Nonparametric methods for change point analysis in high dimensional data

Lupeng Zhang

Nonparametric methods for change point analysis in high dimensional data

Lupeng Zhang

Supervised by Dr. Reza Drikvandi

A thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
Durham University
United Kingdom
March 2026

Abstract

Change point detection has been widely applied across different fields such as finance, engineering, genomics, and other fields. The main objective is to detect significant changes in the distribution of a data sequence. Two types of problems are studied in this thesis: the offline change point problem, where detection is performed after all data have been collected, and the online change point problem, where tests are conducted sequentially as data arrive and timely detection is crucial. Both problems have been well studied in low dimensional contexts. However, classical methods often struggle in high dimensional data where the number of variables is much larger than the number of observations. We develop nonparametric methods for both offline and online change point detection and address key challenges in high dimensional change points.

For the offline change point problem, we introduce distance-based CUSUM statistics for detecting change points in high dimensional observations. Unlike the standard CUSUM statistic which primarily detects linear changes such as shifts in the mean of observations, the distance-based CUSUM statistics are constructed based on pairwise dissimilarity distances between observations. Therefore, they are capable of detecting more general types of change points, including linear and non-linear changes in a data stream, such as changes in the mean, variance, correlation, or other changes in the shape of distribution over time. Moreover, the distance-based

CUSUM method is particularly useful for HDLSS data in which the number of observations is very small but the dimension is very large. Detecting change points in such high dimensional data is an understudied problem. We study the properties of our proposed distance-based CUSUM statistics and employ them to develop a nonparametric test to determine the statistical significance of change point estimates. Our approach does not require normality or any other distribution for the data. We provide theoretical guarantees for our method and demonstrate its empirical performance in comparison with some of the recent methods via extensive simulation studies and two real data applications.

In the online change point problem, nonparametric approaches for high dimensional data streams are currently understudied, and their theoretical analysis is often challenging. We construct a sequential testing algorithm using the proposed distance-based CUSUM statistics with a sliding window approach. Furthermore, we propose a stopping rule based on data-driven thresholds to terminate the testing algorithm as quickly as possible after a change point arrives. We prove that the false alarm rate and average run length are controlled at the nominal level under the null hypothesis of no change point. We provide theoretical guarantees for the consistency of the proposed test and derive the convergence of expected detection delay under the alternative hypothesis that there is a change point in the data stream. We validate the theoretical results and assess the empirical performance of our method through extensive simulation studies. We also compare it with some of the recent methods and demonstrate it on a real data application.

Declaration

The work in this thesis is based on research carried out at the Department of Mathematical Sciences, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text. Some of the work presented in this thesis has been published, accepted in journals, or is available as preprints in public repositories – the relevant publications are listed below

- **Zhang, L.**, Drikvandi, R., and Chen, Y. (2025). Nonparametric online change point detection in high dimensional data streams. Submitted.
- **Zhang, L.** and Drikvandi, R. (2025). Distance-based CUSUM statistics for high dimensional change points. *Statistics and Computing,* 35(6): 1–19. `https://doi.org/10.1007/s11222-025-10752-1`.
- **Zhang, L.** and Drikvandi, R. (2023). High Dimensional Change Points: Challenges and Some Proposals. In *Proceedings of the 5th International Conference on Statistics: Theory and Applications.* Paper No. 142. `https://doi.org/10.11159/icsta23.142`.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Reza Drikvandi. His guidance has shaped every part of my research journey, from my Master's dissertation to this PhD thesis. I am grateful for his support in research, knowledge, academic writing, publication, and completing this thesis, as well as his help in developing my confidence and mindset during my twenties. His patience, advice, insightful life suggestions, and encouragement have made a strong and lasting impact on me. I also would like to thank Prof. Masoud Asgharian from McGill University and Dr. Hyeyoung Maeng from Durham University for assessing my thesis and providing very insightful comments, which greatly improved the quality and presentation of the thesis. I am also very grateful for both examiners' appreciation of my research work, which means a great deal to me.

I would like to thank my parents, my dad Yan Zhang and my mum Qunying Lu, for their endless love, belief in me, and unwavering support. Their understanding and encouragement have been the foundation that made it possible for me to pursue and complete this PhD.

I am grateful to the Department of Mathematical Sciences at Durham University. I have enjoyed my time here since the Master's programme and throughout the PhD, with support in academic training, seminars, excellent office space and facilities, and an extraordinary research atmosphere. I am also grateful for the partial financial support provided by the Scholarships and Student Funding Office.

Finally, I am thankful to my friends who have supported and encouraged me along the way. Their presence has made this long journey lighter and more meaningful.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background and motivations

Change point detection has applications across fields such as finance, marketing, engineering, climate science, genetics, and medical research (see Basseville and Nikiforov, 1993). Its main objective is to detect statistically significant changes in the distribution of a data sequence, which can have a huge real-world impact. For example, in finance, fluctuations in the stock market can lead to substantial losses for investors, and accurate detection can help to reduce this risk. Figure 1.1 shows stock market prices of 496 of the largest US companies during the COVID-19 period, which exhibit noticeable fluctuations. The dataset contains daily closing prices of S&P 500 index stocks from 2020-01-01 to 2020-05-29. This period includes the start of the COVID-19 pandemic and comprises $n = 108$ trading days across $p = 496$ company stocks. This dataset illustrates typical high dimensional data, where the number of variables, say p, is much larger than the number of observations, say n. We will analyze this financial data in Section 5.4.

Change point detection is well studied for low dimensional data. However, it becomes significantly more challenging in high-dimensional settings. Classical methods

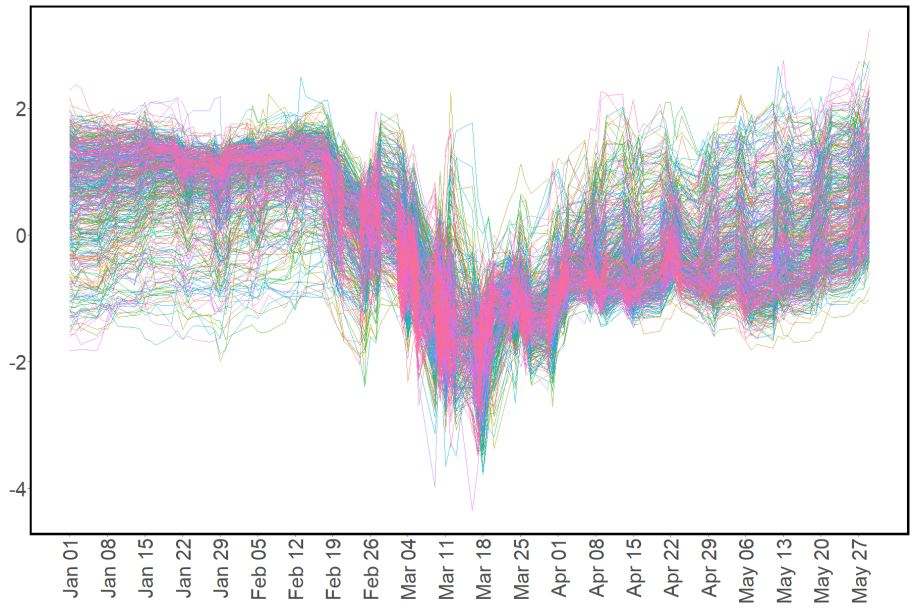


Figure 1.1: Standardized stock prices of 496 companies from 2020-01-01 to 2020-05-29. This period includes the start of the COVID-19 pandemic and comprises $n = 108$ trading days across $p = 496$ company stocks. Some key events cause fluctuations in the US stock market (see Table 5.4). The x-axis shows the trading days, and the y-axis shows the standardized stock prices.

are often inadequate in such scenarios, and we explain the reasons in Section 3.1. This thesis provides several nonparametric solutions to address some crucial challenges in high dimensional change points, making this thesis methodological. We study two types of change point detection problems: offline change point detection and online change point detection. Offline (retrospective) change point detection aims to estimate the number and locations of change points after all data are collected. Online change point detection, also known as sequential change point detection, deals with data streams arriving sequentially, with the objective of detecting change points as fast as possible while controlling the false alarm rate at a pre-specified low level. This differs substantially from the offline setting and requires different statistical techniques. Both offline and online change point detection form the foundation of change point analysis and have been widely studied in recent decades.

The motivations are as follows. For offline change point detection, a common approach for detecting change points is to utilize the standard cumulative sum (CUSUM) statistic, which is shown to effectively detect linear changes such as shifts in the mean of observations (see, e.g., Page, 1954; Aue and Kirch, 2024). However,

the standard CUSUM statistic is not designed to detect other types of changes, such as changes in variance or correlation, changes in the shape of distribution while the mean and variance remain unchanged, or other distributional changes. It is also not easily applied to high dimensional data sequences. Moreover, it remains an understudied problem to detect change points in high dimensional low sample size (HDLSS) data in which the number of observations is very small while the number of variables is very large (see our two real data applications in Section 5.4). The recent methods in the literature do not show a good performance in such situations where the sample size is very small compared to the dimension, as demonstrated in our simulated and real data examples (see Chapter 4, Sections 5.3 and 5.4). These challenges motivate us to develop methods that can address and accurately detect significant changes in the distribution of such high dimensional data streams.

Online change point detection has recently gained increasing attention for modern real-world applications, including sensor networks, cybersecurity, image processing, and many others (e.g., Chen et al., 2022). In the past two decades, online change point detection has been extensively studied for univariate data sequences (see, e.g., Fearnhead and Liu, 2007; Tartakovsky et al., 2014). With growing interest from contemporary industries in high dimensional data, traditional online change point methods become more challenging in high dimensional situations. Recent work has extended classical techniques to high dimensional spaces, such as likelihood ratio tests (e.g., Chen et al., 2022) and CUSUM statistics (e.g., Xie et al., 2023). However, those methods are mostly designed to detect changes in the mean of observations, which limits their ability to detect other types of distributional changes. Moreover, they often rely on certain assumptions about data distributions, such as normality or known pre-change or post-change distributions (e.g., Zou et al., 2015). These challenges motivate us to develop a distribution-free approach capable of detecting more general types of distributional changes in an online setting.

1.2 Contributions

This thesis contributes two nonparametric approaches for high dimensional change point detection. For offline change point detection, we propose distance-based CUSUM statistics based on some appropriate dissimilarity measures for high dimensional observations. In particular, we use the modified version of L_2-norm distance (Hall et al., 2005) and L_1-norm distance, which are especially suitable for HDLSS settings. Due to the asymptotic nature of these dissimilarity measures, our method is also useful for detecting non-sparse high dimensional change points where changes may occur in many variables but with small but significant magnitudes. Distance-based CUSUM statistics can discover both linear changes (e.g., mean shifts) and non-linear changes (e.g., changes in variance, correlation, or higher-order moments), as well as other distributional changes depending on the distance function used. We use the proposed CUSUM statistic to develop a formal nonparametric test based on random permutations to verify the statistical significance of the estimated change point locations. We study the theoretical properties of our distance-based CUSUM statistic, including the asymptotic limit when $p > n \to \infty$, where p is the dimension and n is the sample size. We also show that the proposed method can consistently detect multiple non-sparse change points as $p > n \to \infty$. Our approach does not require normality or any other distribution for high dimensional observations. We use extensive simulation studies and two real data examples to evaluate the empirical performance of our method in finite samples and compare it with some of the recent methods in the literature.

In the online (sequential) change point detection, where observations arrive sequentially, the focus is on fast detection. We employ the proposed distance-based CUSUM statistics in this sequential testing framework and construct a nonparametric method to detect high dimensional online change points. Building on the proposed distance-based CUSUM statistics, the proposed online change point method can detect both linear changes and non-linear changes in high dimensional data streams. In practice, real-world data streams could be large-scale, so we adopt a sliding window approach to manage computational and memory constraints. Another key contribution to the online change point framework is that we construct permutation-

based thresholds and a stopping rule for the online algorithm. We establish theoretical results for our proposed stopping rule under both the null hypothesis and the alternative hypothesis. We show that our procedure controls the family-wise error rate (FWER) of sequential tests. We also derive a lower bound to manage the average run length (ARL). ARL is defined as the expected time of the stopping rule under the null hypothesis (see Tartakovsky et al., 2014). Under the alternative hypothesis, we establish the consistency of the test and the convergence of the expected detection delay (EDD) under some conditions. Again, our online change point method does not require normality or any other distribution for the high dimensional observations. We conduct simulation studies and a real data analysis to evaluate the empirical performance of our proposed method and compare it with some of the recent methods in the literature.

1.3 Organization of the thesis

The thesis is organized as follows. In Chapter 2, we review some related work. In Section 2.1, we begin by introducing classical offline change point detection methods for low dimensional data, including CUSUM-based approaches and likelihood ratio tests. Since these classical methods do not extend directly to high dimensions, we review recent developments for high dimensional change points in Section 2.2, covering both parametric and nonparametric approaches. In Section 2.3, we give an introduction and review related work on online change point detection.

In Chapter 3, we propose a nonparametric approach for detecting a single (offline) change point in high dimensional data and provide theoretical guarantees. In Section 3.1, we discuss the main challenges for high dimensional change points and illustrate them with mathematical derivations and numerical evidence. In Section 3.2, we introduce the nonparametric setting for single change point detection. In Section 3.3, we propose distance-based CUSUM statistics and construct a formal nonparametric test. In addition to the method, we derive a permutation-based confidence interval for the change point estimate. In Section 3.4, we study the asymptotic behavior of distance-based CUSUM statistics when $p > n \to \infty$ and establish

consistency of the change point estimate under some conditions. In Section 3.5, we conclude the proposed method. Furthermore, in Section 3.6, we provide the technical proofs for all theoretical results developed in this chapter.

In Chapter 4, we conduct extensive simulation studies to evaluate the numerical performance of the proposed method for single change point detection. We compare our method with some recent works across various settings. We show that our method can detect more general types of distributional changes than other methods: when a change occurs in the mean of the observations (Section 4.1); when a change occurs in the variance (covariance) of the observations (Section 4.2); and when a change occurs in the shape of the observations while the mean and variance remain unchanged (Section 4.3). Moreover, we show that our method can be tailored to a range of data settings, some of which are challenging for existing methods. These simulations include non-normal data with heavier tails than the normal distribution (Section 4.4), spatially correlated variables (Section 4.5), and data with temporally dependent observations (Section 4.6). We also investigate challenging cases where a change point occurs near the tail of the data sequence (Section 4.7) and where the n/p ratio varies (Section 4.8), including regimes in which n is extremely small relative to p, as is typical in HDLSS data. Finally, in Section 4.9, we report computational times and show that our method is computationally efficient in high dimensions.

In Chapter 5, we extend our method to detect multiple change points. In Section 5.1, we introduce the problem setting for multiple change points. In Section 5.2, we present the algorithm that applies distance-based CUSUM statistics for multiple change point detection using recursive binary segmentation. We establish a consistency result for estimating both the number and the locations of change points under some conditions. In Section 5.3, we conduct simulation studies for the multiple change point scenario with changes in the mean or variance of the observations, the case where change points are closely spaced, and computational time. In Section 5.4, we apply our methods to two real-world high dimensional datasets, the S&P 500 data and the MIT cellphone data, and compare them with some existing methods. In Section 5.5, we discuss other alternatives, such as wild binary segmentation and the PELT method, which can be topics for future research. In Section 5.6, we provide

the proof of the consistency result.

In Chapter 6, we study the online change point problem and propose a nonparametric approach for detecting change points in high dimensional data streams with theoretical guarantees. In Section 6.1, we present the problem setting for online change point with a sliding window scheme. In Section 6.2, we introduce the proposed distance-based CUSUM statistics within an online framework. Moreover, we propose a stopping rule based on some empirical thresholds and construct a sequential testing algorithm. In Section 6.3, we develop theoretical guarantees for the proposed stopping rules. Under H_0, we control the FWER and ARL. Under H_1, we prove test consistency and the convergence of the EDD as $p \to \infty$ with fixed m, and as $p, m \to \infty$ under certain conditions. Here, m denotes the number of sequentially arriving observations. In Section 6.4, we present simulation studies that validate the derived theoretical results (FWER, ARL, detection power, EDD). Moreover, we show competitive performance for a change in mean and variance, and for some non-normal observations, paralleling the simulation results in the offline change point scenario. In Section 6.5, we apply the proposed method to a human activity recognition dataset. The results show that our method can rapidly detect transitions between human activities (standing and walking). We also demonstrate how our method can detect multiple online change points. In Section 6.6, we provide the technical proofs for all theoretical results developed for the proposed online change point method in this chapter.

Finally, in Chapter 7, we summarize the main contributions of this thesis, including the methodology, theoretical results, simulation studies, and real data applications. We highlight the advantages of the proposed nonparametric methods for offline and online change point detection, such as the capability to detect more general types of change points, the distribution-free property, and the suitability for high dimensional data, among others. Furthermore, we outline some directions for future research.

1.4 Notation

We adopt some notation throughout the thesis. For a p-dimensional random vector $\boldsymbol{u} = (u_1, u_2, \ldots, u_p)$, we write the L_q-norm as $\|\boldsymbol{u}\|_q = (\sum_{j=1}^p |u_j|^q)^{1/q}$. Also, we write the L_∞-norm as $\|\boldsymbol{u}\|_\infty = \max_{1\le j\le p} |u_j|$. We define the cardinality of a vector $\boldsymbol{u}$ as $|\boldsymbol{u}|_c = \sum_{j=1}^p \mathbb{1}(u_j \neq 0)$, where $\mathbb{1}(\cdot)$ denotes the indicator function and $|\boldsymbol{u}|_c$ represents the number of nonzero components. We use capital letters for all matrices, and in particular we write an $n \times p$ matrix $\boldsymbol{M}$ as $\boldsymbol{M} = [M_{ij}]_{i=1}^{n}{}_{j=1}^{p}$ with scalar M_{ij} being the element in the i-th row and j-th column. We write $O(\cdot)$, $o(\cdot)$ to denote the common big O and little o notation, respectively. Similarly, we write $O_P(\cdot)$ and $o_P(\cdot)$ to denote the big O in probability and little o in probability, respectively. For two real-valued sequences a_n and b_n, we use $a_n \asymp b_n$ to denote that a_n is of the same order as b_n. We use $\xrightarrow{P}$, $\xrightarrow{D}$, and $\xrightarrow{\text{a.s.}}$ to denote the convergence in probability, the convergence in distribution, and the almost sure convergence, respectively. We also use the symbol $\overset{D}{=}$ to represent the equality in distribution. For two scalars a and b, we use $a \wedge b$ and $a \vee b$ to denote $\min\{a, b\}$ and $\max\{a, b\}$, respectively. For a positive number c, we use $\lfloor c \rfloor$ to denote the largest integer less than or equal to c. We define some further notation later as we develop the methods.

CHAPTER 2

Literature review on high dimensional change points

The literature on change point analysis is extensive. In this chapter, we review state-of-the-art in offline change point detection, from traditional methods in low dimensions (Section 2.1) to recent advances in high dimensions (Section 2.2). For comprehensive reviews of offline change point problems, see Basseville and Nikiforov (1993), Chen and Gupta (2013), and Liu et al. (2022). We introduce online change point detection and review some related work in Section 2.3.

2.1 Classical change point approaches

The change point problem originated about 70 years ago in the context of quality control. It concerns three fundamental questions: whether a change point exists, where it is located, and how many change points there are. Before the era of big data, the problem was thoroughly studied in low dimensions, mostly for univariate data sequences. Classical methods, including the CUSUM method and the likelihood ratio procedure, were developed with complete theoretical justification.

We begin with a general parametric model to introduce the offline change point problem. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a sequence of independent p-dimensional random

vectors with their probability distributions $F_1, F_2, \ldots, F_n$. Suppose that the distributions $F_1(\boldsymbol{\theta}_1), F_2(\boldsymbol{\theta}_2), \ldots, F_n(\boldsymbol{\theta}_n)$ belong to a common parametric family $F(\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta} \in \mathbb{R}^p$, and $\boldsymbol{\theta}$ is unknown, then a general parametric model for the change point problem is to test the following hypothesis

Parametric change point problem:

$$\begin{cases} H_0: \ \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_n = \boldsymbol{\theta}, \\ H_1: \ \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_{\tau_1} \neq \boldsymbol{\theta}_{\tau_1+1} = \cdots = \boldsymbol{\theta}_{\tau_2} \neq \boldsymbol{\theta}_{\tau_2+1} = \cdots = \boldsymbol{\theta}_{\tau_z} \neq \boldsymbol{\theta}_{\tau_z+1} = \cdots = \boldsymbol{\theta}_n, \end{cases} \tag{2.1}$$

where z is the unknown number of change points with $1 \leq z \leq n-1$, and $\tau_1, \tau_2, \ldots, \tau_z$ are the unknown locations of change points. Both the number of change points and their locations in (2.1) need to be estimated.

It is difficult to test (2.1) directly when multiple change points are present ($z > 1$). Consequently, most work begins with the at-most-one-changepoint (AMOC) model, that is, (2.1) with $z = 1$, and then extends to multiple change points. The most common procedure for this extension is called recursive binary segmentation (RBS), proposed by Vostrikova (1981). Consider the data setting in model (2.1). The idea of RBS can be summarized in the following steps:

Step 1. Test the null hypothesis of no change point against the alternative hypothesis of a single change point (AMOC model) as follows

$$\begin{cases} H_0: \ \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_n = \boldsymbol{\theta}, \\ H_1^{\text{AMOC}}: \ \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_\tau \neq \boldsymbol{\theta}_{\tau+1} = \cdots = \boldsymbol{\theta}_n, \end{cases} \tag{2.2}$$

where τ is the single change point location at Step 1. If a significant change point is detected, then proceed to the next step. If no significant change point is detected, then stop.

Step 2. Split the sequence before and after the detected change point in Step 1 into two sub-sequences, and test them using the AMOC model (2.2) for further significant change points.

Step 3. Apply Step 2 recursively until no further sub-sequences contain significant change points or a pre-specified minimum sub-sequence length is reached. Note that the minimum length is often used to prevent greedy detection.

Step 4. Denote the significant change point locations discovered by the above steps and sort them in increasing order. This is because the change point estimates obtained in the above steps are not necessarily ordered.

In univariate data settings, RBS is computationally efficient with time complexity $O(n \log n)$, where n is the number of observations. Also, its simplicity in coding and incorporation into existing change point detection approaches are other reasons for its popularity. However, it has some limitations in detecting multiple change points when the sequence is small. Fryzlewicz (2014) showed that, as $n \to \infty$, RBS is not effective in detecting change points when the length of the sequence is less than $O(n^{3/4})$.

To overcome this limitation, Fryzlewicz (2014) proposed wild binary segmentation (WBS), which is a variant of RBS. Unlike RBS, which detects change points based on the entire data sequence, WBS detects change points based on randomly drawn sub-intervals within the data sequence. In particular, WBS first draws a large number of sub-intervals randomly from the data sequence. For each sub-interval, it performs the detection test, computes the test statistic, and stores the corresponding change point estimate. After this is done for all sub-intervals, WBS selects the best change point estimate as the one corresponding to the maximum test statistic. This procedure is then repeated iteratively until no sub-interval contains a significant change point or a minimum segment length is reached. Fryzlewicz (2014) showed that WBS performs better than RBS in detecting change points when the spacing between them is very small, and also when the jump magnitude, for example the mean shift, is small. We note that WBS improves on RBS through the use of a large number of randomly drawn sub-intervals, ideally as many as possible, which explains the term "wild". For example, Fryzlewicz (2014) recommended using 5000 drawn intervals in their paper. This slightly increases the computational cost in the univariate setting. Both RBS and WBS can be extended to high dimensional data, and we will discuss and compare them in high dimensional settings in Section 5.5.

There are other extensions of RBS and WBS. Fryzlewicz (2020) developed an improved version of WBS with a model selection procedure, which was shown to improve its computational efficiency and to be effective in detecting change points in both infrequent and frequent change point scenarios. Kovács et al. (2023) proposed seeded binary segmentation, which also improves the computational efficiency of WBS. Baranowski et al. (2019) developed the narrowest-over-threshold method, which searches for a change point using the shortest interval for which the test statistic exceeds the threshold, rather than selecting the maximum test statistic over all randomly drawn sub-intervals as in WBS. Both RBS and WBS are heuristic algorithms because each iteration depends on previous decisions. If the first detected change point is unreliable, subsequent estimates may also be affected. There are also other methods for detecting multiple change points, such as optimization-based approaches (e.g., Killick et al., 2012; Maidstone et al., 2017), which detect multiple change points by minimizing a customised cost function, and clustering-based methods (e.g., Matteson and James, 2014). Most of these methods were developed for low dimensional data, and can generally be extended to high dimensional settings. A comprehensive review of multiple change point detection techniques can be found in Truong et al. (2020).

From the above introduction to multiple change point techniques, it is clear that a reliable test for the AMOC model in (2.2) is crucial for both single change point and multiple change point detection. We now review the related work on single change point detection. In parametric change point analysis, the mean shift problem has received the most attention. Extensive studies have investigated this problem under Gaussian models, since the Gaussian assumption is common in practice and facilitates theoretical derivations. Recall the general parametric model in (2.1) with $z = 1$. Here, the parameter of interest is the mean vector. We now introduce the AMOC mean shift problem under the Gaussian model. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a sequence of n independent Gaussian random vectors, where each $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$ is p-dimensional. The parameters are $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \ldots, (\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, where $\boldsymbol{\mu}_i \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{p\times p}$. It is assumed that all covariance matrices are identical, i.e., $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is unknown. Suppose that the mean parameters

$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_n$ are identical to a common mean parameter $\boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is unknown, then the AMOC model for the mean shift problem is to test the following hypothesis

$$\textit{AMOC mean shift problem:} \quad \begin{cases} H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_n = \boldsymbol{\mu}, \\ H_1 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_\tau \neq \boldsymbol{\mu}_{\tau+1} = \cdots = \boldsymbol{\mu}_n, \end{cases} \tag{2.3}$$

where $\tau \in \{1, 2, \ldots, n-1\}$ is an unknown change point location. To test (2.3), the likelihood ratio procedure and the CUSUM method are commonly used. We introduce these standard approaches below.

The likelihood ratio test for the change point problem was first studied by Hinkley (1970). It is known for its robustness in detecting change points in parameters of interest. In the Gaussian model, these parameters may represent the mean vector in (2.3) or the covariance structure (e.g., Chen and Gupta, 2013). We here look at the AMOC mean shift model in (2.3) as an example. Under the null hypothesis of no change point, the likelihood function is

$$L_{H_0}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}) \Big\},$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of the covariance matrix $\boldsymbol{\Sigma}$. The maximum likelihood estimators (MLEs) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i, \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})^\top.$$

Under the alternative hypothesis that there is a change in the mean vector at $\tau \in \{1, 2, \ldots, n-1\}$, let $\boldsymbol{\Sigma}_{\text{pool}}$ denote the pooled covariance matrix computed from the first τ observations and the remaining $n-\tau$ observations. The likelihood function is

$$\begin{aligned} L_{H_1}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_{\text{pool}}) = (2\pi)^{-np/2} |\boldsymbol{\Sigma}_{\text{pool}}|^{-n/2} \exp\Big\{ -\frac{1}{2} \Big(& \sum_{i=1}^{\tau} (\boldsymbol{X}_i - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{\text{pool}}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_1) \\ & + \sum_{i=\tau+1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu}_n)^\top \boldsymbol{\Sigma}_{\text{pool}}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_n) \Big) \Big\}. \end{aligned} \tag{2.4}$$

The MLEs of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_n$, and $\boldsymbol{\Sigma}_{\text{pool}}$ are

$$\widehat{\boldsymbol{\mu}}_1 = \frac{1}{\tau}\sum_{i=1}^{\tau} \boldsymbol{X}_i, \qquad \widehat{\boldsymbol{\mu}}_n = \frac{1}{n-\tau}\sum_{i=\tau+1}^{n} \boldsymbol{X}_i,$$

and

$$\widehat{\boldsymbol{\Sigma}}_{\text{pool}} = \frac{1}{n}\Big(\sum_{i=1}^{\tau}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_1)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_1)^\top + \sum_{i=\tau+1}^{n}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_n)^\top\Big).$$

Given a fixed change point τ in (2.4), the first τ observations come from a normal distribution with mean $\boldsymbol{\mu}_1$, while the remaining $n-\tau$ observations come from a normal distribution with mean $\boldsymbol{\mu}_n$, with $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_n$. This motivates the use of Hotelling's T^2 test, which is based on the likelihood ratio test. The standardized mean difference between the two samples is defined as follows (Chen and Gupta, 2013)

$$\boldsymbol{M}(k) = \sqrt{\frac{k(n-k)}{n}}\Big(\overline{\boldsymbol{X}}_{1:k} - \overline{\boldsymbol{X}}_{k+1:n}\Big),$$

where $\overline{\boldsymbol{X}}_{1:k} = \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{X}_i$ and $\overline{\boldsymbol{X}}_{k+1:n} = \frac{1}{n-k}\sum_{i=k+1}^{n}\boldsymbol{X}_i$ are the sample averages before and after k, with $k \in \{1,\ldots,n-1\}$, respectively. The unbiased pooled covariance estimator based on these two sample averages is

$$\boldsymbol{N}(k) = \frac{1}{n-2}\Big(\sum_{i=1}^{k}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{1:k})(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{1:k})^\top + \sum_{i=k+1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{k+1:n})(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_{k+1:n})^\top\Big).$$

Then, the Hotelling's T^2 test statistic for testing (2.3) is defined as (Chen and Gupta, 2013)

$$T^2(k) = \boldsymbol{M}(k)^\top \boldsymbol{N}(k)^{-1}\boldsymbol{M}(k), \tag{2.5}$$

where $k \in \{1,\ldots,n-1\}$. One can reject H_0 when

$$\max_{1\leq k\leq n-1} T^2(k) > a,$$

where a is a constant threshold determined by the null distribution of $\max_{1\leq k\leq n-1} T^2(k)$, which can be found in Srivastava and Worsley (1986).

The CUSUM method was first proposed by Page (1954) as a simple procedure

to detect an increase in the proportion of defective products during manufacturing. The idea of the CUSUM statistic is to measure the difference between the sample means before and after each candidate change point $k \in \{1, 2, \ldots, n-1\}$ as follows

$$\boldsymbol{C}(k) = \sqrt{\frac{k(n-k)}{n}} \left(\frac{1}{n-k} \sum_{i=k+1}^{n} \boldsymbol{X}_i - \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{X}_i \right), \tag{2.6}$$

where $\boldsymbol{C}(k) \in \mathbb{R}^p$. If a change point occurs, the CUSUM statistic in (2.6) becomes large in magnitude. If there is no change point, the values on either side of the sample mean cancel out, and the CUSUM statistic remains small in magnitude. Thus, a natural estimator of the change point, based on the CUSUM statistic in (2.6), is given by

$$\hat{\tau} = \arg \max_{1 \le k \le n-1} \boldsymbol{C}(k)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{C}(k).$$

Correspondingly, the test statistic is defined as (Liu et al., 2022)

$$T_{\text{CUSUM}} = \boldsymbol{C}(\hat{\tau})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{C}(\hat{\tau}). \tag{2.7}$$

A large value of T_{CUSUM} suggests rejecting H_0 in (2.3), as it indicates a potentially significant shift in the sample means.

