^{1/3}.$$

THEOREM 3.1 (*Main result II*: rate of convergence of the empirical bootstrap for U-statistics). *Suppose that* (M.1), (M.2) *and* (M.2′) *are satisfied. Assume that* $\log(1/\gamma) \le K \log(dn)$ *and* $\log d \le \bar{b} n$ *for some constants* $K, \bar{b} > 0$.

(i) *If* (E.1) *and* (E.1′) *hold, then there exists a constant* $C := C(\underline{b}, \bar{b}, K) > 0$ *such that we have with probability at least* $1 - \gamma$

$$\rho^B(T_n, T_n^*) \le C \varpi_{1,n}. \tag{13}$$

(ii) *If* (E.2) *and* (E.2′) *hold with* $q \ge 4$, *then there exists a constant* $C := C(\underline{b}, \bar{b}, q, K) > 0$ *such that we have with probability at least* $1 - \gamma$

$$\rho^B(T_n, T_n^*) \le C\{\varpi_{1,n} + \varpi_{2,n}^B(\gamma)\}. \tag{14}$$

Theorem 3.1 is nonasymptotic, which implies the asymptotic validity of the EB for U-statistics in the almost sure sense.

COROLLARY 3.2 (Asymptotic validity of the empirical bootstrap for U-statistics in the almost sure sense). *Suppose that* (M.1), (M.2) *and* (M.2′) *are satisfied and* $\log d \le \bar{b} n$ *for some constant* $\bar{b} > 0$.

(i) *Under* (E.1) *and* (E.1′), *we have* $\mathbb{P}(\rho^B(T_n, T_n^*) \le C \varpi_{1,n}$ *for all but finitely many* $n) = 1$, *where* $C > 0$ *is a constant depending only on* $\underline{b}$ *and* $\bar{b}$. *In particular, if* $B_n^2 \log^7(nd) = o(n)$, *then* $\rho^B(T_n, T_n^*) \to 0$ *almost surely.*

(ii) *Under* (E.2) *and* (E.2′) *with* $q > 4$, *we have* $\mathbb{P}(\rho^B(T_n, T_n^*) \le C\{\varpi_{1,n} + \varpi'^B_{2,n}\}$ *for all but finitely many* $n) = 1$, *where*

$$\varpi'^B_{2,n} = \left(\frac{B_n^2 \log^3(nd)\log^{4/q}(n)}{n^{1-4/q}}\right)^{1/3}$$

and $C > 0$ *is a constant depending only on* $\underline{b}, \bar{b}$, *and* q. *In particular, if* $B_n^2 \log^7(nd) = o(n)$ *and* $B_n^2 \log^3(nd)\log^{4/q}(n) = o(n^{1-4/q})$, *then* $\rho^B(T_n, T_n^*) \to 0$ *almost surely.*

3.2. *Randomly reweighted bootstrap with i.i.d. Gaussian weights.* Let $w_1, \dots, w_n$ be i.i.d. $N(1, 1)$ random variables that are also independent of X_1^n and Y. Consider

$$U_n^\diamond = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} w_i w_j h(X_i, X_j). \tag{15}$$

Then $U_n^\diamond$ is the stochastically reweighted version of U_n and it can also be viewed as a random quadratic form in $w_1, \dots, w_n$. Denote $T^\diamond = \sqrt{n}(U_n^\diamond - U_n)/2$ and $T_n^\natural = T_n^\diamond$ in (9). Since the main focus of this paper is to approximate the distribution of the centered U-statistics; that is, $\theta := \mathbb{E}h = 0$, we first consider the bootstrap of the centered U-statistics of the random quadratic form (15) and discuss the effect of centering in the bootstraps in Remark 2.

THEOREM 3.3 (*Main result III*: rate of convergence of the randomly reweighted bootstrap for centered U-statistics). *Assume that* $\theta = 0$. *Suppose that* (M.1) *and* (M.2) *are satisfied. Assume that* $\log(1/\gamma) \le K \log(dn)$ *and* $\log d \le \bar{b}n$ *for some constants* $K, \bar{b} > 0$.

(i) *If* (E.1) *holds, then there exists a constant* $C := C(\underline{b}, \bar{b}, K) > 0$ *such that we have* $\rho^B(T_n, T_n^\diamond) \le C\varpi_{1,n}$ *holds with probability at least* $1 - \gamma$.

(ii) *If* (E.2) *holds with* $q \ge 4$, *then there exists a constant* $C := C(\underline{b}, \bar{b}, q, K) > 0$ *such that we have* $\rho^B(T_n, T_n^\diamond) \le C\{\varpi_{1,n} + \varpi^B_{2,n}(\gamma)\}$ *holds with probability at least* $1 - \gamma$.

From Theorems 3.1 and 3.3, we see that the empirical and the randomly reweighted bootstraps are first-order equivalent, both achieving the same uniform rate of convergence for approximating the probabilities $\mathbb{P}(T_n \in A)$ for $A \in \mathcal{A}^{\mathrm{re}}$. However, unlike the EB, the randomly reweighted bootstrap does not assume the von Mises moment conditions on the diagonal entries.

REMARK 2 (Effect of centering in the randomly reweighted bootstrap). If $\theta \ne 0$, then we can show that the i.i.d. reweighted bootstrap $T_n^\diamond$ is not an asymptotically valid bootstrap approximation for T_n. The reason is that centering is a key structure to maintain in the conditional distribution of $T_n^{\diamond}$. Therefore, in the case, we shall consider the following modified version:

$$U_n^{\flat} = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} w_i w_j h(X_i, X_j) - 2(\bar{w} - 1)U_n \tag{16}$$

and $T_n^{\flat} = \sqrt{n}(U_n^{\flat} - U_n)/2$. Then, $T_n^{\flat} = T_n^{\diamond} - \sqrt{n}(\bar{w} - 1)U_n$. Since w_i are i.i.d. $N(1, 1)$, we have $\sqrt{n}(\bar{w} - 1)U_n | X_1^n \sim N(0, U_n U_n^{\top})$, which is not asymptotically vainishing for $\theta \ne 0$. Therefore, without the centering term $2(\bar{w} - 1)U_n$ in $U_n^{\flat}$, $T_n^{\diamond}$ is not an asymptotically tight sequence for approximating T_n.

THEOREM 3.4 (Rate of convergence of the randomly reweighted bootstrap for noncentered U-statistics). *Suppose that* (M.1) *and* (M.2) *are satisfied. Assume that* $\log(1/\gamma) \le K \log(dn)$ *and* $\log d \le \bar{b}n$ *for some constants* $K, \bar{b} > 0$.

(i) *If* (E.1) *holds, then there exists a constant* $C := C(\underline{b}, \bar{b}, K) > 0$ *such that we have* $\rho^B(T_n, T_n^{\flat}) \le C\varpi_{1,n}$ *holds with probability at least* $1 - \gamma$.

(ii) *If* (E.2) *holds with* $q \ge 4$, *then there exists a constant* $C := C(\underline{b}, \bar{b}, q, K) > 0$ *such that we have* $\rho^B(T_n, T_n^{\flat}) \le C\{\varpi_{1,n} + \varpi_{2,n}^B(\gamma)\}$ *holds with probability at least* $1 - \gamma$.

Theorem 3.4 is valid regardless $\theta \ne 0$ and $\theta = 0$ since in the latter case, $\sqrt{n}(\bar{w} - 1)U_n$ is conditionally negligible compared with $T_n^{\diamond}$. For the EB, centering in the empirical analog $\hat{\Gamma}_n$ of the covariance matrix Γ is automatically fulfilled; see (11). Similar comments apply to the Gaussian multiplier bootstrap $T_n^{\sharp}$ in Section 3.3.

3.3. *Gaussian multiplier bootstrap with jackknife covariance matrix estimator.* The i.i.d. reweighted bootstrap is closely related to the Gaussian multiplier bootstrap with the jackknife estimator of the covariance matrix of T_n. Let $e_1, \dots, e_n$ be i.i.d. $N(0, 1)$ random variables that are independent of X_1^n and Y and

$$T_n^{\sharp} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{1}{n-1} \sum_{j \ne i} h(X_i, X_j) - U_n \right] e_i. \tag{17}$$

Define

$$\hat{\Gamma}_n^{JK} = \frac{1}{(n-1)(n-2)^2} \sum_{i=1}^{n} \sum_{j \ne i} \sum_{k \ne i} \big(h(X_i, X_j) - U_n\big)\big(h(X_i, X_k) - U_n\big)^{\top}. \tag{18}$$

Then, $\hat{\Gamma}_n^{JK}$ is the jackknife estimator of the covariance matrix of T_n [14] and $T_n^{\sharp} | X_1^n \sim N(0, \tilde{\Gamma}_n)$, where

$$\tilde{\Gamma}_n = \frac{(n-2)^2}{n(n-1)} \hat{\Gamma}_n^{JK}. \tag{19}$$

Therefore, it is interesting to view the Gaussian multiplier bootstrap $T_n^\sharp$ as a plug-in estimator of the distribution of T_n by its jackknife covariance matrix estimator. To distinguish the Gaussian wild bootstrap $\hat{L}_0^*$ in (20) (cf. Remark 3), we call $T_n^\sharp$ the *jackknife Gaussian multiplier bootstrap*.

THEOREM 3.5 (*Main result IV*: rate of convergence of the jackknife Gaussian multiplier bootstrap for U-statistics). *Suppose that* (M.1) *and* (M.2) *are satisfied. Assume that* $\log(1/\gamma) \le K \log(dn)$ *and* $\log d \le \bar{b}n$ *for some constants* $K, \bar{b} > 0$.

(i) *If* (E.1) *holds, then there exists a constant* $C := C(\underline{b}, \bar{b}, K) > 0$ *such that we have* $\rho^B(T_n, T_n^\sharp) \le C\varpi_{1,n}$ *holds with probability at least* $1-\gamma$.

(ii) *If* (E.2) *holds with* $q \ge 4$, *then there exists a constant* $C := C(\underline{b}, \bar{b}, q, K) > 0$ *such that we have* $\rho^B(T_n, T_n^\sharp) \le C\{\varpi_{1,n} + \varpi_{2,n}^B(\gamma)\}$ *holds with probability at least* $1-\gamma$.

In the special case $h(x_1, x_2) = (x_1 + x_2)/2$ for $x_1, x_2 \in \mathbb{R}^d$, we have $U_n = \bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is the sample mean vector and $T_n = \sqrt{n}(\bar{X}_n - \theta)/2$ where $\theta = \mathbb{E}(X_1)$. Some algebra shows that $\hat{\Gamma}_n^{JK} = [4(n-1)]^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^\top$, $\hat{\Gamma}_n = n^{-1}(n-1)\hat{\Gamma}_n^{JK}$ in (11), and $\tilde{\Gamma}_n = [(n-2)^2/(n(n-1))]\hat{\Gamma}_n^{JK}$ in (19). Then $T_n^\sharp \sim N(0, \tilde{\Gamma}_n)$ which is the equivalent to the multiplier bootstrap of [24]. Therefore, for i.i.d. samples, Theorems 3.1 and 3.5 are nonlinear generalizations of the empirical and Gaussian multiplier bootstraps considered in [24].

