Electronic Journal of Statistics
Vol. 16 (2022) 1096–1152
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1915

A robust bootstrap change point test for high-dimensional location parameter*

Mengjia Yu and Xiaohui Chen

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, IL 61820, USA
e-mail: mengjia.yu.uiuc@gmail.com; xhchen@illinois.edu

Abstract: We consider the problem of change point detection for high-dimensional distributions in a location family when the dimension can be much larger than the sample size. In change point analysis, the widely used cumulative sum (CUSUM) statistics are sensitive to outliers and heavy-tailed distributions. In this paper, we propose a robust, tuning-free (i.e., fully data-dependent), and easy-to-implement change point test that enjoys strong theoretical guarantees. To achieve the robust purpose in a nonparametric setting, we formulate the change point detection in the multivariate U-statistics framework with anti-symmetric and nonlinear kernels. Specifically, the within-sample noise is canceled out by anti-symmetry of the kernel, while the signal distortion under certain nonlinear kernels can be controlled such that the between-sample change point signal is magnitude preserving. A (half) jackknife multiplier bootstrap (JMB) tailored to the change point detection setting is proposed to calibrate the distribution of our ℓ^∞-norm aggregated test statistic. Subject to mild moment conditions on kernels, we derive the uniform rates of convergence for the JMB to approximate the sampling distribution of the test statistic, and analyze its size and power properties. Extensions to multiple change point testing and estimation are discussed with illustration from numerical studies.

MSC2020 subject classifications: Primary 62F40, 62G35; secondary 62E17.
Keywords and phrases: Bootstrap, change point analysis, Gaussian approximation, high-dimensional data, U-statistics.

Received June 2020.

Contents

arXiv: 1904.03372
*Research partially supported by NSF DMS-1404891, NSF CAREER Award DMS-1752614, and University of Illinois at Urbana-Champaign (UIUC) Research Board Awards (RB17092, RB18099).

1. Introduction

Change point detection problems are commonly seen in many statistical and scientific areas including functional data analysis [6, 3], time series inspection [7, 35, 60], panel data study [19, 51, 34, 8], with applications to fields of biomedical engineering [4, 62], genomics [58], financial revenue returns [5, 20, 8] among many others. Statistical testing and estimation of change points have long history with extensive literature [24, 7, 32, 5, 9, 43, 42]. This paper studies the problem of change point detection for high-dimensional distributions (i.e., $p \gg n$) from a location family with shift parameter. Let $X_i \sim F_i, i = 1, \ldots, n$ be a sequence of independent random vectors taking values in $\mathbb{R}^p$. Our goal is to test whether or not there is a location shift in the distribution functions F_i. Precisely, let $\mathcal{F} = \{F_\theta(x) = F(x - \theta) : \theta \in \mathbb{R}^p\}$ be a location family indexed by the shift parameter θ, where $F = F_0$ is the standard distribution in $\mathcal{F}$ (F_0 is arbitrary). We consider the following hypothesis testing problem:

$$H_0 : X_i \overset{i.i.d.}{\sim} F \text{ versus } H_1 : X_1, \ldots, X_m \overset{i.i.d.}{\sim} F \text{ and } X_{m+1}, \ldots, X_n \overset{i.i.d.}{\sim} F_\theta,$$
$$\text{for some (unknown) } m \in \{1, \ldots, n-1\} \text{ and } \theta \neq 0.$$

An advantage of this model is the flexibility of $\mathcal{F}$ whose mean parameter can be non-existing. Before highlighting the robustness from it, we shall first illustrate below the intuition of constructing a test statistic for separating H_0 and H_1. For brevity, we denote $G = F_\theta$ (i.e., $G(x) = F(x-\theta)$) for a fixed θ, and $Y_j = X_{m+j}, j = 1, \ldots, n-m$. With this notation, we have $X_1, \ldots, X_m$ that are independent and identically distributed (i.i.d.) with distribution F and $Y_1, \ldots, Y_{n-m}$ that are i.i.d. with distribution G such that the change point detection problem boils down to the two-sample testing problem for the shift parameter θ with an unknown change point location m. Since m is unknown, we may take all possible ordered pairs in the whole sample $X_i, i = 1, \ldots, n$, such that the within-sample noise (i.e., in each X and Y samples, separately) cancels out and the between-sample signal is properly preserved under H_1. Note that our change point hypothesis on the location family $\mathcal{F}$ is the same as the location-shift model:

$$X_i = \theta\ \mathbf{1}(i > m) + \xi_i,\ i = 1, \ldots, n,\ \text{where } \xi_i \overset{i.i.d.}{\sim} F \text{ are random vectors in } \mathbb{R}^p\ . \tag{1.1}$$

Viewing θ as the mean-shift, a natural choice for detecting the existence of a change point shift is to consider the noise cancellations in the empirical mean differences:

$$U_n = \sum_{1 \leqslant i < j \leqslant n} (X_i - X_j). \tag{1.2}$$

Under H_0, we have $\mathbb{E}[U_n] = 0$ so that there is no mean-shift signal contained in U_n and the sampling behavior of U_n is purely determined by the random noises $\xi_1, \ldots, \xi_n$. On the other hand, if H_1 is true, then $\mathbb{E}[U_n] = -m(n-m)\theta$. Thus, if the mean difference θ between the two samples is large enough to dominate the random behavior of U_n (due to noise $\{\xi_i\}_{i=1}^n$) under H_0, then the statistic would be able to distinguish H_0 between H_1.

In practice, a main concern of using U_n in (1.2) is its robustness. Specifically, the (empirical) mean functional is not robust in the sense that its influence function is unbounded. Further, in the high-dimensional setting, robustness is a challenging issue since information contained in the data is rather limited. To address this problem, we view the shift signal θ as a more general location parameter in the distribution family $\mathcal{F}$ without referring to the means. This simple observation brings a major advantage that change point detection can be made possible even in cases where the means are undefined (such as the Cauchy distribution). To achieve the robustness purpose in a nonparametric setting, we consider a general nonlinear form of (1.2) in the U-statistics framework. Let $h : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^d$ be an *anti-symmetric* kernel, i.e., $h(x, y) = -h(y, x)$ for all $x, y \in \mathbb{R}^p$. We propose the statistic

$$T_n = T_n(X_1^n) = n^{1/2} \binom{n}{2}^{-1} \sum_{1 \leqslant i < j \leqslant n} h(X_i, X_j) \tag{1.3}$$

to test for H_0 against H_1. Clearly, T_n is a (scaled) U-statistic of order two. The anti-symmetry of the kernel h plays a key role in testing for the change point

in terms of noise cancellations. To see this, under H_0 we have $\mathbb{E}[h(X_1, X_2)] = 0$ and $\mathbb{E}[T_n] = 0$. Observe that

$$T_n = \frac{2}{n^{1/2}(n-1)} \left\{ \sum_{1 \leqslant i < j \leqslant m} h(X_i, X_j) \; + \; \sum_{i=1}^{m} \sum_{j=1}^{n-m} h(X_i, Y_j) \; + \sum_{1 \leqslant i < j \leqslant n-m} h(Y_i, Y_j) \right\}.$$

Thus if H_1 is true, then $\mathbb{E}[T_n] \approx 2n^{-3/2} m(n-m)\theta_h$, where $\theta_h = \mathbb{E}[h(X_1, Y_1)]$ is the change point signal through the kernel h. If θ_h has a suitable lower bound, then we expect that T_n can separate H_0 and H_1. For instance, consider the sign kernel $h(x, y) = \text{sign}(x - y)$, where $\text{sign}(x)$ is the component-wise sign operator of $x \in \mathbb{R}^p$ (i.e., for $j = 1, \ldots, p$, $\text{sign}(x_j) = -1, 0, 1$ if $x_j < 0$, $x_j = 0$, $x_j > 0$, respectively). Then,

$$\theta_{h,j} = \mathbb{E}[\text{sign}(X_{1,j} - Y_{1,j})] = 1 - 2\mathbb{P}(X_{1,j} \leqslant Y_{1,j}) = 1 - 2\mathbb{P}(\Delta_j \leqslant \theta_j),$$

where $\Delta_j = \xi_{1,j} - \xi_{m+1,j}$ is a random variable with symmetric distribution. In particular, if F is the distribution in $\mathbb{R}^p$ with independent components such that each component admits a continuous probability density function $\phi_j, j = 1, \ldots, p$, then under local alternatives (i.e., $\theta \approx \mathbf{0}$) we have $\theta_{h,j} \approx -2\ \phi_j^*(0)\ \theta_j$, where ϕ_j^* is the convolution of the densities of $\xi_{1,j}$ and $-\xi_{m+1,j}$. Hence, θ_h and θ have the same magnitude, implying that signal distortion under the sign kernel is only up to a multiplicative constant.

The mean difference statistic U_n in (1.2) is a special case of T_n with the linear kernel $h(x_1, x_2) = x_1 - x_2$ and $d = p$. The sign kernel $h(x, y) = \text{sign}(x - y)$ considered above is another important anti-symmetric and bounded kernel, which is useful if the means are not robust or undefined. Specifically, for the sign kernel, component-wise median of T_n corresponds to the Hodges-Lehmann estimator for the component-wise population median of the location difference before and after the change point [31]. In the univariate case $p = d = 1$, it is known that the Hodges-Lehmann estimator is a highly robust version of sample mean difference (with the linear kernel) against heavy-tailed distributions, and it has a much higher asymptotic relative efficiency $3/\pi \approx 95\%$ (with respect to the mean) than the sample median at normality [55]. In addition, when the change point location m is known, T_n is also equivalent to the classical nonparametric Mann-Whitney test statistic (see e.g., Chapter 12 in [53]).

Since T_n is a d-dimensional random vector, we need to aggregate its components to make a decision rule for hypothesis testing. We construct the critical regions based on the Kolmogorov-Smirnov (i.e., the ℓ^∞-norm) type aggregation of T_n, namely our change point test statistic is

$$\overline{T}_n := |T_n|_\infty = \max_{1 \leqslant k \leqslant d} |T_{nk}|. \tag{1.4}$$

Then H_0 is rejected if $\overline{T}_n$ is larger than a critical value such as the $(1-\alpha)$ quantile of $\overline{T}_n$. In Section 2, we will introduce a (Gaussian) multiplier bootstrap to calibrate the distribution of $\overline{T}_n$, and we will establish its non-asymptotic validity in the high-dimensional setting in Section 3.

We point out that our test statistic has comparable computational and statistical properties to the widely used cumulative sum (CUSUM) procedures in literature. For a classical treatment of the CUSUM (and other change point) statistics, we refer to [21] as a monograph on the change point analysis. The CUSUM statistics are defined as a sequence of (dependent) random vectors in $\mathbb{R}^p$ of the form

$$Z_n(s) = \left(\frac{s(n-s)}{n}\right)^{1/2} \left(\frac{1}{s}\sum_{i=1}^{s} X_i - \frac{1}{n-s}\sum_{i=s+1}^{n} X_i\right), \quad s = 1, \ldots, n-1. \tag{1.5}$$

It is obvious that the CUSUM statistics have a sequential nature in that the left and right sample averages are examined along all possible change point locations, which is necessary to estimate the location m. However, if the goal is only testing for the existence of a change point, this (local) sequential comparison strategy is not as efficient as a global test (1.3), both computationally and statistically. Consider $d = p$, which is the case for the sign and linear kernels. For a general nonlinear kernel, computational cost is $O(n^2p)$ for T_n (and also for $\overline{T}_n$). If the kernel is linear (i.e., $h(x, y) = x - y$), then the computational cost can be further reduced to $O(np)$ for T_n effortlessly. Thus we call T_n is the global *one-pass* Mann-Whitney type test statistic. In contrast, the computational cost for $\{Z_n(s)\}_{s=1}^{n-1}$ is $O(n^2p)$ which can reduces to $O(np)$ [39] via dynamic programming. Statistically, it has been shown in [61, 38] that a boundary removal procedure is needed for the (bootstrapped) CUSUM change point test to achieve the size validity since the distributions of $Z_n(s)$ are difficult to approximate at the boundary points. On the contrary, the test statistic T_n proposed in this paper does not remove any boundary points because we are able to approximate the distribution of T_n based on majority of the data points in the sample $X_1, \ldots, X_n$. Thus it is expected that $\overline{T}_n$ achieves faster rate of convergence in the error-in-size for the bootstrap calibration. See Remark 2 ahead for a detailed comparison.

1.1. Literature review and our contribution

Single change point inference has been extensively studied in literature such as [21, 29, 33] for univariate or fixed multivariate setting.

Using anti-symmetric kernels in U-statistics for location change can be traced back to [49], which considered a CUSUM-type sequence of two-sample Mann-Whitney statistics with the sign kernel and took the maximum absolute value along the sequence as the test statistic. Asymptotic properties of such statistic for univariate data have been studied in the settings of online and offline change point problems [22, 27, 30, 40]. To the best of our knowledge, the proposed global one-pass Mann-Whitney type change point detection procedure in (1.3) based on a general anti-symmetric kernel without using a CUSUM-type sequence is new in literature, even in the one-dimensional case.

Second, owing to increasing ability to handle large dimensional data, the focus migrates to a more challenging stage in high dimension that allows $p \to \infty$ faster

than n. Therefore, signal aggregation across dimension becomes influential in the designing of statistics and algorithm. For instance, [38, 61, 57] dealt with sparse change (i.e. mean structure changes in a sparse subset of coordinates), while [8, 34, 25] considered ℓ^2-type aggregation for dense change. Taking both cases into account, [25] proposed a scan test statistic aiming at sparser change coupled with their linear statistic in inference. [19] adopted additional weighted CUSUM-type factor along coordinate to make the double-CUSUM statistic more adaptive in detection. The detection rate are also investigated in terms of sparsity and signal magnitude as well as change point location [25, 44, 59]. We show that our result achieves optimal minimax rate, cf. Remark 5. For multiple change point detection which is more challenging and essential in applications, we will discuss a *backward detection* (BD) algorithm without introducing external statistics. We will also discuss an extension to dependent sequence in Remark 6.

Among the change point literature, mean change are widely explored using CUSUM statistics [38, 61, 19, 20], least-square type statistics [8, 10], U-statistics [56] and some other kernel based methods [48, 12, 2]. In practice, when error terms are heavy-tailed, Gaussianity assumption is beyond salvation and becomes too restrictive. This concern especially highlights the potential of robust non-parametric methodology (such as nonlinear projection) to avoid direct measure on mean or higher moments in data distributions. Note that the U-statistic approach, including our method in this paper, is conducting "global" characterization (either one-sample or two-sample) via kernels to have change point signals peak. Such kernel concept is different from kernel density estimator or kernel distance measure for individual observations. Specifically, [48] proposed CUSUM variant statistic based on kernel transferred data points; [12] smoothed left and right mean function using kernel density estimation; [2] applied kernel least-squares criterion to quantify segmentation candidate and estimate change point locations. Compared to aforementioned papers, our U-statistic approach starts from a pure testing point-of-view that does not rely on any tuning of bandwidth or threshold.

The rest of this paper proceeds as follows. The bootstrap calibration for the distribution of $\overline{T}_n$ is described in Section 2. Main results for size validity and power properties of the bootstrap test are derived in Section 3. Extensions to multiple change point scenario are elaborated in Section 4. We report simulation study results in Section 5 and real data examples in Section 6. All proofs with auxiliary lemmas are given in Appendix.

1.2. Notation

For $q > 0$ and a generic vector $x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, we denote $|x|_q = (\sum_{i=1}^p |x_i|^q)^{1/q}$ for the ℓ^q-norm of x and we write $|x| = |x|_2$. For a random variable X, denote $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$. For $\beta > 0$, let $\psi_\beta(x) = \exp(x^\beta) - 1$ be a function defined on $[0, \infty)$ and L_{ψ_β} be the collection of all real-valued random variables X such that $\mathbb{E}[\psi_\beta(|X|/C)] < \infty$ for some $C > 0$. For $X \in L_{\psi_\beta}$, define $\|X\|_{\psi_\beta} = \inf\{C > 0 : \mathbb{E}[\psi_\beta(|X|/C)] \leqslant 1\}$. Then, for $\beta \in [1, \infty)$,

$\|\cdot\|_{\psi_\beta}$ is an Orlicz norm and $(L_{\psi_\beta}, \|\cdot\|_{\psi_\beta})$ is a Banach space [41]. For $\beta \in (0,1)$, $\|\cdot\|_{\psi_\beta}$ is a quasi-norm, i.e., there exists a constant $C(\beta) > 0$ such that $\|X+Y\|_{\psi_\beta} \leqslant C(\beta)(\|X\|_{\psi_\beta} + \|Y\|_{\psi_\beta})$ holds for all $X, Y \in L_{\psi_\beta}$ [1]. Let $\rho(X,Y) = \sup_{t\in\mathbb{R}} |\mathbb{P}(X \leqslant t) - \mathbb{P}(Y \leqslant t)|$ be the Kolmogorov distance between two random variables X and Y. We shall use $C_1, C_2, \ldots$ and $K_1, K_2, \ldots$ to denote positive and finite constants that may have different values. The symbol $\gtrsim$ (or $\asymp, \lesssim$) denotes greater than (or equal to, smaller than) some rates with constants omitted and $\vee$ (or $\wedge$) means the maximum (or minimum) of terms.

Throughout the paper, we assume $n \geqslant 3$ and $d \geqslant 3$ (i.e., $\log n \geqslant 1$ and $\log d \geqslant 1$) to simplify some statements and all inference works for $d = 1, 2$.

2. Bootstrap calibration

To approximate the distribution of $\overline{T}_n$, we propose the following bootstrap procedure. Let $e_1, \ldots, e_n$ be i.i.d. $N(0,1)$ random variables that are independent of X_1^n. Define the bootstrapped U-statistic and test statistic as

$$T_n^\sharp = n^{1/2}\binom{n}{2}^{-1} \sum_{i=1}^n \left\{ \sum_{j=i+1}^n h(X_i, X_j) \right\} e_i \quad \text{and} \quad \overline{T}_n^\sharp := |T_n^\sharp|_\infty = \max_{1\leqslant k\leqslant d} |T_{nk}^\sharp|. \tag{2.1}$$

We reject H_0 if $\overline{T}_n > q_{\overline{T}_n^\sharp | X_1^n}(1-\alpha)$, where

$$q_{\overline{T}_n^\sharp | X_1^n}(1-\alpha) = \inf\left\{ t \in \mathbb{R} : \mathbb{P}(\overline{T}_n^\sharp \leqslant t \mid X_1^n) \geqslant 1-\alpha \right\}$$

is the $(1-\alpha)$ quantile of the conditional distribution of $\overline{T}_n^\sharp$ given X_1^n. Before presenting the rigorous validity of our bootstrap test procedure in terms of the size and power in Section 3, we shall explain the reason why it can (asymptotically) separate H_0 against H_1.

First, suppose H_0 is true, i.e., $X_1, \ldots, X_n$ are i.i.d. with distribution F. Let $g(x) = \mathbb{E}[h(x, X_1)]$ and $f(x_1, x_2) = h(x_1, x_2) - g(x_1) + g(x_2)$. Due to the anti-symmetry of h, we have $f(x_1, x_2) = -f(x_2, x_1)$. Then the Hoeffding decomposition of T_n is

$$T_n = \underbrace{n^{-1/2} \sum_{i=1}^n \frac{2(n-2i+1)}{n-1} g(X_i)}_{L_n} + \underbrace{n^{1/2}\binom{n}{2}^{-1} \sum_{1\leqslant i<j\leqslant n} f(X_i, X_j)}_{R_n}. \tag{2.2}$$

Since f is degenerate, the linear part L_n is expected to be a leading term of T_n, and the distribution of L_n (denote as $\mathcal{L}(L_n)$) can be approximated by its Gaussian analog via matching the first and second moments [17, 13]. Since $\mathbb{E}[L_n] = 0$ and

$$\mathrm{Cov}(L_n) = \frac{4(n+1)}{3(n-1)}\Gamma \approx \frac{4}{3}\Gamma \quad \text{with} \quad \Gamma = \mathrm{Cov}(g(X_1)),$$

we expect that $\mathcal{L}(L_n) \approx \mathcal{L}(Z)$, where $Z \sim N(0, 4\Gamma/3)$, for a large sample size n. Once the Gaussian approximation result for T_n by Z is established, the rest of the work is to compare the distribution of Z and the conditional distribution of $T_n^\sharp$ given X_1^n, both of which are mean-zero Gaussians. Since $\mathrm{Cov}(T_n^\sharp \mid X_1^n) = \frac{4}{n(n-1)^2}\sum_{i=1}^n \sum_{j=i+1}^n \sum_{k=i+1}^n h(X_i, X_j)h(X_i, X_k)^T$, standard concentration inequalities for (one-sample) U-statistics in [13] yield that $\mathrm{Cov}(T_n^\sharp \mid X_1^n) \approx 4\Gamma/3$. Thus we expect that $\mathcal{L}(T_n^\sharp \mid X_1^n) \approx \mathcal{L}(Z) \approx \mathcal{L}(T_n)$, from which the size validity of the bootstrapped change point test based on $\overline{T}_n^\sharp$ follows.

Next, we suppose H_1 is true, i.e., $X_1, \ldots, X_m$ are i.i.d. with distribution F and $Y_1, \ldots, Y_{n-m}$ are i.i.d. with distribution G such that $G(x) = F(x-\theta)$ and $Y_i = X_{i+m}, i = 1, \ldots, n-m$. To study the power property, the main idea is to consider the two-sample Hoeffding decomposition of T_n that is similar to (2.2). Suppose $h(x+c, y+c) = h(x, y)$ is *shift-invariant* in terms of location parameter. Let $\theta_h = \mathbb{E}[h(X_1, Y_1)]$,

$$Gh(x) = \mathbb{E}[h(x, Y_1)] - \theta_h = g(x-\theta) - \theta_h, \quad Fh(y) = \mathbb{E}[h(X_1, y)] - \theta_h = -g(y) - \theta_h,$$

such that $\mathbb{E}[Gh(X_1)] = \mathbb{E}[Fh(Y_1)] = 0$. Define

$$\breve{f}(x, y) = h(x, y) - Gh(x) - Fh(y) - \theta_h,$$

which is degenerate such that $\mathbb{E}[\breve{f}(X_1, Y_1)] = \mathbb{E}[\breve{f}(X_1, y)] = \mathbb{E}[\breve{f}(x, Y_1)] = 0$. Under H_1, we may split the U-statistic sum as

$$\sum_{1 \leqslant i < j \leqslant n} h(X_i, X_j) = \sum_{\substack{1 \leqslant i < j \leqslant m \\ m+1 \leqslant i < j \leqslant n}} h(X_i, X_j) + \sum_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n-m}} h(X_i, Y_j),$$

where the first sum on the r.h.s. of the above equation has mean zero (again, due to the anti-symmetry of h). Thus, to study the power of $\overline{T}_n$ (and its bootstrapped version $\overline{T}_n^\sharp$), it suffices to analyze the second sum on the r.h.s. of the last display above, which is a two-sample U-statistic V_n that admits the following Hoeffding decomposition:

$$\begin{aligned} V_n &= \sum_{i=1}^m \sum_{j=1}^{n-m} h(X_i, Y_j) \\ &= m(n-m)\theta_h + (n-m)\sum_{i=1}^m Gh(X_i) + m\sum_{j=1}^{n-m} Fh(Y_j) + \sum_{i=1}^m \sum_{j=1}^{n-m} \breve{f}(X_i, Y_j). \quad (2.3) \end{aligned}$$

Since the last three sums on the r.h.s. of (2.3) have mean zero, the power of the proposed test is determined by the magnitude of θ_h and the sampling distributions of other terms involving no θ_h. For the latter, all of those distributions can be well estimated and controlled as in H_0 since they do not contain the change point signal. Thus, if $|\theta_h|_\infty$ obeys a minimal signal size requirement, then the power of $\overline{T}_n^\sharp$ would tend to one.

Remark 1. It is interesting to note that our bootstrapped U-statistic $T_n^\sharp$ in (2.1) is closely related to the jackknife multiplier bootstrap (JMB) proposed in [13] for high-dimensional U-statistics and in [15] for infinite-dimensional U-processes with symmetric kernels. In both settings, the (unobserved) Hájek projection process $g(\cdot)$ is estimated by the jackknife procedure and a multiplier bootstrap is applied to the jackknife estimated process. In our change point detection context, since the kernel is anti-symmetric, averaging the empirical Hájek process by jackknife would simply be an estimate of zero. Thus, we may only use half (e.g., a triangular array index subset $i < j$) of the JMB to estimate $g(\cdot)$. In view of this connection, we call our bootstrap method is a JMB tailored to change point detection. □

3. Theoretical properties

Let X, X' be i.i.d. random vectors with distribution F. Recall that $g(x) = \mathbb{E}[h(x, X)]$ and $f(x_1, x_2) = h(x_1, x_2) - g(x_1) + g(x_2)$ in the Hoeffding decomposition (2.2). Then $\mathbb{E}[g(X)] = 0$ and $\mathbb{E}[f(x_1, X')] = \mathbb{E}[f(X, x_2)] = 0$ for all $x_1, x_2 \in \mathbb{R}^p$ (i.e., f is degenerate). Denote $\Gamma = \mathrm{Cov}(g(X)) = \mathbb{E}[g(X)^T g(X)]$. In this section, we will characterize theoretical properties through d (the dimension of h) and θ_h (the expected mean change of $h(X, X+\theta)$) rather than p (the original dimension of data) or θ (the original location shift parameter) since the whole procedure is constructed on top of $h(X, X')$.

