A Unified Data-adaptive Framework for High Dimensional Change Point Detection

Bin Liu*, Cheng Zhou†, Xinsheng Zhang‡, Yufeng Liu§

Abstract

In recent years, change point detection for a high dimensional data sequence has become increasingly important in many scientific fields such as biology and finance. The existing literature develops a variety of methods designed for either a specified parameter (e.g. mean or covariance) or a particular alternative pattern (sparse or dense), but not for both scenarios simultaneously. To overcome this limitation, we provide a general framework for developing tests suitable for a large class of parameters, and also adaptive to various alternative scenarios. In particular, by generalizing the classical cumulative sum (CUSUM) statistic, we construct the U-statistic-based CUSUM matrix $\mathcal{C}$. Two cases corresponding to common or different change point locations across the components are considered. We then propose two types of individual test statistics by aggregating $\mathcal{C}$ based on the adjusted L_p-norm with $p \in \{1, \ldots, \infty\}$. Combining the corresponding individual tests, we construct two types of data-adaptive tests for the two cases, which are both powerful under various alternative patterns. A multiplier bootstrap method is introduced for approximating the proposed test statistics' limiting distributions. With flexible dependence structure across coordinates and mild moment conditions, we show the optimality of our methods theoretically in terms of size and power by allowing the dimension d and the number of parameters q being much larger than the sample size n. An R package called AdaptiveCpt is developed to implement our algorithms. Extensive simulation studies provide further support for our theory. An application to a comparative genomic hybridization (CGH) dataset also demonstrates the usefulness of our proposed methods.

Keyword: Change point detection; Data-adaptive tests; High dimensions; Minimax optimality; Multiplier bootstrap and Gaussian approximation; U-statistics

1 Introduction

In modern statistical applications, high dimensional data are ubiquitous in many scientific fields such as finance, genetics, and engineering. Testing the homogeneity of such data sequences is a challenging yet important problem. In particular, high dimensional data with complex generating mechanism often present structural changes before and after a possible (unknown) change point. It is typically unrealistic

*Department of Statistics, The Chinese University of HongKong, HK; e-mail: liubin0145@gmail.com

†Robotics X Lab, Tencent; e-mail: mikechzhou@tencent.com

‡Department of Statistics, School of Management at Fudan University, China; e-mail: xszhang@fudan.edu.cn

§Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, U.S.A; e-mail:yfliu@email.unc.edu

to assume stationarity for the high dimensional data sequence. As a result, methods designed for stationary data can be invalid. Furthermore, detecting and identifying these change points also have various real applications. For example, in biology (Zhang et al. (2010)), change points exist for the DNA copy number variants among multiple biological samples in the chromosomes; in finance, abrupt economic announcements or changes can disrupt the network among the stocks; in functional Magnetic Resonance Imaging (fMRI) data studies (Zhong and Li (2016)), we can regard the abrupt changes for measurement of blood oxygen level-dependent responses as change points.

Motivated by the broad applications, in this paper, we consider a general framework for high dimensional change point detection problems. More specifically, let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ be a d-dimensional random vector. We are interested in a q-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$ with $\theta_s = \mathbb{E}\Phi_s(\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_m)$, where $\Phi_s(\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_m) : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$ is a measurable, symmetric kernel with order m, and $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_m$ are independent copies with the same distribution as $\boldsymbol{X}$. Denote $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ as n ordered independent observations from $\boldsymbol{X}$ with $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,d})^\top$. We aim to detect whether there is a change point of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ during n observations, where $\boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_{i,q})^\top$ for $1 \le i \le n$. Therefore, we consider the following hypothesis:

$$\begin{aligned} &\mathbf{H}_0 : \theta_{1,s} = \cdots = \theta_{n,s}, \ \text{ for } \ 1 \le s \le q, \quad \text{v.s.} \\ &\mathbf{H}_1 : \exists s \in \{1, \ldots, q\} \text{ and } \widetilde{t}_s \in (0,1), \text{ s.t. } \theta_{1,s} = \cdots = \theta_{\lfloor n\widetilde{t}_s \rfloor, s} \neq \theta_{\lfloor n\widetilde{t}_s \rfloor + 1, s} = \cdots = \theta_{n,s}, \end{aligned} \tag{1.1}$$

where $\widetilde{t}_s$ is the relative change point location for θ_s. In our setting, the dimension d and the number of parameters q can be much larger than the sample size n. Under our testing framework, many existing works are special cases of (1.1) by choosing a specified kernel:

Case 1: For $\theta_s = \mathbb{E}X_s$ with $1 \le s \le d$, the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ is the mean vector by letting $\Phi_s(\boldsymbol{X}) = X_s$ with $m = 1$. Therefore, the high dimensional mean change point detection problem falls into our general framework in (1.1) (Horváth and Hušková (2012); Aston and Kirch (2018); Jirak (2015); Cho et al. (2016); Wang and Samworth (2018)).

Case 2: For $\theta_{i,j} = \mathbb{E}\big((X_i - X'_i)(X_j - X'_j)/2\big)$ with $1 \le i, j \le d$, the parameter $\boldsymbol{\theta} \in \mathbb{R}^{d(d+1)/2}$ is the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$ by letting $\Phi_{i,j}(\boldsymbol{X}, \boldsymbol{X}') = (X_i - X'_i)(X_j - X'_j)/2$ with $m = 2$. Therefore, the general testing framework can deal with the high dimensional covariance change point detection problem (See Aue et al. (2009); Avanesov and Buzun (2018); Wang et al. (2017)).

Case 3: For $\theta_{i,j} = \mathbb{E}\big(\mathrm{sign}(X_i - X'_i)\mathrm{sign}(X_j - X'_j)\big)$ with $1 \le i < j \le d$, the parameter $\boldsymbol{\theta} \in \mathbb{R}^{d(d-1)/2}$ is the Kendall's tau correlation matrix by letting $\Phi_{i,j}(\boldsymbol{X}, \boldsymbol{X}') = \mathrm{sign}(X_i - X'_i)\mathrm{sign}(X_j - X'_j)$ with $m = 2$. In this case, the hypothesis (1.1) includes the high dimensional Kendall's tau correlation matrix change point problem.

In low dimensions, i.e. a case of fixed d with $d < n$, change point detection has been a well-established problem since the early works in Page (1954, 1955). There is also a large number of papers for various testing problems. For example, the methods developed for the mean vector include Srivastava and Worsley (1986); Horváth et al. (1999); Lung-Yut-Fong et al. (2011); papers on the variance or covariance include Inclan and Tiao (1994); Gombay et al. (1996); Berkes et al. (2009). In addition, there

are also some papers based on the non-parametric methods (Csörgő and Horváth (1988); Hušková and Meintanis (2006); Quessy et al. (2013); Matteson and James (2014); Tan et al. (2016)). For a comprehensive review of the low dimensional methods and their theoretical properties, we refer to Csörgő and Horváth (1997); Chen and Gupta (2011); Hušková and Prášková (2014), and the references therein.

Driven by the contemporary statistical applications, high dimensional data are commonly available, where the dimension d and the number of parameters q can be comparable with or even much larger than the sample size n. Examples include genetics, image analysis, and risk management. Compared with the low dimensional case, much less development has been made in the literature on the high dimensional change point detection. Most existing papers focus on the mean vector problem. To review the related literature, we introduce the well-known cumulative sum (CUSUM) (Csörgő and Horváth (1997)) statistic for each coordinate s at the time point k ($1 \leq k \leq n-1$) as:

$$C_s(k) = \sqrt{n}\frac{k}{n}\big(1-\frac{k}{n}\big)\sigma_{s,s}^{-1/2}\Big(\frac{1}{k}\sum_{j=1}^{k} X_{j,s} - \frac{1}{n-k}\sum_{j=k+1}^{n} X_{j,s}\Big), \text{ with } 1 \leq s \leq d, \tag{1.2}$$

where $\sigma_{s,s} = \mathrm{Var}(X_s)$. Based on (1.2), we define the CUSUM matrix $\mathcal{C} = (C_s(k))_{1\leq s\leq d, 1\leq k\leq n-1}$. One challenge for high dimensional change point detection is how to propose a test based on the aggregation of $\mathcal{C}$. To this end, many authors investigate various aggregations of $\mathcal{C}$ under different model assumptions. For independent observations and components, Zhang et al. (2010); Horváth and Hušková (2012) investigated the L_2 aggregations of $\mathcal{C}$ with and without the Gaussian assumption, respectively. Based on the L_2 aggregation, Enikeeva and Harchaoui (2019) proposed linear and scan statistics for the independent and identically distributed (i.i.d) Gaussian distributions with the identity covariance matrix. They also derived a detection boundary under sparse alternatives. Instead of the L_2 aggregation, by taking the element maximum of $\mathcal{C}$, Jirak (2015) proposed the L_∞-based test statistic. Allowing for both temporal and cross-sectional dependence, Jirak (2015) obtained the critical value of the test by the Gumbel distribution. To improve the convergence rate, Jirak (2015) also considered a bootstrap approximation for their test statistic. Based on the hard thresholded L_1 aggregation of $\mathcal{C}$, Cho and Fryzlewicz (2015) investigated a sparse binary segmentation method for the second-order structure change point estimation of high dimensional time series. Cho et al. (2016) proposed a class of double CUSUM statistics by aggregating the ordered CUSUM statistics at each time point. They also explored the high dimensional asymptotic relative efficiency. In different settings, Aston and Kirch (2018); Wang and Samworth (2018) investigated the high dimensional change point detection and estimation using projection methods, and Chen et al. (2015) adopted a graph-based method to detect and identify change points. Recently, Dette and Gösmann (2018) considered the relative change point detection problem in high dimensions.

In view of the existing methods mentioned above, aggregations of $\mathcal{C}$ fall into 2 categories: L_2-type versus L_∞-type tests. For each time point k ($1 \leq k \leq n-1$), on one hand, by accumulating small deviations of all coordinates, the L_2-type tests aim to detect relative dense signals; on the other hand, the L_∞-type tests are more sensitive to sparse signals with strong perturbations on a small number of coordinates. Either L_2 or L_∞-based method only works well for a particular alternative pattern. However, in real applications, the alternative structure (sparse or dense) is usually unknown. Theoretically, Cox and Hinkley (1979) showed that there is no uniformly powerful test under all alternative scenarios.

Hence, it is of great interests to construct data-adaptive tests. Furthermore, Zhang et al. (2010); Horváth and Hušková (2012); Jirak (2015); Enikeeva and Harchaoui (2019) also required strong spatial dependency structures among the components (e.g. the identity covariance matrix) or specific distributional assumptions (e.g. the Gaussian assumption) to guarantee the validity of their theories. It is shown that those methods are no longer applicable once their model assumptions are not satisfied. Moreover, most exiting works only consider the high dimensional change point detection for a particular parameter such as mean. In some real problems, however, many other parameters such as the covariance and Kendall's tau correlation matrices may be of interest. To overcome this limitation, we propose a general framework for solving a wider range of problems and also adaptive to various alternative scenarios, which are robust with respect to the population distributions and the covariance structures. Our main contributions of this paper can be summarized as follows:

- We provide a general framework using U-statistics for change point detection in high dimensions. Our framework covers many existing methods for high dimensional change point problems such as mean or covariance as special cases. Because of the flexibility of U-statistics, our general framework can also solve many other problems such as the Kendall's correlation coefficient or one sample Wilcoxon's rank test, which have not been studied for high dimensional settings.

- We construct two types of new individual test statistics $T_{(s_0,p)}$ and $W_{(s_0,p)}$ ($1 \leq p \leq \infty$), by considering the change point locations and the alternative patterns simultaneously. Our two new statistics aggregate the U-statistic-based CUSUM matrix in a different way, using the (s_0, p)-norm* proposed by Zhou et al. (2018). We investigate the theoretical properties of these two types of statistics in terms of size and power. By definition, $T_{(s_0,p)}$ considers the scenario where the change point occurs at a common time point across the components; $W_{(s_0,p)}$ allows that different coordinates can have different change points of time. With the above flexible aggregations, many existing techniques can be viewed as special cases of our individual test statistics by choosing a particular s_0 or p. Moreover, both $T_{(s_0,p)}$ and $W_{(s_0,p)}$ can capture the alternative pattern by choosing a particular p. In other words, there is at least one test in $T_{(s_0,p)}$ (or $W_{(s_0,p)}$) with $1 \leq p \leq \infty$ powerful under a given alternative scenario. Theoretically, for the two types of individual test statistics, we introduce a high dimensional multiplier bootstrap method in Algorithm 1 to approximate their limiting distributions, respectively. The proposed bootstrap method only requires mild conditions on the covariance structures and the population distributions, and is free of tuning parameters. Fundamentally, it is a nontrivial extension of the low dimensional bootstrap method in Gombay and Horváth (1999); Bücher and Kojadinovic (2016) to the high dimensional case. It is shown that both two types of individual tests can control the type I error under the nominal significance level and reject the null hypothesis with probability tending to one asymptotically.

- Combining the corresponding individual tests in $T_{(s_0,p)}$ (or $W_{(s_0,p)}$) with $1 \leq p \leq \infty$, we further construct the data-adaptive test statistic T_{ad} (or W_{ad}) for the general hypothesis (1.1). The proposed data-adaptive methods T_{ad} and W_{ad} choose the best test within their combinations according to the data and enjoy simultaneously high power across various alternative scenarios.

*For $\boldsymbol{v} = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, $\|\boldsymbol{v}\|_{(s_0,p)} := (\sum_{j=d-s_0+1}^{d} |v_{(j)}|^p)^{1/p}$, where $|v_{(1)}| \leq \cdots \leq |v_{(d)}|$ are the ordered statistic for $|v_1|, \ldots, |v_d|$.

Theoretically, we adopt a low-cost bootstrap method for approximating their limiting distributions. Consequently, the two data-adaptive tests have the prespecified significance level asymptotically. The power results show that our data-adaptive tests enjoy optimality in the sense that the signal-noise ratio for rejecting $\mathbf{H}_0$ reaches the minimax separation rate derived in Enikeeva and Harchaoui (2019) under sparse alternatives. To the best of our knowledge, this is the first power results for U-statistic-based change point detection problems. We also investigate the numerical performance of the proposed methods using both simulated and real datasets. The numerical studies illustrate the wider applicability and better adaptivity of our methods than the existing techniques under various model settings and alternative scenarios. Furthermore, an R package called AdaptiveCpt is available to implement our new data-adaptive tests.

Note that our paper is related to Zhou et al. (2018), which considered the one- or two-sample inference of high dimensional parameters based on U-statistics. However, our paper focuses on change point detection, which requires careful handling of the unknown change point in the data stream as well as nuisance parameters derived from model misspecifications. Furthermore, the construction of the CUSUM matrix $\mathcal{C}$ results in a sequence of dependent random vectors. Hence, our two types of individual test statistics, designed for the change point analysis, require substantial modifications for Gaussian approximations developed for i.i.d cases in Chernozhukov et al. (2017); Zhou et al. (2018), which also really differentiates this work from Zhou et al. (2018).

The rest of this paper is organized as follows. In Section 2, we present our new methodology for the general hypothesis (1.1). In Section 3, we first introduce some definitions and assumptions, and then discuss the theoretical properties of our methods in terms of size and power. In Section 4, we investigate the numerical performance of our proposed methods under various model settings and alternative scenarios. In Section 5, we apply our methods to a comparative genomic hybridization (CGH) dataset for the segmentation of DNA copy-number variation. Some discussions are provided in Section 6. Supplementary materials include both proofs of the theoretical results and additional numerical examples.

2 Methodology

We present our new methodology for the general hypothesis (1.1). In Section 2.1, we introduce some notations used for this paper. In Section 2.2, with known variances, the U-statistic-based CUSUM matrix $\widetilde{\mathcal{C}}$ is constructed. In Section 2.3, we propose two types of oracle individual test statistics $\widetilde{T}_{(s_0,p)}$ and $\widetilde{W}_{(s_0,p)}$ with $1 \leq p \leq \infty$, by considering the change point locations and alternative patterns simultaneously. In order to estimate the unknown variances in $\widetilde{T}_{(s_0,p)}$ and $\widetilde{W}_{(s_0,p)}$, we propose a jackknife estimator in Section 2.4. The high dimensional multiplier bootstrap method is introduced in Section 2.5 to obtain the limiting distributions of $T_{(s_0,p)}$ and $W_{(s_0,p)}$ with $1 \leq p \leq \infty$. In real applications, the alternative patterns are unknown. Consequently, it is desirable to construct data-adaptive methods. In Section 2.6, we combine the individual tests in $T_{(s_0,p)}$ (or $W_{(s_0,p)}$) with $1 \leq p \leq \infty$ and propose the data-adaptive test statistic T_{ad} (or W_{ad}). A low-cost bootstrap method is adopted to efficiently approximate the limiting distribution of T_{ad} (or W_{ad}).