To assess the statistical significance of the estimated change point $\hat{\tau}$, the limiting null distribution of the test statistic T_{CUSUM} in (2.7) has been extensively studied (see Gombay and Horvath, 1990; Csörgö and Horváth, 1997). Let $\{B_1(t), B_2(t), \ldots, B_p(t)\}$ denote independent standard Brownian bridges with mean zero and covariance structure $E\big(B_1(t)B_1(s)\big) = t \wedge s - ts$. Under regular conditions such as i.i.d. Gaussian observations with finite fourth moments, the limiting null distribution of T_{CUSUM} as $n \to \infty$ has been shown as follows (Liu et al., 2022)

$$\frac{k(n-k)}{n^2} T_{\text{CUSUM}} \xrightarrow{D} \sup_{0 \le t \le 1} \sum_{j=1}^{p} B_j^2(t). \tag{2.8}$$

Critical values for the above distribution are provided in Kiefer (1959).

Besides the CUSUM method and the likelihood ratio approach under the Gaussian model, change point problems have also been studied under various other models,

leading to a wide range of alternative methods. For instance, Bayesian formulations treat the number and locations of change points as random and infer them jointly with segment-specific parameters (see Adams and MacKay, 2007). Change point detection has also been studied in linear regression models, where regression parameters may shift at unknown locations (see Zeileis et al., 2003). Information criterion-based methods, such as AIC and BIC, have been used to select both the number and locations of change points (see Yao, 1988). However, these methods are not suitable in high dimensional settings, and we explain the reasons later in Section 3.1. This highlights the need for new techniques specifically developed for high dimensional change point problems, which we introduce next.

2.2 Advances in high dimensional change points

In this section, we review recent developments for high dimensional change points. We begin with methods that adapt the classic CUSUM approach to high dimensional settings. Developments of the CUSUM procedure have been the focus of considerable research over a few decades (see Aue and Kirch, 2024, for a recent comprehensive review of this). Building on foundational work by Page (1954), Csörgö and Horváth (1997) established the limiting theorems for CUSUM-based methods in low dimensional spaces. In addition to the mean shift problem, some studies extend the standard CUSUM procedure to detect a change in the variance of multivariate data. For example, Montgomery (2007) applied CUSUM statistics to squared deviations to identify changes in variance. For change point problems focusing on detecting changes in the covariance or correlation structure in high dimensional data, several recent methods adapt CUSUM-type statistics or likelihood ratio statistics. Dette et al. (2022) proposed the CUSUM statistic based on the dimension-reduced important components to detect a change point in a high dimensional covariance structure. Dörnemann and Dette (2024) developed a min-type procedure that uses a sequential framework of likelihood ratio statistics. To detect changes in the correlation structure, Li and Gao (2024) proposed a two-step procedure employing a sign-flip permutation dimension reduction step and a CUSUM statistic to detect changing components in

a correlation matrix. Their method is effective for detecting change points located both in the middle and at the extreme tail of a data sequence.

The traditional CUSUM approach assumes the dimension of data is fixed and finite. But, due to the recent technological advancements, high dimensional data are becoming increasingly common in different domains. Change point analysis is more challenging in high dimensional situations. In general, detecting a significant change is difficult in the presence of high dimensional noise. Due to the singularity of the inverse covariance matrix in high dimensional settings, which we will show in Section 3.1, the classical CUSUM test in (2.6) is no longer applicable. A common strategy for applying the CUSUM procedure to high dimensional data is to first transform the data to a low dimensional space and then use the standard CUSUM approach. Adopting this strategy, Wang and Samworth (2018) suggested a two-stage procedure based on random projection to estimate change points in the mean of high dimensional data. They assumed sparse change points and normality of the data. Building on this idea, Follain et al. (2022) recently extended the random projection method to estimate high dimensional change points in the presence of missing data. Note that a sparse change point means that only a small number of variables change but with sufficiently large magnitudes. We next introduce the high dimensional mean shift problem, as well as the definition of sparsity in this context.

Considering the following data structure (Liu et al., 2022), for $i = 1, \ldots, n$,

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \Delta\boldsymbol{\mu}\, \mathbb{1}(i > \tau) + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector, $\Delta\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean shift vector if a change point $\tau \in \{1, \ldots, n-1\}$ exists, and $\mathbb{1}(\cdot)$ is the indicator function. The error terms $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})^\top$ are i.i.d. with $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. Under these assumptions, the AMOC model for a high dimensional mean shift problem ($p \gg n$) can be formulated as follows

$$\textit{High dimensional mean shift problem:} \begin{cases} H_0 : \Delta\boldsymbol{\mu} = \mathbf{0}, \\ H_1 : \exists\, \tau \in \{1, \ldots, n-1\} \text{ such that } \Delta\boldsymbol{\mu} \neq \mathbf{0}. \end{cases} \tag{2.9}$$

Note that the high dimensional mean shift model in (2.9) differs from the mean shift model in (2.3), which assumes low dimensions with $p < n$. We here introduce the definition of sparsity in this change point problem. Denote the mean shift vector in (2.3) by $\Delta\boldsymbol{\mu} = (\Delta\mu_1, \ldots, \Delta\mu_p)^\top$. Let $\rho = \{\, j : \ \Delta\mu_j \neq 0 \,\}$ be the set of variables that have undergone a change, and let $|\rho|_c$ denote its cardinality. Enikeeva and Harchaoui (2019) assumed that $|\rho|_c$ depends on p as follows

$$|\rho|_c \asymp p^{\,1-\beta},$$

where $\beta \in (0, 1)$ is the sparsity coefficient. High sparsity corresponds to $\beta \in (1/2, 1)$, while low sparsity (dense) corresponds to $\beta \in (0, 1/2]$. The value $\beta = 1/2$ marks the boundary between these regimes. It is worth mentioning that these sparsity boundary results are derived under the i.i.d. Gaussian model and should be interpreted in that context.

As explained below, we present two examples to illustrate how recent studies have adapted the classical CUSUM method to the high dimensional mean shift problem in (2.9) with sparse change points. The first example concerns detecting sparse change points with $\beta \in (1/2, 1)$. Jirak (2015) proposed an L_∞-norm CUSUM procedure under Gaussian assumptions with a known sparsity regime and polynomial growth $p \ll n^a$ for some $a > 0$. The statistic is

$$T_{L_\infty} = \max_{1 \le k \le n-1} \ \|\boldsymbol{C}(k)\|_\infty,$$

where $\boldsymbol{C}(k)$ is the CUSUM statistic defined in (2.6). Under H_0, T_{L_∞} admits a Gumbel-type limiting null distribution as follows (Liu et al., 2022)

$$\lim_{n\to\infty} \mathbb{P}\bigg(\sqrt{2\log(2p)}\Big(T_{L_\infty} - \Big(\tfrac{1}{2}\sqrt{2\log(2p)} - \tfrac{\log(3\log(2p))}{\sqrt{2\log(2p)}}\Big)\Big) \le x \bigg) = \exp(-e^{-x}).$$

Accordingly, a critical value $-\log\big(-\log(1-\alpha)\big)$ can be used to control the type I error at nominal level α. However, the L_∞-norm CUSUM is powerful mainly under very sparse alternatives. By focusing on the maximum coordinate, it discards information when signals are not extremely sparse, which leads to power loss.

The second example is by Enikeeva and Harchaoui (2019). They proposed a χ^2-type CUSUM test that also assumes normality but allows different sparsity levels. For a fixed k, the coordinates of $\boldsymbol{C}(k)$ (after standardization) are approximately standard normal, so that $\|\boldsymbol{C}(k)\|_2^2$ has an approximate χ_p^2 distribution with mean p and variance $2p$. A linear test statistic is then given by (Enikeeva and Harchaoui, 2019)

$$T_{\text{Linear}} = \max_{1 \le k \le n-1} \frac{\|\boldsymbol{C}(k)\|_2^2 - p}{\sqrt{2p}}.$$

Given a significance level α and the $(1-\alpha)$-quantile of the χ_p^2 distribution, denoted by $q_{\chi_p^2}(1-\alpha)$, one can reject H_0 in (2.9) when

$$T_{\text{Linear}} > \frac{q_{\chi_p^2}(1-\alpha) - p}{\sqrt{2p}}.$$

In addition to the linear statistic, Enikeeva and Harchaoui (2019) also introduced a scan statistic to control the overall type I error across different sparsity levels; see their paper for more details. We note that these methods require a known sparsity level and Gaussian assumptions. Their performance can deteriorate when these assumptions are violated. Such techniques are therefore not suitable for problems with non-sparse change points, where changes occur in many variables but each with small but significant magnitudes.

Apart from the aforementioned methods, other parametric approaches have also been developed. Grundy et al. (2020) proposed a geometric mapping approach that projects high dimensional data onto a two dimensional space using distance and angle measures. Their method assumes normality and can detect changes in the mean or variance of observations. However, it fails to detect changes in higher-order moments of the distribution. Safikhani and Shojaie (2022) suggested a piecewise vector autoregressive model to select high dimensional variables, which is solved using a penalized least squares estimator. Their method is effective when the underlying VAR model is correctly specified. However, misspecification of the autoregressive order or assuming linear dependence when the true structure is non-linear can lead to poor detection performance. Xiao et al. (2019) proposed a robust change point method based on principal component analysis. It performs well when signals are

concentrated in a few principal directions, but its effectiveness can degrade when changes occur in weaker components or when the covariance structure is complex. Hahn et al. (2020) developed a computationally inexpensive Bayesian approach to estimate an optimal projection direction for detecting change points. Their method provides uncertainty quantification and performs well when changes occur in the mean of multivariate normal data. However, it is designed specifically for mean shifts under normality and does not detect other types of changes. Furthermore, Jewell et al. (2022) developed a post-detection procedure that uses the CUSUM statistic to assess uncertainty around an estimated change point. Their method tests whether there is truly a mean shift in a neighbourhood of the detected location, opening a new direction in change point analysis by providing inference after detection.

Nonparametric methods are often more practical than parametric methods, since they do not require specific distributional assumptions or a prescribed sparsity level. In the univariate nonparametric setting, Padilla et al. (2021) proposed a novel change point detection procedure based on the Kolmogorov–Smirnov statistic and showed that it is nearly minimax rate optimal. They also demonstrated that the method is effective in detecting small distributional changes. In high dimensional space, one strategy is to use distance functions, such as the Euclidean distance. However, classical distance-based methods for change point detection, such as Carlstein (1988) and Dümbgen (1991), do not apply to high dimensional data. Recently, several nonparametric methods have been developed for detecting high dimensional change points. A common approach is to construct statistics from dissimilarity distances between pairs of observations. Matteson and James (2014) proposed an L_2-based divergence measure using the energy distance to quantify differences between multivariate distributions. Energy distance was first proposed by Szekely et al. (2005) for classification. Assume that $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^p$ are p-dimensional observations, and $\boldsymbol{X} \sim F$, $\boldsymbol{Y} \sim G$, where F and G are unknown distributions. Then, energy distance based on L_2-norm is given as follows (Szekely et al., 2005)

$$\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \alpha) = 2E(\|\boldsymbol{X} - \boldsymbol{Y}\|_2^\alpha) - E(\|\boldsymbol{X} - \boldsymbol{X}'\|_2^\alpha) - E(\|\boldsymbol{Y} - \boldsymbol{Y}'\|_2^\alpha), \tag{2.10}$$

where $\boldsymbol{X}'$ and $\boldsymbol{Y}'$ are independent copies of $\boldsymbol{X}$ and $\boldsymbol{Y}$, meaning that $\boldsymbol{X}'$ and $\boldsymbol{X}$ are i.i.d., and $\alpha \in (0, 2)$ is a fixed constant. The following theorem provides the theoretical foundation of the energy distance for change point detection.

Theorem 2.2.1. *(Szekely et al. (2005)) For any independent p-dimensional random vectors $\boldsymbol{X} \sim F$, $\boldsymbol{Y} \sim G$ and constant $\alpha \in (0, 2)$, if $E(\|\boldsymbol{X}\|_2^\alpha + \|\boldsymbol{Y}\|_2^\alpha) < \infty$, then $\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \alpha) \in [0, \infty)$, and $\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \alpha) = 0$ if and only if $F = G$.*

Theorem 2.2.1 shows that $\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \alpha) = 0$ when there is no change point, as the two distributions are identical ($F = G$). Moreover, large values of $\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \alpha)$ indicate a change point exists. Motivated by this theorem, Matteson and James (2014) constructed U-statistics based on energy distance. They assumed two independent data sequences $\boldsymbol{X}_{1:h} = \{\boldsymbol{X}_i : i = 1, \ldots, h\}$ and $\boldsymbol{Y}_{1:m} = \{\boldsymbol{Y}_i : i = 1, \ldots, m\}$, with a total of $h + m$ observations. The remaining settings are the same as above, where $\boldsymbol{X}_{1:h}, \boldsymbol{Y}_{1:m} \in \mathbb{R}^p$, $\boldsymbol{X}_{1:h} \sim F$, $\boldsymbol{Y}_{1:m} \sim G$, and F and G are unknown distributions. Then, the empirical divergence measure corresponding to equation (2.10) is defined as follows (Matteson and James, 2014)

$$\begin{aligned}\widehat{\mathcal{D}}(\boldsymbol{X}_{1:h}, \boldsymbol{Y}_{1:m}, \alpha) = \frac{2}{hm}\sum_{i=1}^{h}\sum_{j=1}^{m}\|\boldsymbol{X}_i - \boldsymbol{Y}_j\|_2^\alpha - \binom{h}{2}^{-1}\sum_{1\le i<k\le h}\|\boldsymbol{X}_i - \boldsymbol{X}_k\|_2^\alpha \\ - \binom{m}{2}^{-1}\sum_{1\le j<k\le m}\|\boldsymbol{Y}_j - \boldsymbol{Y}_k\|_2^\alpha, \end{aligned} \tag{2.11}$$

where $\alpha \in (0, 2)$. By the strong law of large numbers and the continuity theorem, Matteson and James (2014) showed the following convergence result

$$\lim_{h\wedge m\to\infty}\widehat{\mathcal{D}}(\boldsymbol{X}_{1:h}, \boldsymbol{Y}_{1:m}, \alpha) \overset{a.s.}{\to} \mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}, \alpha).$$

Suppose there exists a change point τ such that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau \sim F$ and $\boldsymbol{X}_{\tau+1}, \ldots, \boldsymbol{X}_n \sim G$, where $F \neq G$. Then, a change point estimate based on (2.11) is given by

$$\hat{\tau} = \arg\max_{1\le k\le n-1}\frac{k(n-k)}{n}\widehat{\mathcal{D}}(\boldsymbol{X}_{1:k}, \boldsymbol{X}_{k+1:n}, \alpha).$$

Matteson and James (2014) showed the following consistency result for the above

change point estimate.

Theorem 2.2.2. *(Matteson and James (2014)) Let $\gamma \in (0,1)$ represent the fraction of the observations belonging to one of the distributions, such that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\lfloor \gamma n \rfloor} \sim F$ and $\boldsymbol{X}_{\lfloor \gamma n \rfloor + 1}, \ldots, \boldsymbol{X}_n \sim G$. Suppose that the conditions in Theorem 2.2.1 hold. For all $\epsilon > 0$, they showed that*

$$P\Big(\lim_{n \to \infty} \Big| \gamma - \frac{\hat{\tau}}{n} \Big| < \epsilon \Big) = 1.$$

There are some other distance-based methods. Expanding on the work in Matteson and James (2014), Chakraborty and Zhang (2021) developed an L_1-norm-based energy distance for change point detection. Their framework studies how distance and kernel metrics behave in high dimensions and shows that classical energy distances suffer from strong bias when the dimension increases. By using an L_1-norm-based modification, they obtain a distance measure that remains informative in high dimensional settings and improves robustness to heavy-tailed distributions. Some graph-based methods, including those in Chen and Zhang (2015), Garreau and Arlot (2018), and Chu and Chen (2019), also utilize interpoint distances to search for change points by counting the number of edges in a similarity graph. In addition, Drikvandi and Modarres (2025) recently introduced a method for detecting non-sparse high dimensional change points using a new dissimilarity measure and a nonparametric U-statistic. Their method is shown to be effective for change point detection in HDLSS data. They also establish the limiting null distribution of their test statistic and study the optimality of the proposed test, which is a challenging task in nonparametric change point analysis. Li (2020) introduced a distance-based change point method that remains asymptotically distribution-free in high dimensional settings. The method is shown to consistently detect both location and scale changes. It should be mentioned that there are several other methods for high dimensional change point analysis, some of which are reviewed in the recent review paper by Liu et al. (2022).

Compared to these distance-based methods, the novelty of our approach is in introducing novel CUSUM statistics constructed based on pairwise dissimilarity distances between observations, rather than just pairwise observations. This is

different from other methods, such as energy statistics, which directly measure dissimilarity using the Euclidean distance. In contrast, our distance-based CUSUM method measures dissimilarity between observations before and after a change point, which provides an effective way to detect more general types of change points depending on the distance function used. In Chapter 4, we show that our proposed method outperforms some of these methods in detecting change points under different scenarios. Furthermore, the theoretical guarantee of our method is developed under the high dimensional asymptotic regime $p > n \to \infty$, which differs from their low dimensional setup where $n \to \infty$ with fixed p. The latter is not suitable for the high dimensional regime.

2.3 Online change point detection

As modern data arrive faster and become richer, rapid detection has become increasingly important, and online change point detection fits this need well. Online change point detection, also known as sequential change point detection, has received considerable interest in recent years. For sequentially observed data, the goal is to detect any significant change in the distribution of observations as quickly as possible while controlling the false alarm rate at a pre-specified low level. Online change point detection is arguably becoming more popular than offline change point detection for modern real-world applications, such as sensor networks, cybersecurity, image processing, and many others (Chen et al., 2022).

We start by introducing the setup of online change point detection and some key criteria. Consider the same data setting as in the parametric change point model (2.1), with p-dimensional and independent observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ arriving sequentially. We use the index t to denote the time of arrival, where $t = 1, 2, \ldots$. At each time point, we aim to test whether a change point has occurred in the arriving observations as follows

$$\begin{cases} H_0^t : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_t, \\ H_1^t : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_\tau \neq \boldsymbol{\theta}_{\tau+1} = \cdots = \boldsymbol{\theta}_t, \end{cases} \quad \text{for } t = 1, 2, \ldots, \tag{2.12}$$

where τ is the unknown change point. In online change point detection, it is often convenient to write $\tau = \infty$ to denote that there is no change point. If H_0^t is rejected, we infer that a change point is detected and stop the procedure at time t. Note that model (6.1) can be slightly modified when considering other data scenarios, such as the presence of historical data (e.g., Li and Li, 2023) or the use of a sliding window, which is commonly used to reduce computational complexity (e.g., Xie et al., 2023). We will discuss this in Section 6.1.

Unlike the offline change point problem, which mainly focuses on estimating the location and number of change points, online change point detection aims to detect a change point as soon as possible. For example, under H_1^t in (6.1), which implies there is a true change point, online change point detection aims to minimize the delay between the stopping time t and the true change point with a controlled false alarm. This is also a more difficult problem because of the sequential testing framework. Hence, the first challenge is to control false alarms.

One common measure of false alarm control in online change point detection is the average run length (ARL). Suppose a chosen test is performed in an online change point setting, and its test statistic at each time t is denoted by $T(t)$, with threshold a. The ARL is defined as the expected number of observations until the first false alarm occurs, that is,

$$\mathrm{ARL} = \mathbb{E}_\infty \left(\min\{t \in \mathbb{N} : T(t) > a\} \right), \tag{2.13}$$

where the expectation $\mathbb{E}_\infty(\cdot)$ is taken under the null hypothesis of no change point ($\tau = \infty$). In practice, the ARL level is pre-specified. A large ARL makes the procedure less sensitive and the null hypothesis harder to reject, which may lead to a larger detection delay. In contrast, a small ARL makes the procedure more sensitive to change points, but also more likely to raise false alarms. Hence, the choice of ARL is important. Some common choices of ARL in the existing literature are around 1000 to 5000 (e.g., Chen et al., 2022; Li and Li, 2023).

Another criterion for controlling false alarms in sequential testing frameworks is the family-wise error rate (FWER). Consider the test setting in (2.13) and suppose

there are a total of m arriving observations. The FWER is defined as the probability of making the first false detection among the m arriving observations, that is,

$$\text{FWER} = \mathbb{P}_\infty \left(\min \left\{ 1 \leq t \leq m : T(t) > a \right\} \leq m \right), \tag{2.14}$$

where the probability $\mathbb{P}_\infty(\cdot)$ is taken under the null hypothesis of no change point ($\tau = \infty$). Ideally, the FWER in (2.14) is controlled below the nominal level, such as FWER ≤ 0.05. To achieve this, one can consider using the Bonferroni correction (e.g., Austin et al., 2023; Zhang et al., 2025).

Besides false alarm control, another important measure is the expected detection delay (EDD). Consider the same test setting in (2.14) with m arriving observations. Under the alternative hypothesis, which implies that there is a true change point $\tau_0 \in \{1, \ldots, m-1\}$, the EDD can be formulated as

$$\text{EDD}(\tau_0) = \mathbb{E}_{\tau_0} \left(\min\{1 \leq t \leq m : T(t) > a\} - \tau_0 \,|\, \min\{1 \leq t \leq m : T(t) > a\} > \tau_0 \right). \tag{2.15}$$

The main objective of online change point detection is to achieve a small EDD while controlling either the ARL in (2.13) or the FWER in (2.14). We next review some related work on online change point detection.

In the past two decades, online change point detection has been extensively studied for univariate data sequences (see, e.g., Fearnhead and Liu, 2007; Tartakovsky et al., 2014). General reviews of online (sequential) change point problems include Basseville and Nikiforov (1993) and Tartakovsky et al. (2014). With growing interest from contemporary industries in high dimensional data, traditional change point analysis has become more challenging in high dimensional situations. Recent work has extended classical techniques to high dimensional spaces, such as likelihood ratio tests (e.g., Chen et al., 2022) and CUSUM statistics (e.g., Xie et al., 2023). However, these methods are mostly designed to detect changes in the mean of observations, which limits their ability to detect other types of distributional changes. Moreover, they often rely on certain assumptions about data distributions, such as normality or known pre-change or post-change distributions (e.g., Zou et al., 2015). These

motivate us to develop a distribution-free approach capable of detecting more general types of distributional changes in online settings.

The recent literature on online change point detection can be roughly reviewed from two perspectives, namely parametric and nonparametric methods. In the parametric paradigm, some researchers extended the CUSUM statistic of Page (1954) to the online change point setting. As discussed in Section 2.1, the CUSUM method is also very popular in online change point detection. We introduce how CUSUM can be applied here in the univariate case. Consider the same data setting as in the parametric change point model (2.1), with univariate independent observations $X_1, X_2, \ldots$ arriving sequentially. Denote the cumulative sum by $S_t = \sum_{i=1}^{t} X_i$. The online CUSUM procedure stops at time t when

$$S_t - \min_{0 \le i < t} S_i \ge a, \tag{2.16}$$

where a is a pre-specified threshold. If there is a significant positive mean shift relative to the previous minimum cumulative sum at time t, then the statistic in (2.16) becomes large and indicates rejection of the null hypothesis at time t. The online CUSUM statistic can also be written recursively as

$$Z_t = S_t - \min_{0 \le i < t} S_i = \max\{Z_{t-1} + X_t, 0\}, \tag{2.17}$$

where $Z_0 = 0$, and the procedure stops when $Z_t \ge a$. We note that the above CUSUM procedure is one-sided and only detects a positive mean shift. To address this issue, one may use a two-sided or double CUSUM procedure proposed by Waldmann (1996) to detect both negative and positive mean shifts. Building on the work of Page and Wald, the online CUSUM procedure has been well studied in the univariate case. This includes theoretical studies such as Lorden (1971) and Lai (1998). We refer readers to Aue and Kirch (2024) for further details on online CUSUM procedures.

We now turn to some recent advances in online CUSUM procedures. These include the moving window CUSUM (MOSUM) proposed by Horváth et al. (2008), the modified MOSUM (mMOSUM) proposed by Chen and Tian (2010), and the functional online CUSUM (FOCuS) algorithm based on generalized likelihood ratios

proposed by Romano et al. (2023). Also, Kirch and Weber (2018) studied some of those CUSUM-based online detection approaches. These CUSUM methods are shown to effectively detect changes in the mean of observations. Moreover, they assume fixed or finite dimensions and require knowledge of pre-change or post-change distributions, limiting their applicability to modern data stream settings featured by high dimensionality and unknown complicated distributions. Likelihood ratio tests have also been developed for detecting mean shifts in high dimensional observations (e.g., Mei, 2010; Xie and Siegmund, 2013; Chan, 2017). However, their performance often relies on assumptions of normality and independence. In addition, Bayesian tests have been proposed for univariate data sequences (e.g., Adams and MacKay, 2007). In Section 6.4, we show that our proposed method outperforms some of these approaches in the scenarios of correlated normal data as well as some non-normal cases.

Nonparametric approaches are more challenging in the online setting due to the lack of likelihood information or knowledge of the underlying data distribution (e.g., Wang and Xie, 2024). In the univariate nonparametric setting, Yu et al. (2023) developed a CUSUM-based approach that controls both the ARL and the FWER while minimizing the detection delay in a minimax sense. They also showed how their method can be extended to estimate multiple online change points sequentially. In the multivariate nonparametric setting, a popular class of nonparametric methods uses the scan B-statistic based on kernel Maximum Mean Discrepancy (MMD) (see, e.g., Li et al., 2019b). Wei and Xie (2026) proposed a computationally efficient kernel-based CUSUM method for online change point detection. They showed that their method is effective in detecting small changes compared with the scan B-procedure. However, their work mainly focuses on multivariate settings with moderate dimensions. In their simulation settings, the dimension is fixed at $p = 20$ while the sample size is in the thousands. They also noted that extending kernel-based methods to high dimensional settings is challenging due to the curse of dimensionality (see Ramdas et al., 2015; Reddi et al., 2015). Another strategy for handling high dimensional data is to use the interpoint distances among observations. Currently, most existing dissimilarity-based methods are developed for offline change point

detection (e.g., Matteson and James, 2014; Li, 2018; Drikvandi and Modarres, 2025; Zhang and Drikvandi, 2025). Substantial effort is required to extend such methods to the framework of online change points. Some nonparametric online change point methods have recently been developed for specific data applications. For example, Li and Li (2023) proposed an online procedure, building on Avanesov and Buzun (2018), to detect changes in the covariance structure of dynamic network data. Austin et al. (2023) developed a nonparametric likelihood test to monitor the operational performance of network devices. It is worth mentioning that there are a few other methods for sequential change points, which are reviewed in the recent review paper by Wang and Xie (2024).

CHAPTER 3

Distance-based CUSUM for high dimensional change points

In this chapter, we introduce distance-based CUSUM for change points (DCCP), a nonparametric approach for detecting high dimensional non-sparse change points (Zhang and Drikvandi, 2025). In Section 3.1, we discuss the key challenges in high dimensional change points that motivate our methods. In Section 3.2, we present the AMOC model in a nonparametric setting. In Section 3.3, we develop distance-based CUSUM statistics and a permutation test to form a nonparametric procedure for estimating the change point and constructing its confidence interval. Furthermore, we provide a single change point detection algorithm for high dimensional observations ($n \times p$ data matrix with $p \gg n$). In Section 3.4, we derive asymptotic results for the distance-based CUSUM and prove consistency of the estimated change point under the alternative when $p > n \to \infty$, with some conditions. In Section 3.5, we provide some concluding remarks for this chapter. In Section 3.6, we provide the technical proofs for all theoretical results developed in this chapter. Furthermore, we conduct the simulation studies for single and multiple change point detection, as well as the real data applications, in Chapter 4, Sections 5.3 and 5.4, respectively.

3.1 Challenges in high dimensional change points

With the rapid developments in computing mechanisms and data storage, the world has already entered the era of big data. This attracts wide attention because it appears in numerous modern applications. A key feature of such data is the presence of thousands or even millions of variables per subject or individual. So, high dimensional data are an important aspect of big data, where the number of variables greatly exceeds the number of observations ($p \gg n$). An illustrative example is provided in Figure 1.1. Such data examples appear in many fields, including financial applications such as stock returns of hundreds of companies in the S&P 500 index (e.g., Zhang and Drikvandi, 2025), genetic studies with tens of thousands of gene expressions (e.g., Wang and Samworth, 2018), functional magnetic resonance imaging (fMRI) studies in neuroscience (e.g., Zhong et al., 2021), human activity recognition (e.g., Zhang et al., 2025), network traffic detection (e.g., Lévy-Leduc and Roueff, 2009), natural language processing, and astronomy for detecting distant galaxies (e.g., Enikeeva and Harchaoui, 2019).

In change point problems, high dimensional data often provide richer information for analysis but are considerably more challenging to handle, particularly from a theoretical perspective. Due to the curse of dimensionality and related difficulties, classical methods often fail to work in high dimensions. For a general review of high dimensional statistics, see Bühlmann and Van De Geer (2011); Giraud (2021). We next highlight the key challenges specific to change point problems.

For parametric change point methods, obtaining a reliable covariance estimate is difficult when the dimension is larger than the sample size. In high dimensional settings with $n \ll p$, standard estimators become singular or poorly conditioned, which leads to poor results. An accurate covariance estimation requires sample sizes that grow with the dimension, yet in practice high dimensional data often come with small n. Consider the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{X}$ with $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ (all columns centered) and $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p\times p}$. Simple algebra shows the nonexistence of $\widehat{\boldsymbol{\Sigma}}^{-1}$ because

$$\operatorname{rank}(\widehat{\boldsymbol{\Sigma}}) \leq \min\big(\operatorname{rank}(\boldsymbol{X}^\top), \operatorname{rank}(\boldsymbol{X})\big) \leq \min(n, p) = n \ll p.$$

Therefore, $\widehat{\mathbf{\Sigma}}$ is not invertible in high dimensions when $n < p$.

The nonexistence or ill-conditioning of $\widehat{\mathbf{\Sigma}}^{-1}$ makes routine change point methods struggle in high dimensions. For example, the Hotelling's T^2 test in (2.5) is impractical in this case. The classical CUSUM method in (2.8) is also not applicable. Several methods have been developed to estimate covariance matrices in high dimensions (see Cai et al. (2016) for a review). These methods rely on structural assumptions such as sparsity or banding and require tuning, which complicates their use in CUSUM or likelihood-based procedures and in the associated theory. These issues raise the need for developing test statistics for high dimensional change point problems that avoid using covariance matrix inversion.