REMARK 3 (Comparison with the Gaussian wild bootstrap of [18]). In [18], a Gaussian wild bootstrap based on decoupling was proposed. Specifically, let $X_1', \dots, X_n'$ be an independent copy of $X_1, \dots, X_n$. The Hájek projection terms $g(X_i), i = 1, \dots, n$, are estimated by

$$\hat{g}_i = \frac{1}{n}\sum_{j=1}^n h(X_i, X_j') - \frac{1}{n(n-1)}\sum_{1\le j\ne l\le n} h(X_j', X_l').$$

Since $g(X_i) = \mathbb{E}[h(X_i, X')|X_i] - \mathbb{E}[h(X, X')]$, $\hat{g}_i$ can be viewed as an unbiased estimator of $g(X_i)$ conditionally on X_i for $i = 1, \dots, n$. Then the Gaussian wild bootstrap procedure is defined as

$$\hat{L}_0^* = \max_{1\le m\le d} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{g}_{im} e_i, \tag{20}$$

where e_i are i.i.d. $N(0,1)$ random variables. Let $a_{\hat{L}_0^*}(\alpha)$ be the αth conditional quantile of $\hat{L}_0^*$ given $X_1, \dots, X_n$ and $X_1', \dots, X_n'$. Similarly, denote $a_{\bar{T}_n^\sharp}(\alpha)$ as the αth conditional quantile of $\bar{T}_n^\sharp$ given $X_1, \dots, X_n$. Let $K \in (0,1)$ be a constant. Assuming (M.1), (M.2) and in addition $D_2 \le 1$, it was shown in [18] that:

(i) if $B_n^2 \log^7(dn) \le n^{1-K}$ and (E.1) holds, then $\sup_{\alpha\in(0,1)} |\mathbb{P}(\bar{T}_n \le a_{\hat{L}_0^*}(\alpha)) - \alpha| \le Cn^{-K/8}$; (ii) if $B_n^4 \log^7(dn) \le n^{1-K}$ and (E.2) holds with $q = 4$, then $\sup_{\alpha\in(0,1)} |\mathbb{P}(\bar{T}_n \le a_{\hat{L}_0^*}(\alpha)) - \alpha| \le Cn^{-K/12}$. Here, the constant $C > 0$ depends only on $\underline{b}$ in (M.1) in both cases. The following theorem shows that the jackknife Gaussian multiplier bootstrap improves the convergence rate of Gaussian wild bootstrap in [18].

THEOREM 3.6. *Suppose that* (M.1) *and* (M.2) *are satisfied. Let* $K \in (0, 1)$.

(i) *Assume* (E.1). *If* $B_n^2 \log^7(dn) \le n^{1-K}$, *then*

$$\sup_{\alpha\in(0,1)} \bigl|\mathbb{P}\bigl(\bar{T}_n \le a_{\bar{T}_n^\sharp}(\alpha)\bigr) - \alpha\bigr| \le Cn^{-K/6}, \tag{21}$$

where $\bar{T}_n^\sharp = \max_{1\le j\le d} T_{nj}^\sharp$ *and* $C > 0$ *is a constant depending only on* $\underline{b}$.

(ii) *Assume* (E.2) *with* $q \ge 4$. *If* $B_n^2 \log^7(dn) \le n^{1-K}$ *and* $B_n^4 \log^6(d) \le n^{2-4(1+K/6)/q-K}$, *then we have* (21) *with the constant* C *depending only on* $\underline{b}$ *and* q.

In particular, for both subexponential and uniform polynomial kernels, the convergence rate of jackknife Gaussian multiplier bootstrap $T_n^\sharp$ is $O(n^{-K/6})$. The improved dependence on n is due to two reasons. First, we established a GAR for high-dimensional U-statistics with sharper rate (Theorem 2.1). Second, $T_n^\sharp$ does not estimate the individual terms $g(X_i)$ in the Hájek projection which requires a strong control on the maximal deviation $|\hat{g}_i - g(X_i)|_\infty$ over $i = 1, \dots, n$; see Lemma C.4 in [18]. Instead, $T_n^\sharp$ implicitly constructs an estimator $\tilde{\Gamma}_n$ in (19) for the covariance matrix of the linear projection part in the Gaussian approximation. There is a slight trade-off between the moment and scaling limit for uniform polynomial kernels in Theorem 3.6 since the conditions $B_n^2 \log^7(dn) \le n^{1-K}$ and $B_n^4 \log^6(d) \le n^{2-4(1+K/6)/q-K}$ are implied by either $B_n^4 \log^7(dn) \le n^{1-7K/6}$ for $q = 4$ or $B_n^4 \log^7(dn) \le n^{1-K}$ for $q \ge 4(1 + K/6)$. However, in either case, Theorem 3.6 asymptotically permits $d = O(e^{n^c})$ for some $c \in (0, 1/7)$ when $q \ge 4$ and $B_n = O(1)$.

4. Statistical applications. In this section we present two statistical applications for bootstrap methods in Section 3.1. For simplicity, we only present the results for the jackknife Gaussian multiplier bootstrap $T_n^\sharp$ defined in (17). Similar results hold for other bootstraps in Section 3.1. Two additional examples can be found in the SM. Throughout the section, we consider the bootstrap of the sample covariance matrix [i.e., $h(x_1, x_2) = (x_1 - x_2)(x_1 - x_2)^\top/2$ and $\mathbb{R}^d = \mathbb{R}^{p\times p}$]. We define $\bar{T}_n^\sharp = 2n^{-1/2} \max_{1\le m,k\le p} |T_{n,mk}^\sharp|$ by rescaling and denote the αth conditional quantile of $\bar{T}_n^\sharp$ given the data X_1^n as

$$a_{\bar{T}_n^\sharp}(\alpha) = \inf\bigl\{t \in \mathbb{R} : \mathbb{P}_e\bigl(\bar{T}_n^\sharp \le t\bigr) \ge \alpha\bigr\}, \tag{22}$$

where $\mathbb{P}_e$ is the probability taken w.r.t. the i.i.d. $N(0,1)$ random variables $e_1,\dots,e_n$. We can compute the conditional quantile $a_{T_n^\sharp}(\alpha)$ by repeatedly drawing independent samples of the standard Gaussian random variables $e_1,\dots,e_n$.

4.1. *Tuning parameter selection for the thresholded covariance matrix estimator.* Consider the problem of *sparse* covariance matrix estimation. Let $r \in [0,1)$ and

$$\mathcal{G}(r,\zeta_p) = \left\{\Sigma \in \mathbb{R}^{p\times p} : \max_{1\le m\le p} \sum_{k=1}^{p} |\sigma_{mk}|^r \le \zeta_p\right\}$$

be the class of sparse covariance matrices in terms of the strong ℓ^r-ball. Here, $\zeta_p > 0$ may grow with p. Let $\hat{S}_n = \{\hat{s}_{mk}\}_{m,k=1}^p$ be the sample covariance matrix and

$$\hat{\Sigma}(\tau) = \{\hat{s}_{mk}\mathbf{1}\{|\hat{s}_{mk}| > \tau\}\}_{m,k=1}^p, \qquad \tau \ge 0,$$

be the thresholded sample covariance matrix estimator of Σ. A similar matrix class as $\mathcal{G}(r,\zeta_p)$ was introduced in [10] by further requiring that $\max_{1\le m\le p}\sigma_{mm} \le C_0$ for some constant $C_0 > 0$. Here, we do not assume the diagonal entries of Σ are bounded. Performance bounds of the thresholded estimator $\hat{\Sigma}(\tau)$ critically depend on the tuning parameter τ. The oracle choice of the threshold for establishing the rate of convergence under the spectral and Frobenius norms is $\tau_\diamond = |\hat{S}_n - \Sigma|_\infty$. Note that $\tau_\diamond$ is a random variable and its distribution depends on the unknown underlying data distribution F. High probability bounds of $\tau_\diamond$ were given in [10, 20] and asymptotic properties of $\hat{\Sigma}(\tau)$ were analyzed in [10, 13] for i.i.d. sub-Gaussian data and in [20, 21] for heavy-tailed time series with polynomial moments. In both scenarios, the rates of convergence were obtained with the Bonferroni (i.e., the union bound) technique and one-dimensional concentration inequalities. In the problem of the high-dimensional sparse covariance matrix estimation, data-dependent tuning parameter selection is often empirically done with the cross-validation (CV) and its theoretical properties when compared with $\tau_\diamond$ largely remain unknown since the CV threshold does not approximate $\tau_\diamond$. Here, we provide a principled and fully data-dependent way to determine the threshold τ. We first consider sub-Gaussian observations.

DEFINITION 4.1 (Sub-Gaussian random variable). A random variable X is said to be *sub-Gaussian* with mean zero and *variance factor* ν^2, if

$$\mathbb{E}[\exp(X^2/\nu^2)] \le \sqrt{2}. \tag{23}$$

Denote $X \sim$ sub-Gaussian(ν^2). In particular, if $X \sim N(0,\sigma^2)$, then $X \sim$ sub-Gaussian$(4\sigma^2)$.

The upper bound $\sqrt{2}$ in (23) is not essential and it is chosen for conveniently comparing with $\|X\|_{\psi_2}$: if $X \sim \text{sub-Gaussian}(\nu^2)$, then $\nu^2 \geq \|X\|_{\psi_2}$. Clearly, bounded random variables are sub-Gaussian. In addition, random variables with the mixture of sub-Gaussian distributions are also sub-Gaussian. Let K be a positive integer and $\{\pi_k\}_{k=1}^K$ be sub-Gaussian distributions with the variance factors $\{\nu_k^2\}_{k=1}^K$. If a random variable X follows a mixture of K sub-Gaussian distributions $\sum_{k=1}^K \varepsilon_k \pi_k$ with $0 \leq \varepsilon_k \leq 1$ and $\sum_{k=1}^K \varepsilon_k = 1$, then $X \sim \text{sub-Gaussian}(\bar{\nu}^2)$, where $\bar{\nu}^2 = \max\{\nu_1^2, \dots, \nu_K^2\}$. In general, the variance factor for a sub-Gaussian random variable is *not* equivalent to the variance. For a sequence of random variables $X_n, n = 1, 2, \dots$, if $X_n \sim \text{sub-Gaussian}(\nu_n^2)$ and $\sigma_n^2 = \text{Var}(X_n)$, then by Markov's inequality, we always have $\sigma_n^2 \leq \sqrt{2}\nu_n^2$, while ν_n^2 may diverge at faster rate than σ_n^2 as $n \to \infty$. Below we shall give two such examples.

EXAMPLE 4.1 (Mixture of two Gaussian distributions). Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables with the distribution $F_n = (1 - \varepsilon_n)N(0, 1) + \varepsilon_n N(0, a_n^2)$. Suppose that $a_n \geq 1$, $a_n \to \infty$ as $n \to \infty$, and consider $\varepsilon_n = a_n^{-4}$. Then we have $X_n \sim \text{sub-Gaussian}(4a_n^2)$, $\text{Var}(X_n) \asymp 1$, $\|X_n\|_4 \asymp 1$, $\|X_n\|_6 \asymp a_n^{1/3}$, and $\|X_n\|_8 \asymp a_n^{1/2}$. The distribution F_n can be viewed as a ε_n-contaminated one-dimensional normal distribution given by (90) in the SM [19].

EXAMPLE 4.2 (Mixture of two symmetric discrete distributions). Let π_1 be the distribution of a Rademacher random variable Y [i.e., $\mathbb{P}(Y = \pm 1) = 1/2$] and π_2 be the distribution of a discrete random variable Z_n such that $\mathbb{P}(Z_n = \pm a_n) = (2a_n^2)^{-1}$ and $\mathbb{P}(Z_n = 0) = 1 - a_n^{-2}$. Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables with the distribution $F_n = (1 - \varepsilon_n)\pi_1 + \varepsilon_n \pi_2$, where $\varepsilon_n = 2/(a_n^2 - 1)$, $a_n > \sqrt{3}$, and $a_n \to \infty$ as $n \to \infty$. Then $X_n \sim \text{sub-Gaussian}(Ca_n^2)$ for some large enough constant $C > 0$ and elementary calculations show that $\text{Var}(X_n) = 1$, $\|X_n\|_4 = 3^{1/4}$, $\|X_n\|_6 \asymp a_n^{1/3}$, and $\|X_n\|_8 \asymp a_n^{1/2}$.

Therefore, in the statistical applications for sub-Gaussian data, we allow $\nu_n^2 \to \infty$ as $n \to \infty$. Let $\xi_q = \max_{1 \leq k \leq p} \|X_{1k}\|_q$ and recall $\Gamma = \text{Cov}(g(X_1))$.