3.1. Size validity

We first establish the validity of the bootstrap approximation to the distribution of $\overline{T}_n$ under H_0. Let $\underline{b} > 0$ be a constant and $D_n \geqslant 1$ which is allowed to increase with n. We make the following assumptions.

(A1) $\mathbb{E}g_j(X)^2 \geqslant \underline{b}^2$ for all $j = 1, \ldots, d$.
(A2) $\mathbb{E}|h_j(X, X')|^{2+k} \leqslant D_n^k$ for all $j = 1, \ldots, d$ and $k = 1, 2$.
(A3) $\|h_j(X, X')\|_{\psi_1} \leqslant D_n$ for all $j = 1, \ldots, d$.

Condition (A1) is a non-degeneracy requirement for the kernel h. Without (A1), bootstrap may approximate constant observation through a random process so that our method is not valid. Conditions (A2) and (A3) impose moment conditions on the kernel h coupled with the data distribution F. For instance, when the kernel is bounded, we do not explicitly impose additional assumption on the data distribution F. Thus conditions (A2) and (A3) are more robust than the canonical linear kernel when the data distribution has polynomial tails. In our high-dimensional setting, we allow both p and d to increase with n.

Theorem 3.1 (Size validity of bootstrap test under H_0)**.** *Suppose H_0 is true and (A1)-(A3) hold. Let $\gamma \in (0, e^{-1})$ such that $\log(1/\gamma) \leqslant K \log(nd)$ for some constant $K > 0$. Then there exists a constant $C := C(\underline{b}, K)$ depending only on*

$\underline{b}$ and K such that

$$\rho(\overline{T}_n, \overline{T}_n^\sharp \mid X_1^n) := \sup_{t\in\mathbb{R}} \left| \mathbb{P}(\overline{T}_n \leqslant t) - \mathbb{P}(\overline{T}_n^\sharp \leqslant t \mid X_1^n) \right| \leqslant C\varpi_n \tag{3.1}$$

holds with probability at least $1-\gamma$, where

$$\varpi_n = \left\{ \frac{D_n^2 \log^7(nd)}{n} \right\}^{1/6}. \tag{3.2}$$

Consequently, we have

$$\sup_{\alpha\in(0,1)} \left| \mathbb{P}(\overline{T}_n \leqslant q_{\overline{T}_n^\sharp | X_1^n}(\alpha)) - \alpha \right| \leqslant C\varpi_n + \gamma. \tag{3.3}$$

In particular, if $\log d = o(n^{1/7})$, then $\mathbb{P}(\overline{T}_n \leqslant q_{\overline{T}_n^\sharp | X_1^n}(\alpha)) \to \alpha$ uniformly in $\alpha \in (0,1)$ as $n \to \infty$.

Theorem 3.1 constructs non-asymptotic bootstrap validity in theory and guarantees that the α-th quantile of bootstrapped statistic $\overline{T}_n^\sharp | X_1^n$ is always close to the α-th quantile of test statistic $\overline{T}_n$. Moreover, the error bound is uniform over $\alpha \in (0,1)$. The technique for proving Theorem 3.1 extends the Gaussian approximation theory for U-statistics in [13], which focuses on symmetric kernels.

Remark 2 (Comparisons with the CUSUM-based statistics). [38] and [61] propose CUSUM-based bootstrap tests that require the removal of boundary points for detecting change points in high-dimensional mean vectors. Specifically, for the CUSUM statistics (1.5) considered in [61], the test statistic is of the form $S_n = \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} |Z_n(s)|_\infty$ for some boundary removal parameter $\underline{s} \in [1, n/2]$. Accordingly, the Gaussian multiplier bootstrap version of $Z_n(s)$ is defined as:

$$Z_n^\sharp(s) = \left(\frac{n-s}{ns}\right)^{1/2} \sum_{i=1}^{s} e_i(X_i - \overline{X}_s^-) - \left(\frac{s}{n(n-s)}\right)^{1/2} \sum_{i=s+1}^{n} e_i(X_i - \overline{X}_s^+),$$

where $\overline{X}_s^- = s^{-1}\sum_{i=1}^s X_i$ and $\overline{X}_s^+ = (n-s)^{-1}\sum_{i=s+1}^n X_i$ are the left and right sample averages at s, respectively. $Z_n^\sharp(s)$ sequentially inspects the two-sample distributions before and after all possible change point locations in the interval $[\underline{s}, n-\underline{s}]$. Then for the special case of linear kernel $h(x,y) = x-y$ and distribution F satisfying the conditions (A1), (A2), and (A3), the rate of convergence for $\overline{S}_n^\sharp := \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} |Z_n^\sharp(s)|_\infty$ shown in [61] obeys

$$\rho(\overline{S}_n, \overline{S}_n^\sharp \mid X_1^n) \leqslant C \left\{ \frac{D_n^2 \log^7(nd)}{\underline{s}} \right\}^{1/6}$$

with probability at least $1-\gamma$. Comparing the last display with the rate of convergence for $\rho(\overline{T}_n, \overline{T}_n^\sharp \mid X_1^n)$ in (3.1) and (3.2), we see that the JMB method

proposed here has better statistical properties than the Gaussian multiplier bootstrap $\overline{T}_n^\sharp$ without removing any boundary points in computing $\overline{T}_n$ and $\overline{T}_n^\sharp$. Consequently this will reduce the error-in-size (3.3) for our bootstrap calibration $\overline{T}_n^\sharp$. Empirical evidence for our algorithm with smaller error-in-size can be found in Section 5. The main reason for the improved rate is due to the fact that we can approximate the distribution of $\overline{T}_n$ based on the majority of the data points in the entire sample $X_1, \ldots, X_n$. In addition, the proposed change point detector $\overline{T}_n$ and its JMB calibration $\overline{T}_n^\sharp$ can be viewed as a *nonlinear* and *one-pass* version of the CUSUM statistics. □

Remark 3 (Improved size validity of the bootstrap test). Proof of Theorem 3.1 is based on the Gaussian and bootstrap results for linear partial sums in high dimensions [17] and the maximal inequality for degenerate U-statistics [15]. Since the work of [17], there have been substantial progresses being made to improve the rate of convergence of Gaussian approximation for partial sums under various settings. For instance, [18] derived nearly optimal bound for the Gaussian approximation over hyper-rectangles. Tailored to our change point detection setting, if the correlation matrix of L_n is *strongly* non-degenerate (i.e., the smallest eigenvalue of the correlation matrix of L_n is strictly positive), then the rate of Gaussian approximation to L_n can be sharpened to $n^{-1/2}(\log n)(\log d)^{3/2}$. Combining this with the maximal inequality for R_n, we can improve the overall bound for $\rho(\overline{T}_n, \overline{T}_n^\sharp \mid X_1^n)$ to $n^{-1/4}(\log(nd))^{1/2}(\log n)(\log d)^{1/2}$.

Let σ_* be the square root of the smallest eigenvalue of the correlation matrix of $g(X)$. We assume that

(A2') $\mathbb{E}|h_j(X, X')|^4 \leqslant D_n^2$ for all $j = 1, \ldots, d$.
(A3') $\|h_j(X, X')\|_{\psi_2} \leqslant D_n$ for all $j = 1, \ldots, d$.

Theorem 3.2 (Improved size validity of the bootstrap test under H_0). *Suppose H_0 is true, $\sigma_*^2 > 0$, and (A1), (A2') and (A3') hold. Let $\gamma \in (0, e^{-1})$ such that $\log(1/\gamma) \leqslant K \log(nd)$ for some constant $K > 0$. Then there exists a constant $C := C(\underline{b}, \sigma_*, K)$ depending only on $\sigma_*, \underline{b}$ and K such that*

$$\rho(\overline{T}_n, \overline{T}_n^\sharp \mid X_1^n) \leqslant C\varpi_n' \tag{3.4}$$

holds with probability at least $1 - \gamma$, where

$$\varpi_n' = \frac{D_n(\log(nd))^{1/2}(\log n)(\log d)^{1/2}}{n^{1/4}}. \quad \square \tag{3.5}$$

3.2. Power analysis

Next, we analyze the power of the proposed testing under H_1 in terms of the change point signal $\theta_h = \mathbb{E}[h(X, X' + \theta)]$ and its location m. In our U-statistic framework, the test implicitly depends on θ through θ_h, which the signal strength characterization will relate to. As we have discussed earlier, the signal magnitudes between θ and θ_h can be preserved for the robust sign kernel. Under H_1, we assume the following conditions.

(B1) h is *shift-invariant*: $h(x+c, y+c) = h(x,y)$.
(B2) $\mathbb{E}|h_j(X, X'+\theta) - \mathbb{E}[h_j(X, X'+\theta)]|^{2+\ell} \leqslant D_n^\ell$ for all $j = 1, \cdots, d$ and $\ell = 1, 2$.
(B3) $||h_j(X, X'+\theta) - \mathbb{E}[h_j(X, X'+\theta)]||_{\psi_1} \leqslant D_n$ for all $j = 1, \cdots, d$.

Condition (B1) is a natural requirement since the within-sample noise cancellation by h should be invariant under data translation in the location-shift model (1.1). Conditions (B2) and (B3) are in parallel with Condition (A2) and (A3) in the sense that they quantify the moment and tail behaviors of the centered version of the kernel h (w.r.t. the distribution F). In particular, Conditions (B2) and (B3) separate the location-shift signal from the mean-zero noise, and if $\theta = 0$, Conditions (B2) and (B3) reduce to Conditions (A2) and (A3). Our next theorem characterizes the minimal signal strength for detecting the change point under the alternative hypothesis H_1.

Theorem 3.3 (Power of bootstrap test under H_1)**.** *Suppose H_1 is true and (B1)-(B3) hold in addition to (A1)-(A3). Let $\zeta \in (0, e^{-1})$ such that $\log(1/\zeta) \leqslant K \log(nd)$ for some constant $K > 0$. Suppose $m \wedge (n-m) \geqslant K' \log^{5/2}(nd)$ for some large enough $K' > 0$. If*

$$m(n-m)|\theta_h|_\infty > K_0 D_n n^{3/2} \log^{1/2}(nd/\alpha) + C_1(\underline{b}) n^{3/2} \log^{1/2}(\zeta^{-1}) \log^{1/2}(d), \tag{3.6}$$

for some constants K_0 and $C_1(\underline{b})$, then $\mathbb{P}(\overline{T}_n > q_{\overline{T}_n^\sharp | X_1^n}(1-\alpha)) \geqslant 1 - \zeta - C_2(\underline{b})\varpi_n$.

Theorem 3.3 provides the lower bound of signal strength that is related to change point location m and size level α, as well as sample size n and kernel dimension d. Markedly, our theory derives the tail probability control on the maximum of two-sample order-two U-statistics.

Remark 4 (Interpretation of Theorem 3.3). Note the first term on the r.h.s. of (3.6) reflects the Type I error of the bootstrap test (coming from α and ϖ_n in Theorem 3.1), while the second term reflects the connection to the Type II error under H_1 through ζ. If the location shift happens in the middle, i.e., $m \asymp n$, then $m(n-m) \asymp n^2$. In this case, the signal strength has to obey $|\theta_h|_\infty \gtrsim D_n n^{-1/2} \log^{1/2}(nd/\alpha)$, which matches the power result for the bootstrap test based on the CUSUM statistics in [61] (cf. Theorem 3.3 therein). If the location shift occurs at the boundary, for instance $m \wedge (n-m) \asymp n^\beta$ for $\beta < 1/2$, then the signal has to be $|\theta_h|_\infty \gtrsim n^{1/2-\beta}$, which diverges to infinity. Thus, under our framework, detection is possible for local alternative when the change point location satisfies $m \wedge (n-m) \gtrsim D_n n^{1/2} \log^{1/2}(nd)$. □

Remark 5 (Rate optimality for sparse alternative). In [44, Theorem 1], the authors derived the minimax rate of detection boundary for single change point case where F is p-dimensional Gaussian distribution with independent entries. Suppose the location shift only occurs in the first k components with the same size of $\rho > 0$, i.e.

$$\theta = (\underbrace{\rho, \ldots, \rho}_{k \text{ times}}, 0, \ldots, 0)^\top.$$

For sparse regime when $k = |\theta|_0 < \sqrt{p \log\log(8n)}$, let $|\theta_h|_2^2 \approx |\theta|_2^2 = k\rho^2$ under local alternative, then their minimax result reads as

$$\frac{m(n-m)}{n} k\rho^2 \gtrsim \rho^*(p,n,k) \asymp \left(k \log\{\frac{ep \log\log(8n)}{k^2}\} \vee \log\log(8n) \right).$$

Note that, $m(n-m) = (m \wedge (n-m))((m \vee (n-m)) \asymp (m \wedge (n-m))n$. Hence, their result indicates that $\rho \gtrsim (m \wedge (n-m))^{-1/2} \sqrt{\log\{\frac{ep \log\log(8n)}{k^2}\} \vee \frac{1}{k} \log\log(8n)}$. The rate inside square root is up to a logarithm factor through n, p (for example by plugging in $k = 1$). On the other hand, our (3.6) in Theorem 3.3 requires the lower bound $\rho \gtrsim (m \wedge (n-m))^{-1} n^{1/2}$ up to $\log^{1/2}(nd)$. If $m \wedge (n-m)$ is bounded away from boundaries, i.e., $m \asymp n - m \asymp n$, then our result is minimax optimal. □

Remark 6 (Extension of the bootstrap test to time series data). When the noise sequence ξ_i in the location-shift model (1.1) is a stationary time series, we need to modify the bootstrap test statistic to adjust for the temporal dependency because $\mathbb{E}h(X_i, X_j)$ is no longer zero and there is a bias term to be calibrated in the bootstrap test. Nonetheless, if the time series ξ_i is weakly dependent, then the bias term decays to zero when $|i-j|$ increases. This motivates us to consider a *trimmed* version of the bootstrap test by removing summands within close indices in T_n (and thus $T_n^\sharp$). Let the integer $0 \leqslant M < m \wedge (n-m)$ be a trimming parameter. We define a generalized U-statistic as

$$T_n^\natural = n^{1/2} \binom{n}{2}^{-1} \sum_{\substack{i<j \\ |i-j|>M}} h(X_i, X_j) = \frac{2}{n^{1/2}(n-1)} \sum_{i=1}^{n-M-1} \sum_{j=i+M+1}^{n} h(X_i, X_j). \tag{3.7}$$

Under H_0, we expect $h(X_i, X_j)$ behaves similarly to the i.i.d. scenario for large M since the dependency between X_i and X_j is weak. Thus, we have $\mathbb{E}h(X_i, X_j) \approx 0$ for $|i-j| > M$ and $\mathbb{E}T_n^\natural \approx 0$. Under H_1, with $\mathbb{E}h(X_i, X_j) \approx \theta_h$ for $i \leq m < j$ and $|i-j| > M$, we have

$$\begin{aligned} \mathbb{E}T_n^\natural &\approx n^{1/2} \binom{n}{2}^{-1} \left[\sum_{i=1}^{m-M} \sum_{j=m+1}^{n} + \sum_{i=n-M+1}^{m} \sum_{j=i+M+1}^{n} \right] \mathbb{E}h(X_i, X_j) \\ &\approx 2n^{-3/2} \left[m(n-m) - (M+1)M/2 \right] \theta_h. \end{aligned} \tag{3.8}$$

There is a natural trade-off in choosing the trimming parameter M to control the effective signal strength $\mathbb{E}T_n^\natural$ under H_0 and H_1. For larger values of M, calibration of the distribution of $T_n^\natural$ would be more accurate. However, the compromise of signal strength in (3.8) would also be larger. Thus, it would be harder to detect change point (i.e., to separate H_0 from H_1) when the temporal dependence of data is stronger. Similarly as the i.i.d. noise case, we can use the ℓ^∞-norm to construct our test statistic

$$\overline{T}_n^\natural := |T_n^\natural|_\infty = \max_{1 \leqslant k \leqslant d} |T_{nk}^\natural|, \tag{3.9}$$

which separates H_0 from H_1 when temporal dependence exists.

Let $e_1, \ldots, e_{n-M+1}$ be i.i.d. $N(0,1)$ random variables that are independent of X_1^n. Define the bootstrapped test statistic

$$T_n^\flat = n^{1/2}\binom{n}{2}^{-1} \sum_{i=1}^{n-M-1} \left\{ \sum_{j=i+M+1}^{n} h(X_i, X_j) \right\} e_i \tag{3.10}$$

and $\overline{T}_n^\flat := |T_n^\flat|_\infty = \max_{1 \leqslant k \leqslant d} |T_{nk}^\flat|$. We reject H_0 if $\overline{T}_n^\natural > q_{\overline{T}_n^\flat | X_1^n}(1-\alpha)$, the $(1-\alpha)$ quantile of the conditional distribution of $\overline{T}_n^\flat$ given X_1^n.

When ξ_i is an independent noise sequence, we simply set $M = 0$ so that $T_n^\natural$ and $T_n^\flat$ reduce to T_n and $T_n^\sharp$, respectively. In Section 5.6, we shall provide some empirical performance of the trimmed bootstrap test for a vector autoregressive process ξ_i. □

4. Extensions to multiple change points scenario

4.1. Direct extension to multiple change points testing

Recall $X_i \sim F_i, i = 1, \ldots, n$ as a sequence of independent random vectors taking values in $\mathbb{R}^p$. Generally, suppose there are ν change points $m_0 = 0 < m_1 < \cdots < m_\nu < m_{\nu+1} = n$ such that

$$F_{m_k+1}(x) = \cdots = F_{m_{k+1}}(x) = F(x - \theta^{(k)}) \text{ and } F_{m_k} \neq F_{m_{k+1}} \text{ for } k = 0, \ldots, \nu.$$

Without loss of generality, we can assume $\theta^{(0)} = 0$. Consider the alternative hypothesis with multiple change points

$$H_1' : \theta^{(k)} \neq \theta^{(k+1)} \text{ for some } m_k, k = 0, \ldots, \nu \text{ and } \nu \geqslant 1. \tag{4.1}$$

Denote $X_i = \xi_i + \theta^{(k)}$ and due to the shift-invariant property (B1) we have

$$\delta^{(k,k')} = \mathbb{E}h(X_i, X_j) = \mathbb{E}h(\xi_i, \xi_j + (\theta^{(k')} - \theta^{(k)})) \quad \text{for } m_k < i \leqslant m_{k+1}, m_{k'} < j \leqslant m_{k'+1}.$$

Let $s_i = m_{i+1} - m_i$ be the size of data segment that corresponds to the i-th location shift. Then,

$$\mathbb{E}\left[\sum_{1 \leqslant i < j \leqslant n} h(X_i, X_j) \right] = \sum_{0 \leqslant k < k' \leqslant \nu} s_k s_{k'} \delta^{(k,k')} =: \tilde{\Delta}, \tag{4.2}$$

where the standardized signal strength is $|E[T_n]|_\infty = n^{1/2}\binom{n}{2}^{-1}|\tilde{\Delta}|_\infty$. Under the multiple change points alternative, if signal cancellation does not exist, i.e. $|\tilde{\Delta}|_\infty$ is away from 0, then we can directly extend the theory as below.

Lemma 4.1 (Power of the bootstrap test under H_1')**.** *Suppose H_1' is true and (B1)-(B3) hold in addition to (A1)-(A3). Let $\zeta \in (0, e^{-1})$ such that $\log(1/\zeta) \leqslant K \log(\nu^2 nd)$ for some constant $K > 0$. Suppose ν is a constant. If*

$$|\tilde{\Delta}|_\infty > K_0 \nu^2 D_n n^{3/2} \log^{1/2}(nd/\alpha) + C_1(\underline{b}) n^{3/2} \log^{1/2}(\zeta^{-1}) \log^{1/2}(d) + \phi, \quad (4.3)$$

where

$$\phi = K_0' \left\{ n^{3/4} \log^{3/4}(nd/\alpha) \max_{k<k'} (s_k s_{k'})^{1/4} |\delta^{(k,k')}|_\infty + n^{1/2} \log^{1/2}(nd/\alpha) \sum_{k<k'} (s_k s_{k'})^{1/2} |\delta^{(k,k')}|_\infty \right\},$$

then $\mathbb{P}(\overline{T}_n > q_{\overline{T}_n^\sharp | X_1^n}(1-\alpha)) \geqslant 1 - \zeta - C_2(\underline{b}) \varpi_n$ for some constants K_0, K_0' and $C_1(\underline{b}), C_2(\underline{b})$.

Remark 7 (Explanation on ϕ and connection to single change point case). Compared to (3.6) in Theorem 3.3, there is an additional term ϕ in (4.3). It comes from controlling $\mathrm{Cov}(T_n^\sharp \mid X_1^n)$ under the alternative hypothesis. Consider the special case of single change point where $\nu = 1$ in (4.1), we may assume $m = s_0 < s_1 = n - m$. Then $\phi \asymp (m^{1/4} n \log^{3/4}(nd) + m^{1/2} n \log^{1/2}(nd)) |\delta^{(0,1)}|_\infty \lesssim m(n-m)|\delta^{(0,1)}|_\infty = |\tilde{\Delta}|_\infty$ for $m \gtrsim \log^{5/2}(nd)$, i.e., ϕ is dominated by the l.h.s. of (4.3). Then our result under H_1' reads the same as (3.6). □

The l.h.s. of (4.3) is the overall signal strength which does not directly depend on minimum separation of change points $\underline{m} = \min_{0 \leqslant k \leqslant \nu} s_k$ or signal strength like $\bar{\delta} = \max_{0 \leqslant k < k' \leqslant \nu} |\delta^{(k,k')}|_\infty$ or $\bar{\delta}' = \min_{0 \leqslant k < \nu} |\delta^{(k,k+1)}|_\infty$ that is usually assumed under CUSUM-based approach [19, 20, 61]. Taking (1.5) for instance, our framework does not screen out any statistic by visiting each location $i = 1, \ldots, n-1$. Therefore, we allow the product of $s_k s_{k'} \delta^{(k,k')}$ dominates the overall change $\tilde{\Delta}$ even if s_k or $\delta^{(k,k')}$ is fairly small. However, it is inconvenient that signal cancellation in (4.2) cannot be characterized by $\underline{m}$ or $\bar{\delta}$. Another drawback is that $\tilde{\Delta} = 0$ can happen even if $\underline{m} \asymp O(n)$ and $\bar{\delta}$ is large. This issue will be discussed in the next section. Before that, we discuss two special cases derived from Lemma 4.1 based on $\underline{m}$ and $\bar{\delta}$ to make the lemma more informative and instructional. Besides, we can avoid $|\delta^{(k,k')}|_\infty$ being on both sides of (4.3).

1. Suppose $\bar{\delta}$ is upper bounded, for example h is the bounded sign kernel. We have $s_k < n$, which leads to $\max_{0 \leqslant k < k' \leqslant \nu} (s_k s_{k'})^{1/4} \leqslant n^{1/2}$ and $\sum_{k<k'} (s_k s_{k'})^{1/2} \leqslant \nu^2 n$. Since $n \gtrsim \log^7(nd)$, so $\phi \lesssim \nu^2 n^{3/2} \log^{1/2}(nd) \bar{\delta}$, which is nearly the same rate as the first part on the r.h.s. of (4.3). Therefore, ϕ can be dropped.
2. Suppose $\{|\delta^{(k,k')}|_\infty : 0 \leqslant k < k' \leqslant \nu\}$ are at the same magnitude and $|\tilde{\Delta}|_\infty$ is dominated by $s_k s_{k'} |\delta^{(k,k')}|_\infty \gtrsim \underline{m}^2 \bar{\delta}$ for some pair of (k, k'). Then a sufficient condition to control Type II error is to have $\underline{m}^2 \bar{\delta}$ greater than the upper bound of ϕ, namely $n^{3/2} \log^{1/2}(nd) \bar{\delta}$. So we only need $\underline{m} \gtrsim n^{3/4} \log^{1/4}(nd)$. This is weaker than the condition in [19, (B1)] which

requires $\underline{m} \gtrsim n^{6/7}$. One example of such assumption is the setup in [38] where each dimension has at most one change.

In summary, we have the following corollary.

