HIGH-DIMENSIONAL CHANGE POINT DETECTION FOR MEAN AND LOCATION PARAMETERS

BY

MENGJIA YU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Doctoral Committee:

Associate Professor Xiaohui Chen, Chair and Director of Research
Professor Annie Qu, University of California, Irvine
Professor Xiaofeng Shao
Professor Douglas G. Simpson

Abstract

Change point inference refers to detection of structural breaks of a sequence observation, which may have one or more distributional shifts subject to models such as mean or covariance changes. In this dissertation, we consider the offline multiple change point problem that the sample size is fixed in advance or after observation. In particular, we concentrate on high-dimensional setup where the dimension p can be much larger than the sample size n and traditional distribution assumptions can easily fail. The goal is to employ non-parametric approaches to identify change points without involving intermediate estimation to cross-sectional dependence.

In the first part, we consider cumulative sum (CUSUM) statistics that are widely used in the change point inference and identification. We study two problems for high-dimensional mean vectors based on the ℓ^∞-norm of the CUSUM statistics. For the problem of testing for the existence of a change point in an independent sample generated from the mean-shift model, we introduce a Gaussian multiplier bootstrap to calibrate critical values of the CUSUM test statistics in high dimensions. The proposed bootstrap CUSUM test is fully data-dependent and it has strong theoretical guarantees under arbitrary dependence structures and mild moment conditions. Specifically, we show that with a boundary removal parameter the bootstrap CUSUM test enjoys the uniform validity in size under the null and it achieves the minimax separation rate under the sparse alternatives when $p \gg n$. Once a change point is detected, we estimate the change point location by maximizing the ℓ^∞-norm of the generalized CUSUM statistics at two different weighting scales. The first estimator is based on the covariance stationary CUSUM statistics, and we prove its consistency in estimating the location at the nearly parametric rate $n^{-1/2}$ for sub-exponential observations. The second estimator is based on non-stationary CUSUM statistics, assigning less weights on the boundary data points. In the latter case, we show that it achieves the nearly best possible rate of convergence on the order n^{-1}. In both cases, dimension impacts the rate of convergence only through the logarithm factors, and therefore consistency of the CUSUM location estimators is possible when p is much larger than n. In the presence of multiple change points, we propose a principled bootstrap-assisted binary segmentation (BABS) algorithm to dynamically adjust the change point detection rule and recursively estimate their locations. We derive its rate of convergence under suitable signal separation and strength conditions. The results derived are non-asymptotic and we provide extensive simulation studies to assess the finite sample performance. The empirical evidence shows an encouraging agreement with our theoretical results.

In the second part, we analyze the problem of change point detection for high-dimensional distributions in a location family. We propose a robust, tuning-free (i.e., fully data-dependent), and easy-to-implement change point test formulated in the multivariate U-statistics framework with anti-symmetric and nonlinear kernels. It achieves the robust purpose in a non-parametric setting when CUSUM statistics are sensitive to outliers and heavy-tailed distributions. 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 numeric studies.

To my husband, Dawei Ding, and our families.

Acknowledgments

First and foremost, I owe my heartiest appreciation and immeasurable admiration to my advisor, Professor Xiaohui Chen, for providing invaluable guidance throughout my research. His immense knowledge, broad vision and continual inspiration conveyed the spirit of sincerity in regard to research and helped me to transform scientific arguments into structured association. He granted me with generous time and enlightened discussion to address my confusion whenever I needed. I am ineffably indebted to his conscientious patience and encouragement to accomplish my PhD study. I consider it a privilege to have worked under his supervision that pushed me farther than I thought I could go.

I humbly extend my utmost sense of reverence and gratitude to the rest of my committee members: Professor Annie Qu, Professor Xiaofeng Shao and Professor Douglas Simpson, for their constructive comments on this dissertation project and consistent advice during my pursuit of doctoral degree. It was a pleasure to benefit from their collective insights and to make acquaintance with their remarkable personalities. I am rewarded from each of my committee members with their erudition and diligence lifelong.

In addition, I am deeply thankful to my husband Dawei Ding for his unconditional love and endless support in my doctoral study. Without his dedication and consolation, my achievement would never be possible. I am also grateful to our families who always expressed confidence and encouragement to us with no hesitation.

The last but not least gratitude goes to all of the Department of Statistics who directly or indirectly helped me over the last five years. The faculty, staff members and colleagues manifested that we formed a whole greater than the sum of its parts. They kept me going when I was frustrated in work and life. I will always remember their warm and motivating words and treasure the strength they passed to me to fight difficulties in the future.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the era of "Big-Data", structure heterogeneity issue has received enormous attention in various scientific and engineering fields with application to stock market analysis, quality control, genomics detection and many others. Let $X_i \sim F_i, i = 1, \ldots, n$ be a sequence of independent random vectors taking values in $\mathbb{R}^p$. In general, the question of change-point analysis in high-dimension arises in the following statistical testing:

$$H_0 : F_1 = \cdots = F_n,$$

$$H_1 : F_1 = \cdots = F_{m_1} \neq F_{m_1+1} = \cdots = F_{m_2} \neq F_{m_2+1} = \cdots F_{m_\nu} \neq F_{m_\nu+1} = \cdots F_n,$$

for some unknown $\nu \in \mathbb{N}^+$ and change point locations $1 < m_1 < \cdots < m_\nu < n$. The structural stability problem usually can be boiled down to single change point case of $\nu = 1$ (so-called *at-most-one-change* problem) and extension to multiple change points. The stream to tackle single change point case starts from univariate Gaussian distributed $\{F_i\}_{=1}^n$ that have mean shift [32] or variance shift [25] or both [60] and then to a generalization to location and scale parameters shift [102]. Non-parametric approach such as U-statistics or KolmogorovSmirnov type statistic and semi-parametric approach using empirical likelihood are studied in univariate case [84, 24, 82, 112]. In multivariate setup, we refer to [38] for a rigorous and comprehensive study on the likelihood approach and non-parametric approaches for the aforementioned testing as well as in broader regression and time series models together with sequential methods.

Recently, due to the explosive data enrichment in modern applications where the number of variables p is comparable to or even much larger than the sample size n, classical methods are typically inapplicable and the asymptotic theories developed for a fixed dimension do not generally hold. For example, suppose $\{F_i\}_{i=1}^n$ are Gaussian distribution with mean μ_i and covariance Σ. Without loss of generality, we may assume $\mu_1 = \mathbf{0}$. Under the single change point scenario, the log-ratio of the maximized likelihoods between H_1 with a change point at $s = 1, \ldots, n-1$ and H_0 is

$$\log(\Lambda_s) = Z_n(s)^\top \Sigma^{-1} Z_n(s)/2, \tag{1.1}$$

where

$$Z_n(s) = \sqrt{\frac{s(n-s)}{n}} \left(\frac{1}{s} \sum_{i=1}^{s} X_i - \frac{1}{n-s} \sum_{i=s+1}^{n} X_i \right) \tag{1.2}$$

are the *cumulative sum* (CUSUM) statistics [38], a sequence of the normalized mean differences before and after s. Then H_0 is rejected if $\max_{1 \leqslant s < n} \log(\Lambda_s)$ is larger than a critical value. If s is restricted to only one location, i.e., the change point location is known, then the problem reduce to a multivariate two-sample mean test that can be solved by Hotelling's T-squared statistic $T^2 = Z_n(s)^\top \hat{\Sigma}^{-1} Z_n(s)$ where $\hat{\Sigma} = \frac{1}{s-1} \sum_{i=1}^{s} (X_i - \bar{X}_s^-)(X_i - \bar{X}_s^-)^T + \frac{1}{n-s-1} \sum_{i=s+1}^{n} (X_i - \bar{X}_s^+)(X_i - \bar{X}_s^+)^T$ for $p < s < n - p$ and $\bar{X}_s^- = s^{-1} \sum_{i=1}^{s} X_i$ and $\bar{X}_s^+ = (n-s)^{-1} \sum_{i=s+1}^{n} X_i$. In the high-dimensional setting (i.e. $p \gg n$), $\hat{\Sigma}$, estimation of Σ, itself becomes a challenging problem. The spectral norm consistency of Σ (or the inverse Σ^{-1}) is only possible under additional structural assumptions (such as sparsity or low-rankness) on the covariance matrix [16, 17, 22, 23, 30, 11], which may be violated in practical applications. In contrast, tests based on the CUSUM statistics in (1.2) do not involve Σ and they are more robust to the misspecification on covariance structures. Therefore, this motivates us to study the problems of change point testing and estimation based on the high-dimensional CUSUM statistics.

Although extensive research have been conducted on CUSUM approaches, there is still a number of attracting questions in high-dimension including:

1. How to obtain the distribution of the CUSUM statistics when Gaussianity is violated?

2. What range can s take in both testing and estimation?

3. Is it possible to generate mean test to a broader context?

4. How robust an algorithm/approach can be in terms of various aspects?

For the first question, bootstrapping becomes an option to retrieve latent information and calibrate unknown distribution $\{F_i\}_{=1}^n$. A large volume of papers studied the performance of resampling bootstrap [12], weighted bootstrap [50] and multiplier bootstrap [21, 20, 42, 65], where the last technique especially enjoys the convenience of maintaining sequence structure in change point analysis. For the second question, it should be noted that boundary changes that happen close to end points of 1 or n are hard to detect due to insufficient data points in one population. Under H_0, asymptotic theories allow $s/n \in (0, 1)$ [38]. But the connection between signal of change in F_i and detectable range is implicit and not clearly stated. The third and fourth questions are practical concerns when outlier exists, distributional assumption is not satisfied or F_i may not differ in means but in other aspects. To generalize to high-dimensional methods from univariate case, a fundamental problem therein is to integrate changes carefully and determine relative thresholds properly for the sake of testing and estimation.

In this dissertation, we study finite sample change point detection for high-dimensional data using multiplier bootstraps. Specifically, we propose two ℓ^∞-type of statistics based on CUSUM statistics and U-statistics, respectively, to deal with change in mean vector and location parameter. Both of them focus on sparse change that may occur in a small subset of dimension coordinates of X_i at an unknown location m_1. The four motivating questions will be answered in the perspective of methodology and theory.

1.1 Change point in mean vector and our contributions

In the first part of my dissertation, we focus on inference and identification for structural stability in mean vector using CUSUM statistics. CUSUM statistics have been widely used with a long history pioneered by [83] and followed by two lines in literature: sequential detection for *online* change [31, 57, 68, 85] and fixed sample size tests and estimation for *offline* change [104, 24, 90], latter of which is the core point of interest in this dissertation. Let $\{F_i\}$ be a location family such that $F_i(x) = F_1(x - \mu_i)$. In recent development for high-dimensional data, testing and estimation for mean-shift are the most intrinsic questions that have been studied in [43, 65] and in [37, 65, 99, 36], respectively. Among them, [43] restricted on testing for Gaussian distribution with diagonal covariance matrix, [99] considered estimation under the same setting and advanced in extension on temporal and spacial dependence with a brief discussion, [37, 65, 36] considered time series sequence and introduced *boundary removal* concepts without explicit analysis, only [65] provided asymptotic consistency of the bootstrap procedure.

1.1.1 Bootstrapped CUSUM test for single change point

Starting from the simplest single mean-shift model (where data-generating mechanism produces independent X_i with mean change at most once), we propose a testing procedure based on the test statistic $T_n = \max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |Z_n(s)|_\infty$, which aggregates mean-shift signals over sample sequence and targets to the maximum change across coordinates. To achieve distribution approximation of T_n where observations have sub-exponential tails or polynomial tails, we apply Gaussian multiplier bootstrap as a computable data-driven approach. Specifically, we define bootstrapped test statistics that capture the distribution of T_n rather than an upper quantile for a pre-specified significance level. This allows uniform control over size (Type I error). In addition, power can be guaranteed when signal size of change is beyond a lower bound that achieves minimax rate for sparse alternatives. The proposed test is fully data-dependent with no tuning parameter except for the boundary removal parameter $\underline{s}$. Under mild distributional conditions, we demonstrate theoretical requirement on $\underline{s}$ to justify the difficulty of detecting boundary changes using CUSUM statistics.

The multiplier bootstrap calibrates distribution of T_n without referring to extreme value limiting theories that is known to be slow [87, 53, 65]. Aside from mild moment conditions, no structural condition (in contrast with [43, 99]) nor upper bound of *spatial* correlation (in contrast with [36, 65]) on *cross-sectional* covariance matrix is assumed. So our approach is robust to (i) any distributions that have sub-exponential or polynomial tails, (ii) any correlation among coordinates especially heavy dependence (see simulation). In addition, the dimension p only affect the $\underline{s}$ through $\log(np)$. The searching boundary $\underline{s}$ (or $n-\underline{s}$) is no longer polynomially bounded away from the two endpoints unlike the works [37, 36].

1.1.2 CUSUM based estimators for single change point

As a parallel problem to inference, we continue to study the identification issue and propose two consistent change-point estimators based on CUSUM statistics with two different weights over data sequence. One estimator is covariance stationary and related to T_n. The other one provides the best possible convergence rate of n^{-1} up to logarithm factors of np. We establish non-asymptotic consistency for both estimators that allow p grows sub-exponentially fast in n and can search over the whole data sequence. In application, they can also be tailored to truncated versions in order to adapt to the boundary removal concept in testing. Again, they are robust to distributional tails and spatial correlations that are mentioned above.

1.1.3 Algorithm for multiple change points

Despite the simplicity in single change-point problem for independent data, the appealing properties mentioned above can extend to multiple change-point scenarios where temporal dependence may exist. For multiple change-point analysis of independent data, binary segmentation [45] is a natural technique to recursively test and estimate change points in segments before and after a claimed change point. We propose *bootstrap-assisted binary segmentation* (BABS) that combines our test and estimators to consistently search and locate change points until nothing shows significance. To accommodate time-series, a block multiplier bootstrap under single change-point scenario is designed empirically and can be combined with binary segmentation as well.

We would like to emphasize that BABS inherits the robustness from both lines of testing and estimation without introducing additional tuning parameters, and we derive the consistency theoretically. We answered the four questions by

1. dealing with broader distributions with sub-exponential and polynomial tails rather than Gaussian distribution;
2. deriving explicit rate on boundary removal parameter $\underline{s}$;
3. discussing a natural extension to detect covariance change point for sub-Gaussian F_i;
4. relaxing weakly dependent requirement for the covariance of F_i.

To relate our work with [65, Bootstrap I and Bootstrap II], which need estimation of single change point prior to testing, our BABS is designed reversely (or say simultaneously). Arithmetically, testing and estimation are two separate problems that should be able to conduct independently. However, in [65], (long-run) covariance estimators for each coordinates must be normalized according to a possible change point before performing bootstrapping under single alternative H_1. Though, [65, Bootstrap III] proposed a "naive" way to entirely remove the estimation of change point in bootstrap under H_1, their test statistic [65, Eq (1.2)] still relies on (long-run) covariance estimators. Therefore, there is a fundamental difficulty to extend their approach for a global detection of multiple change points except to, for instance, apply wild bootstrap technique [45] with sacrifice on computation cost. Thus, our BABS is more aligned with statistical philosophy with no additional estimation step nor threshold selection that of vital importance to other multiple change point estimation algorithms [37, 99, 36].

1.2 Change point in location parameter and our contributions

In the second part, we consider the class of distribution where mean vector is not well-defined. Recall the location-shift model such that $F_i(x) = F_1(x - \mu_i)$. When it comes to i.i.d. observations whose latent distribution is from a heavy-tailed location family that does not have finite mean, CUSUM-type approaches will fail due to violation on moment conditions. As a consequence, a non-linear projection of data must be imposed to separate H_0 (no change point) and H_1 (at most one location shift over the sequence of data). This motivates us to investigate new technique for high-dimensional robust change-point analysis.

U-statistics, which sum up all permutations of samples filtered by a kernel, produce robust and unbiased estimation of the kernel function after proper normalization. Such function must be selected properly in order to reflect changes in location parameter. In the common choice of symmetric ones, high-order kernels (e.g. in order-four) have been designed to cancel out within-sample means or location parameters [38, 97]. On the other hand, if the single change point is assumed known to be s, the problem falls to investigation of two-sample (before and after s) U-statistics that can separate between-sample difference. Then a natural extension for unknown change point case is to take the test statistic as the maximum of between-sample differences over $s = 1, \ldots, n-1$ [25, 77, 73, 52, 40]. However, such modifications are computationally expensive to cancel out noises by lifting to high-order kernels and unnecessary to screen whole sequence for testing only and no intention on estimation.

To remedy the issue of robustness and efficiency, we propose a novel ℓ^∞-type testing statistic based on one-sample order-two U-statistics coupled with anti-symmetric kernels in Chapter 3. Specifically, our statistic is defined as

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

where $h : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^d$ is an *anti-symmetric* kernel, i.e., $h(x, y) = -h(y, x)$ for all $x, y \in \mathbb{R}^p$, which plays a key role in testing. Under the null H_0, the mean of T_n is always 0, while under the alternative H_1, the within-sample noise cancels out and the between-sample signal is properly preserved by kernel. In the perspective of computational cost, the general order is $O(n^2 p)$ for the example of robust sign-kernel $h(x, y) = sign(x - y)$ (component-wise) and it can be reduced to $O(np)$ by using the one-pass linear-kernel $h(x, y) = x - y$.

Our proposed U-statistics approach is robust, tuning-free and easy-to-implement. Gaussian multiplier bootstrap is designed to calibrate the distribution of test statistic. Subject to mild moment conditions on kernels, we derive the uniform rates of convergence for distribution approximation and the lower bound of location-change signal through kernel. It should be noted that no boundary removal procedure is required since the one-sample U-statistic T_n integrates "two-sample" information without referring each data point. We can answer the four questions raised in the beginning by

1. considering kernel projection to release distributional assumption;
2. performing global testing without boundary removal;
3. leaving possibility of other tests to choices of kernel h;
4. extending to location parameter of F_i which is not necessary to have mean.

In the presence of multiple change points, the U-statistic based test is valid if location shifts accumulate ideally. Since the signal cancellation may nullify power, a block testing is designed to gain single change point structure with sacrifice of sample size. Due to the advantage that our test does not screen any difference before and after each point, it limits to extend the test statistic to a change point estimator as in the CUSUM-based approach. As a consequence, the binary segmentation based forward detection cannot directly apply to our framework. However, a backward searching algorithm (BD) can play a role of estimation without introducing external estimator. Specifically, BD repeatedly merges two consecutive data segments whose union fail to reject H_0. Therefore, it is more powerful especially when there are short sequences with location shifts.

The rest of this dissertation is organized as follows. The CUSUM based test and estimator together with the BABS algorithm for mean change are elaborated in Chapter 2 [106]. The U-statistics based test as well as the BD algorithm is described in Chapter 3 [107]. Main results, numerical studies and proofs are given within each chapter, respectively.

Chapter 2

Change point inference and identification for high-dimensional mean vectors

2.1 Introduction

This paper studies the problems of change point inference and identification for mean vectors of high-dimensional data in finite samples. High-dimensional data are now ubiquitous in many scientific and engineering fields and data heterogeneity is the rule rather than the exception. A central problem of studying the data heterogeneity is to detect structural breaks in the underlying data generation process. Perhaps the two most fundamental questions for abrupt changes are: i) is there a change point in data? ii) if so, when does the change occur? In this work, we consider change point detection and identification for temporally independent data with cross-sectional dependence. Specifically, let $X_1^n = \{X_1, \dots, X_n\}$ be a sequence of independent random vectors in $\mathbb{R}^p$ generated from the mean-shift model:

$$X_i = \mu + \delta_n \mathbf{1}(i > m) + \xi_i, \qquad i = 1, \dots, n, \tag{2.1}$$

where $\mu \in \mathbb{R}^p$ is the population mean parameter, $\delta_n \in \mathbb{R}^p$ is the mean-shift signal parameter, m is the change point location, and $\xi_1, \dots, \xi_n$ are (temporally) independent and identically distributed (i.i.d.) mean-zero noise random vectors in $\mathbb{R}^p$ with common distribution function F. Let $\Sigma = \mathrm{Cov}(\xi_1)$ be the unknown noise covariance matrix that is not necessarily diagonal, and thus we allow cross-sectional (sometimes also referred as spatial) dependence among the components $X_{i1}, \dots, X_{ip}$ for each $i = 1, \dots, n$. Under the mean-shift model, if $\delta_n = 0$ or $m = n$, then $X_1, \dots, X_n$ form a sample of i.i.d. random vectors and no change point occurs. In this paper, our first goal is to test for whether or not there is a change point in the mean vectors $\mu_i = \mathbb{E}(X_i)$, i.e., to test for

$$H_0 : \delta_n = 0 \quad \text{and} \quad H_1 : \delta_n \neq 0 \text{ and there exists an } m \in \{1, \dots, n-1\}, \tag{2.2}$$

where the alternative hypothesis H_1 is parameterized by the change point signal δ_n and location m. If a change point is detected in the mean vectors (i.e., H_0 is rejected), then our second goal is to identify the change point location m.

For i.i.d. Gaussian noise $\xi_i \sim N(0, \Sigma)$, the log-ratio of the maximized likelihoods between H_1 with a change point at $s = 1, \ldots, n-1$ and H_0 without change point is given by

$$\log(\Lambda_s) = Z_n(s)^\top \Sigma^{-1} Z_n(s)/2, \tag{2.3}$$

where

$$Z_n(s) = \sqrt{\frac{s(n-s)}{n}} \left(\frac{1}{s} \sum_{i=1}^{s} X_i - \frac{1}{n-s} \sum_{i=s+1}^{n} X_i \right) \tag{2.4}$$

is a sequence of the normalized mean differences before and after s. Then H_0 is rejected if $\max_{1 \leqslant s < n} \log(\Lambda_s)$ is larger than a critical value. In literature, $\{Z_n(s)\}_{s=1}^{n-1}$ are often called the *cumulative sum* (CUSUM) statistics [38]. Note that the log-ratio statistics of the maximized likelihoods in (2.3) require the knowledge or an estimate of the unknown covariance matrix Σ. In the high-dimensional setting where p is larger (or even much larger) than n, estimation of Σ itself becomes a challenging problem. And the spectral norm consistency of Σ (or the inverse Σ^{-1}) is possible under additional structural assumptions (such as sparsity or low-rankness) on the covariance matrix [16, 17, 22, 23, 30, 11], which may be violated in practical applications. In contrast, tests based on the CUSUM statistics in (2.4) do not involve Σ and they are more robust to the misspecification on covariance structures. Therefore, this motivates us to study the problems of change point testing and estimation based on the high-dimensional CUSUM statistics.

To build a decision rule for change-point detection, we need to cautiously aggregate the (dependent) random vectors $Z_n(s), s = 1, \ldots, n-1$. [43] considers the change point detection on mean vectors under the mean-shift model (2.1) with i.i.d. $\xi_i \sim N(0, \sigma^2 I_p)$. They propose the linear and scan statistics based on the ℓ^2-norm aggregation of the CUSUM statistics and derive the change point detection boundary. [65] considers the ℓ^∞-norm aggregation of the CUSUM statistics and establishes a Gumbel limiting distribution under H_0. [65] also considers the bootstrap approximations to improve the rate of convergence. [99] considers the estimation problem of change points in the high-dimensional mean vectors in reduced dimensions by sparse projections and they derive the rate of convergence for estimating the change point location. In all aforementioned papers [43, 65, 99], strong structural assumptions (i.e., cross-sectional sparsity in the sense that the components $\{X_{ij}\}_{j=1}^p$ of X_i are independent or weakly dependent) are imposed to substantially reduce the intrinsic complexity of the problem. [37] relax the sparsity assumption and consider the estimation problem of change points in the (marginal) variances of high-dimensional time series under a multiplicative model. They propose a *sparsified binary segmentation* (SBS) performing the ℓ^1-norm aggregation on a thresholded version of the CUSUM statistics such that an additional sparsifying step with a tuning parameter is used to avoid noise accumulation in the aggregation. Since the SBS is sensitive to threshold tuning parameters, [36] proposes a double CUSUM test procedure sorting the magnitudes of the p components of CUSUM statistics and it may be viewed as a data-driven alternative for selecting the threshold in [37].

In this paper, besides some mild moment conditions, we do not make any assumption on the cross-sectional dependence structure of the underlying data distribution. We consider the multivariate CUSUM statistics (2.4) in the ℓ^∞-norm aggregated form:

$$T_n = \max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |Z_n(s)|_\infty := \max_{\underline{s} \leqslant s \leqslant n-\underline{s}} \max_{1 \leqslant j \leqslant p} |Z_{nj}(s)|, \tag{2.5}$$

where $\underline{s} \in [1, n/2]$ is a user-specified *boundary removal* parameter. Removing boundary points is necessary in detecting a change point since the distributions of $|Z_n(s)|_\infty$ that are closer to the endpoints are more difficult to approximate because of fewer data points. Then H_0 is rejected if T_n is larger than a critical value such as the $(1-\alpha)$ quantile of T_n. Under H_0, $\{Z_n(s)\}_{s=1}^{n-1}$ is a centered and covariance stationary process in $\mathbb{R}^p$ (i.e., $\mathbb{E}[Z_n(s)] = 0$ and $\mathrm{Cov}(Z_n(s)) = \Sigma$). To approximate the distribution of T_n, extreme value theory is a commonly used technique to derive the Gumbel-type limiting distributions [69, 87]. However, even in $p = 1$ case, the convergence rate of maxima of the CUSUM process $\{Z_n(s)\}_{s=1}^{n-1}$ is known to be very slow [87, 53, 65].

2.1.1 Our contributions

To overcome the fundamental difficulty in calibrating the distribution of T_n, we consider the bootstrap approximation to the *finite sample distribution* of T_n without referring a weak limit of $\{Z_n(s)\}_{s=1}^{n-1}$. In Section 2.2, we propose a Gaussian multiplier bootstrap tailored to the CUSUM test statistics in (2.4). The proposed bootstrap test is fully data-dependent and requires no tuning parameter (except for a pre-specified boundary removal parameter $\underline{s}$). This is in contrast with the thresholding-aggregation method in [37], which requires further data-dependent procedures to choose the threshold and is not easy to justify. We will show in Section 2.3.1 that the bootstrap CUSUM test is a uniformly valid inferential procedure under H_0 where p can grow sub-exponentially fast in n *and* no explicit condition on the dependence structure among the components $\{X_{ij}\}_{j=1}^p$ is needed. This is in contrast with [43, 65, 99] where the components are assumed to be either independent or weakly dependent, and with [37, 36] where the dimension can only grow polynomially fast in sample size. Moreover, we will show that, under a mild signal strength condition, our bootstrap CUSUM test is consistent in the sense that the sum of type I and type II errors is asymptotically vanishing [47, Chapter 6.2]. In addition, the requirement on the signal strength can achieve the minimax separation rate derived in [43] under the sparse alternative (i.e., the change occurs only in a few number of components $X_1, \ldots, X_n$).

If a change point is detected, then we estimate the change point location by maximizing the ℓ^∞-norm of the generalized CUSUM statistics (2.8) at two different weighting scales. The first estimator is based on the covariance stationary CUSUM statistics in (2.4). In Section 2.3.2, we show that it is consistent in estimating the location at the parametric rate $n^{-1/2}$ (up to logarithmic factors) for sub-exponential observations. The second estimator is a non-stationary CUSUM statistics assigning less weights on the boundary data points. In this case, we show that it achieves the best possible rate of convergence on the order n^{-1} (up to logarithmic factors) under some stronger side conditions. In both cases, dimension impacts the rate of convergence only through the logarithmic factors. Thus consistency of the CUSUM location estimators can be achieved when p grows sub-exponentially fast in n.

Our bootstrap change point inference can be naturally extended to handle multiple change points via the generic binary segmentation technique. Once a change point is claimed by our bootstrap test and located in the estimation step, the binary segmentation continues the same testing and estimation procedure on the segments before and after the change until no further change point can be detected by the bootstrap test (cf. Algorithm 1 in Section 2.2.3). Thus, the bootstrap CUSUM test can dynamically adjust the detection rule during the iterations. We derive the rate of convergence of this *bootstrap-assisted binary segmentation* (BABS) for recursively estimating the multiple change points under suitable signal separation and strength conditions. No additional tuning parameter is introduced in BABS.

2.1.2 Literature review

CUSUM statistics [83] are originally introduced in the sequential testing problems to distinguish between the in-control hypothesis $\delta_n = 0$ and the out-control mean-shift hypothesis for a *given* $\delta_n \neq 0$ in model (2.1), aiming to minimize the expected average run length [68, 94, 95, 83, 31, 57, 101, 95, 76, 85]. This paper uses CUSUM statistics for fixed sampled size tests, as in many other statistical change point testing and estimation works [104, 19, 38, 67, 13, 45, 24, 75, 55, 54, 44, 46, 81, 8, 9, 15, 109, 56].

There is a recent surge of literature on change point analysis for high-dimensional data. Change point detection is considered in [43, 65]. Estimation of the number and locations of change points are considered in [37, 65, 99, 36]. Bootstrap inference is considered in [36] (without rigorous statistical guarantees).

Finite sample approximations to the distribution of maxima corresponding to sums of *independent* mean-zero random vectors in high dimensions are studied in [33, 35]. We highlight that validity of our bootstrap CUSUM test for the change point does not (at least directly) follow the Gaussian and bootstrap approximation results in [33, 35]. The reason is that, in the change-point detection context, the extreme-value type test statistic T_n defined in (2.5) is the maximum of a sequence of *dependent* random vectors $Z_n(s), s = \underline{s}, \ldots, n-\underline{s}$. Therefore, the distributional approximation results developed in [33, 35] require considerable modifications tailored to the change point analysis. A main technical innovation of this work is that the CUSUM statistics are affine transformations of the independent data points in an augmented space so that we can make use of the high-dimensional Gaussian and bootstrap approximations without overpaying the price of the increased dimensionality in the embedded larger space.

2.1.3 Organization

The rest of this paper is organized as follows. The bootstrap change point test, estimation of the change point location and extension to multiple change point algorithm are described in Section 2.2. In Section 2.3, we derive the size validity, power properties of the bootstrap test and the rate of convergence for the change point location estimator by the generalized CUSUM statistics, followed by consistency of algorithm designed for multiple change point identification. In Section 2.4, we report extensive simulation results of testing and estimation for a variety of distributions with different dependence structures and moment conditions. In Section 2.5, real data examples are provided. Proofs of the main results in Section 2.3 and additional simulation results are given in Section 2.6.

2.1.4 Notation

For $q > 0$ and a generic vector $x \in \mathbb{R}^p$, 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 [70]. 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. Throughout the paper, we assume $n \geqslant 4$ and $p \geqslant 3$.

2.2 Methodology

2.2.1 Bootstrap CUSUM test

We first introduce a bootstrap procedure to approximate the distribution of T_n. Let $e_1, \dots, e_n$ be i.i.d. $N(0,1)$ random variables independent of $X_1, \dots, X_n$. Let $\bar{X}_s^- = s^{-1} \sum_{i=1}^s X_i$ and $\bar{X}_s^+ = (n-s)^{-1} \sum_{i=s+1}^n X_i$ be the left and right sample averages at s, respectively. Define

$$Z_n^*(s) = \sqrt{\frac{n-s}{ns}} \sum_{i=1}^{s} e_i (X_i - \bar{X}_s^-) - \sqrt{\frac{s}{n(n-s)}} \sum_{i=s+1}^{n} e_i (X_i - \bar{X}_s^+). \tag{2.6}$$

Then the bootstrap test statistic is defined as

$$T_n^* = \max_{\underline{s} \leqslant s \leqslant n - \underline{s}} |Z_n^*(s)|_\infty, \tag{2.7}$$

and the $(1-\alpha)$ conditional quantile of T_n^* given X_1^n

$$q_{T_n^*|X_1^n}(1-\alpha) = \inf\{t \in \mathbb{R} : \mathbb{P}(T_n^* \leqslant t | X_1^n) \geqslant 1 - \alpha\}$$

is used as a critical value of the bootstrap test to approximate the quantiles of T_n. In particular, for any $\alpha \in (0,1)$, we reject H_0 if $T_n > q_{T_n^*|X_1^n}(1-\alpha)$. Note that the Gaussian multiplier bootstrap test statistic T_n^* and its conditional quantile $q_{T_n^*|X_1^n}(1-\alpha)$ are *computable* in the sense that we can repeatedly draw Monte Carlo samples by simulating the multiplier random variables $e_1, \dots, e_n$ to approximate the distribution of T_n^*.

Remark 1 (Comments on centering in the bootstrap CUSUM statistics $Z_n^*(s)$). Alternatively, we can also consider the following version of bootstrap CUSUM statistics

$$\tilde{Z}_n^*(s) = \sqrt{\frac{n-s}{ns}} \sum_{i=1}^{s} e_i X_i - \sqrt{\frac{s}{n(n-s)}} \sum_{i=s+1}^{n} e_i X_i$$

without left and right centering by $\bar{X}_s^-$ and $\bar{X}_s^+$. It can be shown that the bootstrap CUSUM test based on $Z_n^*(s)$ and $\tilde{Z}_n^*(s)$ have the same rate of convergence in the size and power analysis (Theorem 2.1, Corollary 2.2, and Theorem 2.3). However,

$$\begin{aligned}\mathrm{Cov}(Z_n^*(s)|X_1^n) &= \frac{n-s}{ns}\sum_{i=1}^{s}(X_i-\bar{X}_s^-)(X_i-\bar{X}_s^-)^\top + \frac{s}{n(n-s)}\sum_{i=s+1}^{n}(X_i-\bar{X}_s^+)(X_i-\bar{X}_s^+)^\top \\ &\leqslant \frac{n-s}{ns}\sum_{i=1}^{s}X_iX_i^\top + \frac{s}{n(n-s)}\sum_{i=s+1}^{n}X_iX_i^\top = \mathrm{Cov}(\tilde{Z}_n^*(s)|X_1^n),\end{aligned}$$

where we write two square matrices $A \leqslant B$ if $A - B \leqslant 0$ i.e., $A - B$ is negative semi-definite. Hence $\tilde{Z}_n^*(s)$ incurs a larger (conditional) covariance matrix than $Z_n^*(s)$ and it is recommended to use $Z_n^*(s)$ rather than $\tilde{Z}_n^*(s)$. □

Remark 2 (Comparisons with [65] under H_0). In a related work, [65] considers the change point tests for high-dimensional time series based on the following version of the CUSUM statistics

$$B_{nj} = \frac{1}{\hat{\sigma}_j\sqrt{n}}\max_{1\leqslant s\leqslant n}\left|\sum_{i=1}^{s}X_{ij} - \frac{s}{n}\sum_{i=1}^{n}X_{ij}\right|, \quad j = 1,\ldots,p,$$

where $\hat{\sigma}_j^2$ is a consistent estimator for the long-run variance of $\{X_{ij}\}_{i\in\mathbb{N}}$. Then H_0 is rejected if $\tilde{T}_n = \max_{1\leqslant j\leqslant p} B_{nj}$ is larger than a critical value. Under H_0 and the spatial sparsity conditions (Assumption 2.2 in [65]), the author establishes a Gumbel limiting distribution for $\tilde{T}_n$ (after suitable normalizations). To improve the rate of convergence, the author also proposes a parametric bootstrap $\tilde{T}_n^Y = \max_{1\leqslant j\leqslant p} B_{nj}^Y$, where

$$B_{nj}^Y = \frac{1}{\sqrt{n}}\max_{1\leqslant s\leqslant n}\left|\sum_{i=1}^{s}Y_{ij} - \frac{s}{n}\sum_{i=1}^{n}Y_{ij}\right|, \quad j = 1,\ldots,p,$$

and $\{Y_{ij} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\}$ is an array of i.i.d. $N(0,1)$ random variables. Asymptotic bootstrap validity is derived under the same spatial sparsity assumption as in the Gumbel limit. There is an important difference between $\tilde{T}_n^Y$ in [65] and our bootstrap test based on T_n^*. Note that the conditional covariance matrices of $Z_n^*(s)$ given X_1^n are sample analogs of covariance matrices of $Z_n(s)$. We will show in Section 2.3.1 that T_n^* can approximate the distribution of T_n without assuming any kind of spatial (cross-sectional) sparsity conditions. On the contrary, since $\{Y_{ij}\}$ are i.i.d., even when $X_1,\ldots,X_n$ are independent observations, the parametric bootstrap B_{nj}^Y does not mimic the general dependence structure among the components $\{X_{ij}\}_{j=1}^p$. In addition, the bootstrap validity of T_n^* we establish in Theorem 2.1 and Corollary 2.2 below is non-asymptotic and it holds without assuming a Gumbel-type limiting distribution for T_n. □

2.2.2 Estimating the change point location under the alternative hypothesis

If a change point is detected in the mean vectors (i.e., H_0 is rejected), then our next goal is to identify the change point location m. Specifically, we estimate $t_m = m/n, m = 1, \ldots, n$, where the data $X_1, \ldots, X_n$ are observed at evenly spaced time points and their index variables are normalized to $[0, 1]$. We consider the change point location estimator based on the generalized CUSUM statistics [55]

$$Z_{\theta,n}(s) = \left[\frac{s(n-s)}{n}\right]^{1-\theta} \left(\frac{1}{s}\sum_{i=1}^{s} X_i - \frac{1}{n-s}\sum_{i=s+1}^{n} X_i\right), \tag{2.8}$$

where θ is a weighting parameter satisfying $0 \leqslant \theta < 1$. Obviously, the CUSUM statistics $Z_n(s)$ in (2.4) is a special case of $\theta = 1/2$, i.e., $Z_n(s) = Z_{1/2,n}(s)$. Then we estimate m by

$$\hat{m}_\theta = \text{argmax}_{1 \leqslant s < n} |Z_{\theta,n}(s)|_\infty. \tag{2.9}$$

and we use $t_{\hat{m}_\theta} = \hat{m}_\theta/n$ to estimate t_m. It is easily seen that, for smaller values of θ, $Z_{\theta,n}(s)$ assigns less weights on the boundary data points. Therefore, if the true change point location is bounded away from the two endpoints, we expect that $t_{\hat{m}_\theta}$ with a smaller weighting parameter can achieve better rate of convergence. For example, if $t_m \in (0, 1)$ is fixed and $p = 1$, then it is known that the $\{Z_{0,n}(s)\}_{s=1}^{n-1}$ converges weakly to a functional of the Weiner process and the corresponding maximizer $\hat{m}_0$ achieves the rate of convergence of the order n^{-1}, which is clearly the best possible rate and is faster than the parametric rate $n^{-1/2}$ [9, 55]. Instead of considering the whole family of the generalized CUSUM statistics indexed by $\theta \in [0, 1)$, we consider two important cases of $\theta = 1/2$ (covariance stationary) and $\theta = 0$ (non-stationary) in this paper. For $\theta = 1/2$, $Z_{1/2,n}(s)$ is related to the proposed bootstrap CUSUM statistics $Z_n^*(s)$ in (2.6) and the log-ratio statistics in (2.3) under normality with $\Sigma = \sigma^2 \text{Id}_p$. For $\theta = 0$, $Z_{0,n}(s)$ is related to the parametric bootstrap in [65].

Remark 3 (Comments on the boundary). It should be noted that in the bootstrap CUSUM test, we must remove the boundary points from approximating the distribution of T_n. If the boundary points are included in the maxima T_n and T_n^*, then the conditional distribution of T_n^* (given X_1^n) does not provide an accurate approximation to the distribution of T_n. Theorem 2.1 and Theorem 2.3 provide the precise rate of convergence that characterizes the boundary removal parameter $\underline{s}$ to ensure the consistency (in terms of the sum of type I and type II errors) of the bootstrap CUSUM test. On the other hand, the estimation problem in (2.9) does not exclude the endpoints outside the interval $[\underline{s}, n - \underline{s}]$. However, in practice, if the existence of a change point is not known as a priori and it is decided by a test, then the boundary restriction is implicitly imposed for both testing and estimation in empirical applications [9]. □

2.2.3 Bootstrap-assisted binary segmentation for multiple change points

Suppose now there are ν change points $m_0 = 1 < m_1 < \cdots < m_\nu < m_{\nu+1} = n$ and consider the following multiple mean-shifts model:

$$X_i = \mu + \sum_{k=1}^{\nu} \delta_n^{(k)} 1(i > m_k) + \xi_i, \quad i = 1, \ldots, n, \tag{2.10}$$

where $\delta_n^{(k)} \in \mathbb{R}^p$ are non-zero mean-shift vectors and ξ_i are again i.i.d. mean-zero random vectors in $\mathbb{R}^p$. Without less of generality, we may assume $\mu = 0$ and $\delta_n^{(0)} = \delta_n^{(\nu+1)} = 0$. Given a beginning time point b and an ending time point e, we can compute the CUSUM statistics on the initial data segment $\{X_i\}_{i=b}^e$:

$$Z_{n,b,e}(s) = \sqrt{\frac{(s-b+1)(e-s)}{e-b+1}} \left(\frac{1}{s-b+1} \sum_{i=b}^{s} X_i - \frac{1}{e-s} \sum_{i=s+1}^{e} X_i \right).$$

Note that the normalization in $Z_{n,b,e}(s)$ corresponds to the case $\theta = 1/2$ in (2.8). It can be shown that the maximizer of $|\mathbb{E}Z_{n,b,e}(s)|_\infty, s = b, \ldots, e$ always occurs at one of the change points $\{m_k, k = 1, \ldots, \nu\} \cap [b, e]$ (cf. Lemma 2.13). Therefore, under multiple change points model (2.10), we can use $Z_{n,b,e}(s)$ to locate one shift in the interval $[b, e]$. If our bootstrap CUSUM test (calculated based on $\{X_i\}_{i=b}^e$) rejects H_0 at the significance level α, then $\hat{m}_b^e = \mathrm{argmax}_{s=b,\ldots,e} |Z_{n,b,e}(s)|_\infty$ is marked as a change point. In addition, observe that the mean vectors $\mu_i = \mathbb{E}[X_i]$ are piecewise constant such that

$$\mu_{m_k+1} = \cdots = \mu_{m_{k+1}} = \sum_{l=0}^{k} \delta_n^{(l)}.$$

Thus, we may recursively apply the binary segmentation to search along the two directions $[b, \hat{m}_b^e]$ and $[\hat{m}_b^e + 1, e]$ until no further change point would be detected by the subsequent bootstrap tests. The pseudo-code for our bootstrap-assisted binary segmentation algorithm for multiple change points detection, referred as BABS(α, b, e), is summarized in the following Algorithms 1.

2.3 Theoretical results

Denote $\mathbb{P}_0(\cdot)$ and $\mathbb{P}_1(\cdot)$ as the probability computed under H_0 and H_1, respectively.