Another challenge is that traditional methods lack theoretical justification under high dimensional asymptotic settings. Classical results typically assume $n \to \infty$ with fixed p, such as the limiting distribution in (2.8) for the CUSUM method and Theorem 2.2.1 for energy distance-based procedures. These approaches perform well in low dimensional settings; however, modern datasets often violate this assumption. Recall that in genomics, there may be only a few patients while the number of measured genes is in the thousands or even millions. In finance, fluctuations in stock prices across hundreds of companies are often more relevant in the recent three months than in the past ten years. Consequently, asymptotic regimes such as $p > n \to \infty$ or $p \to \infty$ with fixed n (HDLSS data) are important but have been much less studied in change point analysis. In these regimes, the null distribution is difficult to characterize, complicating calibration to a target type I error level $\alpha \in (0, 1)$. This limitation prevents direct extensions of classical methods and motivates the development of test statistics specifically designed for high dimensional change point problems.

Another challenge in high dimensional change point problems is detecting non-sparse change points. Many existing methods assume normality and sparsity, which reduces the problem to a low dimensional structure (see Wang and Samworth, 2018; Enikeeva and Harchaoui, 2019), and then apply classical procedures such as CUSUM or likelihood ratio tests. When these assumptions fail, their test power often drops substantially. In many applications, the change is dense, with many coordinates

shifting by small amounts. As an illustrative example shown in Figure 1.1, during the COVID-19 period, most companies experienced fluctuations in their stock prices. In such dense patterns, coordinate-wise tests (e.g., random projection approaches) and sparsity-based procedures lose information across coordinates and underperform. In Chapter 4, we show that several of the parametric methods perform poorly in these scenarios.

To construct change point tests without relying on sparsity and normality assumptions, many nonparametric methods use distance functions, especially the L_2 norm (Euclidean distance), to build a formal nonparametric test (see Matteson and James, 2014; Chen and Zhang, 2015). However, the L_2 norm distance diverges as the dimension p increases. To see this, let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. p-dimensional standard normal observations. Then $\sum_{l=1}^{p} X_{il}^2 \sim \chi_p^2$ with mean p and variance $2p$. By the central limit theorem, as $p \to \infty$, one can show that

$$\frac{\sum_{l=1}^{p} X_{il}^2 - p}{\sqrt{2p}} \xrightarrow{D} N(0,1).$$

Let $g(x) = \sqrt{x}$. Since the derivative of $g(x)$ is $g'(x) = 1/(2\sqrt{x})$, we have $g'(p) = 1/(2\sqrt{p}) \neq 0$. By the Delta method, Hall et al. (2005) showed that

$$\|\boldsymbol{X}_i\|_2 = \Big(\sum_{l=1}^{p} X_{il}^2 \Big)^{1/2} = \sqrt{p} + O_P(1),$$

as $p \to \infty$. Similarly, because $X_{il} - X_{jl} \sim N(0,2)$, we obtain $\sum_{l=1}^{p} (X_{il} - X_{jl})^2 \sim 2\chi_p^2$. Therefore, one can show that

$$\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2 = \Big(\sum_{l=1}^{p} (X_{il} - X_{jl})^2 \Big)^{1/2} = \sqrt{2p} + O_P(1),$$

as $p \to \infty$.

Thus, pairwise L_2-norm distances grow at a rate $\sqrt{p}$. This distance divergence can make Euclidean distance-based tests unsuitable in the high dimensional regime. For example, the assumptions in Theorem 2.2.1 may fail, since $E(\|\boldsymbol{X}\|_2^{\alpha} + \|\boldsymbol{Y}\|_2^{\alpha})$ can be unbounded as $p \to \infty$. In Figure 3.1, we illustrate the pairwise L_2-norm distances between p-dimensional standard normal observations for $p \in \{5, 10, 100, 1000\}$. It

can be seen that the distances increase with p (at a rate $\sqrt{p}$). Moreover, in Figure 3.2, we report the pairwise L_1 norm (Manhattan) distances. It can be seen that the Manhattan distances exhibit the same divergence phenomenon and increase even faster than the Euclidean distances. Empirically, the L_1 distances grow approximately linearly with p. We later address these divergence issues in Section 3.3.

A related implication of divergent distances is the loss of locality in high dimensional settings. As p increases with fixed n, the minimal pairwise Euclidean distance grows, and most pairwise distances become nearly equal. In this regime, "nearest neighbors" are not meaningful. So, estimators based on local averaging (e.g., k-nearest neighbors and kernel smoothing) do not behave as intended. As shown in Figures 3.1 and 3.2, the histograms shift to the right and tighten, which leaves each point far from the others and at roughly the same distance as p grows.

Kernel-based methods that depend on Euclidean distance are affected for the same reason as above. Gaussian and Laplace kernels map distances to similarities. When distances are nearly equal, the kernel values are almost constant, which reduces performance (see Ramdas et al., 2015). These effects explain why classical distance-based procedures are not suitable in high dimensional settings. These reasons motivate the use of rescaled or modified dissimilarity measures, which we introduce later in Section 3.3.

Another difficulty in high dimensional change point problems is the increasing computational cost. Basic operations with a $p \times p$ matrix, such as multiplication or inversion, require on the order of p^α operations with $\alpha > 2$, so repeated use is expensive when p is in the thousands. For example, covariance-based tests require repeated inversions of $p \times p$ matrices, which is $O(p^3)$ per candidate split and becomes impractical when p is large. Model (variable) selection that compares many subsets quickly becomes infeasible: there are 2^p possible subsets of p variables (each variable is either included or not). Even with $p = 30$, this is $2^{30} = 1{,}073{,}741{,}824$ subsets (see also Giraud, 2021).

Last but not least, most of the methods in the literature can only detect a change in the mean or variance of observations. To detect other distributional (higher-order moments) changes is even more challenging in high dimensional situations. One

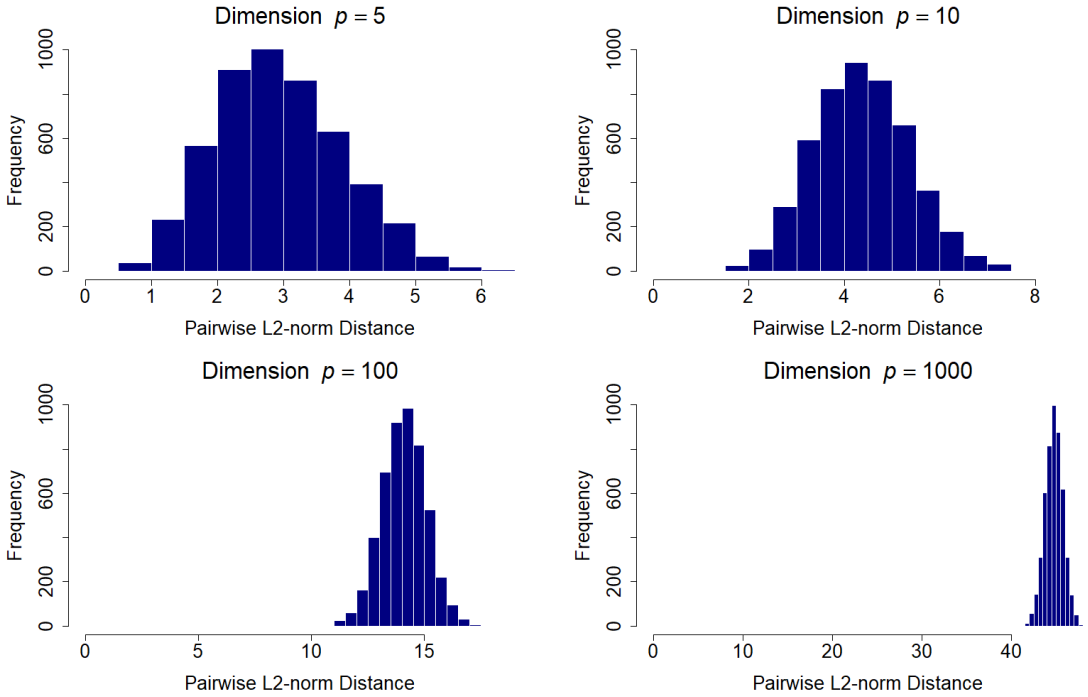


Figure 3.1: Histograms of the pairwise L_2-norm distances among $n = 100$ observations generated from a p-variate standard normal distribution.

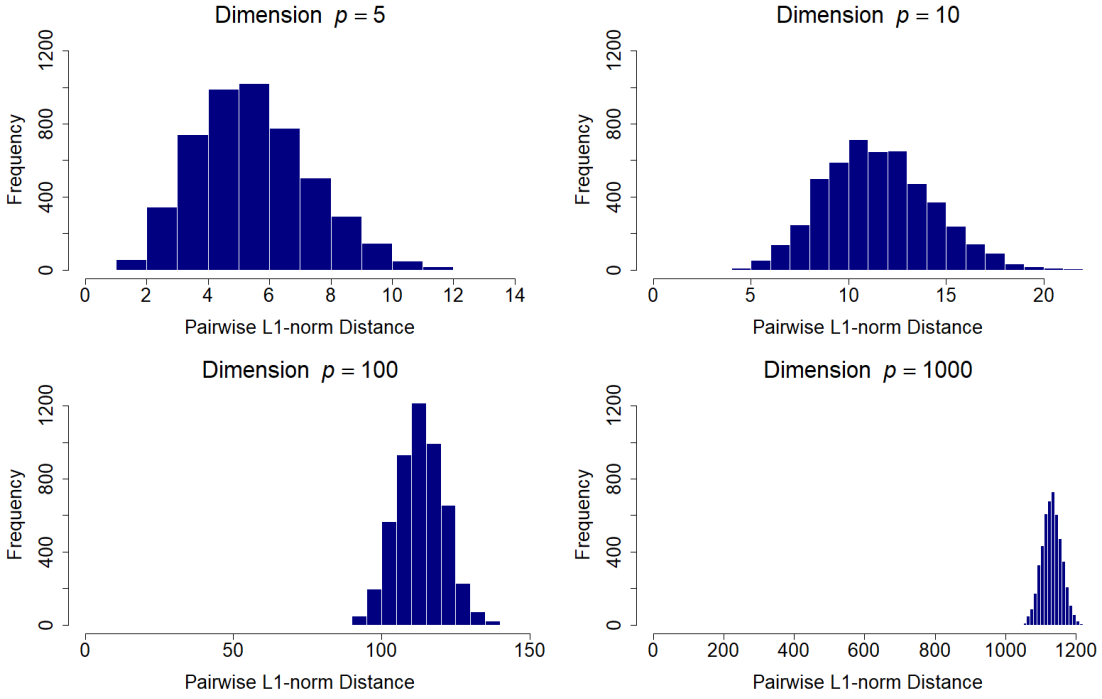
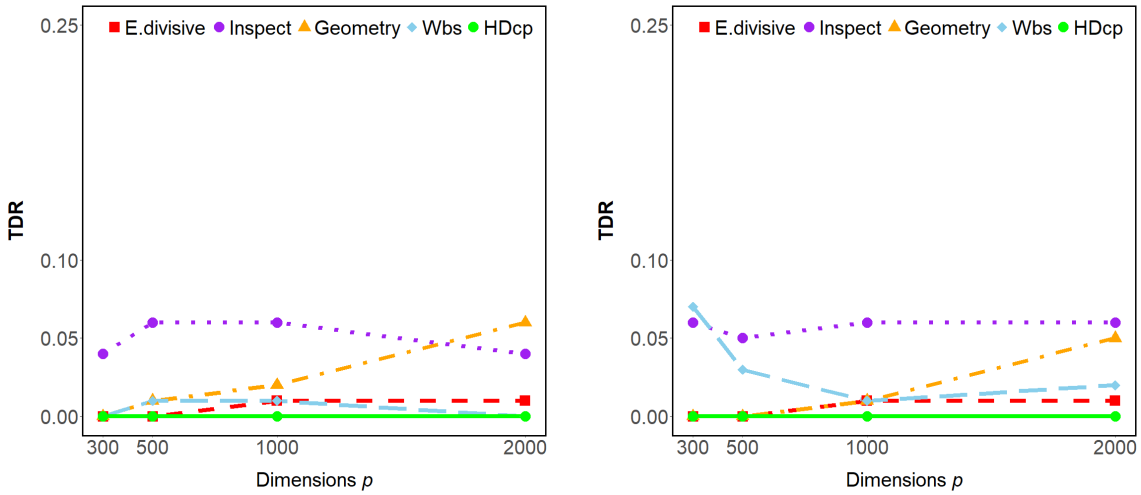


Figure 3.2: Histograms of the pairwise L_1-norm distances among $n = 100$ observations generated from a p-variate standard normal distribution.



(a) Change in the shape of distribution from $N(1,1)$ to Exp(1).

(b) Change in the shape of distribution from $N(1,1)$ to Pois(1).

Figure 3.3: The true detection rate (TDR) over 500 replications for five recent methods in detecting a true change in the shape of distribution while the mean and variance remain the same.

example is when there is a change in the shape of distribution while the mean and variance remain the same. To illustrate this challenging problem, we here assess the performance of five recent methods for high dimensional change points, namely E.divisive by Matteson and James (2014), HDcp by Li et al. (2019a), Inspect by Wang and Samworth (2018), Geometry by Grundy et al. (2020), and WBS by Fryzlewicz (2014) (wbs with observation means), in the two simulation scenarios below.

For the first scenario, we generate 60 observations i.i.d. from a normal distribution $N(1,1)$ and 40 observations i.i.d. from an exponential distribution Exp(1). In the second scenario, we generate 60 observations i.i.d. from $N(1,1)$ and 40 observations i.i.d. from a Poisson distribution Pois(1). In both cases, the distribution changes after location 60, while the mean and variance remain unchanged at 1. Figure 3.3 shows the detection performance of these methods. All these methods perform poorly in detecting such a change in the shape of the distribution.

3.2 AMOC model in the nonparametric setting

Since we want to develop a nonparametric approach, we do not impose any specific assumptions on the underlying distribution of the data sequence. Our objective

is to detect general types of distributional changes rather than a pre-specified one (e.g., a mean shift), which aligns with real-world applications where the nature of the change is often unknown. Accordingly, the AMOC model considered here differs from the parametric model in (2.1). Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a sequence of n random observations, each p-dimensional, with unknown probability distributions $F_1, F_2, \ldots, F_n$, respectively. We here focus on a high dimensional sequence with $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$, $i = 1, \ldots, n$, where $n \ll p$ and furthermore the p variables are potentially correlated. The change point problem can be generally formulated as the following hypothesis test

$$\begin{cases} H_0 : F_1 = F_2 = \cdots = F_n \\ H_1^s : F_1 = \cdots = F_\tau \neq F_{\tau+1} = \cdots = F_n, \end{cases} \tag{3.1}$$

where τ is an unknown change point location, with $\tau \in \{1, \ldots, n-1\}$. If the null hypothesis H_0 in (3.1) is rejected, we infer there is a change in the distribution of observations after location τ. Otherwise, we conclude that there is no significant change point. We will discuss the problem setting of multiple change points given in equation (5.1) in Chapter 5.

3.3 DCCP method and single change point detection algorithm

As discussed in Section 3.1, detecting change points in high dimensional data is challenging, especially in the HDLSS setting. This is because the sample size n is very small compared to the dimension p so there is high dimensional noise to deal with. Instead of searching for the change point location in the $n \times p$ data matrix, an alternative strategy is to work with the $n \times n$ distance matrix of observations, transforming the search to a lower $n \times n$ space as $n \ll p$. Detecting a change point using the distance matrix was recently advocated by Drikvandi and Modarres (2025). They showed that their method is effective for detecting high dimensional non-sparse change points. This approach is unlike the dimensionality reduction techniques such

as random projection (Wang and Samworth, 2018), geometric mapping (Grundy et al., 2020), and principal components (Xiao et al., 2019), as they all directly reduce the dimension of observations. The rationale behind this approach comes from the following theorem in Maa et al. (1996) which shows that, under certain conditions, a change in the distribution of observations can be reflected from a change in the distances between observations, and vice versa.

Theorem 3.3.1. *(Maa et al., 1996) Let S_1 and S_2 be two arbitrary countable sets, and let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \sim F$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots \sim G$ be two independent sequences of p-dimensional random vectors, where F and G are two distributions. If $h(\boldsymbol{X}_i, \boldsymbol{Y}_j)$ is any real-valued non-negative function on $S_1 \times S_2$ such that $h(\boldsymbol{X}_i, \boldsymbol{Y}_j) = 0$ iff $\boldsymbol{X}_i = \boldsymbol{Y}_j$, then*

$$h(\boldsymbol{X}_i, \boldsymbol{X}_j) \overset{D}{=} h(\boldsymbol{Y}_i, \boldsymbol{Y}_j) \overset{D}{=} h(\boldsymbol{X}_i, \boldsymbol{Y}_j) \quad \textit{iff} \quad F = G.$$

The choice of dissimilarity distance can be crucial in high dimensional situations. As previously shown in Section 3.1, the L_2-norm distance $\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2$ suffers from the convergence issue when the dimension p grows. To address this issue, Hall et al. (2005) introduced the modifier $p^{-1/2}$ to the L_2-norm, which results in $p^{-1/2}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2$. This is known as the modified Euclidean distance. Similarly, the modified L_1-norm distance can be defined as $p^{-1}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_1$. To put these together, we define the modified distance $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ as follows

$$d_q(\boldsymbol{X}_i, \boldsymbol{X}_j) = \left(p^{-1}\sum_{l=1}^{p} |X_{il} - X_{jl}|^q\right)^{1/q}, \tag{3.2}$$

where $q = 1, 2$. Note that $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ can be extended to $q > 2$; however, we here establish our methodology with $q = 1, 2$ for the simplicity of exposition. The following theorem shows the modified L_q-norm distance $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ in (3.2) is a metric for $q \geq 1$. The proof, along with our other proofs, is provided in Section 3.6.

Theorem 3.3.2. *Let $q \geq 1$. The modified L_q-norm distance $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ in (3.2) is a metric on the set $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ for $n \geq 3$.*

Both the dissimilarity distances in (3.2) with $q = 1, 2$ satisfy the condition of Theorem 3.3.1, hence they can be used to help discover general changes in the distri-

bution of observations (not just a change in the mean). Let $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]^\top$ represent the entire $n \times p$ data matrix. Using the dissimilarity distance in (3.2), we transform the $n \times p$ data matrix $\boldsymbol{X} = [X_{il}]_{i=1}^{n}{}_{l=1}^{p}$ into the $n \times n$ distance matrix $\boldsymbol{D} = [d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)]_{i=1}^{n}{}_{j=1}^{n}$ with either $q = 1$ or $q = 2$ as follows

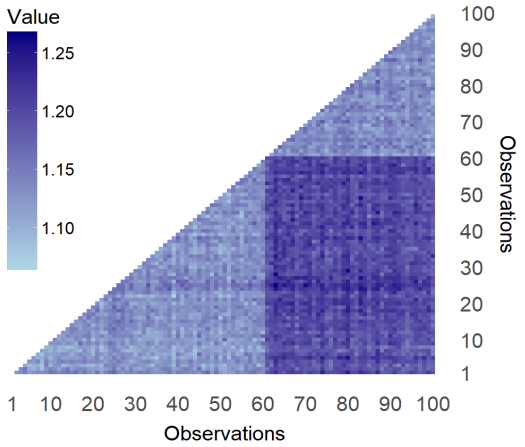
$$\boldsymbol{D} = \begin{bmatrix} d_q(\boldsymbol{X}_1, \boldsymbol{X}_1) & \cdots & d_q(\boldsymbol{X}_1, \boldsymbol{X}_n) \\ d_q(\boldsymbol{X}_2, \boldsymbol{X}_1) & \cdots & d_q(\boldsymbol{X}_2, \boldsymbol{X}_n) \\ \vdots & \ddots & \vdots \\ d_q(\boldsymbol{X}_n, \boldsymbol{X}_1) & \cdots & d_q(\boldsymbol{X}_n, \boldsymbol{X}_n) \end{bmatrix}, \tag{3.3}$$

where $q = 1, 2$.

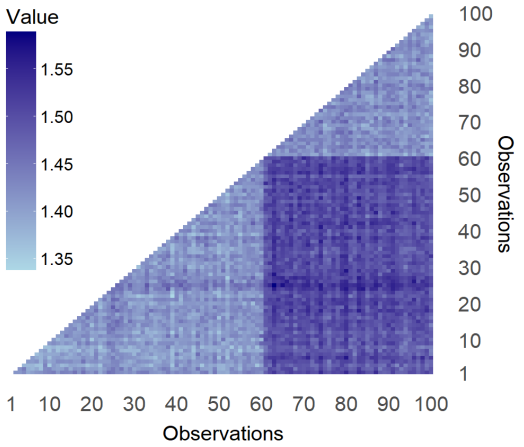
Example 1. To illustrate this process using a simple example with one true change point, we randomly simulate 60 observations from a standard multivariate normal distribution with $p = 1000$ and another 40 observations from the same distribution but with a mean shift of 0.5 for all variables. So there is a true change point at location 60. Figures 3.4a and 3.4b show heatmaps of the lower-triangle of the corresponding distance matrix $\boldsymbol{D}$ with both dissimilarity distances $q = 1$ and $q = 2$. For better visualization, we remove the diagonals from the distance matrix $\boldsymbol{D}$, which are all 0. Moreover, both the modified L_2 norm distance and the modified L_1 norm distance converge to reasonable values even in high dimensional settings ($p = 1000$), in clear contrast to the divergence phenomenon of the classical L_2-norm and L_1-norm shown in Figures 3.1 and 3.2. In Section 3.4, we will show that the distances $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ converge to some finite constants under some conditions. Clearly, the heatmaps show that the observations from different distributions have larger dissimilarity compared to those from the same distribution, which here is due to the change point at location 60.

To capture a systematic trend of change in the values of the distance matrix $\boldsymbol{D}$ due to a potential change point (as in Figures 3.4a and 3.4b), we propose the following distance-based CUSUM statistic for each $i \in \{1, 2, \ldots, n\}$

$$C_i(k) = \frac{\sqrt{k\,(n-k)}}{n}\Big(\frac{1}{n-k}\sum_{j=k+1}^{n} d_q\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) - \frac{1}{k}\sum_{j=1}^{k} d_q\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)\Big), \tag{3.4}$$



(a) Heatmap of the distance matrix $\boldsymbol{D}$ using the modified L_1-norm ($q = 1$).



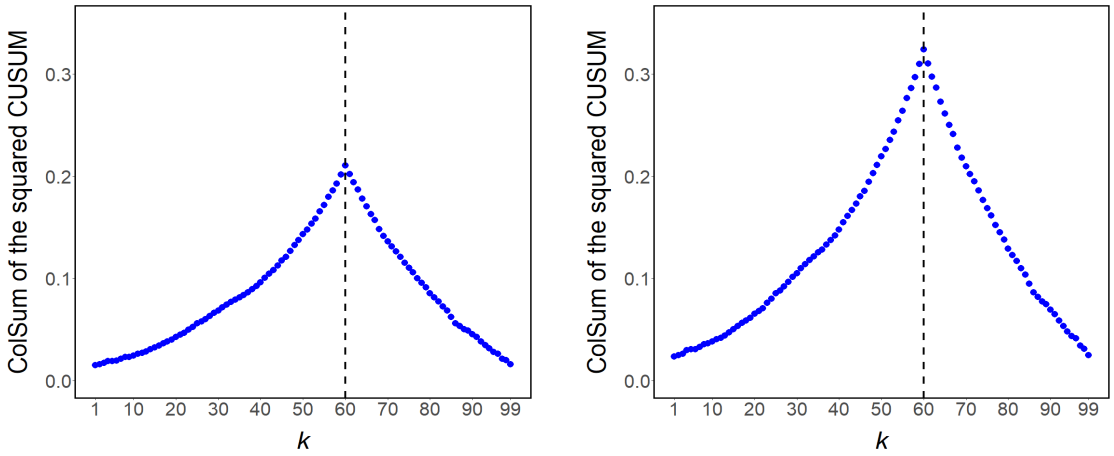
(b) Heatmap of the distance matrix $\boldsymbol{D}$ using the modified L_2-norm ($q = 2$).

Figure 3.4: Illustrative example with a true change point at location $\tau = 60$ with $p = 1000$: figures (a) and (b) show heatmaps of the lower-triangle of symmetric distance matrix $\boldsymbol{D}$ with both $q = 1$ and $q = 2$ respectively.

where k is a candidate search location with $k \in \{1, \ldots, n-1\}$. The distance-based CUSUM statistic (3.4) has two main differences from the standard CUSUM statistic $\boldsymbol{C}(k) = \sqrt{\frac{k(n-k)}{n}} \left(\frac{1}{n-k} \sum_{j=k+1}^{n} \boldsymbol{X}_j - \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{X}_j \right)$. First, it uses the pairwise dissimilarity distances $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ instead of just the observations $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ themselves. Second, the distance-based CUSUM statistic (3.4) is defined for each observation i, $i = 1, \ldots, n$, providing a sequence of n CUSUM values for each $k \in \{1, \ldots, n-1\}$. In fact, the proposed CUSUM statistic (3.4) measures the average distance differences among all observations before and after each candidate search location. The computational cost for calculating the distance-based CUSUM statistic (3.4) is $O(n^2 p)$. However, when n is much smaller than p as in HDLSS settings, the computational cost becomes $O(p)$.

We use the proposed distance-based CUSUM statistics (3.4) to obtain the following $n \times (n-1)$ squared CUSUM matrix

$$\boldsymbol{C} = \begin{bmatrix} C_1^2(1) & \cdots & C_1^2(n-1) \\ C_2^2(1) & \cdots & C_2^2(n-1) \\ \vdots & \ddots & \vdots \\ C_n^2(1) & \cdots & C_n^2(n-1) \end{bmatrix}, \tag{3.5}$$



(a) Column sums (ColSum) of the squared CUSUM matrix $\boldsymbol{C}$ using the modified L_1-norm.

(b) Column sums (ColSum) of the squared CUSUM matrix $\boldsymbol{C}$ using the modified L_2-norm.

Figure 3.5: Illustrative example with a true change point at location $\tau = 60$ with $p = 1000$: figures (a) and (b) visualize column sums of the squared CUSUM matrix $\boldsymbol{C}$, where the true change point location is highlighted by a vertical dashed line.

where all the CUSUM values are squared in order to work with non-negative values. All columns of matrix $\boldsymbol{C}$ in (3.5) contain information about the potential change point location. We suggest estimating the change point location based on the column sums of the squared CUSUM matrix $\boldsymbol{C}$ as follows

$$\hat{\tau} = \underset{1 \leq k \leq n-1}{\arg\max} \Big\{ \frac{1}{n} \sum_{i=1}^{n} {C_i}^2 (k) \Big\}. \tag{3.6}$$

For mathematical reasons, if the values $\frac{1}{n} \sum_{i=1}^{n} {C_i}^2 (k)$ are equal for all $k = 1, \ldots, n-1$, then we set $\hat{\tau} = \emptyset$, where the empty set $\emptyset$ implies no change point is found.

The change point estimate $\hat{\tau}$ in (3.6) is actually the location of the maximum column sums of squared CUSUM matrix $\boldsymbol{C}$. To illustrate this using the previous example, we visualize the column sums of the squared CUSUM matrix $\boldsymbol{C}$ for all $k = 1, \ldots, n-1$ in Figures 3.5a and 3.5b. These plots show that the true change point at location 60 has the largest column sum value. We later prove in Section 3.4 that the change point estimate $\hat{\tau}$ is consistent for the true change point, say τ_0, under some conditions.

To test the statistical significance of change point estimate $\hat{\tau}$ in (3.6), we propose

the following test statistic based on $\hat{\tau}$

$$T(\hat{\tau}) = \frac{1}{n} \sum_{i=1}^{n} C_i^2 \left(\hat{\tau}\right). \tag{3.7}$$

We use a permutation procedure based on the test statistic $T(\hat{\tau})$ to conduct the test of significance. Employing random permutations is helpful here because all the observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are exchangeable under the null hypothesis of no change point in distribution. For this, in each permutation step, we randomly permute the indices of all observations before and after the change point estimate $\hat{\tau}$, while holding the change point location, to get a random permutation sample under H_0. Using formula (3.7), we compute a permutation test statistic based on this permutation sample. Repeating this process S times, we obtain an approximate permutation distribution of the test statistic $T(\hat{\tau})$ as follows

$$G_{T^S_{\text{perm}}}(u) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\left(T^s_{\text{perm}}(\hat{\tau}) \leq u\right) \quad \forall\, u \in \mathbb{R}^+,$$

where $\mathbb{1}(\cdot)$ is the indicator function and $T^s_{\text{perm}}(\hat{\tau})$ denotes the test statistic calculated for the s-th permutation sample. We then calculate the p-value of the permutation test as follows

$$p_{\text{perm}} = 1 - G_{T^S_{\text{perm}}}\left(T_{\text{obs}}(\hat{\tau})\right) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\left(T^s_{\text{perm}}(\hat{\tau}) > T_{\text{obs}}(\hat{\tau})\right), \tag{3.8}$$

where $T_{\text{obs}}(\hat{\tau})$ denotes the test statistic for the observed data. Algorithm 1 summarizes our distance-based CUSUM method for single change point detection.

In addition to the hypothesis test for verifying the estimated change point location, we also construct a confidence interval for the change point location to quantify the uncertainty of estimation. For this, we use a different permutation procedure than the previous permutation procedure conducted under H_0. We here obtain a permutation sample by separately permuting observations before and after the change point estimate $\hat{\tau}$ among themselves. The reason is that the observations before the change point location τ have the same distribution, and the observations

Algorithm 1: DCCP for single change point detection

Input: A data sequence or matrix of observations $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]^\top$.

Output: A significant change point estimate $\hat{\tau}$, or "NA" if there is no significant change point.

Step 1: Calculate the $n \times n$ distance matrix $\boldsymbol{D}$ from the $n \times p$ data matrix $\boldsymbol{X}$.

Step 2: Calculate the squared CUSUM matrix $\boldsymbol{C}$ using the distance matrix $\boldsymbol{D}$.

Step 3: Compute the change point estimate $\hat{\tau}$ in (3.6) and the test statistic $T(\hat{\tau})$ in (3.7).

Step 4: Test the significance of $\hat{\tau}$ using the permutation test.

Step 5: If the p-value p_{perm} in (3.8) is less than α (our default is $\alpha = 0.05$), output $\hat{\tau}$ as a significant change point estimate. Otherwise, output "NA" implying there is no significant change point.

after the change point location τ also have the same distribution. For S random permutations of this, we calculate the change point estimate $\hat{\tau}$ in (3.6) to get S permutation estimates denoted by $\hat{\tau}^* = \left\{\hat{\tau}^{*1}, \hat{\tau}^{*2}, \ldots, \hat{\tau}^{*S}\right\}$. We then construct a $100\,(1-\alpha)\,\%$ confidence interval for change point location τ as follows

$$\left(2\hat{\tau} - \hat{\tau}^*_{(1-\alpha/2)}, 2\hat{\tau} - \hat{\tau}^*_{(\alpha/2)}\right), \tag{3.9}$$

where $\hat{\tau}^*_{(\alpha/2)}$ and $\hat{\tau}^*_{(1-\alpha/2)}$ represent the $(\alpha/2)$-th and $(1-\alpha/2)$-th percentile of $\hat{\tau}^*$ respectively. We later give examples of this change point confidence interval in Section 5.4.