THEOREM 4.1 (*Main result V*: data-driven threshold selection: sub-Gaussian observations). *Let $\nu_n \geq 1$ and X_i be i.i.d. mean zero random vectors in $\mathbb{R}^p$ such that $X_{ik} \sim$ sub-Gaussian(ν_n^2) for all $k = 1, \dots, p$ and $\Sigma \in \mathcal{G}(r, \zeta_p)$. Suppose that there exist constants $C_i > 0, i = 1, \dots, 3$, such that $\Gamma_{(j,k),(j,k)} \geq C_1$, $\xi_6 \leq C_2 \nu_n^{1/3}$ and $\xi_8 \leq C_3 \nu_n^{1/2}$ for all $j, k = 1, \dots, p$. Let $\alpha, \beta, K \in (0, 1)$ and*

$\tau_* = \beta^{-1} a_{\bar{T}_n^\sharp}(1-\alpha)$. If $\nu_n^4 \log^7(np) \le C_4 n^{1-K}$, then we have

$$\|\hat{\Sigma}(\tau_*) - \Sigma\|_2 \le \left[\frac{3+2\beta}{\beta^{1-r}} + \left(\frac{\beta}{1-\beta}\right)^r\right] \zeta_p a_{\bar{T}_n^\sharp}^{1-r}(1-\alpha), \tag{24}$$

$$\frac{1}{p}|\hat{\Sigma}(\tau_*) - \Sigma|_F^2 \le 2\left[\frac{4+3\beta^2}{\beta^{2-r}} + 2\left(\frac{\beta}{1-\beta}\right)^r\right] \zeta_p a_{\bar{T}_n^\sharp}^{2-r}(1-\alpha), \tag{25}$$

with probability at least $1-\alpha - Cn^{-K/6}$ for some constant $C > 0$ depending only on $C_1, \dots, C_4$. In addition, $\mathbb{E}[a_{\bar{T}_n^\sharp}(1-\alpha)] \le C' \xi_4^2 (\log(p)/n)^{1/2}$ and

$$\mathbb{E}[\tau_*] \le C' \beta^{-1} \xi_4^2 (\log(p)/n)^{1/2}, \tag{26}$$

where $C' > 0$ is a constant depending only on α and $C_1, \dots, C_4$.

REMARK 4 (Comments on the conditions in Theorem 4.1). Conditions on the growth rate $\xi_6 \le C_2 \nu_n^{1/3}$ and $\xi_8 \le C_3 \nu_n^{1/2}$ are satisfied by Examples 4.1 and 4.2. The nondegeneracy condition $\Gamma_{(j,k),(j,k)} \ge C_1$ is quite mild. Consider the multivariate cumulants of the joint distribution of the random vector $X = (X_1, \dots, X_p)^\top$ following a distribution F in $\mathbb{R}^p$. Let $\chi(t) = \mathbb{E}[\exp(\iota t^\top X)]$ be the characteristic function of X, where $t = (t_1, \dots, t_p)^\top \in \mathbb{R}^p$ and $\iota = \sqrt{-1}$. Then the *multivariate cumulants* $\kappa_{r_1 r_2 \cdots r_p}^{12\cdots p}$ of the joint distribution of X are the coefficients in the expansion:

$$\log \chi(t) = \sum_{r_1, r_2, \dots, r_p = 0}^{\infty} \kappa_{r_1 r_2 \cdots r_p}^{12\cdots p} \frac{(\iota t_1)^{r_1} (\iota t_2)^{r_2} \cdots (\iota t_p)^{r_p}}{r_1! r_2! \cdots r_p!}.$$

For the covariance matrix kernel, we have

$$\Gamma_{(j,k),(m,l)} = \left(\kappa_{1111}^{jkml} + \sigma_{jm}\sigma_{kl} + \sigma_{jl}\sigma_{km}\right)/4, \tag{27}$$

where κ_{1111}^{jkml} is the joint fourth-order cumulants of F. Therefore, if $\kappa_{1111}^{jkjk} \ge 4C_1$, then $\Gamma_{(j,k),(j,k)} \ge C_1$.

If the data follow a distribution in the elliptic family ([52], Chapter 1), then the condition $\Gamma_{(j,k),(j,k)} \ge C_1$ is equivalent to $\min_{1 \le j \le p} \sigma_{jj} \ge C$ for some constant $C > 0$ depending only on C_1. To see this, for F in the elliptic family, it is known that $\kappa_{1111}^{jkml} = \kappa(\sigma_{jk}\sigma_{ml} + \sigma_{jm}\sigma_{kl} + \sigma_{jl}\sigma_{km})$, where κ is the kurtosis of F. Therefore, $\Gamma_{(j,k),(j,k)} = [(2\kappa+1)\sigma_{jk}^2 + (\kappa+1)\sigma_{jj}\sigma_{kk}]/4$ and $\Gamma_{(j,k),(j,k)} \ge C_1$ if and only if there exists a constant $C > 0$ such that $\sigma_{jj} \ge C$ for all $j = 1, \dots, p$.

There are a number of interesting features of Theorem 4.1. Consider $r = 0$; that is, Σ is truly sparse such that $\max_{1 \le m \le p} \sum_{k=1}^p \mathbf{1}\{\sigma_{mk} \ne 0\} \le \zeta_p$ for $\Sigma \in \mathcal{G}(0, \zeta_p)$. Then we can take $\beta = 1$ (i.e., $\tau_* = a_{\bar{T}_n^\sharp}(1-\alpha)$) and the convergence rates are

$$\|\hat{\Sigma}(\tau_*) - \Sigma\|_2 \le 6\zeta_p \tau_* \quad \text{and} \quad p^{-1}|\hat{\Sigma}(\tau_*) - \Sigma|_F^2 \le 18\zeta_p \tau_*^2.$$

Hence, the tuning parameter can be adaptively selected by bootstrap samples while the rate of convergence is *nearly optimal* in the following sense. Since the distribution of τ_* mimics that of $\tau_\diamond$, $\hat{\Sigma}(\tau_*)$ achieves the same convergence rate as the thresholded estimator $\hat{\Sigma}(\tau_\diamond)$ for the oracle choice of the threshold $\tau_\diamond$ with probability at least $1-\alpha-Cn^{-K/6}$. On the other hand, the bootstrap method is not fully equivalent to the oracle procedure in terms of the constants in the estimation error bounds. Suppose that we know the support Θ of Σ, that is, locations of the nonzero entries in Σ. Then the *oracle estimator* is simply $\breve{\Sigma} = \{\hat{s}_{mk}\mathbf{1}\{(m,k) \in \Theta\}\}_{m,k=1}^p$ and we have

$$\begin{aligned} \|\breve{\Sigma} - \Sigma\|_2 &\le \max_{1\le m\le p} \sum_{k=1}^p |\hat{s}_{mk} - \sigma_{mk}| \mathbf{1}\{(m,k) \in \Theta\} \\ &\le |\hat{S}_n - \Sigma|_\infty \max_{1\le m\le p} \sum_{k=1}^p \mathbf{1}\{(m,k)\in\Theta\} \le \tau_\diamond \zeta_p. \end{aligned}$$

Therefore, the constant of the convergence rate for the bootstrap method does not attain the oracle estimator. However, we shall comment that β is not a tuning parameter since it does not depend on F and the effect of β only appears in the constants in front of the convergence rates (24) and (25).

Assuming that the observations are sub-Gaussian(ν^2) and the variance factor ν^2 is a *fixed* constant, it is known that the threshold value $\tau_\Delta = C(\nu)\sqrt{\log(p)/n}$ achieves the minimax rate for estimating the sparse covariance matrix [13]. Compared with the minimax optimal tuning parameter τ_Δ, our bootstrap threshold τ_* exhibits several advantages for certain sub-Gaussian distributions, which we shall highlight (with stronger side conditions).

First, the bootstrap threshold τ_* does not need the knowledge of ν_n^2 and it allows $\nu_n^2 \to \infty$ as $n \to \infty$. In this case, from (26), the bootstrap threshold $\tau_* = O_{\mathbb{P}}(\xi_4^2\sqrt{\log(p)/n})$, where the constant of $O_{\mathbb{P}}(\cdot)$ depends only on $\alpha, \beta, C_1, \dots, C_4$ in Theorem 4.1, while the universal thresholding rule $\tau_\Delta = C'\nu_n^2\sqrt{\log(p)/n}$. Therefore, if $\xi_4 = o(\nu_n)$, then $\tau_* = o_{\mathbb{P}}(\tau_\Delta)$ and the bootstrap threshold τ_* is less conservative than the minimax threshold. For instance, suppose that $X_{im}, i = 1, \dots, n; m = 1, \dots, p$ have the same marginal distribution in Example 4.1 (continuous case) or Example 4.2 (discrete case). Then we have $\mathbb{E}[\tau_*] = O(\sqrt{\log(p)/n})$ by (26) and $\tau_\Delta = Ca_n^2\sqrt{\log(p)/n}$. Thus $\tau_* = o_{\mathbb{P}}(\tau_\Delta)$ for $a_n \to \infty$.

Second, τ_Δ is nonadaptive to the observations X_1^n since the minimax lower bound is based on the worst case analysis and the matching upper bound is obtained by the Bonferroni inequality, which ignores the dependence structures in F. On the contrary, τ_* takes into account the dependence information of F by conditioning on the observations. Therefore, the bootstrap threshold may better adjust to the dependence structure for some designs of X_i.

EXAMPLE 4.3 (A block diagonal covariance matrix example with reduced rank). Let L, m be two positive integers and $p = Lm$. Let $Z_{il}, i = 1, \dots, n; l = 1, \dots, L$, be i.i.d. mean zero sub-Gaussian(ν_n^2) random variables with unit variance and $Y_{il} = \mathbf{1}_m Z_{il}$, where $\mathbf{1}_m$ is the $m \times 1$ vector containing all ones. Let $X_i = (Y_{i1}^\top, \dots, Y_{iL}^\top)^\top$. Under the assumptions in Theorem 4.1, we can show that $\mathbb{E}[\tau_*] \le C'\beta^{-1}\xi_4^2(\log(L)/n)^{1/2}$. If $\log L = o(\log p)$, then $\tau_* = o_{\mathbb{P}}(\tau_\Delta)$ and $\hat{\Sigma}(\tau_*)$ can gain much tighter performance bounds in (24) and (25) than $\hat{\Sigma}(\tau_\Delta)$. Note that the covariance matrix $\Sigma = \mathrm{Cov}(X_i)$ in this example is block diagonal such that the diagonal blocks of Σ are rank-one matrices $\mathbf{1}_m \mathbf{1}_m^\top$. Therefore, Σ has the simultaneous sparsity (i.e., $\zeta_p = m$) and reduced rank (i.e., $\mathrm{rank}(\Sigma) = L$).

Third, as we shall demonstrate in Theorem 4.2, the Gaussian-type convergence rate of the bootstrap method in Theorem 4.1 can be achieved even for heavy-tailed data with polynomial moments.