Corollary 4.2. *Suppose the conditions in Lemma 4.1 are satisfied.*
(i) If $\bar{\delta} = \max_{0\leqslant k<k'\leqslant\nu} |\delta^{(k,k')}|_\infty$ *is bounded, then* $\mathbb{P}(\overline{T}_n > q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha)) \geqslant 1-\zeta-C_2(\underline{b})\varpi_n$ *when*

$$|\tilde{\Delta}|_\infty = |\sum_{k<k'} s_k s_{k'} \delta^{(k,k')}|_\infty$$
$$> K_0\nu^2 D_n n^{3/2}\log^{1/2}(nd/\alpha) + C_1(\underline{b})n^{3/2}\log^{1/2}(\zeta^{-1})\log^{1/2}(d).$$

(ii) If all $|\delta^{(k,k')}|_\infty$ *are at the same rate and* $|\tilde{\Delta}|_\infty > K_1\underline{m}^2\bar{\delta}'$, *then* ϕ *in* (4.3) *can be dropped when*

$$\underline{m} = \min_{0\leqslant k\leqslant\nu} s_k \geqslant K_2 n^{3/4}\log^{1/4}(nd/\alpha).$$

Consequently, if signals are almost evenly spread (i.e. $\underline{m} \asymp n$*) and* $|\delta^{(k,k')}|_\infty$ *is upper bounded, then* $\mathbb{P}(\overline{T}_n > q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha)) \geqslant 1-\zeta-C_2(\underline{b})\varpi_n$ *when*

$$|\sum_{k<k'} \delta^{(k,k')}|_\infty > K_0\nu^2 D_n n^{-1/2}\log^{1/2}(nd/\alpha) + C_1(\underline{b})n^{-1/2}\log^{1/2}(\zeta^{-1})\log^{1/2}(d).$$

In Remark 4, we have shown that local alternative is detectable when $\underline{m} \gtrsim n^{1/2}\log^{1/2}(nd/\alpha)$. Corollary 4.2 (ii) has a stronger requirement due to extra cost from handling the possible cancellation in analyzing the general case of multiple change points. If there is only one change point, then the interpretation of rates in Lemma 4.1 can be found in Remark 7. A real application for our global test lies in the special case of monotone signals that have order structures $\theta_1 \leqslant \cdots \leqslant \theta_\nu$ [45].

4.2. Modification to block testing

The direct extension of testing H_0 against H_1' depends on $|\tilde{\Delta}|_\infty$, which can be 0 even if each $|\delta^{(k,k')}|_\infty$ are fairly large. The global test will not help under severe signal cancellation. One solution is to localize the test such that the problem can convert to single change point scenario.

Consider performing a block testing in the following way. Divide the sample into B blocks of size L ($n = BL$ for brevity) where $L \leqslant 2\underline{m}$. Then each block contains at most 1 change point. We can apply the original test to the block-vector data $Z_1, \ldots, Z_L \in \mathbb{R}^{Bp}$, where $Z_i = \mathrm{vec}(X_i, \ldots, X_{bL+i}, \ldots, X_{(B-1)L+i})$. Let $h^Z : \mathbb{R}^{Bp} \times \mathbb{R}^{Bp} \to \mathbb{R}^{Bd}$ be the block version extension of h:

$$h^Z(Z_i, Z_j) = (h(X_i, X_j)^\top, \ldots, h(X_{(B-1)L+i}, X_{(B-1)L+j})^\top)^\top.$$

Note that there is no signal cancellation issue. Denote $m_k^Z = (m_k \bmod L)$. Modified theory of power will depend on signal strength as below.

Corollary 4.3. *Suppose the conditions in Lemma 4.1 hold. If*

$$\max_{0\leqslant k\leqslant \nu} m_k^Z(L-m_k^Z)|\delta^{(k,k')}|_\infty > K_0\nu^2 D_n L^{3/2}\log^{1/2}(\frac{nd}{\alpha})+C(\underline{b})L^{3/2}\log^{1/2}(\zeta^{-1})\log^{1/2}(d),$$

then $\mathbb{P}\left(\overline{T}_n > q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha)\right) \geqslant 1-\zeta-C_2(\underline{b})\varpi_n$ *for some constants* K_0 *and* $C_1(\underline{b}), C_2(\underline{b})$.

Note that the rate now depends on L rather than n (except for logarithm factors). The block test sacrifices sample size to gain the single change-point structure. In practice, the block parameter L (or equivalently B) needs to be selected carefully since power depends on the relevant locations of $\{m_k^Z\}_{k=0}^\nu$. One solution is to use $L = 2n^{1/2}\log^{1/2}(nd)$ that is discussed in Remark 4 or $L = 2n^{3/4}\log^{1/4}(nd)$ that is from Corollary 4.2 (ii).

4.3. Discussion on binary segmentation

To deal with multiple change points, binary segmentation (BS) is conceptually straightforward [19, 20, 61]. The main idea is to recursively estimate change points by screening sub-segments before and after each estimated location. However, such process starts from a "global" detection that may miss change points under unfavorable configuration of signal cancellation. To improve BS, [26] proposed wild binary segmentation (WBS) that randomly draw intervals to localize searching for change points. Recently, it has been widely adopted [57, 56] owing to its flexibility and computational efficiency. However, we will not be able to apply BS or WBS based approaches directly because there is no estimator in our framework so far.

One solution is to incorporate an external estimator. For example, consider the U-statistics $T(s) = \sum_{i=1}^s\sum_{j=s+1}^n h(X_i, X_j), s = 1,\ldots,n-1$ where h is the anti-symmetric kernel used in (1.3). It can be shown that for each segment $m_k \leqslant s-1 < s \leqslant m_{k+1}$

$$\mathbb{E}T(s) - \mathbb{E}T(s-1) = \sum_{j=m_{k+1}+1}^n \mathbb{E}h(X_s, X_j) - \sum_{i=1}^{m_k}\mathbb{E}h(X_i, X_s) = const.$$

In other words, within each segment $(m_k, m_{k+1}]$, $\mathbb{E}T_l(s)$ is monotone ($l = 1,\ldots,p$). As a result, $\max_{1\leqslant s\leqslant n-1}|\mathbb{E}T(s)|_\infty$ is always attained at one change point. Therefore, the estimator

$$\hat{m} = \text{argmax}_{1\leqslant s\leqslant n-1}|T(s)|_\infty$$

can play a role in BS type approach. Similar ideas are discussed in [49, 28, 27, 11] as applications using U-statistics for estimation of change points. Though it is fascinating to investigate the consistency of a BS algorithm that combines estimation using $\hat{m}$ and our bootstrapping test using T_n, the focus and main contribution of this paper is to perform a test without visiting each point. So we leave this algorithm as an open question for future analysis.

Another solution is to adopt the randomization idea from WBS to conduct inference in the presence of multiple change points. One can independently sample B_W intervals that are wider than a pre-specified length n' and obtain a set of (scaled) test statistics on each interval. Denote the set as $\mathcal{T}_W(X_1^n)$. For a given level α, we then perform the proposed bootstrap test on the interval whose corresponding (scaled) test statistic achieves the $(1-\alpha)$-th quantile of $\mathcal{T}_W(X_1^n)$. If the bootstrap test rejects H_0 under level α, then it implies a change point in this interval, which in turn concludes H_1' of at least one change point. The WBS-type test is summarized in Algorithm 1.

Algorithm 1 WBS-type testing (α, n') against multiple change points

1: Draw B_W random intervals $[s_b, e_b], b = 1, \ldots, B_W$, where start- and end-points are taken independently and uniformly from $\{1, \ldots, n\}$ such that $e_b - s_b > n'$.
2: Denote $\mathcal{T}_W(X_1^n) = \{\max_{1 \leqslant k \leqslant d} |T_{e_b - s_b}(X_{s_b}^{e_b})|_k, b = 1, \ldots, B_W\}$, where

$$T_{e_b - s_b}(X_{s_b}^{e_b}) = (e_b - s_b)^{1/2} \binom{e_b - s_b}{2}^{-1} \sum_{s_b \leqslant i < j \leqslant e_b} h(X_i, X_j)$$

is our U-statistic on each interval.
3: Let $q_{\overline{T}_W | X_1^n}(1-\alpha)$ be the $(1-\alpha)$-th quantile of $\mathcal{T}_W$ and b' be the corresponding index.
4: Perform our bootstrap test on $[s_{b'}, e_{b'}]$.
5: **if** our bootstrap test is significant at level α **then**
6: reject H_0.
7: **else**
8: reject H_1'.
9: **end if**

Note that the tuning parameter of n' bounds the length of randomly selected intervals from below. If n' is too small, for instance $n' = 1$, then Step 4 is likely to end up with a very small interval $[e_{b'}, s_{b'}]$. Since approximating $\mathrm{Cov}_{jj}(T_{e_b - s_b})$ on small intervals $\{[s_b, e_b]\}$ will not be consistent, it can lead to the failure of size control under H_0. In practice, one may select n' by applying the Algorithm 1 on $\{\epsilon_i X_i\}_{i=1}^n$, where the multipliers $\epsilon_i, i = 1, \ldots, n$ are i.i.d. standard Gaussian random variables that are independent of X_1^n. Since $\mathbb{E}(\epsilon_i X_i \mid X_1^n) = 0$ and $\mathrm{Cov}(\epsilon_i X_i \mid X_1^n) = X_i X_i^T$, the transformed data $\{\epsilon_i X_i\}_{i=1}^n$ can mimic H_0 without any structural assumption. Simulation result for Algorithm 1 is presented in Section A.5 in the Appendix.

4.4. Backward detection approach for change points estimation

As shown in aforementioned forward searching solutions, the drawbacks of BS include cancellation of signals and requirement of change point estimators. Instead of repeatedly splitting intervals after each detection of change point, we can reversely merge consecutive segments in a backward detection way [47, Section 3.2.2]. Then, our test can work as a stopping rule.

Precisely, denote the initial partition of data segments as $b_0^{(0)} = 0 < b_1^{(0)} < b_2^{(0)} < \cdots < b_{\nu_0 - 1}^{(0)} < n = b_{\nu_0}^{(0)}$ and the corresponding data blocks as $\mathcal{B}^{(0)} =$

$\{B_1^{(0)}, B_2^{(0)}, \ldots, B_{\nu_0}^{(0)}\}$, where $B_i^{(0)} = \{X_{b_{i-1}^{(0)}+1}, \ldots, X_{b_i^{(0)}}\}$. For each pair of consecutive blocks $\{B_i^{(0)}, B_{i+1}^{(0)}\}, i = 1, \ldots, \nu_k - 1$, we can compute a *Dissimilarity Index* based on T_n using truncated data sequence, i.e.

$$DI_i = |T_n(B_i^{(0)} \cup B_{i+1}^{(0)})|_\infty = \max_{1 \leqslant k \leqslant d} \left| (b_{i+1}^{(0)} - b_{i-1}^{(0)})^{1/2} \binom{b_{i+1}^{(0)} - b_{i-1}^{(0)}}{2}^{-1} \sum_{b_{i-1}^{(0)}+1 \leqslant i < j \leqslant b_{i+1}^{(0)}} h_k(X_i, X_j) \right|. \quad (4.4)$$

Since each component of T_n is the standardized Hodges-Lehmann type estimator of location shift in each dimension, large DI_i indicates strong dissimilarity between $B_i^{(0)}$ and $B_{i+1}^{(0)}$. Therefore, we can pick the pair of data blocks with the smallest DI and perform our bootstrapped test to decide whether to merge them. If the test fails to reject the null hypothesis of no change point, we merge the two blocks into one. Otherwise, we move on to test the next pair of data blocks with the second smallest DI. The process will continue until no blocks can be merged. The Backward Detection (BD) algorithm is summarized in Algorithm 2.

Algorithm 2 Backward Detection: BD($\mathcal{B}^{(k)}$)

1: Start from data blocks as $\mathcal{B}^{(k)} = \{B_1^{(k)}, B_2^{(k)}, \cdots, B_{\nu_k}^{(k)}\}$
2: Compute the Dissimilarity Index $DI_i = T_n(B_i^{(k)}, B_{i+1}^{(k)})$ as in (4.4) for $i = 1, \ldots, \nu_k - 1$
3: Let $i^* = \text{argmin} DI_i$.
4: **if** our bootstrap test rejects the null for the segment $[b_{i^*-1}^{(k)}, b_{i^*+1}^{(k)}]$ **then**
5: Repeat the test for i^* referring to the next smallest DI_i until all pairs are examined
6: **else**
7: Update $B_i^{(k+1)} = B_i^{(k)}$ for $i < i^*$
8: Merge $B_{i^*}^{(k)}, B_{i^*+1}^{(k)}$ into one block $B_{i^*}^{(k+1)} = B_{i^*}^{(k)} \cup B_{i^*+1}^{(k)}$
9: Set $B_i^{(k+1)} = B_{i+1}^{(k)}$ for $i > i^*$
10: Perform BD($\mathcal{B}^{(k+1)}$)
11: **end if**
12: **return** Estimated blocks $\mathcal{B}$ and corresponding segmentation $\hat{m}_1, \ldots, \hat{m}_{\hat{\nu}}$

Compared to forward detection, BD is able to detect short sequence. Hence, the Backward Detection algorithm will be more powerful compared to the direct extension or the block testing at the beginning of this section. There is no signal cancellation issue for BD. Besides, it can identify change points without introducing new estimators or statistics. However, there is a risk of Type I error inflation since BD recursively performs testing procedure. Let $b_i^{(0)} = iM, i = 1, \ldots, \lfloor n/M \rfloor$, where $\lfloor n/M \rfloor$ is the largest integer not exceeding n/M. Then small M can cause over rejection, while large M may affect estimation accuracy and bring signal cancellation issue back. We should tune the initial partition size M carefully. To the best of our knowledge, there is no theoretical result on the consistency of backward detection in change point estimation. For testing purpose, we can take M as discussed in Section 4.2. Empirical performance are investigated in simulation and real data application.

5. Simulation study

In this section, we first report simulation results of our method in size approximation and power performance under single change point model. Independent random vectors are generated according to the location-shift model (1.1). Comparison with other methods follows. In the end, we evaluate the global test of direct extension and the Backward Detection of estimation for multiple change points.

5.1. Simulation setup

We generate i.i.d. ξ_i from the following distributions.

1. Multivariate Gaussian distribution: $\xi_i \sim N(0, V)$.
2. Multivariate elliptical t-distribution with degree of freedom ν ($\nu > 2$): $\xi_i \sim t_\nu(V)$ with the probability density function [46, Chapter 1]

$$f(x;\nu,V) = \frac{\Gamma(\nu+p)/2}{\Gamma(\nu/2)(\nu\pi)^{p/2}\det(V)^{1/2}}\left(1+\frac{x^\top V^{-1}x}{\nu}\right)^{-(\nu+p)/2}.$$

The covariance matrix of ξ_i is $\Sigma = \frac{\nu}{\nu-2}V$. In our simulation, we use $\nu = 6$.

3. Contaminated Gaussian distribution (i.e., Gaussian mixture model): $\xi_i \sim \text{ctm-G}(\varepsilon,\nu,V) = (1-\varepsilon)N(0,V) + \varepsilon N(0,\nu^2 V)$ with the probability density function

$$f(x;\varepsilon,\nu,V) = \frac{1-\varepsilon}{(2\pi)^{p/2}\det(V)^{1/2}}\exp\left(-\frac{x^\top V^{-1}x}{2}\right) + \frac{\varepsilon}{(2\pi\nu^2)^{p/2}\det(V)^{1/2}}\exp\left(-\frac{x^\top V^{-1}x}{2\nu^2}\right).$$

The covariance matrix of ξ_i is $\Sigma = [(1-\varepsilon)+\varepsilon\nu^2]V$. We set $\varepsilon = 0.2$ and $\nu = 2$.

4. Scale transformation of Cauchy distribution: $\xi_i = V^{1/2}\xi_i'$, where $\xi_i' = (\xi_{i1}', \ldots, \xi_{ip}')^T$ and ξ_{ij}' are i.i.d. standard (univariate) Cauchy distribution.

For each distribution, we consider three spatial dependence structures of V.

(I) Independent: $V = \text{Id}_p$, where Id_p is the $p \times p$ identity matrix.
(II) Strongly dependent: $V = 0.8J + 0.2\text{Id}_p$, where J is the $p \times p$ matrix of all ones.
(III) Moderately dependent: $V_{ij} = 0.8^{|i-j|},\ i,j = 1,\ldots,p$.

Unless explicitly indicated, $B = 200$ bootstrap samples are drawn for each testing procedure and all results are averaged on 500 simulations. We fix the sample size $n = 500$ and dimension $p = 600$ for single change point scenario and focus on the performance of two kernels: the linear kernel $h(x,y) = x - y$ and the sign kernel $h(x,y) = \text{sign}(x-y)$.

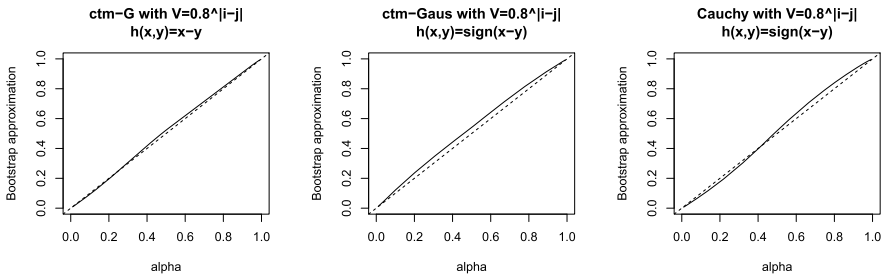


FIG 1. *Selected setups for comparing* $\hat{R}(\alpha)$ *along with* α. *See headlines for corresponding distribution and kernel.*

5.2. Size approximation

Let $\hat{R}(\alpha)$ be the proportion of empirically rejected null hypothesis at significance level $\alpha \in (0,1)$. There are several observations we can draw from Table 1, which shows the empirical uniform error-in-size, $\sup_{\alpha\in(0,1)} |\hat{R}(\alpha)-\alpha|$. First, the dependence structure of V does not influence the errors remarkably. Second, for Gaussian, t_6 and contaminated Gaussian (ctm-G) distributions, the two kernels have very similar errors in size. For the Cauchy distribution which is only applicable for the sign kernel, error-in-size is comparable with the other three distribution settings. Therefore, we conclude that under H_0, the sign kernel gains robustness without losing much accuracy. Three example curves are displayed additional in Figure 1 to visualize the size approximation.

TABLE 1
Uniform error-in-size under H_0.

$\sup_{\alpha\in(0,1)} \lvert\hat{R}(\alpha)-\alpha\rvert$		linear kernel			sign kernel			
		Gaussian	t_6	ctm-G	Gaussian	t_6	ctm-G	Cauchy
I	$V=\mathrm{Id}_p$	0.034	0.086	0.040	0.026	0.066	0.032	0.028
II	$V=0.8J+0.2\mathrm{Id}_p$	0.054	0.020	0.058	0.064	0.040	0.050	0.060
III	$V_{ij}=0.8^{\lvert i-j\rvert}$	0.026	0.048	0.040	0.040	0.036	0.060	0.058

We also compare our test using the linear kernel to the CUSUM counterpart in [61, `BABS`] under the same setting with the boundary removal parameter as $\underline{s} = 40$. Table 2 displays corresponding simulation results. By comparing it to Table 1, we observe that the CUSUM approach suffers from greater size distortion as it has larger uniform errors in general. When we focus on the maximum error within the interval $\alpha \in (0, 0.1]$ (that are common choices in real applications), our linear kernel based algorithm still outperforms. In addition, our test demands no more computational costs and it enjoys flexibility of no tuning parameter.

TABLE 2
Error-in-size $\sup_\alpha |\hat{R}(\alpha) - \alpha|$ *for* $\alpha \in (0,1)$ *and* $\alpha \in (0, 0.1]$

	$\sup_{\alpha\in(0,1)} \lvert\hat{R}(\alpha)-\alpha\rvert$			$\sup_{\alpha\in(0,0.1]} \lvert\hat{R}(\alpha)-\alpha\rvert$					
	CUSUM approach			CUSUM approach			linear kernel		
	Gaussian	t_6	ctm-G	Gaussian	t_6	ctm-G	Gaussian	t_6	ctm-G
I	0.072	0.122	0.096	0.040	0.036	0.064	0.012	0.010	0.020
II	0.066	0.044	0.048	0.026	0.014	0.024	0.008	0.014	0.012
III	0.074	0.092	0.066	0.022	0.038	0.048	0.020	0.018	0.012

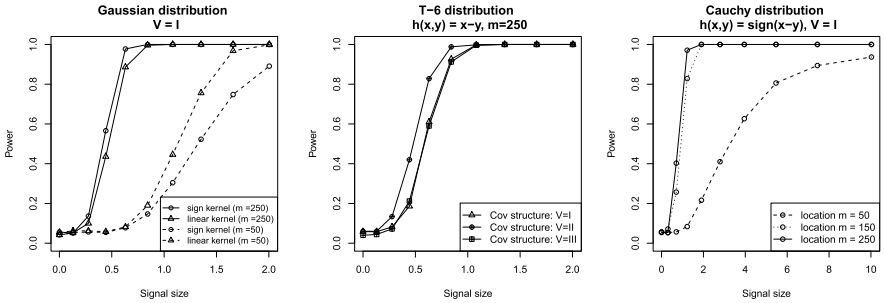


FIG 2. *Selected setups for comparing power curves. See headlines and legends for corresponding distribution, kernel, covariance structures and change point location* m.

5.3. Power of the bootstrap test

Under H_1, the signal vector is chosen as $\theta = (\theta_1, 0, \ldots, 0)^T$ such that $\theta_1 = |\theta|_\infty$. We vary the change point location $m = 50, 150, 250$. Figure 2 displays the power curves for different kernels, change point location m and dependence structure V. The left panel investigates kernel and location impact. Change point at center $m = n/2 = 250$ (solid curves) is easier to detect than that of $m = n/10 = 50$ at boundary (dashed curves) regardless of the choice of kernel. For standard Gaussian distribution, the linear kernel has greater power than the sign kernel when the change occurs at boundary point $m = 50$, but the relation reverses when $m = 250$. The middle panel uses linear kernel as an example to illustrate the observation that the dependence structure V does not significantly influence the power, though our ℓ^∞-type test statistic has advantage in the strong dependence case. The right panel displays the power of the sign kernel for Cauchy distributed data to highlight its robustness to location parameter θ and the impact from change point position m. Regarding to the exact power values, see Table 11 (linear kernel) and 12 (sign kernel) in Appendix.

5.4. Comparison with other methods

We compare our U-statistic approach to other competing algorithms in change point literature. The linear and sign kernels of our approach are used. All of the four competitors, namely [61, `BABS`], [38, `Jirak`], [20, `SBS`] and [57, `Inspect`], are based on CUSUM statistics. Among them, `BABS` and `Jirak` are ℓ^∞-type

bootstrap test for single change point using different weights on $(s(n-s)/s)$ in (1.5), the latter of which needs cross-sectional variance estimation on each dimension and it is sensitive to mean shift near the center of data sequence. The last two competitors target on multiple change point estimation where `SBS` is thresholded ℓ^1-type estimator and `Inspect` is projection based. We adopt their single change point version function in corresponding R packages and convert them to tests using their default threshold computing functions. In our simulation, we set $n = 500, p = 600, \alpha = 0.05, m = 150$, and set boundary removal as 40 for `BABS`, `Jirak` and `SBS`.

Table 3 compares the power of different tests when the signal θ_1 is growing. It is clear that `SBS` and `Inspect` are not suitable in our setting since the location shift parameter is extremely sparse. When the data generating mechanism is not standard multivariate Gaussian (i.e. not Gaussian-I in the table), these two algorithms trigger excessive false alarms when $\theta = 0$ and do not return monotone powers as θ increase. The other two competitors `BABS` and `Jirak` behave similarly and return slightly higher powers than ours in general. Note that these two approaches need to pick boundary removal parameter, which can harm powers if it is too large to include true m in the working interval. The contrasts between linear and sign kernel have been discussed in the previous part. Therefore, Table 3 indicates that our method, which enjoys tuning-free and intermediate-estimation-free properties, is competent in empirical studies.

TABLE 3

Powers for our method using linear and sign kernels, [61, `BABS`], [38, `Jirak`], [20, `SBS`] and [57, `Inspect`].