3.4 Asymptotic results

In this section, we study the asymptotic behavior of CUSUM statistics $C_i\,(k)$ when $p > n \to \infty$, which depends on the asymptotic limit of the dissimilarity distance $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$. We also investigate the consistency of the change point estimate $\hat{\tau} = \underset{1\le k\le n-1}{\arg\max}\left\{\frac{1}{n}\sum_{i=1}^{n} {C_i}^2\,(k)\right\}$ when $p > n \to \infty$ and $p \to \infty$ with fixed n. To develop the theory, we here define the following scalar term for $q = 1$ or $q = 2$

$$\lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} := \sqrt{E\left(d_q^q\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)\right)} \qquad \forall\, i \neq j, \tag{3.10}$$

which later appears in the asymptotic limit of both modified L_1-norm and modified L_2-norm distances.

Let τ_0 be the unknown true change point such that $F_1 = \cdots = F_{\tau_0} \neq F_{\tau_0+1} = \cdots = F_n$ with $\tau_0 \in \{1, \ldots, n-1\}$. We define two sets of indices, one for the observations after location τ_0 and one for the observations before and including location τ_0, as follows

$$A^-_{\tau_0} := \{1, \ldots, \tau_0\}, \quad A^+_{\tau_0} := \{\tau_0 + 1, \ldots, n\}.$$

We can then write for $i \neq j$ that

$$\begin{aligned}
\lambda_{A^-_{\tau_0} A^-_{\tau_0}} &:= \lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} \quad \forall\, i, j \in A^-_{\tau_0}, \\
\lambda_{A^+_{\tau_0} A^+_{\tau_0}} &:= \lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} \quad \forall\, i, j \in A^+_{\tau_0}, \\
\lambda_{A^-_{\tau_0} A^+_{\tau_0}} &:= \lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} \quad \forall\, i \in A^-_{\tau_0}, j \in A^+_{\tau_0} \text{ or } \forall\, i \in A^+_{\tau_0}, j \in A^-_{\tau_0}.
\end{aligned}$$

The following lemma presents some Taylor expansion formulas for $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)$ and $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$, which we will use in our subsequent theoretical results.

Lemma 3.4.1. *Suppose that τ_0 is a true change point location such that $F_1 = \cdots = F_{\tau_0} \neq F_{\tau_0+1} = \cdots = F_n$ with $\tau_0 \in \{1, \ldots, n-1\}$. Consider $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ defined in* (3.2). *We have for $q = 1, 2$*

$$\begin{aligned}
\frac{d_q^{q/2}(\boldsymbol{X}_i, \boldsymbol{X}_j)}{\lambda_{A^-_{\tau_0} A^-_{\tau_0}}} &= 1 + L_{A^-_{\tau_0} A^-_{\tau_0}} + R_{A^-_{\tau_0} A^-_{\tau_0}} \qquad \forall\, i, j \in A^-_{\tau_0}, \\
\frac{d_q^{q/2}(\boldsymbol{X}_i, \boldsymbol{X}_j)}{\lambda_{A^+_{\tau_0} A^+_{\tau_0}}} &= 1 + L_{A^+_{\tau_0} A^+_{\tau_0}} + R_{A^+_{\tau_0} A^+_{\tau_0}} \qquad \forall\, i, j \in A^+_{\tau_0}, \\
\frac{d_q^{q/2}(\boldsymbol{X}_i, \boldsymbol{X}_j)}{\lambda_{A^-_{\tau_0} A^+_{\tau_0}}} &= 1 + L_{A^-_{\tau_0} A^+_{\tau_0}} + R_{A^-_{\tau_0} A^+_{\tau_0}} \quad \forall\, i \in A^-_{\tau_0}, j \in A^+_{\tau_0} \text{ or } \forall\, i \in A^+_{\tau_0}, j \in A^-_{\tau_0},
\end{aligned}$$

where we have $L_{A^-_{\tau_0} A^-_{\tau_0}} = \frac{d_q^q(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^-_{\tau_0} A^-_{\tau_0}}}{2\lambda^2_{A^-_{\tau_0} A^-_{\tau_0}}}$, $L_{A^+_{\tau_0} A^+_{\tau_0}} = \frac{d_q^q(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^+_{\tau_0} A^+_{\tau_0}}}{2\lambda^2_{A^+_{\tau_0} A^+_{\tau_0}}}$ and $L_{A^-_{\tau_0} A^+_{\tau_0}} = \frac{d_q^q(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^-_{\tau_0} A^+_{\tau_0}}}{2\lambda^2_{A^-_{\tau_0} A^+_{\tau_0}}}$ are leading terms of the expansions, while $R_{A^-_{\tau_0} A^-_{\tau_0}} = O\big(L^2_{A^-_{\tau_0} A^-_{\tau_0}}\big)$, $R_{A^+_{\tau_0} A^+_{\tau_0}} = O\big(L^2_{A^+_{\tau_0} A^+_{\tau_0}}\big)$ and $R_{A^-_{\tau_0} A^+_{\tau_0}} = O\big(L^2_{A^-_{\tau_0} A^+_{\tau_0}}\big)$ are remainder terms.

To derive the asymptotic limit with the modified L_2-norm distance $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$,

we make the following assumptions on observations $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$, similar to Hall et al. (2005):

(A1) Assume $\max_{1\le i\le n} \max_{1\le l\le p} E(X_{il}^4) < \infty$.

(A2) Assume $\sum_{l=1}^{p} \sum_{l'=1}^{p} \mathrm{Cov}\big((X_{il} - X_{jl})^2, (X_{il'} - X_{jl'})^2\big) = o(p^2)$ as $p \to \infty$, where $l \neq l'$.

Alternatively, for the modified L_1-norm distance $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)$, we make the following assumptions:

(B1) Assume $\max_{1\le i\le n} \max_{1\le l\le p} E(X_{il}^2) < \infty$.

(B2) Assume $\sum_{l=1}^{p} \sum_{l'=1}^{p} \mathrm{Cov}\left(|X_{il} - X_{jl}|, |X_{il'} - X_{jl'}|\right) = o(p^2)$ as $p \to \infty$, where $l \neq l'$.

With these assumptions, it is clear that $\lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} < \infty$ for $1 \le i, j \le n$. Assumption (A1) says that the fourth moment of random variables X_{il} is uniformly bounded. Assumption (A2) imposes the weak dependence among the random variables, which is trivial if the variables are independent. Hall et al. (2005) made assumptions (A1) and (A2) to prove the high dimensional convergence of the modified L_2-norm distance, and these assumptions are also used for change point analysis by Biswas et al. (2014), Li (2020), and Drikvandi and Modarres (2025), among some others. Note that under assumptions (A1) and (A2), the weak law of large numbers (WLLN) holds on the sequence $\{(X_{il} - X_{jl})^2 : 1 \le l \le p\}$. Hence we have $p^{-1}\left(\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2^2 - E\left(\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2^2\right)\right) \xrightarrow{P} 0$ as $p \to \infty$. Similarly, for the modified L_1-norm distance, assumptions (B1) and (B2) ensure $p^{-1}\left(\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_1 - E\left(\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_1\right)\right) \xrightarrow{P} 0$ as $p \to \infty$. We note that the assumptions for the modified L_2-norm distance are stronger than those for the modified L_1-norm distance.

The following theorem provides the asymptotic convergence of both the modified L_1-norm and the modified L_2-norm as $p \to \infty$. Note that none of the asymptotic limits are exactly 0 even when observations $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ have the same distribution; in fact, the asymptotic limits depend on the scalar term defined in (3.10).

Theorem 3.4.2. *Consider the conditions in Lemma 3.4.1 and the distance function* $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ *defined in* (3.2)*. Under assumptions (A1)-(A2) or (B1)-(B2), as* $p \to \infty$ *we have for* $q = 1, 2$

$$d_q(\boldsymbol{X}_i, \boldsymbol{X}_j) = \begin{cases} \lambda^{2/q}_{A^-_{\tau_0} A^-_{\tau_0}} + o_P(1) & \forall\, i, j \in A^-_{\tau_0}, \\ \lambda^{2/q}_{A^+_{\tau_0} A^+_{\tau_0}} + o_P(1) & \forall\, i, j \in A^+_{\tau_0}, \\ \lambda^{2/q}_{A^-_{\tau_0} A^+_{\tau_0}} + o_P(1) & \forall\, i \in A^-_{\tau_0}, j \in A^+_{\tau_0} \text{ or } \forall\, i \in A^+_{\tau_0}, j \in A^-_{\tau_0}. \end{cases}$$

Remark 1. For the modified L_2-norm $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$, Hall et al. (2005) additionally assume that $p^{-1/2} \|E(\boldsymbol{X}_i) - E(\boldsymbol{X}_j)\|_2^2 \xrightarrow{P} \eta^2_{ij}$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/p \xrightarrow{P} \sigma_i^2$ as $p \to \infty$, where σ_i^2 and η^2_{ij} are finite constants. With this, we can write

$$\begin{aligned} \sigma_i^2 = \sigma_j^2 := \sigma^2_{A^-_{\tau_0}}, \quad \eta^2_{ij} := \eta^2_{A^-_{\tau_0} A^-_{\tau_0}}, \qquad & \forall\, i, j \in A^-_{\tau_0}, \\ \sigma_i^2 = \sigma_j^2 := \sigma^2_{A^+_{\tau_0}}, \quad \eta^2_{ij} := \eta^2_{A^+_{\tau_0} A^+_{\tau_0}}, \qquad & \forall\, i, j \in A^+_{\tau_0}, \\ \sigma_i^2 := \sigma^2_{A^-_{\tau_0}}, \sigma_j^2 := \sigma^2_{A^+_{\tau_0}}, \quad \eta^2_{ij} := \eta^2_{A^-_{\tau_0} A^+_{\tau_0}}, \qquad & \forall\, i \in A^-_{\tau_0}, j \in A^+_{\tau_0}, \\ \sigma_i^2 := \sigma^2_{A^+_{\tau_0}}, \sigma_j^2 := \sigma^2_{A^-_{\tau_0}}, \quad \eta^2_{ij} := \eta^2_{A^+_{\tau_0} A^-_{\tau_0}}, \qquad & \forall\, i \in A^+_{\tau_0}, j \in A^-_{\tau_0}, \end{aligned}$$

where $\eta^2_{A^-_{\tau_0} A^+_{\tau_0}} = \eta^2_{A^+_{\tau_0} A^-_{\tau_0}}$. Hence for $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$, using Theorem 3.4.2 for $q = 2$, we have as $p \to \infty$

$$\begin{aligned} \lambda_{A^-_{\tau_0} A^-_{\tau_0}} &= \left(2\sigma^2_{A^-_{\tau_0}}\right)^{1/2}, \qquad \lambda_{A^+_{\tau_0} A^+_{\tau_0}} = \left(2\sigma^2_{A^+_{\tau_0}}\right)^{1/2}, \\ \lambda_{A^-_{\tau_0} A^+_{\tau_0}} &= \lambda_{A^+_{\tau_0} A^-_{\tau_0}} = \left(\sigma^2_{A^-_{\tau_0}} + \sigma^2_{A^+_{\tau_0}} + \eta^2_{A^-_{\tau_0} A^+_{\tau_0}}\right)^{1/2}. \end{aligned}$$

We note that $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)$ does not exhibit a similar property due to the absolute values in the modified L_1-norm distance.

Remark 2. In theory, the choice between $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)$ and $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$ depends on their asymptotic limits. The above remark states that $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$ can be expressed in terms of the mean and variance of observations only, and therefore $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$ can mainly detect changes in mean or variance of observations. However, the asymptotic behavior of the (modified) L_1-norm distance is more sophisticated and general. It

has been shown to be more advantageous than the (modified) L_2-norm distance for detecting marginal distributions in high dimensional settings (see Zhu and Shao, 2021), and it can detect any changes in the marginal distribution of data under the above assumptions. In Subsection 4.3, we will show via simulation studies that our CUSUM statistic based on the modified L_1-norm distance can detect a change in the marginal distribution while the mean and variance remain unchanged. As previously shown in Section 3.1, this is a challenging problem for which many existing methods struggle to work.

We now provide the asymptotic limit of the distance-based CUSUM statistics (3.4) when $p > n \to \infty$.

Proposition 3.4.3. *Consider the conditions and assumptions in Theorem 3.4.2. As $p > n \to \infty$, we have for $q = 1, 2$ that*

$$C_i(k) = \begin{cases} \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\left(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^{2/q}_{A^-_{\tau_0}A^-_{\tau_0}}\right) + o_P(1) & \forall\, i \in A^-_{\tau_0}, \forall\, k \in A^-_{\tau_0} \setminus \{\tau_0\}, \\ \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\left(\lambda^{2/q}_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}\right) + o_P(1) & \forall\, i \in A^+_{\tau_0}, \forall\, k \in A^-_{\tau_0} \setminus \{\tau_0\}, \\ \frac{\sqrt{\tau_0(n-\tau_0)}}{n}\left(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^{2/q}_{A^-_{\tau_0}A^-_{\tau_0}}\right) + o_P(1) & \forall\, i \in A^-_{\tau_0} \text{ and } k = \tau_0, \\ \frac{\sqrt{\tau_0(n-\tau_0)}}{n}\left(\lambda^{2/q}_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}\right) + o_P(1) & \forall\, i \in A^+_{\tau_0} \text{ and } k = \tau_0, \\ \frac{\tau_0}{n}\sqrt{\frac{n-k}{k}}\left(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^{2/q}_{A^-_{\tau_0}A^-_{\tau_0}}\right) + o_P(1) & \forall\, i \in A^-_{\tau_0}, \forall\, k \in A^+_{\tau_0} \setminus \{n\}, \\ \frac{\tau_0}{n}\sqrt{\frac{n-k}{k}}\left(\lambda^{2/q}_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}\right) + o_P(1) & \forall\, i \in A^+_{\tau_0}, \forall\, k \in A^+_{\tau_0} \setminus \{n\}. \end{cases}$$

Under the null hypothesis H_0 of no change point, it follows from the results of Proposition 3.4.3 that $C_i(k) = o_P(1)$ for all $i = 1, \ldots, n$ and $k = 1, \ldots, n-1$ as $p > n \to \infty$. This is because $\lambda_{A^-_{\tau_0}A^-_{\tau_0}} = \lambda_{A^+_{\tau_0}A^+_{\tau_0}} = \lambda_{A^-_{\tau_0}A^+_{\tau_0}}$ under H_0. The following theorem proves the consistency of our CUSUM test for single change point detection when $p > n \to \infty$, using either of the modified L_1-norm and L_2-norm distances.

Theorem 3.4.4. *Consider the conditions in Lemma 3.4.1 and the distance function $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ defined in* (3.2)*. Under assumptions (A1) - (A2) or (B1) - (B2), we have as $p > n \to \infty$ that*

(a)

$$\max_{1\leq k\leq n-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}C_i^2(k)-\phi(k)\mathbb{1}(k=\tau_0)\Big\}=o_P(1),$$

in which

$$\phi(k)=\begin{cases}\frac{(n-\tau_0)^2\tau_0 k}{n^3(n-k)}\big(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}-\lambda^{2/q}_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2+\frac{(n-\tau_0)^3 k}{n^3(n-k)}\big(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}-\lambda^{2/q}_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 & \forall\, k\in A^-_{\tau_0}\setminus\{\tau_0\},\\ \frac{{\tau_0}^2(n-\tau_0)}{n^3}\big(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}-\lambda^{2/q}_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2+\frac{\tau_0(n-\tau_0)^2}{n^3}\big(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}-\lambda^{2/q}_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 & k=\tau_0,\\ \frac{{\tau_0}^3(n-k)}{n^3 k}\big(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}-\lambda^{2/q}_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2+\frac{{\tau_0}^2(n-\tau_0)(n-k)}{n^3 k}\big(\lambda^{2/q}_{A^-_{\tau_0}A^+_{\tau_0}}-\lambda^{2/q}_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 & \forall k\in A^+_{\tau_0}\setminus\{n\}.\end{cases}$$

(b) the change point estimate $\hat{\tau}=\underset{1\leq k\leq n-1}{\arg\max}\big\{\frac{1}{n}\sum_{i=1}^{n}C_i^2(k)\big\}$ *is consistent for* τ_0 *when* $p>n\to\infty$ *as*

$$\hat{\tau}-\tau_0=o_P(1).$$

Theorem 3.4.4 shows our change point estimate $\hat{\tau}$ is consistent with the true change point τ_0 under the alternative hypothesis when $p>n\to\infty$. Does this result hold for $p\to\infty$ but with fixed n, as in HDLSS data? The following result answers this question.

Theorem 3.4.5. *Consider the conditions in Lemma 3.4.1 and the distance function* $d_q(\boldsymbol{X}_i,\boldsymbol{X}_j)$ *defined in* (3.2)*. Under assumptions (A1) - (A2) or (B1) - (B2), we have as* $p\to\infty$ *that*

$$\max_{1\leq k\leq n-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}C_i^2(k)\Big\}=\phi(k)\mathbb{1}(k=\tau_0)+o_P(1),$$

where $\phi(k)$ *is given in Theorem 3.4.4.*

Remark 3. The results in Theorems 3.4.4 and 3.4.5 can be extended to distance-based CUSUM using other dissimilarity measures. This raises the question of how to choose an appropriate distance for high dimensional data. We recommend selecting a distance that converges in the high dimensional regime ($p\to\infty$) and exhibits distinct asymptotic limits for pairs of observations drawn from the same distribution versus different distributions, as established in Theorem 3.4.2.

3.5 Concluding remarks

In this chapter, we have discussed the main challenges of high dimensional change point analysis. To address these challenges, we have proposed distance-based CUSUM statistics for detecting a single change point in high dimensional observations, which do not require normality or any other distribution for the data. Under the alternative hypothesis, we have studied the asymptotic properties of the proposed distance-based CUSUM statistic as $p > n \to \infty$. Furthermore, we have proved that our change point estimate is consistent with the true change point under suitable conditions. A main advantage of the distance-based CUSUM statistics is their capability in detecting more general types of changes, including linear and non-linear changes such as a change in the mean, variance, correlation, and other distributional changes. Also, due to the nature of the distance functions used, our distance-based CUSUM method is particularly suitable for HDLSS data when the sample size is very small compared to the dimension. In the next chapter, we will demonstrate these advantages via extensive simulation studies.

3.6 Proofs

Proof of Theorem 3.3.2

Consider the definition of $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ in (3.2). We can easily observe $d_q(\boldsymbol{X}_i, \boldsymbol{X}_i) = 0$, $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$ and $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j) = d_q(\boldsymbol{X}_j, \boldsymbol{X}_i)$. To prove the triangle inequality of $d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)$ for observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ with $n \geq 3$, we have

$$
\begin{aligned}
d_q(\boldsymbol{X}_i, \boldsymbol{X}_j) &= \Big(p^{-1} \sum_{l=1}^{p} |X_{il} - X_{jl}|^q \Big)^{1/q} \\
&= p^{-1/q} \Big(\sum_{l=1}^{p} |X_{il} - X_{jl}|^q \Big)^{1/q} \\
&= p^{-1/q} \Big(\sum_{l=1}^{p} |X_{il} - X_{kl} + X_{kl} - X_{jl}|^q \Big)^{1/q} \\
&\leq p^{-1/q} \bigg(\Big(\sum_{l=1}^{p} |X_{il} - X_{kl}|^q \Big)^{1/q} + \Big(\sum_{l=1}^{p} |X_{kl} - X_{jl}|^q \Big)^{1/q} \bigg)
\end{aligned}
$$

$$= p^{-1/q}\Big(\sum_{l=1}^{p}|X_{il}-X_{kl}|^q\Big)^{1/q} + p^{-1/q}\Big(\sum_{l=1}^{p}|X_{kl}-X_{jl}|^q\Big)^{1/q}$$
$$= d_q\left(\boldsymbol{X}_i,\boldsymbol{X}_k\right) + d_q\left(\boldsymbol{X}_k,\boldsymbol{X}_j\right).$$

Hence, $d_q\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right) \leq d_q\left(\boldsymbol{X}_i,\boldsymbol{X}_k\right) + d_q\left(\boldsymbol{X}_k,\boldsymbol{X}_j\right)$ for $n \geq 3$. This completes the proof.

Proof of Lemma 3.4.1

First, applying a Taylor expansion for the function $f(t) = (1+t)^{1/2}$ around $t = 0$, we can write

$$f(t) = 1 + \frac{1}{2}t + R(t), \tag{3.11}$$

where $R(t)$ is a remainder term. For the case of $q = 1$, let $t\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right) = \big(d_1\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right) - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)/\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}$ for all $i,j \in A^-_{\tau_0}$ and $i \neq j$. Using (3.11), we obtain for all $i,j \in A^-_{\tau_0}$ and $i \neq j$ that

$$\frac{d_1^{1/2}\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)}{\lambda_{A^-_{\tau_0}A^-_{\tau_0}}} = 1 + L_{A^-_{\tau_0}A^-_{\tau_0}} + R_{A^-_{\tau_0}A^-_{\tau_0}}, \tag{3.12}$$

where $L_{A^-_{\tau_0}A^-_{\tau_0}} = (d_1\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right) - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}})/2\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}$ is the leading term of the expansion, and $R_{A^-_{\tau_0}A^-_{\tau_0}}$ is the remainder term. To understand the asymptotic behavior of the remainder term, we use the remainder estimation theorem (see Apostol, 1967, Section 7.9) and the fact that $f(t) = (1+t)^{1/2}$ is twice differentiable on some interval containing $t = 0$. This means there exists a positive constant M such that $|f''(t)| \leq M$, and therefore the remainder term satisfies

$$\left|R_{A^-_{\tau_0}A^-_{\tau_0}}\right| \leq \frac{M|t|^2}{2} = \frac{Mt^2}{2}.$$

Hence, we have $R_{A^-_{\tau_0}A^-_{\tau_0}} = O\big(L^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)$.

Similarly, using (3.11) and letting $t\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right) = \big(d_1\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right) - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)/\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}$, we find for all $i,j \in A^+_{\tau_0}$ and $i \neq j$ that

$$\frac{d_1^{1/2}\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)}{\lambda_{A^+_{\tau_0}A^+_{\tau_0}}} = 1 + L_{A^+_{\tau_0}A^+_{\tau_0}} + R_{A^+_{\tau_0}A^+_{\tau_0}},$$

where $L_{A^+_{\tau_0}A^+_{\tau_0}} = (d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}})/2\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}$ is the leading term and $R_{A^+_{\tau_0}A^+_{\tau_0}} = O\big(L^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)$ is the remainder term.

Similarly, using (3.11) and letting $t(\boldsymbol{X}_i, \boldsymbol{X}_j) = \big(d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}\big)/\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}$, we obtain for all $i \in A^-_{\tau_0}, j \in A^+_{\tau_0}$ and $i \in A^+_{\tau_0}, j \in A^-_{\tau_0}$ that (with $i \neq j$)

$$\frac{d_1^{1/2}(\boldsymbol{X}_i, \boldsymbol{X}_j)}{\lambda_{A^-_{\tau_0}A^+_{\tau_0}}} = 1 + L_{A^-_{\tau_0}A^+_{\tau_0}} + R_{A^-_{\tau_0}A^+_{\tau_0}},$$

where $L_{A^-_{\tau_0}A^+_{\tau_0}} = (d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}})/2\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}$ is the leading term and $R_{A^-_{\tau_0}A^+_{\tau_0}} = O\big(L^2_{A^-_{\tau_0}A^+_{\tau_0}}\big)$ is the remainder term.

The proof with $q = 2$ is similar. Let $t(\boldsymbol{X}_i, \boldsymbol{X}_j) = \big(d_2^2(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)/\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}$ and using a similar Taylor expansion as above, we first expand $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)/\lambda_{A^-_{\tau_0}A^-_{\tau_0}}$ in the same manner. Repeating the same Taylor expansion for $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)/\lambda_{A^+_{\tau_0}A^+_{\tau_0}}$ and $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)/\lambda_{A^-_{\tau_0}A^+_{\tau_0}}$, we then get the result of the lemma for the case of $q = 2$. We note that the idea of applying Taylor expansion to the same function $f(t) = (1+t)^{1/2}$ has been used in Zhu et al. (2020). However, our result has two main differences from their paper: (i) our Lemma 3.4.1 applies to both the modified L_1-norm distance and modified L_2-norm distance, whereas their result deals with the L_2 norm distance; and (ii) our method applies Taylor expansion to the modified version of the L_1 norm and L_2 norm distances while their paper considers L_2-norm without modification for growing p.

Proof of Theorem 3.4.2

We first give the proof with the L_1-norm distance $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)$. Using Lemma 3.4.1, for all $i, j \in A^-_{\tau_0}$ and $i \neq j$, we get from (3.12) that

$$d_1^{1/2}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \lambda_{A^-_{\tau_0}A^-_{\tau_0}} + \lambda_{A^-_{\tau_0}A^-_{\tau_0}} L_{A^-_{\tau_0}A^-_{\tau_0}} + \lambda_{A^-_{\tau_0}A^-_{\tau_0}} R_{A^-_{\tau_0}A^-_{\tau_0}}, \tag{3.13}$$

where $L_{A^-_{\tau_0}A^-_{\tau_0}} = (d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}})/2\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}$ is the leading term, and $R_{A^-_{\tau_0}A^-_{\tau_0}} = O\big(L^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)$ is the remainder term. Under Assumptions (B1) – (B2) and using the weak law of large numbers, we have $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} = o_P(1)$ as $p \to \infty$. Hence, since $\lambda_{A^-_{\tau_0}A^-_{\tau_0}} < \infty$ under Assumptions (B1) – (B2), we have $L_{A^-_{\tau_0}A^-_{\tau_0}} = o_P(1)$ as

$p \to \infty$ and consequently $\lambda_{A^-_{\tau_0} A^-_{\tau_0}} L_{A^-_{\tau_0} A^-_{\tau_0}} = o_P(1)$ as $p \to \infty$. Plugging this in (3.13), it then follows that

$$d_1^{1/2}\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = \lambda_{A^-_{\tau_0} A^-_{\tau_0}} + \lambda_{A^-_{\tau_0} A^-_{\tau_0}} R_{A^-_{\tau_0} A^-_{\tau_0}} + o_P(1), \tag{3.14}$$

as $p \to \infty$. Since $L_{A^-_{\tau_0} A^-_{\tau_0}} = o_P(1)$ as $p \to \infty$, we also have $R_{A^-_{\tau_0} A^-_{\tau_0}} = O\big(L^2_{A^-_{\tau_0} A^-_{\tau_0}}\big) = o_P(1)$ as $p \to \infty$. This implies $\lambda_{A^-_{\tau_0} A^-_{\tau_0}} R_{A^-_{\tau_0} A^-_{\tau_0}} = o_P(1)$ as $p \to \infty$ because $\lambda_{A^-_{\tau_0} A^-_{\tau_0}} < \infty$ under Assumptions (B1) – (B2). Inserting this in (3.14), we get the asymptotic limit of $d_1^{1/2}\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)$ as follows

$$d_1^{1/2}\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = \lambda_{A^-_{\tau_0} A^-_{\tau_0}} + o_P(1),$$

as $p \to \infty$. Using the continuous mapping theorem and that $g(u) = u^2$ is a continuous function, we then have

$$d_1\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = \lambda^2_{A^-_{\tau_0} A^-_{\tau_0}} + o_P(1),$$

as $p \to \infty$. We apply the same procedure as above for all $i, j \in A^+_{\tau_0}$, as well as for all $i \in A^-_{\tau_0}, j \in A^+_{\tau_0}$ and $i \in A^+_{\tau_0}, j \in A^-_{\tau_0}$ to find

$$d_1\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = \begin{cases} \lambda^2_{A^-_{\tau_0} A^-_{\tau_0}} + o_P(1) & \forall\, i, j \in A^-_{\tau_0}, \\ \lambda^2_{A^+_{\tau_0} A^+_{\tau_0}} + o_P(1) & \forall\, i, j \in A^+_{\tau_0}, \\ \lambda^2_{A^-_{\tau_0} A^+_{\tau_0}} + o_P(1) & \forall\, i \in A^-_{\tau_0}, j \in A^+_{\tau_0} \text{ or } \forall\, i \in A^+_{\tau_0}, j \in A^-_{\tau_0}, \end{cases}$$

as $p \to \infty$. This completes the proof for the case of $q = 1$.

The proof with the L_2-norm distance $d_2\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)$ is very similar, so we skip it here. We just note that for the case of $q = 2$, under Assumptions (A1) - (A2), the asymptotic convergence of $d_2\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)$ is obtained in the same manner but without applying the continuous mapping theorem in the final step, which makes the proof for $q = 2$ slightly simpler.

Proof of Proposition 3.4.3

We first give the proof for the case of $q = 1$ which is with the L_1-norm distance $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)$ as $p > n \to \infty$ under Assumptions (B1) - (B2). Considering the results of Theorem 3.4.2, the asymptotic limit of $C_i(k)$ in (3.4) obviously depends on the relation between i, k, and τ_0. We recall that $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, n-1\}$ and $\tau_0 \in \{1, \ldots, n-1\}$. We here use the indicator function $\mathbb{1}(\cdot)$ to clarify the different cases of this in our proof. Obviously, $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if $i = j$. First, for all $i \in A^-_{\tau_0}, k \in A^-_{\tau_0} \setminus \{\tau_0\}$, we get as $p > n \to \infty$

$$
\begin{aligned}
&C_i(k)\,\mathbb{1}(i \le k < \tau_0) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{1}{n-k}\sum_{j=k+1}^{n} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k}\sum_{j=1}^{k} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)\Big) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{1}{n-k}\sum_{j=\tau_0+1}^{n} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) + \frac{1}{n-k}\sum_{j=k+1}^{\tau_0} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k}\sum_{j=1}^{k} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)\Big) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{n-\tau_0}{n-k}\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} + \frac{\tau_0-k}{n-k}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} - \frac{k-1}{k}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)\Big) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{n-\tau_0}{n-k}\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} + \frac{\tau_0-k}{n-k}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\Big) + \sqrt{\frac{n-k}{n^2k}}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{n-\tau_0}{n-k}\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \frac{n-\tau_0}{n-k}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\Big) + \sqrt{\frac{1}{nk} - \frac{1}{n^2}}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1) \\
&= \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o(1) + o_P(1),
\end{aligned}
$$

where we note that $\sqrt{\frac{1}{nk} - \frac{1}{n^2}}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}$ is negligible when n grows to infinity. We similarly get as $p > n \to \infty$

$$
\begin{aligned}
&C_i(k)\,\mathbb{1}(k < i \le \tau_0) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{1}{n-k}\sum_{j=k+1}^{n} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k}\sum_{j=1}^{k} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)\Big) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{1}{n-k}\sum_{j=\tau_0+1}^{n} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) + \frac{1}{n-k}\sum_{j=k+1}^{\tau_0} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k}\sum_{j=1}^{k} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)\Big) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{n-\tau_0}{n-k}\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} + \frac{\tau_0-k-1}{n-k}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)\Big) \\
&= \frac{\sqrt{k(n-k)}}{n}\Big(\frac{n-\tau_0}{n-k}\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} + \frac{\tau_0-k}{n-k}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\Big) - \sqrt{\frac{k}{n^2(n-k)}}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)
\end{aligned}
$$

$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{n-\tau_0}{n-k}\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \frac{n-\tau_0}{n-k}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) - \sqrt{\frac{1}{n(n-k)} - \frac{1}{n^2}}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)$$

$$= \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o(1) + o_P(1).$$

Hence, for all $i \in A^-_{\tau_0}, k \in A^-_{\tau_0} \setminus \{\tau_0\}$, we have as $p > n \to \infty$

$$C_i\,(k) = \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1).$$

Repeating the same calculations as above for all $i \in \{1,\ldots,n\}$ and $k \in \{1,\ldots,n-1\}$, it is straightforward to obtain the asymptotic limit of $C_i\,(k)$ when $p > n \to \infty$ as follows

$$C_i(k) = \begin{cases} \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^-_{\tau_0}, \forall\, k \in A^-_{\tau_0} \setminus \{\tau_0\}, \\ \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^+_{\tau_0}, \forall\, k \in A^-_{\tau_0} \setminus \{\tau_0\}, \\ \frac{\sqrt{\tau_0(n-\tau_0)}}{n}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^-_{\tau_0} \text{ and } k = \tau_0, \\ \frac{\sqrt{\tau_0(n-\tau_0)}}{n}\big(\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^+_{\tau_0} \text{ and } k = \tau_0, \\ \frac{\tau_0}{n}\sqrt{\frac{n-k}{k}}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^-_{\tau_0}, \forall\, k \in A^+_{\tau_0} \setminus \{n\}, \\ \frac{\tau_0}{n}\sqrt{\frac{n-k}{k}}\big(\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^+_{\tau_0}, \forall\, k \in A^+_{\tau_0} \setminus \{n\}. \end{cases} \tag{3.15}$$

This completes the proof of the proposition for $q = 1$.