THEOREM 4.2 (Data-driven threshold selection: uniform polynomial moment observations). *Let X_i be i.i.d. mean zero random vectors such that $\|\max_{1\le k\le p}|X_{1k}|\|_8 \le \nu_n$ and $\Sigma \in \mathcal{G}(r, \zeta_p)$. Suppose that there exist constants $C_i > 0, i = 1, \dots, 3$, such that $\Gamma_{(j,k),(j,k)} \ge C_1$, $\xi_6 \le C_2 \nu_n^{1/3}$ and $\xi_8 \le C_3 \nu_n^{1/2}$ for all $j, k = 1, \dots, p$. Let $\alpha, \beta, K \in (0, 1)$ and $\tau_* = \beta^{-1} a_{\bar{T}_n^\sharp}(1-\alpha)$. If $\nu_n^8 \log^7(np) \le C_4 n^{1-7K/6}$, then (24) and (25) hold with probability at least $1 - \alpha - Cn^{-K/6}$ for some constant $C > 0$ depending only on $C_1, \dots, C_4$. In addition, (26) holds for some constant $C' > 0$ depending only on α and $C_1, \dots, C_4$.*

We compare Theorem 4.2 with the threshold obtained by the union bound approach. Assume that $\max_{1\le k\le p}\mathbb{E}|X_{1k}|^q < \infty$ for $q \ge 8$. By the Nagaev inequality [53] applied to the split sample in Remark 5, one can show that

$$\tau_\sharp = C(q)\left\{\frac{p^{4/q}}{n^{1-2/q}}\xi_q^2 + \left(\frac{\log p}{n}\right)^{1/2}\xi_4^2\right\}$$

is the right threshold that gives a large probability bound for $\tau_\diamond = |\hat{S}_n - \Sigma|_\infty$. Consider $q = 8$, $\xi_8 = O(1)$, and the scaling limit $p = n^A$ for $A > 0$. Then the universal threshold $\tau_\sharp = o(1)$ if $0 < A < 3/2$. In contrast, since $\|\max_{1\le k\le p}|X_{1k}|\|_8 \le p^{1/8}\xi_8 = O(p^{1/8})$, it follows from Theorem 4.2 that the bootstrap threshold τ_* is asymptotically valid if $0 < A < 1$ and by (26), $\mathbb{E}[\tau_*] = O(\sqrt{(\log p)/n})$. Therefore, in the least favorable case for the bootstrap, we conclude that: (i) if $A \in (0, 1/2]$, then $\mathbb{E}[\tau_*] \asymp \tau_\sharp$; (ii) if $A \in (1/2, 1)$, then $\mathbb{E}[\tau_*] = o(\tau_\sharp)$ and $\tau_\sharp = o(1)$; (iii) if $A \in [1, 3/2)$, then $\tau_\sharp = o(1)$ while the bootstrap threshold τ_* is not asymptotically valid; (iv) if $A \in [3/2, \infty)$, then neither $\hat{\Sigma}(\tau_*)$ or $\hat{\Sigma}(\tau_\sharp)$ is consistent for estimating Σ. Hence, the bootstrap method gives better convergence rate than the universal thresholding method under the spectral and Frobenius norms when

$A \in (1/2, 1)$. On the other hand, since $\tau_\sharp = o(1)$ when $A \in (0, 3/2)$, the cost of the bootstrap to achieve the Gaussian-like convergence rate $\tau_* = O_{\mathbb{P}}(\sqrt{(\log p)/n})$ for the heavy-tailed distribution F is a stronger requirement on the scaling limit for $A \in (0, 1)$. Moreover, to the best of our knowledge, the minimax lower bound is currently not available to justify $\tau_\sharp$ for observations with polynomial moments. Finally, we remark that bootstrap can adapt to the dependency structure in F. For Example 4.3 with a block diagonal covariance matrix, we only need $L \log^7(nL) = o(n)$, where L can be much smaller than p and the dimension p may still be allowed to be larger or even much larger than the sample size n.

4.2. *Simultaneous inference for covariance and rank correlation matrices.* Another related important problem of estimating the sparse covariance matrix Σ is the consistent recovery of its support, that is, nonzero off-diagonal entries in Σ [43]. Towards this end, a lower bound of the minimum signal strength (Σ-min condition) is a necessary condition to separate the weak signals and true zeros, yet, the Σ-min condition is never verifiable. To avoid this undesirable condition, we can alternatively formulate the recovery problem as a more general hypothesis testing problem:

$$H_0 : \Sigma = \Sigma_0 \quad \text{versus} \quad H_1 : \Sigma \neq \Sigma_0, \tag{28}$$

where Σ_0 is a known $p \times p$ matrix. In particular, if $\Sigma_0 = \mathrm{Id}_{p\times p}$, then the support recovery can be restated as the following simultaneously testing problem: for all $m, k \in \{1, \dots, p\}$ and $m \neq k$,

$$H_{0,mk} : \sigma_{mk} = 0 \quad \text{versus} \quad H_{1,mk} : \sigma_{mk} \neq 0. \tag{29}$$

The test statistic we construct is $\bar{T}_0 = |\hat{S}_n - \Sigma_0|_{\infty,\mathrm{off}}$, which is an ℓ^∞ statistic by taking the maximum magnitudes on the off-diagonal entries. Then H_0 is rejected if $\bar{T}_0 \geq a_{\bar{T}_n^\sharp}(1-\alpha)$.

COROLLARY 4.3 (Asymptotic size of the simultaneous test: sub-Gaussian observations). *Let $\nu_n \geq 1$ and X_i be i.i.d. mean zero random vectors in $\mathbb{R}^p$ such that $X_{ik} \sim$ sub-Gaussian(ν_n^2) for all $k = 1, \dots, p$. Suppose that there exist constants $C_i > 0, i = 1, \dots, 3$, such that $\Gamma_{(j,k),(j,k)} \geq C_1$, $\xi_6 \leq C_2 \nu_n^{1/3}$ and $\xi_8 \leq C_3 \nu_n^{1/2}$ for all $j, k = 1, \dots, p$. Let $\alpha, \beta, K \in (0, 1)$ and $\tau_* = \beta^{-1} a_{\bar{T}_n^\sharp}(1-\alpha)$. If $\nu_n^4 \log^7(np) \leq C_4 n^{1-K}$, then the above test based on $\bar{T}_0$ for (28) has the size $\alpha + O(n^{-K/6})$; that is, the family-wise error rate of the simultaneous test problem (29) is asymptotically controlled at the level α.*

From Corollary 4.3, the test based on $\bar{T}_0$ is asymptotically exact of size α for sub-Gaussian data. A similar result can be established for observations with polynomial moments. Due to the space limit, details are omitted. [15] proposed a similar test statistic for comparing the two-sample large covariance matrices. Their results (Theorem 1 in [15]) are analogous to Corollary 4.3 in this paper in that no structural assumptions in Σ are needed in order to obtain the asymptotic validity of both tests. However, we shall note that their assumptions (C.1), (C.2) and (C.3) on the nondegeneracy are stronger than our condition $\Gamma_{(j,k),(j,k)} \geq C_1$. For sub-Gaussian observations $X_{ik} \sim \text{sub-Gaussian}(\nu_n^2)$, (C.3) in [15] assumed that $\min_{1\leq j\leq k\leq p} \gamma_{jk}/\nu_n^4 \geq c$ for some constant $c > 0$, where $\gamma_{jk} = \mathrm{Var}(X_{1j}X_{1k})$. If $\nu_n^2 \to \infty$, then [15], Theorem 1, requires that γ_{jk} for all $j, k = 1, \dots, p$ have to obey a uniform lower bound that diverges to infinity. For the covariance matrix kernel, since $g(x) = (xx^\top - \Sigma)/2$, we only need that $\min_{j,k} \gamma_{jk} \geq c$ for some fixed lower bound.

Next we comment that a distinguishing feature of our bootstrap test from the ℓ^2 test statistic [17] is that no structural assumptions are made on F and we allow for the strong dependence in Σ. For example, consider again the elliptic distributions ([52], Chapter 1) with the positive-definite $V = \varrho \mathbf{1}_p \mathbf{1}_p^\top + (1-\varrho)\mathrm{Id}_{p\times p}$ such that the covariance matrix Σ is proportion to V. Then we have

$$\operatorname{tr}(V^4) = p\big[1 + (p-1)\varrho^2\big]^2 + p(p-1)\big[2\varrho + (p-2)\varrho^2\big]^2,$$
$$\operatorname{tr}(V^2) = \varrho^2 p^2 + \big(1-\varrho^2\big)p.$$

For any $\varrho \in (0, 1)$, $\operatorname{tr}(V^4)/\operatorname{tr}^2(V^2) \to 1$ as $p \to \infty$. Therefore, the limiting distribution of the ℓ^2 test statistic in [17] is no longer normal and its asymptotic distribution remains unclear.

Finally, the covariance matrix testing problem (28) can be generalized further to nonparametric forms, which can gain more robustness to outliers and the nonlinearity in the dependency structures. Let $U_\diamond = \mathbb{E}[h(X_1, X_2)]$ be the expectation of the random matrix associated with h and U_0 be a known $p \times p$ matrix. Consider the testing problem

$$H_0 : U_\diamond = U_0 \quad \text{versus} \quad H_1 : U_\diamond \neq U_0.$$

Then the test statistic can be constructed as $\bar{T}_0 = |U_n - U_0|_\infty$ [or $\bar{T}_0' = |U_n - U_0|_{\infty,\mathrm{off}}$] and H_0 is rejected if $\bar{T}_0 \geq a_{\bar{T}_n^\sharp}(1-\alpha)$ [or $\bar{T}_0' \geq a_{\bar{T}_n^\sharp}(1-\alpha)$], where the bootstrap samples are generated w.r.t. the kernel h. The above test covers Kendall's tau rank correlation matrix as a special case where h is the bounded kernel.

COROLLARY 4.4 (Asymptotic size of the simultaneous test for Kendall's tau rank correlation matrix). *Let X_i be i.i.d. random vectors with a distribution F in $\mathbb{R}^p$. Suppose that there exists a constant $C_1 > 0$ such that $\Gamma_{(j,k),(j,k)} \geq C_1$ for all $j, k = 1, \dots, p$. Let $\alpha, \beta, K \in (0, 1)$ and $\tau_* = \beta^{-1} a_{\bar{T}_n^\sharp}(1-\alpha)$, where the bootstrap samples are generated with Kendall's tau rank correlation coefficient matrix kernel. If $\log^7(np) \leq C_2 n^{1-K}$, then the test based on $\bar{T}_0'$ has the size $\alpha + O(n^{-K/6})$.*

Therefore, the asymptotic validity of the bootstrap test for large Kendall's tau rank correlation matrix is obtained when $\log p = o(n^{1/7})$ without imposing structural and moment assumptions on F.

5. Proof of the main results. The rest of the paper is organized as follows. In Section 5.1, we first present a useful inequality for bounding the expectation of the sup-norm of the *canonical* U-statistics and then compare with an alternative simple data splitting bound by reducing to the moment bounding exercise for the sup-norm of sums of i.i.d. random vectors. We shall discuss several advantages of using the U-statistics approach by exploring the degeneracy structure. Section 5.2 contains the proof of the Gaussian approximation result and Section 5.3 proves the convergence rate of the bootstrap validity. Proofs of the statistical applications are given in Section 5.4. Additional proofs and technical lemmas are given in the SM [19].

5.1. *A maximal inequality for canonical U-statistics.* Before proving our main results, we first establish a maximal inequality of the canonical U-statistics of order two. The derived expectation bound is useful in controlling the size of the nonlinear and completely degenerate error term in the Gaussian approximation.

THEOREM 5.1 (A maximal inequality for canonical U-statistics). *Let X_1^n be a sample of i.i.d. random variables in a separable and measurable space $(S, \mathcal{S})$. Let $f : S \times S \to \mathbb{R}^d$ be an $\mathcal{S} \otimes \mathcal{S}$-measurable, symmetric and canonical kernel such that $\mathbb{E}|f_m(X_1, X_2)| < \infty$ for all $m = 1, \dots, d$. Let $V_n = [n(n-1)]^{-1} \sum_{1 \le i \ne j \le n} f(X_i, X_j)$, $M = \max_{1 \le i \ne j \le n} \max_{1 \le m \le d} |f_m(X_i, X_j)|$, $D_q = \max_{1 \le m \le d} (\mathbb{E}|f_m(X_1, X_2)|^q)^{1/q}$ for $q > 0$. If $2 \le d \le \exp(bn)$ for some constant $b > 0$, then there exists an absolute constant $K > 0$ such that*

$$\mathbb{E}\big[|V_n|_\infty\big] \le K\big(1 + b^{1/2}\big)\left\{\left(\frac{\log d}{n}\right)^{3/2} \|M\|_4 + \frac{\log d}{n} D_2 + \left(\frac{\log d}{n}\right)^{5/4} D_4\right\}. \tag{30}$$

Note that Theorem 5.1 is nonasymptotic. As immediate consequences of Theorem 5.1, we can derive the rate of convergence of $\mathbb{E}[|V_n|_\infty]$ with kernels under the subexponential and uniform polynomial moment conditions.