Algorithm 1 BABS(α, b, e)

```
if e − b + 1 < 2s̲ then
    STOP
else
    m̂_b^e = argmax_{s=b,...,e} |Z_{n,b,e}(s)|_∞
    if our bootstrap CUSUM test concludes existence of a change in [b, e] then
        add m̂_b^e to the set of estimated change-points;
        BABS(α, b, m̂_b^e);
        BABS(α, m̂_b^e + 1, e).
    else
        STOP
    end if
end if
return estimated change points.
```

2.3.1 Size and power of the bootstrap CUSUM test: one change point

Our first main result (cf. Theorem 2.1) is to establish finite sample bounds for the (random) Kolmogorov distance between T_n and T_n^*:

$$\rho^*(T_n, T_n^*) = \sup_{t \in \mathbb{R}} |\mathbb{P}_0(T_n \leqslant t) - \mathbb{P}_0(T_n^* \leqslant t | X_1^n)|.$$

From this, we can derive the asymptotic bootstrap validity for certain high-dimensional scaling limit for (n, p). In particular, with $\rho^*(T_n, T_n^*) = o_{\mathbb{P}}(1)$, we can show that Type-I error of the bootstrap test is asymptotically controlled at the exact nominal level $\alpha \in (0, 1)$; i.e., $\mathbb{P}_0(T_n > q_{T_n^*|X_1^n}(1-\alpha)) \to \alpha$ (cf. Corollary 2.2).

Let $\underline{b}, \bar{b}, q > 0$. We make the following assumptions.

(A) $\mathrm{Var}(\xi_{ij}) \geqslant \underline{b}$ for all $j = 1, \ldots, p$.

(B) $\mathbb{E}[|\xi_{ij}|^{2+\ell}] \leqslant \bar{b}^\ell$ for $\ell = 1, 2$ and for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

(C) $\|\xi_{ij}\|_{\psi_1} \leqslant \bar{b}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

(D) $\mathbb{E}[\max_{1 \leqslant j \leqslant p}(|\xi_{ij}|/\bar{b})^q] \leqslant 1$ for all $i = 1, \ldots, n$.

Condition (A) is a non-degeneracy assumption. Condition (B) is a mild moment growth condition. Without loss of generality, we may take $\bar{b} \geqslant 1$. Conditions (C) and (D) impose sub-exponential and uniform polynomial moment requirements on the observations, respectively. Define

$$\varpi_{1,n} = \left(\frac{\log^7(np)}{\underline{s}} \right)^{1/6} \quad \text{and} \quad \varpi_{2,n} = \left(\frac{n^{2/q} \log^3(np)}{\gamma^{2/q} \underline{s}} \right)^{1/3}.$$

Theorem 2.1 (**Main result I:** bounds on the Kolmogorov distance between T_n and T_n^* under H_0)**.** *Suppose H_0 is true and assume (A) and (B) hold. Let $\gamma \in (0, e^{-1})$ and suppose that $\log(\gamma^{-1}) \leqslant K \log(pn)$ for some constant $K > 0$.*

(i) If (C) holds, then there exists a constant $C > 0$ only depending on $\underline{b}, \bar{b}, K$ such that

$$\rho^*(T_n, T_n^*) \leqslant C\varpi_{1,n} \tag{2.11}$$

holds with probability at least $1 - \gamma$.

(ii) If (D) holds, then there exists a constant $C > 0$ only depending on $\underline{b}, \bar{b}, K, q$ such that

$$\rho^*(T_n, T_n^*) \leqslant C\{\varpi_{1,n} + \varpi_{2,n}\} \tag{2.12}$$

holds with probability at least $1 - \gamma$.

Based on Theorem 2.1, we have the uniform size validity of the bootstrap CUSUM test.

Corollary 2.2 (Uniform size validity of Gaussian multiplier bootstrap for the CUSUM test)**.** *Suppose H_0 is true and assume (A) and (B) hold. Let $\gamma \in (0, e^{-1})$ and suppose that $\log(\gamma^{-1}) \leqslant K \log(pn)$ for some constant $K > 0$.*

(i) If (C) holds, then there exists a constant $C > 0$ only depending on $\underline{b}, \bar{b}, K$ such that

$$\sup_{\alpha \in (0,1)} |\mathbb{P}_0(T_n \leqslant q_{T_n^*|X_1^n}(\alpha)) - \alpha| \leqslant C\varpi_{1,n} + \gamma. \tag{2.13}$$

Consequently, if $\log^7(np) = o(\underline{s})$, then $\mathbb{P}_0(T_n \leqslant q_{T_n^*|X_1^n}(\alpha)) \to \alpha$ uniformly in $\alpha \in (0, 1)$ as $n \to \infty$.

(ii) If (D) holds, then there exists a constant $C > 0$ only depending on $\underline{b}, \bar{b}, K, q$ such that

$$\sup_{\alpha \in (0,1)} |\mathbb{P}_0(T_n \leqslant q_{T_n^*|X_1^n}(\alpha)) - \alpha| \leqslant C\{\varpi_{1,n} + \varpi_{2,n}\} + \gamma. \tag{2.14}$$

Consequently, if $\max\{\log^7(np), n^{2/q} \log^{3+\varepsilon}(np)\} = o(\underline{s})$ for some $\varepsilon > 0$, then $\mathbb{P}_0(T_n \leqslant q_{T_n^*|X_1^n}(\alpha)) \to \alpha$ uniformly in $\alpha \in (0, 1)$ as $n \to \infty$.

Our second main result is to analyze the power of the bootstrap CUSUM test. We are mainly interested in characterizing the change point signal strength (quantified by the ℓ^∞ norm of δ_n) and the location t_m such that H_0 and H_1 can be asymptotically separated by our bootstrap CUSUM test. Without loss of generality, we may assume that $|\delta_n|_\infty \leqslant 1$.

Theorem 2.3 (**Main result II:** power of Gaussian multiplier bootstrap for CUSUM test under H_1). *Suppose H_1 is true with a change point $m \in [\underline{s}, n - \underline{s}]$ and assume (A) and (B) hold. Let $\zeta \in (0, 1/2)$ and $\gamma \in (0, e^{-1})$ such that $\log(\gamma^{-1}) \leqslant K \log(pn)$ for some constant $K > 0$.*

(i) If (C) holds and

$$|\delta_n|_\infty \geqslant C_1 \sqrt{\frac{\log(\zeta^{-1})\log(np) + \log(np/\alpha)}{nt_m(1-t_m)}} \tag{2.15}$$

for some large enough constant $C_1 := C_1(\bar{b}, \underline{b}, K) > 0$, then there exists a constant $C_2 := C_2(\bar{b}, \underline{b}, K) > 0$ such that

$$\mathbb{P}_1(T_n > q_{T_n^*|X_1^n}(1-\alpha)) \geqslant 1 - \gamma - C_2\varpi_{1,n} - 2\zeta. \tag{2.16}$$

(ii) If (D) holds and $|\delta_n|_\infty$ obeys (2.15) for some large enough constant $C_1 := C_1(\bar{b}, \underline{b}, K, q) > 0$, then there exists a constant $C_2 := C_2(\bar{b}, \underline{b}, K, q) > 0$ such that

$$\mathbb{P}_1(T_n > q_{T_n^*|X_1^n}(1-\alpha)) \geqslant 1 - \gamma - C_2\{\varpi_{1,n} + \varpi_{2,n}\} - 2\zeta. \tag{2.17}$$

Remark 4 (Rate-optimality on the change point detection for sparse alternatives). Under the i.i.d. Gaussian errors $\xi_i \sim N(0, \mathrm{Id}_p)$ in the mean-shift model (2.1), the detection boundary for a change point in a Gaussian sequence is characterized in [43]. Let $a > 0$ and suppose that a change point

$$\delta_n = (\underbrace{a, \dots, a}_{k \text{ times}}, 0, \dots, 0)^\top$$

occurs in the first k components at the location m in the sequence $X_1, \dots, X_n$. Following [43], we consider the scaling limit $p = n^{c_1}$ and $k = p^{1-c_2}$ for some $c_1 > 0$ and $c_2 \in [0, 1)$. If $c_2 \in (1/2, 1)$, then the number of components with a change point is highly sparse. In this case, the minimax separation condition for H_0 and H_1 is given by

$$a = r_p \sqrt{\frac{\log(p)}{nt_m(1-t_m)}}.$$

Specifically, detection is impossible if $\limsup_{p\to\infty} r_p < \sqrt{2c_2 - 1}$ and detection is possible if $\liminf_{p\to\infty} r_p > \sqrt{2c_2/(1-\log 2)}$. On the other hand, choosing $\alpha_n = n^{-c}$ for some constant $c > 0$ in Corollary 2.2 and Theorem 2.3, we see that if

$$a \geqslant C_* \sqrt{\frac{\log(\zeta^{-1})\log(p)}{nt_m(1-t_m)}}$$

for some large constant $C_* > 0$, then our bootstrap CUSUM change point test achieves the minimax separation rate in the high sparsity regime (with stronger side conditions to ensure the bootstrap validity).

Hence, the signal strength requirement for detection in the proposed bootstrap test achieves the minimax optimal rate under the sparse alternatives. On the other hand, it should be noted that, under the dense alternatives $c_2 \in [0, 1/2]$, our bootstrap CUSUM test remains consistent in detecting the change point signal in the sense that the sum of type I and type II errors converges to zero. However, in such case, the bootstrap CUSUM test does not reach the detection boundary and the minimax separation rate [43]. □

Remark 5 (Monotonicity of power in the signal strength). Inspecting the proof of Theorem 2.3, it is seen that the type II error of the bootstrap CUSUM test is bounded by a probability depending on the change point signal strength $|\delta_n|_\infty$ and location m (cf. equation (2.44)). Specifically,

$$\text{Type II error} \leqslant \mathbb{P}_1(\tilde{T}_n \geqslant \tilde{\Delta} - q_{T_n^*|X_1^n}(1-\alpha)),$$

where $\tilde{\Delta} = \sqrt{nt_m(1-t_m)}|\delta_n|_\infty$, $\tilde{T}_n = \max_{\underline{s} \leqslant n \leqslant n-\underline{s}} |Z_n^\xi(s)|$, and $Z_n^\xi(s)$ are the CUSUM statistics computed on the (unknown) ξ_1^n random vectors. Since the distribution of $\tilde{T}_n$ does not depend on δ_n and the conditional quantile $q_{T_n^*|X_1^n}(1-\alpha)$ is bounded by $O(\sqrt{\log(np)})$ with a large probability under H_1, the power of the bootstrap CUSUM test is lower bounded by a quantity that is non-decreasing in $|\delta_n|_\infty$. Simulation examples in Section 2.4 confirm our theoretical observation. In addition, since $t_m(1-t_m)$ is maximized at $t_m = 1/2$, a change point near the middle is easier to detect than it is near the boundary. □

Remark 6 (Choice of boundary removal parameter). There is a trade-off for the choice of boundary removal parameter: the larger $\underline{s}$, the smaller of the error bounds and the more data points are removed from the change point detection (so that the regime allowed by the bootstrap CUSUM test is smaller). In theory, the lower bound of $\underline{s}$ is given in Corollary 2.2 for size validity of the bootstrap CUSUM test. Specifically, if the data distribution has sub-exponential tail (i.e., Condition (C) holds), then we need $\underline{s} \gg \log^7(np)$ for the error-in-size $\varpi_{1,n} = o(1)$; if the data distribution has polynomial tail (i.e., Condition (D) holds) with $q > 0$, then we need $\underline{s} \gg \max\{\log^7(np), n^{2/q}\log^{3+\varepsilon}(np)\}$ for $\varpi_{1,n} + \varpi_{2,n} = o(1)$. This implies that we can choose $\underline{s} = c_1 n^{c_2}$ for some small constants $c_1 > 0, 1 \geqslant c_2 > 0$ in either sub-exponential or polynomial case.

As a leading example, we consider a *fixed* normalized true change point location $t_m = m/n \in (0, 1)$. Then we may choose $\underline{s} = c_0 n$ for some small constant $c_0 > 0$ in order to include the change point in the interval $[\underline{s}, n - \underline{s}]$. Under this framework, asymptotic size validity of the bootstrap CUSUM test is obtained if $p = O(e^{n^c})$ for some $1/7 > c > 0$ under the sub-exponential moment condition on the observations.

□

Remark 7 (Comparisons with [65] under H_1). [65] proposed bootstrap testing procedures for change point under the alternative, which are different from the parametric bootstrap under H_0 (cf. Remark 2). Specifically, the author considers several block versions of the multiplier and empirical bootstraps under H_1 in the time series setting. All of the tests under H_1 in [65] require a minimum signal strength condition (cf. Assumption 4.3 therein):

$$\limsup_{n\to\infty} \frac{\log n}{K_b \min_{j\in\mathcal{S}} \delta_{nj}^2} = 0, \tag{2.18}$$

where $\mathcal{S} = \{j \in \{1,\dots,p\} : \delta_{nj} \neq 0\}$ and K_b is the size of blocks. There are two major differences from the bootstraps in [65]. First, our Gaussian multiplier bootstrap CUSUM test is asymptotically valid and powerful for a change point under both H_0 and H_1. On the contrary, [65, Section 4] designed a series estimators of t_m to adapt to estimation of $\hat{\sigma}_j^2, j = 1,\dots,p$ under H_1, and the estimators must rely on the assumption that each dimension has at most one change point. However, our approach without estimation of t_m can test against more generalized multiple change-point problem where power can still be guaranteed (cf. Lemma 2.6). Second, detection by our bootstrap CUSUM test relies on a lower bound on the signal strength quantified by $|\delta_n|_\infty$, which is much weaker than (2.18). For example, it is possible that the minimum signal strength $\min_{j\in\mathcal{S}} \delta_{nj}^2$ decays to zero faster than $(\log n)/K_b$, while our bootstrap CUSUM test remains valid since it only requires $|\delta_n|_\infty$ satisfies a mild lower bound in (2.15). □

2.3.2 Rate of convergence of the change point location estimator

Our third main result is concerned with the rate of convergence of the change point location estimator $t_{\hat{m}_\theta}$, where $\hat{m}_\theta$ is defined through (2.9) and (2.8). We first consider the case of $\theta = 1/2$ corresponding to the covariance stationary CUSUM statistics.

Theorem 2.4 (**Main result III:** rate of convergence for change point location estimator: $\theta = 1/2$)**.** *Suppose that (B) holds and H_1 is true. Suppose that $\log(\gamma^{-1}) \leqslant K \log(np)$ for some constant $K > 0$.*
(i) If (C) holds, then there exists a constant $C > 0$ depending only on $\bar{b}, K$ such that

$$\mathbb{P}_1\Big(|t_{\hat{m}_{1/2}} - t_m| \leqslant \frac{C\log^2(np)}{\sqrt{nt_m(1-t_m)}|\delta_n|_\infty}\Big) \geqslant 1-\gamma. \tag{2.19}$$

(ii) If (D) holds with $q > 2$, then there exists a constant $C > 0$ depending only on $\bar{b}, K, q$ such that

$$\mathbb{P}_1\Big(|t_{\hat{m}_{1/2}} - t_m| \leqslant \frac{Cn^{1/q}(\log(np) + \gamma^{-1/q})}{\sqrt{nt_m(1-t_m)}|\delta_n|_\infty}\Big) \geqslant 1-\gamma. \tag{2.20}$$

Note that the non-degeneracy Condition (A) is not needed in estimating the change point location. Consider a fixed $t_m \in (0,1)$ as in our leading example. It is seen from Theorem 2.4 that $t_{\hat{m}_{1/2}}$ is consistent for estimating t_m if the signal strength satisfying: i) $|\delta|_\infty \gg n^{-1/2}\log^2(np)$ in the sub-exponential moment case; ii) $|\delta|_\infty \gg n^{-1/2+1/q}\log(np)$ in the polynomial moment case. From Part (i) of Theorem 2.4, it should also be noted that the change point location estimator $t_{\hat{m}_{1/2}}$ does not attain the optimal rate of convergence. In particular, consider the setup where $t_m \in (0,1)$, $p=1$, and $|\delta_n| = c$ is a constant signal. Then the rate of convergence in (2.19) reads $O(\log^2(n)/\sqrt{n})$; that is, up to a logarithmic factor, the change point estimator has the parametric rate of convergence of the order $n^{-1/2}$. In such setup, however, it is known that the best possible rate of convergence for estimating the change point location is of the order n^{-1} [55], which is achieved by maximizing $|Z_{0,n}(s)|$ (i.e., the non-stationary CUSUM statistics). Therefore, it is interesting to study the impact of dimensionality on the rate in the case of $\theta = 0$ when the true change point $t_m \in (0,1)$ is fixed. This is the content of the following theorem. Denote $\underline{\delta}_n = \min_{j \in \mathcal{S}} |\delta_{nj}|$.

Theorem 2.5 (**Main result IV:** rate of convergence for change point location estimator: $\theta = 0$)**.** *Suppose that (B) holds and H_1 is true with a change point m satisfying $c_1 \leqslant t_m \leqslant c_2$ for some constants $c_1, c_2 \in (0,1)$. Suppose that $\log^3(np) \leqslant Kn$ and $\log(\gamma^{-1}) \leqslant K\log(np)$ for some constant $K > 0$.*

(i) If (C) holds, then there exists a constant $C := C(\bar{b}, K, c_1, c_2) > 0$ such that

$$\mathbb{P}_1\left(|t_{\hat{m}_0} - t_m| \leqslant \frac{C\log^4(np)}{n\underline{\delta}_n^2}\right) \geqslant 1-\gamma. \tag{2.21}$$

(ii) If (D) holds for some $q \geqslant 2$, then there exists a constant $C := C(\bar{b}, K, q, c_1, c_2) > 0$ such that

$$\mathbb{P}_1\left(|t_{\hat{m}_0} - t_m| \leqslant \frac{C\log(np)}{n\underline{\delta}_n^2} \cdot \max\left\{1, \frac{n^{2/q}\log(np)}{\gamma^{2/q}}\right\}\right) \geqslant 1-\gamma. \tag{2.22}$$

Based on Theorem 2.5, we see that the dimension impacts the optimal rate of convergence for estimating the change point location only on the logarithmic scale. Compared with Theorem 2.4, we see that faster convergence of $t_{\hat{m}_0}$ than that of $t_{\hat{m}_{1/2}}$ is possible when $t_m \in (0,1)$ is fixed and the dimension grows sub-exponentially fast in the sample size. On the other hand, $t_{\hat{m}_{1/2}}$ is more robust to estimate the change point when its location is near the boundary, i.e., $t_m \to 0$ and $t_m \to 1$ are allowed to maintain the consistency in Theorem 2.4; see our simulation result in Section 2.4 for numeric comparisons.

Remark 8 (Comparison with the sparse projection CUSUM method in [99]). [99] considered a different sparse projection estimator, denoted as $\tilde{m}$, for change point location (even though their estimator is based on the $\theta = 1/2$ normalization). Then, [99, Theorem 1] for single change point in our notation reads: if $X_i \sim N(\mu_i, \sigma^2 I_p)$ are independent and the change point signal δ_n satisfies $\|\delta_n\|_0 \leq k$ and $\|\delta_n\|_2 \geq \vartheta$ such that

$$\frac{\sigma}{\vartheta\tau}\sqrt{\frac{k\log(p\log n)}{n}} \lesssim 1, \tag{2.23}$$

where $\tau = \min\{t_m, 1 - t_m\}$, then with probability tending to one, $\tilde{m}$ obeys

$$n^{-1}|\tilde{m} - m| \lesssim \frac{\sigma^2 \log\log n}{n\vartheta^2}. \tag{2.24}$$

Here, ϑ/σ can be thought as a signal-to-noise ratio.

Let us compare $\tilde{m}$ in [99] with our estimators $\hat{m}_{1/2}$ and $\hat{m}_0$ in the highly sparse regime where $k = 1$, $\vartheta = n^{-c_1}$, and $m = n^{c_2}$ for some $c_1, c_2 \geq 0$. For simplicity, let $\sigma = 1$. For such configuration, $t_{\tilde{m}} = \tilde{m}/n$ attains the nearly minimax-optimal rate of convergence $n^{-1+2c_1} \log\log n$ if (2.23) holds, i.e., we need

$$\sqrt{\log(p \log n)} \lesssim n^{c_2 - c_1 - 1/2},$$

which necessarily requires that $c_2 > c_1 + 1/2$ as $n \to \infty$. It means that the change point location cannot be too close to the boundary $m \gg n^{c_1+1/2}$ in order to obtain the optimal rate for [99]. Thus, if the change point is not close to the boundary, then the sparse projection estimator is nearly optimal at the rate $O(n^{-1+2c_1})$ (where c_1 can be arbitrarily small to zero), and for the boundary scenario, their estimator loses such optimality.

In our Theorem 2.5, it is shown that $t_{\hat{m}_0}$ achieves the nearly optimal rate (up to logarithmic factors) if $m = Cn$ for some constant $C \in (0, 1)$. On the other hand, we showed in our Theorem 3.4 that our estimator $t_{\hat{m}_{1/2}}$ can deal with the *"more boundary"* case as long as

$$\log^2(np) \ll n^{c_2/2 - c_1}.$$

In addition, there are other side differences between our assumptions and the ones in [99], where the latter are more stringent on the data generation mechanism. □

2.3.3 Rate of convergence of bootstrap-assisted binary segmentation

Under the multiple mean-shifts model (2.10), we consider the testing problem for H_0 against the alternative hypothesis with multiple change points

$$H_1^{'} : \delta_n^{(k)} \neq 0 \text{ for some } 1 = m_0 < m_1 < \cdots < m_\nu < m_{\nu+1} = n \text{ and } \nu \geqslant 1. \tag{2.25}$$

We have the following Lemma 2.6 to control the power of our bootstrap CUSUM test based on T_n^* in (2.7) in presence of multiple change points. This is the initial step of the bootstrap-assisted binary segmentation

(Algorithm 1), and the power control is crucial for deriving the overall rate of convergence of recursively estimating the multiple change points in Algorithm 1. Denote $\bar{\delta}_n = \min_{k=1,\ldots,\nu} |\delta_n^{(k)}|_\infty$.

Lemma 2.6 (Power of Gaussian multiplier bootstrap for CUSUM test under H_1')**.** *Suppose $H_1^{'}$ is true under the multiple mean-shift model (2.10) with $\{m_k\}_{k=1}^{\nu} \subset [\underline{s}, n-\underline{s}]$. Assume (A), (B) and $\min_{k=0,\ldots,\nu} |m_{k+1} - m_k| \geqslant D_\nu$ for some $D_\nu > \underline{s}$. Let $\zeta \in (0, 1/2)$ and $\gamma \in (0, e^{-1})$ such that $\log(\gamma^{-1}) \leqslant K \log(np)$ for some constant $K > 0$.*

(i) If (C) holds and

$$\max_{k=1,\ldots,\nu} \frac{D_\nu^2 \bar{\delta}_n}{\sqrt{n^3 t_{m_k}(1-t_{m_k})}} \geqslant C_1 \left[\sqrt{\log(\zeta^{-1})\log(np) + \nu^2 \log(np/\alpha)}\right] \tag{2.26}$$

for some large enough constant $C_1 := C_1(\bar{b}, \underline{b}, K) > 0$, then there exists a constant $C_2 := C_2(\bar{b}, \underline{b}, K) > 0$ such that (2.16) holds accordingly.

(ii) If (D) holds and $|\delta_n|_\infty$ obeys (2.26) for some large enough constant $C_1 := C_1(\bar{b}, \underline{b}, K, q) > 0$, then there exists a constant $C_2 := C_2(\bar{b}, \underline{b}, K, q) > 0$ such that (2.17) holds accordingly.

Remark 9 (Comments on signal strength under multiple change points alternative)*.* The signal strength on the LHS of (2.26) depends on the smallest mean shift $\bar{\delta}_n$ in ℓ^∞-norm and change point locations that are closest to boundary, which is the most difficult situation for CUSUM statistic to detect mean change. If $\nu = 1$, then $D_\nu = \min\{m, n-m\}$ and

$$\frac{m(n-m)}{n} \leqslant D_\nu \leqslant \frac{2m(n-m)}{n}.$$

Thus the LHS of (2.26) has the same order as $n^{1/2}(t_m(1-t_m))^{3/2}|\delta_n|_\infty$, which is stronger than the requirement of lower bound $n^{1/2}(t_m(1-t_m))^{1/2}|\delta_n|_\infty$ in Theorem 2.3. This extra cost comes from handling the possible mean shift cancellation in analyzing the general case of multiple change points. If the single change point is bounded from boundaries (i.e., t_m can be treated as a constant), then Lemma 2.6 gives the same lower bound (2.15) as in Theorem 2.3. □

Now we turn to the bootstrap-assisted binary segmentation algorithm BABS(α, b, e). We make the following assumptions in addition to (A)-(D).

a) $\min_{k=0,\ldots,\nu} |m_{k+1} - m_k| \geqslant D_\nu$, where $D_\nu \geqslant n^\Theta$ for some $\Theta \leqslant 1$.

b) $\min_{k=1,\ldots,\nu} \min_{j \in \mathcal{D}_k} |\delta_{nj}^{(k)}| \geqslant \underline{\delta}_n$, where $\mathcal{D}_k = \{1 \leqslant j \leqslant p : \delta_{nj}^{(k)} \neq 0\}$ and $\underline{\delta}_n \geqslant n^{-\omega}$ for some $\omega \geqslant 0$.

c) $\Theta - \frac{\omega}{2} > \frac{3}{4}$.

d) $n^{\frac{3}{2}\Theta-1-\omega} > C(\alpha)\nu\log^2(np)$.

e) $\epsilon_n < \underline{s} < D_\nu$, where ϵ_n is defined in (2.27).

Assumptions a)-c) are standard signal separation and strength requirements in estimating the multiple change point locations via binary segmentation, see e.g.,Theorem 1 in [45]. Assumption d) ensures that the bootstrap CUSUM test is able to consistently detect the mean-shift signals (cf. Lemma 2.6). Assumption e) is a minimal condition on the boundary removal parameter $\underline{s}$, which is smaller than the separation distance between any consecutive change points and larger than the rate of convergence ϵ_n for consistently estimating all change point locations. Note that the signal strength requirement in estimation depends on $\min_{j\in\mathcal{D}_k}|\delta_{nj}^{(k)}|$ in assumption b), which is typically stronger than $\max_{1\leqslant j\leqslant p}|\delta_{nj}^{(k)}|$ used in the testing problem.

Theorem 2.7 (**Main result V:** rate of convergence of bootstrap-assisted binary segmentation)**.** *Let $\hat{\nu}$ denote the number of change points and $\hat{m}_1 < \cdots < \hat{m}_{\hat{\nu}}$ the change point locations estimated from BABS($\alpha, 1, n$). Assume (A), (B) and a)-e) hold. Let $\gamma \in (0, e^{-1})$ such that $\log(\gamma^{-1}) \leqslant K\log(D_\nu p) \leqslant K\log(np)$ and $\zeta, \varpi_{1,n}$ are defined as in Theorem 2.3. Define*

$$\epsilon_n = \begin{cases} \frac{n^2\log^4(np)}{D_\nu^2\underline{\delta}_n^2}, & \textit{if (C) holds} \\ \frac{n^{2+6/q}(\log^2(np)+\gamma^{-2/q})}{D_\nu^2\underline{\delta}_n^2}, & \textit{if (D) holds} \end{cases}. \tag{2.27}$$

(i) If (C) holds, then there exist a constant $C_0 = C_0(\bar{b}, \underline{b}, K), C_0' = C_0'(\alpha, \bar{b}, K)$ such that

$$\mathbb{P}(\mathcal{S}_n) \geqslant 1 - 2\gamma - \nu(\gamma + 2\zeta + C_0\varpi_{1,n}),$$

where

$$\mathcal{S}_n = \left\{\hat{\nu} = \nu \textit{ and } \max_{k=1,\dots,\nu}|\hat{m}_k - m_k| \leqslant C_0'\epsilon_n\right\}.$$

(ii) If (D) holds, then there exist a constant $C_0 = C_0(\bar{b}, \underline{b}, K, q), C_0' = C_0'(\alpha, \bar{b}, K, q)$ such that

$$\mathbb{P}(\mathcal{S}_n) \geqslant 1 - 2\gamma - \nu\{\gamma + 2\zeta + C_0(\varpi_{1,n} + \varpi_{2,n})\}.$$

Remark 10 (Comparison with other binary segmentation type methods). In [37, 36], binary segmentation was also used to extend their single change-point algorithms to the multiple change-point alternative. [37, Theorem 1] discussed the asymptotic consistency of $\mathbb{P}(\mathcal{S}_n) \to 1$ for SBS algorithm under two cases. With proper tuning on their threshold parameter π_n, they claimed $\epsilon_n = O(\log^{2+c_1}(n))$ for any $c_1 > 0$ when $\Theta = 1$ for $D_\nu = O(n^\Theta)$ and $\epsilon_n = O(n^{2-2\Theta})$ when $\Theta \in (3/4, 1)$, while the rate of BABS for both cases is

$\epsilon_n = O(n^{2-2\Theta}\log^4(np))$, which is nearly the same as [37] up to a logarithmic factor. In [36], the author proposed a binary segmentation via double CUSUM algorithm (DCBS). The DCBS statistic has two CUSUM transforms in both time points with $\theta = 1/2$ in (2.8) and ordered spatial variables (i.e., cross-sectional features) in a different weighting φ. Consider the special case where $\theta = 1/2$ and $\delta_{n,1}^{(l)} = \cdots = \delta_{n,k}^{(l)} = \underline{\delta}_n \neq 0, l = 1, \cdots, \nu$. Then [36, Theorem 3.3] shows that $\mathbb{P}(\mathcal{S}_n) \to 1$ for DCBS where $\epsilon_n = \underline{\delta}_n^{-2} pk^{-2} n^{5-4\Theta}\log^2(n)$ when $\Theta \in (6/7, 1]$, while our BABS needs $\epsilon_n = \underline{\delta}_n^{-2} n^{2-2\Theta}\log^4(np)$. Under sparse alternative where $k < p^{1/2} n^{3/2-\Theta}\log^{-2}(np)$, BABS always has a better rate. However, we allow p to be as sub-exponentially large in n, while [37, Assumption 4] required $pn^{-\log n} \to 0$ and [36, Condition (A2)] fixed p as a polynomial order of n. □

Remark 11 (Extension to change point estimation in covariance matrices). Our BABS algorithm can be tailored to other high-dimensional change point estimation problems beyond the mean vectors. For instance, one can consider the estimation problem of multiple change points in covariance matrices for centered and *component-wise* sub-Gaussian data $X_i \in \mathbb{R}^p$ such that $\|X_{ij}\|_{\psi_2} \leqslant B$. Since $\|X_{ij}\|_{\psi_2} = \|X_{ij}^2\|_{\psi_1}$, the triangle inequality of the ψ_1-norm implies that

$$\|X_{ij}X_{ik}\|_{\psi_1} \leqslant \|X_{ij}^2/2\|_{\psi_1} + \|X_{ik}^2/2\|_{\psi_1} = \|X_{ij}\|_{\psi_2}/2 + \|X_{ik}\|_{\psi_2}/2.$$

Thus, if X_i has bounded sub-Gaussian entries, then the vectorized version of the empirical covariance matrix $\xi_i = \text{vec}(X_i X_i^T) \in \mathbb{R}^{p^2}$ has all entries satisfying the bounded sub-exponential assumption (C) in Section 2.3.1. So Theorem 2.7 (part (i)) implies that with a constant signal strength, $\mathbb{P}(\mathcal{S}_n) \geqslant 1 - O(\nu\varpi_{1,n})$ for $\epsilon_n = O(D_\nu^{-2} n^2 \log^4(np))$.

In [96], the authors studied a similar covariance matrix change points estimation procedure for centered and sub-Gaussian data points $X_i \in \mathbb{R}^p$ such that $\max_{1\leqslant i\leqslant n}\|X_i\|_{\psi_2} \leqslant B$, where $\|X\|_{\psi_2} = \sup_{v\in\mathbb{S}^{p-1}}\|v^T X\|_{\psi_2}$ and $\mathbb{S}^{p-1}$ is the unit sphere in $\mathbb{R}^p$. [96] proposed a binary segmentation in operator norm (BSOP) by recursively maximizing the operator norm of the matrix-valued CUSUM statistics (without the bootstrap calibration). [96, Theorem 1] stated that $\mathbb{P}(\mathcal{S}_n) \geqslant 1 - O(9^p n^{3-cp})$ for $\epsilon_n = O(D_\nu^{-2} n^{5/2}\sqrt{p\log(n)})$ and $p = O(n/\log(n))$. Our BABS improves the BSOP in the following aspects. First, our sub-exponential tail condition is weaker than [96] in view of $\max_{1\leqslant j\leqslant p}\|X_{ij}\|_{\psi_2} \leq \|X_i\|_{\psi_2}$. Second, we allow the dimension p to be sub-exponentially large in the sample size n. This is an essential benefit by working with the ℓ^∞-norm (rather than the operator norm). □

2.4 Simulation studies

In this section, we perform extensive simulation studies to investigate the size and power of the proposed bootstrap change point test, as well as the estimation error of the change point location(s).

2.4.1 Setup

We generate i.i.d. ξ_i in the mean-shifted model (2.1) from three distributions.

1. Multivariate Gaussian distribution: $\xi_i \sim N(0, V)$.

2. Multivariate elliptical t-distribution with degree of freedom ν: $\xi_i \sim t_\nu(V)$ with the probability density function [79, 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 $\Sigma = \nu/(\nu - 2)V$. In our simulation, we use $\nu = 6$.

3. Contaminated Gaussian: $\xi_i \sim \text{ctm-Gaussian}(\varepsilon, \nu, V)$ with the probability density function

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

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

We consider three cross-sectional dependence structures of V for each distribution.

(I) Independent: $V = \mathrm{Id}_p$, where Id_p is the $p \times p$ identity matrix.

(II) Strongly dependent (compound symmetry): $V = 0.8J + 0.2\mathrm{Id}_p$, where J is the $p \times p$ matrix containing all ones.

(III) Moderately dependent (autoregressive): $V_{ij} = 0.8^{|i-j|}$.

In all setups, 200 bootstrap samples are drawn for each simulation and all results are averaged on 1000 simulations.

2.4.2 Simulation results for single change point model.

In this section, we report size validity and power of our bootstrap change point test as well as the error of our CUSUM estimators for change point location.

Size of the bootstrap CUSUM test

We fix the sample size $n = 500$ and vary the dimension $p = 10, 300, 600$. For the bootstrap CUSUM test, we set the boundary removal parameter $\underline{s} = 30, 40$. For any significance level $\alpha \in (0, 1)$, we denote $\hat{R}(\alpha)$ as the rate of empirically rejected null hypothesis in 1000 simulations.

Under H_0, we first compare our bootstrap CUSUM test with two benchmark methods in the following.

(i) The bootstrapped log-ratio of maximized likelihood test (denoted as logLik) corresponds to $\log(\Lambda_s)$ in (2.3) when $p < n$ and

$$\log(\Lambda_s^*) = Z_n^*(s)^T \hat{\Sigma}^{-1} Z_n^*(s)/2,$$

where $\hat{\Sigma}$ is the sample covariance matrix of X_1^n and $Z_n^*(s)$ is the bootstrap CUSUM statistic (2.6). The testing procedure is similar to our bootstrap CUSUM test: reject H_0 if $\log(\Lambda_s)$ is greater than the $(1-\alpha)$-th quantile of bootstrapped $\log(\Lambda_s^*)$.

(ii) The oracle test is based on $\bar{Y}_n = |Y_n|_\infty$, i.e., the Gaussian approximation of T_n, with *known* covariance matrix. By our definition (2.32), Y_n is a Gaussian copy of $\text{vec}(\mathbf{Z}_n) = (Z_n(\underline{s})^\top, \ldots, Z_n(n - \underline{s})^\top)^\top$, whose covariance is close to $\text{Cov}(Z_n^*|X_1^n)$. The oracle test rejects H_0 if T_n is greater than the $(1-\alpha)$-th quantile of $\bar{Y}_n$.

Table 2.1 lists the uniform error-in-size $\sup_{\alpha\in(0,1)} |\hat{R}(\alpha) - \alpha|$ for our test and the two benchmarks. This metric reflects the Kolmogorov distance between distributions of T_n and its bootstrapped analogue T_n^*: the smaller uniform error-in-size, the closer $\rho^*(T_n, T_n^*)$. In Table 2.1, each column corresponds to a combination of one distribution family and one cross-sectional dependence structures, and the rows compare the logLik (when $p < n$), our proposed bootstrap method with $\underline{s} = 30, 40$ and the corresponding benchmark $\bar{Y}_n$ under different settings of $p = 10, 300, 600$. There are several observations we can draw. First, our proposed test has much smaller errors than logLik even when $p = 10$ and 300. Second, the errors of our tests are comparable with those from corresponding $\bar{Y}_n$ in all settings. It shows that our bootstrapped T_n^* is remarkably close to $\bar{Y}_n$ that is generated with given Σ. So our CUSUM test works well as a fully data-dependent approach compared to benchmarks.

We can draw other conclusions for our bootstrapped CUSUM test by comparing results under different choices of boundary removal parameter, distribution family and cross-sectional dependence structure, respectively. In addition to Table 2.1, Table 2.2 provides the empirical Type I error $\hat{R}(\alpha)$ at significance level $\alpha = 0.05$. In general, the scenario that has small uniform error-in-size in Table 2.1 and close to 0.05 error rates in Table 2.2 indicates a precise approximation in size. First, in most cases the errors of $\underline{s} = 40$ in Table 2.1 are smaller than that of $\underline{s} = 30$. Besides, the empirical Type I errors $\hat{R}(0.05)$ are generally close to the nominal size 0.05 for $\underline{s} = 40$. These two tables explain that the greater the boundary removal parameter $\underline{s}$, the better the approximation under H_0. Next, uniform errors-in-size is usually smaller for the Gaussian distribution than t_6 and ctm-Gaussian cases. This is due to one less error occurred in approximation from non-Gaussian CUSUM statistics $Z_n(s)$ to Gaussian analogs $Y_n(s)$. Table 2.2 delivers similar message that Gaussianity helps to control $\hat{R}(0.05)$. Finally, our method is robust to the cross-sectional dependence structure. This is because our procedure manages to capture the coordinate-wise dependency without any estimation steps. In many cases, stronger dependence (II>III>I) is more beneficial for reducing the approximation errors. To sum up, size under H_0 can be better controlled if $\underline{s}$ is large, data is Gaussian distributed or strong cross-sectional dependence exists. As a visualization of the accuracy for size controlling, Figure 2.1 displays three example curves of $\hat{R}(\alpha)$ for our proposed test where $p = 600, \underline{s} = 40$. The rejection rate $\hat{R}(\alpha)$ follows closely along the diagonal line in dash (i.e. the line of $\hat{R}(\alpha) = \alpha$).

$n = 500$		Gaussian			t_6			ctm-Gaussian		
		I	II	III	I	II	III	I	II	III
$p = 10$	logLik	0.150	0.151	0.159	0.122	0.134	0.124	0.136	0.134	0.124
	$\underline{s} = 30$	0.034	0.036	0.041	0.048	0.041	0.039	0.036	0.042	0.021
	$\bar{Y}_n$	0.037	0.041	0.022	0.030	0.030	0.038	0.035	0.029	0.052
	$\underline{s} = 40$	0.042	0.034	0.037	0.043	0.037	0.033	0.041	0.042	0.043
	$\bar{Y}_n$	0.028	0.041	0.030	0.028	0.053	0.044	0.023	0.033	0.035
$p = 300$	logLik	0.837	0.835	0.834	0.680	0.669	0.670	0.691	0.679	0.681
	$\underline{s} = 30$	0.054	0.051	0.050	0.085	0.036	0.049	0.115	0.025	0.065
	$\bar{Y}_n$	0.024	0.046	0.039	0.057	0.064	0.066	0.044	0.067	0.036
	$\underline{s} = 40$	0.046	0.026	0.035	0.058	0.030	0.040	0.057	0.032	0.055
	$\bar{Y}_n$	0.033	0.025	0.066	0.060	0.033	0.064	0.045	0.025	0.061
$p = 600$	$\underline{s} = 30$	0.051	0.035	0.048	0.122	0.044	0.088	0.103	0.030	0.096
	$\bar{Y}_n$	0.030	0.045	0.049	0.048	0.043	0.068	0.039	0.055	0.051
	$\underline{s} = 40$	0.060	0.055	0.046	0.083	0.038	0.087	0.079	0.026	0.057
	$\bar{Y}_n$	0.041	0.041	0.042	0.041	0.025	0.049	0.034	0.038	0.077

Table 2.1: Uniform error-in-size $\sup_{\alpha\in[0,1]} |\hat{R}(\alpha) - \alpha|$ under H_0 compared with benchmarks. Scenarios I-III are $V = I$, $V = 0.8J + 0.2I$ and $V_{ij} = 0.8^{|i-j|}$, respectively.

Next, we compare our method with the tests in [65] and [43] under the setting $n = 500, p = 600, \underline{s} = 40$. In [65], change points are allowed to occur at different locations in each coordinate. To avoid simulation issue in estimating the long-run variance (which is necessary because [65] uses non-stationary CUSUM statistics), we employ the true variance $\hat{\sigma}_h^2 = \sigma_h^2$ in their test statistic [65, Equation (1.2)] and take the

$n = 500$		Gaussian			t_6			ctm-Gaussian		
		I	II	III	I	II	III	I	II	III
$p = 10$	$\underline{s} = 30$	0.051	0.052	0.051	0.044	0.056	0.034	0.043	0.056	0.052
	$\underline{s} = 40$	0.046	0.055	0.052	0.045	0.050	0.048	0.054	0.053	0.040
$p = 300$	$\underline{s} = 30$	0.026	0.054	0.039	0.018	0.034	0.018	0.021	0.043	0.027
	$\underline{s} = 40$	0.045	0.043	0.044	0.024	0.046	0.036	0.026	0.056	0.034
$p = 600$	$\underline{s} = 30$	0.026	0.060	0.027	0.010	0.034	0.020	0.010	0.053	0.019
	$\underline{s} = 40$	0.031	0.038	0.036	0.020	0.044	0.016	0.015	0.042	0.027

Table 2.2: $\hat{R}(0.05)$: empirical Type I error with nominal level 0.05 for our test.

suggested multiplier $\xi_l^2 = 1$ in $\hat{s}_h^2$ [65, Equation (4.5)] when calculating conditional long-run variance in bootstrap. This modified stronger algorithm is denoted by their test statistic B_n. In observation of the fact that long-run variances for all coordinates are the same in our simulation setting, we also enhanced it to another competitor denoted as B_n-enhanced by removing both $\hat{\sigma}_h^2$ and $\hat{s}_h^2$. In [43], the original test without boundary removing and an improved version with $\underline{s} = 40$ are implemented (denoted as their test statistic ψ and ψ-improved, respectively). We set the tuning parameter κ to be 6.6 as suggested by the authors. Similar to Table 2.1 and 2.2, uniform error-in-size is shown in Table 2.3 and empirical Type I error at $\alpha = 0.05$ is shown in Table 2.4.