We then provide the proof for the case of $q = 2$, which is with the L_2-norm distance $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$, although the proof is similar to the case of $q = 1$. For this, under Assumptions (A1) - (A2), we have for all $i \in A^-_{\tau_0}$, $k \in A^-_{\tau_0} \setminus \{\tau_0\}$ that as $p > n \to \infty$

$$C_i\,(k)\,\mathbb{1}(i \le k < \tau_0)$$

$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{1}{n-k}\sum_{j=k+1}^{n} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k}\sum_{j=1}^{k} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j)\big)$$

$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{1}{n-k}\sum_{j=\tau_0+1}^{n} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j) + \frac{1}{n-k}\sum_{j=k+1}^{\tau_0} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k}\sum_{j=1}^{k} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j)\big)$$

$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{n-\tau_0}{n-k}\lambda_{A^-_{\tau_0}A^+_{\tau_0}} + \frac{\tau_0-k}{n-k}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} - \frac{k-1}{k}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)\big)$$

$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{n-\tau_0}{n-k}\lambda_{A^-_{\tau_0}A^+_{\tau_0}} + \frac{\tau_0-k}{n-k}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) + \sqrt{\frac{n-k}{n^2k}}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)$$

$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{n-\tau_0}{n-k}\lambda_{A^-_{\tau_0}A^+_{\tau_0}} - \frac{n-\tau_0}{n-k}\lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) + \sqrt{\frac{1}{nk}-\frac{1}{n^2}}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)$$
$$= \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o(1) + o_P(1),$$

where $\sqrt{\frac{1}{nk}-\frac{1}{n^2}}\lambda_{A^-_{\tau_0}A^-_{\tau_0}}$ is negligible when n grows to infinity. We similarly get as $p > n \to \infty$

$$C_i\,(k)\,\mathbb{1}(k < i \leq \tau_0)$$
$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{1}{n-k}\sum_{j=k+1}^{n} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k}\sum_{j=1}^{k} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j)\big)$$
$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{1}{n-k}\sum_{j=\tau_0+1}^{n} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j) + \frac{1}{n-k}\sum_{j=k+1}^{\tau_0} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k}\sum_{j=1}^{k} d_2\,(\boldsymbol{X}_i, \boldsymbol{X}_j)\big)$$
$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{n-\tau_0}{n-k}\lambda_{A^-_{\tau_0}A^+_{\tau_0}} + \frac{\tau_0-k-1}{n-k}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)\big)$$
$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{n-\tau_0}{n-k}\lambda_{A^-_{\tau_0}A^+_{\tau_0}} + \frac{\tau_0-k}{n-k}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) - \sqrt{\frac{k}{n^2(n-k)}}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)$$
$$= \frac{\sqrt{k\,(n-k)}}{n}\big(\frac{n-\tau_0}{n-k}\lambda_{A^-_{\tau_0}A^+_{\tau_0}} - \frac{n-\tau_0}{n-k}\lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) - \sqrt{\frac{1}{n(n-k)}-\frac{1}{n^2}}\lambda_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)$$
$$= \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o(1) + o_P(1).$$

Hence, for all $i \in A^-_{\tau_0}, k \in A^-_{\tau_0} \setminus \{\tau_0\}$, we have as $p > n \to \infty$

$$C_i\,(k) = \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1).$$

Repeating same calculations as above for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, n-1\}$, it is straightforward to obtain the asymptotic limit of $C_i\,(k)$ when $p > n \to \infty$ as follows

$$C_i(k) = \begin{cases} \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^-_{\tau_0}, \forall\, k \in A^-_{\tau_0} \setminus \{\tau_0\}, \\ \frac{n-\tau_0}{n}\sqrt{\frac{k}{n-k}}\big(\lambda_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^+_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^+_{\tau_0}, \forall\, k \in A^-_{\tau_0} \setminus \{\tau_0\}, \\ \frac{\sqrt{\tau_0(n-\tau_0)}}{n}\big(\lambda_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^-_{\tau_0} \text{ and } k = \tau_0, \\ \frac{\sqrt{\tau_0(n-\tau_0)}}{n}\big(\lambda_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^+_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^+_{\tau_0} \text{ and } k = \tau_0, \\ \frac{\tau_0}{n}\sqrt{\frac{n-k}{k}}\big(\lambda_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^-_{\tau_0}, \forall\, k \in A^+_{\tau_0} \setminus \{n\}, \\ \frac{\tau_0}{n}\sqrt{\frac{n-k}{k}}\big(\lambda_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda_{A^-_{\tau_0}A^+_{\tau_0}}\big) + o_P(1) & \forall\, i \in A^+_{\tau_0}, \forall\, k \in A^+_{\tau_0} \setminus \{n\}. \end{cases}$$

This completes the proof of the proposition for $q = 2$.

Proof of Theorem 3.4.4

(a) For the simplicity of presentation, we here only give the proof with the L_1-norm distance $d_1(\boldsymbol{X}_i, \boldsymbol{X}_j)$, because the proof with the L_2-norm distance $d_2(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is very similar. Under Assumptions (B1) - (B2) and using (3.15), we obtain the asymptotic limit of $\frac{1}{n}\sum_{i=1}^{n} {C_i}^2(k)$ for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, n-1\}$ as $p > n \to \infty$, as follows

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} C_i^2(k)\mathbb{1}(k < \tau_0) \\
&= \frac{1}{n}\Big(\sum_{i=1}^{\tau_0} C_i^2(k)\mathbb{1}(k < \tau_0, i \le \tau_0) + \sum_{i=\tau_0+1}^{n} C_i^2(k)\mathbb{1}(k < \tau_0, i > \tau_0)\Big) \\
&= \frac{(n-\tau_0)^2\tau_0 k}{n^3(n-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 + \frac{(n-\tau_0)^3 k}{n^3(n-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 + o_P(1) \\
&= \phi(k)\mathbb{1}(k < \tau_0) + o_P(1), \\
&\frac{1}{n}\sum_{i=1}^{n} C_i^2(k)\mathbb{1}(k = \tau_0) \\
&= \frac{1}{n}\Big(\sum_{i=1}^{\tau_0} C_i^2(k)\mathbb{1}(k = \tau_0, i \le \tau_0) + \sum_{i=\tau_0+1}^{n} C_i^2(k)\mathbb{1}(k = \tau_0, i > \tau_0)\Big) \\
&= \frac{{\tau_0}^2(n-\tau_0)}{n^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 + \frac{\tau_0(n-\tau_0)^2}{n^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 + o_P(1) \\
&= \phi(k)\mathbb{1}(k = \tau_0) + o_P(1), \\
&\frac{1}{n}\sum_{i=1}^{n} C_i^2(k)\mathbb{1}(k > \tau_0) \\
&= \frac{1}{n}\Big(\sum_{i=1}^{\tau_0} C_i^2(k)\mathbb{1}(k > \tau_0, i \le \tau_0) + \sum_{i=\tau_0+1}^{n} C_i^2(k)\mathbb{1}(k > \tau_0, i > \tau_0)\Big) \\
&= \frac{{\tau_0}^3(n-k)}{n^3 k}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 + \frac{{\tau_0}^2(n-\tau_0)(n-k)}{n^3 k}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 + o_P(1) \\
&= \phi(k)\mathbb{1}(k > \tau_0) + o_P(1),
\end{aligned}
$$

where

$$
\phi(k) = \begin{cases}
\frac{(n-\tau_0)^2\tau_0 k}{n^3(n-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 + \frac{(n-\tau_0)^3 k}{n^3(n-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2, & \forall\, k \in A^-_{\tau_0} \setminus \{\tau_0\}, \\
\frac{{\tau_0}^2(n-\tau_0)}{n^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 + \frac{\tau_0(n-\tau_0)^2}{n^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2, & k = \tau_0, \\
\frac{{\tau_0}^3(n-k)}{n^3 k}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 + \frac{{\tau_0}^2(n-\tau_0)(n-k)}{n^3 k}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2, & \forall\, k \in A^+_{\tau_0} \setminus \{n\}.
\end{cases} \tag{3.16}
$$

Using (3.16), we find that

$$\begin{aligned}
&\phi(k)\mathbb{1}(k=\tau_0)-\phi(k)\mathbb{1}(k<\tau_0)\\
&=\frac{n\tau_0(n-\tau_0)(\tau_0-k)}{n^3(n-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}-\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2+\frac{n(n-\tau_0)^2(\tau_0-k)}{n^3(n-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}-\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2\\
&\geq 0.
\end{aligned}$$

Similarly, we use (3.16) to find that $\phi(k)\mathbb{1}(k=\tau_0)-\phi(k)\mathbb{1}(k>\tau_0)\geq 0$. Putting these two together, we have for all $k\in\{1,\ldots,n-1\}$ that as $p>n\to\infty$

$$\max_{1\leq k\leq n-1}\big\{\phi(k)\big\}=\max_{1\leq k\leq n-1}\big\{\phi(k)\mathbb{1}(k\neq\tau_0)+\phi(k)\mathbb{1}(k=\tau_0)\big\}=\phi(k)\mathbb{1}(k=\tau_0). \quad (3.17)$$

Therefore, we have as $p>n\to\infty$

$$\begin{aligned}
&\max_{1\leq k\leq n-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}C_i^2(k)-\phi(k)\mathbb{1}(k=\tau_0)\Big\}\\
&=\max_{1\leq k\leq n-1}\Big\{\big\{\phi(k)\mathbb{1}(k=\tau_0)+o_P(1)\big\}-\phi(k)\mathbb{1}(k=\tau_0)\Big\}=o_P(1),
\end{aligned}$$

which shows that $\frac{1}{n}\sum_{i=1}^{n}C_i^2(k)$ achieves its maximum at $k=\tau_0$. Illustrative examples of this property are already shown in Figures 3.4 and 3.5 in the main thesis. This completes the proof of part (a).

(b) To prove part (b), we use the result in (3.17). We thus have as $p>n\to\infty$

$$\begin{aligned}
\hat{\tau}&=\arg\max_{1\leq k\leq n-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}C_i^2(k)\Big\}\\
&=\arg\max_{1\leq k\leq n-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}C_i^2(k)\mathbb{1}(k\neq\tau_0)+\frac{1}{n}\sum_{i=1}^{n}C_i^2(k)\mathbb{1}(k=\tau_0)\Big\}\\
&=\arg\max_{1\leq k\leq n-1}\Big\{\phi(k)\mathbb{1}(k\neq\tau_0)+o_P(1)+\phi(k)\mathbb{1}(k=\tau_0)+o_P(1)\Big\}\\
&\xrightarrow{P}\arg\max_{1\leq k\leq n-1}\big\{\phi(k)\mathbb{1}(k\neq\tau_0)+\phi(k)\mathbb{1}(k=\tau_0)\big\}\\
&=\arg\max_{1\leq k\leq n-1}\big\{\phi(k)\big\}\\
&=\tau_0,
\end{aligned}$$

where the last equality is obtained using (3.17). Hence, $\hat{\tau}-\tau_0=o_P(1)$, where $o_P(1)$

follows from the third equality above. This completes the proof of part (b).

Proof of Theorem 3.4.5

Using the same calculation of $\phi(k)$ given in (3.16) as in Theorem 3.4.4, as $p \to \infty$ we have

$$\max_{1 \le k \le n-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} C_i^2(k) \right\} = \phi(k)\, \mathbb{1}(k = \tau_0) + o_P(1),$$

where $\phi(k)$ is given in (3.16) and n is fixed. This differs from part (a) of Theorem 3.4.4, where we showed that, as $p > n \to \infty$,

$$\max_{1 \le k \le n-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} C_i^2(k) - \phi(k)\, \mathbb{1}(k = \tau_0) \right\} = o_P(1).$$

Apart from this adjustment, the proof can be followed in the same way as in Theorem 3.4.4.

CHAPTER 4

Numerical results for a single change point scenario

In this chapter, we conduct extensive simulation studies to evaluate the numerical performance of our proposed DCCP method in Chapter 3 for a single change point scenario. We consider the DCCP method based on both the modified L_1-norm distance, here called DCCP-L_1, and the modified L_2-norm distance, here called DCCP-L_2. We compare our methods with four recent methods for high dimensional change points: the divisive and agglomerative method of Matteson and James (2014), called E.divisive; the random projection approach of Wang and Samworth (2018), called Inspect; the method based on spatial and temporal dependence of data proposed by Li et al. (2019a), called HDcp; and the graph-based method using minimum spanning tree proposed by Chen and Zhang (2015), called MST. We compare all the methods on their performance for detecting a single change point under various high dimensional scenarios, especially for the challenging case when n is very small compared to p. We will present the numerical results for the multiple change point scenario in Section 5.3.

4.1 A change in the mean of observations

We first consider the case of a single change point in the mean of high dimensional observations. In the simulations, we consider sample sizes $n \in \{50, 100\}$ and dimensions $p \in \{500, 1000, 2000\}$. We here generate n random observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from the p-variate normal distribution, where the first $3n/5$ observations are drawn from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and the other $2n/5$ observations are drawn from $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. We set $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 \in \left\{\mathbf{0}_p, \left(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\right), \left(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\right)\right\}$, where $\mathbf{0}_p$ and $\mathbf{1}_p$ denote two p-dimensional vectors of zeros and ones respectively. This means there is a change in the mean of observations for two cases of $\boldsymbol{\mu}_2$. We also consider two covariance structures for variables: the identity matrix $\boldsymbol{I}_p$ imposing an uncorrelated covariance structure, and the autoregressive covariance matrix $\boldsymbol{V}_p := \left[0.5^{|i-j|}_{ij}\right]_{i=1}^{p}{}_{j=1}^{p}$ imposing a correlated covariance structure. Note that the true change point location is $\tau = 3n/5$, except in the case of no change point, that is, when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$.

To measure the performance of all the above change point methods, we use the true discovery rate (TDR) to assess the power in detecting the true change point, as well as the root mean squared error (RMSE) to assess the accuracy of the detected change point. Let R denote the total number of replications, and let $\hat{\tau}^{(r)}$ denote the detected change point in the r-th replication. The TDR is defined as follows

$$\text{TDR} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\{\hat{\tau}^{(r)} = \tau\},$$

and the RMSE is

$$\text{RMSE} = \left(\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tau}^{(r)} - \tau\right)^2\right)^{1/2}.$$

We consider 200 replications for each simulation scenario and use $S = 500$ random permutations for our methods DCCP-L_1 and DCCP-L_2. Table 4.1 reports the type I errors of all the methods under the null hypothesis of no change point. It can be seen that all the methods have a reasonably low type I error rate, close to the nominal level 0.05.

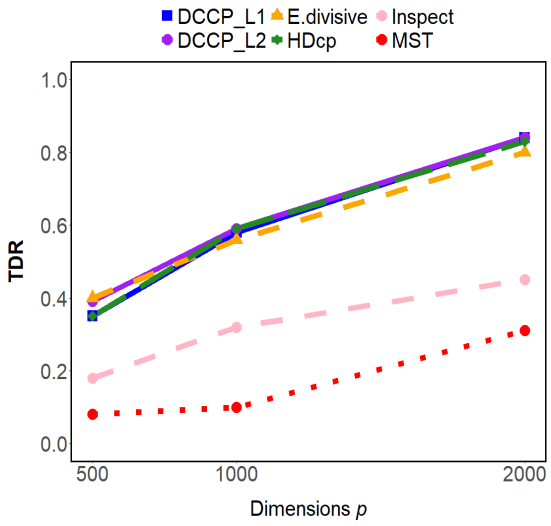
Table 4.2 presents the RMSE results for each method in detecting the true

Table 4.1: Type I error rate for all six methods over 200 replications in the case of no true change point.

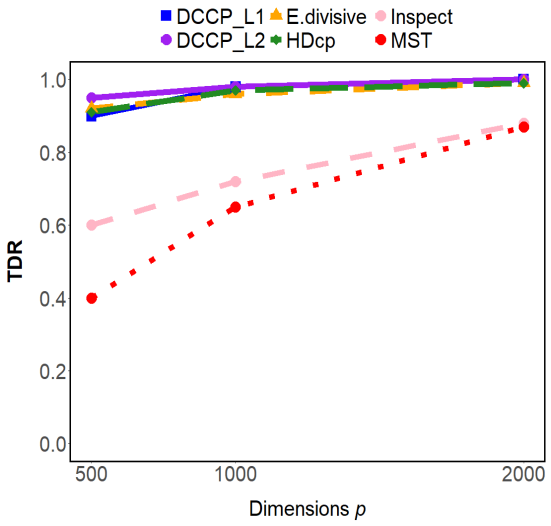
Case	n	p	*Type I error rate of all the methods*					
			DCCP-L_1	DCCP-L_2	E.divisive	HDcp	Inspect	MST
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$ *(same mean, same variance)*	50	500	0.03	0.02	0.04	0.05	0.06	0.03
	50	1000	0.01	0.05	0.06	0.04	0.04	0.06
	50	2000	0.02	0.04	0.02	0.03	0.07	0.04
	100	500	0.04	0.04	0.06	0.02	0.04	0.06
	100	1000	0.04	0.02	0.08	0.04	0.04	0.02
	100	2000	0.03	0.03	0.07	0.04	0.04	0.03

change point in the mean of observations for the case of uncorrelated variables (i.e., $\boldsymbol{\Sigma} = \boldsymbol{I}_p$). The results indicate that our methods, DCCP-L_1 and DCCP-L_2, perform well and show good accuracy, similar to E.divisive and HDcp. For the small sample size $n = 50$, Figures 4.1a and 4.1b visualize the TDR results for all the methods, suggesting that our methods perform reasonably well in detecting a change in the mean of observations compared to the other methods, with the detection power increasing as the dimension gets larger. Note that the MST method is not very competitive in these high dimensional scenarios, as it requires much larger sample sizes, such as $n = 1000$ and $p = 10$, as in their simulation studies. Similarly, while the Inspect method performs better than MST, it is still not very competitive in these non-sparse high dimensional scenarios due to its reliance on sparsity and large sample sizes, such as $n = 2000$, as in their paper.

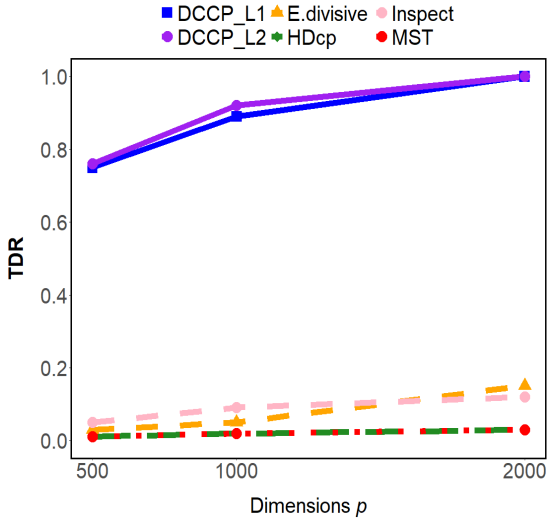
The TDR results for a larger sample size $n = 100$ are reported in Figures 4.2a and 4.2b. We observe that all methods perform better than the results with $n = 50$, which is expected since more information is available in this case. So far, we have conducted simulation studies on detecting a change point in the mean of normal observations. It has been shown that our method performs as well as other methods, with detection power increasing as the dimension grows. This raises a natural question about how our method performs for non-normal observations, such as data drawn from a Student's t distribution. We evaluate this in Section 4.4.



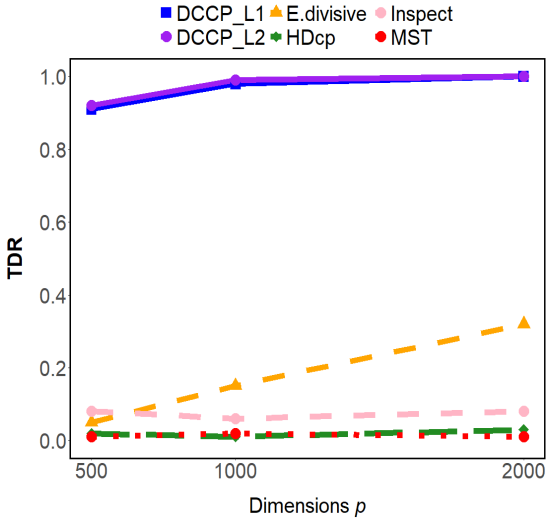
(a) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$



(b) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$

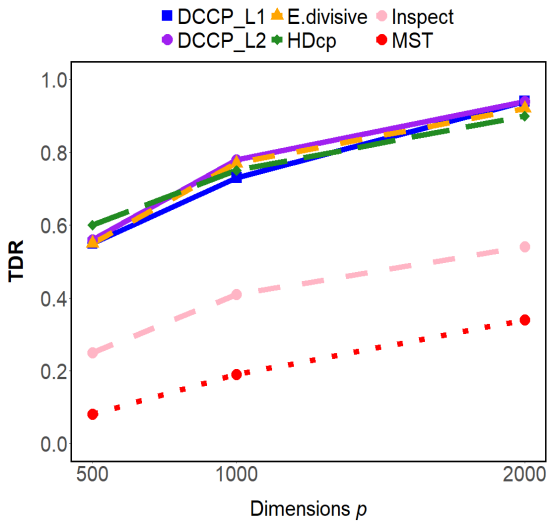


(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{I}_p$

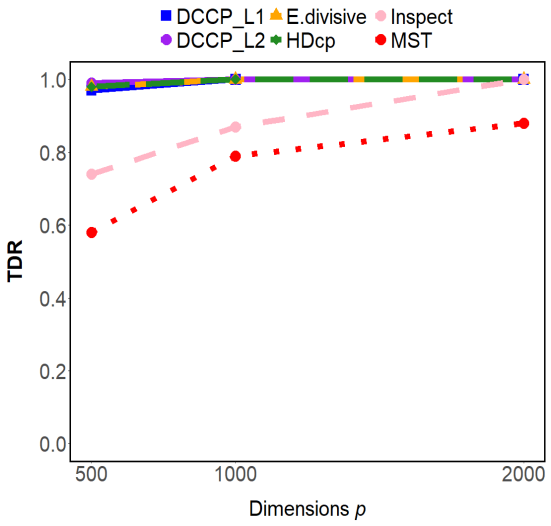


(d) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{I}_p$

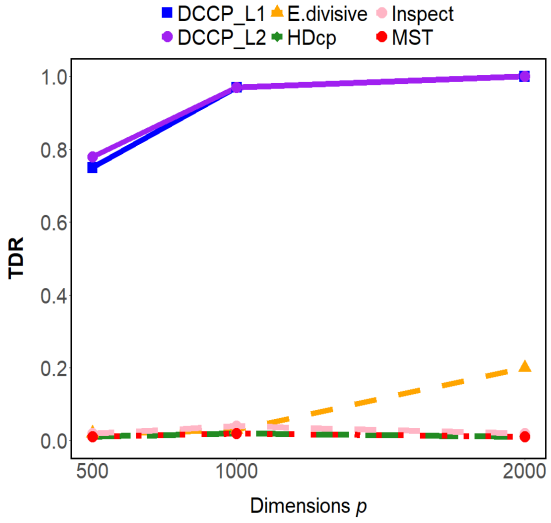
Figure 4.1: True discovery rate (TDR) of all six methods across 200 replications for a single change point detection with sample size $n = 50$: figures (a) and (b) show the results on detecting a change in the mean, and figures (c) and (d) show the results on detecting a change in the variance.



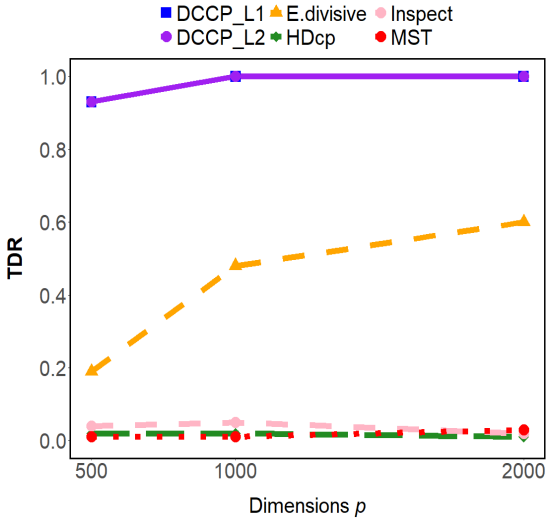
(a) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$



(b) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$



(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{I}_p$



(d) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{I}_p$

Figure 4.2: True discovery rate (TDR) of all six methods across 200 replications for a single change point detection with sample size $n = 100$: figures (a) and (b) show the results on detecting a change in the mean, and figures (c) and (d) show the results on detecting a change in the variance.

4.2 A change in the variance of observations

We next consider the case of a single change in the variance of high dimensional observations. Using the same simulation setting as before, we here generate the first $3n/5$ observations from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and the remaining $2n/5$ observations from $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$ and $\boldsymbol{\Sigma}_2 \in \{1.2\boldsymbol{I}_p, 1.3\boldsymbol{I}_p\}$. Table 4.2 presents the RMSE results for each method in detecting a single change in the variance of observations. Figures 4.1c and 4.1d visualize the TDR results for $n = 50$, and Figures 4.2c and 4.2d show the corresponding results for $n = 100$. These simulation results indicate that our methods, DCCP-L_1 and DCCP-L_2, outperform all the other methods in detecting a change point in the variance of high dimensional observations. As expected, Inspect and HDcp show a poor performance here as they are designed for detecting changes in the mean and not variance. Also, E.divisive tends to improve only when both the dimension of data and the magnitude of variance change get larger. We note that E.divisive requires a large sample size, such as $n = 600$ and a small dimension $p = 5$ in their paper, as previously explained in Section 3.1.

4.3 A change in the distribution while mean and variance remain unchanged

We now consider a challenging problem for many methods in the literature when there is a change in the distribution of observations but the mean and variance of observations remain unchanged (see Zhang and Drikvandi, 2023). For this, we investigate two cases both with $n = 100$ and $p \in \{500, 1000, 2000\}$. One is to generate the first $3n/5$ observations all i.i.d. from a normal distribution $N(1, 1)$ and the remaining $2n/5$ observations all i.i.d. from Exponential distribution Exp(1). Another is to generate the first $3n/5$ observations all i.i.d. from a normal distribution $N(1, 1)$ and the remaining $2n/5$ observations all i.i.d. from a Poisson distribution Pois(1). In these two cases, the distribution of observations changes after location $3n/5$ while both the mean and variance remain unchanged and are equal to 1. The

Table 4.2: Root mean squared errors (RMSE) for all six methods in detecting a single change in the mean or variance of observations over 200 replications.

Case	n	p	*Root mean squared error of all the methods*					
			DCCP-L_1	DCCP-L_2	E.divisive	HDcp	Inspect	MST
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$ *(different mean, same variance)*	50	500	9.29	9.69	7.47	8.75	6.25	16.15
	50	1000	4.51	4.12	3.42	5.79	5.38	13.75
	50	2000	0.61	0.61	1.59	1.60	2.03	9.85
	100	500	3.42	4.97	3.61	3.51	5.70	25.67
	100	1000	0.78	0.71	0.72	0.76	2.41	18.24
	100	2000	0.28	0.26	0.28	0.32	1.24	5.29
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$ *(different mean, same variance)*	50	500	0.38	0.22	1.51	0.24	1.25	5.12
	50	1000	0.14	0.14	0.25	0.54	0.75	1.30
	50	2000	0.00	0.10	0.15	1.90	0.38	0.62
	100	500	0.17	0.10	0.14	0.14	0.72	2.37
	100	1000	0.00	0.00	0.00	0.00	0.54	1.20
	100	2000	0.00	0.00	0.00	0.00	0.10	0.76
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{I}_p$ *(same mean, different variance)*	50	500	1.22	1.05	18.15	18.58	7.02	18.91
	50	1000	0.37	0.33	17.76	18.31	5.52	18.91
	50	2000	0.00	0.00	17.30	18.11	4.71	19.00
	100	500	0.86	0.62	35.12	38.40	15.04	38.83
	100	1000	0.17	0.17	30.55	36.81	10.59	39.00
	100	2000	0.10	0.00	21.49	37.83	11.08	39.00
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{I}_p$ *(same mean, different variance)*	50	500	0.51	0.28	16.83	18.55	6.22	19.00
	50	1000	0.14	0.10	14.96	18.59	6.24	19.00
	50	2000	0.00	0.00	9.72	17.99	5.45	18.91
	100	500	0.26	0.26	21.45	37.83	14.47	39.00
	100	1000	0.00	0.00	9.17	36.26	11.97	39.00
	100	2000	0.00	0.00	1.90	36.57	10.36	38.81

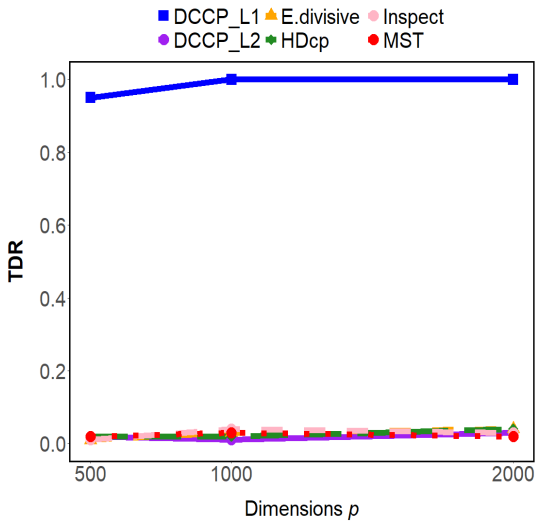
simulation results, shown in Figures 4.3a and 4.3b, indicate that all methods, except our method DCCP-L_1, perform poorly in detecting the change in distribution under this difficult scenario. We note that DCCP-L_1 performs reasonably well here because the asymptotic limit of the modified L_1-norm distance does not simplify to expressions in terms of the mean and variance of observations only, unlike the modified L_2-norm distance (see Remarks 1 and 2).