Before presenting our methodology, we first introduce our setting for the general high dimensional change point model. Denote $\gamma \in \{1, \ldots, q\}$ as the total number of coordinates with a change point. We

set $\Pi_\gamma = \{s_i \in \{1,\ldots,q\} : \text{there is a change point for } \theta_{s_i} \text{ with } 1 \le i \le \gamma\}$ as the set of coordinates with a change point. We assume that the relative change point location $\widetilde{t}_s$ is bounded away from the beginning or the end of the dataset, which is a common assumption in the literature. In other words, there exists $\tau_0 \in (0, 0.5)$ such that $\tau_0 \le \widetilde{t}_s \le 1 - \tau_0$ for $s \in \Pi_\gamma$.

2.1 Notations

We define the L_p norm as $\|\boldsymbol{v}\|_p = (\sum_{j=1}^d |v_j|^p)^{1/p}$ for $\boldsymbol{v} = (v_1,\ldots,v_d)^\top \in \mathbb{R}^d$. For $p = \infty$, $\|\boldsymbol{v}\|_\infty = \max_{1\le j\le d}|v_j|$. For $p = 0$, $\|\boldsymbol{v}\|_0 := \#\{j : v_j \neq 0\}$, where $\#\{\mathcal{S}\}$ denotes the cardinality of a set $\mathcal{S}$. For two real numbered sequences a_n and b_n, we set $a_n = O(b_n)$ if there exits a constant C such that $|a_n| \le C|b_n|$ for a sufficiently large n; $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$; $a_n \asymp b_n$ if there exists constants c and C such that $c|b_n| \le |a_n| \le C|b_n|$ for a sufficiently large n. For a sequence of random variables (r.v.s) $\{\xi_1, \xi_2, \ldots\}$, we set $\xi_n \xrightarrow{\mathbb{P}} \xi$ if ξ_n converges to ξ in probability as $n \to \infty$. We also denote $\xi_n = o_p(1)$ if $\xi_n \xrightarrow{\mathbb{P}} 0$. For a $p \times q$-dimensional matrix $\mathbf{A} = (a_{i,j})$ with $1 \le i \le p$ and $1 \le j \le q$, denote $\text{vec}(\mathbf{A}) = (a_{1,1},\ldots,a_{1,q},\ldots,a_{p,1},\ldots,a_{p,q})^\top$ as the vectorized form of $\mathbf{A}$. We define $\lfloor x \rfloor$ as the largest integer less than or equal to x for $x \ge 0$.

2.2 U-statistic-based CUSUM matrix

We now introduce our methodology for the general hypothesis (1.1). We are interested in the q-dimensional parameter $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_q)^\top$ with $\theta_s = \mathbb{E}\Phi_s(\boldsymbol{X}'_1,\ldots,\boldsymbol{X}'_m)$, where Φ_s is a measurable and symmetric kernel with an order m, and $\boldsymbol{X}'_1,\ldots,\boldsymbol{X}'_m$ are independent copies with the same distributions of $\boldsymbol{X}$. Given the sample observations $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$, for each coordinate s, under $\mathbf{H}_0$, we can estimate θ_s by the following U-statistics

$$\widehat{\theta}_{n,s} = \binom{n}{m}^{-1} \sum_{1\le k_1<\cdots<k_m\le n} \Phi_s(\boldsymbol{X}_{k_1},\ldots,\boldsymbol{X}_{k_m}), \text{ with } s = 1,\ldots,q.$$

Let $\Psi_s(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_m) = \Phi_s(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_m) - \theta_s$ be the centralized kernel. By the well-known Hoeffding's decomposition (Hoeffding (1948)), under $\mathbf{H}_0$, we can decompose $\widehat{\theta}_{n,s} - \theta_s$ as

$$\widehat{\theta}_{n,s} - \theta_s = \frac{m}{n}\sum_{i=1}^n h_{1,s}(\boldsymbol{X}_i) + \binom{n}{m}^{-1} \sum_{1\le k_1<\cdots<k_m\le n} h_{2,s}(\boldsymbol{X}_{k_1},\ldots,\boldsymbol{X}_{k_m}),$$

where

$$h_{1,s}(\boldsymbol{x}) := \mathbb{E}\Psi_s(\boldsymbol{x},\boldsymbol{X}_2,\ldots,\boldsymbol{X}_m), h_{2,s}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m) := \Phi_s(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m) - \sum_{i=1}^m h_{1,s}(\boldsymbol{x}_i) - \theta_s. \quad (2.1)$$

Denote $\lfloor nt \rfloor^* = n - \lfloor nt \rfloor$ for $0 \le t \le 1$. To test whether there is a change point for θ_s during n

observations, for each time point $\lfloor nt \rfloor$, we estimate θ_s before and after $\lfloor nt \rfloor$ and obtain

$$\begin{aligned} \widehat{\theta}_{\lfloor nt \rfloor, s} &= \binom{\lfloor nt \rfloor}{m}^{-1} \sum_{1 \le k_1 < \cdots < k_m \le \lfloor nt \rfloor} \Phi_s(\boldsymbol{X}_{k_1}, \ldots, \boldsymbol{X}_{k_m}), \\ \widehat{\theta}_{\lfloor nt \rfloor^*, s} &= \binom{\lfloor nt \rfloor^*}{m}^{-1} \sum_{(\lfloor nt \rfloor + 1) \le k_1 < \cdots < k_m \le n} \Phi_s(\boldsymbol{X}_{k_1}, \ldots, \boldsymbol{X}_{k_m}). \end{aligned} \tag{2.2}$$

Then based on (2.2), we define the following U-statistic typed CUSUM statistic for each θ_s:

$$\widetilde{C}_s(\lfloor nt \rfloor) = \sqrt{n} \frac{\lfloor nt \rfloor}{n} \frac{\lfloor nt \rfloor^*}{n} \sigma_{s,s}^{-1/2} \left(\widehat{\theta}_{\lfloor nt \rfloor, s} - \widehat{\theta}_{\lfloor nt \rfloor^*, s} \right), \text{ with } s = 1, \ldots, q, \text{ and } 0 \le t \le 1, \tag{2.3}$$

where $\sigma_{s,s} = \mathrm{Var}(h_{1,s}(\boldsymbol{X}_1))$. With known variances $(\sigma_{s,s})_{s=1}^q$, we can use (2.3) to construct the oracle U-statistic-based CUSUM matrix $\widetilde{\mathcal{C}} = (\widetilde{C}_s(\lfloor nt \rfloor))$ with $1 \le s \le q$ and $\tau_0 \le t \le 1 - \tau_0$ as follows:

$$\widetilde{\mathcal{C}} = \begin{pmatrix} \widetilde{C}_1(\lfloor n\tau_0 \rfloor) & \ldots, & \widetilde{C}_1(\lfloor n(1-\tau_0) \rfloor) \\ \widetilde{C}_2(\lfloor n\tau_0 \rfloor) & \ldots, & \widetilde{C}_2(\lfloor n(1-\tau_0) \rfloor) \\ \vdots & \cdots & \vdots \\ \widetilde{C}_q(\lfloor n\tau_0 \rfloor) & \ldots, & \widetilde{C}_q(\lfloor n(1-\tau_0) \rfloor) \end{pmatrix}.$$

Remark 2.1. Note that $\widetilde{C}_s(\lfloor nt \rfloor)$ is the generalization of the CUSUM statistic. For example, for $m = 1$ and $\theta_s = \mathbb{E}X_s$, $\widetilde{C}_s(\lfloor nt \rfloor)$ reduces to the CUSUM statistic in (1.2) for the mean change point detection; for $m = 2$ and $\theta_{i,j} = \mathbb{E}((X_i - X_i')(X_j - X_j')/2)$, $\widetilde{C}_{i,j}(\lfloor nt \rfloor)$ is the CUSUM statistic for the variance/covariance change point detection. More examples can be found in Csörgő and Horváth (1997).

2.3 Two new (s_0, p)-norm-based test statistics

In Section 2.2, we introduce the U-statistic-based CUSUM matrix $\widetilde{\mathcal{C}}$. The challenge for high dimensional change point detection is how to aggregate $\widetilde{\mathcal{C}}$ efficiently. In this section, we introduce two methods for aggregating $\widetilde{\mathcal{C}}$ according to the alternative structures, and the change point locations. To that end, we investigate the following two different scenarios.

2.3.1 Case I: common change point location

We first consider the case where all γ coordinates with a change point have a common change point location. In other words, there exists $t^* \in [\tau_0, 1 - \tau_0]$ such that $\widetilde{t}_s = t^*$ for all $s \in \Pi_\gamma$. For a simple illustration, we generate a 200×100 data matrix $\mathbf{X} = (X_{i,j})$. The rows of $\mathbf{X}$ correspond to the observations, and the columns correspond to the coordinates. We generate a constant mean shift for some columns after a change point. In particular, we generate $\mathbf{X}$ from the following mean shift model:

$$X_{i,j} = \begin{cases} \epsilon_{i,j}, \text{ for } 1 \le i \le n, \text{ and } j \in (\Pi_\gamma)^c, \\ \epsilon_{i,j}, \text{ for } 1 \le i \le \lfloor n\widetilde{t}_j \rfloor, \text{ and } j \in \Pi_\gamma, \\ \delta + \epsilon_{i,j}, \text{ for } \lfloor n\widetilde{t}_j \rfloor + 1 \le i \le n, \text{ and } j \in \Pi_\gamma, \end{cases} \tag{2.4}$$

where $\epsilon_{i,j}$ is i.i.d $N(0,1)$ random variable for $1 \le i \le 200$ and $1 \le j \le 100$. For Model (2.4), we set the change point location at $\widetilde{t}_j = 0.5$ for all $j \in \Pi_\gamma$. We also consider two different alternative

scenarios: the sparse case with a small γ, and the dense case with a large γ. For the sparse case, we set $\gamma = 5$, $\Pi_\gamma = \{10, 30, 50, 70, 90\}$, and the mean jump $\delta = 1$. For the dense case, we set $\gamma = 50$, $\Pi_\gamma = \{1, 3, 5, \ldots, 97, 99\}$, and $\delta = 0.5$.

Figure 1 shows the corresponding heatmap of the CUSUM matrix $\widetilde{\mathcal{C}}$, and the CUSUM charts for coordinates in Π_γ for the sparse and dense cases, respectively. From the heatmap and the CUSUM charts, we can make several observations about $\widetilde{\mathcal{C}}$. The rows of $\widetilde{\mathcal{C}}$ with a change point have very different values. In particular, the column of $\widetilde{\mathcal{C}}$ at the middle of the observations contains the largest test statistic. Furthermore, the CUSUM chart is maximized near the true change point location $t^* = 0.5$ for all entries in Π_γ. Therefore, to efficiently aggregate $\widetilde{\mathcal{C}}$, we need to take the alternative pattern (rows of $\widetilde{\mathcal{C}}$) and the change point location (columns of $\widetilde{\mathcal{C}}$) into consideration simultaneously. In order to construct our new

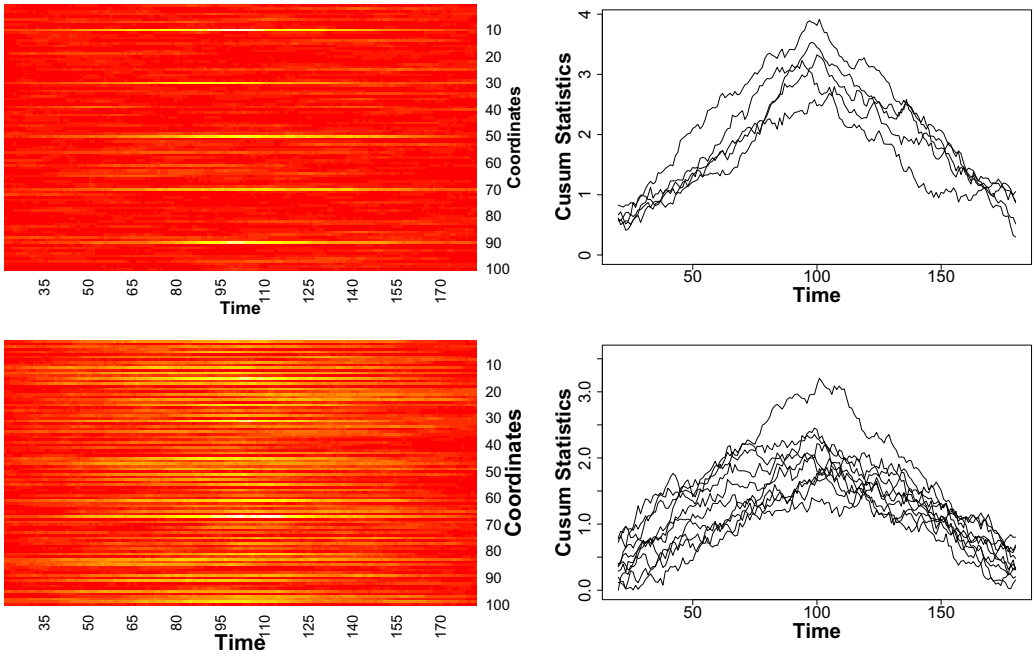

Figure 1: Heatmap and CUSUM charts for sparse and dense alternatives with a common change point location. Top left: Heatmap of $\widetilde{\mathcal{C}}$ for $\gamma = 5$; Top right: CUSUM charts of the corresponding 5 components in Π_5. Bottom left: Heatmap of $\widetilde{\mathcal{C}}$ for $\gamma = 50$; Bottom right: CUSUM charts of 10 randomly selected components in Π_{50}.

test statistics, we consider both the change point location and the alternative pattern. For the change point location, let $\widetilde{\boldsymbol{C}}(\lfloor nt \rfloor) = (\widetilde{C}_1(\lfloor nt \rfloor), \ldots, \widetilde{C}_q(\lfloor nt \rfloor))^\top$ be the column of the CUSUM matrix $\widetilde{\mathcal{C}}$ at the location t with $\tau_0 \leq t \leq 1 - \tau_0$. In this example, there is a common change point location at $t^* = 0.5$. Therefore, $\widetilde{\boldsymbol{C}}(\lfloor nt^* \rfloor)$ contains the most information among all columns of $\widetilde{\mathcal{C}}$. For the alternative pattern, it can be either sparse or dense in real applications. To capture the underlying (unknown) alternative structures, we adopt the (s_0, p)-norm proposed by Zhou et al. (2018) to aggregate $\widetilde{\boldsymbol{C}}(\lfloor nt^* \rfloor)$. The (s_0, p)-norm is essentially the L_p norm of the s_0-largest entries of the corresponding vector. In particular, for $\boldsymbol{v} = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, with $|v_{(1)}| \leq \cdots \leq |v_{(d)}|$ being the ordered statistic for $|v_1|, \ldots, |v_d|$, its (s_0, p)-norm is defined as

$$\|\boldsymbol{v}\|_{(s_0,p)} := \Big(\sum_{j=d-s_0+1}^{d} |v_{(j)}|^p \Big)^{1/p}.$$

By adopting the (s_0, p)-norm, for a given s_0, $\|\widetilde{\boldsymbol{C}}(\lfloor nt^* \rfloor)\|_{(s_0,p)}$ with a small p (e.g. $p = 1, 2$) is more sensitive to dense alternatives with a small change on many entries; $\|\widetilde{\boldsymbol{C}}(\lfloor nt^* \rfloor)\|_{(s_0,p)}$ with a large p (e.g. $p = \infty$) is more sensitive to sparse alternatives with a large change on a few entries. As t^* is typically unknown, we scan all possible locations with $t \in [\tau_0, 1 - \tau_0]$. Hence, for a fixed s_0, with known $\sigma_{s,s}$, we propose the (s_0, p)-norm-based oracle individual test statistic as follows:

$$\widetilde{T}_{(s_0,p)} = \max_{\tau_0 \le t \le 1-\tau_0} \left\|\widetilde{\boldsymbol{C}}(\lfloor nt \rfloor)\right\|_{(s_0,p)}, \text{ for } 1 \le p \le \infty. \tag{2.5}$$

By definition, $\widetilde{T}_{(s_0,p)}$ generalizes the existing methods for aggregating $\widetilde{\mathcal{C}}$ by choosing a proper s_0 or p. For example, by setting $s_0 = d$ and $p = 2$, $\widetilde{T}_{(s_0,p)}$ is the L_2 aggregation in Zhang et al. (2010); by setting $p = \infty$, $\widetilde{T}_{(s_0,p)}$ is the L_∞ aggregation in Jirak (2015); by setting $s_0 = s_0(\pi_{\text{thr}})$ and $p = 1$, $\widetilde{T}_{(s_0,p)}$ is the thresholded L_1 aggregation in Cho and Fryzlewicz (2015), where π_{thr} is a threshold variable. Furthermore, $\widetilde{T}_{(s_0,p)}$ fully takes the alternative pattern and the common change point location into consideration. As shown in our numerical studies, for any given alternative structure, there is at least one test in $\widetilde{T}_{(s_0,p)}$ with $1 \le p \le \infty$ performing well by choosing a proper p. Note that s_0 is a pre-specified parameter. The following sensitivity analysis shows that the individual tests are robust against the choice of s_0 given it is not too small. More discussions about its choice are provided in Section 4.