Based on Table 2.3 and 2.4, we draw the following comparison results. First, the test ψ in [43] suffers from size distortion except for the case of Gaussian distribution with identity covariance structure because the ψ relies heavily on the assumption $X_i \sim N(\mathbf{0}, \sigma^2 \mathrm{Id}_p)$. Second, boundary removal helps ψ to reduce the uniform error-in-size, cf. the ψ-improved. Third, the test B_n in [65] and B_n-enhanced behave similarly and they are comparable with ours. However, note that the two B_n tests receive stronger priori that coordinate-wise (long-run) variances are all equal or assumed given, which is impractical. On the contrary, our proposed method does not involve such estimators of variance. Lastly, in Table 2.4, the test B_n inflates Type-I error than pre-specified 5% while ours behaves conversely.

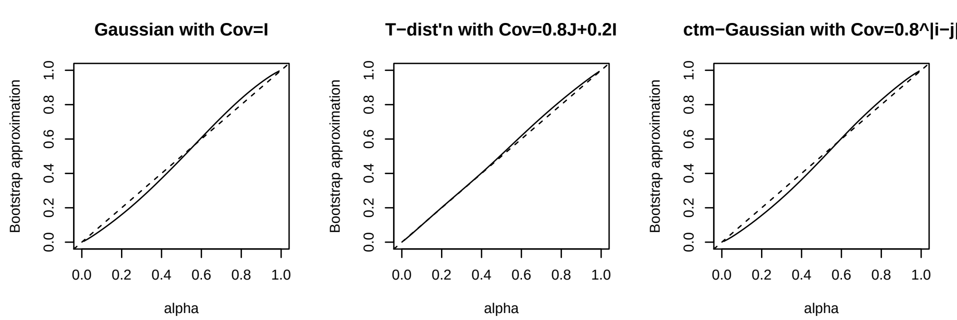


Figure 2.1: Selected setups for comparing $\hat{R}(\alpha)$ with α. Here, $n = 500, p = 600$ and the boundary removal is 40.

$n = 500, p = 600$		Gaussian			t_6			ctm-Gaussian		
		I	II	III	I	II	III	I	II	III
B_n	$\underline{s} = 40$	0.073	0.026	0.049	0.101	0.019	0.087	0.083	0.017	0.068
B_n-enhanced	$\underline{s} = 40$	0.071	0.025	0.063	0.063	0.041	0.068	0.061	0.016	0.067
ψ	$\underline{s} = 1$	0.368	0.988	0.882	0.990	0.990	0.990	0.990	0.990	0.990
ψ-improved	$\underline{s} = 40$	0.186	0.959	0.646	0.990	0.990	0.990	0.990	0.987	0.990
T_n	$\underline{s} = 40$	0.060	0.055	0.046	0.083	0.038	0.087	0.079	0.026	0.057

Table 2.3: Uniform error-in-size $\sup_{\alpha \in [0,1]} |\hat{R}(\alpha) - \alpha|$ under H_0 compared with [65] (B_n and B_n enhanced) and [43] (ψ and ψ improved). T_n stands for our test and the values are copied from the last row of Table 2.1.

$n = 500, p = 600$		Gaussian			t_6			ctm-Gaussian		
		I	II	III	I	II	III	I	II	III
B_n	$\underline{s} = 40$	0.054	0.073	0.063	0.066	0.060	0.059	0.056	0.057	0.063
B_n-enhanced	$\underline{s} = 40$	0.057	0.065	0.061	0.038	0.052	0.054	0.061	0.059	0.058
ψ	$\underline{s} = 1$	0.110	0.998	0.923	1	1	1	1	1	1
ψ-improved	$\underline{s} = 40$	0.055	0.973	0.696	1	1	1	1	0.997	1
T_n	$\underline{s} = 40$	0.031	0.038	0.036	0.020	0.044	0.016	0.015	0.042	0.027

Table 2.4: $\hat{R}(0.05)$: empirical Type I error with nominal level 0.05 for [65] (B_n and B_n enhanced), [43] (ψ and ψ improved), and ours. T_n stands for our test and the values are copied from the last row of Table 2.2.

Power of the bootstrap CUSUM test.

Under $H_1 : \mu_1 = \cdots = \mu_m \neq \mu_{m+1} = \cdots = \mu_n$, we consider the single change point location m at $\{50, 150, 250\}$ (such that $t_m = m/n = 1/10, 3/10, 5/10$ for $n = 500$). Denote k as the number of covariates that have change points, i.e., $\delta_{n,1} = \cdots = \delta_{n,k} \neq 0$. Two types of the mean-shift signal are considered: $k = 1$ for sparse signal and $k = 50$ for dense signal. Since the ℓ^∞-norm targets on sparse signal, we only present the sparse alternative case of $k = 1$ in this section, and results of dense signal for [43] can be found in Section 2.6. To analyze the power under H_1, we fixed observation $n = 500$, $p = 600$, $\underline{s} = 40$ and the significance level $\alpha = 0.05$.

We first compare the power performance of our bootstrap test with the benchmark oracle test (in Section 2.4.2) and investigate the impact from change-point location and distribution. Figure 2.2 shows the empirical power curves v.s. the signal strength $|\delta_n|_\infty = |\delta_{n,1}|$. The left figure displays the impact from change-point location where data follows the ctm-Gaussian distribution (III). We observe that the powers monotonically increase along $|\delta_n|_\infty$ and eventually reach 1 as $|\delta_{n,1}|$ gets large enough. Furthermore, change points closer to boundaries are more difficult to detect at the same signal strength. In addition, note that the power curves of our bootstrap test follows almost identically to the oracle test. The right plot of Figure 2.2 displays the power curves of our method under different distributions and cross-sectional covariance structures. Comparing the diamond-symboled curves, Gaussianity is helpful in terms of power. Similar to the size results in Section 2.4.2, curves in solid line demonstrate that the stronger the cross-sectional dependence, the greater the power. Complete report of powers can be found in Table 2.9 in Section 2.6.

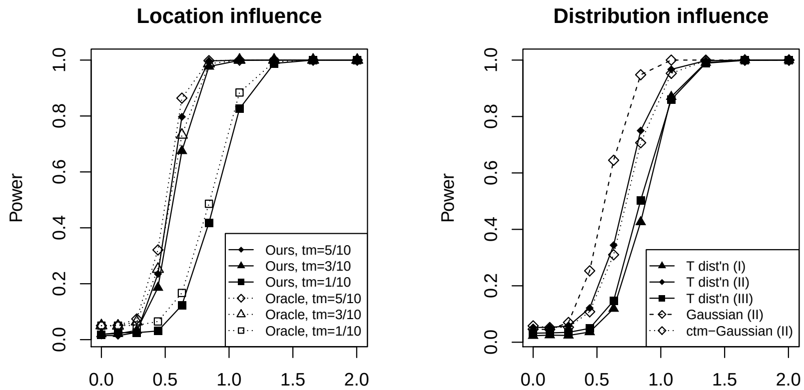
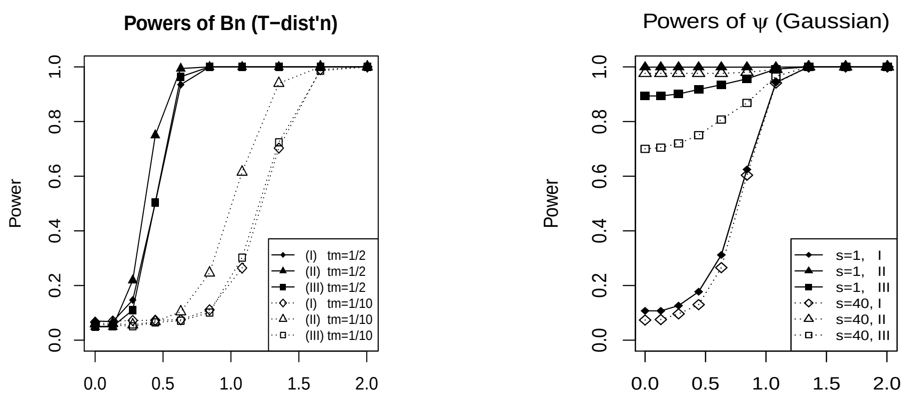


Figure 2.2: Selected power curves in different setups. Left: our method compared with oracle $\bar{Y}_n$ for various t_m. Right: investigated data structure effect where $t_m = 1/10$ fixed. Here, $n = 500$, $p = 600$ and the boundary removal is 40.



(a) Powers of B_n [65] for t_6 distribution.

(b) Powers of original ($\underline{s} = 1$) and improved ψ ($\underline{s} = 40$) [43] for $t_m = 1/2$.

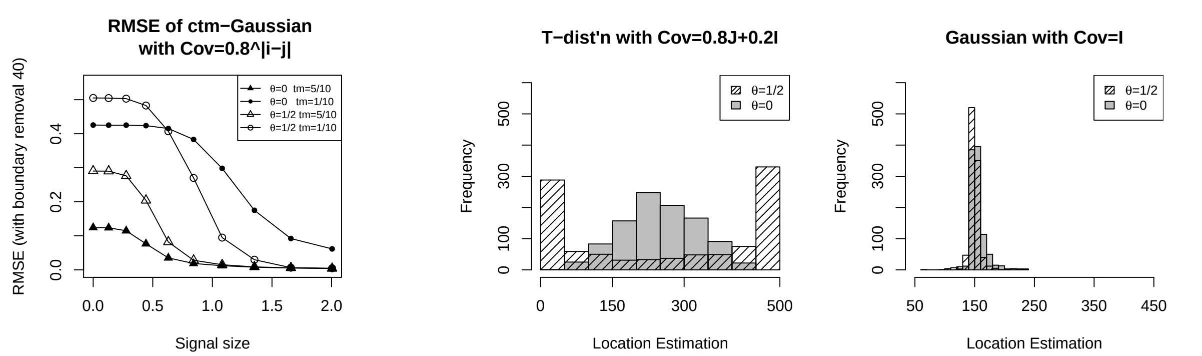
Figure 2.3: Powers of [43], [65] for sparse alternative. Here, $n = 500, p = 600$.

Next, we compare our method with [65] and [43] that are replicated in Figure 2.3. Figure 2.3(a) shows power trends of B_n [65] for t-distributed data under the sparse alternative when $t_m = 1/2$ and $t_m = 1/10$. The test B_n performs better when change point locates at the center than boundary, which is a similar phenomenon as shown in our algorithm. When comparing the solid curves in Figure 2.2-Right and the dashed curves in Figure 2.3(a), we can find that boundary change ($t_m = 1/10$) brings more challenge to B_n than to our test because our powers increase faster than B_n under the same setup. We refer to Table 2.10 for complete power reports of B_n in all scenarios under sparse H_1. We would highlight that although B_n returns slightly higher power than ours when $t_m = 1/2$ (change point is at the center), B_n is computed with true long-run variance and Table 2.4 already indicates that it tend to over reject H_0. Figure 2.3(b) displays power trends of the ψ in [43] ($\underline{s} = 1$) and ψ-improved ($\underline{s} = 40$) for Gaussian distributed data when $t_m = 1/2$. Both ψ tests lose control in power when the independent covariance assumption is violated. Besides, unreported results show invalid power curves in all other cases including t_6 and ctm-Gaussian distributed data. It is unsurprising since we have observed in Table 2.4 that the ψ test suffers from serious size distortion. We refer to Table 2.11 in Section 2.6 for complete power reports of ψ in independent Gaussian scenarios under both sparse and dense H_1.

Performance of the location estimators.

Now we examine the performance of our location estimators $t_{\hat{m}_{1/2}}$ and $t_{\hat{m}_0}$ under sparse alternative where $t_m = 1/10, 3/10, 5/10$. The performance measure is the root-mean-square error (RMSE). Since distribution influence (tail thickness and cross-sectional dependence) has already been thoroughly explored, this part will focus on the comparison among $t_{\hat{m}_{1/2}}$, $t_{\hat{m}_0}$ and other competing methods.

Figure 2.4 shows comparison between $t_{\hat{m}_{1/2}}$ and $t_{\hat{m}_0}$ across different change point location and signal size $|\delta_n|_\infty$. First, Figure 2.4(a) illustrates that locations of change points closer to boundary (i.e., $t_m = 1/10$) are harder to estimate as the RMSEs are uniformly larger than corresponding RMSEs for $t_m = 5/10$. Second, it agrees with Theorem 2.4 and 2.5 that RMSEs of $t_{\hat{m}_0}$ are smaller than that of $t_{\hat{m}_{1/2}}$ when change points are in the middle. As mentioned in Section 2.2.2, $Z_{\theta,n}(s)$ assigns less weights to the boundary points for smaller values of θ. This is also empirically confirmed by Figure 2.4(b): $t_{\hat{m}_{1/2}}$ is more in favor of boundary points when $|\delta_n|_\infty = 0$, and $t_{\hat{m}_0}$ slightly leans to the center when change ($|\delta_n|_\infty = 0.842$) is not in the middle of sequence.

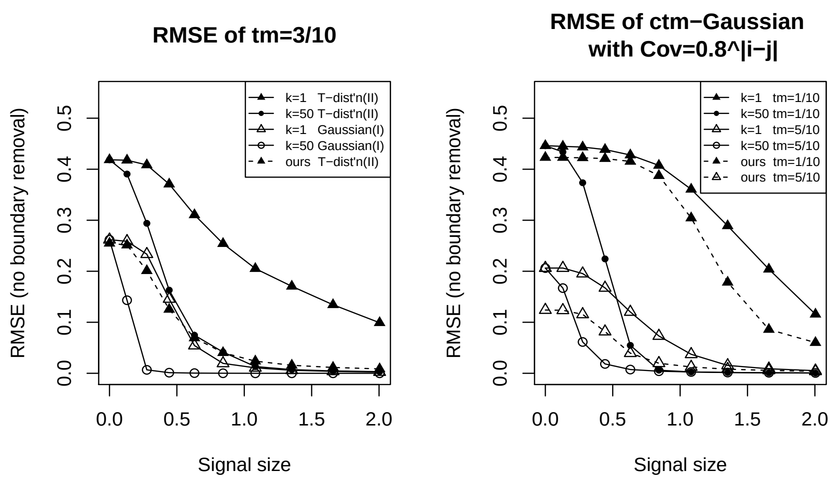


(a) The RMSEs v.s. $|\delta_{n,1}|$ of our two estimators. Here, change point locations are close to the boundary or at the center.

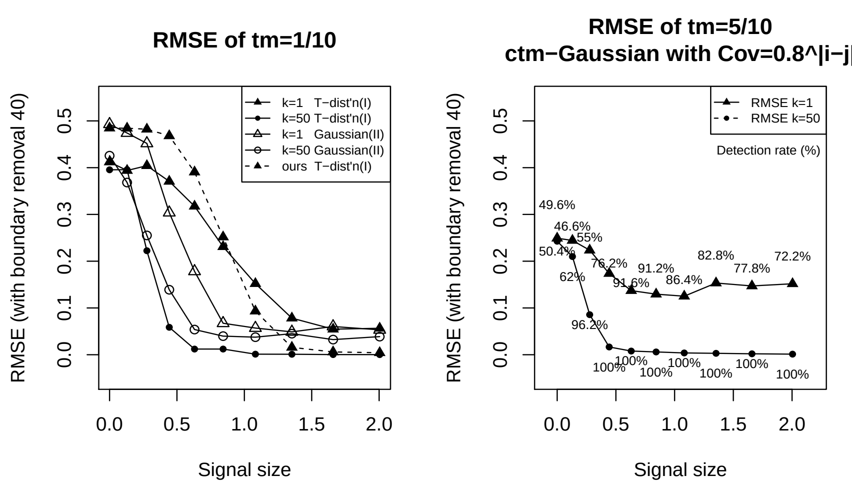
(b) Histograms of our two estimators at $\delta_{n,1} = 0$ (left) $\delta_{n,1} = 0.842$ (right). Here, change point location $m = 0.3n = 150$.

Figure 2.4: RMSEs (left) and histograms (right) of $t_{\hat{m}_{1/2}}$ and $t_{\hat{m}_0}$. Here, $n = 500, p = 600$.

Next, we compare our estimators with [99] and [37]. In [99], a projection based estimator named Inspect is proposed. Theoretical analysis of this algorithm requires the data following Gaussian distribution. In [37], the proposed SBS estimator is the maximizer of threshold ℓ^1-aggregated CUSUM statistics after thresholding. This method is sensitive to threshold tuning parameters selected by bootstrap. Both [99] and [37] allow multiple change points. For now, we first compare with their single change point versions (see R packages `InspectChangepoint` and `hdbinseg`). We also include a truncated version of our location estimator $\hat{m}_\theta = \text{argmax}_{\underline{s} \leqslant s < n-\underline{s}} |Z_{\theta,n}(s)|_\infty$ (cf. Remark 3) for fair comparison. Recall k as the signal density (where $\delta_{n,1} = \cdots = \delta_{n,k} \neq 0$). Both $k = 1, 50$ are considered for these two methods.



(a) RMSEs of [99] (sparse & dense signal) and our non-truncated $t_{\hat{m}_0}$ (sparse signal). Distribution (left) and location (right) effects are investigated.



(b) RMSEs of [37] (sparse & dense signal) and our truncated $t_{\hat{m}_{1/2}}$ (sparse signal). Distribution (left) and location (right) effects are investigated. $\underline{s} = 40$ for all cases.

Figure 2.5: Comparison of location estimators among [99], [37] and ours. Here, $n = 500, p = 600$.

Figure 2.5 illustrates contrasts among the three algorithms. Figure 2.5(a) compares our non-truncated $t_{\hat{m}_0}$ with [99] which has no boundary removal. On the left plot (where $t_m = 3/10$ is fixed), the RMSEs of [99] are large when signal is sparse or data is non-Gaussian distributed with cross-sectional dependent. However, our $t_{\hat{m}_0}$ can identify sparse signal under t_6 distribution with high cross-sectional correlation and performs even comparably to their estimator when there is dense signal. On the right plot (where distribution is fixed as ctm-Gaussian with cross-sectional structure III), both approaches have large RMSEs when the change point is close to boundary. But ours is uniformly better under the corresponding sparse alternative. Figure 2.5(b) compares our truncated $t_{\hat{m}_{1/2}}$ with [37] which also has boundary removal in their R package. We set $\underline{s} = 40$ for both algorithms. Their method works well for the case of $t_m = 1/10$ as shown in the left of Figure 2.5(b), and distribution has less effect on [37] than on [99]. But when signal size is 1 and larger, our $t_{\hat{m}_{1/2}}$ almost dominates solid-symbol curves that are under the same distribution. In the right of Figure 2.5(b), however, [37] returns non-monotone RMSEs when $t_m = 5/10, k = 1$. Even when $\delta = 0$, the discovery rate is as much as 50%. This phenomenon means that large CUSUM values may lead to unreasonable threshold in their method. In Section 2.6, non-monotone RMSE curves are reported in Figure 2.15 for the same case with $\alpha = 0.01$. Even under dense alternative, their searching process fail to discover a change when signal is strong.

To summarize, numerical studies suggest that our estimators are more flexible and stable under various distribution and cross-sectional dependency circumstances regardless of the location of change point. The full RMSEs of our method with sparse signal and selected RMSEs of [99] and [37] are reported in Table 2.12, 2.13 and 2.14 in Section 2.6.

2.4.3 Multiple change points estimation by BABS

In the multiple change-point scenario, we first let the k-th component of $\delta_n^{(k)}$ to have the same mean shift, i.e. $\delta_{n,1}^{(1)} = \delta_{n,2}^{(2)} = \cdots = \delta_{n,\nu}^{(\nu)} = \delta \neq 0$. Since change point estimation can be viewed as a special case of clustering, the accuracy can be measured by the adjusted Rand index (ARI) [86, 63]. We also reported average ARI over all 500 runs.

Simulation of Algorithm 1

We use t_6 distribution with dependence structure (III) as a representative. Let $n = 1000, p = 1200, \underline{s} = 40, \alpha = 0.05$, and the two change points $(m_1, m_2) = (300, 600)$. Some observations can be drawn based on Table 2.5 and Figure 2.6. When signal $\delta = 0.317$ is small, Algorithm 1 cannot locate mean shift accurately. However, as signal gets larger, both the number and the locations of change points can be detected consistently. Meanwhile, ARI is also increasing to 1, which stands for the perfect estimation. We further add one more change where $(m_1, m_2, m_3) = (300, 600, 800)$. Compared with the previous setup, Table 2.6 and Figure 2.7 shows that the estimation is slightly worse under the same signal size. This is because the effective sample size is cut down after each binary segmentation. But our algorithm eventually detect all change points consistently when $\delta = 2$ is large enough.

	δ	0	0.317	0.733	1.282	2.004
Estimated	0	**497**	378	1	0	0
number of	1	3	117	13	0	0
change	**2**	0	**5**	**458**	**464**	**470**
points	3	0	0	26	35	30
	4	0	0	2	1	0
Sum		500	500	500	500	500
ARI		0.994	0.128	0.935	0.978	0.989

Table 2.5: Multiple change point setup with 2 change points $(m_1, m_2) = (300, 600)$: counts of estimated $\hat{\nu}$ and ARI.

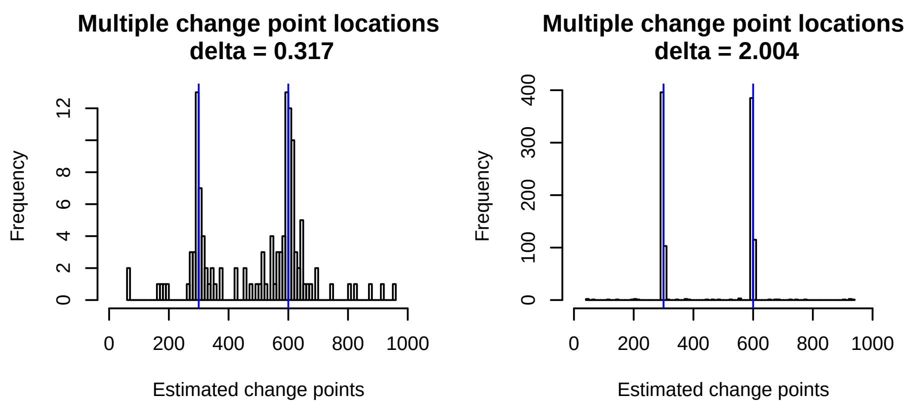


Figure 2.6: Multiple change point setup with 2 change points $(m_1, m_2, m_3) = (300, 600, 800)$: counts of estimated change points over 500 repeats at signal level $\delta = 0.317, 2.004$.

Comparison with [99]

We also apply the setup in [99, Section 5.3]. Here, $n = 2000, p = 200, (m_1, m_2, m_3) = (500, 1000, 1500)$ and data is Gaussian distributed with identity cross-sectional covariance. We consider the complete-overlap mean structure, i.e., changes occur in the same k coordinates, and set $(||\delta_n^{(1)}||_2, ||\delta_n^{(2)}||_2, ||\delta_n^{(3)}||_2) = (\delta, 2\delta, 3\delta)$ for signal strength $\delta \in \{0.4, 0.6\}$. Table 2.7 and Figure 2.8 summarize estimation performance of our `BABS` and the `Inspect` [99]. If $k = 40$, $|\delta_n^{(i)}|_\infty = i \cdot \delta/\sqrt{k}$ is too small for our method to be recognizable. However, when $k = 1$ such that $\delta_n^{(i)}, i = 1, 2, 3$ is sparse (with ℓ^2-norm unchanged), then our algorithm shows superiority in terms of both $\hat{\nu}$ and ARI. Again, our `BABS` algorithm has advantage when the ℓ^∞-norm of signal is bounded

δ		0	0.317	0.733	1.282	2.004
	0	**491**	360	0	0	0
Estimated	1	9	133	5	0	0
number of	2	0	7	141	0	0
change	**3**	0	**0**	**328**	**455**	**474**
points	4	0	0	25	40	25
	5	0	0	1	5	1
Sum		500	500	500	500	500
ARI		0.982	0.106	0.871	0.973	0.989

Table 2.6: Multiple change point setup with 3 change points $(m_1, m_2, m_3) = (300, 600, 800)$: counts of estimated $\hat{\nu}$ and ARI.

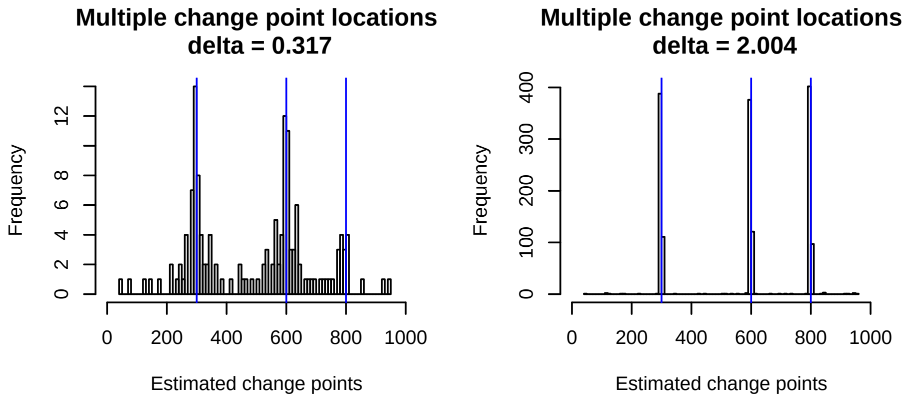


Figure 2.7: Setup 2. Counts of estimated change points over 500 repeats at signal level $\delta = 0.317, 2.004$.

below and likewise when data is non-Gaussian or there is cross-sectional dependence as shown in Section 2.4.2 and 2.4.2. For comparison with [37, 36, 65], we refer to [99, Section 5.3] for a comprehensive simulation study.

2.4.4 Extension to time series: a block multiplier bootstrap

The Gaussian multiplier bootstrap CUSUM test statistic in (2.6) and (2.7) can be extended to a block version where the temporal dependence can be handled. Since the CUSUM test statistic $Z_n(s)$ in (2.4) can be re-written as a block sum, we propose a block version of the bootstrap CUSUM test (2.6) that can accommodate time series data. Precisely, let M, B be positive integers such that $n = MB$. We divide the sample $X_1, \ldots, X_n$ into B blocks of size M. In particular, for $b = 1, \ldots, B$, let $L_b = \{(b-1)M+1, \ldots, bM\}$ be the b-th block. Then, for $s = 1, \ldots, n-1$, we can write $Z_n(s)$ in (2.4) as

$$Z_n(s) = \sqrt{\frac{n-s}{ns}} \sum_{b=1}^{B} \sum_{i \in L_b} X_i \mathbf{1}(i \leqslant s) - \sqrt{\frac{s}{n(n-s)}} \sum_{b=1}^{B} \sum_{i \in L_b} X_i \mathbf{1}(i > s).$$

Parameters in methods		k = 40				k = 1			
		$\delta = 0.4$		$\delta = 0.6$		$\delta = 0.4$		$\delta = 0.6$	
		BABS	Inspect	BABS	Inspect	BABS	Inspect	BABS	Inspect
Estimated number of change points	0	0	0	0	0	0	0	0	0
	1	16	0	0	0	0	0	0	0
	2	329	426	194	182	14	466	0	172
	3	**121**	**71**	**206**	**295**	**399**	**30**	**438**	**305**
	4	28	3	83	17	78	4	56	19
	5	6	0	16	6	7	0	6	4
	6	0	0	1	0	2	0	0	0
Sum		500	500	500	500	500	500	500	500
ARI		0.591	0.711	0.683	0.854	0.941	0.704	0.975	0.883

Table 2.7: Comparison of multiple change points detectors.

For any $\alpha \in (0,1)$, we reject H_0 if the test statistic $T_n = \max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |Z_n(s)|_\infty$ is larger than a critical value. Since the distributions of $Z_n(s)$ under H_0 and H_1 for dependent error processes are different from the i.i.d. errors, we need to accommodate the dependence in calibrating the distributions of the test statistic T_n. The idea is to modify the Gaussian multiplier bootstrap Z_n^* in (2.6) and the bootstrap CUSUM test statistic T_n^* in (2.7) to their block versions. Specifically, to approximate the (finite sample) distribution of T_n, we propose a *block Gaussian multiplier bootstrap* tailored to the CUSUM statistics setting. Let $e_1, \ldots, e_B$ be i.i.d. standard Gaussian random variables. Define

$$Z_n^\sharp(s) = \sqrt{\frac{n-s}{ns}} \sum_{b=1}^{B} e_b V_b^-(s) - \sqrt{\frac{s}{n(n-s)}} \sum_{b=1}^{B} e_b V_b^+(s),$$

where

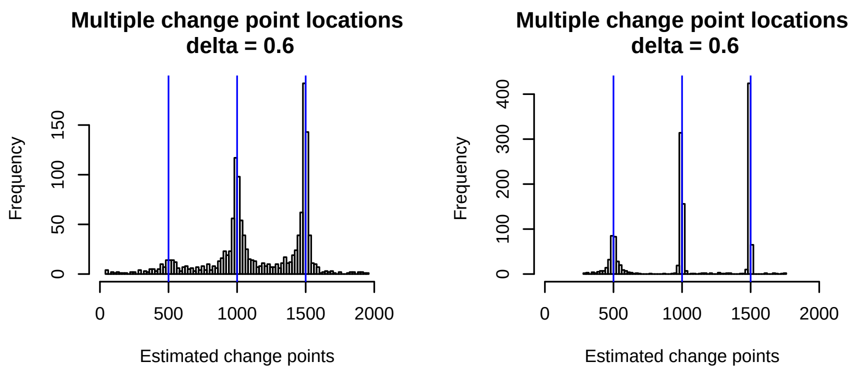
$$V_b^-(s) = \sum_{i \in L_b} (X_i - \bar{X}_s^-) \mathbf{1}(i \leqslant s) \quad \text{and} \quad V_b^+(s) = \sum_{i \in L_b} (X_i - \bar{X}_s^+) \mathbf{1}(i > s).$$

Then the distribution of T_n is approximated by its bootstrap analog given by

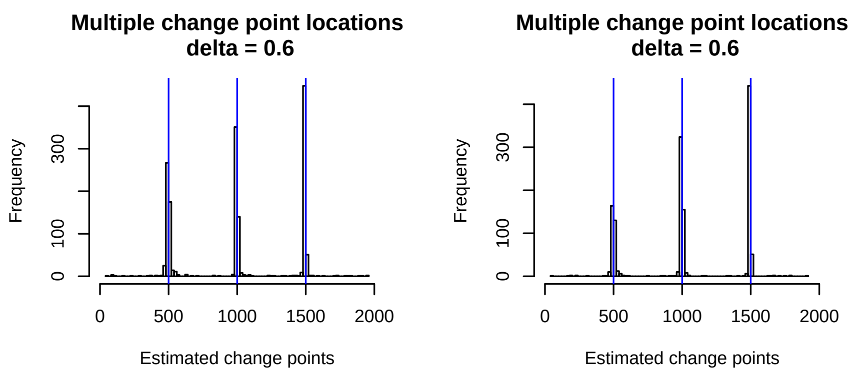
$$T_n^\sharp = \max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |Z_n^\sharp(s)|_\infty.$$

For any $\alpha \in (0,1)$, we reject H_0 if $T_n > q_{T_n^\sharp | X_1^n}(1-\alpha)$, where $q_{T_n^\sharp | X_1^n}(1-\alpha) = \inf\{t \in \mathbb{R} : \mathbb{P}(T_n^\sharp \leqslant t | X_1^n) \geqslant 1-\alpha\}$ is the $(1-\alpha)$ conditional quantile of $T_n^\sharp$ given X_1^n. Note that if the block size $M = 1$ (i.e., $B = n$), then $Z_n^\sharp(s) = Z_n^*(s)$. Thus the bootstrap CUSUM test statistic for independent sequences is a special case of the block CUSUM test statistic. Generally, larger M is needed for stronger temporal dependence, while this would reduce the effective sample size.

We shall study the empirical performance of the block multiplier bootstrap CUSUM test for some dependent process ξ_i. Theoretical supports for the block bootstrap CUSUM test are beyond the scope of this



(a) Dense signal where $k = 40$. Left: Our `BABS` algorithm with $\alpha = 0.05, B = 200, \underline{s} = 40$; Right: `Inspect` algorithm with default setting in their R package `InspectChangepoint`.



(b) Same setting except $k = 1$. Left: `BABS`; Right: `Inspect`.

Figure 2.8: Histograms of estimated change point locations in complete-overlap structure. Parameters $n = 2000, p = 200, (m_1, m_2, m_3) = (500, 1000, 1500), (||\delta_n^{(1)}||_2, ||\delta_n^{(2)}||_2, ||\delta_n^{(3)}||_2) = (0.6, 1.2, 1.8)$.

paper and they can be derived using the recent development of the Gaussian and bootstrap approximation results for high-dimensional time series (see e.g., [110, 108]).

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. In our simulation, we generate a random matrix A with i.i.d. $N(0, 1)$ entries. To ensure the stationarity of ξ_i process, A is normalized such that $\|A\|_2 = 1/1.8 < 1$. In this section, we fix $n = 500, p = 600, \underline{s} = 40$, and consider different block sizes $M = 2, 5, 10, 15$.

We first investigate the performance of the modified block Gaussian multiplier bootstrap CUSUM test. Since distributional impact has already been evaluated in Section 2.4.2, the same setup in Figure 2.1 are selected as example to illustrate the impact of M. In Figure 2.9, the $\hat{R}(\alpha)$ curves show similar (and slightly less accurate) behavior as in temporally independent case. The approximation accuracy also depends on the block size parameter M (which adjusts for the temporal dependence), in addition to cross-sectional dependence.

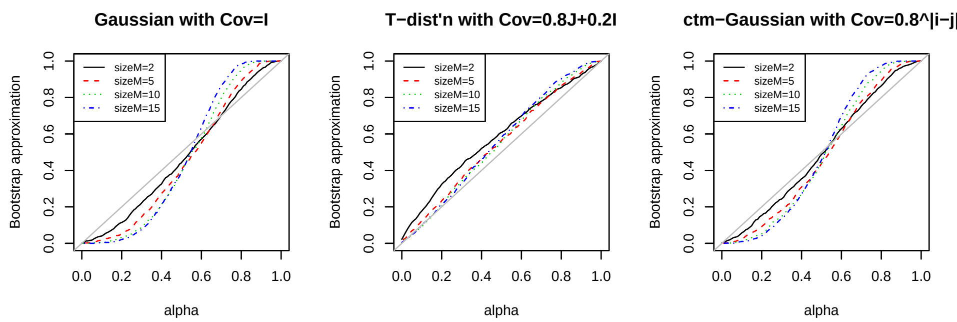
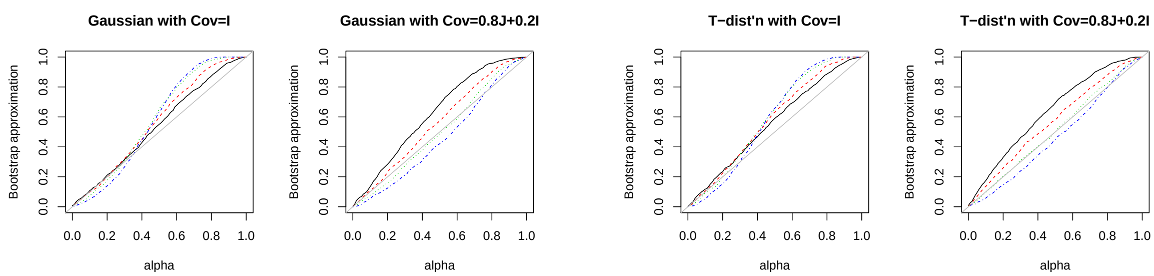


Figure 2.9: Selected plots in comparing bootstrap rejection $\hat{R}(\alpha)$ v.s. level α for our block Gaussian multiplier bootstrap test under H_0. Here, $n = 500$, $p = 600$ and the boundary removal is 40.



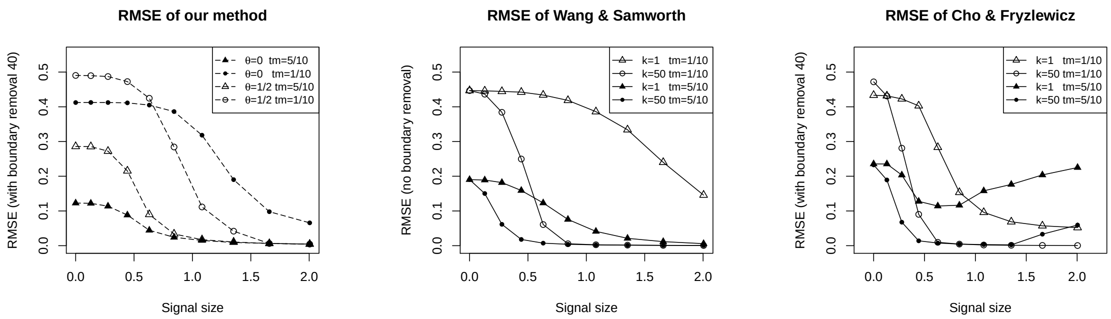
(a) Gaussian distributed η_i.

(b) T_6 distributed η_i.

Figure 2.10: Selected $\hat{R}(\alpha)$ plots for block-wise bootstrap testing in [65] under H_0. Here, $n = 500, p = 600$, the boundary removal is 40, $\hat{\sigma}_h^2 = \sigma_h^2$ and $\xi_l = 1$ is used in conditional long-run variance estimators $\hat{s}_h^2$ in bootstrap. The legend is the same as in Figure 2.9.

In Figure 2.10, similar observation can be found for the block bootstrap algorithm in [65]. Since their performance is sensitive to variance estimators, we substitute long-run variance estimator $\hat{\sigma}_h^2$ by its theoretical value $\sigma_h^2, h = 1, \cdots, p$, i.e. the h-th diagonal element of $\Sigma(0) + \sum_{l=1}^{\infty} A^l \Sigma(0) + \Sigma(0) \sum_{l=1}^{\infty} (A^T)^l$ where $\Sigma(0) = \sum_{l=0}^{\infty} A^l \Sigma (A^T)^l$ is the lag-0 auto-covariance of $\{X_i\}$. And $\xi_l^2 = 1$ is used in $\hat{s}_h^2$, the conditional variance estimator in bootstrap. The size approximation is accurate under spatial independent scenarios $V = Id_p$, while larger block size M is suggested when $V = 0.8J + 0.2Id_p$. Note that, we primarily compare the performances of bootstrap testing instead of estimation of $\hat{\sigma}_h^2$ that is an influential factor in practice.

We also examine our location estimators for the temporal dependence case, and compare with [99] and [37]. Figure 2.11 provides RMSE curves of the three algorithms for ctm-Gaussian data with cross-sectional dependence (III). Similar conclusions can be drawn as in temporal independence case. The RMSEs of our method, [99], and [37] are reported in Table 2.12, 2.13 and 2.14 in Section 2.6.



(a) Our two estimators $t_{\hat{m}_0}$ and $t_{\hat{m}_{1/2}}$ (truncated version with boundary removal $\underline{s} = 40$).

(b) Estimators of [99] (sparse & dense signal cases with no boundary removal).

(c) Estimators of [37] (sparse & dense signal case with boundary removal $\underline{s} = 40$).

Figure 2.11: RMSEs v.s. signal size $|\delta_n|_\infty$ for our algorithm (left), [99] (middle) and [37] (right). Here, $n = 500, p = 600$ and data are from ctm-Gaussian distribution with covariance $V_{i,j} = 0.8^{|i-j|}$.

2.5 Real data applications

2.5.1 Multiple change-point detection using Algorithm 1

The micro-array dataset `aCGH` from the `ecp` R package in [64] consists of $p = 43$ individuals with bladder tumors. There are $n = 2215$ log-intensity-ratio fluorescent measurements of DNA segments that share almost identical change points because the individuals have the same disease. We set $B = 1000, \alpha = 0.05, \underline{s} = 60$. Our `BABS` algorithm finds 27 change points in the copy-number that are marked as red vertical dashed lines in Figure 2.12, which displays the sequences of the first 10 patients. Compared to `Inspect`, which identifies 254 change points by using the default threshold, our discovery is more reasonable and stable under the existence of outliers (e.g. the segment between 1724 and 1836 or the one between 1965 and 2044).

2.5.2 Time-series data using the block multiplier bootstrap test

We apply our block multiplier bootstrap test and location estimators to stock return data that is available on `https://finance.yahoo.com`. The data set (read through the R package `BatchGetSymbols`) contains daily closing prices of $p = 440$ stocks from the S&P500 index during the trading days between Augest 27th, 2007 to August 24th, 2009. Thus, there are $n = 503$ time points. The daily closing prices are transformed to log scales due to their multiplicative nature. We set the boundary removal $\underline{s} = \lfloor 0.05n \rfloor = 25$, bootstrap repeats

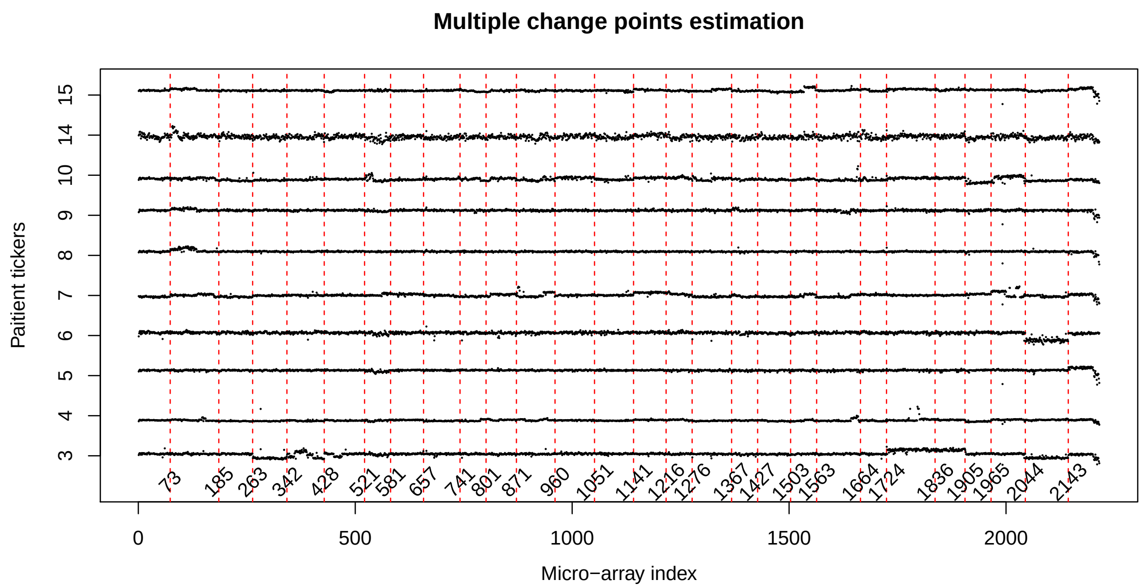


Figure 2.12: Real data study: aCGH data. Here, we set $B = 1000, \alpha = 0.05$ and the boundary removal is 60.