COROLLARY 5.2 (Kernels with subexponential and uniform polynomial moments). *Let B_n, B_n' be two sequences of positive reals and f be a symmetric and canonical kernel. Suppose that $2 \le d \le \exp(bn)$ for some constant $b > 0$:*

(i) *If*

$$\max_{1 \le m \le d} \mathbb{E}\big[\exp\big(|f_m|/B_n\big)\big] \le 2, \tag{31}$$

then there exists a constant $C(b) > 0$ such that

$$\mathbb{E}\big[|V|_\infty\big] \le C(b) B_n \big\{\big(n^{-1} \log d\big)^{3/2} \log(nd) + n^{-1} \log d\big\}. \tag{32}$$

(ii) *Let* $q \geq 4$. *If*

$$\mathbb{E}\Big(\max_{1\leq m\leq d}|f_m|/B_n\Big)^q \vee \max_{1\leq m\leq d}\mathbb{E}\big(|f_m|/B_n'\big)^4 \leq 1, \tag{33}$$

then there exists a constant $C(b) > 0$ *such that*

$$\mathbb{E}\big[|V|_\infty\big] \leq C(b)\big\{B_n n^{-3/2+2/q}(\log d)^{3/2} + B_n' n^{-1}\log d\big\}. \tag{34}$$

REMARK 5 (Comparison of Theorem 5.1 with sums of i.i.d. random vectors by data splitting). We can also bound the expected norm of a U-statistic by the expected norm of sums of i.i.d. random vectors. Assume that $\mathbb{E}|f_k(X_1, X_2)| < \infty$ for all $k = 1, \dots, d$ and let $m = [n/2]$ be the largest integer no greater than $n/2$. As noted in [36], we can write

$$m(V_n - \mathbb{E}V_n) = \frac{1}{n!}\sum_{\text{all }\pi_n} S(X_{\pi_n(1)}, \dots, X_{\pi_n(n)}), \tag{35}$$

where $S(X_1^n) = \sum_{i=1}^m [f(X_{2i-1}, X_{2i}) - \mathbb{E}f]$ and the summation $\sum_{\text{all }\pi_n}$ is taken over all possible permutations $\pi_n : \{1, \dots, n\} \to \{1, \dots, n\}$. By Jensen's inequality and the i.i.d. assumption of X_i, we have

$$\mathbb{E}|V_n - \mathbb{E}V_n|_\infty \leq \frac{1}{m}\mathbb{E}\left|\sum_{i=1}^m [f(X_i, X_{i+m}) - \mathbb{E}f]\right|_\infty, \tag{36}$$

which can be viewed as a data splitting method into two halves. Assuming (31), it follows from Bernstein's inequality [64], Proposition 5.16, that

$$\mathbb{E}|V_n - \mathbb{E}V_n|_\infty \leq K_1 B_n\big(\sqrt{\log(d)/n} + \log(d)/n\big) \tag{37}$$

for some absolute constant $K_1 > 0$. So if $\log d \leq bn^{1-\varepsilon}$ for some $\varepsilon \in (0, 1)$, then $\mathbb{E}|V_n - \mathbb{E}V_n|_\infty \leq C(b)B_n(\log(d)/n)^{1/2} \leq C(b)B_n n^{-\varepsilon/2}$. For the canonical kernel where $\mathbb{E}V_n = 0$, there are two advantages of using the U-statistics approach in Theorem 5.1 over the data splitting method into i.i.d. summands (36) and (37).

First, we can obtain from (32) that $\mathbb{E}|V_n|_\infty \leq C(b)B_n\{n^{1-5\varepsilon/2} + n^{-\varepsilon}\}$. Therefore, sharper rate is obtained by (32) when $\varepsilon \in (1/2, 1)$, which covers the regime of valid Gaussian approximation and bootstrap. Under the scaling limit for the Gaussian approximation validity, that is, $B_n^a \log^7(np)/n \leq Cn^{-K_2}$ for some $K_2 \in (0, 1)$, where $a = 2$ for the subexponential moment kernel and $a = 4$ for the uniform polynomial moment kernel, it is easy to see that $\log d \leq \log(nd) \leq Cn^{(1-K_2)/7}$ so we can take $\varepsilon = (6 + K_2)/7$.

Second and more importantly, the rate of convergence obtained by the Bernstein bound (37) does not lead to a convergence rate for the Gaussian and bootstrap approximations. The reason is that, although (37) is rate-exact for *nondegenerate* U-statistics, where the dependence of the rate in (37) on the sample size is $O(B_n n^{-1/2})$, it is not strong enough to control the size of the nonlinear remainder term $\mathbb{E}[|n^{1/2}V_n|_\infty]$ when $d \to \infty$ (recall that $R_n = n^{1/2}V_n/2$); cf.

Proposition 5.3. On the contrary, our bound in Theorem 5.1 exploits the degeneracy structure of V and the dependence of the rate in (30) on the sample size is $O(B_n n^{-1} + \|M\|_4 n^{-3/2})$. Therefore, Theorem 5.1 is mathematically more appealing in the degenerate case.

For nondegenerate U-statistics $U_n = [n(n-1)]^{-1} \sum_{1 \le i \neq j \le n} h(X_i, X_j)$, the reduction to sums of i.i.d. random vectors in (36) does not give tight asymptotic distributions. To illustrate this point, we consider the case $d = 1$ and let X_i be i.i.d. mean zero random variables with variance σ^2. Let $\zeta_1^2 = \mathrm{Var}(g(X_1))$ and $\zeta_2^2 = \mathrm{Var}(h(X_1, X_2))$. Assume that $\zeta_1^2 > 0$. So ζ_1^2 is the variance of the leading projection term used in the Gaussian approximation and by Jensen's inequality $\zeta_1^2 \le \zeta_2^2$. Note that $\sqrt{n}(U_n - \mathbb{E}U_n) \xrightarrow{D} N(0, 4\zeta_1^2)$ [60], Theorem A, page 192, and by the CLT $\sqrt{2/m} \sum_{i=1}^m [f(X_i, X_{i+m}) - \mathbb{E}f] \xrightarrow{D} N(0, 2\zeta_2^2)$. Since in general $\zeta_2^2 \neq 2\zeta_1^2$, the limiting distribution of the U-statistic is not the same as that in the data splitting method. For example, consider the nondegenerate covariance kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$. Denote $\mu_4 = \mathbb{E}X_1^4$ and $g(x_1) = (x_1^2 - \sigma^2)/2$. Then $\zeta_2^2 = (\mu_4 + \sigma^4)/2$ and $\zeta_1^2 = (\mu_4 - \sigma^4)/4$ so that $\zeta_2^2 > 2\zeta_1^2$ when $\sigma^2 > 0$. In particular, if X_i are i.i.d. $N(0, \sigma^2)$, then $\mu_4 = 3\sigma^4$, $4\zeta_1^2 = 2\sigma^4$, and $2\zeta_2^2 = 4\sigma^4$. Therefore, even though (37) gives better rate in the nondegenerate case, the reduction by splitting the data into the i.i.d. summands is not optimal for the Gaussian approximation purpose, which is the main motivation of this paper. In fact, ζ_2^2 serves no purpose in the limiting distribution of $\sqrt{n}(U_n - \mathbb{E}U_n)$.

5.2. *Proof of results in Section 2.* For $q > 0$ and $\phi \ge 1$, we define

$$D_{g,q} = \max_{1 \le m \le d} \mathbb{E}|g_m(X)|^q,$$

$$M_{g,q}(\phi) = \mathbb{E}\Big[\max_{1 \le m \le d} |g_m(X)|^q \mathbf{1}\Big(\max_{1 \le m \le d} |g_m(X)| > \frac{\sqrt{n}}{4\phi \log d}\Big)\Big],$$

$$M_{Y,q}(\phi) = \mathbb{E}\Big[\max_{1 \le m \le d} |Y_m|^q \mathbf{1}\Big(\max_{1 \le m \le d} |Y_m| > \frac{\sqrt{n}}{4\phi \log d}\Big)\Big]$$

and $M_q(\phi) = M_{g,q}(\phi) + M_{Y,q}(\phi)$. The Gaussian approximation result (GAR) in Proposition 5.3 below relies on the control of $D_{g,3}$ and $M_3(\phi)$. Interestingly, the quantity $M_{g,3}(\phi)$ can be viewed as a stronger version of the Lindeberg condition that allows us to derive the explicit convergence rate of the Gaussian approximation when $d \to \infty$. Denote $\chi_{\tau,ij} = \mathbf{1}(\max_{1 \le m \le d} |h_m(X_i, X_j)| > \tau)$ for $\tau \ge 0$. Let

$$D_q = \max_{1 \le m \le d} \mathbb{E}|h_m(X_1, X_2)|^q,$$

$$M_{h,q}(\tau) = \mathbb{E}\Big[\max_{1 \le i \neq j \le n} \max_{1 \le m \le d} |h_m(X_i, X_j)|^q \chi_{\tau,ij}\Big].$$

For two random vectors X and Y in $\mathbb{R}^d$, we denote

$$\tilde{\rho}^{\mathrm{re}}(X, Y) = \sup_{y\in\mathbb{R}^d} \left|\mathbb{P}(X \le y) - \mathbb{P}(Y \le y)\right|.$$

PROPOSITION 5.3 (A general Gaussian approximation result for U-statistics). *Assume that* (M.1) *holds and* $\log d \le \bar{b}n$ *for some constant* $\bar{b} > 0$. *Then there exist constants* $C_i := C_i(\underline{b}, \bar{b}) > 0, i = 1, 2$ *such that for any real sequence* $\bar{D}_{g,3}$ *satisfying* $D_{g,3} \le \bar{D}_{g,3}$, *we have*

$$\begin{aligned} &\tilde{\rho}^{\mathrm{re}}(T_n, Y) \\ &\quad\le C_1\left\{\left(\frac{\bar{D}_{g,3}^2 \log^7 d}{n}\right)^{1/6} + \frac{M_3(\phi_n)}{\bar{D}_{g,3}}\right. \\ &\qquad\left. + \phi_n\left(\frac{\log^{3/2} d}{n}\left(M_{h,4}(\tau)^{1/4} + \tau\right) + \frac{\log d}{n^{1/2}} D_2^{1/2} + \frac{\log^{5/4} d}{n^{3/4}} D_4^{1/4}\right)\right\}, \end{aligned} \tag{38}$$

where

$$\phi_n = C_2\left(\frac{\bar{D}_{g,3}^2 \log^4 d}{n}\right)^{-1/6}. \tag{39}$$

In addition, $\rho^{\mathrm{re}}(T_n, Y)$ *obeys the same bound in* (38).

With the help of Proposition 5.3, we are now ready to prove Theorem 2.1.

PROOF OF THEOREM 2.1. We may assume that $\varpi_{1,n} \le 1$; otherwise the proof is trivial. Let $\ell_n = \log(nd) > 1$. By (M.2) and Jensen's inequality, we have $D_2 \le B_n^{2/3}$, $D_{g,3} \le D_3 \le B_n$, and $D_4 \le B_n^2$. Write $M_{h,q} = M_{h,q}(0)$. By Proposition 5.3 with $\tau = 0$ and ϕ_n is given by (39), we have

$$\begin{aligned} \rho^{\mathrm{re}}(T_n, Y) &\le C_1\left\{\left(\frac{\bar{D}_{g,3}^2 \log^7 d}{n}\right)^{1/6} + \frac{M_3(\phi_n)}{\bar{D}_{g,3}}\right. \\ &\qquad\left. + \phi_n\left(\frac{\log^{3/2} d}{n} M_{h,4}^{1/4} + \frac{\log d}{n^{1/2}} D_2^{1/2} + \frac{\log^{5/4} d}{n^{3/4}} D_4^{1/4}\right)\right\}, \end{aligned} \tag{40}$$

where $C_1 > 0$ is a constant only depending on $\underline{b}$ and $\bar{b}$.