2.3.2 Case II: different change point locations

In Section 2.3.1, we propose $\widetilde{T}_{(s_0,p)}$ in (2.5) for the general hypothesis (1.1) when all coordinates in Π_γ have a common change point location. In real applications, the change point time $\widetilde{t}_s$ for $s \in \Pi_\gamma$ can be different. To illustrate this scenario briefly, we also generate a 200×100 dataset by Model (2.4). The only difference between Case II and Case I is that the change point location $\widetilde{t}_s$ for θ_s ($s \in \Pi_\gamma$) can be different in the current example. In particular, for the sparse case with $\gamma = 5$, we set the change point locations for the 5 components with a change point at $\widetilde{t}_s = 0.25, 0.35, 0.45, 0.55$, and 0.65, respectively; for the dense case with $\gamma = 50$, we randomly divide the corresponding 50 change point locations into 5 groups, and each group contains 10 components having a common change point location. Specifically, the 5 groups are $\{0.15, \cdots, 0.15\}$ (group 1), $\{0.3, \cdots, 0.3\}$ (group 2), $\{0.45, \cdots, 0.45\}$ (group 3), $\{0.6, \cdots, 0.6\}$ (group 4), and $\{0.75, \cdots, 0.75\}$ (group 5).

Figure 2 shows the heatmap and the CUSUM charts for Case II. Similar to Case I, the rows of $\widetilde{\mathcal{C}}$ with a change point have more different values than those without. Different from Case I, the information for the column of $\widetilde{\mathcal{C}}$ is not centered at a certain location. Moreover, each column $\widetilde{\boldsymbol{C}}(\lfloor nt \rfloor)$ with $\tau_0 \le t \le 1 - \tau_0$ only contains at most 1 out of 5 components in Π_γ for the sparse case, and 10 out of 50 for the dense case. Consequently, it is inefficient to first aggregate $\widetilde{\boldsymbol{C}}(\lfloor nt \rfloor)$ with the (s_0, p)-norm at each time point $\lfloor nt \rfloor$ with $\tau_0 \le t \le 1 - \tau_0$.

To construct powerful test statistics for Case II, we first aggregate the rows of $\widetilde{\mathcal{C}}$ instead of aggregating the columns. In particular, let $\widetilde{B}_s = \max_{\tau_0 \le t \le 1-\tau_0} |\widetilde{C}_s(\lfloor nt \rfloor)|$ be the maximum of the absolute values of CUSUM statistic for θ_s with $1 \le s \le q$. Let $\widetilde{\boldsymbol{B}} = (\widetilde{B}_1, \ldots, \widetilde{B}_q)^\top$ be the aggregated vector. Then we aggregate $\widetilde{\boldsymbol{B}}$ with the (s_0, p)-norm, and propose the (s_0, p)-norm-based oracle individual test statistic as follows:

$$\widetilde{W}_{(s_0,p)} = \left\|\widetilde{\boldsymbol{B}}\right\|_{(s_0,p)}, \text{ for } 1 \le p \le \infty. \tag{2.6}$$

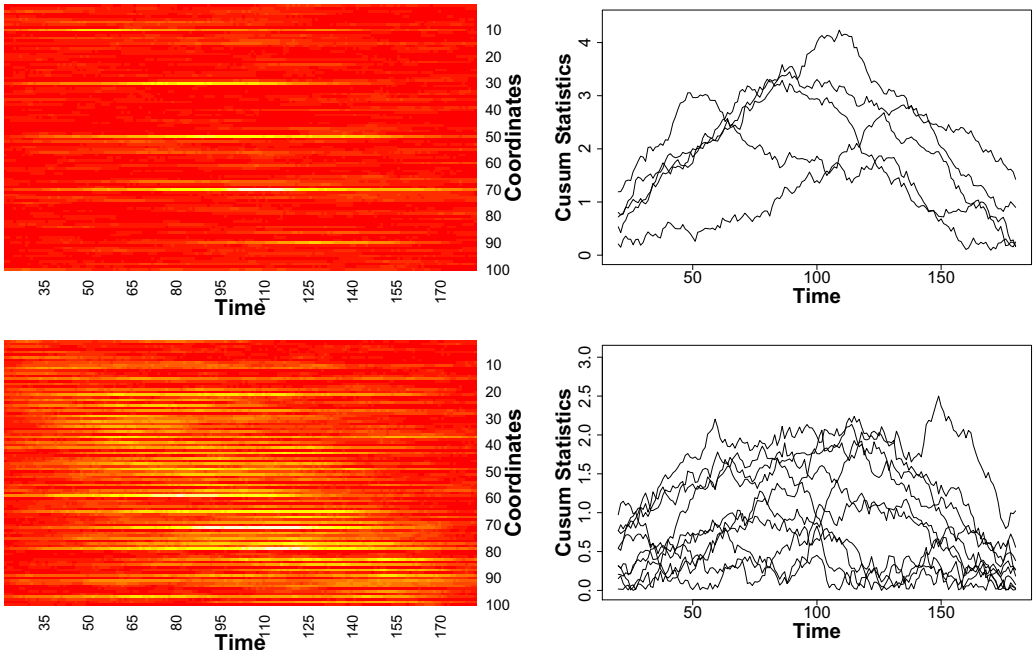

Figure 2: Heatmap and CUSUM charts for sparse and dense alternatives with different change point locations. Top left: Heatmap of $\widetilde{\mathcal{C}}$ for $\gamma = 5$; Top right: CUSUM charts of the corresponding 5 components in Π_5. Bottom left: Heatmap of $\widetilde{\mathcal{C}}$ for $\gamma = 50$; Bottom right: CUSUM charts of 10 randomly selected components in Π_{50}.

By definition, $\widetilde{W}_{(s_0,p)}$ has the following properties: For $p = \infty$, it equals to $\widetilde{T}_{(s_0,p)}$, and also the L_∞ aggregation in Jirak (2015) and Yu and Chen (2017). In addition, by allowing different change point locations among the coordinates, $\widetilde{W}_{(s_0,p)}$ reduces the information loss by first aggregating the rows of $\widetilde{\mathcal{C}}$. Furthermore, $\widetilde{W}_{(s_0,p)}$ can capture the alternative pattern by adopting the (s_0, p)-norm. In particular, for a fixed s_0, $\widetilde{W}_{(s_0,p)}$ with a large p (e.g. $p = \infty$) is sensitive to sparse alternatives with large deviations, and $\widetilde{W}_{(s_0,p)}$ with a small p (e.g. $p = 1, 2$) is powerful against dense alternatives with small deviations. Consequently, for any given alternative pattern, there exists at least one powerful test in $\widetilde{W}_{(s_0,p)}$ with $1 \leq p \leq \infty$. As shown by our numerical studies, $\widetilde{W}_{(s_0,p)}$ has better power performance than $\widetilde{T}_{(s_0,p)}$ when the change point locations for θ_s with $s \in \Pi_\gamma$ are different.

Note that Zhou et al. (2018) constructed their test statistics for one- or two-sample tests based on the (s_0, p)-norm. Considering the particular change point problem in (1.1), we propose our individual test statistics which are very different from Zhou et al. (2018). Specifically, we first construct the U-statistic-based CUSUM matrix $\mathcal{C}$. To aggregate $\mathcal{C}$ efficiently, we take both the change point locations and alternative structures into account, and propose two types of individual test statistics. With the above flexible aggregations, many existing statistics can be viewed as special cases of our individual test statistics. We believe our proposed test statistics are novel in the context of high dimensional change point analysis. In contrast, there is no need to deal with the unknown change point in Zhou et al. (2018). They mainly developed their test statistics using the aggregation of U-statistic-based vectors.

2.4 Jackknife-based variance estimation for $\sigma_{s,s}$

In Section 2.3, we have introduced two types of the (s_0, p)-norm-based oracle individual test statistics $\widetilde{T}_{(s_0,p)}$ and $\widetilde{W}_{(s_0,p)}$. By definition, both statistics depend on $\sigma_{s,s} = \mathrm{Var}(h_{1,s}(\boldsymbol{X}))$. The Hoeffding's projection $h_{1,s}(\boldsymbol{X})$ in (2.1) is unknown. Consequently, we can not apply $\widetilde{T}_{(s_0,p)}$ and $\widetilde{W}_{(s_0,p)}$ directly for the hypothesis (1.1) because of the unknown $\sigma_{s,s}$. In this section, we introduce a jackknife-based variance estimator for $\sigma_{s,s}$.

Note that jackknife is a widely used method for estimating the unknown variances in U-statistics. For example, Zhou et al. (2018) adopted jackknife for estimating the variances in their test statistics. However, since there is no concern about the unknown change point, the estimation in Zhou et al. (2018) is essentially based on i.i.d observations. Different from Zhou et al. (2018), in what follows, we propose jackknife-based pooled variance estimators $\{\widehat{\sigma}_{\lfloor nt\rfloor,s,s}, \tau_0 \le t \le 1-\tau_0, 1 \le s \le q\}$, to deal with the unknown change point, which is an established approach in the context of change point analysis.

We first consider the order $m \ge 2$ for the kernel used in (1.1). For each time point $\lfloor nt\rfloor$ with $t \in [\tau_0, 1-\tau_0]$, we define the jackknife estimator for $h_{1,s}(\boldsymbol{X}_k)$ before and after $\lfloor nt\rfloor$ as:

$$Q_{\lfloor nt\rfloor,s,k} := \binom{\lfloor nt\rfloor - 1}{m-1}^{-1} \sum_{\substack{1\le \ell_1<\cdots<\ell_{m-1}\le \lfloor nt\rfloor \\ \ell_j\neq k, j=1,\ldots,m-1}} \Phi_s(\boldsymbol{X}_k, \boldsymbol{X}_{\ell_1}, \ldots, \boldsymbol{X}_{\ell_{m-1}}), \text{ for } 1 \le k \le \lfloor nt\rfloor,$$
$$Q_{\lfloor nt\rfloor^*,s,k} := \binom{\lfloor nt\rfloor^* - 1}{m-1}^{-1} \sum_{\substack{\lfloor nt\rfloor+1\le \ell_1<\cdots<\ell_{m-1}\le n \\ \ell_j\neq k, j=1,\ldots,m-1}} \Phi_s(\boldsymbol{X}_k, \boldsymbol{X}_{\ell_1}, \ldots, \boldsymbol{X}_{\ell_{m-1}}), \text{ for } \lfloor nt\rfloor+1 \le k \le n.$$

Then based on $Q_{\lfloor nt\rfloor,s,k}$ and $Q_{\lfloor nt\rfloor^*,s,k}$, our jackknife estimator for $\sigma_{s,s}$ is defined as follows:

$$\widehat{\sigma}_{\lfloor nt\rfloor,s,s} = \frac{1}{n}\Big(\sum_{k=1}^{\lfloor nt\rfloor}(Q_{\lfloor nt\rfloor,s,k} - \widehat{\theta}_{\lfloor nt\rfloor,s})^2 + \sum_{k=\lfloor nt\rfloor+1}^{n}(Q_{\lfloor nt\rfloor^*,s,k} - \widehat{\theta}_{\lfloor nt\rfloor^*,s})^2\Big),$$

where $\widehat{\theta}_{\lfloor nt\rfloor,s}$ and $\widehat{\theta}_{\lfloor nt\rfloor^*,s}$ are defined in (2.2).

Next we consider the kernel with an order $m = 1$, where $\widehat{\theta}_{\lfloor nt\rfloor,s}$ and $\widehat{\theta}_{\lfloor nt\rfloor^*,s}$ are reduced to

$$\widehat{\theta}_{\lfloor nt\rfloor,s} = \frac{1}{\lfloor nt\rfloor}\sum_{k=1}^{\lfloor nt\rfloor}\Phi_s(\boldsymbol{X}_k), \quad \text{and} \quad \widehat{\theta}_{\lfloor nt\rfloor^*,s} = \frac{1}{\lfloor nt\rfloor^*}\sum_{k=\lfloor nt\rfloor+1}^{n}\Phi_s(\boldsymbol{X}_k),$$

and $Q_{\lfloor nt\rfloor,s,k}$, $Q_{\lfloor nt\rfloor^*,s,k}$ are reduced to $\Phi_s(\boldsymbol{X}_k)$. Consequently, for $m = 1$, the jackknife estimator in (2.4) reduces to the classical pooled variance estimation in Csörgő and Horváth (1997):

$$\widehat{\sigma}_{\lfloor nt\rfloor,s,s} = \frac{1}{n}\Big(\sum_{k=1}^{\lfloor nt\rfloor}(\Phi_s(\boldsymbol{X}_k) - \widehat{\theta}_{\lfloor nt\rfloor,s})^2 + \sum_{k=\lfloor nt\rfloor+1}^{n}(\Phi_s(\boldsymbol{X}_k) - \widehat{\theta}_{\lfloor nt\rfloor^*,s})^2\Big).$$

Under $\mathbf{H}_0$, our theoretical results show that $\max_{1\le s\le q}\max_{\tau_0\le t\le 1-\tau_0}|\widehat{\sigma}_{\lfloor nt\rfloor,s,s} - \sigma_{s,s}| = o_p(1)$ (see Remark C.1 in the Supplementary Material). We can replace $\sigma_{s,s}$ in (2.3) by its estimator $\widehat{\sigma}_{\lfloor nt\rfloor,s,s}$ for $1 \le s \le q$, and define the following data-driven CUSUM statistic:

$$C_s(\lfloor nt\rfloor) = \sqrt{n}\frac{\lfloor nt\rfloor}{n}\frac{\lfloor nt\rfloor^*}{n}\widehat{\sigma}^{-1/2}_{\lfloor nt\rfloor,s,s}\left(\widehat{\theta}_{\lfloor nt\rfloor,s} - \widehat{\theta}_{\lfloor nt\rfloor^*,s}\right), \quad \text{and} \quad B_s = \max_{\tau_0\le t\le 1-\tau_0}|C_s(\lfloor nt\rfloor)|.$$

Define $\boldsymbol{C}(\lfloor nt \rfloor) = (C_1(\lfloor nt \rfloor), \ldots, C_q(\lfloor nt \rfloor))^\top$, and $\boldsymbol{B} = (B_1, \ldots, B_q)^\top$. Note that there are no unknown parameters in $\boldsymbol{C}(\lfloor nt \rfloor)$ and $\boldsymbol{B}$. Then we define our two types of the (s_0, p)-norm-based individual test statistics as follows:

$$T_{(s_0,p)} = \max_{\tau_0 \le t \le 1-\tau_0} \left\| \boldsymbol{C}(\lfloor nt \rfloor) \right\|_{(s_0,p)}, \quad \text{and} \quad W_{(s_0,p)} = \left\| \boldsymbol{B} \right\|_{(s_0,p)}. \tag{2.7}$$

Throughout this paper, we use $T_{(s_0,p)}$ and $W_{(s_0,p)}$ with $1 \le p \le \infty$ as our individual test statistics for the general hypothesis (1.1).

2.5 Bootstrap procedure for the asymptotic distributions of $T_{(s_0,p)}$ and $W_{(s_0,p)}$

In Section 2.4, we introduce two types of the (s_0, p)-norm-based individual test statistics $T_{(s_0,p)}$ and $W_{(s_0,p)}$ with $1 \le p \le \infty$. As n, d, and q go to infinity, it is difficult to obtain their limiting distributions directly. To overcome this problem, in this section, we introduce the multiplier bootstrap method for approximating the limiting distributions of $T_{(s_0,p)}$ and $W_{(s_0,p)}$.