$\|\theta\|_\infty$	Gaussian-I						Gaussian-II					
	linear	sign	`BABS`	`Jirak`	`SBS`	`Inspect`	linear	sign	`BABS`	`Jirak`	`SBS`	`Inspect`
0	0.030	0.049	0.042	0.061	0.764	0.020	0.042	0.037	0.056	0.052	0.092	0.833
0.28	0.088	0.070	0.087	0.110	0.836	0.021	0.216	0.154	0.209	0.232	0.264	0.724
0.44	0.414	0.342	0.502	0.553	0.928	0.006	0.738	0.619	0.756	0.828	0.744	0.458
0.63	0.890	0.830	0.966	0.967	0.976	0.001	0.996	0.982	0.996	0.999	0.926	0.287
0.84	0.998	0.992	1	1	0.966	0.003	1	1	1	1	0.906	0.205
1.08	1	1	1	1	0.972	0.093	1	1	1	1	0.898	0.183
1.35	1	1	1	1	0.954	0.789	1	1	1	1	0.858	0.287
1.66	1	1	1	1	0.938	0.999	1	1	1	1	0.838	0.997
2.00	1	1	1	1	0.936	1	1	1	1	1	0.834	1
$\|\theta\|_\infty$	ctm-Gaussian-I						t_6-II					
	linear	sign	`BABS`	`Jirak`	`SBS`	`Inspect`	linear	sign	`BABS`	`Jirak`	`SBS`	`Inspect`
0	0.030	0.051	0.020	0.067	0.592	1	0.060	0.068	0.044	0.053	0.060	0.975
0.28	0.036	0.073	0.033	0.076	0.630	1	0.124	0.148	0.109	0.132	0.108	0.942
0.44	0.150	0.189	0.186	0.245	0.752	1	0.418	0.451	0.477	0.537	0.418	0.791
0.63	0.524	0.593	0.675	0.750	0.904	1	0.878	0.912	0.919	0.936	0.856	0.629
0.84	0.940	0.941	0.977	0.987	0.954	1	0.998	1	0.997	1	0.928	0.507
1.08	1	1	0.999	1	0.946	1	1	1	1	1	0.898	0.453
1.35	1	1	1	1	0.938	1	1	1	1	1	0.878	0.609
1.66	1	1	1	1	0.918	1	1	1	1	1	0.846	1
2.00	1	1	1	1	0.902	1	1	1	1	1	0.864	1

For fair comparison, we do not use Cauchy distribution, since all methods, except for our sign kernel method, will fail when there is no well-defined mean

parameter in the heavy tailed distribution. Unreported results show that `SBS` and `Inspect` perform better when the mean change is denser. We also remark that the Double CUSUM Binary Segmentation [19, `DCBS`] cannot detect any change point under our setting when $|\theta|_\infty \leqslant 2$ because the setup is an extremely sparse case, so the table does not include it.

Section A.6 in Appendix presents some further comparison for the size control of `BABS`, `Jirak` and our linear kernel approach under H_0 with fixed p and boundary removal fraction while varying the sample size n.

5.5. Multiple change-point detection

In the multiple change-point scenario, we first let the k-th component of $\theta^{(k)}$ to have the same location shift, i.e. $\theta_1^{(1)} = \theta_2^{(2)} = \cdots = \theta_{n,\nu}^{(\nu)} = \delta \neq 0$. Since change point estimation can be viewed as a special case of clustering, the accuracy can be measured by the adjusted Rand index (ARI) [50, 36]. We also report average ARI over all 500 runs. The bootstrap resampling is 200.

To start with, we consider the direct application of our test using Gaussian distribution and linear kernel as a representative. Let $n = 1000, p = 1200$, $\alpha = 0.05$, and the two change points $(m_1, m_2) = (300, 600)$. The powers are shown in Table 4. Our test works well as there is no signal cancellation.

TABLE 4
Powers under multiple change point scenario using linear kernel. Here, $(m_1, m_2) = (300, 600)$.

	δ	0	0.317	0.733	1.282	2.004
Spacial	I	0.052	0.278	1	1	1
dependent	II	0.064	0.510	1	1	1
structures	III	0.070	0.222	0.996	1	1

Next, we apply the Backward Detection algorithm to estimate change points. We set the initial data blocks as segments of every $M = 100$ data points and take the Gaussian distribution with moderate dependence structure (III) for instance. The estimated change points are summarized in Table 5 (counts and ARIs) and Figure 3 (estimates). When signal $\delta = 0.317$ is small, BD fails to reject H_0 in about half of the time (276 out of 500) and it cannot locate the shifts accurately (small ARIs). However, as signal gets larger, both the number and the locations of change points can be detected consistently (under proper setup of initial data blocks). Meanwhile, ARIs are also increasing to 1, which stands for the perfect estimation. We further add one more change where $(m_1, m_2, m_3) = (300, 600, 800)$. The results in Table 5 and Figure 3 are similar to that of two change point case.

Then, we also use the sign kernel to detect location shift for Cauchy distribution with dependence structure (III). Analogously, initial data blocks are segments of every $M = 100$ data points in sequence. The cases of 2 change points $(m_1, m_2) = (300, 600)$ and 3 change points $(m_1, m_2, m_3) = (300, 600, 800)$ are implemented and the results are shown in Table 6 and Figure 4. Similar con-

TABLE 5
Estimation of multiple change points for $M = 100$: counts and ARIs. Here, the data is Gaussian distributed with dependence structure (III) and the linear kernel is used.

		$(m_1, m_2) = (300, 600)$					$(m_1, m_2, m_3) = (300, 600, 800)$				
δ		0	0.317	0.733	1.282	2.004	0	0.317	0.733	1.282	2.004
Estimated number of change points	0	**497**	276	0	0	0	**494**	270	0	0	0
	1	3	209	0	0	0	6	217	0	0	0
	2	0	**15**	**484**	**492**	**483**	0	13	32	0	0
	3	0	0	16	7	17	0	**0**	**455**	**474**	**483**
	4	0	0	0	1	0	0	0	13	25	17
	5	0	0	0	0	0	0	0	0	1	0
Sum		500	500	500	500	500	500	500	500	500	500
ARI		0.994	0.195	0.933	0.998	0.996	0.988	0.152	0.920	0.995	0.997

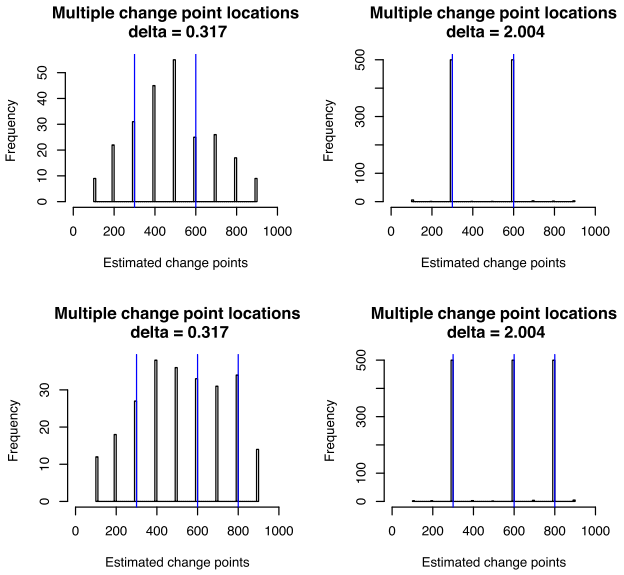


FIG 3. *Multiple change point setup using linear kernel at signal level $\delta = 0.317, 2.004$. Upper: 2 change points $(m_1, m_2) = (300, 600)$. Lower: 3 change points $(m_1, m_2, m_3) = (300, 600, 800)$.*

clusion can be drawn except that stronger signal strength is required as Cauchy distribution has extremely heavy tails.

Lastly, we set $M = 1$ and repeat the experiment using linear kernel and Gaussian distribution with dependence structure (III). The results are summarized in Table 7 and Figure 5. Compared to Table 5 and Figure 4 which correspond to the same setting but $M = 100$, we can easily observe over rejection issue since more change points are concluded than the truth for both cases. However, when signal is large ($\delta = 2.004$), estimated change points still concentrate around the true m_i's. In practice, a threshold $\underline{m}$ can be introduced to force merging two blocks if the cardinality of their union is small.

TABLE 6
Estimation of multiple change points for $M = 100$. Here, the data is Cauchy distributed with dependence structure (III) and the sign kernel is used.

		$(m_1, m_2) = (300, 600)$					$(m_1, m_2, m_3) = (300, 600, 800)$				
δ		0	0.822	2.320	5.050	10.023	0	0.822	2.320	5.050	10.023
Estimated number of change points	0	**465**	44	0	0	0	**460**	36	0	0	0
	1	6	257	0	0	0	11	221	0	0	0
	2	6	**173**	**365**	**470**	**470**	4	172	0	0	0
	3	6	9	18	12	10	3	**50**	**401**	**470**	**477**
	4	5	11	21	15	12	8	9	19	11	8
	5	6	1	59	1	1	5	6	66	1	0
	6	6	5	46	2	7	9	6	14	18	15
Sum		500	500	500	500	500	500	500	500	500	500
ARI		0.930	0.557	0.888	0.986	0.983	0.920	0.495	0.951	0.986	0.989

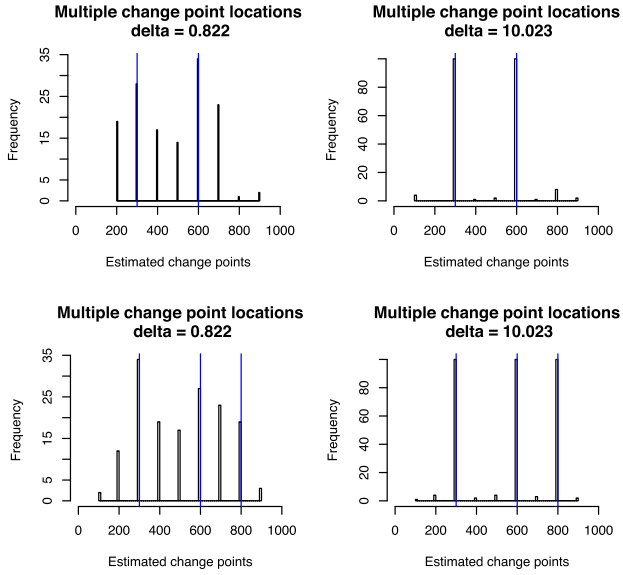


FIG 4. *Multiple change point setup using sign kernel at signal level $\delta = 0.822, 10.023$. Upper: 2 change points $(m_1, m_2) = (300, 600)$. Lower: 3 change points $(m_1, m_2, m_3) = (300, 600, 800)$.*

5.6. Simulation results for time series data

We shall study the empirical performance of the bootstrap test for some dependent process ξ_i. In our simulation, we consider the stationary vector autoregression of order 1 (denote as VAR(1)) error process: $\xi_i = A\xi_{i-1} + \eta_i = \sum_{k=0}^{\infty} A^k \eta_{i-k}$, where $\{\eta_i\}_{i\in\mathbb{Z}}$ is a sequence of i.i.d. mean-zero random vectors in $\mathbb{R}^p$ and A is a $p \times p$ coefficient matrix, where random matrix A is generated with i.i.d. $N(0, 1)$ entries. To ensure the stationarity of ξ_i process, A is normalized such that $\|A\|_2 = 1/1.8 < 1$. The distribution and covariance defined in Section 5.1.

TABLE 7
Estimation of multiple change points for $M = 1$. Here, the data is Gaussian distributed with dependence structure (III) and linear kernel is used.

		$(m_1, m_2) = (300, 600)$					$(m_1, m_2, m_3) = (300, 600, 800)$				
δ		0	0.317	0.733	1.282	2.004	0	0.317	0.733	1.282	2.004
Estimated number of change points	0	**475**	205	0	0	0	477	195	0	0	0
	1	21	230	3	0	0	20	224	0	0	0
	2	3	**59**	**367**	**343**	**344**	3	77	51	0	0
	3	1	5	114	135	133	0	**4**	**324**	**289**	**293**
	4	0	1	16	22	23	0	0	111	167	172
	5	0	0	0	0	0	0	0	13	38	32
	6	0	0	0	0	0	0	0	1	6	2
	8	0	0	0	0	0	0	0	0	0	1
	Sum	500	500	500	500	500	500	500	500	500	500
	ARI	0.950	0.186	0.634	0.785	0.858	0.954	0.160	0.582	0.747	0.834

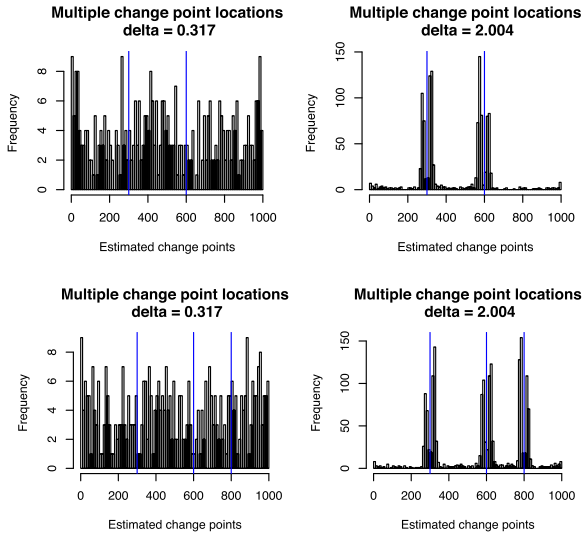


FIG 5. *Multiple change point setup using $M = 1$ and linear kernel at signal level $\delta = 0.317, 2.004$. Upper: 2 change points $(m_1, m_2) = (300, 600)$. Lower: 3 change points $(m_1, m_2, m_3) = (300, 600, 800)$.*

We first use the linear kernel $h(x, y) = x - y$ and consider different trimming parameters $M = 2, 5, 10, 15$. We fix $n = 500, p = 600, B = 200$ and $m = n/2$ under the location-shift model of single change point. Let $\hat{R}(\alpha)$ be the proportion of empirically rejected null hypothesis in 500 simulations. In Table 8 which provides uniform error-in-size $\sup_{\alpha \in [0,1]} |\hat{R}(\alpha) - \alpha|$, we can observe that larger M needs to be selected if stronger dependence (i.e., the compound symmetry structure II) presents regardless of distribution families. This is the trade-off effect through M. We can also find that the best error-in-sizes in each column are comparable to the corresponding values in Table 1. This indicates the effectiveness of our modified approach under temporal dependency. Figure 6 displays

two examples of $\hat{R}(\alpha)$ under H_0 and power under H_1, where the signal vector is chosen as $\theta = (\theta_1, 0, \ldots, 0)^T$ such that $\theta_1 = |\theta|_\infty$.

Next we use sign kernel $h(x, y) = \text{sign}(x - y)$ and consider the trimming parameters $M = 2, 5, 10$. For illustration purpose, we only select the data-generating schemes of Cauchy distribution with Covariance I-III and ctm-Gaussian distribution with Covariance III. The other parameters remain the same as above. The uniform error-in-size for each scenario is give in Table 9. In general, $M = 2$ works the best under each scenario. The non-linear projection by the sign kernel makes the correlation between data pairs weaker. Therefore, it makes the sign kernel more attractive in terms of its robustness against weak temporal dependency. Similarly, two examples are given in Figure 7.

TABLE 8

Uniform error-in-size for linear kernel under H_0, where ξ_i are from VAR(1) process. The columns and rows display the distribution of η_i as defined in Section 5.1 and different trimming parameter M, respectively. The smallest errors in each column are highlighted.

$\sup_{\alpha\in[0,1]} \lvert\hat{R}(\alpha) - \alpha\rvert$	Gaussian			t_6			ctm-Gaussian		
	I	II	III	I	II	III	I	II	III
M=2	**0.058**	0.092	**0.030**	**0.054**	0.082	**0.028**	**0.038**	0.086	**0.054**
M=5	0.076	0.088	0.056	0.092	**0.074**	0.050	0.058	**0.082**	0.086
M=10	0.128	**0.064**	0.102	0.134	0.080	0.080	0.106	0.084	0.136
M=15	0.180	0.066	0.150	0.172	0.086	0.126	0.156	0.094	0.174

TABLE 9

Uniform error-in-size for sign kernel under H_0, where ξ_i are from VAR(1) process. The columns and rows disply the distribution of η_i as defined in Section 5.1 and different trimming parameter M, respectively. The smallest errors in each column are highlighted.

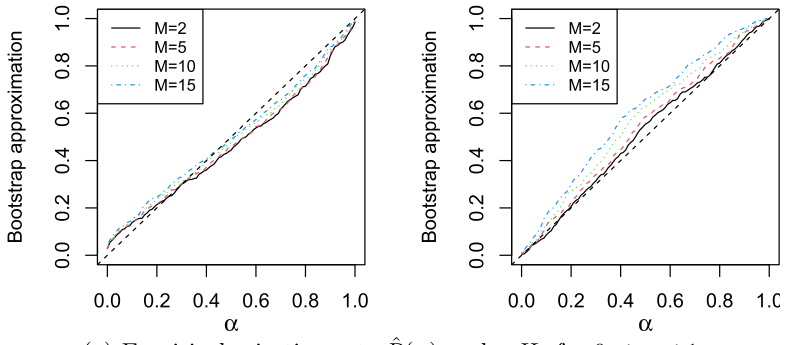
$\sup_{\alpha\in[0,1]} \lvert\hat{R}(\alpha) - \alpha\rvert$	Cauchy (I)	Cauchy (II)	Cauchy (III)	ctm-Gaussian (III)
M=2	**0.068**	**0.057**	**0.068**	**0.060**
M=5	0.094	0.062	0.096	0.088
M=10	0.144	0.078	0.150	0.142

6. Real Data Applications

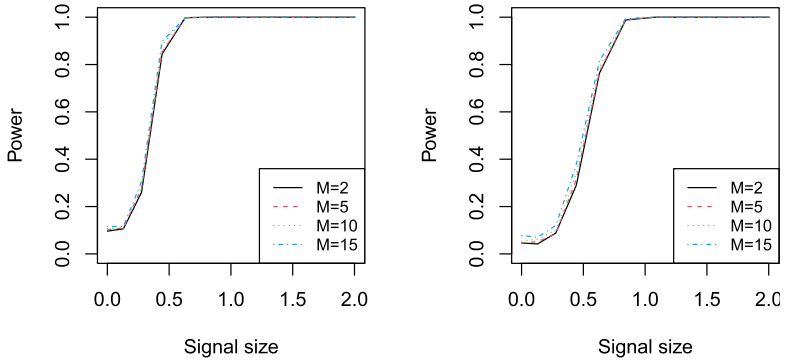
6.1. Single change point: Enron email dataset

Enron Corporation used to be one of the leading American energy companies. In an accounting scandal, Enron share prices decreased from around $80 during the summer of 2000 to pennies at the end of 2001. The bankruptcy was filed on 12/02/2001 and it became the largest bankruptcy reorganization in American history at that time. The Enron email dataset that contains more than 500,000 messages from about 150 users (mostly senior management) was publicly available during the investigation by the Federal Energy Regulatory Commission in 2002. [1]

[1] The raw data is organized in folders (http://www.cs.cmu.edu/~enron/) and its tabular format version is available at https://data.world/brianray/enron-email-dataset. The timeline of major events can be found at http://www.agsm.edu.au/bobm/teaching/BE/Enron/timeline.html.



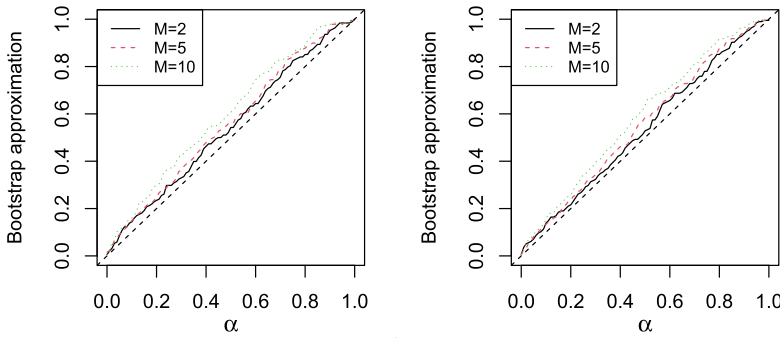
(a) Empirical rejection rate $\hat{R}(\alpha)$ under H_0 for $0 \leqslant \alpha \leqslant 1$.



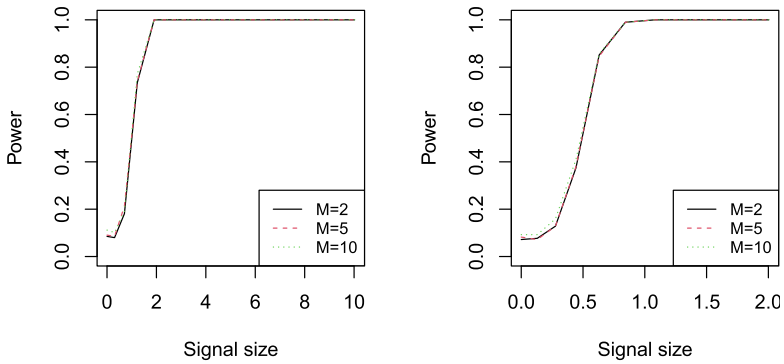
(b) Power under H_1 for $0 \leqslant \theta_1 \leqslant 2$.

FIG 6. *Empirical rejection rate $\hat{R}(\alpha)$ under H_0 and power under H_1 in selected time-series data-generating schemes: (Left) Gaussian distribution with covariance structure II; (Right) ctm-Gaussian distribution with covariance structure III. (Parameters: $n = 500$, $p = 600$, kernel $h(x, y) = x - y$, and trimming parameters $M = 2, 5, 10, 15$.)*

We study the collection of messages sent in 2000-2001. To test for the existence of an abrupt changes in email discussions, our analysis is based on the number of emails sent from each user. In order to exclude the yearly trend and temporal dependence, we apply our method to X_{ij} which is the difference of emails sent from user j on the i-th day for the two years. The leap day (02/29/2000) and the users who were inactive during 2000 or 2001 are removed such that the final data matrix $(X_{ij})_{i=1,\ldots,n;j=1,\ldots,p}$ is of dimension $n = 365$ and $p = 101$. We set bootstrap repetition number $B = 2000$. For the linear kernel, our test statistic has the value $\overline{T}_n = 561.49$ and the 95% quantile of bootstrapped statistic is 117.17. For the sign kernel, our test statistic has the value $\overline{T}_n = 8.95$ and the 95% quantile of bootstrapped statistic is 1.44. Both tests reject the null hypothesis of no abrupt change. As an illustration of the test results, the aggregated trend of $Y_i = \sum_{j=1}^{101} X_{ij}$ in Figure 8 indicates the presence of extensive email communication from the second half of 2000 to the



(a) Empirical rejection rate $\hat{R}(\alpha)$ under H_0 for $0 \leqslant \alpha \leqslant 1$.



(b) Power under H_1 for $0 \leqslant \theta_1 \leqslant 2$.

FIG 7. *Empirical rejection rate $\hat{R}(\alpha)$ under H_0 and power under H_1 in selected time-series data-generating schemes: (Left) Cauchy distribution with covariance structure II; (Right) ctm-Gaussian distribution with covariance structure III. (Parameters: $n = 500$, $p = 600$, kernel $h(x, y) = sign(x - y)$, and trimming parameters $M = 2, 5, 10$.)*

first half of 2001. Our test confirms that there was abnormal email activity in these two years.

6.2. Multiple change point: micro-array dataset

The array comparative genomic hybridization data, `ACGH` [37, R package `ecp`], consists of $p = 43$ patients with bladder tumor. We consider to detect change points among their DNA copy number profiles each of which contains $n = 2215$ log-intensity-ratio fluorescent measurements. We apply the BD algorithm using linear kernel and set bootstrap repeats 1000, significance level $\alpha = 0.01$ and initial data block size $M = 2$. The measurements for the first 10 individuals are shown in Figure 9. Our BD algorithm finds 32 change points that marked in red vertical dashed lines. This number is in a reasonable level as indicated in [57] where the authors only reported 30 most significant ones while their de-

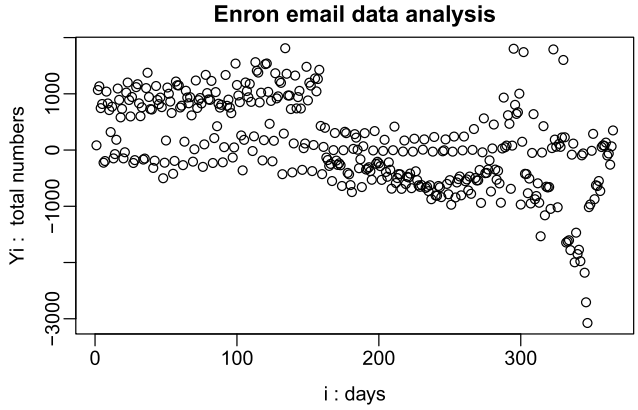


FIG 8. *Trend of* $Y_i = \sum_{j=1}^{101} X_{ij}$ *for Enron email dataset.*

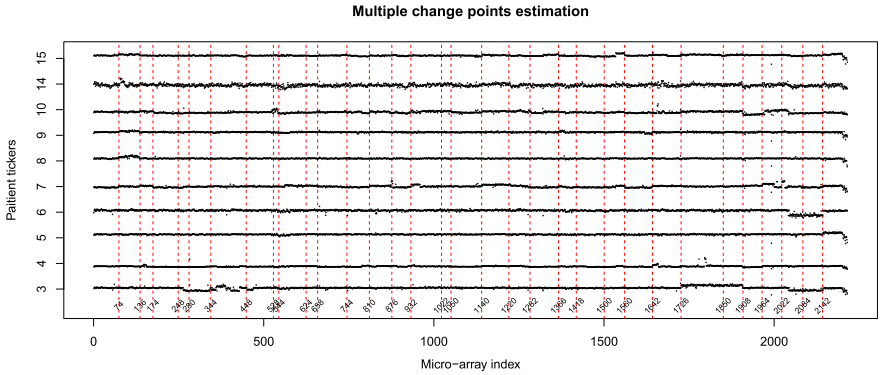


FIG 9. *Real data study: aCGH data. Here, we set* $B = 1000, \alpha = 0.01$ *and the linear kernel.*

fault `Inspect` algorithm found 254 change points. The ARI between ours and the bootstrap-assisted binary segmentation [61, `BABS`] which identifies 27 change points is 0.779. As shown in Table 10, the two methods have overlapped detection that are close loci numbers such as $(73, 74), (342, 344), \ldots, (2143, 2142)$.