Case (E.1). By [63], Lemma 2.2.2, $M_{h,4}^{1/4} \le K_1 B_n \ell_n$. Choosing $\bar{D}_{g,3} = B_n$, we have

$$\begin{aligned} \phi_n \frac{\log^{3/2} d}{n} M_{h,4}^{1/4} &\le C_2 \frac{B_n^{2/3} \ell_n^{11/6}}{n^{5/6}} \le C_2 \varpi_{1,n}, \\ \phi_n \frac{\log d}{n^{1/2}} D_2^{1/2} &\le C_3 \frac{(\log d)^{1/3}}{n^{1/3}} \le C_3 \varpi_{1,n}, \\ \phi_n \frac{\log^{5/4} d}{n^{3/4}} D_4^{1/4} &\le C_4 \frac{B_n^{1/6} (\log d)^{7/12}}{n^{7/12}} \le C_4 \varpi_{1,n}. \end{aligned}$$

Following the proof of [24], Proposition 2.1, we can show that

$$\left(\frac{\bar{D}_{g,3}^2 \log^7 d}{n}\right)^{1/6} + \frac{M_3(\phi_n)}{\bar{D}_{g,3}} \le C_5 \varpi_{1,n}.$$

Then, (5) follows from (40). Here, all constants C_i for $i = 2, \ldots, 5$ depend only on $\underline{b}$ and $\bar{b}$.

Case (E.2). D_2 and D_4 obey the same bounds as case (E.1). Assuming (E.2), $M_{h,4}^{1/4} \le n^{1/2} B_n$. Choosing $\bar{D}_{g,3} = B_n + B_n^2 n^{-1/2+2/q} (\log d)^{-1/2}$, we have

$$\phi_n \frac{\log^{3/2} d}{n} M_{h,4}^{1/4} \le C_6 \frac{B_n^{2/3} \ell_n^{5/6}}{n^{1/3}} \le C_6 \varpi_{1,n}.$$

Following the proof of [24], Proposition 2.1, we can show that

$$\left(\frac{\bar{D}_{g,3}^2 \log^7 d}{n}\right)^{1/6} + \frac{M_3(\phi_n)}{\bar{D}_{g,3}} \le C_7 \{\varpi_{1,n} + \varpi_{2,n}\}.$$

Here, C_6, C_7 are constants depending only on $\underline{b}, \bar{b}$, and q. Then (5) is immediate. □

5.3. *Proof of results in Section* 3. In view of the approximation diagram (9), our first task is to control the random quantity

$$\sup_{A \in \mathcal{A}^{\mathrm{re}}} \left| \mathbb{P}(Y \in A) - \mathbb{P}(Z^X \in A \mid X_1^n) \right|$$

on an event occurring with large probability, which is Step (2) in the approximation diagram (9).

PROPOSITION 5.4 (Gaussian comparison bound for the linear part in U-statistic and its EB version). *Let* $Z^X | X_1^n \sim N(0, \hat{\Gamma}_n)$, *where* $\hat{\Gamma}_n$ *is defined in* (11). *Suppose that* (M.1), (M.2) *and* (M.2′) *are satisfied.*

(i) *If* (E.1) *and* (E.1′) *hold, then there exists a constant* $C(\underline{b}) > 0$ *such that with probability at least* $1 - \gamma$ *we have*

$$\sup_{A \in \mathcal{A}^{\mathrm{re}}} \left| \mathbb{P}(Y \in A) - \mathbb{P}(Z^X \in A \mid X_1^n) \right| \le C(\underline{b}) \varpi_{1,n}^B(\gamma). \tag{41}$$

(ii) *If* (E.2) *and* (E.2′) *hold with* $q \ge 4$, *then there exists a constant* $C(\underline{b}, q) > 0$ *such that with probability at least* $1 - \gamma$ *we have*

$$\sup_{A \in \mathcal{A}^{\mathrm{re}}} \left| \mathbb{P}(Y \in A) - \mathbb{P}(Z^X \in A \mid X_1^n) \right| \le C(\underline{b}, q) \{ \varpi_{1,n}^B(\gamma) + \varpi_{2,n}^B(\gamma) \}. \tag{42}$$

From Proposition 5.4, we are now ready to establish the rate of convergence of the empirical bootstrap for U-statistics. Let $W_{jk} = |n^{-1} \sum_{i=1}^n h_j(X_k, X_i) - V_{nj}|$

for $j = 1, \dots, d$ and $k = 1, \dots, n$. For $q, \tau > 0$, and $\phi \geq 1$, we define

$$\hat{D}_{g,q} = \max_{1 \leq j \leq d} n^{-1} \sum_{k=1}^{n} W_{jk}^{q},$$

$$\hat{D}_{q} = \max_{1 \leq j \leq d} n^{-2} \sum_{k,l=1}^{n} |h_j(X_k, X_l)|^q,$$

$$\hat{M}_{h,q}(\tau) = n^{-2} \sum_{i,k=1}^{n} \max_{1 \leq j \leq d} |h_j(X_i, X_k)|^q \mathbf{1}\Big(\max_{1 \leq j \leq d} |h_j(X_i, X_k)| > \tau\Big),$$

$$\hat{M}_{g,q}(\phi) = n^{-1} \sum_{k=1}^{n} \max_{1 \leq j \leq d} W_{jk}^{q} \mathbf{1}\Big(\max_{1 \leq j \leq d} W_{jk} > \frac{\sqrt{n}}{4\phi \log d}\Big),$$

$$\hat{M}_{Z,q}(\phi) = \mathbb{E}\Big[\max_{1 \leq j \leq d} |Z_j^X|^q \mathbf{1}\Big(\max_{1 \leq j \leq d} |Z_j^X| > \frac{\sqrt{n}}{4\phi \log d}\Big) \Bigm| X_1^n\Big],$$

$\hat{M}_q(\phi) = \hat{M}_{g,q}(\phi) + \hat{M}_{Z,q}(\phi)$, and $Z^X \mid X_1^n \sim N(0, \hat{\Gamma}_n)$.

PROOF OF THEOREM 3.1. In this proof the constants $C_1, C_2, \dots$ depend only on $\underline{b}, \bar{b}, K$ in case (i) and $\underline{b}, \bar{b}, q, K$ in case (ii). First, we may assume that

$$n^{-1} B_n^2 \log^7(nd) \leq c_1 \leq 1 \tag{43}$$

for some sufficiently small constant $c_1 > 0$, where c_1 depends only on $\underline{b}, \bar{b}, K$ in case (i) and on $\underline{b}, \bar{b}, q, K$ in case (ii), since otherwise the proof is trivial by setting the constants $C(\underline{b}, \bar{b}, K)$ in (i) and $C(\underline{b}, \bar{b}, q, K)$ in (ii) large enough. By (9) and the triangle inequality,

$$\begin{aligned} \rho^B(T_n, T_n^*) \leq {} & \rho^{\mathrm{re}}(T_n, Y) + \sup_{A \in \mathcal{A}^{\mathrm{re}}} \big|\mathbb{P}(Y \in A) - \mathbb{P}(Z^X \in A \mid X_1^n)\big| \\ & + \rho^{\mathrm{re}}(Z^X, T_n^* \mid X_1^n), \end{aligned} \tag{44}$$

where $\rho^{\mathrm{re}}(Z^X, T_n^* \mid X_1^n) = \sup_{A \in \mathcal{A}^{\mathrm{re}}} |\mathbb{P}(Z^X \in A \mid X_1^n) - \mathbb{P}(T_n^* \in A \mid X_1^n)|$. Since $\log(1/\gamma) \leq K \log(dn)$, we have $\varpi_{1,n}^B(\gamma) \leq K^{1/3} \varpi_{1,n}$ and $\varpi_{2,n} \leq \varpi_{2,n}^B(\gamma)$ for $\gamma \in (0, e^{-1})$. By Theorem 2.1 and Proposition 5.4, we have: (i) if (E.1) and (E.1′) hold, then with probability at least $1 - 2\gamma/9$ we have

$$\rho^{\mathrm{re}}(T_n, Y) + \sup_{A \in \mathcal{A}^{\mathrm{re}}} \big|\mathbb{P}(Y \in A) - \mathbb{P}(Z^X \in A \mid X_1^n)\big| \leq C(\underline{b}, \bar{b}, K) \varpi_{1,n};$$

(ii) if (E.2) and (E.2′) hold, then with probability at least $1 - 2\gamma/9$ we have

$$\begin{aligned} & \rho^{\mathrm{re}}(T_n, Y) + \sup_{A \in \mathcal{A}^{\mathrm{re}}} \big|\mathbb{P}(Y \in A) - \mathbb{P}(Z^X \in A \mid X_1^n)\big| \\ & \leq C(\underline{b}, \bar{b}, q, K)\{\varpi_{1,n} + \varpi_{2,n}^B(\gamma)\}. \end{aligned}$$

To deal with the third term on the right-hand side of (44), we observe that conditionally on X_1^n, U_n^* is a U-statistics of $\xi_1, \dots, \xi_n$ and Z^X has the conditional covariance matrix $\hat{\Gamma}_n$; cf. (11). So we can apply Proposition 5.3 conditionally.

Case (i). As in the proof of Proposition 5.4, we have with probability at least $1-\gamma/9$:

$$|\hat{\Gamma}_n - \Gamma|_\infty \le C_1\bigl[n^{-1}B_n^2 \log(nd)\log^2(1/\gamma)\bigr]^{1/2}. \tag{45}$$

By (43), (M.1), (M.2) and (M.2′), there exists a constant $C_2 > 0$ such that $\underline{b}/2 \le \hat{\Gamma}_{n,jj} \le C_2 B_n^{2/3} \le C_2 B_n$ for all $j = 1, \dots, d$ holds with probability at least $1-\gamma/9$. Let $\bar{D}_{g,3} = C_3 B_n$, $\bar{D}_2 = C_4 B_n^{2/3}$ and $\bar{D}_4 = C_5 B_n^2 \log(dn)$. By Lemma C.2, each of the three events $\{\hat{D}_{g,3} \ge \bar{D}_{g,3}\}$, $\{\hat{D}_2 \ge \bar{D}_2\}$ and $\{\hat{D}_4 \ge \bar{D}_4\}$ occur with probability at most $\gamma/9$. Let $\phi_n = C_6(n^{-1}\bar{D}_{g,3}^2 \log^4 d)^{-1/6}$ for some $C_6 > 0$ such that $\phi_n \ge 1$. By Jensen's inequality, $\max_{k,j} W_{kj} \le 2n^{-1}\max_{k,j}\sum_{i=1}^n |h_j(X_k, X_i)|$. Then, by the union bound and the assumptions (E.1) and (E.1′), we have

$$\begin{aligned}
\mathbb{P}\bigl(\hat{M}_{g,3}(\phi_n) > 0\bigr) &= \mathbb{P}\Bigl(\max_{1\le j\le d, 1\le k\le n} W_{jk} > \sqrt{n}/(4\phi_n \log d)\Bigr) \\
&\le (dn)\max_{1\le j\le d, 1\le k\le n}\mathbb{P}\Biggl(\frac{1}{n}\sum_{i=1}^n |h_j(X_k, X_i)| > \frac{\sqrt{n}}{8\phi_n \log d}\Biggr) \qquad (46)\\
&\le (2dn)\exp\biggl(-\frac{\sqrt{n}}{8\phi_n(\log d)B_n}\biggr),
\end{aligned}$$

where the last step (46) follows from the triangle inequality on the Orlicz space with the ψ_1 norm $\|n^{-1}\sum_{i=1}^n |h_j(X_k, X_i)|\|_{\psi_1} \le n^{-1}\sum_{i=1}^n \|h_j(X_k, X_i)\|_{\psi_1} \le B_n$. Substituting the value of ϕ_n and using (43), we have

$$\frac{\sqrt{n}}{8\phi_n(\log d)B_n} \ge \frac{C_3^{1/3}\log(nd)}{8C_6 c_1^{1/3}} \ge \frac{C_3^{1/3}}{16C_6 c_1^{1/3}}\biggl[\log(nd) + \frac{1}{K}\log(1/\gamma)\biggr]. \tag{47}$$