We first review the literature on the multiplier bootstrap method both in low dimensions, and high dimensions. In the low dimensional case, the multiplier bootstrap method is well studied. For example, Janssen (1994); Wang and Jing (2004) investigated this method for U-statistics. Gombay and Horváth (1999); Bücher and Kojadinovic (2016) applied the multiplier bootstrap method for U-statistic-based change point detection with independent, and dependent observations, respectively. In the high dimensional setting, Chernozhukov et al. (2017) first introduced the multiplier bootstrap method for sums of independent random vectors. In particular, let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be centered d-dimensional independent random vectors, then the multiplier bootstrap version of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are $\epsilon_1 \boldsymbol{X}_1, \ldots, \epsilon_n \boldsymbol{X}_n$, where $(\epsilon_i)_{i=1}^n$ are i.i.d $N(0,1)$ random variables. According to Chernozhukov et al. (2017), for some set A, we can approximate $\mathbb{P}(n^{-1/2} \sum_{i=1}^n \boldsymbol{X}_i \in A)$ by $\mathbb{P}(n^{-1/2} \sum_{i=1}^n \epsilon_i \boldsymbol{X}_i \in A | \mathcal{X})$ under some regular conditions. As extensions, Chen (2018) and Zhou et al. (2018) generalized the multiplier bootstrap scheme to the high dimensional U-statistic-based vectors with the L_∞-norm, and the (s_0, p)-norm, respectively. Jirak (2015) also applied this method for the high dimensional mean change point detection, and obtained the critical value for their L_∞-based test statistics.

To approximate the limiting distributions of $T_{(s_0,p)}$ and $W_{(s_0,p)}$, we investigate the multiplier bootstrap method for the high dimensional U-process with the (s_0, p)-norm. Specifically, let $\varepsilon_1^b, \ldots, \varepsilon_n^b$ be i.i.d $N(0,1)$ random variables with $b = 1, \ldots, B$. For $1 \le s \le q$ and $\tau_0 \le t \le 1 - \tau_0$, we set

$$\begin{aligned}
\widehat{\theta}^b_{\lfloor nt \rfloor, s} &= \binom{\lfloor nt \rfloor}{m}^{-1} \sum_{1 \le k_1 < \cdots < k_m \le \lfloor nt \rfloor} (\varepsilon^b_{k_1} + \cdots + \varepsilon^b_{k_m}) \big(\Phi_s(\boldsymbol{X}_{k_1}, \ldots, \boldsymbol{X}_{k_m}) - \widehat{\theta}_{\lfloor nt \rfloor, s} \big), \\
\widehat{\theta}^b_{\lfloor nt \rfloor^*, s} &= \binom{\lfloor nt \rfloor^*}{m}^{-1} \sum_{(\lfloor nt \rfloor + 1) \le k_1 < \cdots < k_m \le n} (\varepsilon^b_{k_1} + \cdots + \varepsilon^b_{k_m}) \big(\Phi_s(\boldsymbol{X}_{k_1}, \ldots, \boldsymbol{X}_{k_m}) - \widehat{\theta}_{\lfloor nt \rfloor^*, s} \big),
\end{aligned} \tag{2.8}$$

as the bootstrap version of $\widehat{\theta}_{\lfloor nt \rfloor, s}$ and $\widehat{\theta}_{\lfloor nt \rfloor^*, s}$, where $\widehat{\theta}_{\lfloor nt \rfloor, s}$ and $\widehat{\theta}_{\lfloor nt \rfloor^*, s}$ are defined in (2.2). Then, based on $\widehat{\theta}^b_{\lfloor nt \rfloor, s}$ and $\widehat{\theta}^b_{\lfloor nt \rfloor^*, s}$, we define the corresponding bootstrap version of $C_s(\lfloor nt \rfloor)$ and B_s as:

$$C^b_s(\lfloor nt \rfloor) = \sqrt{n} \frac{\lfloor nt \rfloor}{n} \frac{\lfloor nt \rfloor^*}{n} \widehat{\sigma}^{-1/2}_{\lfloor nt \rfloor, s, s} \left(\widehat{\theta}^b_{\lfloor nt \rfloor, s} - \widehat{\theta}^b_{\lfloor nt \rfloor^*, s} \right), \quad \text{and} \quad B^b_s = \max_{\tau_0 \le t \le 1-\tau_0} |C^b_s(\lfloor nt \rfloor)|. \tag{2.9}$$

Based on (2.9), we define the vector-valued process $\boldsymbol{C}^b(\lfloor nt \rfloor)$ and $\boldsymbol{B}^b$ as:

$$\boldsymbol{C}^b(\lfloor nt \rfloor) = \big(C_1^b(\lfloor nt \rfloor), \ldots, C_q^b(\lfloor nt \rfloor)\big)^\top, \quad \text{and} \quad \boldsymbol{B}^b = \big(B_1^b, \ldots, B_q^b\big)^\top. \tag{2.10}$$

Now, the b-th multiplier bootstrap version of the individual test statistics are defined as:

$$T^b_{(s_0,p)} = \max_{\tau_0 \le t \le 1-\tau_0} \big\|\boldsymbol{C}^b(\lfloor nt \rfloor)\big\|_{(s_0,p)}, \text{ and } W^b_{(s_0,p)} = \big\|\boldsymbol{B}^b\big\|_{(s_0,p)}, \text{ with } 1 \le p \le \infty. \tag{2.11}$$

Given the significance level α, let

$$c^T_{\alpha,(s_0,p)} := \inf\{t \in \mathbb{R} : \mathbb{P}(T_{(s_0,p)} \le t) \ge 1-\alpha\}, \text{ and } c^W_{\alpha,(s_0,p)} := \inf\{t \in \mathbb{R} : \mathbb{P}(W_{(s_0,p)} \le t) \ge 1-\alpha\}$$

be the oracle critical values for $T_{(s_0,p)}$ and $W_{(s_0,p)}$, respectively. Based on the bootstrap samples $\{T^1_{(s_0,p)}, \ldots, T^B_{(s_0,p)}\}$ and $\{W^1_{(s_0,p)}, \ldots, W^B_{(s_0,p)}\}$, we can estimate $c^T_{\alpha,(s_0,p)}$ and $c^W_{\alpha,(s_0,p)}$ by

$$\widehat{c}^T_{\alpha,(s_0,p)} = \inf\Big\{t : \frac{\sum_{b=1}^B \mathbf{1}\{T^b_{(s_0,p)} \le t\}}{B} \ge 1-\alpha\Big\}, \quad \text{and } \widehat{c}^W_{\alpha,(s_0,p)} = \inf\Big\{t : \frac{\sum_{b=1}^B \mathbf{1}\{W^b_{(s_0,p)} \le t\}}{B} \ge 1-\alpha\Big\}.$$

Hence, based on the estimated critical values $\widehat{c}^T_{\alpha,(s_0,p)}$ and $\widehat{c}^W_{\alpha,(s_0,p)}$, we define our two types of the (s_0,p)-norm-based individual tests as:

$$\Psi^T_{\alpha,(s_0,p)} = \mathbf{1}\big\{T_{(s_0,p)} \ge \widehat{c}^T_{\alpha,(s_0,p)}\big\}, \quad \text{and } \Psi^W_{\alpha,(s_0,p)} = \mathbf{1}\big\{W_{(s_0,p)} \ge \widehat{c}^W_{\alpha,(s_0,p)}\big\}.$$

For $T_{(s_0,p)}$, we reject $\mathbf{H}_0$ if and only if $\Psi^T_{\alpha,(s_0,p)} = 1$. Similarly, for $W_{(s_0,p)}$, we reject $\mathbf{H}_0$ if and only if $\Psi^W_{\alpha,(s_0,p)} = 1$. Let $P_{T,(s_0,p)}$ and $P_{W,(s_0,p)}$ be the theoretical P-values for $T_{(s_0,p)}$ and $W_{(s_0,p)}$ respectively. Then based on the bootstrap samples, we can approximate them by

$$\widehat{P}_{T,(s_0,p)} = \frac{\sum_{b=1}^B \mathbf{1}\{T^b_{(s_0,p)} > T_{(s_0,p)}\}}{B+1}, \quad \text{and } \widehat{P}_{W,(s_0,p)} = \frac{\sum_{b=1}^B \mathbf{1}\{W^b_{(s_0,p)} > W_{(s_0,p)}\}}{B+1}. \tag{2.12}$$

Therefore, given the significance level α, for $T_{(s_0,p)}$, we reject $\mathbf{H}_0$ if and only if $\widehat{P}_{T,(s_0,p)} \le \alpha$. Similarly, for $W_{(s_0,p)}$, we reject $\mathbf{H}_0$ if and only if $\widehat{P}_{W,(s_0,p)} \le \alpha$.

2.6 Two types of data-adaptive tests

In Sections 2.3 and 2.5, we introduce two types of the (s_0,p)-norm-based individual test statistics $T_{(s_0,p)}$ and $W_{(s_0,p)}$, and use the multiplier bootstrap method to approximate their limiting distributions. Under a given alternative pattern, both $T_{(s_0,p)}$ and $W_{(s_0,p)}$ have different power performance among various p for a fixed s_0. For example, $T_{(s_0,1)}$ and $W_{(s_0,1)}$ are more sensitive to alternatives with small deviations on a large number of coordinates. In contrast, $T_{(s_0,\infty)}$ and $W_{(s_0,\infty)}$ are more powerful against alternatives with large perturbations on a small number of coordinates. Empirically, the alternative structure is typically unknown. Theoretically, there is also no uniformly powerful test for all alternative patterns (Cox and Hinkley (1979)). Therefore, it is desirable to construct a data-adaptive method which is simultaneously powerful under various alternative scenarios.

As a small P-value leads to rejection of $\mathbf{H}_0$, for $T_{(s_0,p)}$ and $W_{(s_0,p)}$ with $1 \le p \le \infty$, we construct

the corresponding data-adaptive test statistic as their minimum P-value. In particular, our two types of data-adaptive test statistics are as follows:

$$T_{\mathrm{ad}} = \min_{p\in\mathcal{P}} \widehat{P}_{T,(s_0,p)}, \quad \text{and} \quad W_{\mathrm{ad}} = \min_{p\in\mathcal{P}} \widehat{P}_{W,(s_0,p)}, \tag{2.13}$$

where $\widehat{P}_{T,(s_0,p)}$ and $\widehat{P}_{W,(s_0,p)}$ are defined in (2.12), and $\mathcal{P}$ is a candidate subset of p.

In this paper, we require that $\#\{\mathcal{P}\}$ is finite. Our theoretical results in Section 3 require this condition. In practice, if the alternative structure is known, we can choose $\mathcal{P}$ accordingly. For example, we can choose $\mathcal{P} = \{1, 2\}$ for dense alternatives, and $\mathcal{P} = \{\infty\}$ for sparse alternatives. However, if the alternative structure is unknown, we can choose $\mathcal{P}$ consisting both small and large $p \in \{1, \ldots, \infty\}$. For example, we recommend to use $\mathcal{P} = \{1, 2, 3, 4, 5, \infty\}$ in real applications, which has been shown by our numerical studies to enjoy high powers as well as relatively low computational cost. Algorithm 1 describes our procedure to construct T_{ad} and W_{ad}.

Algorithm 1 : A bootstrap procedure to obtain T_{ad} and W_{ad}

Input: Given the data $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, set the values for s_0, τ_0, the bootstrap replication number B, and the candidate subset $\mathcal{P}$.

Step 1: Calculate $T_{(s_0,p)}$ and $W_{(s_0,p)}$ for $p \in \mathcal{P}$ as defined in (2.7).

Step 2: Repeat the procedure (2.8) – (2.11) for B times, and obtain the bootstrap samples $\{T^1_{(s_0,p)}, \ldots, T^B_{(s_0,p)}\}$, and $\{W^1_{(s_0,p)}, \ldots, W^B_{(s_0,p)}\}$ for $p \in \mathcal{P}$.

Step 3: Based on the bootstrap samples in Step 2, calculate empirical P-values $\widehat{P}_{T,(s_0,p)}$ and $\widehat{P}_{W,(s_0,p)}$ for $p \in \mathcal{P}$ as defined in (2.12).

Step 4: Using $\widehat{P}_{T,(s_0,p)}$ and $\widehat{P}_{W,(s_0,p)}$ with $p \in \mathcal{P}$, calculate the two data-adaptive test statistics T_{ad} and W_{ad} as defined in (2.13).

Output: Algorithm 1 provides the multiplier bootstrap samples $\{T^1_{(s_0,p)}, \ldots, T^B_{(s_0,p)}\}$ and $\{W^1_{(s_0,p)}, \ldots, W^B_{(s_0,p)}\}$ with $p \in \mathcal{P}$, and the data-adaptive test statistics T_{ad} and W_{ad}.

Using Algorithm 1, we construct two types of data-adaptive test statistics T_{ad} and W_{ad}. Let $F_{T,\mathrm{ad}}(x)$ and $F_{W,\mathrm{ad}}(x)$ be their distribution functions, respectively. Neither $F_{T,\mathrm{ad}}(x)$ nor $F_{W,\mathrm{ad}}(x)$ is known. Consequently, we can not use T_{ad} or W_{ad} directly for the hypothesis (1.1). To approximate their corresponding P-values, we adopt the low-cost bootstrap method proposed by Zhou et al. (2018). The main idea of the low-cost bootstrap procedure is to utilize the bootstrap samples $\{T^1_{(s_0,p)}, \ldots, T^B_{(s_0,p)}\}$ and $\{W^1_{(s_0,p)}, \ldots, W^B_{(s_0,p)}\}$ efficiently. Specifically, for $b = 1, \ldots, B$, we set the b-th low-cost bootstrap sample for the theoretical P-values of $T_{(s_0,p)}$ and $W_{(s_0,p)}$ as:

$$\widehat{P}^b_{T,(s_0,p)} = \frac{1}{B}\sum_{b'\neq b} \mathbf{1}\{T^{b'}_{(s_0,p)} > T^b_{(s_0,p)}\}, \quad \text{and} \quad \widehat{P}^b_{W,(s_0,p)} = \frac{1}{B}\sum_{b'\neq b} \mathbf{1}\{W^{b'}_{(s_0,p)} > W^b_{(s_0,p)}\}. \tag{2.14}$$

Then, we define the b-th bootstrap sample for the data-adaptive test statistics T_{ad} and W_{ad} as:

$$T^b_{\text{ad}'} = \min_{p\in\mathcal{P}} \widehat{P}^b_{T,(s_0,p)}, \text{ and } W^b_{\text{ad}'} = \min_{p\in\mathcal{P}} \widehat{P}^b_{W,(s_0,p)}. \tag{2.15}$$

Let $P_{T,\text{ad}}$ and $P_{W,\text{ad}}$ be the theoretical P-values for T_{ad} and W_{ad}, respectively. Based on the bootstrap samples $\{T^1_{\text{ad}'}, \ldots, T^B_{\text{ad}'}\}$ and $\{W^1_{\text{ad}'}, \ldots, W^B_{\text{ad}'}\}$, we approximate $P_{T,\text{ad}}$ and $P_{W,\text{ad}}$ by

$$\widehat{P}_{T,\text{ad}} = \frac{1}{B+1}\sum_{b=1}^{B} \mathbf{1}\left\{T^b_{\text{ad}'} \le T_{\text{ad}}\right\}, \quad \text{and } \widehat{P}_{W,\text{ad}} = \frac{1}{B+1}\sum_{b=1}^{B} \mathbf{1}\left\{W^b_{\text{ad}'} \le W_{\text{ad}}\right\}. \tag{2.16}$$

Algorithm 2 describes our method to obtain $\widehat{P}_{T,\text{ad}}$ and $\widehat{P}_{W,\text{ad}}$. Using Algorithm 2, we obtain the esti-

Algorithm 2 : A low-cost bootstrap procedure for approximating P-values for T_{ad} and W_{ad}

Input: Use the bootstrap samples $\{T^1_{(s_0,p)}, \ldots, T^B_{(s_0,p)}\}$ and $\{W^1_{(s_0,p)}, \ldots, W^B_{(s_0,p)}\}$ with $p \in \mathcal{P}$ from Algorithm 1.

Step 1: For $p \in \mathcal{P}$, get the low-cost bootstrap samples $\widehat{P}^b_{T,(s_0,p)}$ and $\widehat{P}^b_{W,(s_0,p)}$ with $1 \le b \le B$ as defined in (2.14).

Step 2: Calculate the b-th bootstrap sample $T^b_{\text{ad}'}$ and $W^b_{\text{ad}'}$ with $1 \le b \le B$ for the data-adaptive test statistics as defined in (2.15)

Step 3: Based on $\{T^1_{\text{ad}'}, \ldots, T^B_{\text{ad}'}\}$, and $\{W^1_{\text{ad}'}, \ldots, W^B_{\text{ad}'}\}$ from Step 2, calculate the empirical P-values $\widehat{P}_{T,\text{ad}}$ and $\widehat{P}_{W,\text{ad}}$ for the data-adaptive tests as defined in (2.16).