TABLE 10
Identified change point locations (loci numbers on genome) in `ACGH` *dataset.*

BABS	73, 185, 263, 342, 428, 521, 581, 657, 741, 801, 871, 960, 1051, 1141, 1216, 1276, 1367, 1427, 1503, 1563, 1664, 1724, 1836, 1905, 1965, 2044, 2143.
BD	74, 136, 174, 248, 280, 344, 448, 528, 544, 624, 658, 744, 810, 876, 932, 1022, 1050, 1140, 1220, 1282, 1366, 1418, 1500, 1560, 1642, 1726, 1850, 1908, 1964, 2022, 2084, 2142.

Appendix A: Proofs and additional numeric results

A.1. Proof of main results

Throughout the whole proofs, we assume $d \geqslant 2$, $n \geqslant 3$ and $n \geqslant \log^7(nd)$ otherwise the rates will automatically hold. The $K_i > 0, i = 1, 2, \ldots$ and $C > 0$ are large constants that may vary part by part.

Proof of Theorem 3.1. Suppose H_0 is true. Without loss of generality, we may assume $\varpi_n \leqslant 1$.

Step 1. Gaussian approximation to T_n.

Denote $\Gamma = \mathrm{Cov}(g(X_1))$. Since the kernel h is anti-symmetric, we have $\mathbb{E}[g(X_1)] = \mathbf{0}$. Thus $\mathbb{E}[L_n] = \mathbf{0}$ and

$$\mathrm{Cov}(L_n) = n\binom{n}{2}^{-2} \sum_{i=1}^{n} (n+1-2i)^2 \, \mathrm{Cov}(g(X_i)) = \frac{4(n+1)}{3(n-1)}\Gamma.$$

By Jensen's inequality, we have $\mathbb{E}|g_j(X_i)|^{2+k} \leqslant D_n^k$ for $k = 1, 2$, and $\|g_j(X_i)\|_{\psi_1} \leqslant D_n$. Then it follows

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{2}{n-1}\right)^{2+k} |n-2i+1|^{2+k} \mathbb{E}|g_j(X_i)|^{2+k} \lesssim D_n^k,$$
$$\left\| \frac{2(n-2i+1)}{n-1} g_j(X_i) \right\|_{\psi_1} \lesssim D_n.$$

In addition, note that $\frac{1}{n}\sum_{i=1}^{n} 4\left(\frac{n-2i+1}{n-1}\right)^2 \Gamma_{jj} = \frac{n+1}{n-1} \cdot \frac{4}{3}\Gamma_{jj} \geqslant \frac{4}{3}\underline{b} > 0$. By Proposition 2.1 in [17] (applied to the max-hyperrectangles), we have

$$\rho(\overline{L}_n, \overline{Z}_n) \leqslant \left\{ \frac{D_n^2 \log^7(nd)}{n} \right\}^{1/6} = \varpi_n,$$

where $\overline{Z}_n = \max_{1\leqslant j\leqslant d} Z_{nj}$ and $Z_n \sim N(0, \frac{4(n+1)}{3(n-1)}\Gamma)$. Let $Z \sim N(0, 4\Gamma/3)$. By the Gaussian comparison inequality (cf. Lemma C.5 in [14]), we have

$$\rho(\overline{Z}_n, \overline{Z}) \lesssim \left(\frac{4}{3n}|\Gamma|_\infty \log^2 d\right)^{1/3}.$$

Since $\Gamma_{jj} \leqslant 1 + \mathbb{E}|g_j(X_1)|^3 \leqslant 1 + D_n \leqslant 2D_n$, it follows from the Cauchy-Schwarz inequality that

$$\rho(\overline{Z}_n, \overline{Z}) \lesssim \left(\frac{D_n \log^2 d}{n}\right)^{1/3} \lesssim \varpi_n.$$

Then by triangle inequality, we have

$$\rho(\overline{L}_n, \overline{Z}) \leqslant \rho(\overline{L}_n, \overline{Z}_n) + \rho(\overline{Z}_n, \overline{Z}) \lesssim \varpi_n. \tag{A.1}$$

Applying Corollary 5.6 in [15] with $k = 2$, we have

$$\mathbb{E}\left(\max_{1\leqslant j\leqslant d} |R_{nj}|\right) \lesssim D_n n^{-1/2} \log d. \tag{A.2}$$

Then for any $t \in \mathbb{R}$ and $a > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\overline{T}_n \leqslant t\right) &\leqslant \mathbb{P}\left(\overline{L}_n \leqslant t + a^{-1}\mathbb{E}[|R_n|_\infty]\right) + \mathbb{P}\left(|R_n|_\infty > a^{-1}\mathbb{E}[|R_n|_\infty]\right)\\ &\leqslant_{(i)} \mathbb{P}\left(\overline{L}_n \leqslant t + a^{-1}\mathbb{E}[|R_n|_\infty]\right) + a\\ &\leqslant_{(ii)} \mathbb{P}\left(\overline{Z} \leqslant t + a^{-1}\mathbb{E}[|R_n|_\infty]\right) + C\varpi_n + a\\ &\leqslant_{(iii)} \mathbb{P}\left(\overline{Z} \leqslant t\right) + Ca^{-1}\mathbb{E}[|R_n|_\infty]\log^{1/2} d + C\varpi_n + a\\ &\leqslant_{(iv)} \mathbb{P}\left(\overline{Z} \leqslant t\right) + CD_n a^{-1} n^{-1/2}\log^{3/2} d + C\varpi_n + a,\end{aligned}$$

where step (i) follows from Markov's inequality, step (ii) from the Gaussian approximation error bound (A.1) for the linear part, step (iii) from Nazarov's inequality (cf. Lemma A.1 in [17]), and step (iv) from the maximal inequality (A.2) for the degenerate term. Likewise, we can deduce the reverse inequality

$$\mathbb{P}\left(\overline{T}_n \leqslant t\right) \geqslant \mathbb{P}\left(\overline{Z} \leqslant t\right) - CD_n a^{-1} n^{-1/2}\log^{3/2} d - C\varpi_n - a.$$

Choosing $a = n^{-1/4} D_n^{1/2}\log^{3/4} d$, we get $\rho(\overline{T}_n, \overline{Z}) \leqslant C\varpi_n$.

Step 2. Bootstrap approximation to T_n. Recall the definition of $T_n^\sharp$ in (2.1), $T_n^\sharp | X_1^n \sim N(\mathbf{0}, 4\hat{\Gamma}_n)$ where

$$\hat{\Gamma}_n = \frac{1}{n(n-1)^2}\sum_{i=1}^n \sum_{j=i+1}^n \sum_{k=i+1}^n h(X_i, X_j)h(X_i, X_k)^T. \tag{A.3}$$

By Lemma A.1, $\mathbb{P}\left(|\hat{\Gamma}_n - \Gamma/3|_\infty \geqslant K_3\left\{\frac{D_n^2\log(nd)}{n}\right\}^{1/2}\right) \leqslant \gamma$. Therefore, [13, Lemma C.1] confirms that with probability greater than $1-\gamma$

$$\rho(\overline{Z}, \overline{T}_n^\sharp | X_1^n) \lesssim \left[|4\hat{\Gamma}_n - 4\Gamma/3|_\infty \log^2(nd)\right]^{1/3} \asymp \left\{\frac{D_n^2\log^5(nd)}{n}\right\}^{1/6} \lesssim \varpi_n.$$

In conclusion, $\rho(\overline{T}_n, \overline{T}_n^\sharp | X_1^n) \leqslant \rho(\overline{T}_n, \overline{Z}) + \rho(\overline{Z}, \overline{T}_n^\sharp \mid X_1^n) \leqslant C(\underline{b}, K)\varpi_n$. □

Proof of Theorem 3.2. This proof is similar to the proof of Theorem 3.1 so that only the key steps are given below. Without loss of generality, we may assume $\varpi_n' \leq 1$.

Step 1. Gaussian approximation to T_n. Let $\Gamma = \mathrm{Cov}(g(X_1))$, then $\mathrm{Cov}(L_n) = \frac{4(n+1)}{3(n-1)}\Gamma$. So the correlation matrix of L_n is the same as the correlation matrix of $g(X_1)$. By (A1), $\sigma_j^2 = \mathrm{Cov}_{jj}(L_n) \geqslant \underline{b}$. By Jensen's inequality, under (A2), we have $\mathbb{E}|g_j(X_i)|^{2+k} \leqslant D_n^k$ for $k = 1, 2$, whereas under (A3'), we have $\|g_j(X_i)\|_{\psi_2} \leqslant D_n$. Therefore,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left|\frac{2(n-2i+1)}{(n-1)\sigma_j} g_j(X_i)\right|^4 \leqslant \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left|g_j(X_i)\right|^4/\sigma_j^4 \leqslant 2D_n^2/\underline{b}^2 \leqslant B_n^2,\\ &\left\|\frac{2(n-2i+1)}{(n-1)\sigma_j} g_j(X_i)\right\|_{\psi_2} \leqslant 2||g_j(X_i)||_{\psi_2}/\sigma_j \leqslant 2D_n/\underline{b}^{1/2} \leqslant B_n,\end{aligned}$$

where $B_n = 2D_n(\underline{b}^{-1} \vee \underline{b}^{-1/2})$. By [18, Corollary 2.1], the Conditions (M) and (E.2) are satisfied so that

$$\rho(\overline{L}_n, \overline{Z}_n) \leqslant K_1 \left(\frac{B_n (\log d)^{3/2} (\log n)}{n^{1/2} \sigma_*^2} \vee \frac{B_n (\log d)^2}{n^{1/2} \sigma_*} \right) \leqslant C_1(\underline{b})(\sigma_*^{-2} \vee \sigma_*^{-1}) \varpi_n', \tag{A.4}$$

where $\overline{Z}_n = \max_{1 \leqslant j \leqslant d} Z_{nj}$ and $Z_n \sim N(0, \frac{4(n+1)}{3(n-1)}\Gamma)$. Let $Z \sim N(0, 4\Gamma/3)$. We still have

$$\rho(\overline{Z}_n, \overline{Z}) \leqslant C_2(\underline{b}) \frac{D_n^{1/3} \log^{2/3} d}{n^{1/3}} \leqslant C_2(\underline{b}) \varpi_n'.$$

Hence, by triangle inequality, $\rho(\overline{L}_n, \overline{Z}) \leqslant \rho(\overline{L}_n, \overline{Z}_n) + \rho(\overline{Z}_n, \overline{Z}) \leqslant C_3(\underline{b}, \sigma_*) \varpi_n'$, where $C_3(\underline{b}, \sigma_*) \leqslant C_1(\underline{b})(\sigma_*^{-2} \vee \sigma_*^{-1}) + C_2(\underline{b})$.

Note that, by choosing $a = n^{-1/4} D_n^{1/2} \log^{3/4} d$, the following approximation still holds

$$\rho(\overline{T}_n, \overline{Z}) \leqslant C_3 D_n a^{-1} n^{-1/2} \log^{3/2} d + C_3 \varpi_n' + a \leqslant (2C_3 + 1) \varpi_n'.$$

Step 2. Bootstrap approximation to T_n. Recall $T_n^\sharp | X_1^n \sim N(\mathbf{0}, 4\hat{\Gamma}_n)$. By [18, Lemma 2.1],

$$\rho(\overline{Z}, \overline{T}_n^\sharp | X_1^n) \leqslant K_2 \frac{V \log d}{\underline{b}^2 \sigma_*^2} \left(1 \vee |\log \frac{V}{\underline{b}^2 \sigma_*^2}| \right),$$

where

$$\mathbb{P}\left(V = |\hat{\Gamma}_n - \Gamma/3|_\infty \leqslant K_3 D_n n^{-1/2} \log^{1/2}(nd) \right) \geqslant 1 - \gamma$$

by Lemma A.1. Without loss of generality, we assume $\varpi_n' \leqslant 1$. Then, with probability greater than $1 - \gamma$, $V \leqslant K_3 n^{-1/4} \log^{-1} n \log^{-1/2} d \leqslant K_3 n^{-1/4}$. If $V \underline{b}^{-2} \sigma_*^{-2} \geqslant 1$, then

$$|\log \frac{V}{\underline{b}^2 \sigma_*^2}| \leqslant C_4(\underline{b}, \sigma_*) \log(n),$$

so that $\rho(\overline{Z}, \overline{T}_n^\sharp | X_1^n) \leqslant C_5(\underline{b}, \sigma_*) V (\log d)(\log n)$ and therefore (3.4) holds. If $V \underline{b}^{-2} \sigma_*^{-2} < 1$, then observing that the function $f(x) = x |\log x| \leqslant e^{-1}(1-t)^{-1} x^t$ for any $0 < t < 1$ on $x \in (0, 1)$, we have

$$\rho(\overline{Z}, \overline{T}_n^\sharp | X_1^n) \leqslant K_2 (\log d)(V \underline{b}^{-2} \sigma_*^{-2})^t.$$

Taking $t = 1/2$ and plugging in $V \leqslant K_3 D_n n^{-1/2} \log^{1/2}(nd)$, we still have (3.4) holds with probability greater than $1 - \gamma$. □

Proof of Theorem 3.3. Denote $T_n = T_n(X_1^n) = n^{1/2} \binom{n}{2}^{-1} \sum_{1 \leqslant i < j \leqslant n} h(X_i, X_j)$ and $T_n^\xi = T_n(\xi_1^n) = n^{1/2} \binom{n}{2}^{-1} \sum_{1 \leqslant i < j \leqslant n} h(\xi_i, \xi_j)$. Define

$$\tilde{\Delta} = n^{-1/2} \binom{n}{2} \{T_n(X_1^n) - T_n(\xi_1^n)\} = \sum_{1 \leqslant i < j \leqslant n} h(X_i, X_j) - h(\xi_i, \xi_j).$$

Note that, $\overline{T}_n^{\xi} = |T_n(\xi_1^n)|_\infty \geqslant 2n^{-1/2}(n-1)^{-1}|\tilde{\Delta}|_\infty - \overline{T}_n$. It follows that

$$\begin{aligned}
\text{Type II error} &= \mathbb{P}\left(\overline{T}_n \leqslant q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha) \mid H_1\right) \\
&\leqslant \mathbb{P}\left(\overline{T}_n^{\xi} \geqslant 2n^{-1/2}(n-1)^{-1}|\tilde{\Delta}|_\infty - q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha) \mid H_1\right) \\
&\leqslant \mathbb{P}\left(\overline{T}_n^{\xi} \geqslant q_{\overline{T}_n^{\xi}}(1-\beta_n) \mid H_1\right) + \\
&\quad \mathbb{P}\left(q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha) + q_{\overline{T}_n^{\xi}}(1-\beta_n) \geqslant 2n^{-1/2}(n-1)^{-1}|\tilde{\Delta}|_\infty \mid H_1\right) \\
&\leqslant \beta_n + \mathbb{P}\left(q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha) + q_{\overline{T}_n^{\xi}}(1-\beta_n) \geqslant 2n^{-3/2}|\tilde{\Delta}|_\infty \mid H_1\right).
\end{aligned}$$

Let $\gamma = \zeta/8$. Now denote

$$\begin{aligned}
\Delta_1 &= \gamma^{-1} D_n \log(d)\{m(n-m)\}^{1/2}, \\
\Delta_2 &= D_n\{m(n-m)\}^{1/2}\{m \wedge (n-m)\}^{1/2}\log^{1/2}(nd), \\
\Delta_3 &= D_n n^{3/2}\log^{1/2}(nd/\alpha), \\
\Delta_4 &= n^{3/2}\log^{1/2}(\gamma^{-1})\log^{1/2}(d).
\end{aligned}$$

We will quantify $|\tilde{\Delta}|_\infty$, $q_{\overline{T}_n^\sharp}(1-\alpha)$ and $q_{\overline{T}_n^{\xi}}(1-\beta_n)$ to conclude that the Type II error is bounded when $|\theta_h|_\infty$ satisfies (3.6).

(1) Quantify $|\tilde{\Delta}|_\infty$. Without loss of generality, we may assume $n_1 = m \leqslant n - m = n_2$. Recall (2.3) where $V_n = V_n(X_1^n)$. Denote $V_n(\xi_1^n)$ in similar way. By shift-invariant assumption and the two-sample projection in Section 2,

$$\begin{aligned}
\tilde{\Delta} &= V_n(X_1^n) - V_n(\xi_1^n) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} h(X_i, Y_j) - h(X_i, Y_j - \theta) \\
&= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} g(Y_j - \theta) - g(Y_j) + \breve{f}(X_i, Y_j) - \breve{f}(X_i, Y_j - \theta) \\
&= n_1 n_2 \theta_h + n_1 \sum_{j=1}^{n_2}\{-g(Y_j) - \theta_h\} + n_1 \sum_{j=1}^{n_2} g(Y_j - \theta) \\
&\qquad + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i, Y_j) - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i, Y_j - \theta). \qquad \text{(A.5)}
\end{aligned}$$

By Lemma A.5, with probability smaller than γ,

$$n_1\Big|\sum_{j=1}^{n_2}[-g(Y_j) - \theta_h]\Big|_\infty \geqslant K_1 D_n n_1 n_2^{1/2}\log^{1/2}(nd) = K_1\Delta_2.$$

Similarly, $n_1|\sum_{j=1}^{n_2} g(Y_j - \theta)|_\infty \geqslant K_2\Delta_2$ with probability smaller than γ. By Lemma A.6,

$$\mathbb{E}\Big|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i, Y_j)\Big|_\infty \leqslant K_3\Delta_1\gamma.$$

From Markov inequality, $\mathbb{P}\left(|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j)|_\infty \geqslant K_3\Delta_1\right) \leqslant \gamma$. Similarly, $|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j-\theta)|_\infty \geqslant K_4\Delta_1$ with probability smaller than γ. Therefore,

$$\begin{aligned}|\tilde{\Delta}|_\infty &\geqslant n_1n_2|\theta_h|_\infty - |n_1\sum_{j=1}^{n_2}[-g(Y_j)-\theta_h]|_\infty - |n_1\sum_{j=1}^{n_2}g(Y_j-\theta)|_\infty \\ &\quad - |\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j)|_\infty - |\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j-\theta)|_\infty \\ &\geqslant n_1n_2|\theta_h|_\infty - (K_1+K_2)\Delta_2 - (K_3+K_4)\Delta_1\end{aligned}$$

with probability no smaller than $1-4\gamma$.

(2) Bound $q_{\overline{T}_n^\sharp}(1-\alpha)$. Recall $T_n^\sharp|X_1^n \sim N_d(\mathbf{0}, 4\hat{\Gamma}_n)$, where $\hat{\Gamma}_n$ is defined in (A.3). By the Bonferroni inequality, $\mathbb{P}\left(\overline{T}_n^\sharp > t|X_1^n\right) \leqslant 2d\left[1-\Phi(t/2\overline{\psi})\right]$, where $\overline{\psi}^2 = \max_{1\leqslant l\leqslant d}\hat{\Gamma}_{n,ll}$. By the Cauchy-Schwarz inequality, for each $l=1,\ldots,d$,

$$\begin{aligned}\left\{\sum_{i<j,k}h_l(X_i,X_j)h_l(X_i,X_k)\right\}^2 &\leqslant \left\{\sum_{i<j,k}h_l^2(X_i,X_j)\right\}\left\{\sum_{i<j,k}h_l^2(X_i,X_k)\right\} \\ &= \left\{\sum_{i<j,k}h_l^2(X_i,X_j)\right\}^2,\end{aligned}$$

which implies

$$\hat{\Gamma}_{n,ll} \leqslant n^{-1}(n-1)^{-2}\sum_{i=1}^n\sum_{i<j}(n-i)h_l^2(X_i,X_j) \leqslant (n-1)^{-2}\sum_{i=1}^n\sum_{i<j}h_l^2(X_i,X_j).$$

By Condition [A2] and [B2], $\mathbb{E}h_l^2(X_i,X_j) \leqslant \mathbb{E}|h_l(X_i,X_j)-\mathbb{E}h_l(X_i,X_j)|^2 + |\mathbb{E}h_l(X_i,X_j)|^2 \leqslant D_n + |\theta_h|_\infty^2\ \mathbf{1}(1\leqslant i\leqslant m<j\leqslant n)$ for any $1\leqslant l\leqslant d$ and $1\leqslant i<j\leqslant n$. From Lemma A.2, it shows that with probability grater than $1-\gamma$,

$$\begin{aligned}\overline{\psi}^2 &\leqslant (n-1)^{-2}\Big\{t^\diamond + \max_{1\leqslant l\leqslant d}\sum_{i=1}^n\sum_{i<j}\mathbb{E}h_l^2(X_i,X_j)\Big\} \\ &\lesssim D_n^2 + |\theta_h|_\infty^2\underbrace{n^{-2}\{n_1n_2 + n_1^{1/2}n_2\log^{1/2}(nd) + n_2\log^3(nd)\log(\gamma^{-1})\}}_{\delta_n}.\end{aligned}$$

Therefore, $\overline{\psi} \leqslant K_5\left[D_n + |\theta_h|_\infty\delta_n^{1/2}\right]$. In addition, for $\Phi^{-1}(1-\alpha/(2d)) = t_\alpha > 0$ (as $d>1$), Gaussian tail bound (Chernoff method) shows $t_\alpha \leqslant [2\log(2d/\alpha)]^{1/2}$. Then, with probability greater than $1-\gamma$,

$$q_{\overline{T}_n^\sharp}(1-\alpha) \leqslant 2\overline{\psi}\Phi^{-1}(1-\alpha/(2d)) \leqslant K_6n^{-3/2}\left(\Delta_3 + |\theta_h|_\infty\left\{n^3\log(2d/\alpha)\delta_n\right\}^{1/2}\right).$$

Since $n_2 \geqslant n/2$ and $n_1 \gtrsim \log^{5/2}(nd)$, the rate of $\{n^3 \log(2d/\alpha)\delta_n\}^{1/2} \lesssim n_1 n_2$ leads to $q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha) \leqslant K_6 n^{-3/2}(\Delta_3 + n_1 n_2 |\theta_h|_\infty)$. For bounded kernel h, a simpler bound of $\overline{\psi} \leqslant K_5 D_n$ directly lead to $q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha) \leqslant K_6 n^{-3/2}\Delta_3$ without assuming $n_1 \gtrsim \log^{5/2}(nd)$.