4.4 A change point with non-normal observations

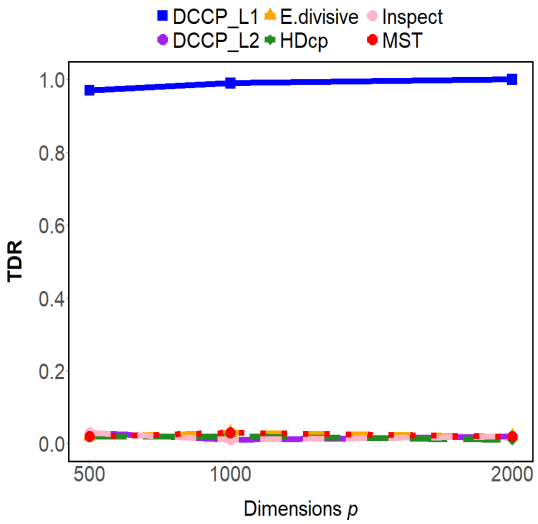
We evaluate the performance of our methods in detecting a single change point in both mean and variance with non-normal observations. We again consider sample size $n = 100$ and dimensions $p \in \{500, 1000, 2000\}$. First, to simulate

Table 4.3: RMSE for all six methods in detecting a single change in the distribution while both the mean and variance remain unchanged over 200 replications.

Case	n	p	*Root mean squared error of all the methods*					
			DCCP-L_1	DCCP-L_2	E.divisive	HDcp	Inspect	MST
	100	500	0.22	38.50	38.20	38.57	22.78	38.70
$N(1,1)$ & Exp(1)	100	1000	0.10	38.64	38.80	38.27	22.92	38.60
	100	2000	0.00	38.62	38.75	38.46	22.19	39.00
	100	500	0.07	39.00	37.97	37.97	22.30	39.00
$N(1,1)$ & Pois(1)	100	1000	0.00	38.48	38.63	38.33	17.73	39.00
	100	2000	0.00	38.82	37.92	38.69	15.57	38.12



(a) For two distributions N(1, 1) and Exp(1).



(b) For two distributions N(1, 1) and Pois(1).

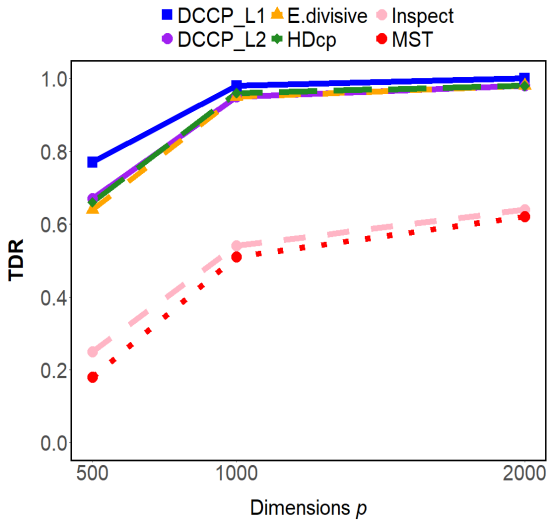
Figure 4.3: True discovery rate (TDR) of all six methods across 200 replications for detecting a change in the distribution while both the mean and variance remain unchanged.

Table 4.4: RMSE for all six methods in detecting a single change point with non-normal observations from a p-variate Student's-t distribution over 200 replications.

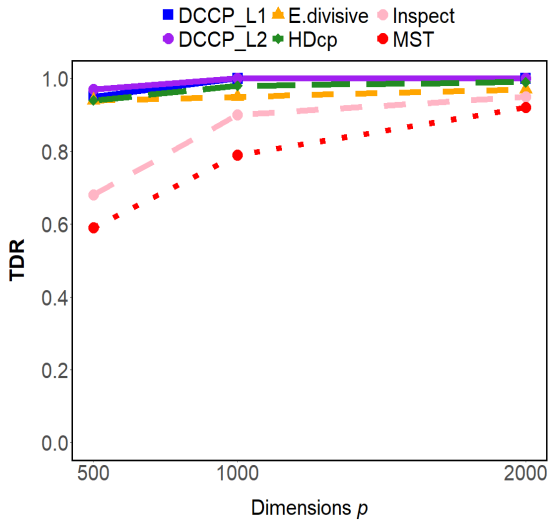
Case	n	p	*Root mean squared error of all the methods*					
			DCCP-L_1	DCCP-L_2	E.divisive	HDcp	Inspect	MST
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$,	100	500	0.67	0.95	4.37	0.96	3.36	16.48
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$, $v_1 = v_2 = 5$	100	1000	0.10	0.22	3.00	0.20	2.03	4.64
(different mean, same variance)	100	2000	0.00	0.14	0.14	0.14	1.22	1.78
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$,	100	500	0.17	0.17	0.87	1.17	0.94	1.26
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$, $v_1 = v_2 = 5$	100	1000	0.03	0.11	3.00	0.08	0.22	0.55
(different mean, same variance)	100	2000	0.00	0.00	5.19	0.04	0.10	0.38
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$,	100	500	2.36	4.39	36.74	39.00	27.96	38.71
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$, $v_1 = 5$, $v_2 = 7$	100	1000	0.93	0.74	33.02	38.89	26.84	39.00
(same mean, different variance)	100	2000	0.63	0.52	23.63	39.33	28.13	39.00
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$,	100	500	1.69	4.02	31.29	39.27	29.68	39.00
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$, $v_1 = 5$, $v_2 = 8$	100	1000	0.58	5.92	22.60	39.14	29.28	38.82
(same mean, different variance)	100	2000	0.22	0.22	7.13	38.82	28.48	38.81

a single change in the mean of non-normal observations, we here generate n random observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from a p-variate Student's-t distribution with standardized covariance matrix where the first $3n/5$ observations are drawn from $t_{v_1}(\boldsymbol{\mu}_1, \boldsymbol{I}_p)$, and the remaining $2n/5$ observations are drawn from $t_{v_2}(\boldsymbol{\mu}_2, \boldsymbol{I}_p)$. We set the degrees of freedom to $v_1 = v_2 = 5$, and we consider $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 \in \big\{\big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big), \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)\big\}$. The RMSE results are shown in Table 4.4.

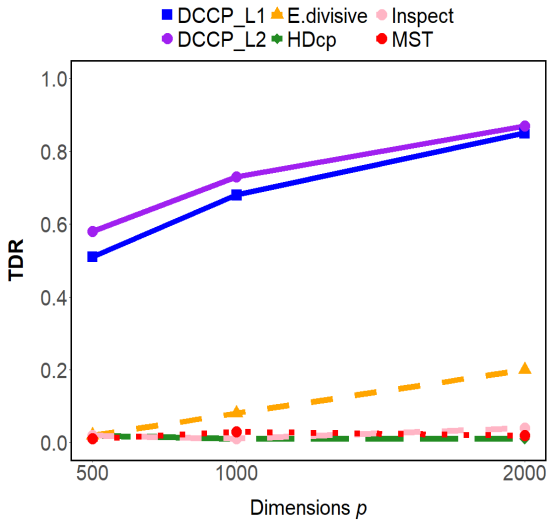
The TDR results, which are reported in Figures 4.4a and 4.4b, show that both of our methods, DCCP-L_1 and DCCP-L_2, perform well compared to the other methods. This is because our approach does not require normality or any other specific distribution, similar to E.divisive, which also performs equally well in this case. Next, to simulate a single change in the variance of non-normal observations, we generate the first $3n/5$ observations from the p-variate Student's-t distribution $t_{v_1}(\boldsymbol{\mu}_1, \boldsymbol{I}_p)$ and the remaining $2n/5$ observations from $t_{v_2}(\boldsymbol{\mu}_2, \boldsymbol{I}_p)$, where we set $v_1 = 5$ and $v_2 \in \{7, 8\}$ to impose different variances, and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$. From the results presented in Figures 4.4c and 4.4d, it can be seen that our methods DCCP-L_1 and DCCP-L_2 perform much better than all the other methods in detecting a change in the variance of non-normal observations, which is in line with the results in Figures 4.2c and 4.2d.



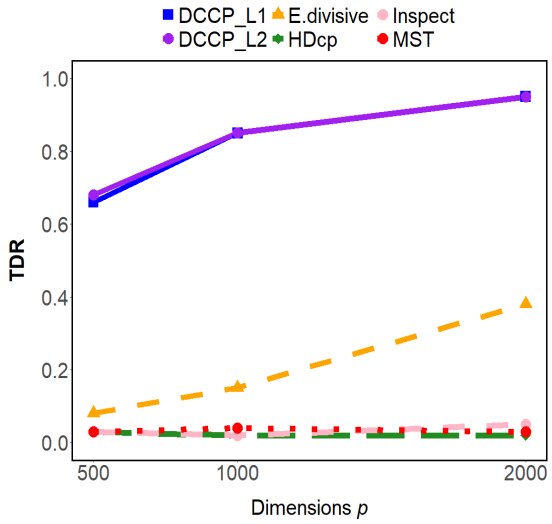
(a) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $v_1 = v_2 = 5$



(b) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $v_1 = v_2 = 5$



(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $v_1 = 5, v_2 = 7$



(d) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $v_1 = 5, v_2 = 8$

Figure 4.4: True discovery rate (TDR) of all six methods across 200 replications for a single change point detection with non-normal observations from a p-variate Student's-t distribution: figures (a) and (b) show the results of detecting a change in the mean, and figures (c) and (d) show the results on detecting a change in the variance.

Table 4.5: RMSE for all six methods in detecting a single change point with correlated variables over 200 replications.

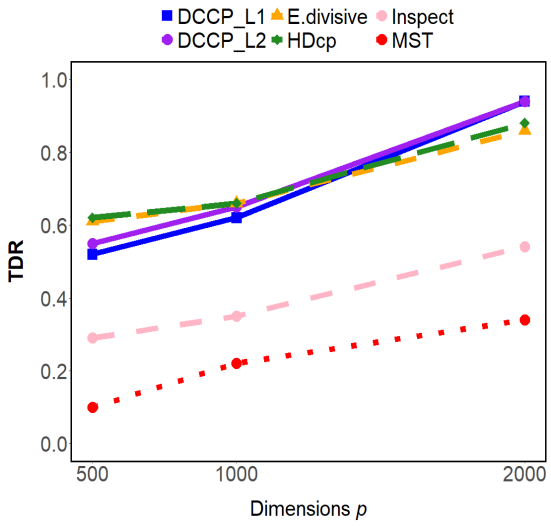
Case	n	p	*Root mean squared error of all the methods*					
			DCCP-L_1	DCCP-L_2	E.divisive	HDcp	Inspect	MST
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$,	100	500	4.90	4.82	3.05	2.03	5.03	26.75
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$	100	1000	1.17	1.23	2.99	0.83	2.23	17.75
(different mean, same variance)	100	2000	0.28	0.29	0.33	0.39	1.44	9.94
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$,	100	500	0.58	0.54	0.62	0.55	1.04	6.42
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$	100	1000	0.26	0.17	0.28	0.21	0.57	2.21
(different mean, same variance)	100	2000	0.00	0.00	0.00	0.00	0.34	1.51
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$,	100	500	1.24	1.12	36.26	37.92	15.23	38.74
$\boldsymbol{\Sigma}_1 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{V}_p$	100	1000	0.38	0.33	33.98	37.97	13.35	39.00
(same mean, different variance)	100	2000	0.12	0.10	27.67	37.51	10.34	38.81
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$,	100	500	0.41	0.36	37.60	37.60	14.37	38.91
$\boldsymbol{\Sigma}_1 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{V}_p$	100	1000	0.14	0.18	16.27	37.55	12.19	38.82
(same mean, different variance)	100	2000	0.07	0.07	6.03	37.52	11.35	38.81

4.5 A change point with correlated variables

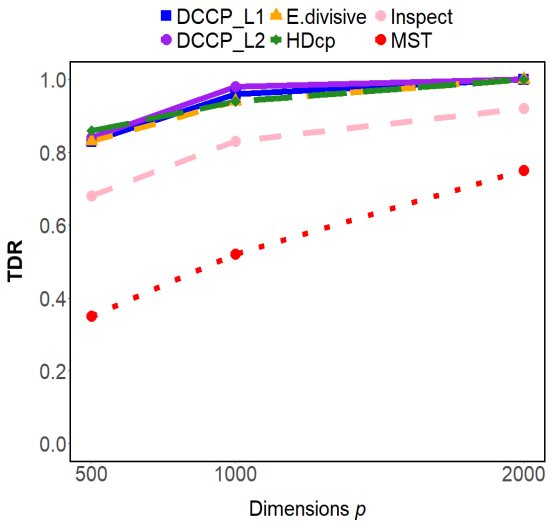
We here evaluate the performance of our methods in detecting a single change point in either the mean or the variance under a correlated covariance structure. We use the same settings as in Sections 4.1 and 4.2, and set $n = 100$. For the case of a change in the mean of observations, we let $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 \in \big\{\big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big), \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)\big\}$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$ where $\boldsymbol{V}_p = \Big[0.5^{|i-j|}_{ij}\Big]_{i=1\,j=1}^{p\quad p}$. For the case of a change in the variance of observations, we let $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{V}_p$, and $\boldsymbol{\Sigma}_2 \in \{1.2\boldsymbol{V}_p, 1.3\boldsymbol{V}_p\}$. The RMSE results are reported in Table 4.5, and the TDR results are reported in Figure 4.5. These simulation results present the same pattern as in the uncorrelated case (Sections 4.1 and 4.2), with performance slightly affected by autoregressive covariance structure, as one may expect.

4.6 A change point with dependent observations

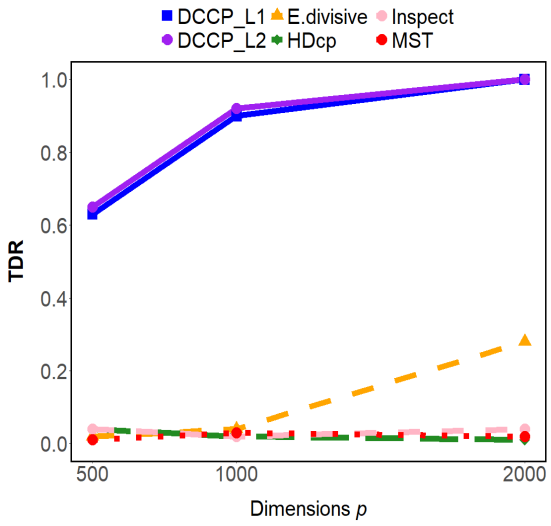
We then evaluate the performance of our methods in detecting a single change point in either the mean or the variance but for weakly dependent observations (in addition to correlated variables). For this, we consider the same simulation setting as before (with $n = 100$) but we here generate the observations from an AR(1)



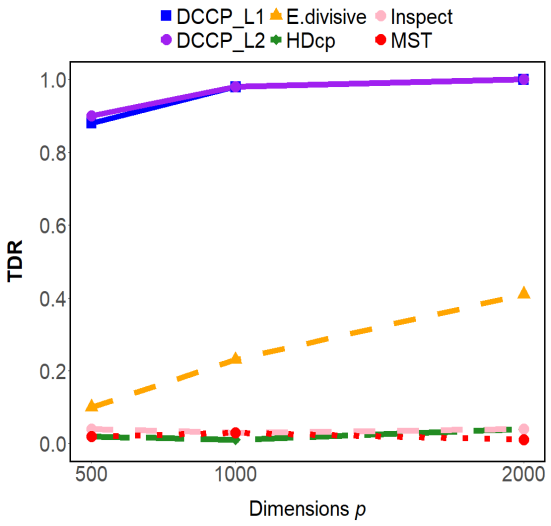
(a) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$



(b) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$



(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{V}_p$



(d) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{V}_p$

Figure 4.5: True discovery rate (TDR) of all six methods across 200 replications for a single change point detection with correlated variables: figures (a) and (b) show the results on detecting a change in the mean, and figures (c) and (d) show the results on detecting a change in the variance.

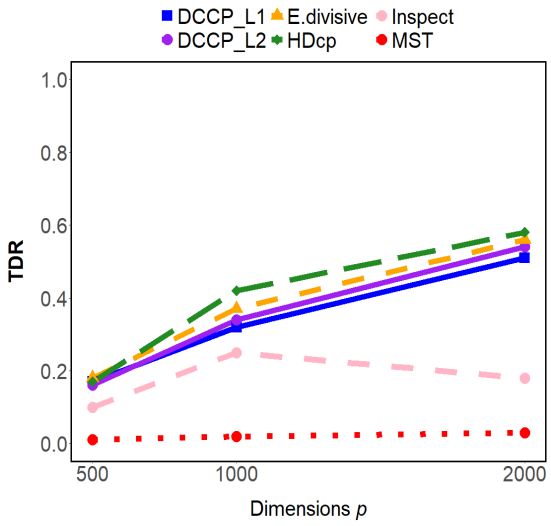
Table 4.6: RMSE for all six methods in detecting a single change point with dependent observations from an AR(1) process over 200 replications.

Case	n	p	*Root mean squared error of all the methods*					
			DCCP-L_1	DCCP-L_2	E.divisive	HDcp	Inspect	MST
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$,	100	500	5.39	6.30	3.72	3.55	7.01	13.04
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$, $\psi = 0.5$	100	1000	2.02	1.85	2.22	1.31	4.30	10.00
(different mean, same variance)	100	2000	0.83	0.83	0.63	0.66	3.17	9.37
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$,	100	500	1.22	1.01	1.93	1.06	1.83	13.86
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$, $\psi = 0.5$	100	1000	0.45	0.28	0.43	0.17	1.26	10.00
(different mean, same variance)	100	2000	0.10	0.00	0.10	0.10	0.78	10.00
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$,	100	500	0.34	0.22	4.90	5.53	7.91	10.00
$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{I}_p$, $\psi = 0.5$	100	1000	0.14	0.10	4.02	5.13	5.75	10.00
(same mean, different variance)	100	2000	0.09	0.05	3.82	4.98	5.52	9.91
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$,	100	500	0.17	0.17	3.22	6.39	8.21	10.00
$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{I}_p$, $\psi = 0.5$	100	1000	0.00	0.00	2.29	8.59	8.00	9.67
(same mean, different variance)	100	2000	0.00	0.00	0.50	6.34	7.78	9.02

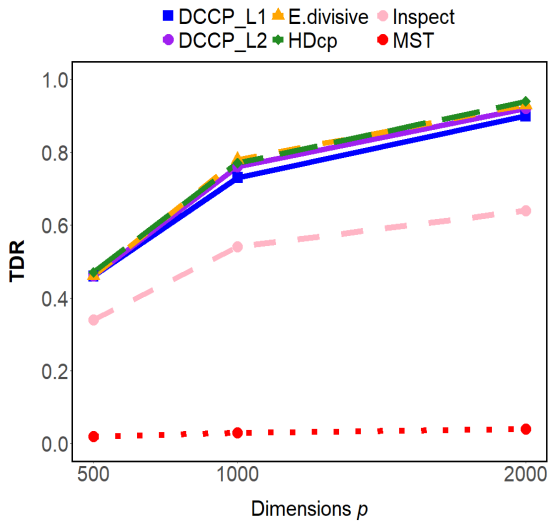
process: $\boldsymbol{X}_i = \psi \boldsymbol{X}_{i-1} + \boldsymbol{\epsilon}_i$, where $\psi = 0.5$ is the autoregressive coefficient and $\boldsymbol{\epsilon}_i$ represents the error term. The RMSE results are presented in Table 4.6. It shows that the accuracy of all methods improves as the dimension p increases, except for MST. The TDR results for normally distributed $\boldsymbol{\epsilon}_i$, which are presented in Figures 4.6a and 4.6b, show that the detection power of a mean shift for all the methods is reduced when the observations are correlated. However, we observe from Figures 4.6c and 4.6d that the detection power of a variance change for these methods is slightly improved when the observations are correlated. This is probably because the dependency structure amplifies deviations in variability, making the change in variance more plausible to discover.

4.7 A change point occurs near the tail of the data sequence

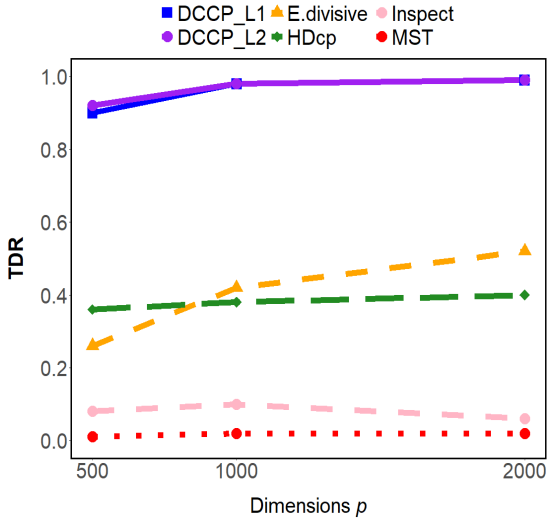
Detecting a change point near the end of the data sequence is challenging because only a limited number of observations are available after the change point. Hence, we are interested in how our methods work in this scenario. We evaluate the performance of our methods in detecting a single change point in either the mean or the variance, with comparison to other methods. We use the same simulation settings as in Sections



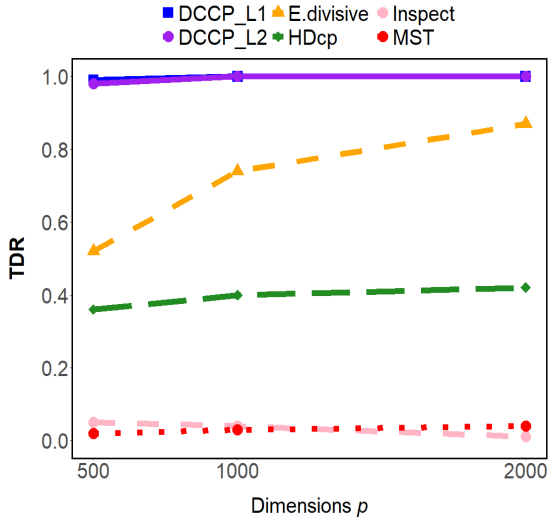
(a) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$



(b) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$



(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{I}_p$



(d) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{I}_p$

Figure 4.6: True discovery rate (TDR) of all six methods across 200 replications for a single change point detection with dependent observations from an AR(1) process: figures (a) and (b) show the results on detecting a change in the mean, and figures (c) and (d) show the results on detecting a change in the variance.

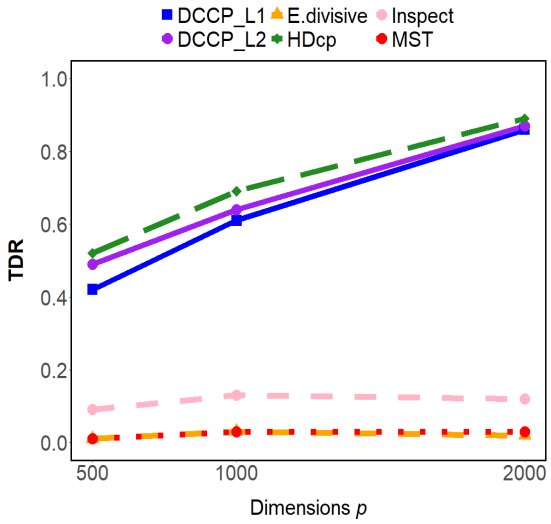
Table 4.7: RMSE for all six methods in detecting a single change point that occurs near the tail of the data sequence ($\tau = 80$ with $n = 100$) over 200 replications.

Case	n	p	*Root mean squared error of all the methods*					
			DCCP-L_1	DCCP-L_2	E.divisive	HDcp	Inspect	MST
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$,	100	500	9.47	7.71	21.84	7.78	15.51	19.28
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$	100	1000	8.00	7.95	21.45	6.91	11.13	18.26
(different mean, same variance)	100	2000	0.59	0.40	21.19	0.46	9.55	17.51
$\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$,	100	500	0.64	0.60	21.95	0.33	6.20	15.45
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$	100	1000	0.24	0.33	20.73	0.20	2.68	12.34
(different mean, same variance)	100	2000	0.00	0.00	20.16	0.00	2.44	10.45
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$,	100	500	0.54	0.50	19.83	17.53	23.03	18.73
$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{I}_p$	100	1000	0.22	0.24	19.66	16.56	19.15	19.03
(same mean, different variance)	100	2000	0.14	0.14	19.80	15.84	17.27	19.37
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$,	100	500	0.28	0.26	20.01	16.79	17.64	18.73
$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{I}_p$	100	1000	0.10	0.10	20.39	14.94	14.03	19.17
(same mean, different variance)	100	2000	0.00	0.00	19.79	13.91	10.46	19.29

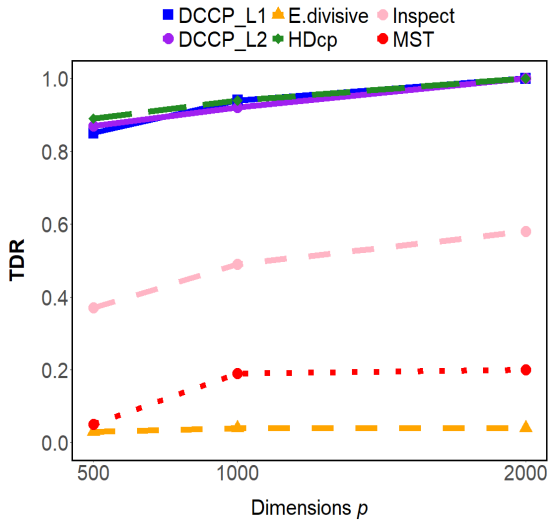
4.1 and 4.2 with fixed $n = 100$, and set the true change point location at $\tau = 80$. The RMSE and TDR results are reported in Table 4.7 and Figure 4.7, respectively. These results indicate that our methods DCCP-L_1 and DCCP-L_2 maintain good performance as p increases, supporting our consistency result in Theorem 3.4.4. We note that the HDcp method also performs competitively in this case; however, E.divisive, Inspect, and MST fail in this challenging scenario. Moreover, if the true change point is located at the extreme tail, such as $\tau = 95$ with $n = 100$, it becomes extremely difficult for any method, including ours, to perform well.

4.8 A change point with varying n/p ratios

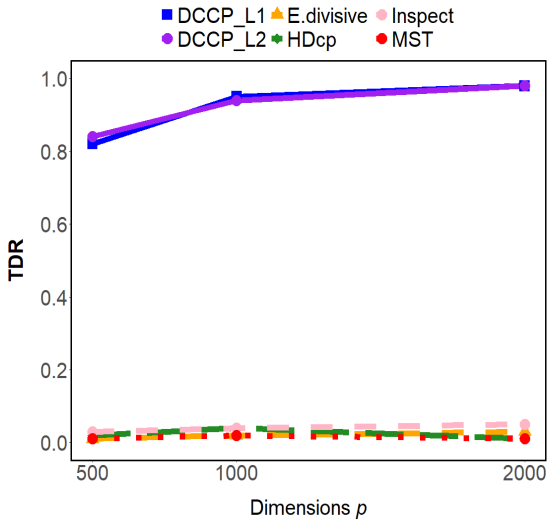
We here present the TDR results for detecting a single change point with varying n/p ratios. We use the same simulation settings as in Section 4.5 with fixed $p = 1000$ and vary n from different values $n \in \{40, 80, 100, 200\}$, which includes the case when n is very small compared to p (i.e., $n = 40$ and $p = 1000$). The results, reported in Figure 4.8, show that our test maintains good performance when n/p varies, also in the case with a very small sample size ($n = 40$). These findings indicate that the distance matrix remains stable with varying n/p ratios, such that the proposed distance-based CUSUM statistics can effectively capture changes in dissimilarity.



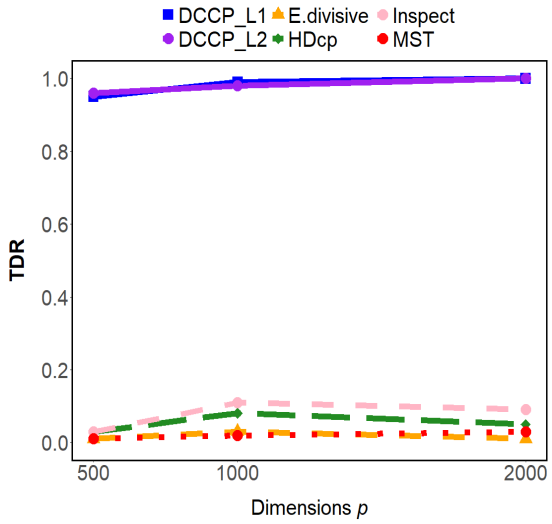
(a) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$



(b) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$



(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{I}_p$



(d) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{I}_p$

Figure 4.7: True discovery rate (TDR) of all six methods across 200 replications for detecting a single change point that occurs near the tail of the data sequence ($\tau = 80$ with $n = 100$): figures (a) and (b) show the results of detecting a change in the mean, and figures (c) and (d) show the results on detecting a change in the variance.