Output: Algorithm 2 provides low-cost bootstrap samples $\{T^1_{\text{ad}'}, \ldots, T^B_{\text{ad}'}\}$ and $\{W^1_{\text{ad}'}, \ldots, W^B_{\text{ad}'}\}$ for the data-adaptive statistics T_{ad} and W_{ad}, and their empirical P-values $\widehat{P}_{T,\text{ad}}$ and $\widehat{P}_{W,\text{ad}}$.

mated P-values for the two types of data-adaptive test statistics. Therefore, given the significance level α, we define the two data-adaptive tests as:

$$\Psi^T_{\alpha,\text{ad}} = \mathbf{1}\{\widehat{P}_{T,\text{ad}} \le \alpha\}, \quad \text{and } \Psi^W_{\alpha,\text{ad}} = \mathbf{1}\{\widehat{P}_{W,\text{ad}} \le \alpha\}.$$

For the data-adaptive test T_{ad}, we reject $\mathbf{H}_0$ if and only if $\Psi^T_{\alpha,\text{ad}} = 1$. Similarly, for W_{ad}, we reject $\mathbf{H}_0$ if and only if $\Psi^W_{\alpha,\text{ad}} = 1$. More detailed illustrations of Algorithms 1 and 2 are provided in the Supplementary Material.

3 Theoretical properties

In this section, we discuss theoretical properties of our two types of the (s_0, p)-norm-based individual tests as well as the data-adaptive tests. In Section 3.1, we introduce some assumptions. Based on that, in Section 3.2, we discuss the size and power properties of the (s_0, p)-norm-based tests. In Section 3.3, we present some theoretical properties of the data-adaptive methods.

Note that both this paper and Zhou et al. (2018) adopt the bootstrap procedure to approximate the corresponding test statistics' limiting distributions. The latter extended the bootstrap procedure of mean

vectors (Chernozhukov et al. (2017)) to U-statistic-based vectors. However, the validity of our bootstrap procedure for change point analysis does not follow directly from Chernozhukov et al. (2017) and Zhou et al. (2018). Specifically, considering the series dependence in the CUSUM matrix $\mathcal{C}$, our two types of individual test statistics designed for the change point detection require substantial modifications for Gaussian approximations developed in Chernozhukov et al. (2017); Zhou et al. (2018). Furthermore, as shown in what follows, the analysis of alternatives also requires careful handling of nuisance parameters derived from model misspecifications, which is also fundamentally different from Zhou et al. (2018).

3.1 Basic assumptions

We introduce some notations and several basic assumptions needed for our theorems. For $\boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^d$, we define

$$\begin{aligned} \boldsymbol{\Psi}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) &:= \big(\Psi_1(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m), \ldots, \Psi_q(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)\big)^\top, \\ \boldsymbol{h}(\boldsymbol{x}) &:= \big(h_{1,1}(\boldsymbol{x}), \ldots, h_{1,q}(\boldsymbol{x})\big)^\top, \end{aligned}$$

where $\Psi_s(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \Phi_s(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) - \theta_s$ is the centralized kernel, and $h_{1,s}(\boldsymbol{x})$ is defined in (2.1). We also set $\mathcal{V}_{s_0} := \{\boldsymbol{v} \in \mathbb{S}^{q-1} : \|\boldsymbol{v}\|_0 \le s_0\}$, where $\mathbb{S}^{q-1} := \{\boldsymbol{v} \in \mathbb{R}^q : \|\boldsymbol{v}\| = 1\}$. With the notations, we then introduce our assumptions as follows:

(A.0) Suppose there exists $\tau_0 \in (0, 0.5)$ such that $\tau_0 \le \widetilde{t}_s \le 1 - \tau_0$ for $s \in \Pi_\gamma$.

(A.1) There exists $0 < \delta < 1/7$ such that $s_0^2 \log(qn) = O(n^\delta)$.

(A.2) For different indices $0 < i_1 < \cdots < i_m < n$, we require

$$\max_{1 \le s \le q} \mathbb{E}\big(\exp(|\Psi_s(\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_m})|/K)\big) \le 2, \text{ for some constant } K > 0.$$

(A.3) There is a constant $b > 0$, such that $\mathbb{E}|\boldsymbol{v}^\top \boldsymbol{h}(\boldsymbol{X})|^2 \ge b$, for all $\boldsymbol{v} \in \mathcal{V}_{s_0}$.

(A.4) For $\ell = 1, 2$, we require $\max_{1 \le s \le q} \mathbb{E}|h_{1,s}(\boldsymbol{X})|^{2+\ell} \le K^\ell$.

Assumption $(\mathbf{A}.\mathbf{0})$ requires that the relative change point location $\widetilde{t}_s$ is strictly bounded away from the beginning or end of data observations, which is a common assumption in the literature (Jirak, 2015; Dette and Gösmann, 2018). It is also a minimum sample size requirement to justify the asymptotic properties of our new tests. Assumption $(\mathbf{A}.\mathbf{1})$ is a technical condition, which describes the scaling relationships among s_0, q, and n. Assumption $(\mathbf{A}.\mathbf{1})$ allows that s_0 and q can go to infinity as $n \to \infty$ as long as $s_0^2 \log(qn) = O(n^\delta)$ for some $0 < \delta < 1/7$. Assumptions $(\mathbf{A}.\mathbf{2})$ – $(\mathbf{A}.\mathbf{4})$ are crucial for proving the results of Gaussian approximations for our two types of (s_0, p)-norm-based individual test statistics. In particular, Assumptions $(\mathbf{A}.\mathbf{2})$ and $(\mathbf{A}.\mathbf{4})$ are moment conditions for the centered kernel $\Psi_s(\cdot)$. Assumption $(\mathbf{A}.\mathbf{2})$ requires that $\Psi_s(\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_m})$ follows sub-exponential distributions. Many bounded kernels such as Wilcoxon or Kendall's tau can satisfy this condition. Assumption $(\mathbf{A}.\mathbf{4})$ requires that $h_{1,s}(\boldsymbol{X})$ has bounded third and forth moments. Assumption $(\mathbf{A}.\mathbf{3})$ requires that the U-statistics are non-degenerate. Moreover, it requires that $\boldsymbol{v}^\top \boldsymbol{h}(\boldsymbol{X})$ is also non-degenerate for any $\boldsymbol{v} \in \mathcal{V}_{s_0}$.

3.2 Theoretical results of the two types of individual test statistics

In this section, we derive the theoretical properties for the two types of the (s_0, p)-norm-based individual test statistics, in terms of their size and power.

3.2.1 Size performance

We first consider their size properties. The following Theorem 3.1 justifies the uniform validity of the bootstrap procedure for $T_{(s_0,p)}$ and $W_{(s_0,p)}$ in Algorithm 1. It is also crucial for the size control.

Theorem 3.1. For $T_{(s_0,p)}$, suppose Assumptions $(\mathbf{A}.\mathbf{0})$ – $(\mathbf{A}.\mathbf{4})$ hold. Under $\mathbf{H}_0$, we have

$$\sup_{z\in(0,\infty)} \left|\mathbb{P}(T_{(s_0,p)} \le z) - \mathbb{P}(T^b_{(s_0,p)} \le z|\mathcal{X})\right| = o_p(1), \text{ as } n, q \to \infty.$$

Similarly, for $W_{(s_0,p)}$, suppose Assumptions $(\mathbf{A}.\mathbf{0})$ – $(\mathbf{A}.\mathbf{4})$ hold. Then under $\mathbf{H}_0$, we have

$$\sup_{z\in(0,\infty)} \left|\mathbb{P}(W_{(s_0,p)} \le z) - \mathbb{P}(W^b_{(s_0,p)} \le z|\mathcal{X})\right| = o_p(1), \text{ as } n, q \to \infty.$$

Under Theorem 3.1, the following Corollary 3.1 shows that our two types of the (s_0, p)-norm-based individual tests $\Psi^T_{\alpha,(s_0,p)}$ and $\Psi^W_{\alpha,(s_0,p)}$ defined in (2.7) have the asymptotic level of α.

Corollary 3.1. For $T_{(s_0,p)}$, suppose Assumptions $(\mathbf{A}.\mathbf{0})$ – $(\mathbf{A}.\mathbf{4})$ hold. Under $\mathbf{H}_0$, we have

$$\mathbb{P}(\Psi^T_{\alpha,(s_0,p)} = 1) \to \alpha, \text{ and } \widehat{P}_{T,(s_0,p)} - P_{T,(s_0,p)} \xrightarrow{\mathbb{P}} 0, \text{ as } n, q, B \to \infty.$$

Similarly, for $W_{(s_0,p)}$, suppose Assumptions $(\mathbf{A}.\mathbf{0})$ – $(\mathbf{A}.\mathbf{4})$ hold. Then under $\mathbf{H}_0$, we have

$$\mathbb{P}(\Psi^W_{\alpha,(s_0,p)} = 1) \to \alpha, \text{ and } \widehat{P}_{W,(s_0,p)} - P_{W,(s_0,p)} \xrightarrow{\mathbb{P}} 0, \text{ as } n, q, B \to \infty,$$

where $P_{T,(s_0,p)} := 1 - F_{T_{(s_0,p)}}(T_{(s_0,p)})$ and $P_{W,(s_0,p)} := 1 - F_{W_{(s_0,p)}}(W_{(s_0,p)})$ are the theoretical P-values for $T_{(s_0,p)}$ and $W_{(s_0,p)}$, respectively, and $\widehat{P}_{T,(s_0,p)}$ and $\widehat{P}_{W,(s_0,p)}$ defined in (2.16) are the corresponding approximations.

3.2.2 Power performance

We next discuss the oracle power properties. For simplicity, we first assume $\sigma_{s,s} = \mathrm{Var}(h_{1,s}(\boldsymbol{X}))$ is known. With known $(\sigma_{s,s})_{s=1}^q$, we then define the oracle individual tests for $\widetilde{T}_{(s_0,p)}$ and $\widetilde{W}_{(s_0,p)}$ as

$$\Psi^{\widetilde{T}}_{\alpha,(s_0,p)} = \mathbf{1}\{\widetilde{T}_{(s_0,p)} \ge \widehat{c}^{\widetilde{T}}_{\alpha,(s_0,p)}\}, \text{ and } \Psi^{\widetilde{W}}_{\alpha,(s_0,p)} = \mathbf{1}\{\widetilde{W}_{(s_0,p)} \ge \widehat{c}^{\widetilde{W}}_{\alpha,(s_0,p)}\},$$

where $\widetilde{T}_{(s_0,p)}$ is defined in (2.5), $\widetilde{W}_{(s_0,p)}$ is defined in (2.6), and $\widehat{c}^{\widetilde{T}}_{\alpha,(s_0,p)}$ and $\widehat{c}^{\widetilde{W}}_{\alpha,(s_0,p)}$ are the estimated critical values for $\widetilde{T}_{(s_0,p)}$ and $\widetilde{W}_{(s_0,p)}$, respectively, by the corresponding multiplier bootstrap procedure. To analyze the power for $\Psi^{\widetilde{T}}_{\alpha,(s_0,p)}$ and $\Psi^{\widetilde{W}}_{\alpha,(s_0,p)}$, we need to introduce some additional notations and assumptions. Note that under $\mathbf{H}_1$, there is a change point of θ_s at the location $\lfloor n\widetilde{t}_s \rfloor$ for $s \in \Pi_\gamma$. We

set $\theta_s^{(1)}$ and $\theta_s^{(2)}$ as the parameters for the coordinate $s \in \Pi_\gamma$ before $\lfloor n\widetilde{t}_s \rfloor$, and after $\lfloor n\widetilde{t}_s \rfloor$, respectively. Then we define the oracle signal to noise ratio vector $\boldsymbol{D} = (D_1, \ldots, D_q)^\top$ (associated with $\widetilde{t}_s$) with

$$D_s := \begin{cases} 0, & \text{for } s \in (\Pi_\gamma)^c \\ \big|\widetilde{t}_s(1-\widetilde{t}_s)(\theta_s^{(2)} - \theta_s^{(1)})\sigma_{s,s}^{-1/2}\big|, & \text{for } s \in \Pi_\gamma. \end{cases} \tag{3.1}$$

Under $\mathbf{H}_1$, for each coordinate $s \in \Pi_\gamma$, its change point location $\lfloor n\widetilde{t}_s \rfloor$ divides the data into two parts with different population distributions. More specifically, for each coordinate s, given the specified kernel $\Phi_s(\cdot)$, let $\boldsymbol{X}^{\Phi_s}$ be the corresponding sub-vector of $(X_1, \ldots, X_d)^\top$ according to the specified kernel. Let $\boldsymbol{X}_i^{\Phi_s}$ be the i-th observation of the sub-vector $\boldsymbol{X}^{\Phi_s}$ for $1 \le i \le n$. We assume that for $s \in \Pi_\gamma$, $(\boldsymbol{X}_i^{\Phi_s})_{1 \le i \le \lfloor n\widetilde{t}_s \rfloor} \overset{\text{i.i.d}}{\sim} F_{\widetilde{t}_s}(\boldsymbol{x})$, and $(\boldsymbol{X}_i^{\Phi_s})_{\lfloor n\widetilde{t}_s \rfloor + 1 \le i \le n} \overset{\text{i.i.d}}{\sim} G_{\widetilde{t}_s}(\boldsymbol{x})$, where $F_{\widetilde{t}_s}(\boldsymbol{x})$ and $G_{\widetilde{t}_s}(\boldsymbol{x})$ denote two different distribution functions. To simplify notations, we write $\boldsymbol{X}_i^{\Phi_s}$ as $\boldsymbol{X}_i$ for a given kernel Φ_s. With basic notations, for $s \in \Pi_\gamma$, we define

$$\beta_s^{(j)} := \mathbb{E}\Phi_s(\boldsymbol{X}_1', \ldots, \boldsymbol{X}_j', \boldsymbol{X}_{j+1}', \ldots, \boldsymbol{X}_m'), \quad \text{for } 1 \le j \le m-1,$$

where $\boldsymbol{X}_\ell' \overset{\text{i.i.d}}{\sim} F_{\widetilde{t}_s}(\boldsymbol{x})$ for $1 \le \ell \le j$, and $\boldsymbol{X}_\ell' \overset{\text{i.i.d}}{\sim} G_{\widetilde{t}_s}(\boldsymbol{x})$ for $j+1 \le \ell \le m$. To simplify notations, we write $\beta_s^{(0)} := \theta_s^{(1)}$, and $\beta_s^{(m)} := \theta_s^{(2)}$.

Note that for the mean change point problem with a kernel $\Phi_s = X_s$, the sub-vector $\boldsymbol{X}^{\Phi_s}$ is X_s for $1 \le s \le d$; for the Kendall's tau correlation problem with $\Phi_{i,j} = \text{sign}(X_i - X_i')\text{sign}(X_j - X_j')$, the sub-vector $\boldsymbol{X}^{\Phi_{i,j}}$ is $(X_i, X_j)^\top$ for $1 \le i < j \le d$.

We regard $(\beta_s^{(j)})_{j=1}^{m-1}$ defined in (3.2.2) as the parameters (cross terms) with mixing distributions. For a given kernel $\Phi_s(\cdot)$ with an order m, there are $m-1$ cross terms under $\mathbf{H}_1$. For example, for $m = 1$, there is no cross term; for $m = 2$, there is only one cross term, etc.

To analyze the power properties of the individual tests, we need the following assumption on the parameters. In particular, we require:

(A.5) Suppose there exits $M_0 > 0$ such that $\max\big(\max_{s \in \Pi_\gamma} \max_{0 \le j \le m} |\beta_s^{(j)}|, \max_{s \in (\Pi_\gamma)^c} |\theta_s|\big) \le M_0$.

After introducing the assumption on the parameters, to derive the power results, we also require the following Assumption $(\mathbf{A}.\mathbf{1})'$ on the scaling relationships among s_0, q, and n:

(A.1)′ We require $\log q = o(n^{1/3})$ and $n = o(\exp(\log^{\delta_1}(q)))$ for some $0 < \delta_1 < 1$. Moreover, we assume $\exists \delta_2 > 0$ such that $s_0 = \log^{\delta_2}(q)$.

Under basic assumptions, Theorem 3.2 presents the power results for $\widetilde{T}_{(s_0,p)}$ and $\widetilde{W}_{(s_0,p)}$.

Theorem 3.2. Suppose that Assumptions $(\mathbf{A}.\mathbf{0})$, $(\mathbf{A}.\mathbf{1})'$, $(\mathbf{A}.\mathbf{2})$ – $(\mathbf{A}.\mathbf{5})$ hold. Let $\epsilon_n = o(1)$ satisfy $\epsilon_n \sqrt{\log(q(n - 2\lfloor n\tau_0 \rfloor))} \to \infty$ as $n, q \to \infty$. If $\mathbf{H}_1$ holds with

$$\sqrt{n}\big\|\boldsymbol{D}\big\|_{(s_0,p)} \ge s_0(1+\epsilon_n)C_0^{-1}\big(A_0\sqrt{2\log(q(n-2\lfloor n\tau_0 \rfloor))} + \sqrt{2\log(\alpha^{-1})}\big), \tag{3.2}$$

then for both $\widetilde{T}_{(s_0,p)}$ and $\widetilde{W}_{(s_0,p)}$, we have

$$\mathbb{P}(\Psi_{\alpha,(s_0,p)}^{\widetilde{T}} = 1) \to 1, \text{ and } \mathbb{P}(\Psi_{\alpha,(s_0,p)}^{\widetilde{W}} = 1) \to 1, \text{ as } n, q, B \to \infty,$$

where A_0 and C_0 are universal positive constants only depending on M_0, b and K.