$B = 200$ and the block sizes $M = 1, 2, 5, 10$. Table 2.8 shows the (conditional) quantiles of the bootstrapped CUSUM test statistics for different block sizes. In particular, our CUSUM test statistic returns the value $T_n = 38.699$, while the 99%-quantile of block bootstrapped CUSUM test statistic for $M = 10$ is only 8.861. Therefore, the null hypothesis H_0 is rejected. In addition, $\hat{m}_{1/2} = 265$ indicates that the turning point was September 12th, 2008, the last trading day before Lehman Brothers Holdings Inc. declared bankruptcy on September 15th, 2008. Figure 2.13 plots the top 5 stocks with the largest change values on September 12th, 2008.

	$M = 1$	$M = 2$	$M = 5$	$M = 10$
$q_{0.90}$	2.350	3.380	5.219	6.981
$q_{0.95}$	2.470	3.860	5.464	7.327
$q_{0.99}$	2.782	4.580	5.746	8.861

Table 2.8: Quantiles of bootstrapped statistics.

A binary segmentation extension based on the block multiplier bootstrap test is considered as well. We implement this extension with $\hat{m}_{1/2}$ and $\hat{m}_0$ separately. The detected change points are shown in Figure 2.14(a) and 2.14(b), respectively. We set $\underline{s} = 25$, $B = 400$, $M = 5$ and $\alpha = 0.05$ for both scenarios. Overall, the two estimators share common time points on detection when mean-shift signals are large enough. There are seven overlapping change points identified by both algorithms, and the one with $\hat{m}_{1/2}$ additionally locates June 12th, 2008 and July 21st, 2008 but misses October 17th, 2008 compared to $\hat{m}_0$. That is, in

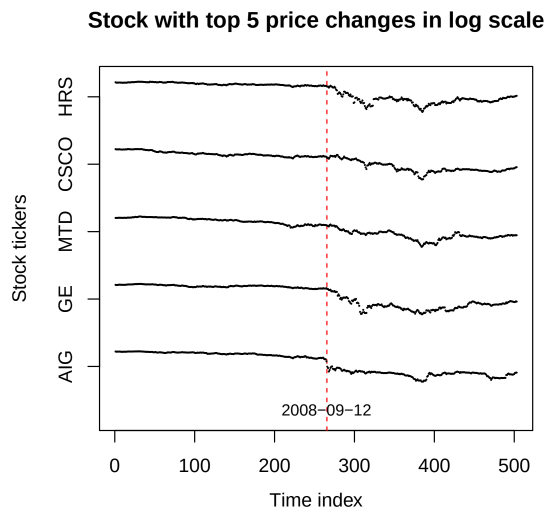
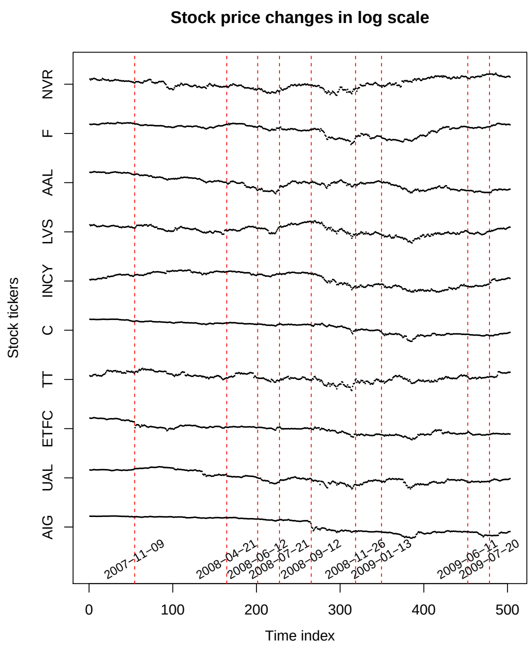


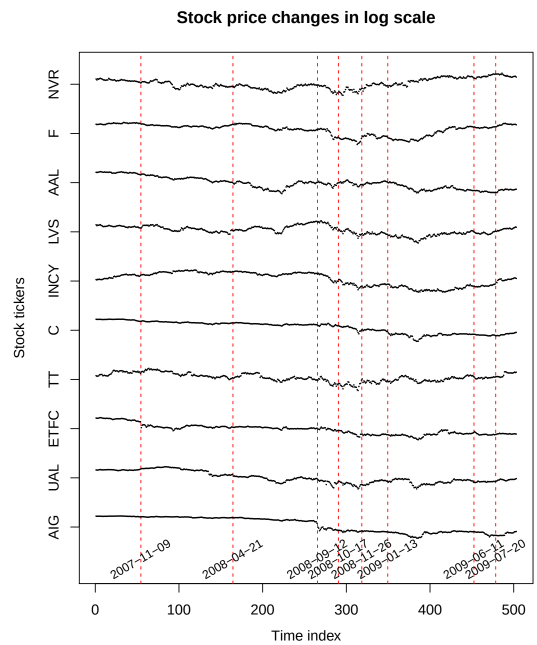
Figure 2.13: Selected stock trends.

the interval between April 21, 2008 and September 12th, 2008, $\hat{m}_{1/2}$ is sensitive to change points on the boundary points (i.e., June 12th, 2008 and July 21st, 2008). However, $\hat{m}_0$ is sensitive to the middle change point, namely October 17th, 2008, in the interval between September 12th, 2008 and November 26th, 2008. This exactly reflects our observation in the Simulation section.

We would like to make two notes based on this example. First, the estimator $\hat{m}_{1/2}$ is compatible with our test statistic T_n using stationary weight $\theta = 1/2$, while $\hat{m}_0$ is also a reasonable choice since prices are likely to be integrated. Therefore, we evaluate each of them in this stock-price data set. Second, neither of the two algorithms identifies change point between January 13th, 2009 and June 11st, 2009, though there seems to be fluctuations in means. One possible reason is that there exists non-synchronous change points (e.g. around time index at 380), which are not estimable. But this is a common issue in multiple change-point analysis and it is necessary to make some minimal separation or spacing assumptions (e.g. [45, 37, 36, 11]). In practice, such problem is worthy of further elaboration.



(a) Detected change points using $\hat{m}_{1/2}$ as location estimator.



(b) Detected change points using $\hat{m}_0$ as location estimator.

Figure 2.14: Finance data analyzed by extension of BABS using block CUSUM bootstrap test and two different location estimators.

2.6 Proofs and auxiliary numerical results

2.6.1 Proof of main results in Section 2.3

In this section, we prove the main theoretical results in Section 2.3. We first present a useful maximal inequality for weighted partial sums of independent and centered random vectors.

Lemma 2.8 (Talagrand's inequality for weighted partial sums of independent and centered random vectors)**.** *Let $X_1, \ldots, X_n$ be independent and centered random vectors in $\mathbb{R}^p$ and $\{a_{is}\}_{i,s=1}^n$ be an $n \times n$ matrix of real numbers. Define*

$$W_{n,sj} = \sum_{i=1}^n a_{is} X_{ij}, \quad Z_n = \max_{1 \leqslant s \leqslant n} \max_{1 \leqslant j \leqslant p} |W_{n,sj}|,$$

$$M = \max_{1 \leqslant s,i \leqslant n} \max_{1 \leqslant j \leqslant p} |a_{is} X_{ij}|, \quad \sigma^2 = \max_{1 \leqslant s \leqslant n} \max_{1 \leqslant j \leqslant p} \sum_{i=1}^n a_{is}^2 \mathbb{E}(X_{ij}^2).$$

(i) Let $\beta \in (0,1]$ and suppose that $\|X_{ij}\|_{\psi_\beta} < \infty$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Then, $\forall \eta \in (0,1]$, there exists a constant $C > 0$ depending only on β and η such that we have for $t > 0$

$$\mathbb{P}(Z_n \geqslant (1+\eta)\mathbb{E}[Z_n] + t) \leqslant \exp\left(-\frac{t^2}{3\sigma^2}\right) + 3\exp\left\{-\left(\frac{t}{C\|M\|_{\psi_\beta}}\right)^\beta\right\}. \tag{2.28}$$

(ii) Let $s \geqslant 1$ and suppose that $\mathbb{E}|X_{ij}|^s < \infty$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Then, $\forall \eta \in (0,1]$, there exists a constant $C > 0$ depending only on s and η such that we have for $t > 0$

$$\mathbb{P}(Z_n \geqslant (1+\eta)\mathbb{E}[Z_n] + t) \leqslant \exp\left(-\frac{t^2}{3\sigma^2}\right) + C\frac{\mathbb{E}[M^s]}{t^s}. \tag{2.29}$$

2.6.2 Proof of Theorem 2.1

Proof of Theorem 2.1. Suppose that H_0 is true. We may assume $\log^7(np) \leqslant \underline{s}$ for otherwise (2.11) and (2.12) trivially hold by choosing the constant $C > 0$ large enough therein. For $s = 1, \ldots, n-1$, let

$$a_{is} = \begin{cases} \sqrt{\frac{n-s}{ns}} & \text{if } 1 \leqslant i \leqslant s \\ -\sqrt{\frac{s}{n(n-s)}} & \text{if } s+1 \leqslant i \leqslant n \end{cases} \tag{2.30}$$

and $\mathbf{a}_s = (a_{1s}, \ldots, a_{ns})^\top$. Denote $\mathbf{X} = (X_1, \ldots, X_n)$ as the $p \times n$ data matrix and $A = (\mathbf{a}_{\underline{s}}, \ldots, \mathbf{a}_{n-\underline{s}})$. Then we can write

$$Z_n(s) = \sum_{i=1}^n a_{is} X_i = \mathbf{X}\mathbf{a}_s. \tag{2.31}$$

Since $\mathbb{E}[Z_n(s)] = 0$ under H_0, without loss of generality, we may assume $\mu_i \equiv 0$. Note that, for any $1 \leqslant s \leqslant s' \leqslant n-1$, we have

$$\mathrm{Cov}(Z_n(s), Z_n(s')) = \sum_{i=1}^{n} a_{is}a_{is'}\Sigma = \Sigma\sqrt{\frac{s(n-s')}{s'(n-s)}}.$$

Step 1: Gaussian approximation for CUSUM statistic. Let

$$\mathbf{Z}_n = (Z_n(\underline{s}), \ldots, Z_n(n-\underline{s})) = \mathbf{X}(\mathbf{a}_{\underline{s}}, \ldots, \mathbf{a}_{n-\underline{s}}) = \mathbf{X}A$$

be the CUSUM transformation of $\mathbf{X}$. Let $\mathrm{vec}(\mathbf{Z}_n)$ be the column stacked version of $\mathbf{Z}_n$, i.e. $\mathrm{vec}(\mathbf{Z}_n) = (Z_n(\underline{s})^\top, \ldots, Z_n(n-\underline{s})^\top)^\top$ is the $[(n-2\underline{s}+1)p] \times 1$ vector associated with $\mathbf{Z}_n$. Then we can write

$$\mathrm{vec}(\mathbf{Z}_n) = \mathrm{vec}(\mathbf{X}A) = (A^\top \otimes I_p)\mathrm{vec}(\mathbf{X}),$$

where $\otimes$ is the Kronecker product of two matrices. Since $\mathbb{E}[\mathrm{vec}(\mathbf{X})] = 0$ and $\mathrm{Cov}(\mathrm{vec}(\mathbf{X})) = \Gamma$, where Γ is the block diagonal matrix of size $(pn)\times(pn)$ with Σ being the diagonal sub-matrices, we have $\mathbb{E}[\mathrm{vec}(\mathbf{Z}_n)] = 0$ and $\mathrm{Cov}(\mathrm{vec}(\mathbf{Z}_n)) = (A^\top \otimes I_p)\Gamma(A \otimes I_p)$. Let

$$Y_n \sim N(0, (A^\top \otimes I_p)\Gamma(A \otimes I_p)). \tag{2.32}$$

be a joint mean-zero Gaussian random vector in $\mathbb{R}^{(n-2\underline{s}+1)p}$ with the same covariance matrix as $\mathrm{vec}(\mathbf{Z}_n)$. Denote $\bar{Y}_n = |Y_n|_\infty$. Since $\mathrm{Cov}(Z_n(s)) = \Sigma$ for all $s = 1, \ldots, n-1$, it follows from (A) that $\mathrm{Var}(Z_{nj}(s)) \geqslant \underline{b}$ for all $1 \leqslant j \leqslant p$ and $\underline{s} \leqslant s \leqslant n-\underline{s}$. Since $1 \leqslant \underline{s} \leqslant n/2$, we have

$$\begin{aligned}
\sum_{i=1}^{n} |a_{is}|^3 &= \sqrt{\frac{n}{s(n-s)}} - \frac{2}{n}\sqrt{\frac{s(n-s)}{n}} \leqslant \sqrt{\frac{n}{\underline{s}(n-\underline{s})}} \leqslant \sqrt{\frac{2}{\underline{s}}}, \\
\sum_{i=1}^{n} |a_{is}|^4 &= \frac{n}{s(n-s)} - \frac{3}{n} \leqslant \frac{n}{\underline{s}(n-\underline{s})} \leqslant \frac{2}{\underline{s}}.
\end{aligned}$$

Set $B_n = (2\bar{b}^2\underline{s}^{-1}n)^{1/2}$. By assumption (B), we have for $\ell = 1, 2$,

$$n^{-1}\sum_{i=1}^{n} |n^{1/2}a_{is}|^{2+\ell}\mathbb{E}|X_{ij}|^{2+\ell} \leqslant B_n^\ell.$$

Note that $\underline{s}^{1/2}|a_{is}| \leqslant 1$ for all $s = \underline{s}, \ldots, n-\underline{s}$.

Part (i). If (C) holds, then we have

$$\mathbb{E}\left[\exp\left(n^{1/2}|a_{is}||X_{ij}|/B_n\right)\right] \leqslant \mathbb{E}\left[\exp\left(\underline{s}^{1/2}|a_{is}||X_{ij}|/\bar{b}\right)\right] \leqslant 2.$$

By [35, Proposition 2.1], there exists a constant $C_1 > 0$ depending only on $\underline{b}$ and $\bar{b}$ such that

$$\rho(T_n, \bar{Y}_n) \leqslant C(\underline{b})\left(\frac{B_n^2\log^7(pn)}{n}\right)^{1/6} \leqslant C_1\varpi_{1,n}. \tag{2.33}$$

Part (ii). If (D) holds, then we have

$$\mathbb{E}\left\{\max_{1\leqslant j\leqslant p}\max_{\underline{s}\leqslant s\leqslant n-\underline{s}}(|n^{1/2}a_{is}||X_{ij}|/B_n)^q\right\} \leqslant \left[\underline{s}^{1/2}(\max_{\underline{s}\leqslant s\leqslant n-\underline{s}}|a_{is}|)\right]^q \mathbb{E}\left[\max_{1\leqslant j\leqslant p}(|X_{ij}|/\bar{b})^q\right] \leqslant 1$$

for all $i = 1, \ldots, n$. By [35, Proposition 2.1], there exists a constant $C_1 > 0$ depending only on $\underline{b}, \bar{b}, q$ such that

$$\begin{aligned}\rho(T_n, \bar{Y}_n) &\leqslant C(\underline{b}, q)\left\{\left(\frac{B_n^2\log^7(pn)}{n}\right)^{1/6} + \left(\frac{B_n^2\log^3(pn)}{n^{1-2/q}}\right)^{1/3}\right\} \\ &\leqslant C_1\{\varpi_{1,n} + \varpi_{2,n}\}.\end{aligned} \tag{2.34}$$

Step 2: Gaussian comparison for $\bar{Y}_n$ and bootstrap CUSUM statistic T_n^*. Let

$$\begin{aligned}\hat{S}_{n,s}^- &= \frac{1}{s}\sum_{i=1}^{s}(X_i - \bar{X}_s^-)(X_i - \bar{X}_s^-)^\top, \\ \hat{S}_{n,s}^+ &= \frac{1}{n-s}\sum_{i=s+1}^{n}(X_i - \bar{X}_s^+)(X_i - \bar{X}_s^+)^\top,\end{aligned} \tag{2.35}$$

be the sample covariance matrices based on the left and right observations at s, respectively. Then

$$Z_n^*(s)|X_1^n \sim N\left(0, \frac{n-s}{n}\hat{S}_{n,s}^- + \frac{s}{n}\hat{S}_{n,s}^+\right).$$

Let

$$\mathbf{a}_{is}^* = \begin{cases} \sqrt{\frac{n-s}{ns}}(X_i - \bar{X}_s^-) & \text{if } 1 \leqslant i \leqslant s \\ -\sqrt{\frac{s}{n(n-s)}}(X_i - \bar{X}_s^+) & \text{if } s+1 \leqslant i \leqslant n \end{cases}$$

and $A_s^* = (\mathbf{a}_{1s}^*, \ldots, \mathbf{a}_{ns}^*)$. Let $\mathbf{e} = (e_1, \ldots, e_n)^\top$. Then we can write

$$Z_n^*(s) = \sum_{i=1}^{n}\mathbf{a}_{is}^* e_i = A_s^*\mathbf{e}.$$

Let

$$Z_n^* = \begin{pmatrix} Z_n^*(\underline{s}) \\ \vdots \\ Z_n^*(n-\underline{s}) \end{pmatrix} = \begin{pmatrix} A_{\underline{s}}^*\mathbf{e} \\ \vdots \\ A_{n-\underline{s}}^*\mathbf{e} \end{pmatrix} = \begin{pmatrix} A_{\underline{s}}^* \\ \vdots \\ A_{n-\underline{s}}^* \end{pmatrix} \mathbf{e} := \boldsymbol{A}^*\mathbf{e}.$$

Since $\mathbf{e} \sim N(0, \mathrm{Id}_n)$, it follows that $Z_n^*|X_1^n \sim N(0, \boldsymbol{A}^*\boldsymbol{A}^{*\top})$. Next, we compute an explicit expression for the covariance matrix of Z_n^* given X_1^n. Some routine algebra show that for any $\underline{s} \leqslant s \leqslant s' \leqslant n - \underline{s}$,

$$\begin{aligned}\mathrm{Cov}(Z_n^*(s), Z_n^*(s')|X_1^n) = \Sigma\sqrt{\frac{s(n-s')}{s'(n-s)}} + \frac{R_1}{n}\sqrt{\frac{(n-s)(n-s')}{ss'}} + \frac{R_2}{n}\sqrt{\frac{s(n-s')}{s'(n-s)}} \\ + \frac{R_3}{n}\sqrt{\frac{s(n-s')}{s'(n-s)}} + \frac{R_4}{n}\sqrt{\frac{ss'}{(n-s)(n-s')}} - \frac{R_5}{n}\sqrt{\frac{s(n-s')}{s'(n-s)}},\end{aligned}$$

where $R_1 = \sum_{i=1}^s [(X_i - \bar{X}_s^-)(X_i - \bar{X}_{s'}^-)^\top - \Sigma]$, $R_2 = \sum_{i=1}^s [(X_i - \bar{X}_s^+)(X_i - \bar{X}_{s'}^-)^\top - \Sigma]$, $R_3 = \sum_{i=s'+1}^n [(X_i - \bar{X}_s^+)(X_i - \bar{X}_{s'}^-)^\top - \Sigma]$, $R_4 = \sum_{i=s'+1}^n [(X_i - \bar{X}_s^+)(X_i - \bar{X}_{s'}^+)^\top - \Sigma]$, and $R_5 = \sum_{i=1}^n [(X_i - \bar{X}_s^+)(X_i - \bar{X}_{s'}^-)^\top - \Sigma]$. Let

$$\begin{aligned}\hat{\Delta}_1 &= \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \left| \frac{1}{s}\sum_{i=1}^s (X_iX_i^\top - \Sigma)\right|_\infty, \quad \hat{\Delta}_3 = \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} |\bar{X}_s^-|_\infty, \\ \hat{\Delta}_2 &= \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \left| \frac{1}{n-s}\sum_{i=s+1}^n (X_iX_i^\top - \Sigma)\right|_\infty, \quad \hat{\Delta}_4 = \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} |\bar{X}_s^+|_\infty.\end{aligned} \tag{2.36}$$

Then there exists a universal constant $K_1 > 0$ such that

$$|\mathrm{Cov}(Z_n^*|X_1^n) - \mathrm{Cov}(Y_n)|_\infty \leqslant K_1\hat{\Delta},$$

where

$$\hat{\Delta} = \max\{\hat{\Delta}_1, \hat{\Delta}_2\} + \max\{\hat{\Delta}_3^2, \hat{\Delta}_4^2\}. \tag{2.37}$$

Let $\bar{\Delta}$ be a positive real number and $E = \{\hat{\Delta} \leqslant \bar{\Delta}\}$. By [27, Lemma C.1], there exists a constant $C_2 > 0$ depending only on $\underline{b}$ such that on the event E, we have

$$\rho^*(\bar{Y}_n, \bar{Z}_n^*) \leqslant C_2\bar{\Delta}^{1/3}\log^{2/3}(np).$$

Part (i). If (C) holds, then we choose

$$\bar{\Delta} = C_3\underline{s}^{-1/2}\log^{3/2}(np) \tag{2.38}$$

for some large enough constant $C_3 := C_3(\underline{b}, \bar{b}, K) > 0$. By Lemma 2.9, we have $\mathbb{P}(E) \geqslant 1 - \gamma$. Then there exists a constant $C_4 := C_4(\underline{b}, \bar{b}, K) > 0$ such that

$$\rho^*(\bar{Y}_n, \bar{Z}_n^*) \leqslant C_4 \varpi_{1,n} \tag{2.39}$$

holds with probability at least $1-\gamma$. Combining (2.33) and (2.39), we obtain (2.11).

<u>*Part (ii)*</u>. If (D) holds, then we choose

$$\bar{\Delta} = C_5\{\underline{s}^{-1/2} \log^{1/2}(np) + \gamma^{-2/q} \underline{s}^{-1} n^{2/q} \log(np)\} \tag{2.40}$$

for some large enough constant $C_5 := C_5(\underline{b}, \bar{b}, K, q) > 0$. By Lemma 2.9, we have $\mathbb{P}(E) \geqslant 1-\gamma$. Then there exists a constant $C_6 := C_6(\underline{b}, \bar{b}, K, q) > 0$ such that

$$\rho^*(\bar{Y}_n, \bar{Z}_n^*) \leqslant C_6\{\varpi_{1,n} + \varpi_{2,n}\}. \tag{2.41}$$

holds with probability at least $1-\gamma$. Combining (2.34) and (2.41), we obtain (2.12). □

Lemma 2.9 (Bound on $\max_{1\leqslant i\leqslant 4} \hat{\Delta}_i$)**.** *Suppose H_0 is true and assume (A) and (B) hold. Let $\gamma \in (0, e^{-1})$ and suppose that $\log(\gamma^{-1}) \leqslant K \log(pn)$ for some constant $K > 0$. Let $\hat{\Delta}_i, i = 1, \ldots, 4$ be defined in (2.36).*
(i) If (C) holds and $\log^5(np) \leqslant \underline{s}$, then there exists a constant $C > 0$ depending only on $\bar{b}, K$ such that

$$\mathbb{P}(\max_{1\leqslant i\leqslant 4} \hat{\Delta}_i \leqslant C\underline{s}^{-1/2} \log^{3/2}(np)) \geqslant 1-\gamma.$$

(ii) If (D) holds with $q \geqslant 4$, then there exists a constant $C > 0$ depending only on $\bar{b}, K, q$ such that

$$\mathbb{P}(\max_{1\leqslant i\leqslant 4} \hat{\Delta}_i \leqslant C\{\underline{s}^{-1/2} \log^{3/2}(np) + \gamma^{-2/q} \underline{s}^{-1} n^{2/q} \log(np)\}) \geqslant 1-\gamma.$$

Proof of Corollary 2.2. Under H_0, we write $\mathbb{P} = \mathbb{P}_0$. Let Y_n be a joint Gaussian random vector defined in (2.32) and $\bar{Y}_n = |Y_n|_\infty$. Let $\rho_\ominus(\alpha) = \mathbb{P}(\{T_n \leqslant q_{T_n^*|X_1^n}(\alpha)\} \ominus \{T_n \leqslant q_{\bar{Y}_n}(\alpha)\})$ and $A \ominus B = (A \setminus B) \cup (B \setminus A)$ be the symmetric difference of two events A and B. Note that

$$\begin{aligned} |\mathbb{P}(T_n \leqslant q_{T_n^*|X_1^n}(\alpha)) - \alpha| &\leqslant |\mathbb{P}(T_n \leqslant q_{T_n^*|X_1^n}(\alpha)) - \mathbb{P}(T_n \leqslant q_{\bar{Y}_n}(\alpha))| + \rho(T_n, \bar{Y}_n) \\ &\leqslant \mathbb{P}(\{T_n \leqslant q_{T_n^*|X_1^n}(\alpha)\} \ominus \{T_n \leqslant q_{\bar{Y}_n}(\alpha)\}) + \rho(T_n, \bar{Y}_n) \\ &= \rho_\ominus(\alpha) + \rho(T_n, \bar{Y}_n). \end{aligned}$$

By [27, Lemma C.3], there exists a constant $C > 0$ only depending on $\underline{b}$ such that for any real number $\bar{\Delta} > 0$, we have

$$\rho_{\ominus}(\alpha) \leqslant 2[\rho(T_n, \bar{Y}_n) + C\bar{\Delta}^{1/3}\log^{2/3}(np) + \mathbb{P}(\hat{\Delta} > \bar{\Delta})],$$

where $\hat{\Delta}$ is defined in (2.37).

<u>*Part (i)*</u>. Assume (C) and choose $\bar{\Delta}$ in (2.38). By Lemma 2.9, we have $\mathbb{P}(\hat{\Delta} > \bar{\Delta}) < \gamma/2$. Combining with (2.33), we get (2.13). Uniform convergence of $\mathbb{P}(T_n \leqslant q_{T_n^*|X_1^n}(\alpha))$ to α follows by choosing $\gamma = n^{-1}$.

<u>*Part (ii)*</u>. Assume (D) and choose $\bar{\Delta}$ in (2.40). By Lemma 2.9, we have $\mathbb{P}(\hat{\Delta} > \bar{\Delta}) < \gamma/2$. Combining with (2.34), we get (2.14). Uniform convergence of $\mathbb{P}(T_n \leqslant q_{T_n^*|X_1^n}(\alpha))$ to α follows by choosing $\gamma = [\log(np)]^{-q\varepsilon/2}$. □

2.6.3 Proof of Theorem 2.3

Proof of Theorem 2.3. Under H_1, without loss of generality, we may assume $\mu = 0$. Then

$$\xi_i = \begin{cases} X_i, & \text{if } 1 \leqslant i \leqslant m \\ X_i - \delta_n, & \text{if } m+1 \leqslant i \leqslant n \end{cases}. \tag{2.42}$$

Observe that the CUSUM statistic (computed on $X_1, \ldots, X_n$) in (2.4) can be written as

$$Z_n(s) = Z_n^{\xi}(s) + \Delta_s,$$

where

$$Z_n^{\xi}(s) = \sqrt{\frac{s(n-s)}{n}} \left\{ \frac{1}{s}\sum_{i=1}^{s} \xi_i - \frac{1}{n-s}\sum_{i=s+1}^{n} \xi_i \right\}$$

and

$$\Delta_s = \begin{cases} -\sqrt{\frac{s}{n(n-s)}}(n-m)\delta_n, & \text{if } 1 \leqslant s \leqslant m \\ -\sqrt{\frac{n-s}{ns}}m\delta_n, & \text{if } m+1 \leqslant s \leqslant n \end{cases} \tag{2.43}$$

is the mean shift. Note that $|\Delta_s|_\infty$ reaches its maximum at $s = m$, i.e.,

$$\max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |\Delta_s|_\infty = \max_{1 \leqslant s \leqslant n} |\Delta_s|_\infty = |\Delta_m|_\infty = \sqrt{\frac{m(n-m)}{n}}|\delta_n|_\infty := \tilde{\Delta}.$$

Let $\tilde{T}_n = \max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |Z_n^{\xi}(s)|_\infty$. Then we have

$$T_n = \max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |Z_n(s)|_\infty = \max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |Z_n^{\xi}(s) + \Delta_s|_\infty \geqslant \tilde{\Delta} - \tilde{T}_n,$$

from which it follows that the type II error of our bootstrap test obeys

$$\text{Type II error} = \mathbb{P}_1(T_n \leqslant q_{T_n^*|X_1^n}(1-\alpha)) \leqslant \mathbb{P}_1(\tilde{T}_n \geqslant \tilde{\Delta} - q_{T_n^*|X_1^n}(1-\alpha)). \tag{2.44}$$

Let $\beta_n \in (0,1)$ and $\hat{\Delta} := \hat{\Delta}(X_1^n) = q_{T_n^*|X_1^n}(1-\alpha) + q_{\tilde{T}_n}(1-\beta_n)$. Clearly, $\hat{\Delta}$ is a random quantity that is $\sigma(X_1,\ldots,X_n)$-measurable. Then,

$$\begin{aligned}\mathbb{P}_1(\tilde{T}_n \geqslant \tilde{\Delta} - q_{T_n^*|X_1^n}(1-\alpha)) &\leqslant \mathbb{P}_1(\tilde{T}_n \geqslant \hat{\Delta} - q_{T_n^*|X_1^n}(1-\alpha)) + \mathbb{P}_1(\hat{\Delta} > \tilde{\Delta}) \\ &\leqslant \beta_n + \mathbb{P}_1(\hat{\Delta} > \tilde{\Delta}). \end{aligned} \tag{2.45}$$

Observe that the distribution of $\tilde{T}_n$ does not depend on δ_n. Hence, $\tilde{T}_n$ has the same distributions as T_n under H_0.

Part (i). Assume (C). Let $Y_n \sim N(0, (A^\top \otimes I_p)\Gamma(A \otimes I_p))$ be a joint mean-zero Gaussian random vector in $\mathbb{R}^{(n-2\underline{s}+1)p}$, where A is defined in (2.30). Denote $\bar{Y}_n = |Y_n|_\infty$. By the Gaussian approximation (2.33), there exists a constant $C_1 := C_1(\underline{b}, \bar{b}, K) > 0$ such that

$$\mathbb{P}(\tilde{T}_n > t) \leqslant \mathbb{P}(\bar{Y}_n > t) + C_1 \varpi_{1,n}$$

holds for all $t \in \mathbb{R}$. By [92, Lemma 2.2.2], $\|\bar{Y}_n\|_{\psi_2} \leqslant C_2 \log^{1/2}(np)$, where $C_2 > 0$ is a constant depending only on $\bar{b}$. So we have $\forall t > 0$,

$$\mathbb{P}(\bar{Y}_n > t) \leqslant 2\exp[-(t/\|\bar{Y}_n\|_{\psi_2})^2] \leqslant 2\exp[-C_2^{-2}\log^{-1}(np)t^2].$$

Choosing $t = C_3[\log(\zeta^{-1})\log(np)]^{1/2}$ for some large enough constant $C_3 > 0$, we get $\mathbb{P}(\bar{Y}_n > t) \leqslant 2\zeta^{C_3^2/C_2^2} \leqslant 2\zeta$. Now, take $\beta_n = C_1\varpi_{1,n} + 2\zeta$. Since $q_{\tilde{T}_n}(1-\beta_n) = \inf\{t \in \mathbb{R} : \mathbb{P}(\tilde{T}_n > t) < \beta_n\}$, we deduce that

$$q_{\tilde{T}_n}(1-\beta_n) \leqslant C_3[\log(\zeta^{-1})\log(np)]^{1/2}. \tag{2.46}$$

Next, we deal with $q_{T_n^*|X_1^n}(1-\alpha)$. Recall that $\hat{S}_{n,s}^-$ and $\hat{S}_{n,s}^+$ are defined in (2.35). By the Bonferroni inequality, we have

$$\mathbb{P}_1(\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} |Z_n^*(s)|_\infty > t | X_1^n) \leqslant 2np[1-\Phi(t/\bar{\psi})],$$

where

$$\bar{\psi}^2 = \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \left\{ \frac{n-s}{n}\hat{S}_{n,s,jj}^- + \frac{s}{n}\hat{S}_{n,s,jj}^+ \right\},$$

and $\hat{S}^-_{n,s,jj}$, $\hat{S}^+_{n,s,jj}$ are the (j,j)-th diagonal entry of $\hat{S}^-_{n,s}$, $\hat{S}^+_{n,s}$ respectively. Then it follows that there exists a universal constant $K_1 > 0$ such that

$$q_{T_n^*|X_1^n}(1-\alpha) \leqslant \bar{\psi}\Phi^{-1}(1-\alpha/(2np)).$$

Next, we bound the quantiles $t_{n,\alpha} := \Phi^{-1}(1-\alpha/(2np))$. Recall that $n \geqslant 4$, $p \geqslant 3$, and $\alpha \in (0,1)$. Since $\Phi^{-1}(\cdot)$ is a strictly increasing function, we have $t_{n,\alpha} \geqslant \Phi^{-1}(23/24) > 1.73$. By the standard Gaussian tail bound $1-\Phi(x) < \phi(x)/x$ for all $x > 0$, we deduce that

$$\frac{\alpha}{2np} = 1 - \Phi(t_{n,\alpha}) < \frac{\phi(t_{n,\alpha})}{t_{n,\alpha}} < 0.25 \exp\left(-\frac{t_{n,\alpha}^2}{2}\right).$$

Therefore, $t_{n,\alpha} < \sqrt{2\log(np/(2\alpha))}$ and

$$q_{T_n^*|X_1^n}(1-\alpha) \leqslant K_1 \bar{\psi} \log^{1/2}(np/\alpha) \tag{2.47}$$

for some universal constant $K_1 > 0$. Define

$$\hat{S}^{\xi,-}_{n,s} = \frac{1}{s}\sum_{i=1}^{s}(\xi_i - \bar{\xi}^-_s)(\xi_i - \bar{\xi}^-_s)^\top,$$
$$\hat{S}^{\xi,+}_{n,s} = \frac{1}{n-s}\sum_{i=s+1}^{n}(\xi_i - \bar{\xi}^+_s)(\xi_i - \bar{\xi}^+_s)^\top,$$

where $\bar{\xi}^-_s = s^{-1}\sum_{i=1}^s \xi_i$ and $\bar{\xi}^+_s = (n-s)^{-1}\sum_{i=s+1}^n \xi_i$. By Lemma 2.10, we have

$$\bar{\psi}^2 \leqslant 2 \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \left\{\frac{n-s}{n}\hat{S}^{\xi,-}_{n,s,jj} + \frac{s}{n}\hat{S}^{\xi,+}_{n,s,jj}\right\} + 4|\delta_n|_\infty^2.$$

Since

$$\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \frac{n-s}{n}\hat{S}^{\xi,-}_{n,s,jj} \leqslant \max_{1\leqslant j\leqslant p}\Sigma_{jj} + \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p}\left|\frac{1}{s}\sum_{i=1}^{s}(\xi_{ij}^2 - \Sigma_{jj})\right| + \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p}\left|\bar{\xi}^-_{sj}\right|^2,$$

it follows from Lemma 2.9 that there exists a constant $C_4 > 0$ depending only on $\bar{b}, K$ such that

$$\mathbb{P}\left(\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s}\frac{1}{s}(\xi_{ij}^2 - \Sigma_{jj})\right| \geqslant C_4 \underline{s}^{-1/2}\log^{3/2}(np)\right) \leqslant \gamma/4$$

and the same probability bound holds for $\max_{\underline{s}\leqslant s\leqslant n-\underline{s}}\max_{1\leqslant j\leqslant p}|\bar{\xi}^-_{sj}|^2$. Combining with (2.47), we deduce that there exists a constant $C_5 > 0$ depending only on $\bar{b}, K$ such that

$$\mathbb{P}_1\left(q_{T_n^*|X_1^n}(1-\alpha) \geqslant C_5 \max\{|\delta_n|_\infty, 1\} \log^{1/2}(np/\alpha)\right) \leqslant \gamma.$$

Then (2.16) follows from the last inequality together with (2.15), (2.45), and (2.46).

Part (ii). Assume (D). By the Gaussian approximation (2.34), there exists a constant $C_1 := C_1(\underline{b}, \bar{b}, K, q) > 0$ such that

$$\mathbb{P}(\tilde{T}_n > t) \leqslant \mathbb{P}(\bar{Y}_n > t) + C_1\{\varpi_{1,n} + \varpi_{2,n}\}$$

holds for all $t \in \mathbb{R}$. By the same argument as in Part (i), we have (2.47) and (2.46) hold with $\beta_n = C_1\{\varpi_{1,n} + \varpi_{2,n}\} + 2\zeta$. By Lemma 2.9, there exists a constant $C_2 > 0$ depending only on $\bar{b}, K, q$ such that

$$\mathbb{P}\left(\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \left|\sum_{i=1}^{s} \frac{1}{s}(\xi_{ij}^2 - \Sigma_{jj})\right| \geqslant C_2\{\underline{s}^{-1/2} \log^{1/2}(np) + \gamma^{-2/q} \underline{s}^{-1} n^{2/q} \log(np)\}\right) \leqslant \gamma/4$$

and the same probability bound holds for $\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} |\bar{\xi}_{sj}^{-}|^2$. Then the rest of the proof follows similar lines as in Part (i). ☐

Lemma 2.10 (Bound on $\bar{\psi}$)**.** *Assume that $X_1, \ldots, X_n$ are independent random vectors that are generated from the model (2.1). Then we have*

$$\bar{\psi}^2 \leqslant 2 \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \left\{\frac{n-s}{n} \hat{S}_{n,s,jj}^{\xi,-} + \frac{s}{n} \hat{S}_{n,s,jj}^{\xi,+}\right\} + 4|\delta_n|_\infty^2. \tag{2.48}$$

2.6.4 Proof of Theorem 2.4

Proof of Theorem 2.4. In this proof, we use $K_1, K_2, \ldots$ to denote universal constants. Note that $\mathbb{E}[Z_n(s)] = \Delta_s$, where Δ_s is defined in (2.43). Therefore, $|\mathbb{E}[Z_n(\cdot)]|_\infty$ reaches its maximum at m and we have

$$|\mathbb{E}[Z_n(m)]|_\infty - |\mathbb{E}[Z_n(s)]|_\infty = \begin{cases} \sqrt{\frac{m(n-m)}{n}}\left(1 - \sqrt{\frac{s(n-m)}{m(n-s)}}\right)|\delta_n|_\infty, & \text{if } 1 \leqslant s \leqslant m \\ \sqrt{\frac{m(n-m)}{n}}\left(1 - \sqrt{\frac{m(n-s)}{s(n-m)}}\right)|\delta_n|_\infty, & \text{if } m < s \leqslant n \end{cases}.$$

If $1 \leqslant s \leqslant m$, then

$$\begin{aligned}
1-\sqrt{\frac{s(n-m)}{m(n-s)}} &= \frac{(\sqrt{m(n-s)}-\sqrt{s(n-m)})(\sqrt{m(n-s)}+\sqrt{s(n-m)})}{\sqrt{m(n-s)}(\sqrt{m(n-s)}+\sqrt{s(n-m)})} \\
&= \frac{n(m-s)}{m(n-s)+\sqrt{m(n-m)s(n-s)}} \\
&\geqslant \frac{n(m-s)}{mn+\frac{n}{2}\sqrt{m(n-m)}} \\
&\geqslant \frac{m-s}{\sqrt{m}(\sqrt{m}+\sqrt{n-m})}.
\end{aligned}$$

So we get

$$\begin{aligned}
\sqrt{\frac{m(n-m)}{n}}\Big(1-\sqrt{\frac{s(n-m)}{m(n-s)}}\Big)|\delta_n|_\infty &\geqslant \sqrt{\frac{m(n-m)}{n}}\frac{m-s}{\sqrt{m}(\sqrt{m}+\sqrt{n-m})}|\delta_n|_\infty \\
&= \sqrt{n}\frac{\sqrt{1-t_m}}{\sqrt{t_m}+\sqrt{1-t_m}}(t_m-t_s)|\delta_n|_\infty.
\end{aligned}$$

Likewise, if $m < s \leqslant n$, then

$$1-\sqrt{\frac{m(n-s)}{s(n-m)}} \geqslant \frac{s-m}{\sqrt{n-m}(\sqrt{m}+\sqrt{n-m})}$$

and

$$\sqrt{\frac{m(n-m)}{n}}\Big(1-\sqrt{\frac{m(n-s)}{s(n-m)}}\Big)|\delta_n|_\infty \geqslant \sqrt{n}\frac{\sqrt{t_m}}{\sqrt{t_m}+\sqrt{1-t_m}}(t_s-t_m)|\delta_n|_\infty.$$

Hence, for any $1 \leqslant s \leqslant n$, we have

$$|\mathbb{E}[Z_n(m)]|_\infty - |\mathbb{E}[Z_n(s)]|_\infty \geqslant \sqrt{n}\,\frac{\sqrt{t_m}\wedge\sqrt{1-t_m}}{\sqrt{t_m}+\sqrt{1-t_m}}|t_m-t_s||\delta_n|_\infty. \tag{2.49}$$

By the triangle inequality,

$$\begin{aligned}
|Z_n(s)|_\infty - |Z_n(m)|_\infty &\leqslant |Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty + |\mathbb{E}[Z_n(s)]|_\infty \\
&\quad + |Z_n(m)-\mathbb{E}[Z_n(m)]|_\infty - |\mathbb{E}[Z_n(m)]|_\infty \\
&\leqslant 2\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}Z_n(s)|_\infty + |\mathbb{E}[Z_n(s)]|_\infty - |\mathbb{E}[Z_n(m)]|_\infty.
\end{aligned}$$

Combining the last inequality with (2.49) and using $\sqrt{t_m}+\sqrt{1-t_m} \leqslant \sqrt{2}$, we get

$$\sqrt{\frac{n}{2}}(\sqrt{t_m}\wedge\sqrt{1-t_m})|t_m-t_s||\delta_n|_\infty \leqslant 2\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty + |Z_n(m)|_\infty - |Z_n(s)|_\infty.$$

Replacing s by $\hat{m}_{1/2}$ and noticing that $|Z_n(m)|_\infty \leqslant |Z_n(\hat{m}_{1/2})|_\infty$, we obtain that

$$
\begin{aligned}
|t_m - t_{\hat{m}_{1/2}}| \leqslant & \frac{2\sqrt{2}\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty}{\sqrt{n}(\sqrt{t_m}\wedge\sqrt{1-t_m})|\delta_n|_\infty} \\
\leqslant & \frac{2\sqrt{2}\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty}{\sqrt{nt_m(1-t_m)}|\delta_n|_\infty},
\end{aligned} \tag{2.50}
$$

where the last step follows from the inequality $t(1-t) \leqslant t\wedge(1-t) \leqslant 2t(1-t)$ for all $t \in [0,1]$. Next, we bound $\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty$. Recall that $Z_n(s)-\mathbb{E}[Z_n(s)] = \sum_{i=1}^n a_{is}(X_i-\mu_i)$, where a_{is} is defined in (2.30).