Therefore, $\mathbb{P}(\hat{M}_{g,3}(\phi_n) > 0) \le \gamma/9$ by choosing $c_1 > 0$ small enough. Next, we deal with $\hat{M}_{Z,3}(\phi_n)$. Since conditional on X_1^n, $Z^X \sim N(0, \hat{\Gamma}_n)$. On the event $\{\hat{\Gamma}_{n,jj} \le C_2 B_n, \forall j = 1, \dots, d\}$, we have $\|Z_j^X\|_{\psi_2} \le \sqrt{8C_2B_n/3}B_n^{1/2}$ and $\|Z_j^X\|_{\psi_1} \le C_7 B_n^{1/2}$ for all $j = 1, \dots, d$, where $C_7 = \sqrt{8C_2/(3\log 2)}$. Integration-by-parts yields

$$\hat{M}_{Z,3}(\phi_n) = \int_t^\infty \mathbb{P}\Bigl(\max_j |Z_j^X| > u^{1/3} \mid X_1^n\Bigr)\,du + t\mathbb{P}\Bigl(\max_j |Z_j^X| > t^{1/3} \mid X_1^n\Bigr),$$

where $t = (\sqrt{n}/4\phi_n \log d)^3$. Since for any $u > 0$

$$\mathbb{P}\Bigl(\max_j |Z_j^X| > u^{1/3} \mid X_1^n\Bigr) \le (2d)\exp\bigl(-u^{1/3}/(C_7 B_n^{1/2})\bigr),$$

we have by elementary calculations that

$$\int_t^\infty \mathbb{P}\Big(\max_j |Z_j^X| > u^{1/3} \mid X_1^n\Big)\, du \le C_8\, dt \Bigg[\sum_{\ell=1}^3 (B_n^{1/2} t^{-1/3})^\ell\Bigg] \exp\Big(-\frac{t^{1/3}}{C_7 B_n^{1/2}}\Big).$$

Since $B_n^{1/2} t^{-1/3} \log(nd) \le 4C_6 C_3^{-1/3} [n^{-1} B_n^2 \log^4(nd)]^{1/3} \le 4C_6 C_3^{-1/3} c_1^{1/3}$, it follows from (46) and (47) that

$$\begin{aligned}\hat{M}_{Z,3}(\phi_n) &\le C_9 dt \exp\Big(-\frac{t^{1/3}}{C_7 B_n^{1/2}}\Big) \le C_9 d n^{3/2} \exp\Big(-\frac{\log(nd)}{4C_7 C_6 C_3^{-1/3} c_1^{1/3}}\Big)\\ &\le C_9 n^{-1/2}\end{aligned}$$

for $c_1 > 0$ small enough. For the term $\hat{M}_{h,4}(\tau)$, we note that

$$\mathbb{P}\big(\hat{M}_{h,4}(\tau) > 0\big) = \mathbb{P}\Big(\max_{1\le i,k\le n} \max_{1\le j\le d} \big|h_j(X_i, X_k)\big| > \tau\Big)$$

and by (E.1) and (E.1′) $\|h_j(X_j, X_k)\|_{\psi_1} \le B_n$. So we have

$$\mathbb{P}\big(\hat{M}_{h,4}(\tau) > 0\big) \le \big(2dn^2\big) \exp(-\tau/B_n).$$

Choose $\tau = C_{10} n^{1/2}/[\phi_n \log d]$. Then, by (46) and (47), we have $\mathbb{P}(\hat{M}_{h,4}(\tau) > 0) \le \gamma/9$. Now, by Proposition 5.3 conditional on X_1^n with $M_{h,4}(\tau) \le n^2 \mathbb{E}[\max_{1\le j\le d} |h_j(X, X')|^4 \mathbf{1}(\max_{1\le j\le d} |h_j(X, X')| > \tau)]$, we conclude that

$$\begin{aligned}\rho^{\mathrm{re}}\big(Z^X, T_n^* \mid X_1^n\big) &\le C_{11}\bigg\{\Big(\frac{B_n^2 \log^7 d}{n}\Big)^{1/6} + \frac{1}{n^{1/2} B_n} + \phi_n \frac{\log d}{n^{1/2}} B_n^{1/3}\\ &\qquad + \phi_n \frac{\log^{5/4} d}{n^{3/4}} B_n^{1/2} \log^{1/4}(dn) + \frac{\log^{1/2} d}{n^{1/2}}\bigg\}\\ &\le C_{12} \varpi_{1,n}\end{aligned}$$

holds with probability at least $1 - 7\gamma/9$. So, (13) follows.

Case (ii). In addition to (43), we may assume that

$$\frac{B_n^2 \log^3(nd)}{\gamma^{2/q} n^{1-2/q}} \le c_2 \le 1 \tag{48}$$

for some small enough constant $c_2 > 0$. As in Case (i), there exists a constant $C_1 > 0$ such that $\underline{b}/2 \le \hat{\Gamma}_{n,jj} \le C_1 B_n$ for all $j = 1, \ldots, d$ holds with probability at least $1 - \gamma/9$. Let $\bar{D}_{g,3} = C_2[B_n + n^{-1+3/q} B_n^3 \gamma^{-3/q} (\log d)]$, $\bar{D}_2 = C_3[B_n^{2/3} + n^{-1+2/q} B_n^2 \gamma^{-2/q} (\log d)]$, and $\bar{D}_4 = C_4[B_n^2 + n^{-1+4/q} B_n^2 \gamma^{-4/q} (\log d)]$. By Lemma C.2, each of the three events $\{\hat{D}_{g,3} \ge \bar{D}_{g,3}\}$, $\{\hat{D}_2 \ge \bar{D}_2\}$, and $\{\hat{D}_4 \ge \bar{D}_4\}$ occur with probability at most $\gamma/9$. Note that

$$\begin{aligned}\phi_n &:= C_5\big(n^{-1} \bar{D}_{g,3}^2 \log^4 d\big)^{-1/6}\\ &\le C_5 C_2^{-1/3} \min\{n^{1/6} B_n^{-1/3} \log^{-2/3} d, n^{1/2-1/q} B_n^{-1} \gamma^{1/q} \log^{-1} d\}.\end{aligned} \tag{49}$$

By (46), the union bound, (49) and choosing C_2 large enough, we have

$$\begin{aligned}\mathbb{P}\big(\hat{M}_{g,3}(\phi_n) > 0\big) &\le n \max_{1\le k\le n} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \max_j \big|h_j(X_k, X_i)\big| > \frac{\sqrt{n}}{8\phi_n \log d}\right) \\ &\le \frac{(8B_n\phi_n \log d)^q}{n^{q/2-1}} \le \frac{\gamma}{9},\end{aligned}$$

where the second last step follows from (E.2), (E.2′) and the triangle inequality $\|n^{-1}\sum_{i=1}^{n} \max_j |h_j(X_k, X_i)|\|_q \le n^{-1}\sum_{i=1}^{n} \|\max_j |h_j(X_k, X_i)|\|_q \le B_n$. Bound on the term $\hat{M}_{Z,3}(\phi_n)$ is the same as in Case (i). Choose $\tau = C_6 n^{1/2+1/q}/[\phi_n \log d]$ for some $C_6 > 0$. Then we have

$$\mathbb{P}\big(\hat{M}_{h,4}(\tau) > 0\big) \le C_6^{-q} n^2 \frac{(B_n\phi_n \log d)^q}{n^{q/2+1}} \le \frac{\gamma}{9}.$$

Then we have by elementary calculations that

$$\begin{aligned}\rho^{\mathrm{re}}\big(Z^X, T_n^* \mid X_1^n\big) &\le C_7\left\{\left(\frac{\bar{D}_{g,3}^2 \log^7 d}{n}\right)^{1/6} + \frac{1}{n^{1/2}B_n} + \phi_n \frac{\log d}{n^{1/2}} \bar{D}_2^{1/2}\right. \\ &\qquad \left. + \phi_n \frac{\log^{5/4} d}{n^{3/4}} \bar{D}_4^{1/4} + \frac{\log^{1/2} d}{n^{1/2-1/q}}\right\} \\ &\le C_8\big\{\varpi_{1,n} + \varpi_{2,n}^{B}(\gamma)\big\}\end{aligned}$$

with probability at least $1 - 7\gamma/9$. The proof is now complete. □

PROOF OF COROLLARY 3.2. Let $\gamma_n = [n \log^2(n)]^{-1}$. Then $\sum_{n=4}^{\infty} \gamma_n \le \int_3^{\infty} [x \log^2(x)]^{-1}\,dx = \log^{-1}(3) < \infty$. Applying Theorem 3.1 with $\gamma = \gamma_n$ and by the Borel–Cantelli lemma, we have $\mathbb{P}(\rho^B(T_n, T_n^*) > C\varpi_{1,n} \text{ i.o.}) = 0$ for part (i) and $\mathbb{P}(\rho^B(T_n, T_n^*) > C\{\varpi_{1,n} + \varpi'^{B}_{2,n}(\gamma)\} \text{ i.o.}) = 0$ for part (ii), from which the corollary follows. □

The proof of the validity of the randomly reweighted bootstrap with i.i.d. Gaussian weights (Section 3.2) and Gaussian multiplier bootstrap with the jackknife covariance matrix estimator (Section 3.3) can be found in Section C in the SM [19].

5.4. *Proof of results in Section 4.*

PROOF OF THEOREM 4.1. Let $\tau_\diamond = \beta^{-1}|\hat{S}_n - \Sigma|_\infty$. By the sub-Gaussian assumption and Lemma F.1, it is easy to verify that there is a large enough constant $C > 0$ depending only on C_2, C_3 such that

$$\max_{\ell=1,2} \mathbb{E}\big[|h_{mk}|^{2+\ell}/\big(C\nu_n^{2\ell}\big)\big] \vee \mathbb{E}\big[\exp\big(|h_{mk}|/\nu_n^2\big)\big] \le 2, \tag{50}$$

where h is the covariance matrix kernel. Since $\Gamma_{(j,k),(j,k)} \geq C_1$ for all $j, k = 1, \dots, p$ and $\nu_n^4 \log^7(np) \leq C_4 n^{1-K}$, we have by Theorem 3.6 that $|\hat{S}_n - \Sigma|_\infty \leq a_{\bar{T}_n^\sharp}(1-\alpha)$ with probability at least $1 - \alpha - Cn^{-K/6}$, where $C > 0$ is a constant depending only on $C_i, i = 1, \dots, 4$. Therefore, $\mathbb{P}(\tau_\diamond \leq \tau_*) \geq 1 - \alpha - Cn^{-K/6}$ and the rest of the proof is restricted to the event $\{\tau_\diamond \leq \tau_*\}$. By the decomposition,

$$\begin{aligned}\|\hat{\Sigma}(\tau_*) - \Sigma\|_2 &\leq \|\hat{\Sigma}(\tau_*) - T_{\tau_*}(\Sigma)\|_2 + \|T_{\tau_*}(\Sigma) - \Sigma\|_2 \\ &\leq I + II + III + \tau_*^{1-r}\zeta_p,\end{aligned}$$

where $T_\tau(\Sigma) = \{\sigma_{mk}\mathbf{1}\{|\sigma_{mk}| > \tau\}\}_{m,k=1}^p$ is the resulting matrix of the thresholding operator on Σ at the level τ and

$$I = \max_m \sum_k |\hat{s}_{mk}|\mathbf{1}\{|\hat{s}_{mk}| > \tau_*, |\sigma_{mk}| \leq \tau_*\},$$

$$II = \max_m \sum_k |\sigma_{mk}|\mathbf{1}\{|\hat{s}_{mk}| \leq \tau_*, |\sigma_{mk}| > \tau_*\},$$

$$III = \max_m \sum_k |\hat{s}_{mk} - \sigma_{mk}|\mathbf{1}\{|\hat{s}_{mk}| > \tau_*, |\sigma_{mk}| > \tau_*\}.$$

Note that on the event $\{\tau_\diamond \leq \tau_*\}$, $\max_{m,k} |\hat{s}_{mk} - \sigma_{mk}| \leq \beta\tau_*$. Since $\Sigma \in \mathcal{G}(r, \zeta_p)$, we can bound

$$III \leq (\beta\tau_*)(\tau_*^{-r}\zeta_p) = \beta\tau_*^{1-r}\zeta_p.$$

By the triangle inequality,

$$\begin{aligned}II &\leq \max_m \sum_k |\hat{s}_{mk} - \sigma_{mk}|\mathbf{1}\{|\sigma_{mk}| > \tau_*\} + \max_m \sum_k |\hat{s}_{mk}|\mathbf{1}\{|\hat{s}_{mk}| \leq \tau_*, |\sigma_{mk}| > \tau_*\} \\ &\leq (\beta\tau_*)(\tau_*^{-r}\zeta_p) + \tau_*(\tau_*^{-r}\zeta_p) = (1+\beta)\tau_*^{1-r}\zeta_p.\end{aligned}$$