Theorem 3.2 shows that with probability tending to one, the oracle individual tests $\Psi^{\widetilde{T}}_{\alpha,(s_0,p)}$ and $\Psi^{\widetilde{W}}_{\alpha,(s_0,p)}$ can detect the change points as long as the signal-noise ratio vector $\boldsymbol{D}$ satisfies Condition (3.2). By (3.1) and (3.2), we see that it is more likely to reject $\mathbf{H}_0$ if the change point location $\widetilde{t}_s$ with $s \in \Pi_\gamma$ gets closer to the middle of the data observations. Note that the scaling relationships among s_0, n, and q in Assumption $(\mathbf{A}.\mathbf{1})'$ are weaker than those in Assumption $(\mathbf{A}.\mathbf{1})$, which allows larger q to reject $\mathbf{H}_0$.

As an important remark, we discuss the value of C_0 in (3.2), which characterizes the lower bound of the signal-noise ratio vector $\boldsymbol{D}$ for rejecting $\mathbf{H}_0$. In particular, by (3.2), it is more likely to reject $\mathbf{H}_0$ when $C_0 > 1$ than $C_0 \leq 1$. Essentially, the value of C_0 relies on the relationships among the parameters $\theta_s^{(1)}$, $\theta_s^{(2)}$, and the cross terms $(\beta_s^{(j)})_{j=1}^{m-1}$. To see this, for $s \in \Pi_\gamma$, by letting $\boldsymbol{\beta}_s = \big(\beta_s^{(0)}, \ldots, \beta_s^{(m)}\big)^\top$, we define $\delta_s(t; \widetilde{t}_s, \boldsymbol{\beta}_s)$ as:

$$\delta_s(t; \widetilde{t}_s, \boldsymbol{\beta}_s) = \lim_{n\to\infty} \frac{\lfloor nt \rfloor \lfloor nt \rfloor^*}{n^2} \mathbb{E}\big(\widehat{\theta}_{\lfloor nt \rfloor, s} - \widehat{\theta}_{\lfloor nt \rfloor^*, s}\big), \quad \text{with } t \in [\tau_0, 1-\tau_0]. \tag{3.3}$$

The function $\delta_s(t; \widetilde{t}_s, \boldsymbol{\beta}_s)$ characterizes the expected signal jump at each time point t. On one hand, for $s \in \Pi_\gamma$, if the signal jump $|\theta_s^{(1)} - \theta_s^{(2)}|$ is large enough such that the true change point location $\widetilde{t}_s$ maximizes $|\delta_s(t; \widetilde{t}_s, \boldsymbol{\beta}_s)|$, we have $C_0 \leq 1$ as shown in our proof. On the other hand, if the values of cross terms $(\beta_s^{(j)})_{j=1}^{m-1}$ are much bigger than those of the true parameters $\theta_s^{(1)}$ and $\theta_s^{(2)}$ such that $\widetilde{t}_s$ fails to maximize $|\delta_s(t; \widetilde{t}_s, \boldsymbol{\beta}_s)|$, we have $C_0 > 1$. In this case, by (3.2), it is more likely to reject $\mathbf{H}_0$. From this aspect, the cross terms under $\mathbf{H}_1$ provide useful information for the change point detection, although they may have a negative effect on the identification of the true change point location $\widetilde{t}_s$.

Note that for the kernel $\Phi_s(\cdot)$ with an order $m = 1$, the function $\delta_s(t; \widetilde{t}_s, \boldsymbol{\beta}_s)$ reduces to

$$\delta_s(t; \widetilde{t}_s, \theta_s^{(1)}, \theta_s^{(2)}) = \begin{cases} t(1-\widetilde{t}_s)\big(\theta_s^{(1)} - \theta_s^{(2)}\big), & t \in [\tau_0, \widetilde{t}_s], \\ (1-t)\widetilde{t}_s\big(\theta_s^{(1)} - \theta_s^{(2)}\big), & t \in [\widetilde{t}_s, 1-\tau_0]. \end{cases}$$

In this case, $\widetilde{t}_s = \arg\max_t |\delta_s(t; \widetilde{t}_s, \boldsymbol{\beta}_s)|$ always holds for all $s \in \Pi_\gamma$. Consequently, as a special case, we can set the constant $C_0 = 1$ in (3.2) for the mean change point detection.

3.3 Theoretical properties of the two types of data-adaptive test statistics

In Section 3.2, we present theoretical properties of our two types of the (s_0, p)-norm-based individual test statistics $T_{(s_0,p)}$ and $W_{(s_0,p)}$. In this section, we discuss the size and power properties of the data-adaptive tests $\Psi^T_{\alpha,\mathrm{ad}}$ and $\Psi^W_{\alpha,\mathrm{ad}}$ introduced in (2.6). To present the theorems below, it is necessary to make one additional Assumption $(\mathbf{A}.\mathbf{1})''$. Since Assumption $(\mathbf{A}.\mathbf{1})''$ is more technically involved and does not add much to our understanding of the main results, we include it in the supplementary material.

The following Theorem 3.3 justifies the validity of the low-cost bootstrap procedure in Algorithm 2. It also shows that our data-adaptive tests have the asymptotic level of α.

Theorem 3.3. For T_{ad}, suppose Assumptions $(\mathbf{A}.\mathbf{0})$, $(\mathbf{A}.\mathbf{1})''$, $(\mathbf{A}.\mathbf{2})$ – $(\mathbf{A}.\mathbf{4})$ hold. Under $\mathbf{H}_0$, we have

$$\mathbb{P}(\Psi^T_{\alpha,\mathrm{ad}} = 1) \to \alpha, \quad \text{and } \widehat{P}_{T,\mathrm{ad}} - \widetilde{P}_{T,\mathrm{ad}} \xrightarrow{\mathbb{P}} 0, \text{ as } n, q, B \to \infty.$$

Similarly, for W_{ad}, suppose Assumptions $(\mathbf{A}.\mathbf{0})$, $(\mathbf{A}.\mathbf{1})''$, $(\mathbf{A}.\mathbf{2})$ – $(\mathbf{A}.\mathbf{4})$ hold. Under $\mathbf{H}_0$, we have

$$\mathbb{P}(\Psi^W_{\alpha,\mathrm{ad}} = 1) \to \alpha, \quad \text{and} \ \ \widehat{P}_{W,\mathrm{ad}} - \widetilde{P}_{W,\mathrm{ad}} \xrightarrow{\mathbb{P}} 0, \ \text{as} \ n, q, B \to \infty.$$

After analyzing the size, we now discuss the power. Similar to Theorem 3.2, we first assume that $\sigma_{s,s} = \mathrm{Var}(h_{1,s}(\boldsymbol{X}_1))$ is known. The following Theorem 3.4 shows that under some regular conditions, our two types of data-adaptive tests can reject the null hypothesis with probability tending to one.

Theorem 3.4. Suppose Assumptions $(\mathbf{A}.\mathbf{0})$, $(\mathbf{A}.\mathbf{1})'$, $(\mathbf{A}.\mathbf{2})$ – $(\mathbf{A}.\mathbf{5})$ hold. Let $\epsilon_n = o(1)$ satisfy $\epsilon_n \sqrt{\log(q(n - 2\lfloor n\tau_0 \rfloor))} \to \infty$ as $n, q \to \infty$. If $\mathbf{H}_1$ holds with

$$\sqrt{n}\big\|\boldsymbol{D}\big\|_{(s_0,p)} \geq s_0(1+\epsilon_n)C_0^{-1}\big(A_0\sqrt{2\log(q(n-2\lfloor n\tau_0\rfloor))} + \sqrt{2\log(\#\{\mathcal{P}\}/\alpha)}\big), \tag{3.4}$$

then for both T_{ad} and W_{ad}, we have

$$\mathbb{P}(\Psi^T_{\alpha,\mathrm{ad}} = 1) \to 1, \quad \text{and} \ \ \mathbb{P}(\Psi^W_{\alpha,\mathrm{ad}} = 1) \to 1, \ \text{as} \ n, q, B \to \infty,$$

where A_0 is a positive constant only depending on M_0, b and K, and C_0 is a universal positive constant.

So far we have been focusing on the case with known $\sigma_{s,s}$. In real applications, $\sigma_{s,s}$ is typically unknown. In that case, we can show that the results still hold if the variance estimator satisfies some conditions. As discussed in Shao and Zhang (2010), inappropriate estimation of $\sigma_{s,s}$ can lead to non-monotone power. Therefore, we need to control the estimation error of $\widehat{\sigma}_{s,s}$ under $\mathbf{H}_1$. Suppose the variance estimators $\widehat{\sigma}_{s,s}$ with $1 \leq s \leq q$ satisfy

$$\max_{1\leq s\leq q} |\widehat{\sigma}_{s,s} - \sigma_{s,s}| = o_p\Big(\frac{1}{\sqrt{\log(qn)}}\Big). \tag{3.5}$$

Then, under the same assumptions, Theorems 3.2 and 3.4 still hold by replacing $\sigma_{s,s}$ by $\widehat{\sigma}_{s,s}$.

In practice, to obtain appropriate estimators satisfying (3.5), we can first get the change point location's estimator for each coordinate by

$$\widehat{t}_s = \underset{t\in[\tau_0,1-\tau_0]}{\arg\max} \Big|\sqrt{n}\frac{\lfloor nt\rfloor}{n}\frac{\lfloor nt\rfloor^*}{n}\Big(\widehat{\theta}_{\lfloor nt\rfloor,s} - \widehat{\theta}_{\lfloor nt\rfloor^*,s}\Big)\Big|, \ \text{for} \ 1 \leq s \leq q.$$

Then we put $\widehat{t}_s$ in $\widehat{\sigma}_{\lfloor nt\rfloor,s,s}$ defined in (2.4) and obtain the final variance estimation as

$$\widehat{\sigma}_{s,s} = \frac{1}{n}\Big(\sum_{k=1}^{\lfloor n\widehat{t}_s\rfloor}\big(Q_{\lfloor n\widehat{t}_s\rfloor,s,k} - \widehat{\theta}_{\lfloor n\widehat{t}_s\rfloor,s}\big)^2 + \sum_{k=\lfloor n\widehat{t}_s\rfloor+1}^{n}\big(Q_{\lfloor n\widehat{t}_s\rfloor^*,s,k} - \widehat{\theta}_{\lfloor n\widehat{t}_s\rfloor^*,s}\big)^2\Big), \ \text{for} \ 1 \leq s \leq q. \tag{3.6}$$

As shown by our extensive numerical studies, our new tests with the above estimators in (3.6) can control the size well under $\mathbf{H}_0$ and have "reasonable" power performance under $\mathbf{H}_1$. Furthermore, the following Proposition 3.5 shows that $\{\widehat{\sigma}_{s,s}\}_{1\leq s\leq q}$ are uniformly consistent for the kernel with $m = 1$.

Proposition 3.5. Assume that Assumptions $(\mathbf{A}.\mathbf{0})$, $(\mathbf{A}.\mathbf{1})'$, $(\mathbf{A}.\mathbf{2})$ – $(\mathbf{A}.\mathbf{5})$ hold. Under $\mathbf{H}_1$, suppose

additionally that

$$\limsup_{n,q\to\infty} \frac{\log(n)}{n\delta_{\min}^2} = 0$$

holds, where $\delta_{\min} = \min_{s\in\Pi_\gamma} |\theta_s^{(2)} - \theta_s^{(1)}|$. Then, for the kernel with an order $m = 1$, there exists a sufficiently small universal constant $C_1 > 0$ such that

$$\max_{1\le s\le q} |\widehat{\sigma}_{s,s} - \sigma_{s,s}| = o_p(n^{-C_1}). \tag{3.7}$$

Proposition 3.5 shows that the variance estimators are uniformly consistent as long as the minimum signal jump does not converge too fast to 0 as n and q tend to infinity. It is a mild modification of the results in Section 3 of Jirak (2015). It is worth mentioning that it is challenging to extend the result in (3.7) directly to the general case with an order $m \ge 2$. One difficulty is that, to prove (3.7), we need to justify the estimation consistency of $\widehat{t}_s$ for $\widetilde{t}_s$ with $s \in \Pi_\gamma$. However, this requires $\delta_s(t; \widetilde{t}_s, \boldsymbol{\beta}_s)$ in (3.3) obtains its maximum absolute value at the true change point location $\widetilde{t}_s$, which involves complicated discussions about the relationships among $\theta_s^{(1)}$, $\theta_s^{(2)}$, and the cross terms $\{\beta_s^{(j)}\}_{j=1}^{m-1}$. Since this work is beyond the scope of this paper, we leave it as a future research direction.

3.4 More discussions about the theoretical results

In this section, we compare our proposed methods with several other tests from a theoretical viewpoint. In particular, we investigate the high dimensional efficiency as well as the detection boundary.

3.4.1 High dimensional efficiency

Note that Aston and Kirch (2018) introduced the asymptotic concept of high dimensional efficiency to compare different tests' detection powers. Specifically, recall the signal to noise ratio $\boldsymbol{D} = (D_1, \ldots, D_q)^\top$ as defined in (3.1). Define $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_q)^\top$ with $\Delta_s = \theta_s^{(2)} - \theta_s^{(1)}$. Then, the high dimensional efficiency is defined as a rate at which $\|\boldsymbol{D}\|_2$ or $\|\boldsymbol{\Delta}\|_2$ is allowed to converge to zero such that the asymptotical power is strictly above the nominal level α. Note that $D_s = \widetilde{t}_s(1-\widetilde{t}_s)\sigma_{s,s}^{-1/2}\Delta_s$. Since $\|\boldsymbol{D}\|_2 \asymp \|\boldsymbol{\Delta}\|_2$, according to Aston and Kirch (2018) (Definition 2.1), any tests using either $\boldsymbol{D}$ or $\boldsymbol{\Delta}$ have the same high dimensional efficiency. To compare different methods in a unified form, in this section, we use $\boldsymbol{\Delta}$ to define the high dimensional efficiency.

With the above new concept, Aston and Kirch (2018) mainly investigated the properties of projection-based tests, where they first used a vector $\boldsymbol{p}$ to project $\{\boldsymbol{X}_i\}_{i=1}^n$ into a univariate data sequence $\{\boldsymbol{p}^\top \boldsymbol{X}_i\}_{i=1}^n$, then they used $\{\boldsymbol{p}^\top \boldsymbol{X}_i\}_{i=1}^n$ to construct CUSUM statistics to detect the change point. As shown by Aston and Kirch (2018), choosing $\boldsymbol{p} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}$ obtains the highest high dimensional efficiency, which is defined as the oracle projection. Furthermore, the following Table 1 summarizes the efficiency of different tests for high dimensional mean change point detection.

According to Table 1, the oracle projection with known $\boldsymbol{\Delta}$ and $\boldsymbol{\Sigma}$ has the highest efficiency. If the underlying covariance structure is the identity, the two types of L_2-based tests, designed for the dense alternative, have an efficiency loss of an order $q^{1/4}$ as compared to the oracle projection. Furthermore, as shown in Aston and Kirch (2018) (Corollary 3.8), they perform similarly to the case of a random

Table 1: High dimensional efficiency for change point tests

Methods	Efficiency	Methods	Efficiency
Oracle Projection (Aston and Kirch (2018))	$\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}\|_2$	L_∞- based (Jirak (2015))	$\frac{\min_{s\in\Pi_\gamma}\|\Delta_s\|}{\log^{1/2}(n)}$
L_2-based (Horváth and Hušková (2012))	$\frac{1}{q^{1/4}}\|\boldsymbol{\Delta}\|_2$	Double CUSUM (Cho et al. (2016))	$\frac{\gamma^{\varphi-1}\sum_{s\in\Pi_\gamma}\|\Delta_s\|}{q^\varphi\log(n)}$
L_2-based (Enikeeva and Harchaoui (2019))	$\frac{\|\boldsymbol{\Delta}\|_2}{q^{1/4}(\log(q\log(n)))^{1/4}}$	(s_0,p)-norm-based	$\frac{\|\boldsymbol{\Delta}\|_{(s_0,p)}}{s_0\log^{1/2}(q)}$

projection if the covariance structure is non-identity. Our simulation studies further show that both of them suffer from size distortions as the covariance matrix is far away from identity. Under sparse alternatives with $\gamma \asymp 1$, we note that, compared to the oracle case, Jirak (2015), Cho et al. (2016) with $\varphi = 0$, and our (s_0, p)-norm-based tests with a fixed s_0 have only an efficiency loss of an order $\log(n)$ or $\log(q)$. Furthermore, as shown in what follows, our new tests reach the detection boundary for sparse change point alternatives, suggesting that our new tests are more efficient than other methods in the sparse setting. For instance, suppose we have data with only two components having a change point, i.e., $\gamma = 2$. Considering Table 1, our individual test with $s_0 = 2$ and $p = 2$ has the highest efficiency among all candidates.