Part (i). Assume (C). By [1, Theorem 4], we have $\forall t>0$,

$$
\begin{aligned}
& \mathbb{P}(\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty \geqslant 2\mathbb{E}[\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty]+t) \\
& \leqslant \exp\left(-\frac{t^2}{3\tau^2}\right)+3\exp\left(-\frac{t}{K_1\|M\|_{\psi_1}}\right),
\end{aligned} \tag{2.51}
$$

where

$$
\begin{aligned}
\tau^2 &= \max_{1\leqslant s\leqslant n}\max_{1\leqslant j\leqslant p}\sum_{i=1}^n a_{is}^2\mathbb{E}(X_{ij}-\mu_{ij})^2, \\
M &= \max_{1\leqslant i,s\leqslant n}\max_{1\leqslant j\leqslant p}|a_{is}(X_{ij}-\mu_{ij})|.
\end{aligned}
$$

Since $\sum_{i=1}^n a_{is}^2 = 1$, we have $\tau^2 \leqslant \bar{b}$. Since $\max_{1\leqslant i,s\leqslant n}|a_{is}| \leqslant 1$, by [92, Lemma 2.2.2], we have $\|M\|_2 \leqslant 2\|M\|_{\psi_1} \leqslant 2\bar{b}\log(np)$. By [35, Lemma E.1], we have

$$
\mathbb{E}[\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty] \leqslant K_2\{\tau\log^{1/2}(np)+\|M\|_2\log(np)\}. \tag{2.52}
$$

Choosing $t = K_3\bar{b}\log(np)\log(\gamma^{-1})$ in (2.51) for some large enough universal constant $K_3>0$, we deduce that there exists a constant $C := C(\bar{b},K)>0$ such that

$$
\mathbb{P}(\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty \geqslant C\log^2(np)) \leqslant 2\gamma.
$$

Combining the last inequality with (2.50), we obtain (2.19).

Part (ii). Assume (D) with $q \geqslant 2$. By [2, Theorem 2], we have $\forall t > 0$,

$$
\begin{aligned}
&\mathbb{P}(\max_{1\leqslant s\leqslant n} |Z_n(s) - \mathbb{E}[Z_n(s)]|_\infty \geqslant 2\mathbb{E}[\max_{1\leqslant s\leqslant n} |Z_n(s) - \mathbb{E}[Z_n(s)]|_\infty] + t) \\
&\leqslant \exp\left(-\frac{t^2}{3\tau^2}\right) + C(q)\frac{\mathbb{E}[M^q]}{t^q},
\end{aligned} \tag{2.53}
$$

where τ^2 and M have the same definitions as in Part (i). By [92, Lemma 2.2.2], we have $\|M\|_2 \leqslant \|M\|_q \leqslant n^{1/q}\bar{b}$. Note that $\tau^2 \leqslant \bar{b}$ and $\mathbb{E}[\max_{1\leqslant s\leqslant n} |Z_n(s) - \mathbb{E}[Z_n(s)]|_\infty]$ obeys the bound in (2.52). Hence, choosing $t = C(q)\{\bar{b}^{1/2}\log(\gamma^{-1}) + n^{1/q}\bar{b}^{1/q}\gamma^{-1/q}\}$ in (2.53), we get

$$
\mathbb{P}\left(\max_{1\leqslant s\leqslant n} |Z_n(s) - \mathbb{E}[Z_n(s)]|_\infty \geqslant C(q)n^{1/q}(\log(np) + \gamma^{-1/q})\right) \leqslant 2\log^{-q}(np).
$$

Combining the last inequality with (2.50), we obtain (2.20). □

2.6.5 Proof of Theorem 2.5

Proof of Theorem 2.5. Without loss of generality, we may assume $\delta_{nj} \leqslant 0$ for all $1 \leqslant j \leqslant p$. In addition, we may assume that

$$
\underline{\delta}_n \gg \frac{\log^2(np)}{n^{1/2}} \tag{2.54}
$$

in Part (i), and

$$
\underline{\delta}_n \gg \frac{\log^{1/2}(np)}{n^{1/2}} \vee \frac{\log(np)}{\gamma^{1/q}n^{1/2-1/q}} \tag{2.55}
$$

in Part (ii), because otherwise (2.21) and (2.22) trivially hold by choosing the constant $C > 0$ large enough. Denote $h(t_m) = t_m \wedge (1 - t_m)$. To simplify the notation, we write

$$
\tilde{Z}_n(s) := Z_{0,n}(s) = \sum_{i=1}^{s} X_i - \frac{s}{n}\sum_{i=1}^{n} X_i = \sqrt{\frac{s(n-s)}{n}}Z_n(s),
$$

where $Z_n(s)$ is defined in (2.4), and $\tilde{m} = \hat{m}_0$. Let j^* be an index in $\{1, \ldots, p\}$ such that $\max_{1\leqslant j\leqslant p} \tilde{Z}_{nj}(m) = \tilde{Z}_{nj^*}(m)$. It is clear that j^* is a random variable depending on m. By Lemma 2.11, $\tilde{Z}_{nj^*}(m) = \max_{1\leqslant j\leqslant p} |\tilde{Z}_{nj}(m)| \geqslant$

0 holds with probability greater than $1-\gamma/36$. For $r \geqslant 1$, observe that

$$\begin{aligned}
&\{|t_{\tilde{m}} - t_m| > r/n\} \\
\subset &\{\tilde{m} - m > r\} \bigcup \{\tilde{m} - m < -(r-1)\} \\
\subset &\{\max_{s \leqslant m+r} |\tilde{Z}_n(s)|_\infty < |\tilde{Z}_n(\tilde{m})|_\infty\} \bigcup \{\max_{s \leqslant m-(r-1)} |\tilde{Z}_n(s)|_\infty = |\tilde{Z}_n(\tilde{m})|_\infty\} \\
\subset &\{\max_{s \geqslant m+r} |\tilde{Z}_n(s)|_\infty \geqslant |\tilde{Z}_n(m)|_\infty\} \bigcup \{\max_{s \leqslant m-(r-1)} |\tilde{Z}_n(s)|_\infty \geqslant |\tilde{Z}_n(m)|_\infty\}.
\end{aligned}$$

Thus we have $\mathbb{P}(|t_{\tilde{m}} - t_m| > r/n) \leqslant I + II$, where

$$\begin{aligned}
I =& \mathbb{P}\left(\max_{s \geqslant m+r} |\tilde{Z}_n(s)|_\infty \geqslant |\tilde{Z}_n(m)|_\infty\right) \leqslant \mathbb{P}\left(\max_{s \geqslant m+r} |\tilde{Z}_n(s)|_\infty \geqslant \tilde{Z}_{nj^*}(m)\right), \\
II =& \mathbb{P}\left(\max_{s \leqslant m-(r-1)} |\tilde{Z}_n(s)|_\infty \geqslant |\tilde{Z}_n(m)|_\infty\right) \leqslant \mathbb{P}\left(\max_{s \leqslant m-(r-1)} |\tilde{Z}_n(s)|_\infty \geqslant \tilde{Z}_{nj^*}(m)\right).
\end{aligned}$$

Because of the symmetry, we only deal with I since II obeys the same bound. Let

$$r = \begin{cases} C(\bar{b}, K, c_1, c_2)\frac{\log^4(np)}{\underline{\delta}_n^2} & \text{in Part (i)} \\ C(\bar{b}, K, q, c_1, c_2)\frac{\log(np)}{\underline{\delta}_n^2} \max\left\{1, \frac{n^{2/q}\log(np)}{\gamma^{2/q}}\right\} & \text{in Part (ii)} \end{cases}$$

and $\mathcal{G}$ be the event where $\max_{s \geqslant m+r} |\tilde{Z}_n(s)|_\infty$ is attained at the coordinates of $j \in \mathcal{S}$, i.e.,

$$\mathcal{G} = \left\{\max_{s \geqslant m+r} \max_{1 \leqslant j \leqslant p} \left|\tilde{Z}_{nj}(s)\right| = \max_{s \geqslant m+r} \max_{j \in \mathcal{S}} \left|\tilde{Z}_{nj}(s)\right|\right\}. \tag{2.56}$$

By Lemma 2.11, $\mathbb{P}(\mathcal{G}) \geqslant 1-\gamma/18$. From now on, our analysis will be restricted to events $\mathcal{G}$ and $\tilde{Z}_{nj^*}(m) \geqslant 0$, where the union event holds for probability greater than $1-\gamma/12$. Note that $|x| \geqslant y \geqslant 0$ implies that either $x - y \geqslant 0$ or $x + y \leqslant 0$. Then we have

$$\begin{aligned}
I \leqslant& \mathbb{P}\left(\max_{s \geqslant m+r} \max_{j \in \mathcal{S}} |\tilde{Z}_n(s)|_\infty \geqslant \tilde{Z}_{nj^*}(m)\right) + \gamma/12 \\
\leqslant& \mathbb{P}\left(\bigcup_{s \geqslant m+r} \bigcup_{j \in \mathcal{S}} \left\{\tilde{Z}_{nj}(s) - \tilde{Z}_{nj^*}(m) \geqslant 0\right\}\right) \\
&+ \mathbb{P}\left(\bigcup_{s \geqslant m+r} \bigcup_{j \in \mathcal{S}} \left\{\tilde{Z}_{nj}(s) + \tilde{Z}_{nj^*}(m) \leqslant 0\right\}\right) + \gamma/12 \\
\leqslant& \mathbb{P}\left(\max_{s \geqslant m+r} \left[\max_{j \in \mathcal{S}} \tilde{Z}_{nj}(s) - \max_{j \in \mathcal{S}} \tilde{Z}_{nj}(m)\right] \geqslant 0\right) \\
&+ \mathbb{P}\left(\min_{s \geqslant m+r} \left[\min_{j \in \mathcal{S}} \tilde{Z}_{nj}(s) + \max_{j \in \mathcal{S}} \tilde{Z}_{nj}(m)\right] \leqslant 0\right) + \gamma/12 \\
:=& III + IV + \gamma/12.
\end{aligned}$$

Since $\max_i a_i - \max_i b_i \leqslant \max_i(a_i - b_i)$ and $\min_i a_i - \min_i b_i \geqslant \min_i(a_i - b_i)$ hold for any sequences $\{a_i\}$ and $\{b_i\}$, we have

$$III \leqslant \mathbb{P}\left(\max_{s\geqslant m+r}\max_{j\in\mathcal{S}}\left[\tilde{Z}_{nj}(s) - \tilde{Z}_{nj}(m)\right] \geqslant 0\right),$$
$$IV \leqslant \mathbb{P}\left(\min_{s\geqslant m+r}\min_{j\in\mathcal{S}}\left[\tilde{Z}_{nj}(s) + \tilde{Z}_{nj}(m)\right] \leqslant 0\right).$$

For each $s \geqslant m + r$ and $j = 1, \ldots, p$, since $\tilde{Z}_{nj}(s) - \tilde{Z}_{nj}(m) \geqslant 0$ if and only if

$$\begin{aligned}&\tilde{Z}_{nj}(s) - \mathbb{E}[\tilde{Z}_{nj}(s)] - \tilde{Z}_{nj}(m) + \mathbb{E}[\tilde{Z}_{nj}(m)] \\ &\qquad \geqslant \mathbb{E}[\tilde{Z}_{nj}(m)] - \mathbb{E}[\tilde{Z}_{nj}(s)] = -\frac{s-m}{n} m\delta_{nj}.\end{aligned}$$

Since $\xi_i = X_i - \mathbb{E}(X_i)$ and $\delta_{nj} \leqslant 0$ for all $1 \leqslant j \leqslant p$, we have

$$\begin{aligned}\left\{\tilde{Z}_{nj}(s) - \tilde{Z}_{nj}(m) \geqslant 0\right\} &\subset \left\{\left|\sum_{i=m+1}^{s}\xi_{ij} - \frac{s-m}{n}\sum_{i=1}^{n}\xi_{ij}\right| \geqslant (s-m)h(t_m)|\delta_{nj}|\right\} \\ &\subset \left\{\left|\sum_{i=m+1}^{s}\xi_{ij}\right| \geqslant \frac{1}{2}(s-m)h(t_m)|\delta_{nj}|\right\} \bigcup \left\{\frac{1}{n}\left|\sum_{i=1}^{n}\xi_{ij}\right| \geqslant \frac{1}{2}h(t_m)|\delta_{nj}|\right\}.\end{aligned}$$

Then we have $III \leqslant V + VI$, where

$$\begin{aligned}V =&\mathbb{P}\left(\max_{s\geqslant m+r}\max_{j\in\mathcal{S}}\left|\frac{1}{s-m}\sum_{i=m+1}^{s}\xi_{ij}\right| \geqslant \frac{h(t_m)}{2}\underline{\delta}_n\right) \\ \leqslant&\mathbb{P}\left(\max_{r\leqslant s'\leqslant n-m}\max_{j\in\mathcal{S}}\left|\frac{1}{s'}\sum_{i=1}^{s'}\xi_{ij}\right| \geqslant \frac{h(t_m)}{2}\underline{\delta}_n\right), \qquad (2.57)\end{aligned}$$

$$VI =\mathbb{P}\left(\max_{j\in\mathcal{S}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right| \geqslant \frac{h(t_m)}{2}\underline{\delta}_n\right). \qquad (2.58)$$

Here the second inequality for bounding V is due to $\xi_1, \ldots, \xi_n$ are i.i.d. Similarly, for each $s \geqslant m + r$ and $j = 1, \ldots, p$, $-\tilde{Z}_{nj}(s) - \tilde{Z}_{nj}(m) \geqslant 0$ if and only if

$$\begin{aligned}&-\tilde{Z}_{nj}(s) + \mathbb{E}[\tilde{Z}_{nj}(s)] - \tilde{Z}_{nj}(m) + \mathbb{E}[\tilde{Z}_{nj}(m)] \\ &\qquad \geqslant \mathbb{E}[\tilde{Z}_{nj}(m)] + \mathbb{E}[\tilde{Z}_{nj}(s)] = -\frac{2n-s-m}{n} m\delta_{nj}.\end{aligned}$$

Then we have

$$\begin{aligned}
&\left\{-\tilde{Z}_{nj}(s)-\tilde{Z}_{nj}(m)\geqslant 0\right\}\\
\subset&\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right|\geqslant\frac{h(t_m)}{2}|\delta_{nj}|\right\}\cup\left\{\frac{1}{2n-s-m}\left|\sum_{i=m+1}^{n}\xi_{ij}+\sum_{i=s+1}^{n}\xi_{ij}\right|\geqslant\frac{h(t_m)}{2}|\delta_{nj}|\right\}.
\end{aligned}$$

Since $(2n-s-m)^{-1}\leqslant(n-m)^{-1}$, we have

$$\begin{aligned}
&\left\{\frac{1}{2n-s-m}\left|\sum_{i=m+1}^{n}\xi_{ij}+\sum_{i=s+1}^{n}\xi_{ij}\right|\geqslant\frac{h(t_m)}{2}|\delta_{nj}|\right\}\\
\subset&\left\{\frac{1}{n-m}\left|\sum_{i=m+1}^{n}\xi_{ij}\right|\geqslant\frac{h(t_m)}{4}|\delta_{nj}|\right\}\cup\left\{\frac{1}{n-m}\left|\sum_{i=s+1}^{n}\xi_{ij}\right|\geqslant\frac{h(t_m)}{4}|\delta_{nj}|\right\}.
\end{aligned}$$

Then we obtain that

$$IV=\mathbb{P}\left(\max_{s\geqslant m+r}\max_{j\in\mathcal{S}}\left[-\tilde{Z}_{nj}(s)-\tilde{Z}_{nj}(m)\right]\geqslant 0\right)\leqslant VI+2VII,$$

where

$$VII=\mathbb{P}\left(\max_{s\geqslant m}\max_{j\in\mathcal{S}}\left|\frac{1}{n}\sum_{i=s+1}^{n}\xi_{ij}\right|\geqslant\frac{h(t_m)}{8}\underline{\delta}_n\right). \tag{2.59}$$

So now we have $I\leqslant V+2VI+2VII+\gamma/12$.

Part (i). Suppose (C) holds. To bound V, applying Lemma 2.8, we have for any $u>0$

$$\begin{aligned}
&\mathbb{P}\left(\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s'}\frac{1}{s'}\xi_{ij}\right|\geqslant 2\mathbb{E}\left[\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s'}\frac{1}{s'}\xi_{ij}\right|\right]+u\right)\\
&\quad\leqslant\exp\left(-\frac{u^2}{3\tau_1^2}\right)+3\exp\left(-\frac{u}{K_1\|M_1\|_{\psi_1}}\right),
\end{aligned}$$

where

$$\tau_1^2=\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\sum_{i=1}^{s'}\frac{1}{s'^2}\mathbb{E}(\xi_{ij}^2)\quad\text{and}\quad M_1=\max_{1\leqslant i\leqslant n}\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\frac{1}{s'}|\xi_{ij}|.$$

Note that $\tau_1^2\leqslant r^{-1}\bar{b}$ and

$$\begin{aligned}
\|M_1\|_2\leqslant K_2\|M_1\|_{\psi_1}&=K_2\left\|\max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant p}\max_{r\leqslant s'\leqslant n-m}({s'}^{-1})|\xi_{ij}|\right\|_{\psi_1}\\
&\leqslant K_2r^{-1}\log(np)\max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant p}\|\xi_{ij}\|_{\psi_1}\leqslant K_2\bar{b}r^{-1}\log(np).
\end{aligned}$$

Using Lemma E.1 in [35], we have

$$\mathbb{E}\left[\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s'}\frac{1}{s'}\xi_{ij}\right|\right] \leqslant K_3\left\{\sqrt{\log(np)}\tau_1+\log(np)\|M_1\|_2\right\}$$
$$\leqslant C_1(\bar{b})\left\{r^{-1/2}\log^{1/2}(np)+r^{-1}\log^2(np)\right\}$$
$$\leqslant C_1(\bar{b})r^{-1/2}\log^2(np).$$

Therefore, we have

$$\mathbb{P}\left(\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s'}\frac{1}{s'}\xi_{ij}\right|\geqslant 2C_1(\bar{b})r^{-1/2}\log^2(np)+u\right)$$
$$\leqslant \exp\left(-\frac{ru^2}{3\bar{b}}\right)+3\exp\left(-\frac{ru}{K_3\bar{b}\log(np)}\right).$$

Let $u^*=C^*(\bar{b},K)r^{-1/2}\log^2(np)$. Then it follows from the assumption $\log(1/\gamma)\leqslant K\log(np)$ that

$$V\leqslant \mathbb{P}\left(\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\frac{1}{s'}\sum_{i=1}^{s'}\xi_{ij}\right|\geqslant u^*\right)\leqslant \gamma/12.$$

Similarly, to bound VI, by Lemma 2.8, we have for any $u>0$

$$\mathbb{P}\left(\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right|\geqslant 2\mathbb{E}\left[\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right|\right]+u\right)$$
$$\leqslant \exp\left(-\frac{u^2}{3\tau_2^2}\right)+3\exp\left(-\frac{u}{K_1\|M_2\|_{\psi_1}}\right),$$

where

$$\tau_2^2=\max_{1\leqslant j\leqslant p}\sum_{i=1}^{n}\frac{1}{n^2}\mathbb{E}(\xi_{ij}^2)\leqslant \bar{b}n^{-1}\quad\text{and}\quad M_2=\max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant p}\frac{1}{n}|\xi_{ij}|.$$

Note that $\|M_2\|_2\leqslant\sqrt{2}\|M_2\|_{\psi_1}\leqslant K_4\bar{b}n^{-1}\log(np)$. Since $\log^3(np)\leqslant Kn$, we have

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right|\right]\leqslant C_2(\bar{b})\left\{n^{-1/2}\log^{1/2}(np)+n^{-1}\log^2(np)\right\}$$
$$\leqslant C_2(\bar{b},K)n^{-1/2}\log^{1/2}(np).$$

Let $u^\diamond = C^\diamond(\bar{b}, K)n^{-1/2}\log^{1/2}(np)$. Then it yields that

$$VI \leqslant \mathbb{P}\left(\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right| \geqslant u^\diamond\right) \leqslant \gamma/12.$$

For VII, notice that

$$\begin{aligned}
&\mathbb{P}\left(\max_{n-1\geqslant s\geqslant m}\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=s+1}^{n}\xi_{ij}\right| \geqslant 2\mathbb{E}\left[\max_{n-1\geqslant s\geqslant m}\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=s+1}^{n}\xi_{ij}\right|\right] + u\right) \\
&\quad\leqslant \exp\left(-\frac{u^2}{3\tau_3^2}\right) + 3\exp\left(-\frac{u}{K_1\|M_3\|_{\psi_1}}\right),
\end{aligned}$$

where

$$\begin{aligned}
\tau_3^2 &= \max_{n-1\geqslant s\geqslant m}\max_{1\leqslant j\leqslant p}\sum_{i=s+1}^{n}\frac{1}{n^2}\mathbb{E}(\xi_{ij}^2) \leqslant \bar{b}n^{-2}(n-m) \leqslant \bar{b}n^{-1}, \\
M_3 &= \max_{n-1\geqslant s\geqslant m}\max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant p}|\frac{1}{n}\xi_{ij}| = M_2, \\
\mathbb{E}&\left[\max_{n-1\geqslant s\geqslant m}\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=s+1}^{n}\xi_{ij}\right|\right] \leqslant C_3(\bar{b})\left\{n^{-1/2}\log^{1/2}(np) + n^{-1}\log^2(np)\right\}.
\end{aligned}$$

Then it follows that

$$VII \leqslant \mathbb{P}\left(\max_{n-1\geqslant s\geqslant m}\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=s+1}^{n}\xi_{ij}\right| \geqslant u^\diamond\right) \leqslant \gamma/12.$$

Now combining these estimates into (2.57), (2.58), and (2.59), we conclude that $I \leqslant \gamma/2$ holds under the assumption (2.54) and choosing a large enough constant $C(\bar{b}, K, c_1, c_2) > 0$ in the definition of r. Same bound holds for II and (2.21) follows.

Part (ii). Suppose (D) holds. To bound V, applying Lemma 2.8, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s'}\frac{1}{s'}\xi_{ij}\right| \geqslant 2\mathbb{E}\left[\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s'}\frac{1}{s'}\xi_{ij}\right|\right] + u\right) \\
&\quad\leqslant \exp\left(-\frac{u^2}{3\tau_1^2}\right) + C(q)\frac{\mathbb{E}[M_1^q]}{u^q}
\end{aligned}$$

holds for all $u > 0$. Note that $\tau_1^2 \leqslant r^{-1}\bar{b}$, and for $q \geqslant 2$ we have $\|M_1\|_2 \leqslant \|M_1\|_q$ with

$$\|M_1\|_q^q = \mathbb{E}\left[\max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant p}(\max_{s'\geqslant r}\frac{1}{s'})|\xi_{ij}|^q\right] \leqslant r^{-q}\sum_{i=1}^{n}\mathbb{E}\left[\max_{1\leqslant j\leqslant p}|\xi_{ij}|^q\right] \leqslant \bar{b}^q r^{-q} n.$$

Thus,

$$\mathbb{E}\left[\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s'}\frac{1}{s'}\xi_{ij}\right|\right] \leqslant K_5\left\{\sqrt{\log(np)}\tau_1+\log(np)\|M_1\|_2\right\}$$
$$\leqslant C_4(\bar{b},q)\left\{r^{-1/2}\log^{1/2}(np)+r^{-1}n^{1/q}\log(np)\right\}.$$

Let

$$u^* = C^*(\bar{b},K,q)r^{-1/2}\log^{1/2}(np)\max\{1,\gamma^{-1/q}n^{1/q}\log^{1/2}(np)\}.$$

Then, we have

$$V \leqslant \mathbb{P}\left(\max_{r\leqslant s'\leqslant n-m}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{s'}\frac{1}{s'}\xi_{ij}\right| \geqslant u^*\right) \leqslant \gamma/12.$$

To bound VI, note that

$$\mathbb{P}\left(\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right| \geqslant 2\mathbb{E}\left[\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right|\right]+t\right) \leqslant \exp\left(-\frac{t^2}{3\tau_2^2}\right)+C(q)\frac{\mathbb{E}[M_2^q]}{t^q},$$

where τ_2^2 and M_2 are defined the same as in in Part (i). Since $\tau_2^2 \leqslant n^{-1}\bar{b}$ and

$$\|M_2\|_q^q = \frac{1}{n^q}\mathbb{E}\left[\max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant p}|\xi_{ij}|^q\right] \leqslant \frac{1}{n^q}\sum_{i=1}^{n}\mathbb{E}\left[\max_{1\leqslant j\leqslant p}|\xi_{ij}|^q\right] \leqslant \bar{b}^q n^{1-q},$$
$$\mathbb{E}\left[\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij}\right|\right] \leqslant K_6\left\{\sqrt{\log(np)}\tau_2+\log(np)\|M_2\|_2\right\}$$
$$\leqslant C_5(\bar{b},q)\left\{n^{-1/2}\log^{1/2}(np)+n^{1/q-1}\log(np)\right\}.$$

As $\log(\gamma^{-1}) \leqslant K\log(np)$, we can take

$$u^\diamond = C^\diamond(\bar{b},K,q)n^{-1/2}\log^{1/2}(np)\max\{1,\gamma^{-1/q}n^{1/q-1/2}\log^{1/2}(np)\}$$

so that

$$VI \leqslant \mathbb{P}\left(\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{n}\xi_{ij}\right| \geqslant u^\diamond\right) \leqslant \gamma/12.$$

For VII, notice that

$$\mathbb{P}\left(\max_{s\geqslant m}\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=s+1}^{n}\xi_{ij}\right|\geqslant 2\mathbb{E}\left[\max_{s\geqslant m}\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=s+1}^{n}\xi_{ij}\right|\right]+t\right)$$
$$\leqslant \exp\left(-\frac{t^2}{3\tau_3^2}\right)+C(q)\frac{\mathbb{E}[M_3^q]}{t^q},$$

where $\tau_3^2 \leqslant \bar{b}n^{-1}$ and M_3 are defined the same as in Part (i). Then we have

$$VII \leqslant \mathbb{P}\left(\max_{s\geqslant m}\max_{1\leqslant j\leqslant p}\left|\frac{1}{n}\sum_{i=s+1}^{n}\xi_{ij}\right|\geqslant t^{\diamond}\right)\leqslant \gamma/12.$$

Hence, $I \leqslant \gamma/2$ under the assumption (2.55), and choosing a large enough constant $C(\bar{b}, K, q, c_1, c_2) > 0$ in the definition of r. By a similar argument, II obeys the same bound as I. □

Lemma 2.11. *Suppose that (B) holds and H_1 is true with a change point m satisfying $c_1 \leqslant t_m \leqslant c_2$ for some constants $c_1, c_2 \in (0,1)$. Suppose that $\log^3(np) \leqslant Kn$ and $\log(\gamma^{-1}) \leqslant K\log(np)$ for some constant $K > 0$. Let $\mathcal{G}$ be defined in (2.56). Then $\mathbb{P}(\mathcal{G}) \geqslant 1-\gamma/18$,*
(i) if (C) and (2.54) hold, where $r = C(\bar{b}, K, c_1, c_2)\underline{\delta}_n^{-2}\log^4(np)$; or
(ii) if (D) and (2.55) hold, where $r = C(\bar{b}, K, q, c_1, c_2)\underline{\delta}_n^{-2}\log(np)\max\{1, \gamma^{-2/q}n^{2/q}\log(np)\}$.
In addition, if $\delta_{nj} \leqslant 0$ for all $1 \leqslant j \leqslant p$, then $\mathbb{P}\left(\max_{1\leqslant j\leqslant p}\left|\tilde{Z}_{nj}(m)\right| = \tilde{Z}_{nj^*}(m)\right) \geqslant 1-\gamma/36$ in both (i) and (ii), where $j^* \in \{1,\dots,p\}$ is defined as $\tilde{Z}_{nj^*}(m) = \max_{1\leqslant j\leqslant p}\tilde{Z}_{nj}(m)$.

2.6.6 Proof of Theorem 2.7

Proof of Theorem 2.7. Part (i) Assuming (C). Suppose there are undetected change-points in $[b,e]$ which satisfies

$$m_{k_0} \leqslant b < m_{k_0+1} < \cdots < m_{k_0+q} < e \leqslant m_{k_0+q+1}$$

for $0 \leqslant k_0 \leqslant \nu - q$. Let $\lambda_1 = \lambda_2 = \bar{b}\log^2(np)$,

$$\mathcal{A}_n = \left\{|(e-b+1)^{-1/2}\sum_{i=b}^{e}\xi_i|_\infty < K_2\lambda_2, \forall 1\leqslant b\leqslant e\leqslant n\right\}$$

and

$$\mathcal{B}_n = \left\{\max_{1\leqslant b\leqslant s\leqslant e\leqslant n}|Z_{n,b,e}(s) - EZ_{n,b,e}(s)|_\infty < K_1\lambda_1\right\}.$$

By Lemma 2.12, $\mathbb{P}(\mathcal{A}_n\cap\mathcal{B}_n) \geqslant 1-2\gamma$. In the following, we will only consider the case of $\mathcal{A}_n\cap\mathcal{B}_n$ where the random component of $\{X_i\}_{i=1}^n$, namely the behavior of $\{\xi_i\}_{i=1}^n$, is well characterized.

Let $\hat{m}_{b,e} = \text{argmax}_{b \leqslant s \leqslant e} |Z_{n,b,e}(s)|_\infty$ be the location where the maximum CUSUM statistic within interval $[b, e]$ is reached and let $h = \text{argmax}_{j=1,\ldots,p} |Z_{n,b,e}(\hat{m}_{b,e})|_j$ be its dimension. According to Algorithm 1, if $\hat{m}_{b,e}$ is large enough to pass the bootstrap test and it is close to one of $\{m_{k_0+1}, \ldots, m_{k_0+q}\}$, then one change point is consistently identified. The remaining proof follows the structure of proof in [45, Theorem 1] to complete this claim with modifications on the key steps: i) $\hat{m}_{b,e} \in \mathcal{S}_n$; ii) Power of bootstrap test is guaranteed such that $\mathbb{P}(\mathcal{S}_n)$ is bounded below. Note that, the main differences between the proof of [45, Theorem 1] and our argument are in the following aspects. First, [45] considers one-dimensional observations while we extend it to p-dimensional case. The set $\mathcal{A}_n$ and $\mathcal{B}_n$ are adapted to sub-exponential random vectors. Second, [45, Lemma A.5] provides stopping conditions when the following (2.60) and (2.61) fail, that is the search stops if (i) no change point in $[b, e]$ or (ii) the only 1 or 2 change points left in $[b, e]$ are within the distance of $\underline{s}$ from b or e (therefore they should be classified as the same change point as b or e). But under our context, testing procedure substitutes subjective pick of threshold so that we do not need to control the magnitude of $|Z_{n,b,e}(s)|_\infty$.

Apply [45, Theorem 1], we can make the following statements on set $\mathcal{B}_n \cap \mathcal{A}_n$. When

$$b < m_{k_0+r} - C_3 D_\nu < m_{k_0+r} + C_3 D_\nu < e \text{ for some } 1 \leqslant r \leqslant q \tag{2.60}$$

and

$$\max\left\{\min\{m_{k_0+1} - b, b - m_{k_0}\}, \min\{m_{k_0+q+1} - e, e - m_{k_0+q}\}\right\} \leqslant C_4 \epsilon_n \tag{2.61}$$

hold for some constants C_3, C_4, by Lemma A.3 in [45],

$$|\hat{m}_{b,e} - m_{k_0+r}| \leqslant C_5 \epsilon_n' = C_5 \lambda_2^2 n^2 D_\nu^{-2} (\delta_h^{(p_0+r)})^{-2}$$

i.e. $\hat{m}_{b,e}$ falls within the distance of $C_5 \bar{b}^2 \epsilon_n$ of a previously undetected change point m_{k_0+r}.

Note that $h \in \mathcal{D}_{k_0+r}$, i.e. $\delta_h^{(k_0+r)} \neq 0$. Otherwise, if the h^{th}-dimension has a mean-shift within the working interval $[b, e]$, then it contradicts with $|Z_{n,b,e}(s)|_\infty = |Z_{n,b,e}(\hat{m}_{b,e})|_h$. If there is no mean-shift in the h^{th}-dimension within $[b, e]$, then $|Z_{n,b,e}(\hat{m}_{b,e})|_h \leqslant C_6(\alpha, \bar{b})\lambda_1$. But by Lemma 2.13,

$$\begin{aligned}
\max_{b+\underline{s} \leqslant s \leqslant e-\underline{s}} |EZ_{n,b,e}(s)|_\infty &\geqslant \max_{l=k_0,\ldots,k_0+q+1} \frac{CD_\nu^2 \bar{\delta}_n}{\sqrt{(e-b)(m_l-b)(e-m_l)}} \\
&\geqslant \frac{CD_\nu^2 \underline{\delta}_n}{\sqrt{nD_\nu(n-D_\nu)}} \\
&\geqslant C_7(\bar{b}) D_\nu^{3/2} n^{-1} \underline{\delta}_n
\end{aligned}$$

The second line is due to $D_\nu \geqslant \epsilon_n \gtrsim b - m_{k_0} \geqslant m_{k_0+1} - m_{k_0} \geqslant D_\nu$ by (2.61) and Assumption a), b), e). It leads to contradiction. When $\delta_h^{(p_0+r)} = 0$ but $\delta_h^{(p_0+r+1)} \neq 0$, we just need to record $r+1$ as r.

Next, according to Lemma 2.6, in order to have Type-II error (of performed test in current interval $[b,e]$) is less than $\gamma + 2\zeta + C_2\varpi_{1,(e-b)}$, we only need to show $|\tilde{\Delta}|_\infty = |\mathbb{E}Z_{n,b,e}(s)|_\infty \geqslant C\nu \log^{1/2}((e-b)p)$ for some $s \in [b+\underline{s}, e-\underline{s}]$, $C_1 = C_1(\underline{b}, \bar{b}, K)$ and $C_2 = C_2(\underline{b}, \bar{b}, K)$. Here, $\varpi_{1,(e-b)} = (\log^7((e-b)p)/\underline{s})^{1/6}$ with $\underline{s}$ determined by n, not $(e-b)$. Then by Assumption d),

$$\max_{b+\underline{s}\leqslant s\leqslant e-\underline{s}} |EZ_{n,b,e}(s)|_\infty \geqslant C_7 D_\nu^{3/2} n^{-1} \underline{\delta}_n \geqslant C_7 C^2 n^{\frac{3}{2}\Theta - 1 - \omega} > C_8(\alpha, \bar{b})\nu \log^{1/2}((e-b)p),$$

which indicates the change point is significant to pass the bootstrap test with probability no less than $1 - \gamma - 2\zeta - C_0^{'}\varpi_{1,n}$ where $\varpi_{1,n} \geqslant \varpi_{1,(e-b)}$ by Assumption e).

As a consequence, the procedure then moves on to operate on the intervals $[b, \hat{m}_{b,e}]$ and $[\hat{m}_{b,e}, e]$ where both (2.60) and (2.61) still hold. Therefore, all change points will be detected one by one until the conditions in Lemma A.5 in [45] are met. Assumption c) is implicitly called in [45, Theorem 1].

Part (ii) Assuming (D). Let $\lambda_1 = \lambda_2 = C(q)\bar{b}n^{3/q}(\log(np) + \gamma^{-1/q})$. Similarly, Lemma 2.12 shows $\mathbb{P}(\mathcal{A}_n \cap \mathcal{B}_n) \geqslant 1 - 2\gamma$. According to the proof of Theorem 2.3, Type-II error is less than $\gamma + 2\zeta + C_0^{'}(\varpi_{1,(e-b)} + \varpi_{2,(e-b)})$ when $|\tilde{\Delta}|_\infty = |\mathbb{E}Z_{n,b,e}(s)|_\infty \geqslant C\log((e-b)p)$ for some $s \in [b+\underline{s}, e-\underline{s}]$. Using the same arguments, we conclude that for each step in binary segmentation the bootstrap test is passed with probability no less than $1 - \gamma - 2\zeta - C_0^{'}(\varpi_{1,n} + \varpi_{2,n})$ and estimated location $\hat{m}_{b,e}$ falls within distance of $C\epsilon_n = C(\bar{b}, q)n^{2+6/q}D_\nu^{-2}(\underline{\delta}_n)^{-2}(\log^2(np) + \gamma^{-2/q})$ until the stopping conditions are met. □

Lemma 2.12. *$\mathbb{P}(\mathcal{A}_n) \geqslant 1-\gamma$ and $\mathbb{P}(\mathcal{B}_n) \geqslant 1-\gamma$ for γ defined the same as Theorem 2.7.*

Proof of Lemma 2.6. The structure of this proof is similar to the one for Theorem 2.3. Without less of generality we may assume $\mu = 0$. For $\xi_i = X_i - \mu_i$ where $\mu_i = \mathbb{E}[X_i]$ has the form of

$$\mu_{m_k+1} = \cdots = \mu_{m_{k+1}} = \sum_{l=0}^{k} \delta^{(l)},$$

the CUSUM statistic computed on $X_1, \ldots, X_n$ can be decomposed as $Z_n(s) = Z_n^\xi(s) + \Delta_s$, where $Z_n^\xi(s) = \sqrt{\frac{s(n-s)}{n}}\left\{\frac{1}{s}\sum_{i=1}^s \xi_i - \frac{1}{n-s}\sum_{i=s+1}^n \xi_i\right\}$ is defined the same as in the proof of Theorem 2.3 but $\Delta_s = \mathbb{E}Z_n(s)$ is extended to multiple change point model as in (2.69). Again, by Lemma 2.13, $|\Delta_s|_\infty$ reaches its maximum at one of the change points, i.e.,

$$\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} |\Delta_s|_\infty = \max_{k=1,\ldots,\nu} |\Delta_{m_k}|_\infty \gtrsim \max_{l=1,\ldots,\nu} \frac{CD_\nu^2 \bar{\delta}_n}{\sqrt{nm_l(n-m_l)}} =: \tilde{\Delta}. \tag{2.62}$$

The Type II error obeys the same bound in arguments of (2.44) and (2.45).

Part (i). Assume (C). Take $\beta_n = C_1\varpi_{1,n} + 2\zeta$, $q_{\tilde{T}_n}(1-\beta_n)$ holds for the same bound in (2.46) . Note that $q_{T_n^*|X_1^n}(1-\alpha)$ is bounded the same as in (2.47), but ψ changes to

$$\bar{\psi}^2 \leqslant 2 \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \left\{ \frac{n-s}{n} \hat{S}^{\xi,-}_{n,s,jj} + \frac{s}{n} \hat{S}^{\xi,+}_{n,s,jj} \right\} + 4\nu^2 \bar{\delta}_n^2. \tag{2.63}$$

The sketch of proof for (2.63) is shown in Part (iii). Therefore, based on the same probability bounds of $\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \left|\frac{1}{s}\sum_{i=1}^s (\xi_{ij}^2 - \Sigma_{jj})\right|$ and $\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \left|\bar{\xi}_{sj}^-\right|^2$, we deduce that

$$\mathbb{P}\left(q_{T_n^*|X_1^n}(1-\alpha) \geqslant C_5 \max\{\nu\bar{\delta}_n, 1\} \log^{1/2}(np/\alpha) \right) \leqslant \gamma.$$

Therefore, (2.16) follows.

Part (ii). Assume (D). Arguments are exactly the same as Part (ii) in Theorem 2.3.