Let $\eta \in (0, 1)$. We have $I \leq IV + V + VI$, where

$$IV = \max_m \sum_k |\sigma_{mk}|\mathbf{1}\{|\hat{s}_{mk}| > \tau_*, |\sigma_{mk}| \leq \tau_*\},$$

$$V = \max_m \sum_k |\hat{s}_{mk} - \sigma_{mk}|\mathbf{1}\{|\hat{s}_{mk}| > \tau_*, |\sigma_{mk}| \leq \eta\tau_*\},$$

$$VI = \max_m \sum_k |\hat{s}_{mk} - \sigma_{mk}|\mathbf{1}\{|\hat{s}_{mk}| > \tau_*, \eta\tau_* < |\sigma_{mk}| \leq \tau_*\}.$$

Clearly, $IV \leq \tau_*^{1-r}\zeta_p$. On the indicator event of V, we observe that

$$\beta\tau_* \geq |\hat{s}_{mk} - \sigma_{mk}| \geq |\hat{s}_{mk}| - |\sigma_{mk}| > (1-\eta)\tau_*.$$

Therefore, $V = 0$ if $\eta + \beta \leq 1$. For VI, we have

$$VI \leq (\beta\tau_*)(\eta\tau_*)^{-r}\zeta_p.$$

Collecting all terms, we conclude that

$$\|\hat{\Sigma}(\tau_*) - \Sigma\|_2 \le (3 + 2\beta + \eta^{-r}\beta)\zeta_p \tau_*^{1-r} + V.$$

Then (24) follows from the choice $\eta = 1 - \beta$. The Frobenius norm rate (25) can be established similarly. Details are omitted.

Next, we prove (26). Let $\hat{g}_i = (n-1)^{-1}\sum_{j\neq i} h(X_i, X_j) - U_n$ and denote $\Phi(\cdot)$ as the c.d.f. of the standard Gaussian random variable. By the union bound, we have for all $t > 0$

$$\mathbb{P}_e\left(\frac{2}{\sqrt{n}}\left|\sum_{i=1}^n \hat{g}_i e_i\right|_\infty \ge t\right) \le 2p^2\left[1 - \Phi\left(\frac{t}{\bar{\psi}}\right)\right],$$

where $\bar{\psi} = \max_{1\le m,k\le p} |\psi_{mk}|$ and $\psi_{mk}^2 = 4n^{-1}\sum_{i=1}^n \hat{g}_{i,mk}^2$. Let $\tilde{\tau} = n^{-1/2}\beta^{-1}\bar{\psi}\Phi^{-1}(1-\alpha/(2p^2))$; then $\tau_* \le \tilde{\tau}$. Since $\Phi^{-1}(1-\alpha/(2p^2)) \asymp (\log p)^{1/2}$, we have $\mathbb{E}[\tau_*] \le C'\beta^{-1}\mathbb{E}[\bar{\psi}](\log(p)/n)^{1/2}$, where $C' > 0$ is a constant only depending on α. Now we bound $\mathbb{E}[\bar{\psi}]$. By Jensen's inequality,

$$\psi_{mk}^2 \le \frac{16}{n(n-1)} \sum_{1\le i\neq j\le n} h_{mk}^2(X_i, X_j).$$

Let $\ell = [n/2]$. By the data splitting argument in (36), [23], Lemma 9, and Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left\{\max_{m,k} \frac{1}{n(n-1)}\left|\sum_{1\le i\neq j\le n} [h_{mk}^2(X_i, X_j) - \mathbb{E}h_{mk}^2]\right|\right\} \\
&\quad \le \frac{1}{\ell}\mathbb{E}\left\{\max_{m,k}\left|\sum_{i=1}^{\ell}[h_{mk}^2(X_i, X_{i+\ell}) - \mathbb{E}h_{mk}^2]\right|\right\} \\
&\quad \le \frac{K_1}{\ell}\left\{(\log p)^{1/2}\left[\max_{m,k}\sum_{i=1}^{\ell}\mathbb{E}h_{mk}^4(X_i, X_{i+\ell})\right]^{1/2}\right. \\
&\qquad \left. + (\log p)\left[\mathbb{E}\max_{m,k}\max_{1\le i\le \ell} h_{mk}^4(X_i, X_{i+\ell})\right]^{1/2}\right\}.
\end{aligned}$$

By Pisier's inequality ([63], Lemma 2.2.2) we have

$$\left\|\max_{m,k}\max_{i\le\ell}\left|h_{mk}(X_i, X_{i+\ell})\right|\right\|_4 \le K_2 \nu_n^2 \log(np).$$

Hence it follows from (50) that

$$\mathbb{E}[\bar{\psi}^2] \le C\left\{\xi_4^4 + \xi_8^4\left(\frac{\log p}{n}\right)^{1/2} + \nu_n^4 \frac{\log^3(np)}{n}\right\}.$$

Since $\nu_n^4 \log^7(np) \le C_4 n^{1-K}$ and $\xi_8 \le C_3 \nu_n^{1/2}$, we have $\mathbb{E}[\bar{\psi}] \le C\xi_4^2$, where $C > 0$ is constant depending only on $C_i, i = 1, \dots, 4$. Then we conclude that $\mathbb{E}[\tau_*] \le C(\alpha, C_1, \dots, C_4)\beta^{-1}\xi_4^2(\log(p)/n)^{1/2}$. $\square$

Acknowledgments. The author would like to thank two anonymous referees, an Associate Editor and the Co-Editor Tailen Hsing for their careful readings and many constructive comments that lead to significant improvements of this paper. The author is also grateful to Stephen Portnoy (UIUC), Xiaofeng Shao (UIUC) and Wei Biao Wu (University of Chicago) for their helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement to "Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications" (DOI: 10.1214/17-AOS1563SUPP; .pdf). This supplemental file contains additional proofs, technical lemmas and simulation results.

REFERENCES

[1] ADAMCZAK, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** 1000–1034. MR2424985

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

[3] ARCONES, M. A. and GINÉ, E. (1993). Limit theorems for U-processes. *Ann. Probab.* **21** 1494–1542. MR1235426

[4] BENTKUS, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *J. Statist. Plann. Inference* **113** 385–402. MR1965117

[5] BENTKUS, V., GÖTZE, F. and VAN ZWET, W. R. (1997). An Edgeworth expansion for symmetric statistics. *Ann. Statist.* **25** 851–896. MR1439326

[6] BENTKUS, V. Y. (1985). Lower bounds for the rate of convergence in the central limit theorem in Banach spaces. *Litovsk. Mat. Sb.* **25** 10–21. MR0823198

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

[8] BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (1986). The Edgeworth expansion for U-statistics of degree two. *Ann. Statist.* **14** 1463–1484. MR0868312

[9] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[10] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[11] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

[12] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[13] CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

[14] CALLAERT, H. and VERAVERBEKE, N. (1981). The order of the normal approximation for a studentized U-statistic. *Ann. Statist.* **9** 194–200. MR0600547

[15] CHANG, J., ZHOU, W., ZHOU, W.-X. and WANG, L. (2017). Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *Biometrics* **73** 31–41. MR3632349

[16] CHEN, L. H. Y., FANG, X. and SHAO, Q.-M. (2013). From Stein identities to moderate deviations. *Ann. Probab.* **41** 262–293. MR3059199

[17] CHEN, S. X., ZHANG, L.-X. and ZHONG, P.-S. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105** 810–819. MR2724863

[18] CHEN, X. (2016). Gaussian approximation for the sup-norm of high-dimensional matrix-variate U-statistics and its applications. Preprint. Available at arXiv:1602.00199.

[19] CHEN, X. (2018). Supplement to "Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications." DOI:10.1214/17-AOS1563SUPP.

[20] CHEN, X., XU, M. and WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994–3021. MR3161455

[21] CHEN, X., XU, M. and WU, W. B. (2016). Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Trans. Signal Process.* **64** 6459–6470. MR3566612

[22] 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

[23] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields* **162** 47–70. MR3350040

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

[25] DASGUPTA, A., LAHIRI, S. N. and STOYANOV, J. (2014). Sharp fixed n bounds and asymptotic expansions for the mean and the median of a Gaussian sample maximum, and applications to the Donoho–Jin model. *Stat. Methodol.* **20** 40–62. MR3205720

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

[27] DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

[28] DE LA PEÑA, V. C. H. and GINÉ, E. (1999). *Decoupling: From Dependence to Independence, Randomly Stopped Processes. U-Statistics and Processes. Martingales and Beyond.* Springer, New York. MR1666908

[29] EINMAHL, U. and LI, D. (2008). Characterization of LIL behavior in Banach space. *Trans. Amer. Math. Soc.* **360** 6677–6693. MR2434306

[30] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

[31] FAN, J., LIAO, Y. and MINCHEVA, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410

[32] GINÉ, E., LATAŁA, R. and ZINN, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability, II (Seattle, WA, 1999). Progress in Probability* **47** 13–38. Birkhäuser, Boston, MA. MR1857312

[33] GÖTZE, F. (1987). Approximations for multivariate U-statistics. *J. Multivariate Anal.* **22** 212–229. MR0899659

[34] GREGORY, G. G. (1977). Large sample theory for U-statistics and tests of fit. *Ann. Statist.* **5** 110–123. MR0433669

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

[36] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[37] HOUDRÉ, C. and REYNAUD-BOURET, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic Inequalities and Applications. Progress in Probability* **56** 55–69. Birkhäuser, Basel. MR2073426

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

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

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

[41] JANSSEN, P. (1994). Weighted bootstrapping of U-statistics. *J. Statist. Plann. Inference* **38** 31–41. MR1256846

[42] KLEIN, T. and RIO, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33** 1060–1077. MR2135312

[43] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

[44] LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. MR2933663

[45] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes. Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015

[46] LEHMANN, E. L. (1999). *Elements of Large-Sample Theory*. Springer, New York. MR1663158

[47] LO, A. Y. (1987). A large sample study of the Bayesian bootstrap. *Ann. Statist.* **15** 360–375. MR0885742

[48] MAI, Q., ZOU, H. and YUAN, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42. MR2899661

[49] MASON, D. M. and NEWTON, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.* **20** 1611–1624. MR1186268

[50] MASSART, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863–884. MR1782276

[51] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[52] MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. MR0652932

[53] NAGAEV, S. V. (1979). Large deviations of sums of independent random variables. *Ann. Probab.* **7** 745–789. MR0542129

[54] PENG, J., WANG, P., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

[55] PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York. MR0388499

[56] PORTNOY, S. (1986). On the central limit theorem in $\mathbf{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields* **73** 571–583. MR0863546

[57] PRÆSTGAARD, J. and WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301

[58] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[59] RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134. MR0600538

[60] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

[61] SHAO, Q.-M. and WANG, Q. (2013). Self-normalized limit theorems: A survey. *Probab. Surv.* **10** 69–93. MR3161676

[62] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006

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

[64] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

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

[66] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

[67] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

[68] ZHANG, C.-H. (1999). Sub-Bernoulli functions, moment inequalities and strong laws for nonnegative and symmetrized U-statistics. *Ann. Probab.* **27** 432–453. MR1681165

[69] ZHANG, D. and WU, W. B. (2017). Gaussian approximation for high-dimensional time series. *Ann. Statist.* **45** 1895–1919. MR3718156

[70] ZHANG, X. and CHENG, G. (2014). Bootstrapping high dimensional time series. Available at arXiv:1406.1037.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT URBANA–CHAMPAIGN
725 S. WRIGHT STREET
CHAMPAIGN, ILLINOIS 61820
USA
E-MAIL: xhchen@illinois.edu