3.4.2 Detection boundary under sparse alternatives

We note that the rates in (3.2) and (3.4) are rate-optimal for high dimensional change point detection problems. In particular, Enikeeva and Harchaoui (2019) considered the minimax rate optimality of the high dimensional mean change point detection for independent Gaussian random vectors with the identity covariance matrix. Specifically, let $\gamma = q^{1-\beta}$ with $\beta \in (0, 1)$ denote the sparsity of coordinates with a change point. Suppose there is a common change point location at $\lfloor nt^* \rfloor$ for some $t^* \in (0, 1)$, and the signal strength satisfies $|\theta_s^{(2)} - \theta_s^{(1)}| = a_{n.q}$ for all $s \in \Pi_\gamma$. Then, under the high sparsity setting with $\beta \in (1/2, 1)$, they obtained the minimax separation rate of an order

$$a_{n.q} = C\Big(\frac{\log(q)}{nt^*(1-t^*)}\Big)^{1/2}. \tag{3.8}$$

In other words, there is no α-level test can correctly reject $\mathbf{H}_0$ uniformly over $a_{n.q} = C_*\Big(\frac{\log(q)}{nt^*(1-t^*)}\Big)^{1/2}$ for some very small C_*. Considering (3.2) and (3.4), to reject the null hypothesis, we require

$$\|\boldsymbol{D}\|_{(s_0,p)} \geq Cs_0\Big(\frac{\log(q)}{nt^*(1-t^*)}\Big)^{1/2}. \tag{3.9}$$

Therefore, considering (3.8) and (3.9), with a fixed s_0, both our two types of individual and data-adaptive tests obtain the rate optimality for the sparse alternatives.

Lastly, we note that the L_2-based test in Enikeeva and Harchaoui (2019) almost reaches the detection boundary for dense alternatives ($\beta \in [0, 1/2]$)) with a rate of an order $a_{n,q} = O\big((\log(q\log(n)))^{1/2}/(\sqrt{n}q^{1/4})\big)$. Although our tests do not reach that rate, they are still consistent for dense alternatives and have comparable powers with that of Enikeeva and Harchaoui (2019) in terms of finite sample performance.

4 Simulation studies

In this section, we examine the empirical performance of our proposed methods in terms of size and power, and compare them with several existing state-of-art techniques.

4.1 Model settings

We consider the mean change point problem. In particular, by letting $\theta_s = \mu_s = \mathbb{E}X_s$ for $1 \le s \le d$, we consider the following hypothesis:

$$\mathbf{H}_0 : \mu_{1,s} = \cdots = \mu_{n,s}, \text{ for } 1 \le s \le d, \quad \text{v.s.}$$
$$\mathbf{H}_1 : \exists s \in \{1, \ldots, d\} \text{ and } \widetilde{t_s} \in (0, 1), \text{ s.t. } \mu_{1,s} = \cdots = \mu_{\lfloor n\widetilde{t_s}\rfloor,s} \neq \mu_{\lfloor n\widetilde{t_s}\rfloor+1,s} = \cdots = \mu_{n,s}.$$

To show the adaptivity of our methods, we compare the proposed new tests with several existing methods for the high dimensional mean change point detection, including the classical Hotelling's T^2 test (Hotelling; Srivastava and Worsley (1986)); L_2-type tests: H&H (Horváth and Hušková, 2012) and E&H (Enikeeva and Harchaoui, 2019); the L_∞-type test (Jirak; Jirak (2015)); the double CUSUM methods DC-0, DC-0.5, and DC-COM, where DC-COM is the combination of DC-0 and DC-0.5 (Cho et al. (2016)); the projection-based methods with an oracle projection (Ora-Pro), random projection (Ran-Pro), and a fixed angle projection $2\pi/5$-Pro, where the angle between $2\pi/5$-Pro and Ora-Pro is $2\pi/5$ radian. More details can be found in Aston and Kirch (2018).

We consider the alternative scenario where $(\mu_s)_{s\in\Pi_\gamma}$ have a common change point location t^*, where $t^* \in \{0.3, 0.4, 0.5\}$. We set the sample size $n = 200$. The dimension d is in $\{100, 200, 300\}$, and the bootstrap replication number B is 500. To show the broad applicability of our methods, we generate data from different distributions with various covariance structures as in the following five models:

Model 1: We generate data from multivariate Gaussian distributions with the identity covariance matrix $\mathbf{\Sigma} = \mathbf{I}_d$.

Model 2: We generate data from multivariate Gaussian distributions with blocked diagonal $\mathbf{\Sigma}^*$, where $\mathbf{\Sigma}^\star = (\sigma^\star_{ij}) \in \mathbb{R}^{d\times d}$ with $\sigma^\star_{ii} \overset{\text{i.i.d.}}{\sim} \mathrm{U}(1, 2)$, $\sigma^\star_{ij} = 0.5$ for $5(k-1)+1 \le i \ne j \le 5k$ $(k = 1, \ldots, \lfloor d/5\rfloor)$, and $\sigma^\star_{ij} = 0$ otherwise.

Model 3: We generate data from multivariate Gaussian distributions with banded $\mathbf{\Sigma}'$, where $\mathbf{\Sigma}' = (\sigma'_{ij}) \in \mathbb{R}^{d\times d}$ with $\sigma'_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le d$.

Model 4: We generate data from multivariate Gaussian distributions with non-sparse $\mathbf{\Sigma}$ as in Zhou et al. (2018). The details are provided in the supplementary material.

Model 5 : We generate data from multivariate student $t(5, \mathbf{I}_d)$ distributions.

For all models, we consider a wide range of alternative patterns including both sparse and dense settings to compare the performance of these methods. Specifically, we set $\gamma = 5$, and $\gamma = 100$ for

sparse and dense settings, respectively. We select Π_γ randomly from $\{1, \ldots, d\}$. Let $\delta_s = \mu_s^{(2)} - \mu_s^{(1)}$ be the mean shift for $s \in \Pi_\gamma$ which follows $\mathrm{U}(u_1, u_2)$. Under $\mathbf{H}_0$, we set $u_1 = u_2 = 0$. Under the sparse alternative, we set $u_1 = 0$, and $u_2 = c_1\sqrt{\log(d)/n}$ with a constant $c_1 > 0$. Under the dense alternative, we set $u_1 = 0$, and $u_2 = c_2\sqrt{1/n}$ for some constant $c_2 > 0$. To avoid trivial powers, we set different values of c_1 and c_2 for those five models. Specifically, from Models 1 to 5, we set $c_1 = 5.5, 6.5, 5.5, 6.5, 6.5$ and $c_2 = 3.5, 4.5, 4, 4.5, 5$, respectively.

4.2 Size performance

We first consider the size performance. We set the significance level $\alpha = 0.05$. For both $T_{(s_0,p)}$ and $W_{(s_0,p)}$, we choose $\mathcal{P} = \{1, 2, 3, 4, 5, \infty\}$, and obtain the corresponding P-values by Algorithm 1. For both T_{ad} and W_{ad}, we obtain the P-values via the low-cost bootstrap procedure in Algorithm 2. For Hotelling, we approximate the distribution under $\mathbf{H}_0$ using the approximation method in Srivastava and Worsley (1986). For H&H, following the suggestion by Horváth and Hušková (2012), we set the critical value as 0.894. For E&H, we set the threshold as the $1 - \alpha/2n$ quantile of $\chi^2(d)$ distribution. It is worth mentioning that the above theoretical critical values for either E&H or H&H are obtained based on the assumption that the cross-sectional structures among coordinates are independent. To compare these two methods in a more equal fashion, we adopt the Generalised Dynamic Factor Model (GDFM) based bootstrap algorithm proposed in Cho et al. (2016), to obtain their critical values empirically, which are referred to E&H-Boot and H&H-Boot, respectively. As discussed in Cho et al. (2016), GDFM-based bootstrap utilizes the representation property of the GDFM and is able to handle the cross-correlations as well as within-series correlations of the panel data. For Jirak, we obtain the critical value by the multiplier bootstrap method. For $T_{\mathrm{DC}}^{\varphi}$, we use the algorithm proposed by Cho et al. (2016). Note that the projection-based methods in Aston and Kirch (2018) rely on the information of the signal jump $\boldsymbol{\delta}$. We do not report their size results here.

Table 2 demonstrates the empirical sizes for all these methods under Models 1 – 5. For Hotelling, it is designed for multivariate normal distributions with $d < n$. In those settings, its sizes are under the nominal level. For the L_∞-typed method Jirak, it is a bit over-sized for all settings. Note that the L_2-typed tests H&H and E&H are designed for either independent coordinates or Gaussian distributions. With naive asymptotic critical values, they suffer from serious size distortions for Models 2 – 5, indicating that they are no longer applicable to those settings. Furthermore, their empirical size performance can be improved in most cases by adopting the bootstrap algorithm. The double CUSUM-based tests can control the size well when the dimension is not very large ($d = 100, 200$). Their sizes become conservative with a relatively large dimension ($d = 300$). As for our data-adaptive tests T-AD and W-AD with a fixed $s_0 = d/2$, they can control the size correctly across various model settings and dimensions.

4.3 Power comparisons with several other existing state-of-art methods

We next compare the proposed new data-adaptive methods with other techniques in terms of change point detection. Table 3 shows the power results of those methods for Model 1 under various change point locations, dimensions, and alternative structures. To present the results more directly, for each setting (t^*, γ, d), we mark tests having top three powers as bold and those having bottom three powers

Table 2: Empirical sizes (%) for tests under Models 1 – 5 with $\tau_0 = 0.2$, $\alpha = 0.05$, $n = 200$, and $B = 500$, based on 2000 replications.

Settings	d	T-AD	W-AD	Hotelling	Jirak	H&H	H&H-Boot	E&H	E&H-Boot	DC-0	DC-0.5	DC-COM
model 1	100	4.65	4.45	2.65	8.60	8.65	8.84	1.25	5.62	4.50	5.65	4.55
	200	5.15	5.05	NA	7.55	8.25	10.44	1.50	3.82	6.25	6.45	6.20
	300	4.00	3.65	NA	8.70	7.45	0.8	1.10	2.11	0.00	0.50	0.00
model 2	100	4.70	5.70	3.50	8.20	21.95	18.17	10.80	10.04	4.70	11.75	4.60
	200	4.50	4.35	NA	8.70	21.30	15.36	7.95	5.52	6.45	7.10	6.45
	300	3.85	3.80	NA	8.80	22.65	6.63	8.90	7.13	0.10	2.85	0.10
model 3	100	5.85	5.55	2.65	7.10	28.85	21.29	13.40	6.93	5.30	8.00	5.45
	200	5.05	4.55	NA	7.35	28.94	7.83	14.35	2.81	7.45	3.85	7.55
	300	5.45	5.15	NA	8.65	28.80	8.63	13.85	6.02	0.20	3.40	0.20
model 4	100	4.55	4.80	3.35	7.90	22.75	20.58	9.45	8.63	3.70	8.35	3.70
	200	5.10	5.25	NA	7.15	22.90	20.88	8.55	6.12	6.20	9.40	6.35
	300	4.40	4.05	NA	7.65	23.55	6.43	8.20	6.93	0.00	1.95	0.00
model 5	100	2.85	3.05	44.50	6.15	50.55	3.82	76.50	30.22	5.75	14.15	5.70
	200	3.10	3.15	NA	6.70	70.45	0.9	87.15	11.75	5.70	4.85	5.15
	300	2.90	3.20	NA	7.50	80.00	0.3	92.95	14.46	0.70	5.55	0.85

Table 3: Empirical powers (%) for Model 1 with different change point locations, dimensions, and alternative structures, based on 2000 replications. Bold and italic items correspond to tests with top three and bottom three powers excluding projection methods for each (t^*, γ, d).

Location	Type	Methods	$\gamma = 5$			$\gamma = 100$		
			$d = 100$	$d = 200$	$d = 300$	$d = 100$	$d = 200$	$d = 300$
$t^* = 0.5$	L_2	Hotelling	*34.45*	NA	NA	73.75	NA	NA
	L_2	E&H-Boot	*70.68*	*60.82*	*38.16*	**99.18**	**93.42**	**74.69**
	L_2	H&H-Boot	87.45	*80.04*	54.12	**99.61**	**99.18**	**88.68**
	L_∞	Jirak	**90.49**	**91.69**	**92.74**	*58.79*	*46.65*	*40.45*
	DC	DC-0	75.50	82.05	*41.15*	*17.95*	*17.15*	*0.30*
	DC	DC-0.5	*54.30*	*37.25*	*6.05*	97.65	86.20	39.35
	DC	DC-COM	76.20	82.50	42.20	*18.75*	*17.55*	*0.30*
	Projection	Ora-Pro	99.70	99.80	99.65	100.00	100.00	100.00
	Projection	Ran-Pro	9.75	7.40	5.45	11.75	8.05	7.35
	Projection	$2\pi/5$-Pro	51.6	56.55	58.05	76.55	80.10	77.70
	Adaptive	T-AD	**90.95**	**92.00**	**92.00**	**99.75**	**92.25**	**75.55**
	Adaptive	W-AD	**89.95**	**91.45**	**91.70**	97.85	76.95	57.35

as italic. Note that the projection-based methods require the knowledge of the signal jump $\boldsymbol{\delta}$ and the underlying covariance structure $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Sigma}^{-1}$). We only report their power results as an upper benchmark (Ora-Pro) and a lower benchmark (Ran-Pro) of other methods. According to Table 3, our proposed tests have the following advantages over their competitors:

- Under the sparse alternatives, the L_∞-based test Jirak has the best power performance. The two types of L_2-based methods H&H-Boot and E&H-Boot generally have the lowest powers. For the double CUSUM-based tests, as shown in Cho et al. (2016), we prefer DC-0 for sparse alternatives and DC-0.5 for dense alternatives, respectively. In this case, DC-0 and DC-COM have better performance than DC-0.5. As for our new methods T-AD and W-AD, they have comparable power performance with the best test (Jirak) in sparse settings.

- Under the dense alternatives, the candidate methods behave differently to the sparse settings. Specifically, the L_2- based tests H&H-Boot and E&H-Boot perform better than the L_∞-based test Jirak. DC-0.5 has higher powers than DC-0 and DC-COM, which is consistent with the theoretical analysis in Cho et al. (2016). Our proposed methods, especially for T-AD, still enjoy comparable power performance with the L_2-based tests in most cases.

- We note that the three projection-based methods depend on the knowledge of the signal jump $\boldsymbol{\delta}$ and the underlying covariance matrix $\boldsymbol{\Sigma}$ or its inverse $\boldsymbol{\Sigma}^{-1}$. In particular, Ora-Pro has the best power performance across all settings by knowing $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ in advance. As shown in Aston and Kirch (2018), Ora-Pro has the highest high dimensional efficiency (see Definition 2.1 therein) theoretically for the mean change point detection, which serves as an upper benchmark of other methods. In practice, however, it is of great difficulty to estimate $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ simultaneously and construct efficient tests based on them. Furthermore, as a lower benchmark, Ran-Pro has the worst performance by using a random projection.

Power results for Models 2-5 are reported in the Supplementary Material, and are similar to the Model 1 results, indicating good performance across a range of data distributions and various covariance structures. In comparisons with the Wang and Samworth method (Wang and Samworth (2018)) for change point identification of high dimensional mean vectors, our algorithm is comparable under Gaussian distributions and better performing for the t_5. This suggests using our algorithm after first detecting the existence of a change point with a data-adaptive test (see Supplementary A.5).

Note that the proposed data-adaptive testing procedure involves the selection of s_0 and $\mathcal{P}$. As shown by our sensitivity analysis (Sections A.2 and A.3 of the Supplementary Material), the proposed new method is robust against the choice of s_0, given s_0 is not too small. In practice, we recommend the use of $s_0 = d/2$. As for the choice of $\mathcal{P}$, it is shown that $\mathcal{P} = \{1, 2, 3, 4, 5, \infty\}$ is relatively fast to execute and has good power performance under various alternative structures, and is recommended to use.