(3) Bound $q_{\overline{T}_n^\xi}(1-\beta_n)$. Note that $\overline{T}_n^\xi$ has the same distribution as $\overline{T}_n|H_0$. By the approximation in Theorem 3.1 Step 1, we have $\rho(\overline{T}_n^\xi, \overline{Z}) \leqslant C_1\varpi_n$ holds for $Z \sim N_d(0, 4\Gamma/3)$ with probability grater than $1-\gamma$. Since $||\overline{Z}||_{\psi_2} \leqslant C_2(\underline{b}) \log^{1/2}(d)$ by [54, Lemma 2.2.2] and $\mathbb{P}(\overline{Z} > t) \leqslant 2\exp\left\{-(\frac{t}{||\overline{Z}||_{\psi_2}})^2\right\} \leqslant 2\exp\left\{-C_2(\underline{b})^{-2}\log^{-1}(d)t^2\right\}$. Choosing $t = C_3(\underline{b})\log^{1/2}(\gamma^{-1})\log^{1/2}(d)$ for large enough $C_3(\underline{b})$, we have $\mathbb{P}(\overline{Z} > t) \leqslant 2\gamma$. Hence, $\mathbb{P}(\overline{T}_n^\xi > t) \leqslant \mathbb{P}(\overline{Z} > t) + C_1\varpi_n$. Let $\beta_n = 2\gamma + C_1\varpi_n$. Then with probability grater than $1-\gamma$,

$$q_{\overline{T}_n^\xi}(1-\beta_n) \leqslant C_3(\underline{b})\log^{1/2}(\gamma^{-1})\log^{1/2}(d) = C_3(\underline{b})n^{-3/2}\Delta_4.$$

Combining Step (1)-(3), when $m(n-m)|\theta_h|_\infty > 2(K_3+K_4)\Delta_1 + 2(K_1+K_2)\Delta_2 + K_6\Delta_3 + C_3(\underline{b})\Delta_4$,

$$|\tilde{\Delta}|_\infty \geqslant \frac{1}{2}n^{3/2}\left\{q_{\overline{T}_n^\sharp}(1-\alpha) + q_{\overline{T}_n^\xi}(1-\beta_n)\right\}$$

with probability no smaller than $1-6\gamma$. That is, the Type II error is less than $6\gamma + \beta_n = 8\gamma + C_1\varpi_n$, where we set $\zeta = 8\gamma$. As $(\Delta_1 \vee \Delta_2) \lesssim \Delta_3$, the conclusion of Theorem 3.3 immediately follows for some large enough $K \geqslant 2\sum_{i=1}^6 K_i$. □

Proof of Lemma 4.1. Let

$$\tilde{\Delta} = \sum_{1\leqslant i<j\leqslant n} h(X_i, X_j) - h(\xi_i, \xi_j) = \sum_{k<k'} \tilde{\Delta}^{(k,k')},$$

where

$$\tilde{\Delta}^{(k,k')} = \sum_{\substack{m_k < i \leqslant m_{k+1} \\ m_{k'} < j \leqslant m_{k'+1}}} h(X_i, X_j) - h(\xi_i, \xi_j).$$

Similar to the proof of Theorem 3.3, we shall quantify $|\tilde{\Delta}|_\infty$, $q_{\overline{T}_n^\sharp}(1-\alpha)$ and $q_{\overline{T}_n^\xi}(1-\beta_n)$ to conclude that the Type II error is bounded when $|\delta|_\infty$ satisfies (4.3).

(1) Quantify $|\tilde{\Delta}|_\infty$.

$$\begin{aligned}\tilde{\Delta}^{(k,k')} = s_k s_{k'}\delta^{(k,k')} &+ s_k \sum_{m_{k'} < j \leqslant m_{k'+1}} \{-g(X_j - \theta^{(k)}) - \delta^{(k,k')}\} \\ &+ s_k \sum_{m_{k'} < j \leqslant m_{k'+1}} g(X_j - (\theta^{(k')} - \theta^{(k)})) \\ &+ \sum_{\substack{m_k < i \leqslant m_{k+1} \\ m_{k'} < j \leqslant m_{k'+1}}} \breve{f}(X_i, X_j) - \sum_{\substack{m_k < i \leqslant m_{k+1} \\ m_{k'} < j \leqslant m_{k'+1}}} \breve{f}(X_i, X_j - \theta^{(k)}).\end{aligned}$$

Applying the results in Step (1) to $\sum_{k<k'}\tilde{\Delta}^{(k,k')}$, we have each of the following inequalities satisfied with probability greater than $1-\gamma$:

$$
\begin{aligned}
&|\sum_{k<k'} s_k \sum_{m_{k'}<j\leqslant m_{k'+1}}\{-g(X_j-\theta^{(k)})-\delta^{(k,k')}\}|_\infty \\
&\quad\leqslant \sum_{k<k'} K_1 D_n (s_k s_{k'})^{1/2} n^{1/2}\log^{1/2}(nd) \leqslant K_1\nu^2 D_n n^{3/2}\log^{1/2}(nd);\\
&|\sum_{k<k'} s_k \sum_{m_{k'}<j\leqslant m_{k'+1}} g(X_j-(\theta^{(k')}-\theta^{(k)}))|_\infty \\
&\quad\leqslant \sum_{k<k'} K_2 D_n (s_k s_{k'})^{1/2} n^{1/2}\log^{1/2}(nd) \leqslant K_2\nu^2 D_n n^{3/2}\log^{1/2}(nd);\\
&|\sum_{k<k'} \sum_{\substack{m_k<i\leqslant m_{k+1}\\ m_{k'}<j\leqslant m_{k'+1}}} \breve{f}(X_i,X_j)|_\infty + |\sum_{k<k'} \sum_{\substack{m_k<i\leqslant m_{k+1}\\ m_{k'}<j\leqslant m_{k'+1}}} \breve{f}(X_i,X_j-\theta^{(k)})|_\infty \\
&\quad\leqslant \sum_{k<k'} K_3\gamma^{-1} D_n (s_k s_{k'})^{1/2}\log d \leqslant K_3\nu^2 D_n n^{3/2}\log^{1/2}(nd).
\end{aligned}
$$

Combining all pairs of (k,k') for $0\leqslant k<k'\leqslant\nu$, it follows

$$
\begin{aligned}
|\tilde{\Delta}|_\infty &= |\sum_{k<k'}\tilde{\Delta}^{(k,k')}|_\infty \\
&\geqslant |\sum_{k<k'} s_k s_{k'}\delta^{(k,k')}|_\infty - (K_1+K_2+K_3)\nu^2 D_n n^{3/2}\log^{1/2}(nd)
\end{aligned}
$$

with probability greater than $1-3\gamma$.

(2) Bound $q_{\overline{T}_n^\sharp}(1-\alpha)$. Under H_1', $T_n^\sharp|X_1^n \sim N_d(\mathbf{0}, 4\hat{\Gamma}_n)$, where $\hat{\Gamma}_n$ is defined the same as in (A.3). To control the magnitude of $|\sum_{1\leqslant i<j\leqslant n} h_l^2(X_i,X_j)|$, note that

$$
\sum_{1\leqslant i<j\leqslant n} = \sum_{\substack{m_k<i\leqslant m_{k+1}\\ m_{k'}<j\leqslant m_{k'+1}\\ 0\leqslant k<k'\leqslant\nu}} + \sum_{\substack{m_k<i<j\leqslant m_{k+1}\\ 0\leqslant k\leqslant\nu}}.
$$

So we can modify Lemma A.2 from the following two cases. For the case of $\mathcal{C}_{k,k'} = \{m_k < i \leqslant m_{k+1} \leqslant m_{k'} < j \leqslant m_{k'+1}\}$ where i,j are in different segments, $\mathbb{E}h_l^2(X_i,X_j)\leqslant D_n + |\delta_l^{(k,k')}|^2$, based on modified Lemma A.2 we have

$$
\begin{aligned}
\mathbb{P}\Big(\max_{1\leqslant l\leqslant d}|\sum_{\mathcal{C}_{k,k'}} h_l^2(X_i,X_j) - \mathbb{E}h_l^2(X_i,X_j)| \geqslant \\
\max_{k<k'} K_4(D_n^2 + |\delta^{(k,k')}|_\infty^2)(s_k s_{k'})^{1/2} n^{1/2}\log^{1/2}(nd)\Big) \leqslant \gamma.
\end{aligned}
$$

For the case of $\mathcal{C}_k = \{m_k < i < j \leqslant m_{k+1}\}$ where i,j are in the same segments, $|\mathbb{E}h_l(X_i,X_j)|^2 \leqslant D_n$ and

$$
\mathbb{P}\left(\max_{1\leqslant l\leqslant d}|\sum_{\mathcal{C}_k} h_l^2(X_i,X_j) - \mathbb{E}h_l^2(X_i,X_j)| \geqslant K_5 D_n^2 n^{3/2}\log^{1/2}(nd)\right) \leqslant \gamma.
$$

Take $t^\diamond = D_n^2 n^{3/2}\log^{1/2}(nd) + \max_{k<k'}(s_k s_{k'})^{1/2}|\delta^{(k,k')}|_\infty^2 n^{1/2}\log^{1/2}(nd)$. Then, adding all $\mathcal{C}_k$ and $\mathcal{C}_{k,k'}$ together,

$$\begin{aligned}\overline{\psi}^2 &= \max_{1\leqslant l\leqslant d}\hat{\Gamma}_{n,ll}\\ &\leqslant (n-1)^{-2}K_6\Big\{t^\diamond + \max_{1\leqslant l\leqslant d}\sum_{i=1}^n\sum_{i<j}\mathbb{E}h_l^2(X_i,X_j)\Big\}\\ &\leqslant K_6\left\{D_n^2 + n^{-3/2}\log^{1/2}(nd)\max_{k<k'}(s_k s_{k'})^{1/2}|\delta^{(k,k')}|_\infty^2\right.\\ &\qquad\left. + n^{-2}\sum_{k<k'}s_k s_{k'}|\delta^{(k,k')}|_\infty^2\right\}\end{aligned}$$

holds with probability greater than $1-(\nu+1)(\nu+2)\gamma/2$. Therefore, $q_{\overline{T}_n^\natural}(1-\alpha)\leqslant K_7\overline{\psi}t_\alpha$, where $t_\alpha = \Phi^{-1}(1-\alpha/(2d))\leqslant 2\log^{1/2}(nd/\alpha)$ and

$$\begin{aligned}\overline{\psi}\leqslant K_6\left\{D_n + n^{-3/4}\log^{1/4}(nd)\max_{k<k'}(s_k s_{k'})^{1/4}|\delta^{(k,k')}|_\infty\right.\\ \left. + n^{-1}\sum_{k<k'}(s_k s_{k'})^{1/2}|\delta^{(k,k')}|_\infty\right\}.\end{aligned}$$

(3) Bound $q_{\overline{T}_n^\xi}(1-\beta_n)$. Since $\overline{T}_n^\xi$ does not depend on H_1', it obeys the same bound

$$q_{\overline{T}_n^\xi}(1-\beta_n)\leqslant C(\underline{b})\log^{1/2}(\gamma^{-1})\log^{1/2}(d) = C(\underline{b})\log^{1/2}(\gamma^{-1})\log^{1/2}(d)$$

with probability grater than $1-\gamma$ for $\beta_n = 2\gamma + C_1\varpi_n$.

Combining Step (1)-(3), when

$$\begin{aligned}&|\sum_{k<k'}s_k s_{k'}\delta^{(k,k')}|_\infty\\ >\ &K_0\nu^2 D_n n^{3/2}\log^{1/2}(nd/\alpha)\\ &+ C(\underline{b})n^{3/2}\log^{1/2}(\gamma^{-1})\log^{1/2}(d)\\ &+ K_0'\log^{1/2}(nd/\alpha)\left\{n^{3/4}\log^{1/4}(nd)\max_{k<k'}(s_k s_{k'})^{1/4}|\delta^{(k,k')}|_\infty\right.\\ &\qquad\left. + n^{1/2}\sum_{k<k'}(s_k s_{k'})^{1/2}|\delta^{(k,k')}|_\infty\right\},\end{aligned}$$

the Type II error will be smaller than $\beta_n + \{4+(\nu+1)(\nu+2)/2\}\gamma$ for $\beta_n = 2\gamma + C_1\varpi_n$. Substitute γ by $\{4+(\nu+1)(\nu+2)/2\}^{-1}\zeta$, we reach the conclusion of theorem. □

A.2. Proof of lemmas in theorems

Lemma A.1 (Bounding $|\hat{\Gamma}_n - \Gamma/3|_\infty$ under H_0.)**.** *Suppose all the conditions in Theorem 3.1 hold. Let* $\Gamma = Cov(g(X_1))$ *and* $\hat{\Gamma}_n$ *be defined as in (A.3). Then with probability greater than* $1-\gamma$,

$$|\hat{\Gamma}_n - \Gamma/3|_\infty \leqslant K_0 \left(\frac{D_n^2 \log(nd)}{n}\right)^{1/2}.$$

Proof of Lemma A.1. Note $\Gamma = \mathrm{Cov}(\mathbb{E}[h(X, X_1)|X]) = \mathbb{E}[h(X_1, X_2)h(X_1, X_3)^T]$ and let $\Gamma_2 = \mathbb{E}[h(X_1, X_2)h(X_1, X_2)^T]$. Then

$$\begin{aligned}\mathbb{E}\hat{\Gamma}_n &= \frac{1}{n(n-1)^2}\sum_{i=1}^{n}(n-i)(n-i-1)\Gamma + \frac{1}{n(n-1)^2}\sum_{i=1}^{n}(n-i)\Gamma_2 \\ &= \frac{n-2}{3(n-1)}\Gamma + \frac{1}{2(n-1)}\Gamma_2.\end{aligned}$$

Note that, the summation in $\hat{\Gamma}_n$ can split into two parts

$$\sum_{i=1}^{n}\sum_{j,k>i} = \sum_{i=1}^{n}\sum_{j\neq k>i} + \sum_{i=1}^{n}\sum_{j=k>i}.$$

In Steps 1 and 2 below, we will deal with

$$\hat{\Gamma}_{n1} = \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq k>i} h(X_i, X_j)h(X_i, X_k)^T \text{ and}$$

$$\hat{\Gamma}_{n2} = \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j=k>i} h(X_i, X_j)h(X_i, X_k)^T,$$

where $\hat{\Gamma}_n = \hat{\Gamma}_{n1} + \hat{\Gamma}_{n2}$. Then conclusion will be made in Step 3.

Step 1: Term $\hat{\Gamma}_{n1} = \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq k>i} h(X_i, X_j)h(X_i, X_k)^T$.

Define $H(x_1, x_2, x_3)$ to be $h(x_1, x_2)h(x_1, x_3)^T$. To symmetrize H, let $H'(X_i, X_j, X_k) = \sum_{\pi_3}\tilde{H}(X_{\pi_3(i)}, X_{\pi_3(j)}, X_{\pi_3(k)})$, where

$$\tilde{H}(X_i, X_j, X_k) = \begin{cases} H(X_i, X_j, X_k), & \text{if } i < j \neq k, \\ \mathbf{0}, & \text{otherwise} \end{cases},$$

and π_3 is a permutation of $\{i, j, k\}$. Then,

$$\begin{aligned}\hat{\Gamma}_{n1} &= \frac{1}{n(n-1)^2}\sum_{i<j\neq k} H(X_i, X_j, X_k) = \frac{1}{n(n-1)^2}\sum_{i\neq j\neq k}\tilde{H}(X_i, X_j, X_k) \\ &= \frac{1}{6n(n-1)^2}\sum_{i\neq j\neq k} H'(X_i, X_j, X_k)\end{aligned}$$

is a U-statistics of order 3 and $\mathbb{E}\hat{\Gamma}_{n1} = \frac{n-2}{3(n-1)}\Gamma$. Let

$$W_n = \frac{(n-3)!}{n!}\sum_{i\neq j\neq k} H'(X_i, X_j, X_k) = \frac{6(n-1)}{n-2}\hat{\Gamma}_{n1}.$$

Apply Lemma E.1 in [13] to H' for $\alpha = 1/2, \eta = 1$ and $\delta = 1/2$,

$$\mathbb{P}\left(\frac{n}{3}|W_n - \mathbb{E}W_n|_\infty \geqslant 2\mathbb{E}Z_1 + t\right) \leqslant \exp\left(-\frac{t^2}{3\overline{\zeta}_n^2}\right) + 3\exp\left[-\left(\frac{t}{K_1||M||_{\psi_{1/2}}}\right)^{1/2}\right], \quad \text{(A.6)}$$

where

$$\begin{aligned}
&\mathbb{E}W_n = \mathbb{E}H'(X_1, X_2, X_3) = 2\Gamma,\\
&Z_1 = \max_{1\leqslant m_1,m_2\leqslant d}\left|\sum_{i=0}^{[\frac{n}{3}]-1}\left[\overline{H'}_{m_1,m_2}(X_{3i+1}^{3i+3}) - \mathbb{E}\overline{H'}_{m_1,m_2}\right]\right|,\\
&\overline{\zeta}_n^2 = \max_{1\leqslant m_1,m_2\leqslant d}\sum_{i=0}^{[\frac{n}{2}]-1}\mathbb{E}H'^2_{m_1,m_2}(X_{3i+1}^{3i+3}),\\
&M = \max_{1\leqslant m_1,m_2\leqslant d}\max_{0\leqslant i\leqslant[\frac{n}{3}]-1}\left|H'_{m_1,m_2}(X_{3i+1}^{3i+3})\right|.
\end{aligned}$$

and $\overline{H'}_{m_1,m_2}(x_1,x_2,x_3) = H'_{m_1,m_2}(x_1,x_2,x_3)\mathbf{1}_{\{\max_{m_1,m_2}|H'_{m_1,m_2}(x_1,x_2,X_3)|\leqslant\tau\}}$ for $\tau = 8\mathbb{E}M$. By Cauchy-Schwarz and Condition (A2),

$$\begin{aligned}
\mathbb{E}H'^2_{m_1,m_2}(X_{3i+1}^{3i+3}) &\leqslant 2\mathbb{E}H^2_{m_1,m_2}(X_{3i+1}^{3i+3})\\
&\leqslant \left(\mathbb{E}h^4_{m1}(X_{3i+1}, X_{3i+2})\right)^{1/2}\left(\mathbb{E}h^4_{m2}(X_{3i+1}, X_{3i+3})\right)^{1/2} \leqslant D_n^2.
\end{aligned}$$

So $\overline{\zeta}_n \leqslant n^{1/2}D_n$. From (i) [54, Lemma 2.2.2], (ii) the fact of $||X^2||_{\psi_{1/2}} = ||X||^2_{\psi_1}$ and (iii) Condition (A3), we obtain

$$\begin{aligned}
&||M||_{\psi_{1/2}}\\
&= ||\max_{m_1,m_2}\max_i h_{m_1}(X_{3i+1}, X_{3i+2})h_{m_2}(X_{3i+1}, X_{3i+3})||_{\psi_{1/2}}\\
&\leqslant_{(i)} K_2\log^2(nd)\max_{m_1,m_2}\max_i||h_{m_1}(X_{3i+1}, X_{3i+2})h_{m_2}(X_{3i+1}, X_{3i+3})||_{\psi_{1/2}}\\
&\leqslant K_2'\log^2(nd)\max_{m_1}\max_i||h^2_{m_1}(X_{3i+1}, X_{3i+2})||_{\psi_{1/2}}\\
&=_{(ii)} K_2'\log^2(nd)\max_{m_1}\max_i||h_{m_1}(X_{3i+1}, X_{3i+2})||^2_{\psi_1}\\
&\leqslant_{(iii)} K_2'\log^2(nd)D_n^2,
\end{aligned}$$

where the ranges of indices are $1 \leqslant m_1, m_2 \leqslant d$ and $0 \leqslant i \leqslant \frac{n}{3} - 1$. By [16, Lemma 8],

$$\mathbb{E}Z_1 \leqslant K_3\left\{\sqrt{\log d}\,\overline{\zeta}_n + \log d\,||M||_{\psi_{1/2}}\right\} \leqslant K_4[n\log(nd)D_n^2]^{1/2}.$$

Therefore, (A.6) leads to

$$\mathbb{P}\big(|\hat{\Gamma}_{n1}-\mathbb{E}\hat{\Gamma}_{n1}|_\infty \geqslant 4K_4 n^{-1/2}D_n\log^{1/2}(nd)+t\big) \leqslant \exp\left(-\frac{nt^2}{3D_n^2}\right)+3\exp\left[-\frac{\sqrt{nt}}{K_1{K_2}^{1/2}\log(nd)D_n}\right].$$

Recall $K\log(nd)\geqslant\log(1/\gamma)\geqslant 1$ and $n\gtrsim D_n^2\log^7(nd)$. Choose

$$t^*=K_5\sqrt{\frac{D_n^2\log(nd)}{n}}$$

for some large enough $K_5>0$. Then,

$$\mathbb{P}\left(|\hat{\Gamma}_{n1}-\mathbb{E}\hat{\Gamma}_{n1}|_\infty\geqslant t^*\right)\leqslant\gamma^{\frac{K_5^2}{3K}}+3\gamma^{\frac{K_5^{1/2}}{KK_1K_2^{1/2}}}\leqslant\gamma/2.$$

Step 2: Term $\hat{\Gamma}_{n2}=\frac{1}{n(n-1)^2}\sum_{i=1}^n\sum_{j=k>i}h(X_i,X_j)h(X_i,X_k)^T$.

Let $H(x_1,x_2)=h(x_1,x_2)h(x_1,x_2)^T$. Denote $W_n'=\frac{(n-2)!}{n!}\sum_{i\neq j}H(X_i,X_j)=2(n-1)\hat{\Gamma}_{n2}$. By Lemma E.1 in [13],

$$\mathbb{P}\left(\frac{n}{2}|W_n'-\mathbb{E}W_n'|_\infty\geqslant 2\mathbb{E}Z_1'+t\right) \leqslant\exp\left(-\frac{t^2}{3\overline{\zeta'}_n^2}\right)+3\exp\left[-\left(\frac{t}{K_6||M'||_{\psi_{1/2}}}\right)^{1/2}\right],$$

where

$$\begin{aligned}
&\mathbb{E}W_n'=\mathbb{E}[H(X_1,X_2)]=\Gamma_2,\\
&Z_1'=\max_{1\leqslant m_1,m_2\leqslant d}\left|\sum_{i=0}^{[\frac{n}{2}]-1}\left[\overline{H}_{m_1,m_2}(X_{2i+1}^{2i+2})-\mathbb{E}\overline{H}_{m_1,m_2}\right]\right|,\\
&\overline{\zeta'}_n^2=\max_{1\leqslant m_1,m_2\leqslant d}\sum_{i=0}^{[\frac{n}{2}]-1}\mathbb{E}H_{m_1,m_2}^2(X_{2i+1}^{2i+2}),\\
&M'=\max_{1\leqslant m_1,m_2\leqslant d}\max_{0\leqslant i\leqslant[\frac{n}{2}]-1}\left|H_{m_1,m_2}(X_{2i+1}^{2i+2})\right|.
\end{aligned}$$

In addition, $\overline{H}_{m_1,m_2}(x_1,x_2)=H_{m_1,m_2}(x_1,x_2)\mathbf{1}_{\{\max_{m_1,m_2}|H_{m_1,m_2}(x_1,x_2)|\leqslant\tau\}}$ for $\tau=8\mathbb{E}M'$. Similarly,

$$\mathbb{E}H_{m_1,m_2}^2(X_{2i+1}^{2i+2})\leqslant\left(\mathbb{E}h_{m1}^4(X_{2i+1}^{2i+2})\right)^{1/2}\left(\mathbb{E}h_{m2}^4(X_{2i+1}^{2i+2})\right)^{1/2}\leqslant D_n^2.$$

So $\overline{\zeta'}_n\leqslant n^{1/2}D_n$. In addition,

$$||M'||_{\psi_{1/2}}=||\max_{1\leqslant m_1,m_2\leqslant d}\max_{0\leqslant i\leqslant\frac{n}{2}-1}h_{m_1}(X_{2i+1}^{2i+2})h_{m_2}(X_{2i+1}^{2i+2})||_{\psi_{1/2}}$$

$$\leqslant K_7 \log^2(nd) \max_{1\leqslant m_1\leqslant d} \max_{0\leqslant i\leqslant \frac{n}{2}-1} ||h_{m_1}(X_{2i+1}, X_{2i+2})||_{\psi_1}^2$$
$$\leqslant K_7 \log^2(nd) D_n^2.$$

Then by [16, Lemma 8], we have $\mathbb{E}Z_1' \leqslant K_8[n\log(nd)D_n^2]^{1/2}$. Similar to Step 1, taking $t'^* = K_9\sqrt{\frac{D_n^2\log(nd)}{n}}$ for some large enough $K_9 > 0$, we end up with

$$\mathbb{P}\left(|W_n' - \mathbb{E}W_n'|_\infty \geqslant t'^*\right) \leqslant \gamma/2,$$

i.e. $\mathbb{P}\left(|\hat{\Gamma}_{n2} - \Gamma_2|_\infty \geqslant (n-1)^{-1}\cdot t'^*\right) \leqslant \gamma/2$.