Part (iii). The result of (2.63) comes from the proof of Lemma 2.10 with a modification to multiple mean-shifts model (2.10). Recall that $\bar{X}_s^- = s^{-1}\sum_{i=1}^s X_i$, $\bar{X}_s^+ = (n-s)^{-1}\sum_{i=s+1}^n X_i$, and $\bar{\xi}_s^-, \bar{\xi}_s^+, \bar{\mu}_s^-, \bar{\mu}_s^+$ are similarly defined by replacing X_1^n with ξ_1^n and μ_1^n, respectively. Then, elementary calculations yield $\bar{X}_s^- = \bar{\xi}_s^- + \bar{\mu}_s^-$ and $\bar{X}_s^+ = \bar{\xi}_s^+ + \bar{\mu}_s^+$, where

$$\bar{\mu}_s^- = \frac{1}{s}\sum_{l=0}^k (s-m_l)\delta^{(l)} \text{ and } \bar{\mu}_s^+ = \sum_{l=0}^k \delta^{(l)} + \frac{1}{n-s}\sum_{l=k+1}^{\nu} (n-m_l)\delta^{(l)} \text{ for } m_k < s \leqslant m_{k+1}.$$

Note that

$$\begin{cases} \frac{1}{\sqrt{s}}\sum_{i=1}^s e_i(X_i - \bar{X}_s^-) = \underbrace{\frac{1}{\sqrt{s}}\sum_{i=1}^s e_i(\xi_i - \bar{\xi}_s^-)}_{:=A} + \frac{1}{\sqrt{s}}\sum_{i=1}^s e_i(\mu_i - \bar{\mu}_s^-) \\ \frac{1}{\sqrt{n-s}}\sum_{i=s+1}^n e_i(X_i - \bar{X}_s^+) = \underbrace{\frac{1}{\sqrt{n-s}}\sum_{i=s+1}^n e_i(\xi_i - \bar{\xi}_s^+)}_{:=B} + \frac{1}{\sqrt{n-s}}\sum_{i=s+1}^n e_i(\mu_i - \bar{\mu}_s^+) \end{cases},$$

where

$$\begin{cases} \sum_{i=1}^s (\mu_{ij} - \bar{\mu}_{sj}^-)^2 & = \sum_{i=1}^s \mu_{ij}^2 - s(\bar{\mu}_{sj}^-)^2 & \leqslant \sum_{i=1}^s \mu_{ij}^2 & \leqslant s\nu^2\bar{\delta}_n^2 \\ \sum_{i=s+1}^n (\mu_{ij} - \bar{\mu}_{sj}^+)^2 & = \sum_{i=s+1}^n \mu_{ij}^2 - (n-s)(\bar{\mu}_{sj}^+)^2 & \leqslant \sum_{i=s+1}^n \mu_{ij}^2 & \leqslant (n-s)\nu^2\bar{\delta}_n^2 \end{cases}$$

as $|\mu_s|_\infty = |\sum_{l=0}^k \delta^{(l)}|_\infty \leqslant \nu\bar{\delta}_n$ for $s = m_k+1, \ldots, m_{k+1}$. Similarly, we have

$$\begin{aligned}
\bar{\psi}^2 &= \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \Big\{ \frac{n-s}{n}\mathrm{Cov}_e\big(A_j + \frac{1}{\sqrt{s}}\sum_{i=1}^s e_i(\mu_{ij} - \bar{\mu}_{sj}^-)\big) \\
&\qquad + \frac{s}{n}\mathrm{Cov}_e\big(B_j + \frac{1}{\sqrt{n-s}}\sum_{i=s+1}^n e_i(\mu_{ij} - \bar{\mu}_{sj}^+)\big)\Big\} \\
&\leqslant 2\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \Big\{ \frac{n-s}{n}\mathrm{Cov}_e(A_j) + \frac{s}{n}\mathrm{Cov}_e(B_j)\Big\} \\
&\quad +2\max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j\leqslant p} \left[\frac{1}{s}\sum_{i=1}^s(\mu_{ij} - \bar{\mu}_{sj}^-)^2 + \frac{1}{n-s}\sum_{i=s+1}^n(\mu_{ij} - \bar{\mu}_{sj}^+)^2\right].
\end{aligned}$$

Then (2.63) is immediate. □

Lemma 2.13. *$argmax_{s=1,\ldots,n-1}|\mathbb{E}Z_n(s)|_\infty \subset \{m_k, k = 1, \ldots, \nu\}$. Consequently, we automatically have $argmax_{s=b,\ldots,e}|\mathbb{E}Z_{n,b,e}(s)|_\infty \subset \{m_k, k=1,\ldots,\nu\} \cap [b,e]$. Moreover, there exist some constant C such that*

$$\max_{l=1,\ldots,\nu}|\Delta_{m_l}|_\infty \geqslant \max_{l=1,\ldots,\nu}\frac{CD_\nu^2\bar{\delta}_n}{\sqrt{nm_l(n-m_l)}} \geqslant \max_{l=1,\ldots,\nu} Cn^{-5/2}\sqrt{m_l(n-m_l)}D_\nu^2\bar{\delta}_n.$$

2.6.7 Proof of auxiliary lemmas

This section contains auxiliary lemmas for Section 2.6.

Proof of Lemma 2.8. For $i = 1, \ldots, n$, let

$$Y_i = \begin{pmatrix} a_{1i}X_{i1} & \ldots & a_{1i}X_{ip} \\ \vdots & \ddots & \vdots \\ a_{ni}X_{i1} & \ldots & a_{ni}X_{ip} \end{pmatrix}.$$

Then $Y_1, \ldots, Y_n$ is a sequence of independent mean-zero random matrices in $\mathbb{R}^{n\times p}$. Note that $W_n = \sum_{i=1}^n Y_i$ and $Z_n = |W_n|_\infty$. Then (2.28) is an immediate consequence of Lemma E.3 in [35]. □

Proof of Lemma 2.9. Part (i). Assume (C). Write

$$\hat{\Delta}_1 = \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j,k\leqslant p}\left|\sum_{i=1}^n b_{is}(X_{ij}X_{ik} - \sigma_{jk})\right|,$$

where

$$b_{is} = \begin{cases} s^{-1}, & \text{if } 1\leqslant i\leqslant s \\ 0, & \text{if } s+1\leqslant i\leqslant n \end{cases}.$$

By Part (i) of Lemma 2.8, there exists a universal constant $K_1 > 0$ such that for all $t > 0$,

$$\mathbb{P}(\hat{\Delta}_1 \geqslant 2\mathbb{E}[\hat{\Delta}_1] + t) \leqslant \exp\left(-\frac{t^2}{3\tau^2}\right) + 3\exp\left[-\left(\frac{t}{K_1\|M\|_{\psi_{1/2}}}\right)^{1/2}\right],$$

where

$$\begin{aligned}
\tau^2 &= \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j,k\leqslant p} \sum_{i=1}^{n} b_{is}^2 \mathbb{E}(X_{ij}X_{ik} - \sigma_{jk})^2, \\
M &= \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j,k\leqslant p} \max_{1\leqslant i\leqslant s} |s^{-1}(X_{ij}X_{ik} - \sigma_{jk})|.
\end{aligned}$$

Note that

$$\begin{aligned}
\|M\|_2 \leqslant K_2\|M\|_{\psi_{1/2}} \leqslant &\frac{K_2}{\underline{s}}\| \max_{1\leqslant i\leqslant n} \max_{1\leqslant j,k\leqslant p} |X_{ij}X_{ik}|\|_{\psi_{1/2}} \leqslant \frac{K_2}{\underline{s}}\| \max_{1\leqslant i\leqslant n} \max_{1\leqslant j\leqslant p} X_{ij}^2\|_{\psi_{1/2}} \\
= &\frac{K_2}{\underline{s}}\| \max_{1\leqslant i\leqslant n} \max_{1\leqslant j\leqslant p} |X_{ij}|\|_{\psi_1} \leqslant \frac{K_2}{\underline{s}}\bar{b}^2 \log^2(np).
\end{aligned}$$

By the Cauchy-Schwarz inequality and assumption (B), we have

$$\tau^2 \leqslant \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} \max_{1\leqslant j,k\leqslant p} \sum_{i=1}^{s} \frac{1}{s^2}\mathbb{E}(X_{ij}X_{ik})^2 \leqslant \bar{b}^2\underline{s}^{-1}.$$

By [35, Lemma E.1], there exists a universal constant $K_3 > 0$ such that

$$\begin{aligned}
\mathbb{E}[\hat{\Delta}_1] &\leqslant K_3\left\{\tau\log^{1/2}(np^2) + \|M\|_2\log(np^2)\right\} \\
&\leqslant K_3\left\{\bar{b}\log^{1/2}(np)\underline{s}^{-1/2} + \bar{b}^2\log^3(np)\underline{s}^{-1}\right\}.
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
&\mathbb{P}\left(\hat{\Delta}_1 \geqslant 2K_3\left\{\bar{b}\log^{1/2}(np)\underline{s}^{-1/2} + \bar{b}^2\log^3(np)\underline{s}^{-1}\right\} + t\right) \\
&\qquad \leqslant \exp\left(-\frac{t^2\underline{s}}{3\bar{b}^2}\right) + 3\exp\left(-\frac{t^{1/2}\underline{s}^{1/2}}{K_4\bar{b}\log(np)}\right).
\end{aligned}$$

Choose $t = C_1\underline{s}^{-1/2}\log^{1/2}(np)$ for some large enough constant $C_1 := C_1(\bar{b}, K) \geqslant 1$. Using $\log(\gamma^{-1}) \leqslant K\log(np)$ and $\log^5(np) \leqslant \underline{s}$, we obtain that

$$\mathbb{P}\left(\hat{\Delta}_1 \geqslant C\underline{s}^{-1/2}\log^{3/2}(np)\right) \leqslant \gamma/4. \tag{2.64}$$

Since $X_1, \dots, X_n$ are i.i.d. under H_0, $\hat{\Delta}_1$ and $\hat{\Delta}_2$ share the same distribution and therefore $\hat{\Delta}_2$ also obeys the bound (2.64). $\hat{\Delta}_3$ and $\hat{\Delta}_4$ can be dealt similarly. Indeed, by Lemma 2.8, there exists a universal constant $K_5 > 0$ such that for all $t > 0$,

$$\mathbb{P}(\hat{\Delta}_3 \geqslant 2\mathbb{E}[\max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |\sum_{i=1}^{n} b_{is} X_i|_\infty \,] + t) \leqslant \exp\left(-\frac{t^2}{3\tilde{\tau}^2}\right) + 3\exp\left(-\frac{t}{K_5 \|\tilde{M}\|_{\psi_1}}\right),$$

where $\tilde{\tau}^2 = \max_{\underline{s} \leqslant s \leqslant n-\underline{s}, 1 \leqslant j \leqslant p} \sum_{i=1}^{n} b_{is}^2 \mathbb{E} X_{ij}^2$ and $\tilde{M} = \max_{1 \leqslant i \leqslant n, \underline{s} \leqslant s \leqslant n-\underline{s}, 1 \leqslant j \leqslant p} |b_{is} X_{ij}|$. By (B), $\tilde{\tau}^2 \leqslant \bar{b}\underline{s}^{-1}$. By (C) and [92, Lemma 2.2.2], there exists a universal constant $K_6 > 0$ such that $\|\tilde{M}\|_{\psi_1} \leqslant K_6 \bar{b} \log(np) \underline{s}^{-1}$. By [35, Lemma E.1], there exists a universal constant $K_7 > 0$ such that

$$\begin{aligned}
\mathbb{E}[\max_{\underline{s} \leqslant s \leqslant n-\underline{s}} |\sum_{i=1}^{n} b_{si} X_i|_\infty \,] &\leqslant K_7 \left\{\tilde{\tau} \log^{1/2}(np) + \|\tilde{M}\|_2 \log(np)\right\} \\
&\leqslant K_7 \left\{\bar{b} \log^{1/2}(np) \underline{s}^{-1/2} + \bar{b} \log^2(np) \underline{s}^{-1}\right\}.
\end{aligned}$$

So it follows that

$$\begin{aligned}
&\mathbb{P}\left(\hat{\Delta}_3 \geqslant 2K_7 \left\{\bar{b} \log^{1/2}(np) \underline{s}^{-1/2} + \bar{b} \log^2(np) \underline{s}^{-1}\right\} + t\right) \\
&\qquad \leqslant \exp\left(-\frac{t^2 \underline{s}}{3\bar{b}}\right) + 3\exp\left(-\frac{t\underline{s}}{K_8 \bar{b} \log(np)}\right).
\end{aligned}$$

Using $t = C\underline{s}^{-1/2} \log^{1/2}(np) \log(\gamma^{-1})$, we get

$$\mathbb{P}\left(\hat{\Delta}_3 \geqslant C\underline{s}^{-1/2} \log^{3/2}(np)\right) \leqslant \gamma/4.$$

Part (ii). Assume (D). By Part (ii) of Lemma 2.8, there exists a constant $C_1 := C_1(q) > 0$ such that for all $t > 0$,

$$\mathbb{P}(\hat{\Delta}_1 \geqslant 2\mathbb{E}[\hat{\Delta}_1] + t) \leqslant \exp\left(-\frac{t^2}{3\tau^2}\right) + C_1 \frac{\mathbb{E}[M^{q/2}]}{t^{q/2}},$$

where τ^2 and M have the same definition as in Part (i). As in Part (i), $\tau^2 \leqslant \bar{b}^2 \underline{s}^{-1}$. Note that

$$\mathbb{E}[M^{q/2}] \leqslant C_2(q) \underline{s}^{-q/2} \{\mathbb{E}[\max_{1 \leqslant i \leqslant n} \max_{1 \leqslant j \leqslant p} |X_{ij}|^q] + \max_{1 \leqslant j,k \leqslant p} |\sigma_{jk}|^{q/2}\} \leqslant C_2(q) \bar{b}^q n \underline{s}^{-q/2}.$$

and

$$\begin{aligned}
\mathbb{E}[\hat{\Delta}_1] &\leqslant K_1 \{\tau \log^{1/2}(np^2) + \|M\|_{q/2} \log(np^2)\} \\
&\leqslant K_1 \{\bar{b}\underline{s}^{-1/2} \log^{1/2}(np) + C_2(q)^{2/q} \bar{b}^2 n^{2/q} \log(np) \underline{s}^{-1}\}.
\end{aligned}$$

Therefore, there exists a constant $C_3(q) > 0$ such that

$$\begin{aligned}\mathbb{P}\left(\hat{\Delta}_1 \geqslant 2K_1\left\{\bar{b}\underline{s}^{-1/2}\log^{1/2}(np) + C_2(q)^{2/q}\bar{b}^2 n^{2/q}\log(np)\underline{s}^{-1}\right\} + t\right) \\ \leqslant \exp\left(-\frac{t^2\underline{s}}{3\bar{b}^2}\right) + C_3(q)\frac{n\bar{b}^q}{t^{q/2}\underline{s}^{q/2}}.\end{aligned}$$

Now, choosing $t = C_4\{\underline{s}^{-1/2}\log^{1/2}(np) + \gamma^{-2/q}\underline{s}^{-1}n^{2/q}\}$ for some large enough constant $C_4 := C_4(\bar{b}, K, q) \geqslant 1$. Using $\log(\gamma^{-1}) \leqslant K\log(np)$ and $\log^3(np) \leqslant n$, we obtain that

$$\mathbb{P}\left(\hat{\Delta}_1 \geqslant C_5\{\underline{s}^{-1/2}\log^{3/2}(np) + \gamma^{-2/q}\underline{s}^{-1}n^{2/q}\log(np)\}\right) \leqslant \gamma/4. \tag{2.65}$$

Other terms $\hat{\Delta}_i, i = 2, 3, 4$ can be similarly handled and details are omitted. □

Proof of Lemma 2.10. Recall that $\bar{X}_s^- = s^{-1}\sum_{i=1}^s X_i$, $\bar{X}_s^+ = (n-s)^{-1}\sum_{i=s+1}^n X_i$, and $\bar{\xi}_s^-, \bar{\xi}_s^+$ are similarly defined by replacing X_1^n with ξ_1^n. Then, elementary calculations yield

$$\bar{X}_s^- = \begin{cases} \bar{\xi}_s^-, & \text{if } 1 \leqslant s \leqslant m-1 \\ \bar{\xi}_s^- + \frac{s-m}{s}\delta_n, & \text{if } m \leqslant s \leqslant n-1 \end{cases}$$

and

$$\bar{X}_s^+ = \begin{cases} \bar{\xi}_s^+ + \frac{n-m}{n-s}\delta_n, & \text{if } 1 \leqslant s \leqslant m-1 \\ \bar{\xi}_s^+ + \delta_n, & \text{if } m \leqslant s \leqslant n-1 \end{cases}.$$

Let

$$Z_n^{\xi,*}(s) = \sqrt{\frac{n-s}{n}}\underbrace{s^{-1/2}\sum_{i=1}^s e_i(\xi_i - \bar{\xi}_s^-)}_{:=A} - \sqrt{\frac{s}{n}}\underbrace{(n-s)^{-1/2}\sum_{i=s+1}^n e_i(\xi_i - \bar{\xi}_s^+)}_{:=B}$$

be the bootstrap CUSUM statistic computed on the transformed data ξ_1^n. For $m+1 \leqslant s \leqslant n-1$, we define $b_{is} = -(s-m)/s$ if $1 \leqslant i \leqslant m$, $b_{is} = m/s$ if $m+1 \leqslant i \leqslant s$, and $b_{is} = 0$ otherwise. For $1 \leqslant s \leqslant m-1$, we define $b'_{is} = -(n-m)/(n-s)$ if $s+1 \leqslant i \leqslant m$, $b'_{is} = (m-s)/(n-s)$ if $m+1 \leqslant i \leqslant n$, and $b'_{is} = 0$ otherwise. Then routine algebra show that

$$\frac{1}{\sqrt{s}}\sum_{i=1}^s e_i(X_i - \bar{X}_s^-) = \begin{cases} A, & \text{if } 1 \leqslant s \leqslant m \\ A + \frac{\delta_n}{\sqrt{s}}\sum_{i=1}^s b_{is}e_i, & \text{if } m+1 \leqslant s \leqslant n \end{cases}$$

and

$$\frac{1}{\sqrt{n-s}}\sum_{i=s+1}^{n} e_i(X_i-\bar{X}_s^-)=\begin{cases} B+\frac{\delta_n}{\sqrt{n-s}}\sum_{i=s+1}^{n} b'_{is}e_i, & \text{if } 1\leqslant s\leqslant m-1\\ B, & \text{if } m\leqslant s\leqslant n\end{cases}.$$

Denote $\mathrm{Cov}_e(\cdot)$ as the covariance operator taken w.r.t. the random variables $e_1,\ldots,e_n$ By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\bar{\psi}^2 &= \max_{\underline{s}\leqslant s\leqslant n-\underline{s}}\max_{1\leqslant j\leqslant p}\Big\{\frac{n-s}{n}\mathrm{Cov}_e(A_j+\frac{\delta_{nj}}{\sqrt{s}}\sum_{i=1}^{s} b_{is}e_i\mathbf{1}_{(m+1\leqslant s\leqslant n-1)})\\ &\qquad +\frac{s}{n}\mathrm{Cov}_e(B_j+\frac{\delta_{nj}}{\sqrt{n-s}}\sum_{i=s+1}^{n} b'_{is}e_i\mathbf{1}_{(1\leqslant s\leqslant m-1)})\Big\}\\ &\leqslant 2\max_{\underline{s}\leqslant s\leqslant n-\underline{s}}\max_{1\leqslant j\leqslant p}\Big\{\frac{n-s}{n}\mathrm{Cov}_e(A_j)+\frac{s}{n}\mathrm{Cov}_e(B_j)\Big\}\\ &\quad +2|\delta_n|_\infty^2\max_{\underline{s}\leqslant s\leqslant n-\underline{s}}\left[\frac{1}{s}\sum_{i=1}^{s} b_{is}^2\mathbf{1}_{(m+1\leqslant s\leqslant n-1)}\right]\\ &\quad +2|\delta_n|_\infty^2\max_{\underline{s}\leqslant s\leqslant n-\underline{s}}\left[\frac{1}{n-s}\sum_{i=s+1}^{n} b'^2_{is}\mathbf{1}_{(1\leqslant s\leqslant m-1)}\right].\end{aligned}$$

Note that

$$\begin{aligned}\sum_{i=1}^{s} b_{is}^2 &= \frac{m(s-m)}{s}\leqslant s \qquad \text{for } m+1\leqslant s\leqslant n-1,\\ \sum_{i=s+1}^{n} b'^2_{is} &= \frac{(n-m)(m-s)}{n-s}\leqslant n-s \qquad \text{for } 1\leqslant s\leqslant m-1.\end{aligned}$$

Then (2.48) is immediate. □

Proof of Lemma 2.11. Part (i). If (C) holds, First, note that we can write $\tilde{Z}_n(s)=\sum_{i=1}^{n} v_{is}X_i$, where

$$v_{is}=\begin{cases}\frac{n-s}{n} & \text{if } 1\leqslant i\leqslant s\\ -\frac{s}{n} & \text{if } s+1\leqslant i\leqslant n\end{cases}.$$

Since $\xi_i=X_i-\mathbb{E}(X_i)$, we have $\tilde{Z}_n(s)-\mathbb{E}[\tilde{Z}_n(s)]=\sum_{i=1}^{n} v_{is}\xi_i$. By Lemma 2.8, we have

$$\begin{aligned}&\mathbb{P}\left(\max_{s\geqslant m+r}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{n} v_{is}\xi_{ij}\right|\geqslant 2\mathbb{E}\left[\max_{s\geqslant m+r}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{n} v_{is}\xi_{ij}\right|\right]+t\right)\\ &\quad\leqslant \exp\left(-\frac{t^2}{3\tau^2}\right)+3\exp\left(-\frac{t}{K_1\|M\|_{\psi_1}}\right),\end{aligned}$$

where

$$\tau^2 = \max_{s \geqslant m+r} \max_{1 \leqslant j \leqslant p} \sum_{i=1}^{n} v_{is}^2 \mathbb{E}(\xi_{ij}^2) \quad \text{and} \quad M = \max_{1 \leqslant i \leqslant n} \max_{s \geqslant m+r} \max_{1 \leqslant j \leqslant p} |v_{is}\xi_{ij}|.$$

Note that

$$\begin{aligned}
\tau^2 &\leqslant \bar{b} \max_{s \geqslant m+r} \sum_{i=1}^{n} v_{is}^2 = \frac{\bar{b}}{n} \max_{s \geqslant m+r} (n-s)s \leqslant \frac{\bar{b}n}{4}, \\
\|M\|_2 &= \left\| \max_{1 \leqslant i \leqslant n} \max_{1 \leqslant j \leqslant p} (\max_{s \geqslant m+r} |v_{is}|) |\xi_{ij}| \right\|_2 \leqslant K_2 \left\| \max_{1 \leqslant i \leqslant n} \max_{1 \leqslant j \leqslant p} (\max_{s \geqslant m+r} |v_{is}|) |\xi_{ij}| \right\|_{\psi_1} \\
&\leqslant K_2 \log(np) \max_{1 \leqslant i \leqslant n} \max_{1 \leqslant j \leqslant p} (\max_{s \geqslant m+r} |v_{is}|) \left\| \xi_{ij} \right\|_{\psi_1} \leqslant K_2 \bar{b} \log(np).
\end{aligned}$$

Using Lemma E.2 in [35] and $\log^3(np) \leqslant Kn$, we have

$$\begin{aligned}
\mathbb{E} \left[\max_{s \geqslant m+r} \max_{1 \leqslant j \leqslant p} \left| \sum_{i=1}^{n} v_{is} \xi_i \right| \right] &\leqslant K_3 \left\{ \sqrt{\log(np)} \tau + \log(np) \|M\|_2 \right\} \\
&\leqslant K_3 \left\{ \sqrt{\bar{b} n \log(np)} + \bar{b} \log^2(np) \right\} \\
&\leqslant C_1(\bar{b}, K) \sqrt{n \log(np)}.
\end{aligned}$$

Thus we get

$$\begin{aligned}
&\mathbb{P} \left(\max_{s \geqslant m+r} \max_{1 \leqslant j \leqslant p} \left| \sum_{i=1}^{n} v_{is} \xi_i \right| \geqslant C_1(\bar{b}, K) \sqrt{n \log(np)} + t \right) \\
&\leqslant \exp\left(-\frac{4t^2}{3\bar{b}n} \right) + 3 \exp\left(-\frac{t}{K_4 \bar{b} \log(np)} \right).
\end{aligned}$$

Choosing $t = C_2(\bar{b}, K)\sqrt{n \log(\gamma^{-1})}$ and using $\log(\gamma^{-1}) \leqslant K \log(np)$, we have

$$\mathbb{P}\left(\max_{s \geqslant m+r} \max_{1 \leqslant j \leqslant p} \left| \tilde{Z}_{nj}(s) - \mathbb{E}[\tilde{Z}_{nj}(s)] \right| \geqslant t^{\dagger} \right) \leqslant \gamma/36,$$

where $t^{\dagger} = C_3(\bar{b}, K)\sqrt{n \log(np)}$. Note that for any two sequences $\{a_i\}$ and $\{b_i\}$, we have by the elementary inequality $|\max_i |a_i| - \max_i |b_i|| \leqslant \max_i ||a_i| - |b_i|| \leqslant \max_i |a_i - b_i|$ that

$$\begin{aligned}
&\mathbb{P} \left(\left| \max_{s \geqslant m+r} \max_{j \in \mathcal{S}} \left| \tilde{Z}_{nj}(s) \right| - \max_{s \geqslant m+r} \max_{j \in \mathcal{S}} \left| \mathbb{E} \tilde{Z}_{nj}(s) \right| \right| \geqslant t^{\dagger} \right) \leqslant \gamma/24, \\
&\mathbb{P} \left(\max_{s \geqslant m+r} \max_{j \in \mathcal{S}^c} \left| \tilde{Z}_{nj}(s) \right| \geqslant t^{\dagger} \right) \leqslant \gamma/36.
\end{aligned}$$

Since

$$\max_{s\geqslant m+r}\max_{j\in\mathcal{S}}\left|\mathbb{E}\tilde{Z}_{nj}(s)\right| = \max_{s\geqslant m+r}\max_{j\in\mathcal{S}}\left|\frac{(n-s)m}{n}\delta_{nj}\right| = nt_m(1-t_m-t_r)|\delta_n|_\infty,$$

it follows that

$$|\delta_n|_\infty \geqslant \frac{C_4\log^{1/2}(np)}{n^{1/2}}$$

for some large enough constant $C_4 = C_4(\bar{b}, K, c_1, c_2) > 0$ implies that

$$\begin{aligned}\mathbb{P}(\mathcal{G}^c) \leqslant& \mathbb{P}\left(\max_{s\geqslant m+r}\max_{j\in\mathcal{S}}\left|\tilde{Z}_{nj}(s)\right| - \max_{s\geqslant m+r}\max_{j\in\mathcal{S}^c}\left|\tilde{Z}_{nj}(s)\right| \leqslant 0\right)\\ \leqslant& \mathbb{P}\left(\max_{s\geqslant m+r}\max_{j\in\mathcal{S}}\left|\tilde{Z}_{nj}(s)\right| \leqslant \max_{s\geqslant m+r}\max_{j\in\mathcal{S}}\left|\mathbb{E}\tilde{Z}_{nj}(s)\right| - t^\dagger\right)\\ &+ \mathbb{P}\left(-\max_{s\geqslant m+r}\max_{j\in\mathcal{S}^c}\left|\tilde{Z}_{nj}(s)\right| \leqslant -t^\dagger\right)\\ \leqslant& \gamma/36 + \gamma/36 = \gamma/18.\end{aligned}$$

In addition, following the same arguments,

$$\mathbb{P}\left(\max_{1\leqslant j\leqslant p}\left|\tilde{Z}_{nj}(m) - \mathbb{E}[\tilde{Z}_{nj}(m)]\right| \geqslant t^\dagger\right) \leqslant \gamma/36.$$

If $\delta_{nj} \leqslant 0, \forall 1 \leqslant j \leqslant p$, then $\mathbb{E}[\tilde{Z}_{nj}(m)] = nt_m(1-t_m)|\delta_{nj}| \geqslant 0$, and

$$\min_{1\leqslant j\leqslant p}\tilde{Z}_{nj}(m) \geqslant -t^\dagger, \quad \max_{1\leqslant j\leqslant p}\tilde{Z}_{nj}(m) \geqslant nt_m(1-t_m)|\delta_n|_\infty - t^\dagger \geqslant t^\dagger$$

with probability greater than $1-\gamma/36$ when $C_4 \geqslant 2C_3$. In other words, $\left|\max_{1\leqslant j\leqslant p}\tilde{Z}_{nj}(m)\right| \geqslant \left|\min_{1\leqslant j\leqslant p}\tilde{Z}_{nj}(m)\right|$, which implies $\max_{1\leqslant j\leqslant p}\left|\tilde{Z}_{nj}(m)\right| = \max_{1\leqslant j\leqslant p}\tilde{Z}_{nj}(m) = \tilde{Z}_{nj^*}(m) \geqslant 0$. Therefore,

$$\mathbb{P}\left(\max_{1\leqslant j\leqslant p}\left|\tilde{Z}_{nj}(m)\right| = \tilde{Z}_{nj^*}(m)\right) \geqslant 1-\gamma/36.$$

Part (ii). If (D) holds, By Lemma 2.8, we have

$$\begin{aligned}&\mathbb{P}\left(\max_{s\geqslant m+r}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^n v_{is}\xi_i\right| \geqslant 2\mathbb{E}\left[\max_{s\geqslant m+r}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^n v_{is}\xi_i\right|\right] + t\right)\\ &\leqslant \exp\left(-\frac{t^2}{3\tau^2}\right) + K_5\frac{\mathbb{E}\left[M^q\right]}{t^q},\end{aligned}$$

where τ^2 and M have the same definition as in Part (i). Since $\tau^2 \leqslant n\bar{b}/4$, we have

$$\|M\|_q^q = \mathbb{E}\left(\max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant p}(\max_{s\geqslant m+r}|v_{is}|^q)|\xi_{ij}|^q\right) \leqslant \sum_{i=1}^n \mathbb{E}\left(\max_{1\leqslant j\leqslant p}|\xi_{ij}|^q\right) \leqslant n\bar{b}^q,$$

which implies that $\|M\|_2 \leqslant \|M\|_q = n^{1/q}\bar{b}$ for $q \geqslant 2$. Using Lemma E.2 in [35] and $\log^3(np) \leqslant Kn$, we have

$$\mathbb{E}\left[\max_{s\geqslant m+r}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^n v_{is}\xi_i\right|\right] \leqslant K_4\left\{\sqrt{\bar{b}n\log(np)} + \bar{b}n^{1/q}\log(np)\right\}.$$

Thus we get

$$\begin{aligned}&\mathbb{P}\left(\max_{s\geqslant m+r}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^n v_{is}\xi_i\right| \geqslant C_5(\bar{b},K)\left\{\sqrt{n\log(np)} + n^{1/q}\log(np)\right\} + t\right)\\ &\leqslant \exp\left(-\frac{4t^2}{3\bar{b}n}\right) + K_5\frac{n\bar{b}^q}{t^q}.\end{aligned}$$

Choosing $t = C_6(\bar{b},K,q)\{\sqrt{n\log(\gamma^{-1})} + \gamma^{-1/q}n^{1/q}\}$ and using $\log(\gamma^{-1}) \leqslant K\log(np)$, we have

$$\mathbb{P}\left(\max_{s\geqslant m+r}\max_{1\leqslant j\leqslant p}\left|\tilde{Z}_{nj}(s) - \mathbb{E}[\tilde{Z}_{nj}(s)]\right| \geqslant t^\dagger\right) \leqslant \gamma/36,$$

where $t^\dagger = C_7(\bar{b},K,q)\{\sqrt{n\log(np)} + \gamma^{-1/q}n^{1/q}\log(np)\}$. If

$$|\delta_n|_\infty \geqslant \frac{C_8\log^{1/2}(np)}{n^{1/2}}\max\left\{1, \gamma^{-1/q}n^{1/q-1/2}\log^{1/2}(np)\right\},$$

for some large enough constant $C_8 = C_8(\bar{b},K,q,c_1,c_2) > 0$, then it follows from the same argument as in Part (i) that $\mathbb{P}(\mathcal{G}^c) \leqslant \gamma/18$. In addition, $\mathbb{P}\left(\max_{1\leqslant j\leqslant p}\left|\tilde{Z}_{nj}(m)\right| = \tilde{Z}_{nj^*}(m)\right) \geqslant 1-\gamma/36$. □

Proof of Lemma 2.12. *Part (i). Assume (C).* Consider $\mathcal{A}_n$ first. Apply our Lemma [A.1] to $a_{i,s_1,s_2} = \frac{1}{\sqrt{s_2-s_1+1}}\mathbf{1}_{\{s_1\leqslant i\leqslant s_2\}}$ (i.e. $s=(s_1,s_2)\in\mathbb{N}^2$) and $X_{ij} = \xi_{ij}$, we have for $\forall t > 0$

$$\begin{aligned}&\mathbb{P}(\max_{1\leqslant s_1\leqslant s_2\leqslant n}|\sum_{i=1}^n a_{i,s_1,s_2}\xi_i|_\infty \geqslant 2\mathbb{E}[\max_{1\leqslant s_1\leqslant s_2\leqslant n}|\sum_{i=1}^n a_{i,s_1,s_2}\xi_i|_\infty] + t)\\ &\leqslant \exp\left(-\frac{t^2}{3\tau^2}\right) + 3\exp\left(-\frac{t}{K_1\|M\|_{\psi_1}}\right),\end{aligned} \tag{2.66}$$

where

$$\begin{aligned}\tau^2 &= \max_{1\leqslant s_1\leqslant s_2\leqslant n}\max_{1\leqslant j\leqslant p}\sum_{i=1}^n a_{i,s_1,s_2}^2\mathbb{E}\xi_{ij}^2,\\ M &= \max_{1\leqslant i\leqslant n}\max_{1\leqslant s_1\leqslant s_2\leqslant n}\max_{1\leqslant j\leqslant p}|a_{i,s_1,s_2}\xi_{ij}|.\end{aligned}$$

Since $\tau^2 \leqslant \bar{b}$, and by [92, Lemma 2.2.2], we have $\|M\|_2 \leqslant 2\|M\|_{\psi_1} \leqslant C\bar{b}\log(n^3p) \leqslant C'\bar{b}\log(np)$. By [35, Lemma E.1], we have

$$\mathbb{E}[\max_{1\leqslant s_1\leqslant s_2\leqslant n}|\sum_{i=1}^n a_{i,s_1,s_2}\xi_i|_\infty] \leqslant K_2\{\tau\log^{1/2}(np) + \|M\|_2\log(np)\}. \tag{2.67}$$

Choosing $t = K_3\bar{b}\log(np)\log(\gamma^{-1})$ in (2.66) for some large enough universal constant $K_3 > 0$, we deduce that there exists a constant $C := C(\bar{b}, K) > 0$ such that

$$\mathbb{P}(\max_{1\leqslant s_1\leqslant s_2\leqslant n}|a_{i,s_1,s_2}\sum_{i=1}^s\xi_i|_\infty \geqslant K\bar{b}\log^2(np)) \leqslant \gamma.$$

So $\mathbb{P}(\mathcal{A}_n) \geqslant 1-\gamma$.

For $\mathcal{B}_n$, replacing $a_{i,s}$ by $a_{i,s_1,s_2} = \frac{1}{\sqrt{s_2-s_1+1}}\mathbf{1}_{\{s_1\leqslant i\leqslant s_2\}}$ in our [45] and [46], we have

$$\mathbb{P}(\max_{1\leqslant s_1\leqslant s\leqslant s_2\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty \geqslant K\bar{b}\log^2(np)) \leqslant \gamma.$$

for $\log(\gamma^{-1}) \leqslant K\log(np)$. Then, $\mathbb{P}(\mathcal{B}_n) \geqslant 1-\gamma$.

Part (ii). Assume (D). Assume (D) with $q \geqslant 2$. Similarly, apply Lemma [A.1] to a_{i,s_1,s_2} and $X_{ij} = \xi_{ij}$, we have for $\forall t > 0$

$$\begin{aligned}&\mathbb{P}(\max_{1\leqslant s_1\leqslant s_2\leqslant n}|\sum_{i=1}^n a_{i,s_1,s_2}\xi_i|_\infty \geqslant 2\mathbb{E}[\max_{1\leqslant s_1\leqslant s_2\leqslant n}|\sum_{i=1}^n a_{i,s_1,s_2}\xi_i|_\infty] + t)\\ &\qquad\leqslant \exp\left(-\frac{t^2}{3\tau^2}\right) + C(q)\frac{\mathbb{E}[M^q]}{t^q},\end{aligned} \tag{2.68}$$

where τ^2 and M have the same definitions as in Part (i). By [92, Lemma 2.2.2], we have $\|M\|_2 \leqslant \|M\|_q \leqslant n^{3/q}\bar{b}$. Note that $\tau^2 \leqslant \bar{b}$ and $\mathbb{E}[\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty]$ obeys the bound in (2.67). Hence, choosing $t = C(q)\{\bar{b}^{1/2}\log(\gamma^{-1}) + \bar{b}n^{3/q}\gamma^{-1/q}\}$ in (2.68), we get

$$\mathbb{P}\left(\max_{1\leqslant s\leqslant n}|Z_n(s)-\mathbb{E}[Z_n(s)]|_\infty \geqslant C(q)\bar{b}n^{3/q}(\log(np)+\gamma^{-1/q})\right) \leqslant 2\log^{-q}(np).$$

$\square$

Proof of Lemma 2.13. Denote the CUSUM mean computed on X_1^n as

$$\Delta_s = \mathbb{E}Z_n(s) = -\sum_{k=1}^{\nu}\left[\sqrt{\frac{s}{n(n-s)}}(n-m_k)\ \delta^{(k)}\ \mathbf{1}_{\{s\leqslant m_k\}} + \sqrt{\frac{(n-s)}{ns}}m_k\ \delta^{(k)}\ \mathbf{1}_{\{s>m_k\}}\right].$$

In particular, for $m_k < s \leqslant m_{k+1}$ of $k = 0,\ldots,\nu$,

$$\Delta_s = -\sqrt{\frac{s(n-s)}{n}}\left(\frac{m_0\delta^{(0)}+\cdots+m_k\delta^{(k)}}{s} + \frac{(n-m_{k+1})\delta^{(k+1)}+\cdots+(n-m_{\nu+1})\delta^{(\nu+1)}}{n-s}\right). \tag{2.69}$$

Let $\Delta(s)$ be defined as the same expression of (2.69) but on the whole real numbers $s \in (1, n)$.

Step 1: Suppose time of changes $\nu = 2$ and data is univariate $p = 1$.

$$f(s) := \frac{d}{ds}\Delta(s) = \begin{cases} -\frac{1}{2}\sqrt{\frac{n}{s(n-s)^3}}\left((n-m_1)\delta^{(1)}+(n-m_2)\delta^{(2)}\right), s<m_1 \\ \frac{1}{2}\sqrt{\frac{n}{s(n-s)}}\left(\frac{m_1}{s}\delta^{(1)} - \frac{n-m_2}{n-s}\delta^{(2)}\right), m_1<s<m_2 \\ \frac{1}{2}\sqrt{\frac{n}{s(n-s)^3}}\left(m_1\delta^{(1)}+m_2\delta^{(2)}\right), s>m_2 \end{cases} \tag{2.70}$$

(i) Suppose $\mathrm{sign}(\delta^{(1)}) = \mathrm{sign}(\delta^{(2)}) \neq 0$, then the sign of the first derivative of $\Delta(s)$ is summarized as the table below, where s_0 satisfies $s_0/(n-s_0) = (m_1\delta^{(1)})/((n-m_2)\delta^{(2)})$ when $|\delta^{(1)}|/|\delta^{(2)}| \in (\frac{n-m_2}{n-m_1}, \frac{m_2}{m_1})$.

<table>
<tr><th colspan="2">$\mathrm{sign}(f(s))$</th><th>$s<m_1$</th><th colspan="2">$m_1<s<m_2$</th><th>$m_2<s$</th></tr>
<tr><td rowspan="4">$\frac{|\delta^{(1)}|}{|\delta^{(2)}|}$</td><td>$(0, \frac{n-m_2}{n-m_1})$</td><td>$-\mathrm{sign}(\delta^{(1)})$</td><td colspan="3">$\mathrm{sign}(\delta^{(1)})$</td></tr>
<tr><td rowspan="2">$(\frac{n-m_2}{n-m_1}, \frac{m_2}{m_1})$</td><td rowspan="2">$-\mathrm{sign}(\delta^{(1)})$</td><td>$m_1<s<s_0$</td><td>$s_0<s<m_2$</td><td rowspan="2">$\mathrm{sign}(\delta^{(1)})$</td></tr>
<tr><td>$\mathrm{sign}(\delta^{(1)})$</td><td>$-\mathrm{sign}(\delta^{(1)})$</td></tr>
<tr><td>$(\frac{m_2}{m_1}, +\infty)$</td><td colspan="3">$-\mathrm{sign}(\delta^{(1)})$</td><td>$\mathrm{sign}(\delta^{(1)})$</td></tr>
</table>

Observing that $\Delta(s)$ is continuous and always with sign of $-\mathrm{sign}(\delta^{(1)})$. The maximum of $|\Delta(s)|$ locates at either m_1 or m_2.

(ii) Suppose $\mathrm{sign}(\delta^{(1)}) = -\mathrm{sign}(\delta^{(2)}) \neq 0$, then based on the summarized $\mathrm{sign}(f(s))$ in the table below and the fact that $\Delta(s)$ is continuous and always with sign $-\mathrm{sign}(\delta^{(1)})$, we claim the maximum of $|\Delta(s)|$ locates at either $s = m_1$ or $s = m_2$.

<table>
<tr><th colspan="2">$\text{sign}(f(s))$</th><th>$s < m_1$</th><th>$m_1 < s < m_2$</th><th>$m_2 < s$</th></tr>
<tr><td rowspan="3">$\frac{|\delta^{(1)}|}{|\delta^{(2)}|}$</td><td>$(0, \frac{n-m_2}{n-m_1})$</td><td colspan="2">$\text{sign}(\delta^{(1)})$</td><td>$-\text{sign}(\delta^{(1)})$</td></tr>
<tr><td>$(\frac{n-m_2}{n-m_1}, \frac{m_2}{m_1})$</td><td>$-\text{sign}(\delta^{(1)})$</td><td>$\text{sign}(\delta^{(1)})$</td><td>$-\text{sign}(\delta^{(1)})$</td></tr>
<tr><td>$(\frac{m_2}{m_1}, +\infty)$</td><td>$-\text{sign}(\delta^{(1)})$</td><td colspan="2">$\text{sign}(\delta^{(1)})$</td></tr>
</table>

Step 2: Generalization to multiple change points $\nu \geqslant 2$ and multiple dimension $p \geqslant 1$. Based on arguments in Step 1, for any $s \in (m_k, m_{k+1})$ of $k = 1, \ldots, \nu - 1$, the maximum of $Z_n(s)$ happens at boundaries of intervals, i.e. $\{m_1, \ldots, m_\nu\}$. Notice that when $s \in (m_k, m_{k+1})$ for $k = 0, \nu$, it reduces to the case of single change point where $Z_n(s)$ reaches its maximum at m_1 or m_ν. In addition,

$$\underset{s=1,\ldots,n-1}{\text{argmax}} |\Delta_s|_\infty \subset \bigcup_{j=1,\ldots,p} \underset{s}{\text{argmax}} |\Delta_j(s)| \subset \{m_k, k = 1, \ldots, \nu\}$$

Therefore, $\text{argmax}_{s=1,\ldots,n-1} |\Delta_s|_\infty \subset \{m_k, k = 1, \ldots, \nu\}$ for $\nu \geqslant 1$ and $p \geqslant 1$.

Step 3: Characterization of signal size $\max_{k=1,\ldots,\nu} |\Delta_{m_k}|_\infty$. Suppose $p = 1$. Let $\nu \geqslant 2$ otherwise it reduces to single change point case which obeys the lower bound in Lemma (see (2.43) in the proof of Theorem 2.3). Denote $L = \sum_{k=1}^{l} m_k \delta^{(k)}, R = \sum_{k=l+1}^{\nu} (n - m_k)\delta^{(k)}$ for some $1 \leqslant l \leqslant \nu$. Then by (2.69),

$$\begin{cases} \Delta_{m_l} = -\frac{1}{\sqrt{n}} \underbrace{\left[\sqrt{\frac{n - m_l}{m_l}} L + \sqrt{\frac{m_l}{n - m_l}} R\right]}_{:=A} \\ \Delta_{m_{l+1}} = -\frac{1}{\sqrt{n}} \underbrace{\left[\sqrt{\frac{n - m_{l+1}}{m_{l+1}}} L + \sqrt{\frac{m_{l+1}}{n - m_{l+1}}} R\right]}_{:=B} \end{cases}.$$

Since A and B cannot be 0 at the same time, we will show that (i) when L and R have the same sign, $|A|$ itself is large and (ii) when L and R have the opposite sign, $|A - B|$ is still large even if signal cancellation exists. Therefore, $\max\{O(|A|), O(|B|)\}) = O(|A - B|)$ can be bounded below in either case, and so does $\max_{l=1,\ldots,\nu} |\Delta_{m_l}|$.