So far, we have focused the attention on change point detection with a common change point location. In practice, the change points for different coordinates can be different. In such cases, it is demonstrated in Supplementary A.4 that W_{ad} has better power performance than T_{ad}. This suggests that it is better to aggregate rows of the CUSUM matrix first, in the case of multiple change points.

Lastly, in addition to the high dimensional mean change point problem, we also consider the covariance matrix change point detection by letting $\theta_{i,j} = \sigma_{i,j} = \mathbb{E}X_iX_j$ for $1 \le i < j \le d$. In this case, two methods are considered: Pearson's sample covariance matrix with the kernel $\Phi_{i,j}(\boldsymbol{X}, \boldsymbol{X'}) = (X_i - X_i')(X_j - X_j')/2$, and the Kendall's tau correlation matrix with the kernel $\Phi_{i,j}(\boldsymbol{X}, \boldsymbol{X'}) = \mathrm{sign}(X_i - X_i')\mathrm{sign}(X_j - X_j')$. It is shown that our proposed methods still enjoy adaptive performance

for large scale covariance matrix change point detection. Furthermore, for data with light-tailed distributions such as Gaussian distributions, the Pearson covariance-based tests have higher powers than those of Kendall's tau-based tests. In contrast, the Kendall's tau-based tests are more robust than those of the Pearson's when the data are heavy-tailed. See Supplementary A.6 for more details.

5 Real applications

In this section, we apply our proposed new method to a genomic dataset of bladder tumor profiles. The motivation of this real example comes from comparative genomic hybridization (CGH) in genomic studies, where CGH is used to obtain the number of copies of genes in a DNA profile. Elevated or lowered copy-number in the tumor generally corresponds to the associated disease (Nicolas et al. (2006)).

We obtain the dataset from the R package "ecp". This dataset contains observations on 43 profiles ($d = 43$) from a DNA with a length of 2215 ($n = 2215$). According to Matteson and James (2014), this dataset records the relative hybridization intensity with respect to a normal genome reference and is normalized on a logarithmic scale. The goal is to find which regions of DNA contain abnormal DNA copy-numbers for the bladder tumor. As shown in Figure 3, abnormal DNA copy-numbers can be profile-specific or be shared across many profiles. In other words, the alternative patterns can be either sparse or dense in this data example. This suggests that it is appropriate to use the data-adaptive method to identify the change points. Furthermore, for each profile, abnormal DNA copy-numbers can happen for more than one time. To locate the segments precisely, we combine our data-adaptive testing procedure with the binary segmentation method in Vostrikova (1981). More specifically, for each search interval (s, e), we use our data-adaptive testing procedure to detect the existence of a change point. If $\mathbf{H}_0$ is rejected, we identify the new change point b from the individual test which has the minimum P-value. In particular,

$$b = \underset{s+h\le k\le e-h}{\arg\max} \left\|\boldsymbol{C}(k)\right\|_{(s_0,p^*)}, \text{ with } p^* = \underset{p\in\mathcal{P}}{\arg\min}\, \widehat{P}_{T,(s_0,p)}.$$

Then the interval (s, e) is split into two subintervals (s, b) and (b, e) and we conduct the above procedure on (s, b) and (b, e) separately. This algorithm is stopped until no subinterval can detect a change point. In this data exploration, we set the significance level $\alpha = 0.05$. We choose the parameters as $s_0 = 43$, $\mathcal{P} = \{1, 2, 3, 4, 5, \infty\}$, and $h = 40$. Each subinterval-based test is based on 500 bootstrap replications. Using our new algorithm, 43 change points are identified. The segmentation results are reported in Figure 3, which are similar to the previous studies in Matteson and James (2014). This indicates that our method works well in this real example.

6 Discussion

This paper provides a unified data-adaptive framework for change point detection in high dimensions, where the dimension d and number of parameters q can be much larger than the sample size n. To that end, we first construct U-statistic-based CUSUM matrix $\mathcal{C}$. To aggregate $\mathcal{C}$ efficiently, we take both the change point location and the alternative pattern into consideration, and propose two types of

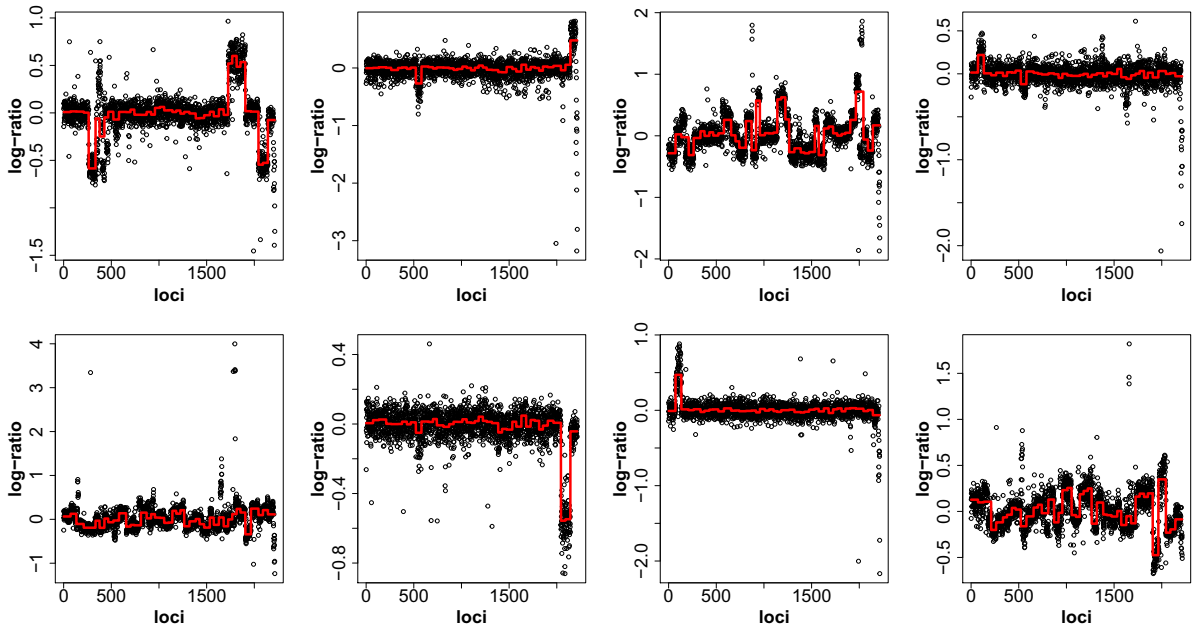

Figure 3: CGH data segmentation for the first 8 profiles. Red lines correspond to medians for the identified segments of each profile.

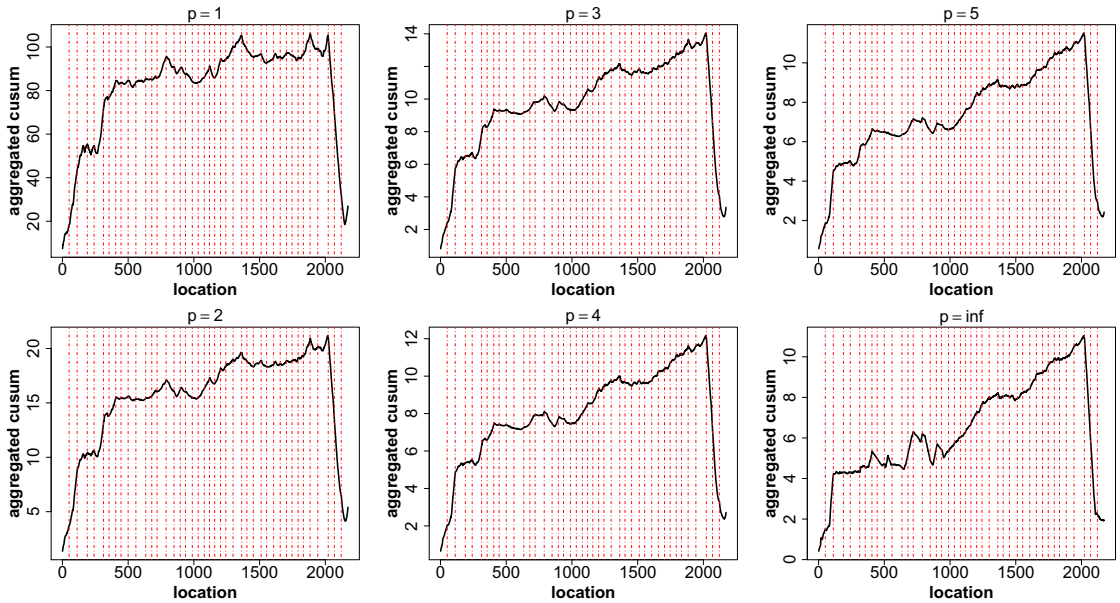

Figure 4: Plots for $\|\boldsymbol{C}(k)\|_{(s_0,p)}$ for the interval $(1, 2215)$ with $s_0 = 43$ and $p \in \{1, 2, 3, 4, 5, \infty\}$. The x-axis denotes the location of DNA. The y-axis is the value of $\|\boldsymbol{C}(k)\|_{(s_0,p)}$. The red dotted lines correspond to the identified 43 change points.

(s_0, p)-norm-based individual test statistics $T_{(s_0,p)}$ and $W_{(s_0,p)}$ with $1 \leq p \leq \infty$. Under this framework, many existing statistics are special cases of $T_{(s_0,p)}$ or $W_{(s_0,p)}$ by choosing a proper s_0, p, and the kernel $\boldsymbol{\Phi}(\cdot)$. We introduce s_0 to boost the power for the individual tests with smaller p (e.g. $p = 1, 2$). We also use p to capture the alternative pattern. Thus, there is at least one test in $T_{(s_0,p)}$ (or $W_{(s_0,p)}$) with

$1 \leq p \leq \infty$ to be powerful for any given alternative structure. In real applications, the alternative structure is unknown. To construct a data-adaptive method, we combine the corresponding individual tests in $T_{(s_0,p)}$ (or $W_{(s_0,p)}$) with $1 \leq p \leq \infty$. It is shown that our two types of data-adaptive tests are simultaneously powerful under various alternatives. For approximating the individual and data-adaptive test statistics' limiting distributions, we adopt the multiplier and the low-cost bootstrap methods in Algorithms 1 and 2, respectively. With mild moment conditions, we justify the validity of our methods in terms of size and power. An R package called AdaptiveCpt is developed to implement our new tests. Extensive simulation studies show that the proposed data-adaptive techniques outperform the existing methods under various model settings and alternative structures.

There are several future directions to be investigated along our proposed work. Our theoretical results are based on the independent assumption for the n observations. In real applications, the observations can be dependent, e.g. financial data. In this case, we can possibly generalize our methods to the high dimensional time series with some dependency structure. Another possible extension is to consider more complex data models such as data with missing entries. Furthermore, if there are more than one change points in each coordinate, we can apply our methods recursively by binary or circular binary methods (Vostrikova (1981); Olshen et al. (2004)). More explorations can be pursued in the future.

Acknowledgement

The authors thank the Joint Editor Professor Simon Wood, the Associate Editor, and two referees, whose helpful comments and suggestions led to a much improved presentation. This research is supported in part by a fellowship from China Scholarship Council (B. Liu), the National Natural Science Foundation of China 11571080 (Zhang), and US National Science Foundation grants IIS1632951, DMS-1821231 and National Institute of Health grant R01GM126550 (Y. Liu).

References

ASTON, J. A. D. and KIRCH, C. (2018). High dimensional efficiency with applications to change point tests. *Electronic Journal of Statistics* **12** 1901–1947.

AUE, A., HÖRMANN, S., HORVÁTH, L. and REIMHERR, M. (2009). Break detection in the covariance structure of multivariate time series. *The Annals of Statistics* **37** 4046–4087.

AVANESOV, V. and BUZUN, N. (2018). Change-point detection in high-dimensional covariance structure. *Electronic Journal of Statistics* **12** 3254–3294.

BERKES, I., GOMBAY, E. and HORVÁTH, L. (2009). Testing for changes in the covariance structure of linear processes. *Journal of Statistical Planning and Inference* **139** 2044–2063.

BÜCHER, A. and KOJADINOVIC, I. (2016). Dependent multiplier bootstraps for non-degenerate U-statistics under mixing conditions with applications. *Journal of Statistical Planning and Inference* **170** 83–105.

CHEN, H., ZHANG, N. ET AL. (2015). Graph-based change-point detection. *The Annals of Statistics* **43** 139–176.

CHEN, J. and GUPTA, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science and Business Media.

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

CHERNOZHUKOV, V., CHETVERIKOV, D., KATO, K. ET AL. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability* **45** 2309–2352.

CHO, H. and FRYZLEWICZ, P. (2015). Multiple change point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 475–507.

CHO, H. ET AL. (2016). Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics* **10** 2000–2038.

COX, D. R. and HINKLEY, D. V. (1979). *Theoretical statistics*. CRC Press.

CSÖRGŐ, M. and HORVÁTH, L. (1988). Invariance principles for changepoint problems. *Journal of Multivariate Analysis* **27** 151–168.

CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit theorems in change-point analysis*. John Wiley and Sons Inc.

DETTE, H. and GÖSMANN, J. (2018). Relevant change points in high dimensional time series. *Electronic Journal of Statistics* **12** 2578–2636.

ENIKEEVA, F. and HARCHAOUI, Z. (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics* **47** 2051–2079.

GOMBAY, E. and HORVÁTH, L. (1999). Change-points and bootstrap. *Environmetrics* **10** 725–736.

GOMBAY, E., HORVÁTH, L. and HUŠKOVÁ, M. (1996). Estimators and tests for change in variances. *Statistics and Risk Modeling* **14** 145–160.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* **19** 293–325.

HORVÁTH, L. and HUŠKOVÁ, M. (2012). Change-point detection in panel data. *Journal of Time Series Analysis* **33** 631–648.

HORVÁTH, L., KOKOSZKA, P. and STEINEBACH, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis* **68** 96–119.

HUŠKOVÁ, M. and MEINTANIS, S. G. (2006). Change point analysis based on empirical characteristic functions. *Metrika* **63** 145–168.

HUŠKOVÁ, M. and PRÁŠKOVÁ, Z. (2014). Comments on: Extensions of some classical methods in change point analysis. *Test* **23** 265–269.

INCLAN, C. and TIAO, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association* **89** 913–923.

JANSSEN, P. (1994). Weighted bootstrapping of U-statistics. *Journal of Statistical Planning and Inference* **38** 31–41.

JIRAK, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics* **43** 2451–2483.

LUNG-YUT-FONG, A., LÉVY-LEDUC, C. and CAPPÉ, O. (2011). Homogeneity and change-point detection tests for multivariate data using rank statistics. *Preprint arXiv:1107.1971* .

MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association* **109** 334–345.

NICOLAS, S., CELINE, V., FABIEN, R. ET AL. (2006). Regional copy number-independent deregulation of transcription in cancer. *Nature Genetics* **38** 1386–1396.

OLSHEN, A. B., VENKATRAMAN, E., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.

PAGE, E. (1955). Control charts with warning lines. *Biometrika* **42** 243–257.

PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–115.

QUESSY, J.-F., SAÏD, M. and FAVRE, A.-C. (2013). Multivariate Kendall's tau for change-point detection in copulas. *Canadian Journal of Statistics* **41** 65–82.

SHAO, X. and ZHANG, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association* **105** 1228–1240.

SRIVASTAVA, M. and WORSLEY, K. J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. *Journal of the American Statistical Association* **81** 199–204.

TAN, C. C., SHI, X. P., SUN, X. Y. and WU, Y. H. (2016). On nonparametric change point estimator based on empirical characteristic functions. *Science China Mathematics* **59** 2463–2484.

VOSTRIKOVA, L. Y. (1981). Detecting disorder in multidimensional random process. *Soviet Math. Dokl* **24** 55–59.

WANG, D., YU, Y. and RINALDO, A. (2017). Optimal covariance change point detection in high dimension. *Preprint arXiv:1712.09912* .

WANG, Q. and JING, B.-Y. (2004). Weighted bootstrap for U-statistics. *Journal of multivariate analysis* **91** 177–198.

WANG, T. and SAMWORTH, R. J. (2018). High-dimensional changepoint estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 57–83.

YU, M. and CHEN, X. (2017). Finite sample change point inference and identification for high-dimensional mean vectors. *Preprint arXiv:1711.08747* .

ZHANG, N. R., SIEGMUND, D., JI, H. P. and LI, J. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645.

ZHONG, P.-S. and LI, J. (2016). Test for temporal homogeneity of means in high-dimensional longitudinal data. *Preprint arXiv:1608.07482* .

ZHOU, C., ZHOU, W.-X., ZHANG, X.-S. and LIU, H. (2018). A unified framework for testing high dimensional parameters: a data-adaptive approach. *Preprint arXiv:1808.02648* .