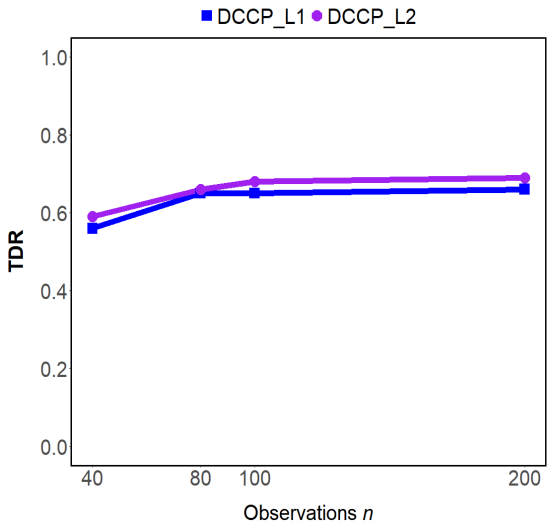
Table 4.8: Average computation time, over 200 replications, for our methods DCCP-L_1 and DCCP-L_2 in detecting a change in the mean of observations where $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$.

n	p	S	Number of true change points	*Computation time (in seconds)*	
				DCCP-L_1	DCCP-L_2
100	500	200	1	21.89	22.45
100	1000	200	1	23.78	23.25
100	2000	200	1	26.21	25.87
100	500	500	1	47.89	47.56
100	1000	500	1	49.92	49.53
100	2000	500	1	52.21	53.32

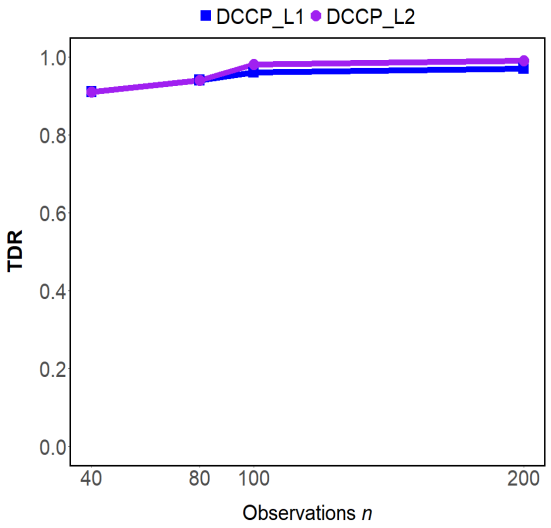
We note that if n is too small, like $n = 10$ or $n = 20$, then it would be extremely difficult for any method, including ours, to expect to work well. This is due to the required minimum sample size or space between change points, as shown in Section 5.2. We also note that while our method still performs reasonably well when the sample size n is small, many existing methods in the literature often require a much larger sample size to work well.

4.9 Computation time

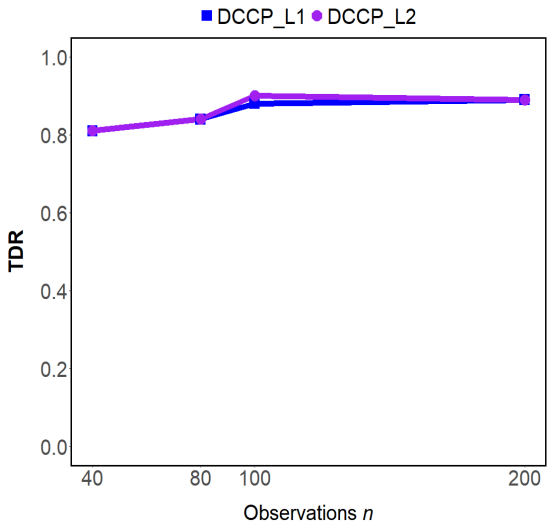
We report the computation time of our methods DCCP-L_1 and DCCP-L_2 for detecting a change point in one of our simulation studies in Section 4.1 with a true change in the mean of observations, where we recall $n = 100$, $p \in \{500, 1000, 2000\}$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$. We record the average computation time over 200 replications using a PC (2.20 GHz, 64 GB RAM). The computation times (in seconds) based on $S \in \{200, 500\}$ random permutations are reported in Table 4.8. The results show that our method has a very reasonable practical runtime in HDLSS settings when p is very large compared to n. Specifically, the computation time for both methods increases modestly with the number of variables p and remains efficient for very large dimensions. For example, it takes about 25 seconds for our method to run when $p = 2000$ with 200 permutations, and it takes around 50 seconds to run when $p = 2000$ with 500 permutations. So our method is computationally efficient and relatively quick to run in these high dimensional settings.



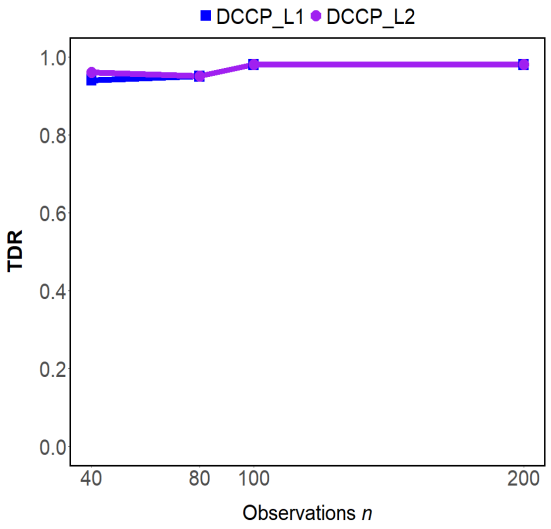
(a) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$



(b) $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$



(c) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{V}_p$



(d) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = 1.3\boldsymbol{V}_p$

Figure 4.8: True discovery rate (TDR) of DCCP-L_1 and DCCP-L_2 across 200 replications for a single change point detection with varying n/p ratios ($p = 1000$): figures (a) and (b) show the results of detecting a change in the mean, and figures (c) and (d) show the results on detecting a change in the variance.

4.10 Concluding remarks

This chapter presents thorough simulation studies to assess the numerical performance of our method for single change point detection. We have shown that our method can effectively detect non-sparse high dimensional change points where changes may happen in many variables but with small significant magnitudes. Moreover, the results have shown that the method detects a wide range of distributional changes, including a change in the mean, a change in the variance (or covariance), and a change in the shape of distribution while the mean and variance remain unchanged. The last case is particularly challenging, and many existing methods struggle to detect it. Across different data settings, DCCP has been shown to perform well with spatially correlated variables, temporally dependent observations, and non-normal data with heavy tails. Furthermore, numerical results demonstrate our method remains consistent in several demanding scenarios. In particular, it maintains good accuracy when the change point occurs near the end of the sequence and across a wide range of n/p ratios, even for very small samples such as $n = 40$ with $p = 1000$. We also observe that the computation times are reasonable and remain efficient even for very large dimensions (e.g., $p = 2000$). A possible future topic of this work is to adapt the methodology to more complex data structures, such as sequences with missing values (e.g., financial time series) or contaminated observations (e.g., sensor data with outliers), which frequently occur in modern applications.

CHAPTER 5

Extension to multiple change points

In this chapter, we utilize the proposed distance-based CUSUM statistic to detect multiple change points in high dimensional data sequences. In Section 5.1, we introduce the problem setting for multiple change point detection. In Section 5.2, we present a multiple change point detection algorithm and provide theoretical guarantees showing that our method can consistently find multiple non-sparse change points when $p > n \to \infty$ under certain conditions. In Section 5.3, we conduct simulation studies for the multiple change point scenario. In Section 5.4, we present two real data applications: S&P 500 data and MIT cellphone data, both of which are high dimensional datasets. In Section 5.5, we discuss other alternatives that can be considered for incorporation into our proposed framework for multiple change point detection. In Section 5.6, we provide the proofs for the theoretical result.

5.1 Multiple change point problem setting

In this section, we extend the proposed distance-based CUSUM statistic to detect multiple change points for high dimensional data sequences. We assume the number of change points and their locations are both unknown. The problem of multiple

change point detection can be formulated as the following hypothesis test

$$\begin{cases} H_0 : F_1 = F_2 = \cdots = F_n \\ H_1^m : F_1 = \cdots = F_{\tau_1} \neq F_{\tau_1+1} = \cdots = F_{\tau_2} \neq F_{\tau_2+1} = \cdots = F_{\tau_z} \neq F_{\tau_z+1} = \cdots = F_n, \end{cases} \tag{5.1}$$

where $1 \leq \tau_1 < \tau_2 < \cdots < \tau_z < n$ are unknown change point locations and z is the number of change points, which is also unknown. If H_0 in (5.1) is rejected, the main objective will be to find the z change point estimates $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_z$. As introduced in Section 2.1, we employ the RBS approach of Vostrikova (1981), incorporating our distance-based CUSUM procedure for single change point detection. We also incorporate the WBS approach of Fryzlewicz (2014) with the proposed distance-based CUSUM, which will be discussed in Section 5.5.

The RBS process begins by applying our Algorithm 1 for single change point to the entire data sequence $[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]$. If a significant change point is detected, the data sequence is split into two segments, one before and including the detected change point, and another after the detected change point. The process recursively continues with each segment for further change points. We provide the full details in the following.

5.2 DCCP for multiple change point detection with theoretical results

Let s_m and e_m denote the starting and ending indices of a sub-sequence, which we define as $[\boldsymbol{X}_{s_m}, \boldsymbol{X}_{s_m+1}, \ldots, \boldsymbol{X}_{e_m}]$. We first apply Algorithm 1 to detect a significant change point $\hat{\gamma}_m$ within this sub-sequence. If a significant change point is found, the sub-sequence $[\boldsymbol{X}_{s_m}, \boldsymbol{X}_{s_m+1}, \ldots, \boldsymbol{X}_{e_m}]$ is split into two new sub-sequences $[\boldsymbol{X}_{s_m}, \boldsymbol{X}_{s_m+1}, \ldots, \boldsymbol{X}_{\hat{\gamma}_m}]$ and $[\boldsymbol{X}_{\hat{\gamma}_m+1}, \boldsymbol{X}_{\hat{\gamma}_m+2}, \ldots, \boldsymbol{X}_{e_m}]$. We then apply Algorithm 1 to these new sub-sequences to check for additional significant change points. We continue this recursive search until no further change points are detected or a minimum segment length is reached. We set the default minimum segment length to 10, which is also commonly used in the literature (e.g., Drikvandi and Modarres, 2025).

Note that $(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_z) = \text{sort}\,(\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_z)$ in increasing order, since the change point estimates $\hat{\gamma}_i$ from RBS are not necessarily ordered. Algorithm 2 summarizes our distance-based CUSUM method for multiple change point detection.

Algorithm 2: DCCP for multiple change point detection with RBS

Input: A data sequence or matrix of observations $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]^\top$.

Output: A list of ordered significant change point estimates $\{\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_z\}$, or "NA" if there is no significant change point.

Step 1: Apply Algorithm 1 for single change point detection to the data sequence $\boldsymbol{X}$. If there is no significant change point, output "NA". Otherwise, denote the detected change point by $\hat{\gamma}_1$ and go to the next step.

Step 2: Split the data sequence $\boldsymbol{X}$ into two sub-sequences before and after the detected change point $\hat{\gamma}_1$. Apply Algorithm 1 to each of the two sub-sequences to search for more change points.

Step 3: Repeat Step 2 until no further sub-sequences contain significant change points or the minimum segment length (our default is 10) is reached.

Step 4: Denote all detected change points by $\{\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_z\}$ and return $(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_z) = \text{sort}\,(\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_z)$ in increasing order as a list of ordered significant change points.

We here expand on the technical details of our method for multiple change point detection. For a data sequence $[\boldsymbol{X}_{s_m}, \boldsymbol{X}_{s_m+1}, \ldots, \boldsymbol{X}_{e_m}]$, the distance-based CUSUM in (3.4) can be written as follows

$$C_i(k, s_m, e_m) = \frac{\sqrt{(k - s_m + 1)\,(e_m - k)}}{e_m - s_m + 1}\Big(\frac{1}{e_m - k}\sum_{j=k+1}^{e_m} d_q(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{k - s_m + 1}\sum_{j=s_m}^{k} d_q(\boldsymbol{X}_i, \boldsymbol{X}_j)\Big), \tag{5.2}$$

where $k \in \{s_m, s_m + 1, \ldots, e_m - 1\}$. Accordingly, we calculate the average column sums of the squared CUSUM matrix $\boldsymbol{C}$ for the data sequence $[\boldsymbol{X}_{s_m}, \boldsymbol{X}_{s_m+1}, \ldots, \boldsymbol{X}_{e_m}]$, that is $\frac{1}{e_m - s_m + 1}\sum_{i=s_m}^{e_m} {C_i}^2\,(k, s_m, e_m)$. Then, the change point estimate is as follows

$$\hat{\gamma}_m = \underset{s_m \le k \le e_m - 1}{\arg\max}\Big\{\frac{1}{e_m - s_m + 1}\sum_{i=s_m}^{e_m} {C_i}^2\,(k, s_m, e_m)\Big\},$$

where the case $s_1 = 1$ and $e_1 = n$ indicates that the first time search segment is the

entire data sequence, as $e_1 - s_1 + 1 = n$. To conduct the test of significance, we similarly apply the permutation test with the following test statistic based on $\hat{\gamma}_m$

$$T_{s_m,e_m}(\hat{\gamma}_m) = \frac{1}{e_m - s_m + 1} \sum_{i=s_m}^{e_m} C_i{}^2 \left(\hat{\gamma}_m, s_m, e_m\right). \tag{5.3}$$

Example 2. To illustrate this using a simple example, we consider a scenario with two true change points. We simulate 30 1000-dimensional observations all i.i.d. from normal distribution $N(0, 0.5)$, another 30 observations all i.i.d. from the same distribution but with a mean shift of 1, and another 40 observations all i.i.d. from the same distribution but with a mean shift of 3. This means that the two true change points are at locations $\tau_1 = 30$ and $\tau_2 = 60$. Figure 5.1 visualizes Algorithm 2 for this example, showing how the algorithm can detect multiple change points in high dimensional observations.

In the following theorem, we prove the consistency of our CUSUM method for multiple change point detection under some conditions.

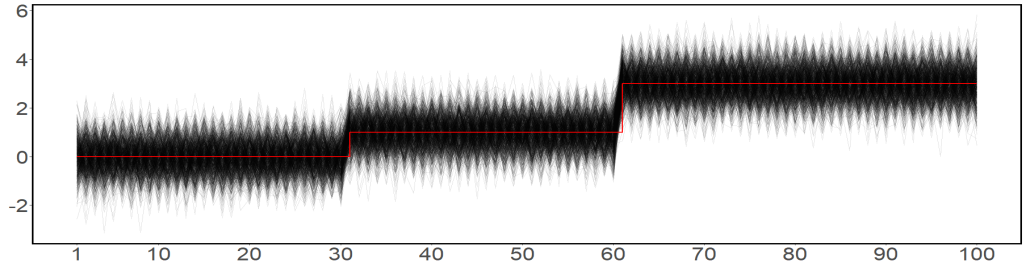
Theorem 5.2.1. *Suppose that there are z true change points $\tau_1^0, \tau_2^0, \ldots, \tau_z^0, 1 \leq \tau_1^0 < \tau_2^0 < \ldots < \tau_z^0 < n$, so that $F_1 = \cdots = F_{\tau_1^0} \neq F_{\tau_1^0+1} = \cdots = F_{\tau_2^0} \neq F_{\tau_2^0+1} = \cdots = F_{\tau_z^0} \neq F_{\tau_z^0+1} = \cdots = F_n$. Assume the minimum space between change points satisfies $\min_{1\leq i\leq z-1} \left|\tau_{i+1}^0 - \tau_i^0\right| \geq Mn^\varepsilon$ for some $M > 0$ and $\varepsilon \leq 1$. Under Assumptions (A1)-(A2) or (B1)-(B2), Algorithm 2 returns the change point estimates $(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_z)$ satisfying*

$$\left\|(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_z) - \left(\tau_1^0, \tau_2^0, \ldots, \tau_z^0\right)\right\|_\infty = o_P(1),$$

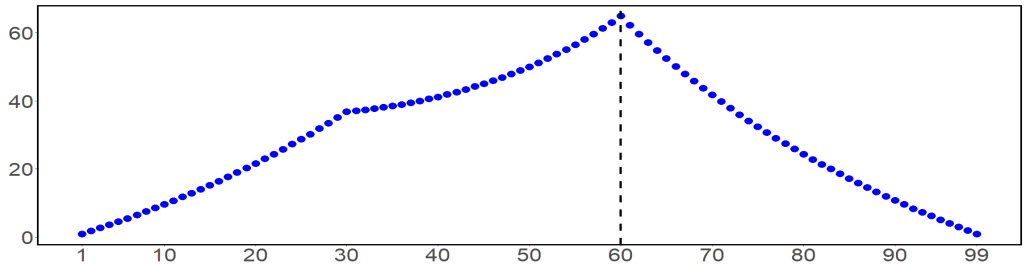
where $p > n \to \infty$.

We note that the condition $\min_{1\leq i\leq z-1} \left|\tau_{i+1}^0 - \tau_i^0\right| \geq Mn^\varepsilon$ ensures that there are not too many change points to discover, since it implies $z \leq \left(\min_{1\leq i\leq z-1} \left|\tau_{i+1}^0 - \tau_i^0\right| / M\right)^{1/\varepsilon}$ as $0 \leq z < n$ (see also Vostrikova, 1981; Fryzlewicz, 2014). Also, in the proof of Theorem 5.2.1 in Section 5.6, we demonstrate that the algorithm does not return additional change points asymptotically when $p > n \to \infty$.

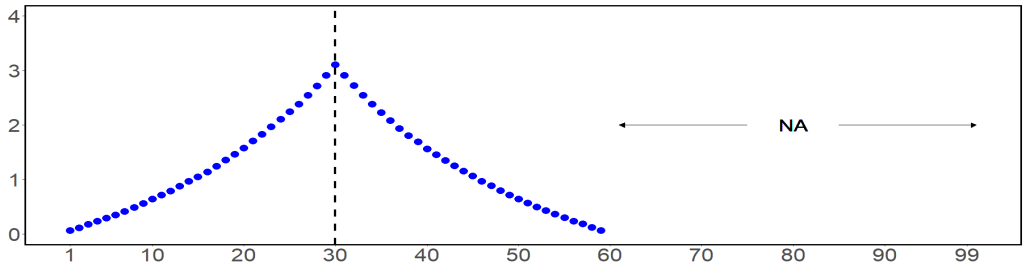
As noted above, RBS requires that the minimum spacing between change points is not too small. This naturally raises the question of how the method performs



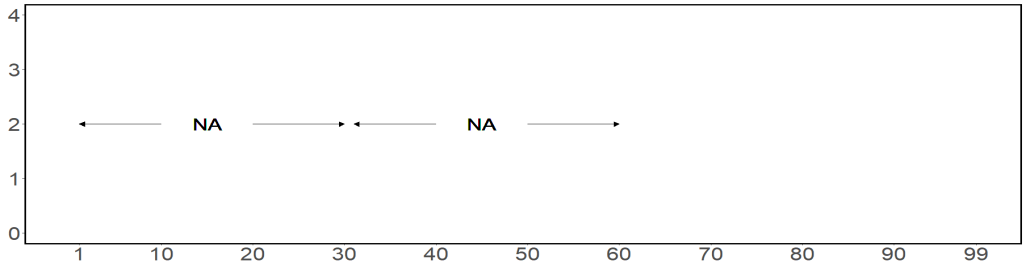
(a) High dimensional observations with $n = 100$ and $p = 1000$, and two true change points at locations 30 and 60. The red line shows the true mean of the observations. The x-axis shows the observation index, and the y-axis shows the values of the standardized variables.



(b) A significant change point is detected at location 60 in the sequence $[X_1, \ldots, X_{100}]$. The x-axis shows the candidate search location k, and the y-axis shows the column sums of squared CUSUM statistics.



(c) A significant change point is detected at location 30 in the sequence $[X_1, \ldots, X_{60}]$. No change point is detected in the sequence $[X_{61}, \ldots, X_{100}]$. The x-axis shows the candidate search location k, and the y-axis shows the column sums of squared CUSUM statistics.



(d) No change point is detected in the sequences $[X_1, \ldots, X_{30}]$ and $[X_{31}, \ldots, X_{60}]$.

Figure 5.1: Illustrative example of Algorithm 2 with two true change points. Significantly detected change points are marked by vertical dashed lines. If no significant change point is detected in a segment, it is labelled as "NA".

when change points are closely spaced under the alternative hypothesis. We examine this in Subsection 5.3.2.

We have developed an R package called `distCUSUM` for the implementation of our distance-based CUSUM method, using both the modified L_1-norm and modified L_2-norm functions. The R package will be made available online on GitHub at `https://github.com/lupengzhang/distCUSUM`. The R package returns significant change point locations and corresponding p-values. It can also be applied with any other distance function specified by the user. Our default distance function is the modified L_1-norm because it requires weaker assumptions compared to the modified L_2-norm and generally has a better empirical performance, as shown in our simulation and data application results.

5.3 Numerical results for multiple change points

5.3.1 Multiple change points in the mean or variance of observations

We here conduct simulations for the case when there are multiple change points. Considering $n = 100$ and $p \in \{500,\ 1000,\ 2000\}$, we simulate high dimensional data with three true change points at locations $\tau_1 = n/5$, $\tau_2 = 2n/5$, and $\tau_3 = 4n/5$. For this, we generate the first $n/5$ observations from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, the next $n/5$ observations from $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, the next $2n/5$ observations from $N(\boldsymbol{\mu}_3,\ \boldsymbol{\Sigma}_3)$, and the last $n/5$ observations from $N(\boldsymbol{\mu}_4,\ \boldsymbol{\Sigma}_4)$. For the case of three changes in the mean of observations, we let $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4 = \boldsymbol{V}_p$. For the case of three changes in the variance of observations, we let $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 = \mathbf{0}_p$, and $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$ and $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$.

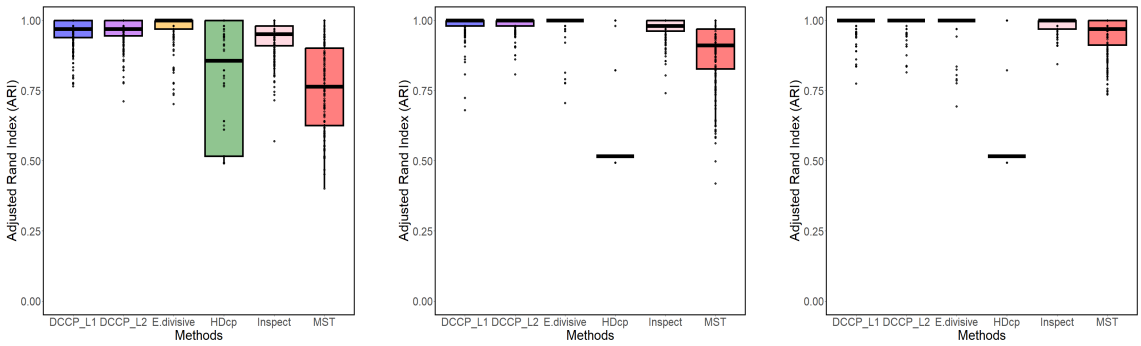
We use Algorithm 2 in Section 5.2, with the minimum segment size of 10. We measure the accuracy of the change point estimates using the adjusted Rand index (ARI) by Rand (1971), since change point estimation can be viewed as a special case of classification. The ARI measure is also adopted in other change point studies (e.g.,

Matteson and James, 2014; Wang and Samworth, 2018). The ARI results over 200 replications are visualized by the box plots in Figure 5.2. Also, Table 5.1 presents the results on frequency and the average number of true change points detected for all the methods. It can be seen that, for detecting multiple change points in the mean, the performance of our methods DCCP-L_1 and DCCP-L_2 is as good as E.divisive, while our methods outperform E.divisive in detecting multiple change points in the variance. Again, the other methods are not competitive in these high dimensional settings. For instance, Inspect requires sparsity and performs better when the sample size n is very large. In our simulations, it is observed that HDcp cannot exactly identify the true change points in these high dimensional scenarios, which explains its low accuracy as exhibited in Figure 5.2. From Table 5.1, one can see that our methods DCCP-L_1 and DCCP-L_2 have an average number of truly detected change points much closer to 3 (the number of true change points). This is in line with our consistency result in Theorem 5.2.1.

5.3.2 Closely located multiple change points

We present the frequency and average number of truly detected change points when the true change points are closely located. The simulation settings are the same as in Subsection 5.3.1, except that the three true change points are set at $\tau_1 = 40$, $\tau_2 = 50$, and $\tau_3 = 60$. In Subsection 5.3.1, we set three true change point locations at 20, 40, and 80. The results, reported in Table 5.2, show a slight decrease in performance of our methods DCCP-L_1 and DCCP-L_2. This is expected as the binary segmentation requires the assumption that the minimum space of change points is not too small (see the assumption in Theorem 5.2.1). However, our methods achieve an average number of truly detected change points much closer to 3 (the number of true change points) when p increases. This supports that the consistency of our test in detecting both the locations and the number of multiple change points still holds here as p increases.

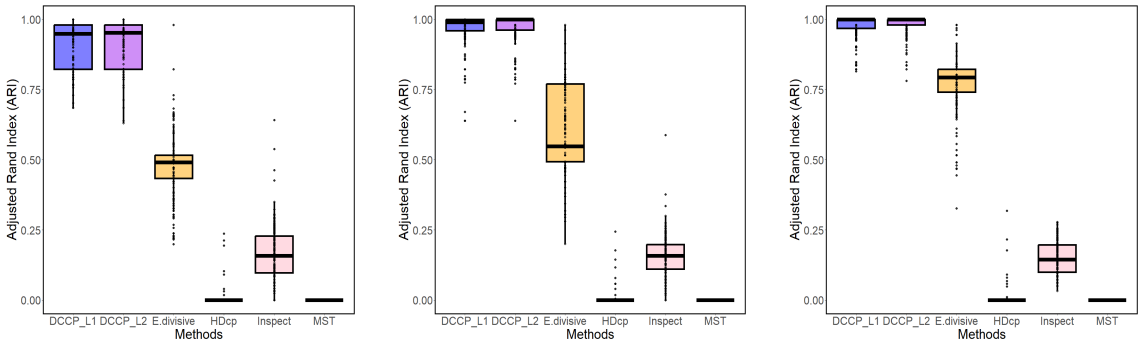
Moreover, while other methods show a slight drop in performance in this case, E.divisive, HDcp, Inspect, and MST are also affected, with their performance often decreasing more severely. This is because closely located change points are difficult



(a) $p = 500$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$

(b) $p = 1000$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$

(c) $p = 2000$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$



(d) $p = 500$, $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$

(e) $p = 1000$, $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$

(f) $p = 2000$, $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$

Figure 5.2: Adjusted Rand index (ARI) of all six methods over 200 replications for multiple change point detection with three true change points ($\tau_1 = 20$, $\tau_2 = 40$, and $\tau_3 = 80$), $n = 100$, and $p \in \{500, 1000, 2000\}$. Note that figures (a), (b) and (c) are for the changes in the mean, and figures (d), (e) and (f) are for the changes in the variance.

Table 5.1: Frequency and average number of the truly detected change points across 200 replications for all six methods in the case when there are three true change points ($\tau_1 = 20$, $\tau_2 = 40$, and $\tau_3 = 80$) in the data.

Case	n	p	Number of true change points	Methods	Frequency of total true change points detected				Average number of true change points detected
					0	1	2	3	
different mean, same variance $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$	100	500	3	DCCP-L_1	0.01	0.10	0.42	0.47	2.35
	100	500	3	DCCP-L_2	0.01	0.11	0.39	0.49	2.36
	100	500	3	E.divisive	0.00	0.05	0.41	0.54	2.49
	100	500	3	HDcp	0.03	0.37	0.29	0.31	1.88
	100	500	3	Inspect	0.03	0.34	0.42	0.21	1.81
	100	500	3	MST	0.27	0.44	0.26	0.03	1.05
	100	1000	3	DCCP-L_1	0.00	0.01	0.21	0.78	2.77
	100	1000	3	DCCP-L_2	0.00	0.01	0.17	0.82	2.81
	100	1000	3	E.divisive	0.00	0.01	0.14	0.85	2.84
	100	1000	3	HDcp	0.01	0.73	0.12	0.14	1.39
	100	1000	3	Inspect	0.01	0.13	0.42	0.44	2.29
	100	1000	3	MST	0.11	0.35	0.37	0.17	1.60
	100	2000	3	DCCP-L_1	0.00	0.00	0.01	0.99	2.99
	100	2000	3	DCCP-L_2	0.00	0.00	0.01	0.99	2.99
	100	2000	3	E.divisive	0.00	0.00	0.01	0.99	2.99
	100	2000	3	HDcp	0.00	0.99	0.01	0.00	1.01
	100	2000	3	Inspect	0.00	0.02	0.28	0.70	2.68
	100	2000	3	MST	0.02	0.19	0.37	0.42	2.19
same mean, different variance $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$	100	500	3	DCCP-L_1	0.00	0.08	0.71	0.21	2.13
	100	500	3	DCCP-L_2	0.01	0.06	0.66	0.27	2.19
	100	500	3	E.divisive	0.66	0.34	0.00	0.00	0.34
	100	500	3	HDcp	0.98	0.02	0.00	0.00	0.02
	100	500	3	Inspect	0.93	0.07	0.00	0.00	0.07
	100	500	3	MST	0.98	0.02	0.00	0.00	0.02
	100	1000	3	DCCP-L_1	0.00	0.01	0.46	0.53	2.52
	100	1000	3	DCCP-L_2	0.00	0.01	0.45	0.54	2.53
	100	1000	3	E.divisive	0.46	0.41	0.13	0.00	0.67
	100	1000	3	HDcp	0.99	0.01	0.00	0.00	0.01
	100	1000	3	Inspect	0.95	0.05	0.00	0.00	0.05
	100	1000	3	MST	0.99	0.01	0.00	0.00	0.01
	100	2000	3	DCCP-L_1	0.00	0.00	0.20	0.80	2.80
	100	2000	3	DCCP-L_2	0.00	0.00	0.19	0.81	2.81
	100	2000	3	E.divisive	0.14	0.49	0.37	0.00	1.23
	100	2000	3	HDcp	0.99	0.01	0.00	0.00	0.01
	100	2000	3	Inspect	0.97	0.03	0.00	0.00	0.03
	100	2000	3	MST	0.99	0.01	0.00	0.00	0.01

for these methods as well. For example, E.divisive uses hierarchical clustering to detect multiple change points, and when two change points are close, the clustering step can merge them into one. In Figure 5.3, our methods also achieve higher ARI values, indicating that our change point estimates more accurately identify multiple change points than the other methods.

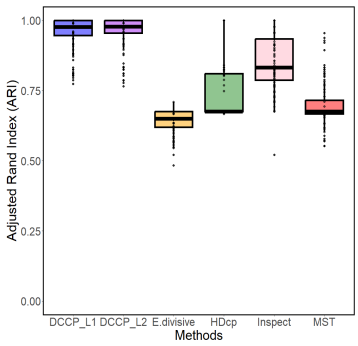
5.3.3 Computation time

We report the computation time of our methods DCCP-L_1 and DCCP-L_2 for detecting three change points in one of our simulation studies in Subsection 5.3.1. This simulation is in addition to the computation time for detecting a single change point in Section 4.9. We do so with true changes in the mean of observations, where we recall $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4 = \boldsymbol{V}_p$. We record the average computation time over 200 replications using a PC (2.20 GHz, 64 GB RAM). The computation times (in seconds) based on $S \in \{200, 500\}$ random permutations are reported in Table 5.3. The results show that our method runs in a few minutes for detecting three change points. Computation time for both methods increases modestly with the number of variables p and remains efficient even in very large dimensions ($p = 2000$).