First, suppose L and R have the opposite sign. WLOG, let $L > 0$.

$$\begin{aligned} |A - B| &= \left(\sqrt{\frac{n - m_l}{m_l}} - \sqrt{\frac{n - m_{l+1}}{m_{l+1}}}\right) |L| + \left(\sqrt{\frac{m_l}{n - m_l}} - \sqrt{\frac{m_{l+1}}{n - m_{l+1}}}\right) |R| \\ &= \frac{n(m_{l+1} - m_l)/\sqrt{m_{l+1} m_l}}{\sqrt{m_{l+1}(n - m_l)} \sqrt{m_{l1}(n - m_{l+1})}} |L| + \frac{n(m_{l+1} - m_l)/\sqrt{(n - m_{l+1})(n - m_l)}}{\sqrt{m_{l+1}(n - m_l)} \sqrt{m_{l1}(n - m_{l+1})}} |R| \\ &\geqslant \frac{m_{l+1} - m_l}{2m_{l+1}} \frac{n}{\sqrt{m_l(n - m_l)}} |L| + \frac{m_{l+1} - m_l}{2(n - m_l)} \frac{n}{\sqrt{m_{l+1}(n - m_{l+1})}} |R| \end{aligned}$$

Denote $\underline{\delta} = \min_{k=1,\ldots,\nu} |\delta^{(k)}|$. Claim that we can find an l such that $\max\{|L|, |R|\} \geqslant C_1 D_v \underline{\delta}$ for some constant C_1. Otherwise, we just need to move l to $l+1$ such that L and R change $m_{l+1}\delta^{(l+1)} \geqslant D_\nu \underline{\delta}$ and $(n - m_{l+1})\delta^{(l+1)} \geqslant D_\nu \underline{\delta}$, respectively, in opposite direction so that they still have opposite signs. (Move l to $l-1$ if l is the last change point.) Note that $\max\{\frac{m_{l+1}-m_l}{2m_{l+1}}, \frac{m_{l+1}-m_l}{2(n-m_l)}\} \geqslant \frac{D_\nu}{n}$. Consequently, we have $|A - B| \geqslant C_1 \frac{D_\nu^2 \underline{\delta}}{\sqrt{m_l(n-m_l)}}$. Next, suppose L and R have the same sign. We still have $\max\{|L|, |R|\} \geqslant C_2 D_v \underline{\delta}$ for some constant C_2. Otherwise, we move l to $l+1$ and it reduces to the case above. Then, $|A| \geqslant C_2(\sqrt{\frac{n-m_l}{m_l}} \wedge \sqrt{\frac{m_l}{n-m_l}}) D_v \underline{\delta} \geqslant C_2 \frac{D_\nu^2 \underline{\delta}}{\sqrt{m_l(n-m_l)}}$. Therefore, when $p = 1$, there exists some constant C such that $\max_{l=1,\ldots,\nu} |\Delta_{m_l}| \geqslant \max_{l=1,\ldots,\nu} \frac{C D_\nu^2 \underline{\delta}}{\sqrt{n m_l(n-m_l)}}$.

For the case of multiple dimension $p > 1$, recall $\bar{\delta}_n = \min_{k=1,\ldots,\nu} |\delta_n^{(k)}|_\infty$. Since for the dimension $h = \mathrm{argmax}_{j=1,\ldots,p} |\delta_{n,j}^{(l+1)}|$, when l moves to $l+1$ the terms, L_h and R_h changes in opposite direction with distances at least $D_v |\delta_h^{(l+1)}| \geqslant D_v \bar{\delta}_n$. Then, $\max\{|L|_\infty, |R|_\infty\} \geqslant C_3 D_v \bar{\delta}_n$ still holds. So based on similar argument, we can conclude that

$$\max_{l=1,\ldots,\nu} |\Delta_{m_l}|_\infty \geqslant \max_{l=1,\ldots,\nu} \frac{C D_\nu^2 \bar{\delta}_n}{\sqrt{n m_l(n - m_l)}}.$$

□

2.6.8 Additional simulation results

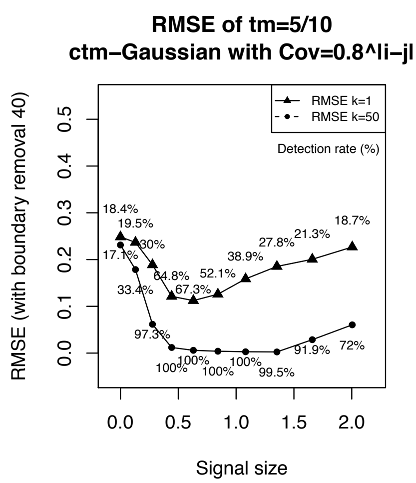


Figure 2.15: Same setting as in 2.5(b) except $\alpha = 0.01$.

Our method	Gaussian			t_6			ctm-Gaussian		
$\lvert\delta\rvert_\infty = \delta_1$	I	II	III	I	II	III	I	II	III
$t_m = 5/10$									
0	0.031	0.038	0.036	0.020	0.044	0.016	0.015	0.042	0.027
0.13	0.035	0.043	0.034	0.020	0.044	0.022	0.014	0.047	0.027
0.28	0.098	0.295	0.128	0.042	0.137	0.038	0.026	0.126	0.050
0.44	0.662	0.884	0.677	0.296	0.559	0.279	0.235	0.567	0.280
0.63	0.989	1	0.993	0.831	0.939	0.851	0.797	0.958	0.843
0.84	1	1	1	0.990	1	0.997	0.997	1	0.996
$t_m = 3/10$									
0	0.042	0.056	0.034	0.024	0.044	0.023	0.020	0.049	0.018
0.13	0.043	0.060	0.034	0.022	0.047	0.027	0.021	0.047	0.018
0.28	0.087	0.209	0.082	0.03	0.109	0.034	0.033	0.093	0.029
0.44	0.502	0.756	0.513	0.181	0.477	0.192	0.186	0.431	0.187
0.63	0.966	0.996	0.972	0.652	0.919	0.711	0.675	0.890	0.690
0.84	1	1	1	0.981	0.997	0.979	0.977	0.999	0.987
1.08	1	1	1	1	1	0.999	0.999	1	1
$t_m = 1/10$									
0	0.039	0.046	0.037	0.023	0.052	0.032	0.018	0.058	0.017
0.13	0.036	0.045	0.035	0.025	0.053	0.033	0.025	0.052	0.016
0.28	0.047	0.070	0.0400	0.024	0.057	0.035	0.025	0.053	0.020
0.44	0.094	0.253	0.091	0.036	0.121	0.049	0.031	0.108	0.032
0.63	0.370	0.645	0.432	0.119	0.344	0.147	0.123	0.310	0.109
0.84	0.861	0.948	0.861	0.426	0.75	0.503	0.418	0.707	0.406
1.08	0.998	1	0.997	0.870	0.967	0.860	0.826	0.953	0.847
1.35	1	1	1	0.991	0.998	0.989	0.988	0.998	0.992

Table 2.9: Power report of our method for sparse alternative where $\alpha = 0.05$, $t_m = 5/10, 3/10, 1/10$. Here, $n = 500, p = 600$ and the boundary removal is 40.

B_n	Gaussian			t_6			ctm-Gaussian		
$\lvert\delta\rvert_\infty = \delta_1$	I	II	III	I	II	III	I	II	III
				$t_m = 5/10$					
0	0.053	0.062	0.043	0.070	0.049	0.050	0.061	0.053	0.060
0.13	0.063	0.077	0.055	0.068	0.049	0.051	0.058	0.058	0.071
0.28	0.219	0.418	0.210	0.147	0.219	0.110	0.111	0.210	0.109
0.44	0.821	0.946	0.809	0.502	0.750	0.504	0.433	0.706	0.429
0.84	1	1	1	1	1	1	0.998	1	1
				$t_m = 3/10$					
0	0.061	0.052	0.053	0.056	0.053	0.061	0.067	0.056	0.056
0.13	0.053	0.064	0.063	0.061	0.055	0.061	0.072	0.059	0.063
0.28	0.110	0.232	0.110	0.081	0.132	0.073	0.076	0.121	0.080
0.44	0.553	0.828	0.561	0.263	0.537	0.257	0.245	0.469	0.257
0.84	1	1	1	0.996	1	0.996	0.987	1	0.988
				$t_m = 1/10$					
0	0.054	0.073	0.063	0.066	0.060	0.059	0.056	0.057	0.063
0.13	0.062	0.069	0.062	0.070	0.061	0.055	0.049	0.058	0.069
0.44	0.058	0.084	0.061	0.073	0.068	0.065	0.050	0.054	0.071
0.84	0.189	0.520	0.243	0.111	0.246	0.101	0.079	0.229	0.098
1.08	0.684	0.925	0.730	0.264	0.615	0.302	0.210	0.556	0.252
1.35	0.986	1	0.985	0.703	0.940	0.724	0.620	0.903	0.670
1.66	1	1	1	0.99	1	0.986	0.962	0.997	0.968

Table 2.10: Power report of B_n in [65] for sparse alternative where $\alpha = 0.05$, $t_m = 5/10, 3/10, 1/10$. Here, $n = 500, p = 600$ and the boundary removal is 40.

ψ	$\underline{s} = 1$			$\underline{s} = 40$		
$\lvert\delta\rvert_\infty$	$t_m = 5/10$	$t_m = 3/10$	$t_m = 1/10$	$t_m = 5/10$	$t_m = 3/10$	$t_m = 1/10$
			sparse H_1: Gaussian (I)			
0	0.107	0.101	0.116	0.074	0.059	0.701
0.13	0.107	0.105	0.119	0.075	0.062	0.705
0.28	0.126	0.115	0.125	0.096	0.075	0.710
0.44	0.177	0.154	0.136	0.130	0.106	0.720
0.63	0.312	0.248	0.157	0.265	0.215	0.735
0.84	0.625	0.483	0.195	0.604	0.478	0.764
1.08	0.943	0.839	0.307	0.941	0.850	0.800
1.35	1	0.997	0.534	1	0.997	0.843
1.66	1	1	0.846	1	1	0.923
2	1	1	0.992	1	1	0.990
			dense H_1: Gaussian (I)			
0	0.110	0.116	0.120	0.055	0.067	0.074
0.13	0.755	0.608	0.233	0.735	0.573	0.169
0.28	1	1	0.977	1	1	0.973

Table 2.11: Power report of ψ in [43] for both sparse and dense Gaussian alternative where $\alpha = 0.05$, $t_m = 5/10, 3/10, 1/10$ and spatial dependence structure (I). Here, $n = 500, p = 600$.

$\lvert\delta\rvert_\infty = \delta_1$	Temporal Independent: Gaussian (I) $\theta = 0$	Gaussian (I) $\theta = 1/2$	t_6 (II) $\theta = 0$	t_6 (II) $\theta = 1/2$	ctm-Gaussian (III) $\theta = 0$	ctm-Gaussian (III) $\theta = 1/2$	Dependent TS: ctm-Gaussian (III) $\theta = 0$	TS: ctm-Gaussian (III) $\theta = 1/2$
$\underline{s} = 1, t_m = 5/10$								
0	0.116	0.382	0.167	0.407	0.124	0.431	0.124	0.429
0.13	0.115	0.380	0.164	0.405	0.124	0.430	0.124	0.429
0.28	0.101	0.346	0.118	0.354	0.115	0.422	0.116	0.423
0.44	0.054	0.175	0.067	0.229	0.082	0.350	0.088	0.362
0.63	0.026	0.033	0.037	0.099	0.039	0.178	0.044	0.207
0.84	0.015	0.015	0.020	0.035	0.020	0.054	0.024	0.060
1.35	0.005	0.005	0.008	0.009	0.008	0.010	0.011	0.011
2.00	0.003	0.002	0.004	0.004	0.004	0.004	0.004	0.004
$\underline{s} = 1, t_m = 1/10$								
0	0.416	0.545	0.436	0.552	0.423	0.575	0.416	0.573
0.13	0.416	0.544	0.435	0.550	0.423	0.575	0.416	0.572
0.28	0.416	0.539	0.433	0.539	0.422	0.573	0.416	0.571
0.44	0.413	0.498	0.414	0.488	0.421	0.567	0.414	0.562
0.63	0.394	0.353	0.349	0.350	0.415	0.525	0.410	0.525
0.84	0.318	0.133	0.262	0.187	0.387	0.417	0.388	0.433
1.35	0.081	0.006	0.116	0.032	0.178	0.065	0.193	0.112
2.00	0.039	0.002	0.058	0.005	0.060	0.004	0.062	0.005
$\underline{s} = 40, t_m = 5/10$								
0	0.119	0.280	0.168	0.296	0.124	0.290	0.123	0.286
0.13	0.118	0.279	0.162	0.292	0.124	0.290	0.122	0.285
0.28	0.103	0.245	0.116	0.237	0.115	0.276	0.114	0.272
0.44	0.055	0.126	0.068	0.132	0.077	0.204	0.088	0.215
0.63	0.022	0.033	0.037	0.066	0.035	0.082	0.044	0.090
0.84	0.013	0.017	0.020	0.026	0.019	0.028	0.024	0.034
1.35	0.005	0.006	0.008	0.009	0.008	0.009	0.009	0.010
2.00	0.002	0.002	0.003	0.004	0.004	0.004	0.005	0.004
$\underline{s} = 40, t_m = 1/10$								
0	0.422	0.484	0.434	0.503	0.426	0.505	0.412	0.490
0.13	0.422	0.483	0.433	0.502	0.425	0.505	0.412	0.490
0.28	0.422	0.476	0.430	0.489	0.425	0.503	0.412	0.487
0.44	0.418	0.431	0.409	0.400	0.424	0.483	0.411	0.473
0.63	0.396	0.306	0.355	0.257	0.415	0.407	0.405	0.425
0.84	0.326	0.125	0.268	0.115	0.383	0.270	0.386	0.284
1.35	0.072	0.005	0.116	0.010	0.175	0.030	0.190	0.042
2.00	0.039	0.003	0.057	0.004	0.062	0.005	0.066	0.005

Table 2.12: RMSE of our estimators $\hat{m}_0$ and $\hat{m}_{1/2}$. Both truncated and non-truncated versions are implemented under $t_m = 5/10$ and $t_m = 1/10$.

	Temporal Independent						Dependent	
	Gaussian (I)		t_6 (II)		ctm-Gaussian (III)		TS: ctm-Gaussian (III)	
$\|\delta\|_\infty$	$k=1$	$k=50$	$k=1$	$k=50$	$k=1$	$k=50$	$k=1$	$k=50$
$t_m = 5/10$								
0	0.166	0.166	0.364	0.364	0.206	0.206	0.190	0.190
0.13	0.164	0.079	0.363	0.340	0.206	0.167	0.189	0.150
0.28	0.139	0.005	0.354	0.251	0.195	0.061	0.182	0.061
0.44	0.080	0.001	0.315	0.132	0.167	0.018	0.159	0.018
0.63	0.034	0.000	0.262	0.061	0.120	0.007	0.123	0.007
0.84	0.018	0.000	0.211	0.027	0.073	0.004	0.076	0.004
1.35	0.006	0.000	0.132	0.008	0.015	0.001	0.021	0.001
2.00	0.003	0.000	0.075	0.003	0.005	0.001	0.006	0.001
$t_m = 1/10$								
0	0.425	0.425	0.555	0.555	0.446	0.446	0.447	0.447
0.13	0.424	0.394	0.555	0.544	0.445	0.434	0.446	0.436
0.28	0.420	0.211	0.555	0.504	0.443	0.374	0.444	0.384
0.44	0.405	0.004	0.548	0.424	0.439	0.224	0.442	0.250
0.63	0.351	0.001	0.527	0.309	0.428	0.055	0.434	0.061
0.84	0.255	0.000	0.487	0.197	0.407	0.006	0.418	0.006
1.35	0.031	0.000	0.390	0.047	0.289	0.002	0.334	0.002
2.00	0.003	0.000	0.293	0.006	0.116	0.001	0.146	0.001

Table 2.13: RMSE of estimator in [99].

	Temporal Independent						Dependent	
	Gaussian (I)		t_6 (II)		ctm-Gaussian (III)		TS: ctm-Gaussian (III)	
$\|\delta\|_\infty$	$k=1$	$k=50$	$k=1$	$k=50$	$k=1$	$k=50$	$k=1$	$k=50$
$t_m = 5/10$								
0	0.207	0.221	0.246	0.240	0.249	0.243	0.243	0.229
0.13	0.201	0.113	0.246	0.205	0.244	0.210	0.237	0.208
0.28	0.187	0.011	0.172	0.134	0.224	0.086	0.234	0.098
0.44	0.131	0.004	0.107	0.072	0.174	0.017	0.185	0.024
0.63	0.116	0.003	0.072	0.036	0.137	0.008	0.137	0.009
0.84	0.119	0.002	0.061	0.028	0.130	0.006	0.123	0.006
1.35	0.126	0.001	0.066	0.023	0.153	0.003	0.135	0.003
2	0.140	0.001	0.069	0.032	0.152	0.001	0.161	0.001
$t_m = 1/10$								
0	0.379	0.381	0.433	0.431	0.440	0.430	0.430	0.419
0.13	0.392	0.316	0.444	0.400	0.436	0.374	0.409	0.416
0.28	0.365	0.099	0.426	0.321	0.415	0.326	0.414	0.314
0.44	0.355	0.005	0.360	0.194	0.405	0.126	0.394	0.159
0.63	0.267	0.002	0.202	0.106	0.348	0.026	0.363	0.017
0.84	0.164	0.001	0.092	0.043	0.272	0.005	0.278	0.005
1.35	0.056	0.000	0.034	0.028	0.105	0.002	0.105	0.002
2	0.040	0.000	0.036	0.023	0.078	0.001	0.061	0.001

Table 2.14: RMSE of estimator in [37].

Chapter 3

Robust bootstrap change point test for high-dimensional location parameter

3.1 Introduction

Change point detection problems are commonly seen in many statistical and scientific areas including functional data analysis [7, 4], time series inspection [8, 62, 105], panel data study [36, 88, 61, 10], with applications to fields of biomedical engineering [5, 111], genomics [100], financial revenue returns [6, 37, 10] among many others. Statistical testing and estimation of change points have long history and extensive literature [41, 8, 59, 6, 11, 72, 71]. 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 $X_i \sim F_i, i = 1, \dots, n$ 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:

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

The greatest 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 $Y_j = X_{m+j}, j = 1, \dots, n-m$. With this notation, we have $X_1, \dots, X_m$ that are independent and identically distributed (i.i.d.) with distribution F and $Y_1, \dots, Y_{n-m}$ 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 $X_i, i = 1, \dots, n$, 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:

$$X_i = \theta\ \mathbf{1}(i > m) + \xi_i,\ \ i = 1, \ldots, n,\ \text{ where } \xi_i \overset{i.i.d.}{\sim} F \text{ are random vectors in } \mathbb{R}^p\ . \tag{3.1}$$

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:

$$U_n = \sum_{1 \leqslant i < j \leqslant n} (X_i - X_j). \tag{3.2}$$

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 and H_1.

In practice, a main concern for using U_n in (3.2) 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 (3.2) 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

$$T_n = T_n(X_1^n) = n^{1/2} \binom{n}{2}^{-1} \sum_{1 \leqslant i < j \leqslant n} h(X_i, X_j) \tag{3.3}$$

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 $j = 1, \dots, p$, $\text{sign}_j(x) = \text{sign}(x_j) = -1, 0, 1$ 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 $\phi_j, j = 1, \dots, p$, 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 (3.2) 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 [58]. 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 [93]. In addition, when the change point location m is known, T_n is also equivalent to the classical nonparametric Mann-Whitney U test statistic (see e.g., Chapter 12 in [91]).

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

$$\overline{T}_n := |T_n|_\infty = \max_{1 \leqslant k \leqslant d} |T_{nk}|. \tag{3.4}$$

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 3.2, 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 3.3.

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 [38] 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

$$Z_n(s) = \left(\frac{s(n-s)}{n}\right)^{1/2} \left(\frac{1}{s}\sum_{i=1}^{s} X_i - \frac{1}{n-s}\sum_{i=s+1}^{n} X_i\right), \quad s = 1, \ldots, n-1. \tag{3.5}$$

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 (3.3), 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 $O(n^2 p)$ 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 $O(n^2 p)$ which can reduces to $O(np)$ [66] via dynamic programming. Statistically, it has been shown in [106, 65] 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 13 ahead for a detailed comparison.

3.1.1 Literature review and our contribution

Single change point inference has been extensively studied in literature such as [38, 51, 60] for univariate or fixed multivariate setting. 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, [65, 106, 99] dealt with sparse change (i.e. mean structure changes in a sparse subset of coordinates), while [10, 61, 43] considered ℓ^2-type aggregation for dense change. To take both cases into account, [43] proposed a scan test statistic aiming to sparser change coupled with their linear statistic in inference, and [36] 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 [43, 74, 103]. We show that our result achieves optimal minimax rate, c.f. Remark 15. For multiple change point detection which is more challenging and essential in real application, we will discuss a *backward detection* (BD) algorithm without introducing external statistics.

Among the change point literature, mean change are widely explored using CUSUM statistics [65, 106, 36, 37], least-square type statistics [10, 14], U-statistics [98] and some other kernel based methods [82, 26, 3]. In practice where error terms are heavy-tailed, Gaussianity assumption is beyond salvation and becomes too restrictive. This concern especially highlights the potential of robust nonparametric methodology (such as nonlinear projection) to avoid direct measure on mean or higher moments. 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, [82] proposed CUSUM variant statistic based on kernel transferred data points; [26] smoothed left and right mean function using kernel density estimation; [3] 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 chapter proceeds as follows. The bootstrap calibration for the distribution of $\overline{T}_n$ is described in Section 3.2. Main results for size validity and power properties of the bootstrap test are derived in Section 3.3. Extension to multiple change point scenario are elaborated in Section 3.4. We report simulation study results in Section 3.5 and real data examples in Section 3.6. All proofs with auxiliary lemmas are given in Section 3.7.

3.1.2 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 [70]. 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 chapter, we assume $d \geqslant 2$ to simplify some statements and all inference works for $d = 1$.

3.2 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

$$T_n^\sharp = n^{1/2}\binom{n}{2}^{-1}\sum_{i=1}^{n}\left\{\sum_{j=i+1}^{n} h(X_i,X_j)\right\}e_i \quad \text{and} \quad \overline{T}_n^\sharp := |T_n^\sharp|_\infty = \max_{1\leqslant k\leqslant d}|T_{nk}^\sharp|. \tag{3.6}$$

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 3.3, 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

$$T_n = \underbrace{n^{-1/2}\sum_{i=1}^{n}\frac{2(n-2i+1)}{n-1}g(X_i)}_{L_n} + \underbrace{n^{1/2}\binom{n}{2}^{-1}\sum_{1\leqslant i<j\leqslant n} f(X_i,X_j)}_{R_n}. \tag{3.7}$$

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 [35, 27]. 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 [27] 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, 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 (3.7). 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 $\overline{T}^\sharp_n$), 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:

$$V_n = \sum_{i=1}^{m} \sum_{j=1}^{n-m} h(X_i, Y_j) = m(n-m)\theta_h + (n-m) \sum_{i=1}^{m} Gh(X_i) + m \sum_{j=1}^{n-m} Fh(Y_j) + \sum_{i=1}^{m} \sum_{j=1}^{n-m} \breve{f}(X_i, Y_j). \quad (3.8)$$

Since the last three sums on the r.h.s. of (3.8) 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 $\overline{T}^\sharp_n$ would tend to one.

Remark 12. It is interesting to note that our bootstrapped U-statistic $T^\sharp_n$ in (3.6) is closely related to the jackknife multiplier bootstrap (JMB) proposed in [27] for high-dimensional U-statistics and in [28] 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. □

3.3 Theoretical properties

Let $X, X' \in \mathbb{R}^p$ 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 (3.7). 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 $\Gamma = \text{Cov}(g(X)) = \mathbb{E}[g(X)^T g(X)]$. 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 $h(X, X')$.

3.3.1 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. Condition (A2) and (A3) impose mild moment conditions on the kernel h coupled with the distribution F. 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*

$$\rho(\overline{T}_n, \overline{T}_n^\sharp \mid X_1^n) := \sup_{t \in \mathbb{R}} \left| \mathbb{P}(\overline{T}_n \leqslant t) - \mathbb{P}(\overline{T}_n^\sharp \leqslant t \mid X_1^n) \right| \leqslant C \varpi_n \tag{3.9}$$

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

$$\varpi_n = \left\{ \frac{D_n^2 \log^7(nd)}{n} \right\}^{1/6}. \tag{3.10}$$

Consequently, we have

$$\sup_{\alpha\in(0,1)}\left|\mathbb{P}(\overline{T}_n \leqslant q_{\overline{T}^\sharp_n|X_1^n}(\alpha)) - \alpha\right| \leqslant C\varpi_n + \gamma. \tag{3.11}$$

In particular, if $\log d = o(n^{1/7})$*, then* $\mathbb{P}(\overline{T}_n \leqslant q_{\overline{T}^\sharp_n|X_1^n}(\alpha)) \to \alpha$ *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 $\overline{T}^\sharp_n|X_1^n$ 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 extend the Gaussian approximation theory for U-statistics in [27], which focus on symmetric kernels.

Remark 13 (Comparisons with the CUSUM-based statistics). [65] and [106] 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 in (2.4) considered in [106], 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 $\overline{S}^\sharp_n := \max_{\underline{s}\leqslant s\leqslant n-\underline{s}} |Z_n^\sharp(s)|_\infty$ shown in [106] obeys

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

with probability at least $1-\gamma$. Comparing the last display with the rate of convergence for $\rho(\overline{T}_n, \overline{T}^\sharp_n \mid X_1^n)$ in (3.9) and (3.10), we see that the JMB method proposed here has better statistical properties than the Gaussian multiplier bootstrap $\overline{T}^\sharp_n$ without removing any boundary points in computing $\overline{T}_n$ and $\overline{T}^\sharp_n$. Consequently this will reduce the error-in-size (3.11) for our bootstrap calibration $\overline{T}^\sharp_n$. Empirical evidence for our algorithm with smaller error-in-size can be found in Section 3.5. 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 $\overline{T}^\sharp_n$ can be viewed as a *nonlinear* and *one-pass* version of the CUSUM statistics. □

3.3.2 Power analysis

Next, we analyze the power of 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 (3.1). Condition (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, Condition (B2) and (B3) separate the location-shift signal from the mean-zero noise, and if $\theta = 0$, then Condition (B2) and (B3) reduce Condition (A2) and (A3). Our next theorem characterizes the minimal signal strength for detecting the change point under the alternative hypothesis H_1.

Theorem 3.2 (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*

$$m(n-m)|\theta_h|_\infty > K_0 D_n n^{3/2} \log^{1/2}(nd/\alpha) + C_1(\underline{b}) n^{3/2} \log^{1/2}(\zeta^{-1}) \log^{1/2}(d), \tag{3.12}$$

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

Remark 14 (Interpretation of Theorem 3.2). Note the first term on r.h.s. of (3.12) 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 [106] (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 15 (Rate optimality for sparse alternative). In [74, Theorem 1], the authors derive 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 $s = |\theta|_0 < \sqrt{p \log\log(8n)}$, let $|\theta_h|_2^2 \approx |\theta|_2^2 = k\rho^2$ under local alternative, then their minimax result reads as

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

which indicates $\rho \gtrsim (m \wedge n-m)^{-1/2}$ up to a logarithm factor. Note that $m(n-m) \asymp n(m \wedge n-m)$ since $(m \vee n-m)$ is between $[n/2, n]$, so our (3.12) in Theorem 3.2 provides the lower bound

$$\rho \gtrsim C(D_n, \underline{b}, \alpha)(m \wedge n-m)^{-1} n^{1/2} \log^{1/2}(nd).$$

If $(m \wedge n-m)$ is bounded away from boundaries, i.e. $m \asymp n-m \asymp n$, then our result is minimax optimal. □

3.4 Extension to multiple change points scenario

3.4.1 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

$$H_1^{'} : \theta^{(k)} \neq \theta^{(k+1)} \text{ for some } m_k, k = 0, \ldots, \nu \text{ and } \nu \geqslant 1. \tag{3.13}$$

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

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

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

$$\mathbb{E}\left[\sum_{1\leqslant i<j\leqslant n} h(X_i, X_j)\right] = \sum_{0\leqslant k<k'\leqslant \nu} s_k s_{k'} \delta^{(k,k')} =: \tilde{\Delta}, \tag{3.14}$$

where the standardized signal strength is $|E[T_n]|_\infty = n^{1/2}\binom{n}{2}^{-1}|\tilde{\Delta}|_\infty$. Under the multiple 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 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 $\log(1/\zeta) \leqslant K \log(\nu^2 nd)$ for some constant $K > 0$. Suppose ν is a constant. If*

$$|\tilde{\Delta}|_\infty > K_0 \nu^2 D_n n^{3/2} \log^{1/2}(nd/\alpha) + C_1(\underline{b}) n^{3/2} \log^{1/2}(\zeta^{-1}) \log^{1/2}(d) + \phi, \tag{3.15}$$

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 16 (Explanation on ϕ and connection to single change point case). Compared to (3.12) in Theorem 3.2, there is an additional content ϕ in (3.15). 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 (3.13), 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 (3.15). Then our result under H_1' reads the same as (3.12). □

The l.h.s. of (3.15) 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 [36, 37, 106]. Taking (3.5) 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 (3.14) 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 3.3 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 (3.15).

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 r.h.s. of (3.15). 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 [36, (B1)] which requires $\underline{m} \gtrsim n^{6/7}$. One example of such assumption is the setup in [65] that each dimension has at most one change.

In summary, we have the following corollary.

Corollary 3.4. *Suppose the conditions in Lemma 3.3 are satisfied.*
(i) If $\bar{\delta} = \max_{0\leqslant k<k'\leqslant\nu}|\delta^{(k,k')}|_\infty$ *is bounded, then* $\mathbb{P}(\overline{T}_n > q_{\overline{T}^\sharp_n|X_1^n}(1-\alpha)) \geqslant 1 - \zeta - C_2(\underline{b})\varpi_n$ *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* (3.15) *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* $\mathbb{P}(\overline{T}_n > q_{\overline{T}^\sharp_n|X_1^n}(1-\alpha)) \geqslant 1 - \zeta - C_2(\underline{b})\varpi_n$ *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 14, we have shown that local alternative H_1 is detectable when $\underline{m} \gtrsim n^{1/2}\log^{1/2}(nd/\alpha)$. Corollary 3.4 (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 3.3 can be found in Remark 16. A real application for our global test lies in the special case of monotone signals that have order structures $\theta_1 \leqslant \cdots \leqslant \theta_\nu$ [78].

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

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

Note that there is no signal cancellation issue. Modified theory of power will depend on signal strength as below.

Corollary 3.5. *Denote* $m_k^Z = (m_k \bmod M)$*. Suppose the conditions in Lemma 3.3 hold. If*

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

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 *and* $C_1(\underline{b}), C_2(\underline{b})$.

Note that the rate now depends on M 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 M (or equivalently B) need to be selected carefully since power depends on the relevant locations of $\{m_k^Z\}_{k=0}^{\nu}$. One solution is to use $M = 2n^{1/2} \log^{1/2}(nd)$ that is discussed in Remark 14 or $M = 2n^{3/4} \log^{1/4}(nd)$ that is from Corollary 3.4 (ii).

3.4.3 Discussion on binary segmentation in change points estimation

To deal with multiple change points, binary segmentation (BS) is conceptually straightforward [36, 37, 106]. 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, [45] proposed wild binary segmentation (WBS) that randomly draw intervals to localize searching for change points. Recently, it has been widely adopted [99, 98] owing to its flexibility and computational efficiency. However, we will not be able to apply BS or WBS based approaches because there is no estimator in our framework so far.

A reasonable 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 (3.3). 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 word, within each segment $(m_k, m_{k+1}]$, $\mathbb{E}T_l(s)$ is monotone $(l = 1, \ldots, p)$. So $\max_{1\leqslant s\leqslant n-1} |\mathbb{E}T(s)|_\infty$ is always attained at one change point. Therefore, the estimator

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

can play a role in BS type approach. Similar ideas are discussed in [84, 49, 48, 18] 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.

3.4.4 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 [80, 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 $\mathcal{B}^{(0)} = \{B_1^{(0)}, B_2^{(0)}, \cdots, B_{\nu_0}^{(0)}\}$, 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.

$$DI_i = |T_n(B_i^{(0)} \cup B_{i+1}^{(0)})|_\infty = \max_{1\leqslant k\leqslant d} \left| (b_{i+1}^{(0)} - b_{i-1}^{(0)})^{1/2} \binom{b_{i+1}^{(0)} - b_{i-1}^{(0)}}{2}^{-1} \sum_{b_{i-1}^{(0)}+1\leqslant i<j\leqslant b_{i+1}^{(0)}} h_k(X_i, X_j) \right|. \tag{3.16}$$

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 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($\mathcal{B}^{(k)}$)

Start from data blocks as $\mathcal{B}^{(k)} = \{B_1^{(k)}, B_2^{(k)}, \cdots, B_{\nu_k}^{(k)}\}$
Compute the Dissimilarity Index $DI_i = T_n(B_i^{(k)}, B_{i+1}^{(k)})$ as in (3.16) for $i = 1, \ldots, \nu_k - 1$
Let $i^* = \text{argmin} DI_i$.
if our bootstrap test rejects the null for the segment $[b_{i^*-1}^{(k)}, b_{i^*+1}^{(k)}]$ **then**
    Repeat the test for i^* referring to the next smallest DI_i until all pairs are examined
else
    Update $B_i^{(k+1)} = B_i^{(k)}$ for $i < i^*$
    Merge $B_{i^*}^{(k)}, B_{i^*+1}^{(k)}$ into one block $B_{i^*}^{(k+1)} = B_{i^*}^{(k)} \cup B_{i^*+1}^{(k)}$
    Set $B_i^{(k+1)} = B_{i+1}^{(k)}$ for $i > i^*$
    Perform BD($\mathcal{B}^{(k+1)}$)
end if
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 worry on signal cancellation issue. 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 3.4.2. Empirical performance are investigated in simulation and real data application.

3.5 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 (3.1). 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.

3.5.1 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 [79, 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 (i.e., Gaussian mixture model): $\xi_i \sim \text{ctm-G}(\varepsilon, \nu, V) = (1-\epsilon)N(0, V) + \epsilon N(0, \nu^2 V)$ 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: $\xi_i = V^{1/2}\eta_i$, where $\eta_i = (\eta_{i1}, \ldots, \eta_{ip})^T$ and η_{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: $V_{ij} = 0.8^{|i-j|}$, $i, j = 1, \ldots, p$.

In all setups, $B = 200$ 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 $h(x, y) = \text{sign}(x - y)$.

3.5.2 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 3.1, 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 3.1 to help visualizing the size approximation.

$\sup_{\alpha\in(0,1)} \lvert\hat{R}(\alpha)-\alpha\rvert$		linear kernel			sign kernel			
		Gaussian	t_6	ctm-G	Gaussian	t_6	ctm-G	Cauchy
I	$V=\text{Id}_p$	0.034	0.086	0.040	0.026	0.066	0.032	0.028
II	$V=0.8J+0.2\text{Id}_p$	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

Table 3.1: Uniform error-in-size under H_0.

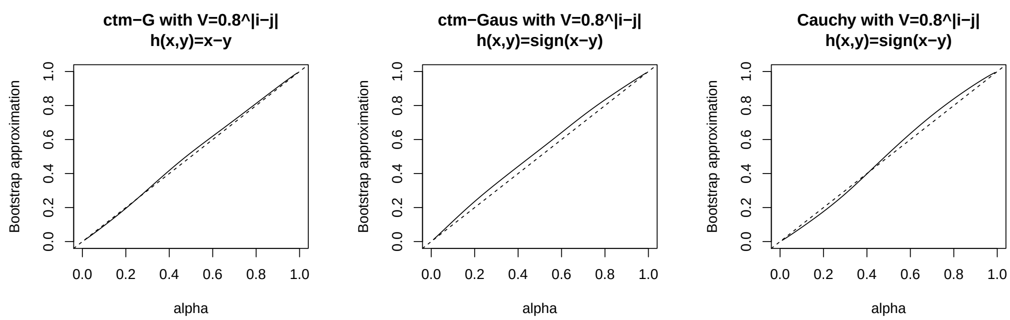


Figure 3.1: Selected setups for comparing $\hat{R}(\alpha)$ along with α. See headlines for corresponding distribution and kernel.

We also compare our test with the linear kernel to the CUSUM counterpart in [106, BABS] under the same setting with the boundary removal parameter as $\underline{s}=40$. Table 3.2 displays corresponding simulation results. By comparing it to Table 3.1, 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.

	$\sup_{\alpha\in(0,1)} \lvert\hat{R}(\alpha)-\alpha\rvert$			$\sup_{\alpha\in(0,0.1]} \lvert\hat{R}(\alpha)-\alpha\rvert$					
	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

Table 3.2: Error-in-size $\sup_\alpha \lvert\hat{R}(\alpha)-\alpha\rvert$ for $\alpha\in(0,1)$ and $\alpha\in(0,0.1]$.

3.5.3 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 3.2 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) whichever kernel is selected. 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 reverse 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 the exact power values, please refer to Table 3.9 (linear kernel) and 3.10 (sign kernel) in Section 3.7.

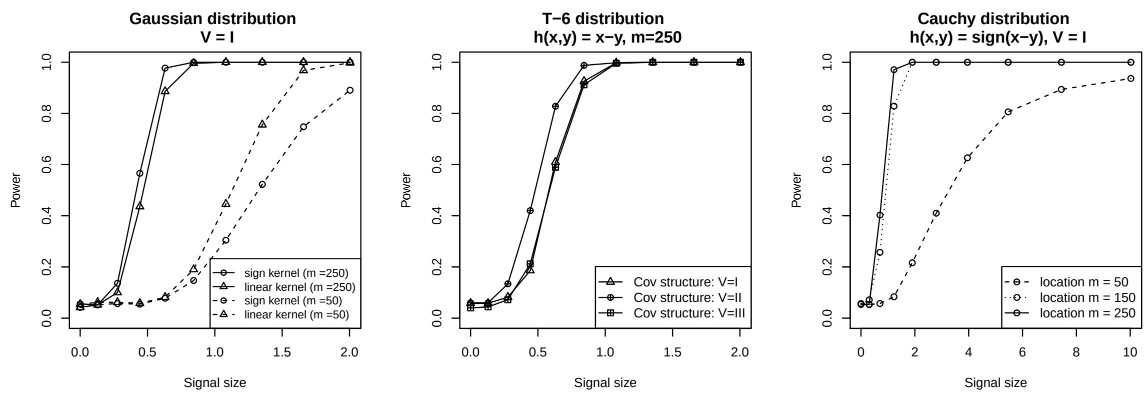


Figure 3.2: Selected setups for comparing power curves. See headlines and legends for corresponding distribution, kernel, covariance structures and change point location m.

3.5.4 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 [106, `BABS`], [65, `Jirak`], [37, `SBS`] and [99, `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 (3.5), 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 thresholdded ℓ^1-type estimator and `Inspect` is projection based. We adopt their single change point version function in R packages and convert them to tests using their default threshold computing functions. In our simulation,

$\alpha = 0.05, m = 150$ are fixed and we set boundary removal as 40 for `BABS`, `Jirak` and `SBS`.

Table 3.3 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 3.3 indicates that our method, which enjoys tuning-free and intermediate-estimation-free properties, is competent in empirical studies.

For fair comparison, we do not use Cauchy distribution since all methods except for our sign kernel one 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 [36, `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.

$\|\theta\|_\infty$	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

Table 3.3: Powers for our method using linear and sign kernels, [106, `BABS`], [65, `Jirak`], [37, `SBS`] and [99, `Inspect`].

3.5.5 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) [86, 63]. 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 3.4. Our test works well as there is no signal cancellation.

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

Table 3.4: Powers under multiple change point scenario using linear kernel. Here, $(m_1, m_2) = (300, 600)$.

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 3.5 (counts and ARIs) and Figure 3.3 (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 corresponding results in Table 3.5 and Figure 3.3 are similar to that of two change point case.

		$(m_1, m_2) = (300, 600)$					$(m_1, m_2, m_3) = (300, 600, 800)$				
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

Table 3.5: Estimation of multiple change points for $M = 100$. Here, the data is Gaussian distributed with dependence structure (III) and *linear* kernel is used.

Then, we also use the sign kernel to detect location shift for Cauchy distribution with dependence

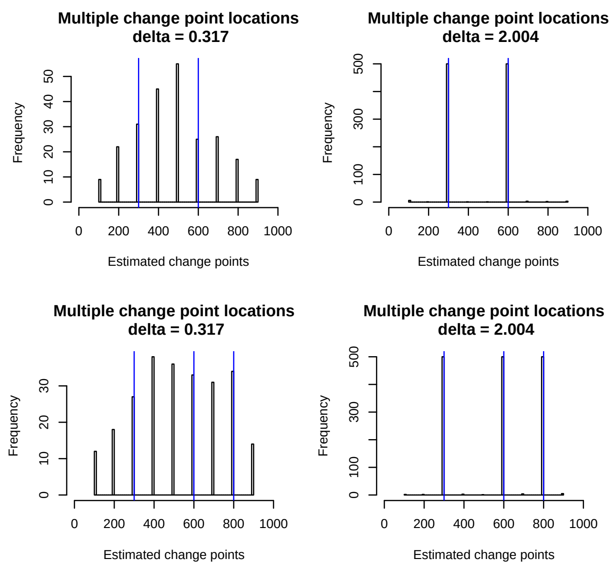


Figure 3.3: Multiple change point setup using *linear* kernel at signal level $\delta = 0.822, 10.023$. Upper: 2 change points $(m_1, m_2) = (300, 600)$. Lower: 3 change points $(m_1, m_2, m_3) = (300, 600, 800)$.

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 3.6 and Figure 3.4. Similar conclusion can be drawn except that stronger signal strength is required as Cauchy distribution has extremely heavy tails.