Step 3: Approximating $\hat{\Gamma}_n$ to $\Gamma/3$. By Cauchy-Schwarz inequality and Condition (A2),

$$\begin{aligned}
|\Gamma|_\infty &= \max_{1\leqslant m_1,m_2\leqslant d} |\mathbb{E}h_{m1}(X_1,X_2)\mathbb{E}h_{m2}(X_1,X_3)| \\
&\leqslant \max_{1\leqslant m_1\leqslant d} |\mathbb{E}h_{m1}^2(X_1,X_2)| \leqslant \max_{1\leqslant m_1\leqslant d} |\mathbb{E}h_{m1}^4(X_1,X_2)|^{1/2} \leqslant D_n, \\
|\Gamma_2|_\infty &= \max_{1\leqslant m_1,m_2\leqslant d} |\mathbb{E}h_{m1}(X_1,X_2)\mathbb{E}h_{m2}(X_1,X_2)| \\
&\leqslant \max_{1\leqslant m_1\leqslant d} |\mathbb{E}h_{m1}^2(X_1,X_2)| \leqslant D_n.
\end{aligned}$$

Notice that

$$|\hat{\Gamma}_n - \Gamma/3|_\infty \leqslant |\hat{\Gamma}_n - \mathbb{E}\hat{\Gamma}_n|_\infty + |\mathbb{E}\hat{\Gamma}_n - \Gamma/3|_\infty,$$

where

$$|\mathbb{E}\hat{\Gamma}_n - \Gamma/3|_\infty \leqslant \frac{1}{3(n-1)}|\Gamma|_\infty + \frac{1}{2(n-1)}|\Gamma_2|_\infty \leqslant n^{-1}D_n \leqslant K_{10}\sqrt{\frac{D_n^2\log(nd)}{n}}.$$

Combine Step 1 and 2 and take $t_0 = K_0\sqrt{\frac{D_n^2\log(nd)}{n}}$ for some $K_0 > K_{10}+K_9+K_5$ large enough, we have

$$\mathbb{P}\left(|\hat{\Gamma}_n - \Gamma/3|_\infty \geqslant t_0\right) \leqslant \gamma. \qquad \square$$

Lemma A.2 (Bounding $\max_{1\leqslant l\leqslant d}|\sum_{i=1}^n\sum_{i<j} h_l^2(X_i,X_j) - \mathbb{E}h_l^2(X_i,X_j)|$ under H_1.)**.** *Suppose all the conditions in Theorem 3.1 and Theorem 3.3 hold. Let $\gamma \in (0, e^{-1})$ such that $\log(\gamma^{-1}) \leqslant K\log(nd)$ and suppose $n_1 = m \leqslant n-m = n_2$. Then the following holds with probability greater than $1-\gamma$ for some large enough constant $K^\diamond$*

$$\max_{1\leqslant l\leqslant d} |\sum_{i=1}^n\sum_{i<j} h_l^2(X_i,X_j) - \mathbb{E}h_l^2(X_i,X_j)| \leqslant K^\diamond t^\diamond,$$

where $t^\diamond = D_n^2 n^{\frac{3}{2}}\log^{\frac{1}{2}}(nd) + |\theta_h|_\infty^2[n_1^{\frac{1}{2}}n_2\log^{\frac{1}{2}}(nd) + n_2\log^3(nd)\log(\gamma^{-1})]$.

Proof of Lemma A.2. Note that $h_l^2(x,y) = h_l^2(y,x)$ and the summation breaks down to

$$\sum_{i=1}^{n}\sum_{i<j} = \sum_{i=1}^{m}\sum_{j=i+1}^{m} + \sum_{i=1}^{m}\sum_{j=m+1}^{n} + \sum_{i=m+1}^{n}\sum_{j=i+1}^{n}.$$

Apply [13, Lemma E.1] to $\hat{\Gamma}_1 = \frac{1}{n_1(n_1-1)}\sum_{1\leqslant i<j\leqslant n_1} h(X_i,X_j)h(X_i,X_j)^T$, calculation (similar to Lemma A.1 Step 2) shows

$$\mathbb{P}\Big(|\hat{\Gamma}_1 - \mathbb{E}\hat{\Gamma}_1|_\infty \geqslant K_1[D_n n_1^{-1/2}\log^{1/2}(d) + D_n^2 n_1^{-1}\log^3(n_1 d)] + t\Big) \leqslant \exp\left(-\frac{n_1 t^2}{3D_n^2}\right) + 3\exp\left[-\left(\frac{\sqrt{n_1 t}}{K_2 D_n \log(n_1 d)}\right)\right].$$

Take $t_1 = K_3[D_n n_1^{-1/2}\log^{1/2}(nd) \vee D_n^2 n_1^{-1}\log^3(nd)\log(\gamma^{-1})]$. It follows that

$$\frac{n_1 {t_1}^2}{D_n^2} \gtrsim D_n^2\log(nd) \gtrsim \log(\gamma^{-1}) \text{ and}$$
$$\frac{\sqrt{n_1 t_1}}{D_n\log(n_1 d)} \gtrsim \left(\frac{\log^3(nd)\log(\gamma^{-1})}{\log^2(n_1 d)}\right)^{1/2} \gtrsim \log(\gamma^{-1}).$$

So $\mathbb{P}\left(|\hat{\Gamma}_1 - \mathbb{E}\hat{\Gamma}_1|_\infty \geqslant t_1\right) \leqslant \gamma/3$ for some large enough K_3. Therefore, the diagonal part obeys the same bound such that the first term $\sum_{i=1}^m\sum_{j=i+1}^m h_l^2(X_i,X_j)$ has a tail bound

$$\mathbb{P}\left(\binom{m}{2}^{-1}\max_{1\leqslant l\leqslant d}|\sum_{i=1}^{m}\sum_{j=i+1}^{m} h_l^2(X_i,X_j) - \mathbb{E}h_l^2(X_i,X_j)|_\infty \geqslant t_1\right) \leqslant \gamma/3.$$

Next, apply the two-sample tail bound Lemma A.4 to the middle term. Thus,

$$\mathbb{P}\left(\frac{1}{m(n-m)}\max_{1\leqslant l\leqslant d}|\sum_{i=1}^{m}\sum_{j=m+1}^{n} h_l^2(X_i,X_j) - \mathbb{E}h_l^2(X_i,X_j)|_\infty \geqslant t_2\right) \leqslant \gamma/3$$

holds for $t_2 = K_4 B_n^2[{n_1}^{-1/2}\log^{1/2}(nd) \vee {n_1}^{-1}\log^3(nd)\log(1/\gamma)]$, where $B_n = D_n + |\theta_h|_\infty$. At last for the third term, applying [13, Lemma E.1] to $\hat{\Gamma}_2 = \frac{1}{n_2(n_2-1)}\sum_{1\leqslant i<j\leqslant n_2} h(Y_i,Y_j)h(Y_i,Y_j)^T$, we have

$$\mathbb{P}\Big(|\hat{\Gamma}_2 - \mathbb{E}\hat{\Gamma}_2|_\infty \geqslant K_5(D_n^2 n_2^{-1}\log(n_2 d))^{1/2} + t\Big) \leqslant \exp\left(-\frac{n_2 t^2}{3D_n^2}\right) + 3\exp\left[-\left(\frac{\sqrt{n_2 t}}{K_6 D_n\log(n_2 d)}\right)\right].$$

Since $n_2 = n - m \geqslant n/2$ and $n \gtrsim D_n^2\log^7(nd)$, it suffices to take $t_3 = K_7 D_n n^{-1/2}\log^{1/2}(nd)$ such that

$$\frac{n_2 {t_3}^2}{D_n^2} \gtrsim \log(nd) \quad \text{and} \quad \frac{\sqrt{n_2 t_3}}{D_n\log(n_2 d)} \gtrsim D_n^{-1/2} n^{1/4}\log^{-3/4}(nd) \gtrsim \log(\gamma^{-1}).$$

Then, the third term has a tail bound

$$\mathbb{P}\left(\binom{n-m}{2}^{-1}\max_{1\leqslant l\leqslant d}|\sum_{i=m+1}^{n}\sum_{j=i+1}^{n}h_l^2(X_i,X_j)-\mathbb{E}h_l^2(X_i,X_j)|_\infty\geqslant t_3\right)\leqslant\gamma/3.$$

Since there exists a large enough constant $K^\diamond$ such that

$$\begin{aligned}&(n_1^2t_1)\vee(n_1n_2t_2)\vee(n_2^2t_3)\\&\leqslant K^\diamond\left\{D_n^2n^{\frac{3}{2}}\log^{\frac{1}{2}}(nd)+|\theta_h|_\infty^2[n_1^{\frac{1}{2}}n_2\log^{\frac{1}{2}}(nd)+n_2\log^3(nd)\log(\gamma^{-1})]\right\}=:t^\diamond,\end{aligned}$$

we conclude $\mathbb{P}\left(\max_{1\leqslant l\leqslant d}|\sum_{i=1}^{n}\sum_{i<j}h_l^2(X_i,X_j)-\mathbb{E}h_l^2(X_i,X_j)|\geqslant 3t^\diamond\right)\leqslant\gamma$. □

A.3. Lemma for tail probability of the maximum of two-sample U-statistics

Let $X_1^{n_1}$ and $Y_1^{n_2}$ be two random samples taking values in a measurable space $(S,\mathcal{S})$. Suppose $X_i\sim F$ are independent with $Y_j\sim G$. Let $h:S^2\to\mathbb{R}^d$ be a measurable function and

$$T_n=\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}h(X_i,Y_j)$$

be the two-sample U-statistics. WLOG, we may first assume $n_1\leqslant n_2$. Consider a permutation π_{n_2} on $Y_1^{n_2}$ and the sum of first n_1 pairs $\sum_{i=1}^{n_1}h(X_i,Y_{\pi_{n_2}(i)})$

$$\begin{array}{ccc:c}X_1&\cdots&X_{n_1}&\\\downarrow&&\downarrow&\\Y_{\pi_{n_2}(1)}&\cdots&Y_{\pi_{n_2}(n_1)}&Y_{\pi_{n_2}(n_1+1)}\ \cdots\ Y_{\pi_{n_2}(n_2)}\end{array}$$

The symmetry leads to $\sum_{\pi_{n_2}}\sum_{i=1}^{n_1}h(X_i,Y_{\pi_{n_2}(i)})=(n_2-1)!\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}h(X_i,Y_j)$, i.e.

$$\frac{1}{n_2!}\sum_{\pi_{n_2}}\sum_{i=1}^{n_1}h(X_i,Y_{\pi_{n_2}(i)})=\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}h(X_i,Y_j).$$

This representation reduce the bounds on $Z=n_1|T_n-\theta_h|_\infty$ to those of $|V|_\infty=|\sum_{i=1}^{n_1}h(X_i,Y_i)-\theta_h|_\infty$, where $\theta_h=\mathbb{E}h(X_1,Y_1)$. Define

$$\begin{aligned}\overline{h}(x,y)&=h(x,y)\mathbf{1}\{\max_{1\leqslant k\leqslant d}|h_k(x,y)|\leqslant\tau\},\tau>0\\Z_1&=\max_{1\leqslant k\leqslant d}\left|\sum_{i=1}^{n_1}\overline{h}_k(X_i,Y_i)-\mathbb{E}\bar{h}_k\right|\\M&=\max_{1\leqslant k\leqslant d}\max_{1\leqslant i\leqslant n_1}|h_k(X_i,Y_i)|\end{aligned}$$

$$\overline{\zeta}_{n_1}^2 \quad = \quad \max_{1\leqslant k\leqslant d}\sum_{i=1}^{n_1}\mathbb{E}h_k^2(X_i,Y_i)$$

By similar argument of Lemma E.1 in [13], we have the following result.

Lemma A.3 (Sub-exponential inequality for the maxima of centered two-sample U-statistics)**.** *Let $X_1,\cdots X_{n_1}$ and $Y_1,\cdots Y_{n_2}$ be two independent sets of iid random vectors from F and G, respectively. Suppose $n_1\leqslant n_2$ and $||h_k(X_1,Y_1)||_{\psi_\alpha}<\infty$ for $\alpha\in(0,1]$ and all $k=1,\cdots,d$. Let $\tau=8\mathbb{E}[M]$, then for any $0<\eta\leqslant 1$ and $\delta>0$, there exists a constant $C(\alpha,\eta,\delta)>0$ such that*

$$\mathbb{P}(Z\geqslant(1+\eta)\mathbb{E}Z_1+t)\leqslant\exp\left(-\frac{t^2}{2(1+\delta)\overline{\zeta}_{n_1}^2}\right)+3\exp\left[-\left(\frac{t}{C(\alpha,\eta,\delta)||M||_{\psi_\alpha}}\right)^{\alpha}\right] \tag{A.7}$$

holds for all $t>0$.

Proof of Lemma A.3. See Lemma E.1 in [13]. □

By Lemma A.3, we can have the following result.

Lemma A.4 (Tail bound of the maxima of two-sample U-statistics in second order)**.** *Let $X_1,\cdots X_{n_1}$ and $Y_1,\cdots Y_{n_2}$ be two independent sets of iid random vectors from F and G, respectively. Let $\underline{n}=\min\{n_1,n_2\}$, $\overline{n}=\max\{n_1,n_2\}$ and $\zeta\in(0,1)$ be a constant s.t. $\log(\zeta^{-1})\leqslant K\log(\overline{n}d)$. Suppose $||h_k(X_1,Y_1)-\mathbb{E}h_k(X_1,Y_1)||_{\psi_1}\leqslant D_n$ and $\mathbb{E}|h_k(X_1,Y_1)-\mathbb{E}h_k(X_1,Y_1)|^{2+\ell}\leqslant D_n^\ell$ for all $k=1,\cdots,d$ and $\ell=1,2$. Denote $B_n=D_n+|\theta_h|_\infty$, where $\theta_h=\mathbb{E}h(X_1,Y_1)$. Then,*

$$\mathbb{P}(\max_{1\leqslant k\leqslant d}|\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}h_k^2(X_i,Y_j)-\mathbb{E}h_k^2(X_i,Y_j)|\geqslant t^*)\leqslant\zeta \tag{A.8}$$

holds for $t^*=K_0B_n^2\{\underline{n}^{-1/2}\log^{1/2}(\overline{n}d)+\underline{n}^{-1}\log^3(\overline{n}d)\log(1/\zeta)\}$.

Proof of Lemma A.4. Without loss of generality, we may assume $D_n\geqslant 1$. Let $H_k(x,y)=h_k^2(x,y)$, $k=1,\ldots,d$, and define Z, Z_1, M and $\overline{\zeta}_{n_1}^2$ for H accordingly. Apply Lemma A.3 to $H(x,y)$ and follow the fact $||M||_2\lesssim||M||_{\psi_{1/2}}=||\sqrt{M}||_{\psi_1}^2$, we have

$$\mathbb{P}(Z\geqslant 2\mathbb{E}Z_1+t)\leqslant\exp\left(-\frac{t^2}{3\overline{\zeta}_{n_1}^2}\right)+3\exp\left[-\left(\frac{\sqrt{t}}{K_1||\sqrt{M}||_{\psi_1}}\right)\right].$$

Note that $||h_k(X_1,Y_1)||_{\psi_1}\leqslant||h_k(X_1,Y_1)-\mathbb{E}h_k(X_1,Y_1)||_{\psi_1}+||\mathbb{E}h_k(X_1,Y_1)||_{\psi_1}\leqslant D_n+||\theta_{h,k}||_{\psi_1}=B_n$ and $\mathbb{E}h_k^4(X_1,Y_1)\lesssim\mathbb{E}|h_k(X_1,Y_1)-\theta_{h,k}|^4+|\theta_{h,k}|^4\leqslant D_n^2+|\theta_h|_\infty^4\lesssim B_n^4$. By Lemma 2.2.2 in [54],

$$\begin{aligned}||\sqrt{M}||_{\psi_1}^2&=||\max_{1\leqslant k\leqslant d}\max_{1\leqslant i\leqslant n_1}|h_k(X_i,Y_i)||_{\psi_1}^2\\&\leqslant K_3(\log(n_1d)\max_{k,i}||h_k(X_i,Y_i)||_{\psi_1})^2\end{aligned}$$

$$= K_3 \log^2(n_1 d) B_n^2.$$

Since $\overline{\zeta}_{n_1}^2 = \max_{1\leqslant k\leqslant d} \sum_{i=1}^{n_1} \mathbb{E}h_k^4(X_i, Y_i) \leqslant n_1 B_n^4$, by Lemma 8 in [16] and Jensen inequality,

$$\mathbb{E}Z_1 \leqslant K_4[\log^{1/2}(d)\overline{\zeta}_{n_1} + \log(d)||M||_2] \leqslant K_5(B_n^2 n_1^{1/2}\log^{1/2}(n_1 d) + B_n^2 \log^3(n_1 d)).$$

Therefore,

$$\begin{aligned}\mathbb{P}\Bigg(&\max_{1\leqslant k\leqslant d} |\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} h_k^2(X_i, Y_j) - \mathbb{E}h_k^2| \\ &\geqslant K_5 B_n^2[n_1^{-1/2}\log^{1/2}(d) + n_1^{-1}\log^3(n_1 d)] + t\Bigg) \\ &\leqslant \exp\left(-\frac{n_1 t^2}{3B_n^4}\right) + 3\exp\left[-\left(\frac{\sqrt{n_1 t}}{K_1 K_3 B_n \log(n_1 d)}\right)\right]\end{aligned}$$

Recall $\underline{n} = n_1$ and $\overline{n} = n_2$.
(i) If $\underline{n} \geqslant K_6 \log^5(\overline{n}d)\log^2(1/\zeta)$, then take $t_1^* = KB_n^2 \underline{n}^{-1/2}\log^{1/2}(\overline{n}d)$ such that

$$\frac{n_1 t_1^{*2}}{B_n^4} = \log(\overline{n}d) \gtrsim \log(1/\zeta) \text{ and } \frac{\sqrt{n_1 t_1^*}}{B_n \log(n_1 d)} \geqslant \underline{n}^{1/4}\log^{-3/4}(\overline{n}d) \gtrsim \log(1/\zeta).$$

(ii) If $\underline{n} \leqslant K_6 \log^5(\overline{n}d)\log^2(1/\zeta)$, then take $t_2^* = KB_n^2 \underline{n}^{-1}\log^3(\overline{n}d)\log(1/\zeta)$ such that

$$\begin{aligned}&\frac{n_1 t_2^{*2}}{B_n^4} \geqslant \underline{n}^{-1}\log^6(\overline{n}d)\log^2(1/\zeta) \gtrsim \log(1/\zeta) \qquad \text{and} \\ &\frac{\sqrt{n_1 t_2^*}}{B_n \log(n_1 d)} = \log^{1/2}(\overline{n}d)\log^{1/2}(1/\zeta) \gtrsim \log(1/\zeta).\end{aligned}$$

Observing $B_n^2[n_1^{-1/2}\log^{1/2}(d) + n_1^{-1}\log^3(n_1 d)] \lesssim t_1^* + t_2^* =: t^*$. Hence,

$$\mathbb{P}(\max_{1\leqslant k\leqslant d} |\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} h_k^2(X_i, Y_j) - \mathbb{E}h_k^2| \geqslant t^*) \leqslant \zeta. \qquad \square$$

A.4. Lemma for two-sample Hoeffding decomposition

Lemma A.5 (Tail bound of the maxima of the first order projection)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random vectors from F and Y is independently draw from G. Suppose $\theta_h = \mathbb{E}h(X_1, Y)$, $||h_k(X_1, Y) - \theta_{hk}||_{\psi_1} \leqslant D_n$ and $\mathbb{E}|h_k(X_1, Y) - \theta_{hk}|^{2+\ell} \leqslant D_n^\ell$ for all $k = 1, \ldots, d$ and $\ell = 1, 2$. Let $\zeta \in (0, 1)$ be a constant s.t. $\log(\zeta^{-1}) \leqslant K\log(nd)$. Define the projection $Gh(x) = \mathbb{E}h(x, Y) - \theta_h$. Then,*

$$\mathbb{P}\left(|\sum_{i=1}^{n} Gh(X_i)|_\infty \geqslant KD_n\{n^{1/2}\log^{1/2}(nd) \vee \log^2(nd)\}\right) \leqslant \zeta.$$

Therefore when $n \gtrsim \log^3(nd)$,

$$\mathbb{P}\left(|\sum_{i=1}^{n} Gh(X_i)|_\infty \geqslant KD_n n^{1/2}\log^{1/2}(nd)\right) \leqslant \zeta.$$

Proof of Lemma A.5. Let $Z = \max_{1\leqslant k\leqslant d}|\sum_{i=1}^n [Gh_k(X_i)]|$, $\sigma^2 = \max_{1\leqslant k\leqslant d}\sum_{i=1}^n \mathbb{E}[Gh_k(X_i)]^2$ and $M = \max_{1\leqslant i\leqslant n}\max_{1\leqslant k\leqslant d}|Gh_k(X_i)|$. By [1, Theorem 4],

$$\mathbb{P}\left(Z \geqslant 2\mathbb{E}Z + t\right) \leqslant \exp\left(-\frac{t^2}{3\sigma^2}\right) + 3\exp\left(-\frac{t}{K_1||M||_{\psi_1}}\right).$$

By Jensen inequality, $\mathbb{E}|Gh_k(X_i)|^2 = \mathbb{E}|\mathbb{E}[h_k(X_i,Y)-\theta_{hk}|X_i]|^2 \leqslant \mathbb{E}|h_k(X_i,Y)-\theta_{hk}|^2 \leqslant D_n$ and $||Gh_k(X_i)||_{\psi_1} \leqslant ||h_k(X_i,Y)-\theta_{hk}||_{\psi_1} \leqslant D_n$. So $\sigma^2 \leqslant nD_n$. By [1, Lemma 2.2.2] and [16, Lemma 8],

$$||M||_{\psi_1} \leqslant K_2\log(nd)\max_{i,k}||Gh_k(X_i)||_{\psi_1} \leqslant K_2 D_n\log(nd) \quad \text{and}$$

$$\mathbb{E}Z \leqslant K_3\{\sigma\sqrt{\log d} + ||M||_{\psi_1}\log d\} \leqslant K_4\{\sqrt{n\log(d)D_n} + \log(nd)\log(d)D_n\}.$$

Take $t^* = K_5 D_n\{n^{1/2}\log^{1/2}(nd) \vee \log^2(nd)\}$, simple calculation shows $\mathbb{P}(Z \geqslant t^*) \leqslant \zeta$. □

Lemma A.6 (Maximal inequality for canonical two-sample U-statistics)**.** *Let* $X_1,\ldots,X_{n_1}$ *and* $Y_1,\ldots,Y_{n_2}$ *be two independent sets of iid random vectors from* F *and* G*, respectively. Let* $\theta_h = \mathbb{E}h(X_1,Y_1)$*,* $n_1 \leqslant n_2$ *and* $d \geqslant 2$*. Suppose* $||h_m(X_1,Y_1)-\theta_{h,m}||_{\psi_1} \leqslant D_n$ *and* $\mathbb{E}|h_m(X_1,Y_1)-\theta_{h,m}|^{2+\ell} \leqslant D_n^\ell$ *for all* $m = 1,\ldots,d$ *and* $\ell = 1,2$*. We have*

$$\begin{aligned}&\mathbb{E}|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j)|_\infty\\ \leqslant& KD_n\log(d)\Big\{\log(d)\log(n_2 d) + (n_1 n_2)^{1/2}\\ &\qquad + [n_2\log(d)\log^2(n_2 d)]^{1/2} + [n_1 n_2^2\log(d)]^{1/4}\Big\}.\end{aligned}$$

Proof of Lemma A.6. The structure of this proof is similar to the one-sample version in [13, Thm 5.1]. By constructing randomization from iid Rademacher random variables (i.e. $\mathbb{P}(\epsilon_i = \pm 1) = \frac{1}{2}$ for all ϵ_i and ϵ'_j, $i = 1,\ldots,n_1, j = 1,\ldots,n_2$), [23, Thm 3.5.3] shows

$$\mathbb{E}|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j)|_\infty \leqslant K_1\mathbb{E}|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j)\epsilon_i\epsilon'_j|_\infty$$

Fix an $m = 1,\ldots,d$. Let Λ^m be a (n_1+n_2)-by-(n_1+n_2) matrix with zero diagonal blocks, where $\Lambda^m_{ij} = \breve{f}_m(X_i,Y_{j-n_1})$ if $1 \leqslant i \leqslant n_1, n_1+1 \leqslant j \leqslant n_1+n_2$ and

$\Lambda^m_{ij} = 0$, *otherwise*. Apply Hanson-Wright inequality [52, Thm 1] conditioning on $X_1^{n_1}$ and $Y_1^{n_2}$,

$$\mathbb{P}\left(\epsilon^T \Lambda^m \epsilon | X_1^{n_1} Y_1^{n_2}\right) \leqslant 2\exp[-K_2 \min\{\frac{t^2}{|\Lambda^m|_F^2}, \frac{t}{||\Lambda^m||_2}\}],$$

where $\epsilon^T = (\epsilon_1, \ldots, \epsilon_{n_1}, \epsilon'_1, \ldots, \epsilon'_{n_2})$ and $t > 0$. Denote $V_1 = \max_{1\leqslant m\leqslant d} |\Lambda^m|_F$ and
$V_2 = \max_{1\leqslant m\leqslant d} ||\Lambda^m||_2$. Let

$$t^* = \max\{V_1\sqrt{\frac{\log d}{K_2}}, V_2\frac{\log d}{K_2}\},$$

such that

$$\begin{aligned}\mathbb{E}[\max_{1\leqslant m\leqslant d} |\epsilon^T \Lambda^m \epsilon| | X_1^{n_1}, Y_1^{n_2}] &= \int_0^\infty \mathbb{P}\left(\max_{1\leqslant m\leqslant d} |\epsilon^T \Lambda^m \epsilon| \geqslant t | X_1^{n_1}, Y_1^{n_2}\right) dt \\ &\leqslant t^* + 2d\int_{t^*}^\infty \max\{\exp(-\frac{K_2 t^2}{V_1^2}), \exp(-\frac{K_2 t}{V_2})\}.\end{aligned}$$