		$(m_1, m_2) = (300, 600)$					$(m_1, m_2, m_3) = (300, 600, 800)$				
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

Table 3.6: Estimation of multiple change points for $M = 100$. Here, the data is Cauchy distributed with dependence structure (III) and *sign* kernel is used.

Last, we set $M = 1$ and repeat the experiment using linear kernel and Gaussian distribution with dependence structure (III). The results are summarized in Table 3.7 and Figure 3.5. Compared to Table 3.5 and Figure 3.4 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

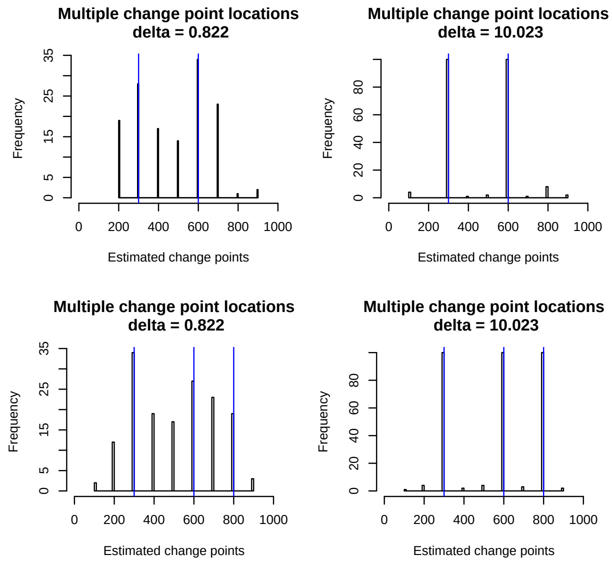


Figure 3.4: Multiple change point setup using *sign* kernel at signal level $\delta = 0.822, 10.023$. Upper: 2 change points $(m_1, m_2) = (300, 600)$. Lower: 3 change points $(m_1, m_2, m_3) = (300, 600, 800)$.

be introduced to force merging two blocks if the cardinality of their union is small.

		$(m_1, m_2) = (300, 600)$					$(m_1, m_2, m_3) = (300, 600, 800)$				
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

Table 3.7: Estimation of multiple change points for $M = 1$. Here, the data is Gaussian distributed with dependence structure (III) and *linear* kernel is used.

3.6 Real Data Applications

3.6.1 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

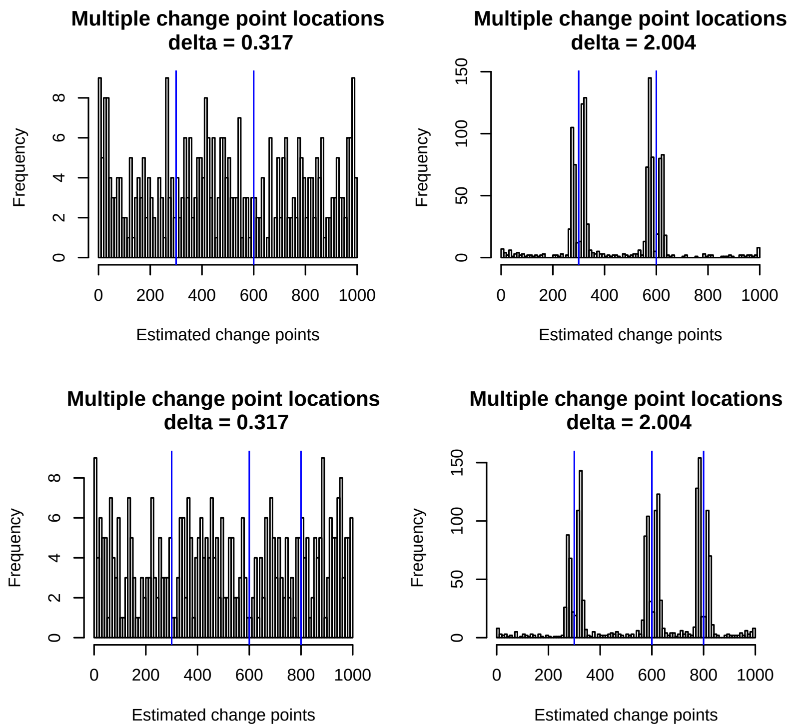


Figure 3.5: Multiple change point setup using $M = 1$ and *linear* kernel at signal level $\delta = 0.317, 2.004$. Upper: 2 change points $(m_1, m_2) = (300, 600)$. Lower: 3 change points $(m_1, m_2, m_3) = (300, 600, 800)$.

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] .

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 3.6 indicates the presence of extensive email communication from the second

[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`.

half of 2000 to the first half of 2001. Our test confirms that there was abnormal email activity in these two years.

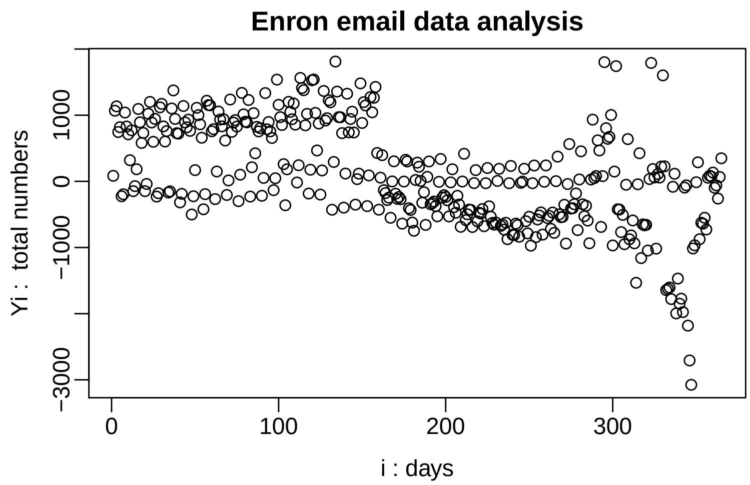


Figure 3.6: Trend of $Y_i = \sum_{j=1}^{101} X_{ij}$ for Enron email dataset.

3.6.2 Multiple change point: Micro-array dataset

The array comparative genomic hybridization data, `ACGH` [64, 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 3.7. Our BD algorithm finds 32 change points that are marked in red vertical dashed lines. This number is in a reasonable level as indicated in [99] where the authors only reported 30 most significant ones while their default `Inspect` algorithm found 254 change points. The ARI between ours and the bootstrap-assisted binary segmentation [106, `BABS`] which identifies 27 change points is 0.779. As shown in Table 3.8, the two methods have overlapped detection that are close loci numbers such as $(73, 74), (342, 344), (521, 528), \ldots, (2143, 2142)$.

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.

Table 3.8: Identified change point locations (loci numbers on genome) in `ACGH` dataset.

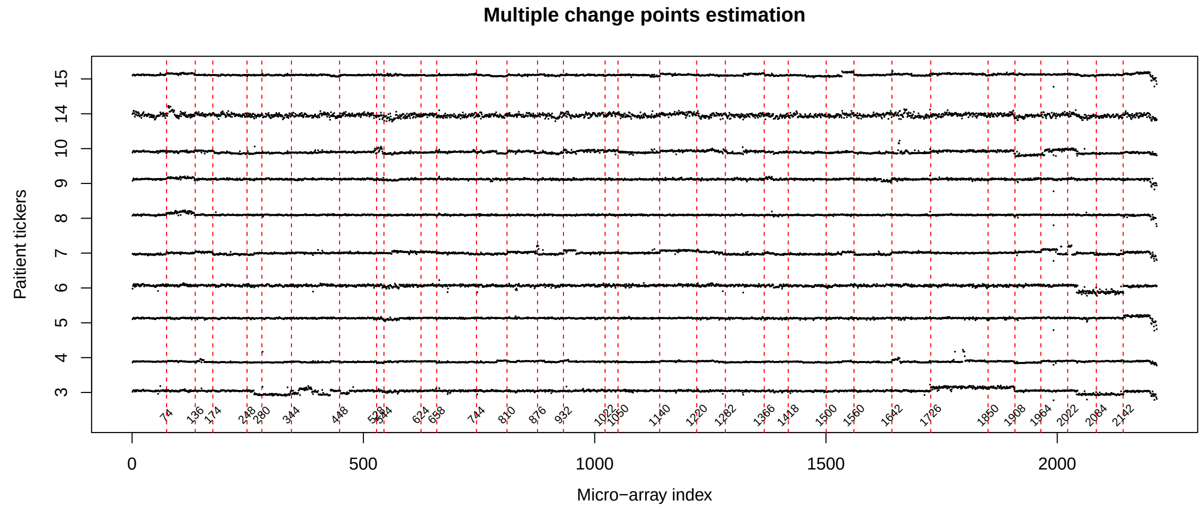


Figure 3.7: Real data study: aCGH data. Here, we use $B = 1000, \alpha = 0.01$ and the linear kernel.

3.7 Proofs and additional numerical results

3.7.1 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

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

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 [35] (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 [29]), 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

$$\rho(\overline{L}_n, \overline{Z}) \leqslant \rho(\overline{L}_n, \overline{Z}_n) + \rho(\overline{Z}_n, \overline{Z}) \lesssim \varpi_n. \tag{3.17}$$

Applying Corollary 5.6 in [28] with $k = 2$, we have

$$\mathbb{E}\left(\max_{1\leqslant j\leqslant d} |R_{nj}|\right) \lesssim D_n n^{-1/2} \log d. \tag{3.18}$$

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 (3.17) for the linear part, step (iii) from Nazarov's inequality (cf. Lemma A.1 in [35]), and step (iv) from the maximal inequality (3.18) 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 (3.6), $T_n^\sharp | X_1^n \sim N(\mathbf{0}, 4\hat{\Gamma}_n)$ where

$$\hat{\Gamma}_n = \frac{1}{n(n-1)^2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=i+1}^{n} h(X_i, X_j) h(X_i, X_k)^T. \tag{3.19}$$

By Lemma 3.6, $\mathbb{P}\left(|\hat{\Gamma}_n - \Gamma/3|_\infty \geqslant K_3 \left\{\frac{D_n^2 \log(nd)}{n}\right\}^{1/2}\right) \leqslant \gamma$. Therefore, [27, 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. Denote

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

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

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

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 (3.12).

(1) Quantify $|\tilde{\Delta}|_\infty$. Without loss of generality, we may assume $n_1 = m \leqslant n-m = n_2$. Recall (3.8) 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 3.2,

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

By Lemma 3.10, 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 3.11,

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

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,

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

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 (3.19). 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 $l = 1, \ldots, d$,

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

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 3.7, 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 $1-\gamma$,

$$q_{\overline{T}_n^\sharp}(1-\alpha) \leqslant 2\overline{\psi}\Phi^{-1}(1-\alpha/(2d)) \leqslant K_6 n^{-3/2}\left(\Delta_3 + |\theta_h|_\infty\left\{n^3\log(2d/\alpha)\delta_n\right\}^{1/2}\right).$$

Since $n_2 \geqslant n/2$ and $n_1 \gtrsim \log^{5/2}(nd)$, the rate of $\left\{n^3\log(2d/\alpha)\delta_n\right\}^{1/2} \lesssim n_1n_2$ leads to $q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha) \leqslant K_6n^{-3/2}(\Delta_3+n_1n_2|\theta_h|_\infty)$. For bounded kernel h, a simpler bound of $\overline{\psi} \leqslant K_5D_n$ directly lead to $q_{\overline{T}_n^\sharp|X_1^n}(1-\alpha) \leqslant K_6n^{-3/2}\Delta_3$ 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 Step1, 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 [92, 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 3.2 immediately follows for some large enough

$K \geqslant 2\sum_{i=1}^{6} K_i$.

$\square$

Proof of Lemma 3.3. 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 3.2, 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 (3.15).

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

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

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

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

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 (3.19). 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 3.7 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 3.7 we have

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

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,

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

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

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

(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

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

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. □

3.7.2 Proof of lemmas in theorems

Lemma 3.6 (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 (3.19). 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 3.6. Note $\Gamma = \mathrm{Cov}(g(X)) = \mathrm{Cov}(\mathbb{E}[h(X,X_1)|X]) = \mathbb{E}[h(X_1,X_2)h(X_1,X_3)^T]$ 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 the following Step 1 and 2, we will deal with $\hat{\Gamma}_{n1} = \frac{1}{n(n-1)^2}\sum_{i=1}^n\sum_{j\neq k>i} h(X_i,X_j)h(X_i,X_k)^T$ and $\hat{\Gamma}_{n2} = \frac{1}{n(n-1)^2}\sum_{i=1}^n\sum_{j=k>i} h(X_i,X_j)h(X_i,X_k)^T$ respectively, 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 [27] to H' for $\alpha = 1/2, \eta = 1$ and $\delta = 1/2$,

$$\mathbb{P}\left(\frac{n}{3}|W_n - \mathbb{E}W_n|_\infty \geqslant 2\mathbb{E}Z_1 + t\right) \leqslant \exp\left(-\frac{t^2}{3\bar{\zeta}_n^2}\right) + 3\exp\left[-\left(\frac{t}{K_1||M||_{\psi_{1/2}}}\right)^{1/2}\right], \tag{3.20}$$

where

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

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),

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

So $\overline{\zeta}_n \leqslant n^{1/2} D_n$. From (i) [92, Lemma 2.2.2], (ii) the fact of $||X^2||_{\psi_{1/2}} = ||X||^2_{\psi_1}$ and (iii) Condition (A3), we obtain

$$\begin{aligned}
||M||_{\psi_{1/2}} &= || \max_{1 \leqslant m_1, m_2 \leqslant d} \max_{0 \leqslant i \leqslant \frac{n}{3}-1} h_{m_1}(X_{3i+1}, X_{3i+2}) h_{m_2}(X_{3i+1}, X_{3i+3}) ||_{\psi_{1/2}} \\
&\leqslant_{(i)} K_2 \log^2(nd) \max_{1 \leqslant m_1, m_2 \leqslant d} \max_{0 \leqslant i \leqslant \frac{n}{3}-1} ||h_{m_1}(X_{3i+1}, X_{3i+2}) h_{m_2}(X_{3i+1}, X_{3i+3})||_{\psi_{1/2}} \\
&\leqslant K_2' \log^2(nd) \max_{1 \leqslant m_1 \leqslant d} \max_{0 \leqslant i \leqslant \frac{n}{3}-1} ||h^2_{m_1}(X_{3i+1}, X_{3i+2})||_{\psi_{1/2}} \\
&=_{(ii)} K_2' \log^2(nd) \max_{1 \leqslant m_1 \leqslant d} \max_{0 \leqslant i \leqslant \frac{n}{3}-1} ||h_{m_1}(X_{3i+1}, X_{3i+2})||^2_{\psi_1} \\
&\leqslant_{(iii)} K_2' \log^2(nd) D_n^2.
\end{aligned}$$

By [34, 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, (3.20) leads to

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

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,

$$\mathbb{P}\left(|\hat{\Gamma}_{n1} - \mathbb{E}\hat{\Gamma}_{n1}|_\infty \geqslant t^*\right) \leqslant \gamma^{\frac{K_5^2}{3K}} + 3\gamma^{\frac{K_5^{1/2}}{K K_1 K_2^{1/2}}} \leqslant \gamma/2.$$

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 $H(x_1, x_2)$ be defined as $h(x_1, x_2) h(x_1, x_2)^T$. 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 [27],

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

where

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

and $\overline{H}_{m_1,m_2}(x_1, x_2) = H_{m_1,m_2}(x_1, x_2)\mathbf{1}_{\{\max_{m_1,m_2} |H_{m_1,m_2}(x_1,x_2)| \leqslant \tau\}}$ 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,

$$
\begin{aligned}
||M'||_{\psi_{1/2}} &= || \max_{1\leqslant m_1, m_2 \leqslant d} \max_{0 \leqslant i \leqslant \frac{n}{2}-1} h_{m_1}(X_{2i+1}^{2i+2}) h_{m_2}(X_{2i+1}^{2i+2}) ||_{\psi_{1/2}}\\
&\leqslant K_7 \log^2(nd) \max_{1\leqslant m_1 \leqslant d} \max_{0 \leqslant i \leqslant \frac{n}{2}-1} ||h_{m_1}(X_{2i+1}, X_{2i+2})||_{\psi_1}^2\\
&\leqslant K_7 \log^2(nd) D_n^2.
\end{aligned}
$$

Then by [34, 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

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

□

Lemma 3.7 (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 3.2 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 3.7. Note that the summation breaks down to

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

and $h_l^2(x,y) = h_l^2(y,x)$. Apply [27, 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 3.6 Step2) shows

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

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

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

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 3.9 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, apply [27, Lemma E.1] to $\hat{\Gamma}_2 = \frac{1}{n_2(n_2-1)}\sum_{1\leqslant i<j\leqslant n_2} h(Y_i,Y_j)h(Y_i,Y_j)^T$ for the third term, we have

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

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

$$\begin{aligned}&(n_1^2 t_1) \vee (n_1 n_2 t_2) \vee (n_2^2 t_3)\\ \leqslant &K^\diamond \left\{D_n^2 n^{\frac{3}{2}} \log^{\frac{1}{2}}(nd) + |\theta_h|_\infty^2 [n_1^{\frac{1}{2}} n_2 \log^{\frac{1}{2}}(nd) + n_2 \log^3(nd)\log(\gamma^{-1})]\right\} =: t^\diamond,\end{aligned}$$

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

3.7.3 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

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

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 $\sum_{i=1}^{n_1} h(X_i, Y_{\pi_{n_2}(i)})$

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

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

$$\begin{aligned}
\overline{h}(x, y) &= h(x, y) \mathbf{1}\{\max_{1 \leqslant k \leqslant d} |h_k(x, y)| \leqslant \tau\}, \tau > 0 \\
Z_1 &= \max_{1 \leqslant k \leqslant d} \left| \sum_{i=1}^{n_1} \overline{h}_k(X_i, Y_i) - \mathbb{E} \bar{h}_k \right| \\
M &= \max_{1 \leqslant k \leqslant d} \max_{1 \leqslant i \leqslant n_1} |h_k(X_i, Y_i)| \\
\overline{\zeta}_{n_1}^2 &= \max_{1 \leqslant k \leqslant d} \sum_{i=1}^{n_1} \mathbb{E} h_k^2(X_i, Y_i)
\end{aligned}$$

By similar argument of Lemma E.1 in [27], we have the following result.

Lemma 3.8 (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

$$\mathbb{P}(Z \geqslant (1+\eta)\mathbb{E}Z_1 + t) \leqslant \exp\left(-\frac{t^2}{2(1+\delta)\overline{\zeta}_{n_1}^2}\right) + 3\exp\left[-\left(\frac{t}{C(\alpha,\eta,\delta)||M||_{\psi_\alpha}}\right)^\alpha\right] \tag{3.21}$$

holds for all $t > 0$.

Proof of Lemma 3.8. See Lemma E.1 in [27]. □

By Lemma 3.8, we can have the following result.

Lemma 3.9 (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,*

$$\mathbb{P}(\max_{1\leqslant k\leqslant d}|\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} h_k^2(X_i,Y_j) - \mathbb{E}h_k^2(X_i,Y_j)| \geqslant t^*) \leqslant \zeta \tag{3.22}$$

holds for $t^* = K_0 B_n^2\{\underline{n}^{-1/2}\log^{1/2}(\overline{n}d) + \underline{n}^{-1}\log^3(\overline{n}d)\log(1/\zeta)\}$.

Proof of Lemma 3.9. 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 3.8 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 [92],

$$||\sqrt{M}||_{\psi_1}^2 = ||\max_{1\leqslant k\leqslant d}\max_{1\leqslant i\leqslant n_1}|h_k(X_i,Y_i)|||_{\psi_1}^2 \leqslant K_3(\log(n_1 d)\max_{k,i}||h_k(X_i,Y_i)||_{\psi_1})^2 = K_3\log^2(n_1 d)B_n^2.$$

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 [34] 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,

$$\mathbb{P}\left(\max_{1\leqslant k\leqslant d}|\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}h_k^2(X_i,Y_j)-\mathbb{E}h_k^2|\geqslant K_5B_n^2[n_1^{-1/2}\log^{1/2}(d)+n_1^{-1}\log^3(n_1d)]+t\right)$$
$$\leqslant \exp\left(-\frac{n_1t^2}{3B_n^4}\right)+3\exp\left[-\left(\frac{\sqrt{n_1t}}{K_1K_3B_n\log(n_1d)}\right)\right]$$

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

$$\frac{n_1t_1^{*2}}{B_n^4}=\log(\overline{n}d)\gtrsim\log(1/\zeta)\text{ and }\frac{\sqrt{n_1t_1^*}}{B_n\log(n_1d)}\geqslant\underline{n}^{1/4}\log^{-3/4}(\overline{n}d)\gtrsim\log(1/\zeta).$$

(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

$$\frac{n_1t_2^{*2}}{B_n^4}\geqslant\underline{n}^{-1}\log^6(\overline{n}d)\log^2(1/\zeta)\gtrsim\log(1/\zeta)\text{ and }\frac{\sqrt{n_1t_2^*}}{B_n\log(n_1d)}=\log^{1/2}(\overline{n}d)\log^{1/2}(1/\zeta)\gtrsim\log(1/\zeta).$$

Observing $B_n^2[n_1^{-1/2}\log^{1/2}(d)+n_1^{-1}\log^3(n_1d)]\lesssim t_1^*+t_2^*=:t^*$. Hence,

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

□

3.7.4 Lemma for two-sample Hoeffding decomposition

Lemma 3.10 (Tail bound of the maxima of the first order projection)**.** *Let $X_1,\dots,X_n$ 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 $k=1,\dots,d$ 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,*

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

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

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

Proof of Lemma 3.10. 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 [34, Lemma 8],

$$\begin{aligned}&||M||_{\psi_1}\leqslant K_2\log(nd)\max_{i,k}||Gh_k(X_i)||_{\psi_1}\leqslant K_2D_n\log(nd)\quad\text{ and}\\&\mathbb{E}Z\leqslant K_3\{\sigma\sqrt{\log d}+||M||_{\psi_1}\log d\}\leqslant K_4\{\sqrt{n\log(d)D_n}+\log(nd)\log(d)D_n\}.\end{aligned}$$

Take $t^*=K_5D_n\{n^{1/2}\log^{1/2}(nd)\vee\log^2(nd)\}$, simple calculation shows $\mathbb{P}(Z\geqslant t^*)\leqslant\zeta$. □

Lemma 3.11 (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*

$$\begin{aligned}&\mathbb{E}|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j)|_\infty\\\leqslant&KD_n\log(d)\left\{\log(d)\log(n_2d)+(n_1n_2)^{1/2}+[n_2\log(d)\log^2(n_2d)]^{1/2}+[n_1n_2^2\log(d)]^{1/4}\right\}.\end{aligned}$$

Proof of Lemma 3.11. The structure of this proof is similar to the one-sample version in [27, Thm 5.1]. By constructing randomization from iid Rademacher random variables (i.e. $\mathbb{P}(\epsilon_i=\pm1)=\frac{1}{2}$ for all ϵ_i and ϵ_j', $i=1,\ldots,n_1, j=1,\ldots,n_2$), [39, 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 [89, 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

$$\begin{aligned}\mathbb{E}[\max_{1\leqslant m\leqslant d}|\epsilon^T\Lambda^m\epsilon||X_1^{n_1},Y_1^{n_2}] &= \int_0^\infty \mathbb{P}\left(\max_{1\leqslant m\leqslant d}|\epsilon^T\Lambda^m\epsilon| \geqslant t|X_1^{n_1},Y_1^{n_2}\right)dt\\ &\leqslant t^* + 2d\int_{t^*}^\infty \max\{\exp(-\frac{K_2t^2}{V_1^2}), \exp(-\frac{K_2t}{V_2})\}.\end{aligned}$$

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

$$2d\int_{t^*}^\infty \exp(-\frac{K_2t^2}{V_1^2})dt \leqslant \frac{V_1}{\sqrt{2K_2}}\int_{\sqrt{2\log d}}^\infty \exp(-\frac{s^2}{2})ds \leqslant \frac{V_1}{\sqrt{K_2\log d}} \leqslant K_2V_1.$$

Similarly,

$$2d\int_{t^*}^\infty \exp(-\frac{K_2t}{V_2})dt \leqslant 2V_2/K_2.$$

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

$$\begin{aligned}\mathbb{E}|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}(X_i,Y_j)\epsilon_i\epsilon_j'|_\infty &\leqslant K_1\mathbb{E}[t^*+K_2V_1+2V_2/K_2] \leqslant K_3(\log d)\mathbb{E}V_1\\ &\leqslant K_3(\log d)(\mathbb{E}\max_{1\leqslant m\leqslant d}|\Lambda^m|_F^2)^{1/2}. \qquad (3.23)\end{aligned}$$

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,

$$\begin{aligned}&\mathbb{E}[\max_{1\leqslant m\leqslant d}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_m^2(X_i,Y_j)] = \mathbb{E}[\max_{1\leqslant m\leqslant d}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0^m(X_i,Y_j)+\breve{f}_1^m(X_i)+\breve{f}_2^m(Y_j)+\mathbb{E}\breve{f}_m^2]\\ &\leqslant \mathbb{E}[|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0^m(X_i,Y_j)|_\infty] + n_2\mathbb{E}[|\sum_{i=1}^{n_1}\breve{f}_1^m(X_i)|_\infty] + n_1\mathbb{E}[|\sum_{j=1}^{n_2}\breve{f}_2^m(Y_j)|_\infty] + n_1n_2\max_{1\leqslant m\leqslant d}\mathbb{E}\breve{f}_m^2. \qquad (3.24)\end{aligned}$$

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

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

Denote $M = \max_{i,j,m}|\breve{f}_m(X_i, Y_j)|$. Following arguments in beginning and the symmetrization inequality [92, Lemma 2.3.1], we have

$$\mathbb{E}|\sum_{i=1}^{n_1}\breve{f}_1(X_i)|_\infty \leqslant \sqrt{\log d}\;\mathbb{E}\sqrt{\max_m \sum_{i=1}^{n_1}\breve{f}_1^m(X_i)^2} \leqslant K_4\sqrt{\log d}\sqrt{n_1\max_m \mathbb{E}\breve{f}_m^4 + \log d||M||_4^4}, \tag{3.25}$$

$$\mathbb{E}|\sum_{j=1}^{n_2}\breve{f}_2(Y_j)|_\infty \leqslant \sqrt{\log d}\;\mathbb{E}\sqrt{\max_m \sum_{j=1}^{n_2}\breve{f}_2^m(Y_j)^2} \leqslant K_5\sqrt{\log d}\sqrt{n_2\max_m \mathbb{E}\breve{f}_m^4 + \log d||M||_4^4}, \tag{3.26}$$

$$\mathbb{E}|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0(X_i, Y_j)|_\infty \leqslant \log d\;\mathbb{E}\sqrt{\max_m \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0^m(X_i, Y_j)^2} \leqslant K_6\log d\sqrt{I}||M||_2. \tag{3.27}$$

The last step of (3.25) comes from [27, Equation (58)]. The (3.26) follows the same procedure. And the first step of (3.27) is dealt the same way as (3.23) with

$$\begin{aligned}\mathbb{E}\sqrt{\max_m \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\breve{f}_0^m(X_i, Y_j)^2} \leqslant 2\Big[&\mathbb{E}\sqrt{\max_m \sum_{i,j}\breve{f}_m^4(X_i, Y_j)} + \mathbb{E}\sqrt{\max_m \sum_{i,j}(\mathbb{E}[\breve{f}_m^2(X_i, Y_j')|X_1^{n_1}])^2} \\ &+\mathbb{E}\sqrt{\max_m \sum_{i,j}(\mathbb{E}[\breve{f}_m^2(X_i', Y_j)|Y_1^{n_2}])^2} + \mathbb{E}\sqrt{\max_m \sum_{i,j}(\mathbb{E}\breve{f}_m^2(X_i, Y_j))^2}\Big] \\ \leqslant& K_6\sqrt{I}\sqrt{\mathbb{E}M^2}.\end{aligned}$$

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 $\max_m \mathbb{E}\breve{f}_m^4 \leqslant D_n^2$ and $||M||_4 \lesssim ||M||_{\psi_1} \leqslant K_7 D_n \log(n_1 n_2 d) \leqslant 2K_7 D_n \log(n_2 d)$. Besides, we have $D_q = \max_m[\mathbb{E}|\breve{f}_m(X, Y)|^q]^{1/q} \lesssim D_n$. Plug (3.25)-(3.27) in (3.24) and the solution of quadratic inequality for I gives

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

Therefore, the square-root of I is less than the square-root of each term on RHS. Plug the result in 3.23. A simplified result is obtained in the statement of Lemma 3.11. □

3.7.5 Additional tables

	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

Table 3.9: 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				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 3.10: Powers report of our method using *sign* kernel. Here, $n = 500, p = 600, \alpha = 0.05$ and change point locations are $t_m = m/n = 5/10, 3/10, 1/10$.

References

[1] Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, (34):1000–1034, 2008.

[2] Radosław Adamczak. A few remarks on the operator norm of random toeplitz matrices. *Journal of Theoretical Probability*, 23(1):85–108, 2010.

[3] Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20(162):1–56, 2019.

[4] John AD Aston and Claudia Kirch. Detecting and estimating changes in dependent functional data. *Journal of Multivariate Analysis*, 109:204–220, 2012.

[5] John AD Aston, Claudia Kirch, et al. Evaluating stationarity via change-point alternatives with applications to fmri data. *The Annals of Applied Statistics*, 6(4):1906–1948, 2012.

[6] John AD Aston, Claudia Kirch, et al. High dimensional efficiency with applications to change point tests. *Electronic Journal of Statistics*, 12(1):1901–1947, 2018.

[7] Alexander Aue, Robertas Gabrys, Lajos Horváth, and Piotr Kokoszka. Estimation of a change-point in the mean function of functional data. *Journal of Multivariate Analysis*, 100(10):2254–2269, 2009.

[8] Alexander Aue, Siegfried Hörmann, Lajos Horváth, Matthew Reimherr, et al. Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087, 2009.

[9] Jushan Bai. Estimation of a change point in multiple regression models. *The Review of Economics and Statistics*, 79(4):551–563, 1997.

[10] Jushan Bai. Common breaks in means and variances for panel data. *Journal of Econometrics*, 157(1):78–92, 2010.

[11] Matteo Barigozzi, Haeran Cho, and Piotr Fryzlewicz. Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics*, 2018.

[12] Ian Barnett and Jukka-Pekka Onnela. Change point detection in correlation networks. *Scientific reports*, 6:18893, 2016.

[13] István Berkes, Robertas Gabrys, Lajos Horváth, and Piotr Kokoszka. Detecting changes in the mean of functional observations. *J. R. Stat. Soc. Ser. B*, 71(5):475–507, 2009.

[14] Monika Bhattacharjee, Moulinath Banerjee, and George Michailidis. Change point estimation in panel data with temporal and cross-sectional dependence. *arXiv preprint arXiv:1904.11101*, 2019.

[15] Pallab K. Bhattacharya. Maximum likelihood estimation of a change-point in the distribution of independent random variables: General multiparameter case. *Journal of Multivariate Analysis*, 23(2):183–208, 1987.

[16] Peter J. Bickel and Elizaveta Levina. Covariance Regularization by Thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

[17] Peter J. Bickel and Elizaveta Levina. Regularized Estimation of Large Covariance Matrices. *Annals of Statistics*, 36(1):199–227, 2008.

[18] Vincent Brault, Sarah Ouadah, Laure Sansonnet, and Céline Lévy-Leduc. Nonparametric multiple change-point estimation for analyzing large hi-c data matrices. *Journal of Multivariate Analysis*, 165:143–165, 2018.

[19] B.E. Brodsky and B.S. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Springer-Science+Business Media, B.V., 1993.

[20] Axel Bücher, Jean-David Fermanian, and Ivan Kojadinovic. Combining cumulative sum change-point detection tests for assessing the stationarity of univariate time series. *Journal of Time Series Analysis*, 40(1):124–150, 2019.

[21] Axel Bücher, Ivan Kojadinovic, Tom Rohmer, and Johan Segers. Detecting changes in cross-sectional dependence in multivariate time series. *Journal of Multivariate Analysis*, 132:111–128, 2014.

[22] Tony Cai, Cun-Hui Zhang, and Harrison Zhou. Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[23] Tony Cai and Harrison Zhou. Optimal Rates of Convergence for Sparse Covariance Matrix Estimation. *Annals of Statistics*, 40(5):2389–2420, 2012.

[24] Edward Carlstein. Nonparametric change-point estimation. *The Annals of Statistics*, pages 188–197, 1988.

[25] Jie Chen and Arjun K Gupta. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical association*, 92(438):739–747, 1997.

[26] Likai Chen, Weining Wang, and Weibiao Wu. Inference of break-points in high-dimensional time series. *Available at SSRN 3378221*, 2019.

[27] Xiaohui Chen. Gaussian and bootstrap approximations for high-dimensional u-statistics and their applications. *Annals of Statistics*, 46(2):642–678, 2018.

[28] Xiaohui Chen and Kengo Kato. Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. 2017. arXiv:1708.02705.

[29] Xiaohui Chen and Kengo Kato. Randomized incomplete u-statistics in high dimensions. *The Annals of Statistics, accepted (available at arXiv:1712.00771)*, 2018+.

[30] Xiaohui Chen, Mengyu Xu, and Wei Biao Wu. Covariance and precision matrix estimation for high-dimensional time series. *Annals of Statistics*, 41(6):2994–3021, 2013.

[31] H Chernoff and S Zacks. Estimating the current mean of a normal distribution which is subjected to changes in time. *Annals of Mathematical Statistics*, 35(3):999–1018, 1964.

[32] Herman Chernoff and Shelemyahu Zacks. Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35(3):999–1018, 1964.

[33] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41(6):2786–2819, 2013.

[34] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probab. Theory Related Fields*, 162:47–70, 2015.

[35] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4):2309–2352, 2017.

[36] Haeran Cho et al. Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics*, 10(2):2000–2038, 2016.

[37] Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B*, 77:475–507, 2015.

[38] M. Csörgö and L Horváth. *Limit Theorems in Change-Point Analysis*. New York, 1997.

[39] Victor de la Peña and Evarist Giné. *Decoupling: From Dependence to Independence*. Springer, 1999.

[40] Herold Dehling and Roland Fried. Asymptotic distribution of two-sample empirical u-quantiles with applications to robust tests for shifts in location. *Journal of Multivariate Analysis*, 105(1):124–140, 2012.

[41] H Dette, GM Pan, and Q Yang. Estimating a change point in a sequence of very high-dimensional covariance matrices. *arXiv preprint arXiv:1807.10797*, 2018.

[42] Holger Dette, Josua Gösmann, et al. Relevant change points in high dimensional time series. *Electronic Journal of Statistics*, 12(2):2578–2636, 2018.

[43] Farida Enikeeva, Zaid Harchaoui, et al. High-dimensional change-point detection under sparse alternatives. *Annals of Statistics*, 47(4):2051–2079, 2019.

[44] Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

[45] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42(6):2243–2281, 2014.

[46] Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. *arXiv preprint arXiv:1612.04740*, 2016.

[47] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

[48] Edit Gombay. U-statistics for change under alternatives. *Journal of Multivariate Analysis*, 78(1):139–158, 2001.

[49] Edit Gombay and Lajos Horváth. An application of u-statistics to change-point analysis. *Acta Scientiarum Mathematicarum*, 60(1):345–358, 1995.

[50] Edit Gombay and Lajos Horváth. Change-points and bootstrap. *Environmetrics: The official journal of the International Environmetrics Society*, 10(6):725–736, 1999.

[51] Edit Gombay and Lajos Horváth. Rates of convergence for u-statistic processes and their bootstrapped versions. *Journal of Statistical Planning and Inference*, 102(2):247–272, 2002.

[52] Edit Gombay and Shuangquan Liu. A nonparametric test for change in randomly censored data. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 113–121, 2000.

[53] Peter Hall. On convergence rates of suprema. *Probability Theory and Related Fields*, 89:447–455, 1991.

[54] Zaıd Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

[55] Samir Ben Hariz, Jonathan Wylie, and Qiang Zhang. Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. *Annals of Statistics*, 35(4):1802–1826, 2007.

[56] Flore Harlé, Florent Chatelain, Cédric Gouy-Pailler, and Sophie Achard. Bayesian model for multiple change-points detection in multivariate time series. *IEEE Trans. Signal Processing*, 64(16):4351–4362, 2016.

[57] David Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17, 1970.

[58] J.L. Hodges and E.L. Lehmann. Estimates of location based on rank tests. *Annals of Mathematical Statistics*, 34(2):598–611, 1963.

[59] Mark Holmes, Ivan Kojadinovic, and Jean-François Quessy. Nonparametric tests for change-point detection à la gombay and horváth. *Journal of Multivariate Analysis*, 115:16–32, 2013.

[60] Lajos Horváth. The maximum likelihood method for testing changes in the parameters of normal observations. *The Annals of statistics*, pages 671–680, 1993.

[61] Lajos Horváth and Marie Hušková. Change-point detection in panel data. *Journal of Time Series Analysis*, 33(4):631–648, 2012.

[62] Lajos Horváth, Piotr Kokoszka, and Josef Steinebach. Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis*, 68(1):96–119, 1999.

[63] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[64] Nicholas A James and David S Matteson. ecp: An r package for nonparametric multiple change point analysis of multivariate data. *arXiv preprint arXiv:1309.3295*, 2013.

[65] Moritz Jirak. Uniform change point tests in high dimension. *Annals of Statistics*, 43:2451–2483, 2015.

[66] Rebecca Killick and Idris Eckley. changepoint: An r package for changepoint analysis. *Journal of statistical software*, 58(3):1–19, 2014.

[67] Piotr Kokoszka and Remigijus Leipus. Change-point estimation in arch models. *Bernoulli*, 6(3):475–507, 2000.

[68] Tze Leung Lai. Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 11(2):303–351, 2001.

[69] M. R. Leadbetter and Holger Lindgren, Georg an Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer-Verlag, 1983.

[70] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer. New York, 1991.

[71] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Oracle estimation of a change point in high dimensional quantile regression. *Journal of the American Statistical Association*, (just-accepted), 2017.

[72] Sokbae Lee, Myung Hwan Seo, and Youngki Shin. The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):193–210, 2016.

[73] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems*, pages 3366–3374, 2015.

[74] Haoyang Liu, Chao Gao, and Richard J. Samworth. Minimax rates in sparse, high-dimensional change-point detection, 2019.

[75] Clive R Loader et al. Change point estimation using nonparametric regression. *The Annals of Statistics*, 24(4):1667–1678, 1996.

[76] G. Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

[77] David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

[78] Kentaro Minami. Estimating piecewise monotone signals. *Electronic Journal of Statistics*, 14(1):15081576, 2020.

[79] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics, 1982.

[80] Yue S Niu, Ning Hao, and Heping Zhang. Multiple change-point detection: A selective overview. *Statistical Science*, pages 611–623, 2016.

[81] Hernando Ombao, Rainer von Sachs, and Wensheng Guo. Slex analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc.*, 100:519–531, 2005.

[82] Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric change point detection and localization. *arXiv preprint arXiv:1905.10019*, 2019.

[83] E.S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3):523–527, 1955.

[84] A. N. Pettitt. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(2):126–135, 1979.

[85] Peihua Qiu. *Introduction to statistical process control*. CRC Press, 2013.

[86] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[87] Sidney I. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Applied Probability. Springer-Verlag, 1987.

[88] Michael Robbins, Colin Gallagher, Robert Lund, and Alexander Aue. Mean shift testing in correlated data. *Journal of Time Series Analysis*, 32(5):498–511, 2011.

[89] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[90] Xiaofeng Shao and Xianyang Zhang. Testing for change points in time series. *Journal of the American Statistical Association*, 105(491):1228–1240, 2010.

[91] Ada van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[92] Ada van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

[93] Daniel Vogel and Martin Wendler. Studentized u-quantile processes under dependence with applications to change-point analysis. *Bernoulli*, 23(4B):3114–3144, 2017.

[94] Abraham Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16:117–186, 1945.

[95] Abraham Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19:326–339, 1948.

[96] Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal covariance change point localization in high dimension. *arXiv:1712.09912*, 2017.

[97] Runmin Wang and Xiaofeng Shao. Dating the break in high-dimensional data. *arXiv preprint arXiv:2002.04115*, 2020.

[98] Runmin Wang, Stanislav Volgushev, and Xiaofeng Shao. Inference for change points in high dimensional data. *arXiv preprint arXiv:1905.08446*, 2019.

[99] Tengyao Wang and Richard J Samworth. High dimensional change point estimation via sparse projection. *Jounral of Royal Statistical Society, Series B (to appear). arXiv:1606.06246v2*, 2017.

[100] Yao Wang, Chunguo Wu, Zhaohua Ji, Binghong Wang, and Yanchun Liang. Non-parametric change-point method for differential gene expression detection. *PloS one*, 6(5):e20060, 2011.

[101] William Woodall and Matoteng Ncube. Multivariate cusum quality-control procedures. *Technometrics*, 27(3):285–292, 1985.

[102] Dongdong Xiang, Shulin Gao, Wendong Li, Xiaolong Pu, and Wen Dou. A new nonparametric monitoring of data streams for changes in location and scale via cucconi statistic. *Journal of Nonparametric Statistics*, 31(3):743–760, 2019.

[103] Yao Xie and David Siegmund. Sequential multi-sensor change-point detection. In *2013 Information Theory and Applications Workshop (ITA)*, pages 1–20. IEEE, 2013.

[104] Yi-Ching Yao. Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *Annals of Statistics*, 15(3):1321–1328, 1987.

[105] Chun Yip Yau and Zifeng Zhao. Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):895–916, 2016.

[106] Mengjia Yu and Xiaohui Chen. Finite sample change point inference and identification for high-dimensional mean vectors. *arXiv preprint arXiv:1711.08747*, 2017.

[107] Mengjia Yu and Xiaohui Chen. A robust bootstrap change point test for high-dimensional location parameter. *arXiv preprint arXiv:1904.03372*, 2019.

[108] Danna Zhang and Wei Biao Wu. Gaussian approximation for high-dimensional time series. *Annals of Statistics*, 45(5):1895–1919, 2017.

[109] Nancy R Zhang, David O Siegmund, Hanlee Ji, and Jun Z Li. Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645, 2010.

[110] Xianyang Zhang and Guang Cheng. Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli*, 24(4A):2640–2675, 2018.

[111] Ping-Shou Zhong and Jun Li. Test for temporal homogeneity of means in high-dimensional longitudinal data. *arXiv preprint arXiv:1608.07482*, 2016.

[112] Yan Zhou, Liya Fu, and Baoxue Zhang. Two non parametric methods for change-point detection in distribution. *Communications in Statistics-Theory and Methods*, 46(6):2801–2815, 2017.