Apply the tail bound of standard Gaussian random variables $1-\Phi(x) \leqslant \phi(x)/x$ for $x > 0$, and note that $d \geqslant 2$, we have

$$2d\int_{t^*}^\infty \exp(-\frac{K_2 t^2}{V_1^2})dt \leqslant \frac{V_1}{\sqrt{2K_2}}\int_{\sqrt{2\log d}}^\infty \exp(-\frac{s^2}{2})ds \leqslant \frac{V_1}{\sqrt{K_2 \log d}} \leqslant K_2 V_1.$$

Similarly,

$$2d\int_{t^*}^\infty \exp(-\frac{K_2 t}{V_2})dt \leqslant 2V_2/K_2.$$

By Jensen's inequality and the fact $V_2 \leqslant V_1$, we have

$$\begin{aligned}\mathbb{E}|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \breve{f}(X_i, Y_j)\epsilon_i \epsilon'_j|_\infty &\leqslant K_1 \mathbb{E}[t^* + K_2 V_1 + 2V_2/K_2] \leqslant K_3(\log d)\mathbb{E}V_1 \\ &\leqslant K_3(\log d)(\mathbb{E}\max_{1\leqslant m\leqslant d} |\Lambda^m|_F^2)^{1/2}. \qquad (A.9)\end{aligned}$$

Our last task is to bound

$$I \overset{def}{=} \mathbb{E}\max_{1\leqslant m\leqslant d} |\Lambda^m|_F^2 = \mathbb{E}[\max_{1\leqslant m\leqslant d}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \breve{f}_m^2(X_i, Y_j)].$$

Consider Hoeffding decomposition of $\breve{f}_m^2$,

$$\breve{f}_0^m(x_1, y_1) = \breve{f}_m^2(x_1, y_1) - \breve{f}_1^m(x_1) - \breve{f}_2^m(y_1) - \mathbb{E}\breve{f}_m^2,$$

where $\breve{f}_1^m(x_1) = \mathbb{E}\breve{f}_m^2(x_1, Y) - \mathbb{E}\breve{f}_m^2$ and $\breve{f}_2^m(y_1) = \mathbb{E}\breve{f}_m^2(X, y_1) - \mathbb{E}\breve{f}_m^2$ for $X \sim F \perp\!\!\!\perp Y \sim G$ are two random vectors independent from $X_1^{n_1}, Y_1^{n_2}$, and all x_1, y_1 from the measurable space of F and G, respectively. Then,

$$\mathbb{E}[\max_{1\leqslant m\leqslant d}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \breve{f}_m^2(X_i, Y_j)]$$

$$
\begin{aligned}
=&\mathbb{E}[\max_{1\leqslant m\leqslant d}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0^m(X_i,Y_j)+\breve{f}_1^m(X_i)+\breve{f}_2^m(Y_j)+\mathbb{E}\breve{f}_m^2]\\
\leqslant&\mathbb{E}[|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0^m(X_i,Y_j)|_\infty]+n_2\mathbb{E}[|\sum_{i=1}^{n_1}\breve{f}_1^m(X_i)|_\infty]\\
&+n_1\mathbb{E}[|\sum_{j=1}^{n_2}\breve{f}_2^m(Y_j)|_\infty]+n_1n_2\max_{1\leqslant m\leqslant d}\mathbb{E}\breve{f}_m^2. \qquad \text{(A.10)}
\end{aligned}
$$

Note that, conditioning on $X_1^{n_1}$, Hoeffding inequality shows for $t>0$

$$
\mathbb{P}\left(|\sum_{i=1}^{n_1}\breve{f}_1^m(X_i)\epsilon_i|>t|X_1^{n_1}\right)\leqslant 2\exp\left(-\frac{t^2}{2\sum_{i=1}^{n_1}\breve{f}_1^m(X_i)^2}\right).
$$

Denote $M=\max_{i,j,m}|\breve{f}_m(X_i,Y_j)|$. Following arguments in beginning and the symmetrization inequality [54, Lemma 2.3.1], we have

$$
\begin{aligned}
\mathbb{E}|\sum_{i=1}^{n_1}\breve{f}_1(X_i)|_\infty&\leqslant\sqrt{\log d}\,\mathbb{E}\sqrt{\max_m\sum_{i=1}^{n_1}\breve{f}_1^m(X_i)^2}\\
&\leqslant K_4\sqrt{\log d}\sqrt{n_1\max_m\mathbb{E}\breve{f}_m^4+\log d||M||_4^4}, \qquad \text{(A.11)}\\
\mathbb{E}|\sum_{j=1}^{n_2}\breve{f}_2(Y_j)|_\infty&\leqslant\sqrt{\log d}\,\mathbb{E}\sqrt{\max_m\sum_{j=1}^{n_2}\breve{f}_2^m(Y_j)^2}\\
&\leqslant K_5\sqrt{\log d}\sqrt{n_2\max_m\mathbb{E}\breve{f}_m^4+\log d||M||_4^4}, \qquad \text{(A.12)}
\end{aligned}
$$

$$
\mathbb{E}|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0(X_i,Y_j)|_\infty\leqslant\log d\,\mathbb{E}\sqrt{\max_m\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0^m(X_i,Y_j)^2}\leqslant K_6\log d\sqrt{I}||M||_2. \qquad \text{(A.13)}
$$

The last step of (A.11) comes from [13, Equation (58)]. The (A.12) follows the same procedure. And the first step of (A.13) is dealt the same way as (A.9) with

$$
\begin{aligned}
\mathbb{E}\sqrt{\max_m\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0^m(X_i,Y_j)^2}\leqslant 2\Big[&\mathbb{E}\sqrt{\max_m\sum_{i,j}\breve{f}_m^4(X_i,Y_j)}\\
&+\mathbb{E}\sqrt{\max_m\sum_{i,j}(\mathbb{E}[\breve{f}_m^2(X_i,Y_j')|X_1^{n_1}])^2}\\
&+\mathbb{E}\sqrt{\max_m\sum_{i,j}(\mathbb{E}[\breve{f}_m^2(X_i',Y_j)|Y_1^{n_2}])^2}\\
&+\mathbb{E}\sqrt{\max_m\sum_{i,j}(\mathbb{E}\breve{f}_m^2(X_i,Y_j))^2}\Big]
\end{aligned}
$$

$$\leqslant K_6\sqrt{I}\sqrt{\mathbb{E}M^2}.$$

Since $||h_m(X_1,Y_1)-\theta_{h,m}||_{\psi_1} \leqslant D_n$ and $\mathbb{E}|h_m(X_1,Y_1)-\theta_{h,m}|^{2+\ell} \leqslant D_n^\ell$, we know $\max_m \mathbb{E}\check{f}_m^4 \leqslant D_n^2$ and $||M||_4 \lesssim ||M||_{\psi_1} \leqslant K_7 D_n \log(n_1 n_2 d) \leqslant 2K_7 D_n \log(n_2 d)$. In addition, we have $D_q = \max_m[\mathbb{E}|\check{f}_m(X,Y)|^q]^{1/q} \lesssim D_n$. Plug (A.11)-(A.13) in (A.10) and the solution of quadratic inequality for I gives

$$I \leqslant K_8\Big\{||M||_2^2\log^2 d + n_1 n_2 D_2 + n_2\sqrt{\log d}\sqrt{n_1 D_4 + \log d||M||_4^4} + n_1\sqrt{\log d}\sqrt{n_2 D_4 + \log d||M||_4^4}\Big\}.$$

Therefore, the square-root of I is less than the square-root of each term on RHS. Plug the result in (A.9). A simplified result is obtained in the statement of Lemma A.6. □

A.5. Additional simulation and tables

TABLE 11

Powers report of our method using linear kernel. Here, $n = 500, p = 600, \alpha = 0.05$ and change point locations are $t_m = m/n = 5/10, 3/10, 1/10$.

	Gaussian			t_6			ctm-Gaussian		
$\|\theta\|_\infty$	I	II	III	I	II	III	I	II	III
$t_m = 5/10$									
0	0.042	0.050	0.032	0.058	0.060	0.040	0.052	0.050	0.048
0.28	0.100	0.178	0.082	0.082	0.134	0.072	0.066	0.102	0.070
0.44	0.436	0.628	0.390	0.186	0.420	0.212	0.154	0.356	0.200
0.63	0.886	0.970	0.896	0.610	0.828	0.590	0.554	0.810	0.578
0.84	0.996	1	0.996	0.926	0.988	0.912	0.918	0.990	0.910
$t_m = 3/10$									
0	0.030	0.042	0.066	0.038	0.060	0.026	0.030	0.072	0.060
0.28	0.088	0.216	0.108	0.068	0.124	0.036	0.036	0.156	0.082
0.44	0.414	0.738	0.384	0.222	0.418	0.178	0.150	0.440	0.200
0.63	0.890	0.996	0.908	0.594	0.878	0.634	0.524	0.846	0.570
0.84	0.998	1	0.998	0.930	0.998	0.960	0.940	0.996	0.940
$t_m = 1/10$									
0	0.054	0.060	0.050	0.064	0.058	0.060	0.054	0.054	0.064
0.63	0.082	0.210	0.086	0.078	0.126	0.082	0.058	0.118	0.086
0.84	0.190	0.472	0.224	0.144	0.278	0.120	0.116	0.240	0.120
1.08	0.446	0.768	0.446	0.268	0.492	0.252	0.208	0.470	0.230
1.35	0.756	0.966	0.770	0.486	0.762	0.516	0.444	0.760	0.462
2.00	0.998	1.000	0.998	0.954	0.996	0.960	0.962	0.994	0.956

In this section, we test the performance of the WBS-type procedure (Algorithm 1). Let $n = 500, p = 600, B = 200, B_W = 200, n' = 0.2n = 100$ and data be i.i.d. Gaussian distributed with covariance structure III. The two change points are $(m_1, m_2) = (150, 300)$ and only the k-th component of the k-th change point has signal $\theta_1^{(1)} = \theta_2^{(2)} = \delta \neq 0$. The powers along δ for each $\alpha = 0.01, 0.05, 0.1$ are shown in the rows of Table 13. We find that when $\delta = 0$, the power is close to the nominal levels, respectively. Besides, the power grows as δ increases.

TABLE 12
Powers report of our method using sign kernel. Here, $n = 500, p = 600, \alpha = 0.05$ and change point locations are $t_m = m/n = 5/10, 3/10, 1/10$.

	Gaussian			t_6			ctm-Gaussian				Cauchy		
$\|\theta\|_\infty$	I	II	III	I	II	III	I	II	III	$\|\theta\|_\infty$	I	II	III
$t_m = 5/10$													
0	0.056	0.043	0.048	0.066	0.062	0.066	0.067	0.032	0.055	0	0.054	0.062	0.039
0.28	0.136	0.289	0.147	0.110	0.229	0.099	0.105	0.204	0.083	0.71	0.403	0.651	0.432
0.44	0.566	0.870	0.624	0.452	0.738	0.479	0.364	0.674	0.397	1.23	0.971	1	0.981
0.63	0.977	1	0.971	0.915	0.996	0.913	0.854	0.980	0.872	1.91	1	1	1
0.84	1	1	1	0.998	1	1	0.988	1	0.998	2.79	1	1	1
$t_m = 3/10$													
0	0.049	0.037	0.047	0.039	0.068	0.056	0.051	0.049	0.055	0	0.055	0.035	0.065
0.28	0.070	0.154	0.068	0.058	0.148	0.078	0.073	0.104	0.083	0.71	0.257	0.386	0.280
0.44	0.342	0.619	0.342	0.218	0.451	0.230	0.189	0.427	0.240	1.23	0.829	0.969	0.876
0.63	0.830	0.982	0.848	0.663	0.912	0.706	0.593	0.872	0.628	1.91	1	1	1
0.84	0.992	1	0.996	0.975	1	0.973	0.941	0.994	0.945	2.79	1	1	1
$t_m = 1/10$													
0	0.042	0.046	0.065	0.053	0.046	0.046	0.050	0.048	0.050	0	0.057	0.059	0.080
0.63	0.078	0.139	0.082	0.063	0.107	0.078	0.060	0.110	0.075	1.91	0.216	0.394	0.243
0.84	0.147	0.309	0.155	0.097	0.231	0.132	0.104	0.218	0.110	2.79	0.410	0.680	0.433
1.08	0.305	0.580	0.336	0.214	0.458	0.248	0.183	0.423	0.222	3.95	0.627	0.873	0.647
1.35	0.523	0.796	0.588	0.405	0.706	0.439	0.367	0.660	0.351	5.47	0.806	0.931	0.806
2.00	0.891	0.992	0.931	0.794	0.964	0.834	0.815	0.950	0.828	10.02	0.937	0.980	0.933

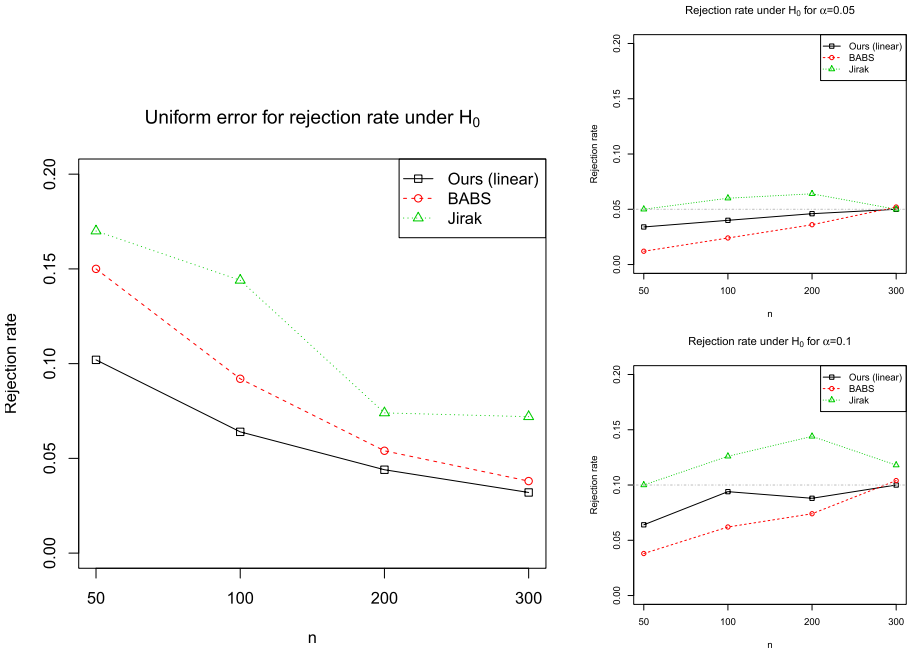
TABLE 13
Power of U-statistics based WBS-type testing.

Power	δ				
	0	0.317	0.733	1.282	2.004
$\alpha = 0.01$	0.012	0.046	0.806	0.900	0.908
$\alpha = 0.05$	0.032	0.100	0.818	0.898	0.926
$\alpha = 0.1$	0.088	0.198	0.882	0.926	0.938

A.6. Additional comparisons with `BABS` *and* `Jirak`

We further compare the size control of [61, `BABS`], [38, `Jirak`] and our linear kernel approach under H_0. As suggested, we fix $p = 100$ and vary n from 50 to 300. The bootstrap repeat is $B = 200$, ξ_i are i.i.d. Gaussian with dependence structure III, and each simulation repeats 500 times. The boundary removal parameters in `BABS` and `Jirak` are both $0.1n$.

From Figure 10a, we can find that all three methods have a decreasing trend when n grows, but our U-statistic approach has the lowest uniform error-in-size $\sup_{\alpha\in[0,1]} |\hat{R}(\alpha) - \alpha|$ under each choice of n. This confirms that our U-statistic test performs better than the others for small n. From Figure 10b where empirical rejection rates at $\alpha = 0.05, 0.1$ are provided, we may observe that the difference among three methods diminishes for $n = 300$. However, our approach is closer to the corresponding nominal significance level except for $n = 50$. Therefore, the simulation indicates that the no-boundary-removal property in our proposed test is beneficial to size control under small sample size.



(a) Uniform error-in-size, $\sup_{\alpha\in[0,1]}|\hat{R}(\alpha)-\alpha|$ under H_0.

(b) Empirical $\hat{R}(\alpha)$ at $\alpha = 0.05, 0.1$.

FIG 10. *Comparison of size control among* `BABS`, `Jirak` *and our method using linear kernel.*

Acknowledgments

This work is completed in part with the high-performance computing resource provided by the Illinois Campus Cluster Program at UIUC. The authors are grateful to the editor, associate editor, and referee for their insightful comments.

References

[1] ADAMCZAK, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability* **13** 1000-1034. MR2424985

[2] ARLOT, S., CELISSE, A. and HARCHAOUI, Z. (2019). A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research* **20** 1–56. MR4048973

[3] ASTON, J. A. and KIRCH, C. (2012). Detecting and estimating changes in dependent functional data. *Journal of Multivariate Analysis* **109** 204–220. MR2922864

[4] Aston, J. A., Kirch, C. et al. (2012). Evaluating stationarity via change-point alternatives with applications to fMRI data. *The Annals of Applied Statistics* **6** 1906–1948. MR3058688

[5] Aston, J. A., Kirch, C. et al. (2018). High dimensional efficiency with applications to change point tests. *Electronic Journal of Statistics* **12** 1901–1947. MR3815301

[6] Aue, A., Gabrys, R., Horváth, L. and Kokoszka, P. (2009). Estimation of a change-point in the mean function of functional data. *Journal of Multivariate Analysis* **100** 2254–2269. MR2560367

[7] Aue, A., Hörmann, S., Horváth, L., Reimherr, M. et al. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics* **37** 4046–4087. MR2572452

[8] Bai, J. (2010). Common breaks in means and variances for panel data. *Journal of Econometrics* **157** 78–92. MR2652280

[9] Barigozzi, M., Cho, H. and Fryzlewicz, P. (2018). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics* **206** 187-225. MR3840788

[10] Bhattacharjee, M., Banerjee, M. and Michailidis, G. (2019). Change Point Estimation in Panel Data with Temporal and Cross-sectional Dependence. *arXiv preprint arXiv:1904.11101*. MR4119175

[11] Brault, V., Ouadah, S., Sansonnet, L. and Lévy-Leduc, C. (2018). Nonparametric multiple change-point estimation for analyzing large Hi-C data matrices. *Journal of Multivariate Analysis* **165** 143–165. MR3768758

[12] Chen, L., Wang, W. and Wu, W. (2019). Inference of Break-Points in High-Dimensional Time Series. *Available at SSRN 3378221*.

[13] Chen, X. (2018). Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *The Annals of Statistics* **46** 642–678. MR3782380

[14] Chen, X. and Kato, K. (2019). Randomized incomplete U-statistics in high dimensions. *The Annals of Statistics* **47** 3127-3156. MR4025737

[15] Chen, X. and Kato, K. (2020). Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. *Probability Theory and Related Fields* **176** 1097-1163. MR4087490

[16] 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

[17] Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability* **45** 2309-2352. MR3693963

[18] Chernozhukov, V., Chetverikov, D. and Koike, Y. (2020). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *arXiv:2012.09513*.

[19] Cho, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics* **10** 2000-2038. MR3522667

[20] Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of*

the Royal Statistical Society: Series B **77** 475-507. MR3310536
[21] Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. New York: Wiley. MR2743035
[22] Csörgo, M. and Horváth, L. (1988). Invariance principles for change-point problems. *Journal of Multivariate Analysis* **27** 151-168. MR0971179
[23] de la Peña, V. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Springer. MR1666908
[24] Dette, H., Pan, G. and Yang, Q. (2018). Estimating a change point in a sequence of very high-dimensional covariance matrices. *arXiv:1807.10797*.
[25] Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics* **47** 2051-2079. MR3953444
[26] Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics* **42** 2243-2281. MR3269979
[27] Gombay, E. (2001). U-statistics for change under alternatives. *Journal of Multivariate Analysis* **78** 139–158. MR1856269
[28] Gombay, E. and Horváth, L. (1995). An application of U-statistics to change-point analysis. *Acta Scientiarum Mathematicarum* **60** 345–358. MR1348699
[29] Gombay, E. and Horváth, L. (2002). Rates of convergence for U-statistic processes and their bootstrapped versions. *Journal of Statistical Planning and Inference* **102** 247–272. MR1896486
[30] Hawkins, D. M. and Deng, Q. (2010). A Nonparametric Change-Point Control Chart. *Journal of Quality Technology* **42** 165-173.
[31] Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics* **34** 598-611. MR0152070
[32] Holmes, M., Kojadinovic, I. and Quessy, J.-F. (2013). Nonparametric tests for change-point detection à la Gombay and Horváth. *Journal of Multivariate Analysis* **115** 16–32. MR3004542
[33] Horváth, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *The Annals of Statistics* **21** 671–680. MR1232511
[34] Horváth, L. and Hušková, M. (2012). Change-point detection in panel data. *Journal of Time Series Analysis* **33** 631–648. MR2944843
[35] Horváth, L., Kokoszka, P. and Steinebach, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis* **68** 96–119. MR1668911
[36] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2** 193-218.
[37] James, N. A. and Matteson, D. S. (2015). ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software* **62** 1-25. MR3180567
[38] Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics* **43** 2451-2483. MR3405600
[39] Killick, R. and Eckley, I. (2014). changepoint: An R package for changepoint analysis. *Journal of statistical software* **58** 1–19.

[40] Kirch, C. and Stoehr, C. (2019). Sequential change point tests based on U-statistics.
[41] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces: isoperimetry and processes*. New York: Springer-Verlag. MR1102015
[42] Lee, S., Liao, Y., Seo, M. H. and Shin, Y. (2018). Oracle estimation of a change point in high-dimensional quantile regression. *Journal of the American Statistical Association* **113** 1184–1194. MR3862349
[43] Lee, S., Seo, M. H. and Shin, Y. (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** 193–210. MR3453652
[44] Liu, H., Gao, C. and Samworth, R. J. (2021). Minimax rates in sparse, high-dimensional change point detection. *The Annals of Statistics* **49** 1081–1112. MR4255120
[45] Minami, K. (2020). Estimating piecewise monotone signals. *Electronic Journal of Statistics* **14** 1508–1576. MR4082476
[46] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. MR0652932
[47] Niu, Y. S., Hao, N. and Zhang, H. (2016). Multiple change-point detection: A selective overview. *Statistical Science* **31** 611–623. MR3598742
[48] Padilla, O. H. M., Yu, Y., Wang, D. and Rinaldo, A. (2019). Optimal nonparametric change point detection and localization. *arXiv:1905.10019*. MR4255296
[49] Pettitt, A. N. (1979). A Non-Parametric Approach to the Change-Point Problem. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28** 126–135. MR0539082
[50] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66** 846–850.
[51] Robbins, M., Gallagher, C., Lund, R. and Aue, A. (2011). Mean shift testing in correlated data. *Journal of Time Series Analysis* **32** 498–511. MR2835683
[52] Rudelson, M., Vershynin, R. et al. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability* **18**. MR3125258
[53] van der Vaart, A. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. MR1652247
[54] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. New York: Springer. MR1385671
[55] Vogel, D. and Wendler, M. (2017). Studentized U-quantile processes under dependence with applications to change-point analysis. *Bernoulli* **23** 3114-3144. MR3654800
[56] Wang, R., Volgushev, S. and Shao, X. (2019). Inference for Change Points in High Dimensional Data. *arXiv:1905.08446*.
[57] Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of Royal Statistical Society: Series*

B (Statistical Methodology) **80** 57-83. MR3744712

[58] Wang, Y., Wu, C., Ji, Z., Wang, B. and Liang, Y. (2011). Non-parametric change-point method for differential gene expression detection. *PloS one* **6** e20060.

[59] Xie, Y. and Siegmund, D. (2013). Sequential multi-sensor change-point detection. In *2013 Information Theory and Applications Workshop (ITA)* 1–20. IEEE. MR3099117

[60] Yau, C. Y. and Zhao, Z. (2016). Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** 895–916. MR3534355

[61] Yu, M. and Chen, X. (2021). Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **83** 247-270. MR4250275

[62] Zhong, P.-S. and Li, J. (2016). Test for Temporal Homogeneity of Means in High-dimensional Longitudinal Data. *arXiv:1608.07482*.