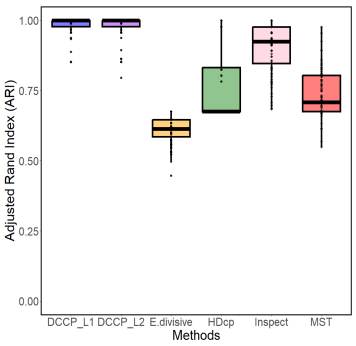
5.4 Real data applications

5.4.1 Application I: S&P 500 data

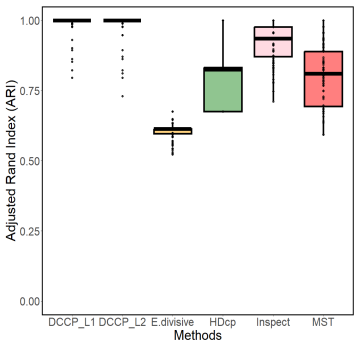
As our first data application, we apply our distance-based CUSUM method to analyze the S&P 500 data from the US stock market return, which tracks the stock performance for 500 of the largest companies listed on stock exchanges in the US. The dataset is available online at https://www.finance.yahoo.com and can be obtained using the R package *BatchGetSymbols* for different time periods. Our analysis here focuses on daily closing prices of S&P 500 index stocks spanning from 2020-01-01 to 2020-05-29. This period includes the start of the COVID-19 pandemic and comprises $n = 108$ trading days across $p = 496$ company stocks. The reason we select this time



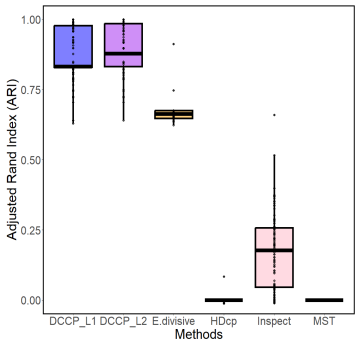
(a) $p = 500$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$



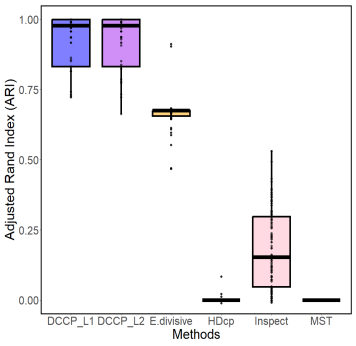
(b) $p = 1000$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$



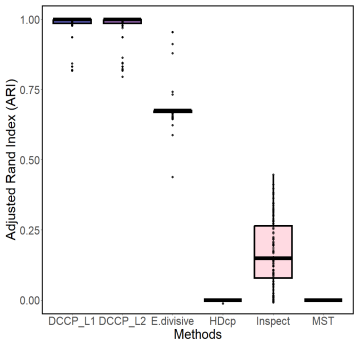
(c) $p = 2000$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$



(d) $p = 500$, $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$



(e) $p = 1000$, $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$



(f) $p = 2000$, $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$

Figure 5.3: Adjusted Rand index (ARI) of all six methods over 200 replications in the case when the space between change points is small ($\tau_1 = 40$, $\tau_2 = 50$, and $\tau_3 = 60$), with $n = 100$ and $p \in \{500, 1000, 2000\}$. Note that figures (a), (b) and (c) are for the changes in the mean, and figures (d), (e) and (f) are for the changes in the variance.

Table 5.2: Frequency and average number of the truly detected change points across 200 replications for all six methods in the case when the space between change points is small ($\tau_1 = 40$, $\tau_2 = 50$, and $\tau_3 = 60$).

Case	n	p	Number of true change points	Methods	Frequency of total true change points detected				Average number of true change points detected
					0	1	2	3	
different mean, same variance $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$	100	500	3	DCCP-L_1	0.01	0.17	0.42	0.40	2.21
	100	500	3	DCCP-L_2	0.01	0.12	0.45	0.42	2.28
	100	500	3	E.divisive	0.04	0.92	0.04	0.00	1.00
	100	500	3	HDcp	0.06	0.72	0.17	0.05	1.21
	100	500	3	Inspect	0.05	0.51	0.37	0.07	1.46
	100	500	3	MST	0.27	0.67	0.06	0.00	0.79
	100	1000	3	DCCP-L_1	0.00	0.02	0.25	0.73	2.71
	100	1000	3	DCCP-L_2	0.00	0.01	0.25	0.74	2.73
	100	1000	3	E.divisive	0.00	0.90	0.08	0.02	1.08
	100	1000	3	HDcp	0.00	0.56	0.34	0.10	1.54
	100	1000	3	Inspect	0.00	0.43	0.44	0.13	1.70
	100	1000	3	MST	0.15	0.68	0.16	0.01	1.03
	100	2000	3	DCCP-L_1	0.00	0.00	0.04	0.96	2.96
	100	2000	3	DCCP-L_2	0.00	0.00	0.02	0.98	2.98
	100	2000	3	E.divisive	0.00	0.85	0.08	0.07	1.22
	100	2000	3	HDcp	0.00	0.36	0.53	0.11	1.75
	100	2000	3	Inspect	0.00	0.35	0.51	0.14	1.79
	100	2000	3	MST	0.10	0.52	0.30	0.08	1.36
same mean, different variance $\boldsymbol{\Sigma}_1 = 0.7\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$, $\boldsymbol{\Sigma}_3 = 1.3\boldsymbol{V}_p$, $\boldsymbol{\Sigma}_4 = 1.5\boldsymbol{V}_p$	100	500	3	DCCP-L_1	0.02	0.20	0.63	0.15	1.91
	100	500	3	DCCP-L_2	0.01	0.22	0.55	0.22	1.98
	100	500	3	E.divisive	0.62	0.38	0.00	0.00	0.38
	100	500	3	HDcp	0.98	0.02	0.00	0.00	0.02
	100	500	3	Inspect	0.99	0.01	0.00	0.00	0.01
	100	500	3	MST	0.99	0.01	0.00	0.00	0.01
	100	1000	3	DCCP-L_1	0.00	0.03	0.54	0.43	2.40
	100	1000	3	DCCP-L_2	0.00	0.02	0.52	0.46	2.44
	100	1000	3	E.divisive	0.38	0.62	0.00	0.00	0.62
	100	1000	3	HDcp	0.99	0.01	0.00	0.00	0.01
	100	1000	3	Inspect	0.98	0.02	0.00	0.00	0.02
	100	1000	3	MST	0.98	0.02	0.00	0.00	0.02
	100	2000	3	DCCP-L_1	0.00	0.00	0.25	0.75	2.75
	100	2000	3	DCCP-L_2	0.00	0.00	0.26	0.74	2.74
	100	2000	3	E.divisive	0.31	0.61	0.08	0.00	0.77
	100	2000	3	HDcp	0.99	0.01	0.00	0.00	0.01
	100	2000	3	Inspect	0.97	0.03	0.00	0.00	0.03
	100	2000	3	MST	0.99	0.01	0.00	0.00	0.01

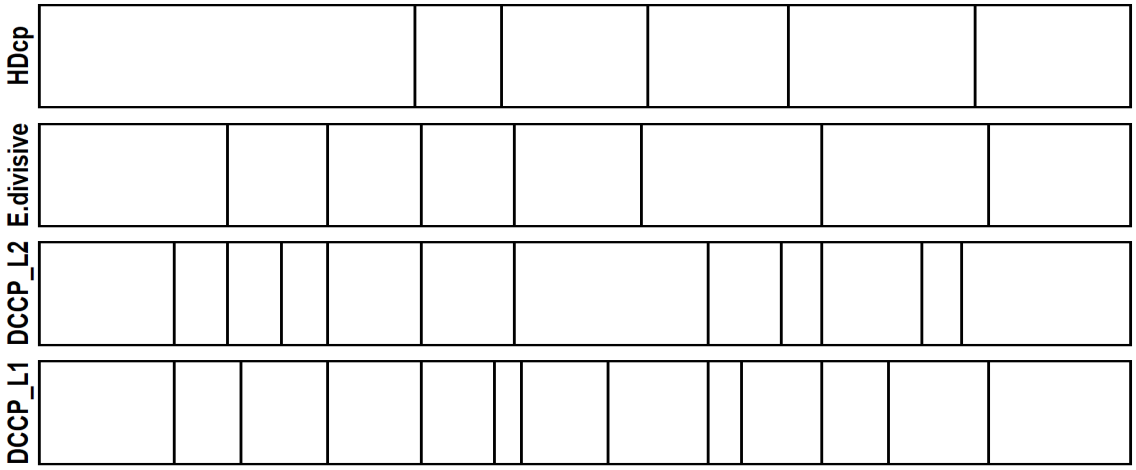
Table 5.3: Average computation time, over 200 replications, for our methods DCCP-L_1 and DCCP-L_2 in detecting three change points with changes in the mean of observations where $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, and $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$.

n	p	S	Number of true change points	*Computation time (in seconds)*	
				DCCP-L_1	DCCP-L_2
100	500	200	3	70.96	74.88
100	1000	200	3	71.18	69.98
100	2000	200	3	78.32	77.21
100	500	500	3	172.41	174.98
100	1000	500	3	173.30	171.77
100	2000	500	3	195.09	192.46

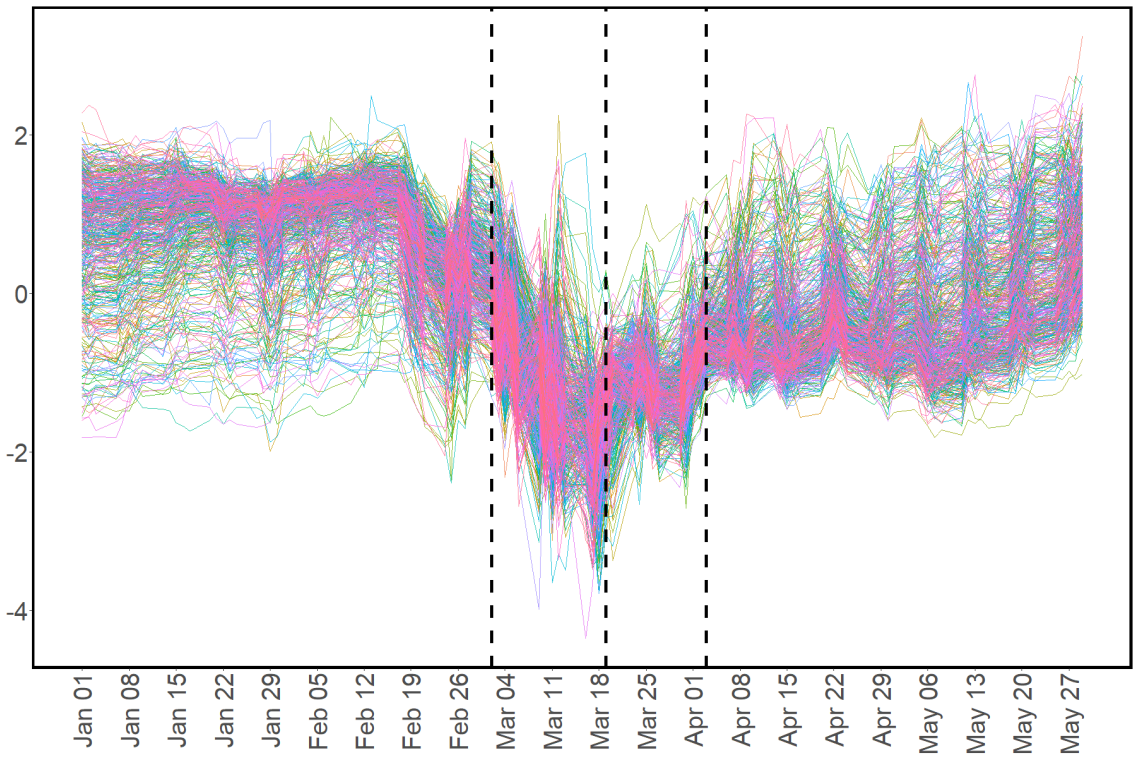
period is due to its high relevance in illustrating the impact of some key events on the US stock market during COVID-19. These events are listed and explained in Table 5.4.

We here apply our CUSUM methods DCCP-L_1 and DCCP-L_2, with the minimum segment size of 10, to this dataset to assess their performance in detecting the resultant market fluctuations. We also include E.divisive and HDcp for comparisons. Figure 5.4 (the bottom plot) visualizes the standardized stock prices of 496 companies over $n = 108$ trading days. The significant change points detected by each of the four methods are shown at the top of Figure 5.4. To explain the results, DCCP-L_1 identified twelve significant change points at locations 10, 18, 27, 37, 45, 49, 58, 68, 72, 80, 88, and 99. Also, DCCP-L_2 returned eleven significant change points at locations 10, 15, 21, 27, 37, 48, 68, 76, 80, 91, and 95. This is while E.divisive, based on the minimum segment size of 10, detected seven significant change points at locations 15, 27, 37, 48, 67, 80, and 99. Also, HDcp identified five change points at locations 36, 47, 65, 78, and 98. Comparing the detected change points by each method against the COVID-19 events of Table 5.4 highlighted by three vertical dashed lines in Figure 5.4, we can see that our method DCCP-L_1 correctly captured all those events in Table 5.4.

For the S&P 500 data, we also calculate and report confidence intervals for the detected change point locations using our methods DCCP-L_1 and DCCP-L_2. Specifically, we apply the confidence interval formula (3.9) in Section 3.3, with 1000 permutation samples. Figure 5.5 presents the 95% confidence intervals for the



(a) Significant change points detected by DCCP-L_1, DCCP-L_2, E.divisive, and HDcp.



(b) The COVID-19 events in Table 5.4 are highlighted by vertical dashed lines. The x-axis shows the trading days, and the y-axis shows the standardized stock prices.

Figure 5.4: Standardized stock prices of 496 companies from 2020-01-01 to 2020-05-29 are shown in subfigure (b), and subfigure (a) shows the significant change points detected by the four methods.

Table 5.4: List of events affecting the US stock market at the early stage of COVID-19.

Time period	Event
2020-03-02	The S&P 500 index experienced significant fluctuations and entered a bear market by approximately 20% decline (Statista Research Department, 2022).
2020-03-19	Stock markets fell after COVID-19, with investor fears over the economic growth impact (Drikvandi and Modarres, 2025).
2020-04-03	The S&P 500 index commenced its recovery phase in early April (Statista Research Department, 2022).

detected change point locations with both of our methods, DCCP-L_1 and DCCP-L_2. The confidence intervals seem to have a reasonable length.

5.4.2 Application II: MIT cellphone data

As our second data application, we apply our CUSUM method to analyze the Massachusetts Institute of Technology (MIT) cellphone data, which concerns human interactions through cellphone activities (Eagle and Pentland, 2006). The dataset can be accessed via the website at http://realitycommons.media.mit.edu/realitymining.html. It encompasses records from 96 participants, including students and staff, who used mobile phones with pre-installed software to record call logs from 2004-09-15 to 2005-05-04. This period, $n = 232$ days, includes the winter and spring vacations according to the MIT 2004-2005 academic calendar. To analyze this dataset, we first construct a 96×96 matrix to measure the cellphone activities of all individual pairs for each of 232 days. In particular, each matrix entry (i, j) is set to 1 if at least one phone call is recorded between individuals i and j on that day. Otherwise, it is set to 0. Note that each matrix is an adjacency matrix. We consider the upper triangle elements of each adjacent matrix as a 4560-dimensional vector, representing all the unique pairs of individuals' cellphone activities. Hence, we obtain a 232×4560 high dimensional data matrix. The winter and spring vacation dates are listed in Table 5.5, alongside some other potentially important events.

We apply our methods DCCP-L_1 and DCCP-L_2 to this data matrix and compare the results with E.divisive and HDcp. Using the minimum segment size of 10, DCCP-

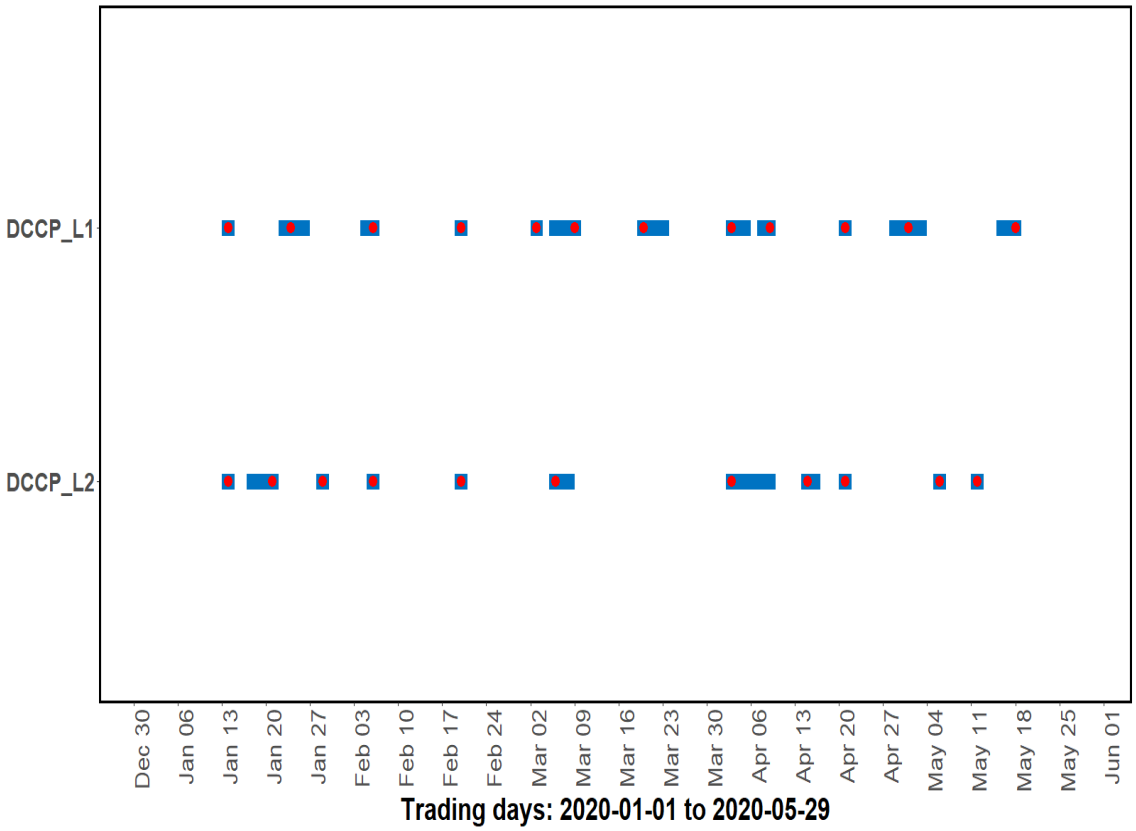


Figure 5.5: S&P 500 data: the intervals in blue show the 95% confidence intervals for the detected change points using the proposed methods DCCP-L_1 and DCCP-L_2. The estimated change points are also highlighted with dots in red.

Table 5.5: List of events potentially affecting individuals' cellphone activities, provided by MIT Registrar's office 2004-2005 academic calendar.

Time	Events
2004-10-21	Sponsor meeting.
2004-12-17	Winter vacation.
2004-12-24	Christmas holiday.
2005-01-01	New Year holiday.
2005-03-20	Spring vacation.
2005-04-16	Tax day.

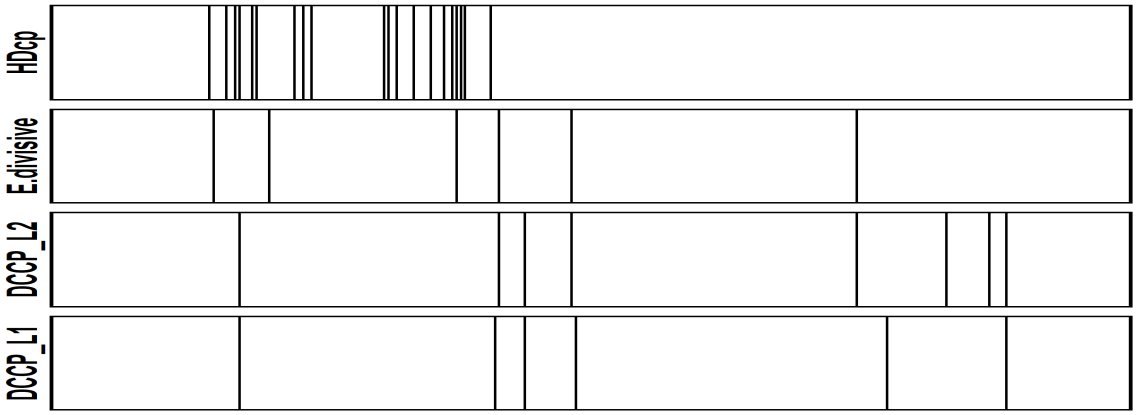
L_1 returns six significant change points at locations 34, 94, 101, 113, 186, and 214, while DCCP-L_2 returns eight significant change points at locations 34, 95, 101, 112, 179, 200, 210, and 214. The results of E.divisive and HDcp can also be seen in Figure 5.6. Comparing the detected change points by each method in Figure 5.6 against the events of Table 5.5 highlighted by vertical dashed lines, it can be seen that our methods, both DCCP-L_1 and DCCP-L_2, correctly captured almost all these relevant events in Table 5.5, especially the Christmas holiday and Spring vacation. We note that E.divisive returns a few similar change points to DCCP-L_2, while HDcp detects more change points before the winter holiday but returns no significant change points afterwards.

For the MIT cellphone data, we also calculate and report confidence intervals for the detected change point locations using our methods DCCP-L_1 and DCCP-L_2. Specifically, we apply the confidence interval formula (3.9) in Section 3.3, with 1000 permutation samples. Figure 5.7 presents the 95% confidence intervals for the detected change point locations with both of our methods DCCP-L_1 and DCCP-L_2. The confidence intervals for the first change point estimated by DCCP-L_1 and DCCP-L_2 are broader than those for the other change points, as more observations can be permuted when the next change point is relatively far. The remaining intervals are narrower and show consistent lengths.

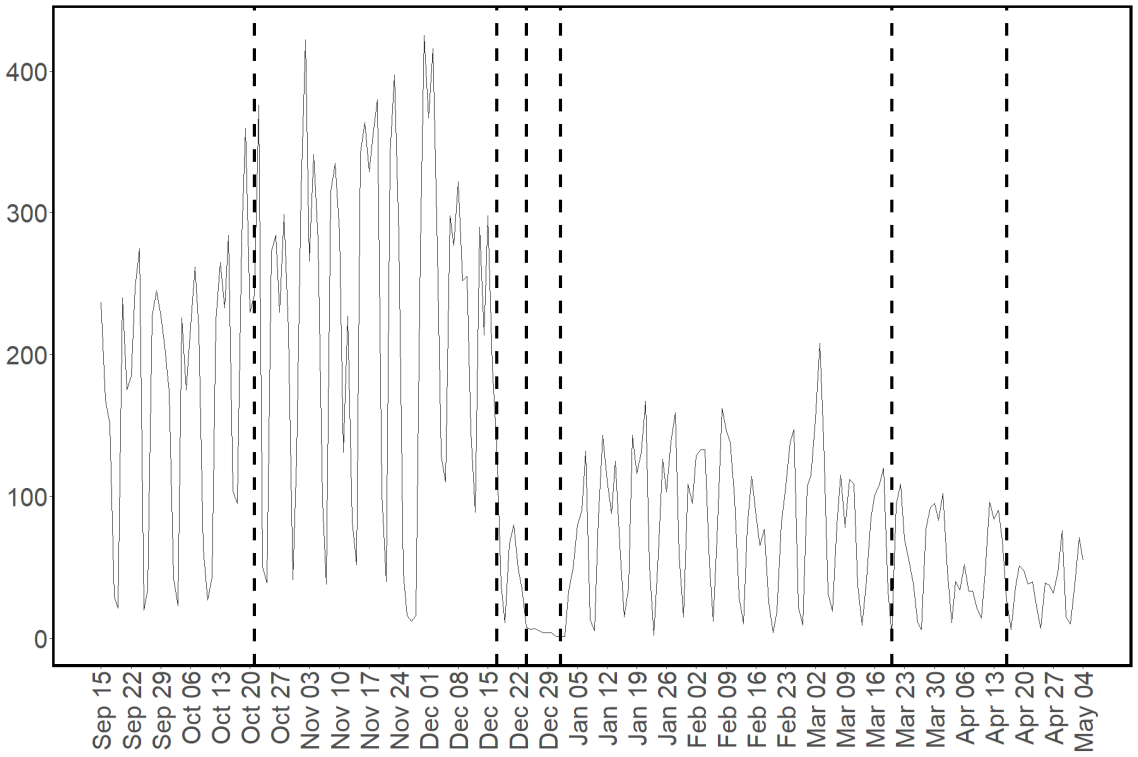
5.5 Wild binary segmentation and PELT

As discussed in Section 2.1, the RBS procedure is widely used for multiple change point detection because of its computational efficiency and simplicity. In this section, we also incorporate WBS (Fryzlewicz, 2014) as an alternative segmentation scheme in our method. To apply this procedure in our approach, we randomly draw W sub-intervals $\{1 \leq w \leq W : s_m \leq s_m^w < e_m^w \leq e_m\}$ from the sequence $[\boldsymbol{X}_{s_m}, \boldsymbol{X}_{s_m+1}, \ldots, \boldsymbol{X}_{e_m}]$. For each sub-interval $[\boldsymbol{X}_{s_m^w}, \boldsymbol{X}_{s_m^w+1}, \ldots, \boldsymbol{X}_{e_m^w}]$, using CUSUM formula in (5.2), we obtain the change point estimate $\hat{\gamma}_m^w$ as follows

$$\hat{\gamma}_m^w = \underset{s_m^w \leq k \leq e_m^w - 1}{\arg\max} \ \frac{1}{e_m^w - s_m^w + 1} \sum_{i=s_m^w}^{e_m^w} C_i^2\big(k, s_m^w, e_m^w\big).$$



(a) Significant change points detected by DCCP-L_1, DCCP-L_2, E.divisive, and HDcp.



(b) The MIT cellphone events in Table 5.5 are highlighted by vertical dashed lines. The x-axis shows the days, and the y-axis shows the sum of the daily cellphone activities.

Figure 5.6: Sum of the daily cellphone activities of all individuals from 2004-09-15 to 2005-05-04 are shown in subfigure (b), and subfigure (a) shows the significant change points detected by the four methods.

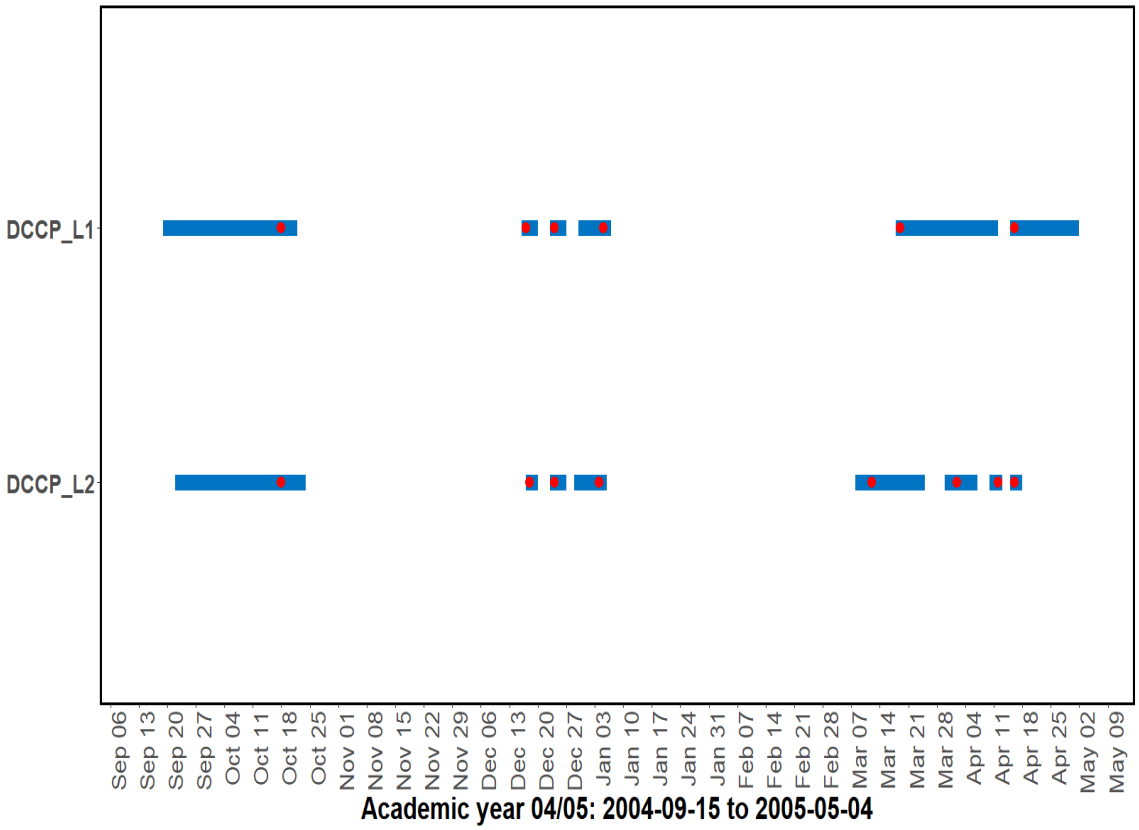


Figure 5.7: MIT cellphone data: the intervals in blue show the 95% confidence intervals for the detected change points using the proposed methods DCCP-L_1 and DCCP-L_2. The estimated change points are also highlighted with the dots in red.

After repeating this calculation, for all W randomly drawn sub-intervals we obtain a collection of change point estimates $\{\hat{\gamma}_m^1, \ldots, \hat{\gamma}_m^W\}$ and their corresponding test statistics $\{T_{s_m^1, e_m^1}(\hat{\gamma}_m^1), \ldots, T_{s_m^W, e_m^W}(\hat{\gamma}_m^W)\}$ using (5.3). The change point estimate $\hat{\gamma}_m$ under the WBS procedure is chosen as the candidate with the largest test statistic across all W sub-intervals as follows

$$\hat{\gamma}_m = \underset{\{\hat{\gamma}_m^1, \ldots, \hat{\gamma}_m^W\}}{\arg\max} \left\{ T_{s_m^1, e_m^1}(\hat{\gamma}_m^1), \ldots, T_{s_m^W, e_m^W}(\hat{\gamma}_m^W) \right\}.$$

If the detected change point is significant, we split the data before and after $\hat{\gamma}_m$ and repeat the procedure recursively until no further significant change point is found. We note that by replacing "Step 1" in Algorithm 2 with the above calculations, Algorithm 2 can be updated to the WBS approach.

In our experiments, WBS was more computationally expensive, while it did not necessarily improve much over RBS in our simulation. Despite Fryzlewicz (2014) showed that it performs as well as RBS, their results are based on univariate data with the classic CUSUM statistic. Our distance-based CUSUM differs from that setting, and moreover it is very computationally intensive to compute test statistics over many random sub-intervals in high dimensional data. For these reasons, we prefer RBS as the default segmentation scheme for DCCP. As shown in Subsection 5.3.2, DCCP still performs well with only a slight drop when the minimum spacing assumption is violated.

Besides segmentation-based approaches like RBS and WBS, another optimization-based approach called the pruned exact linear time (PELT) method (Killick et al., 2012) has also been proposed, which we briefly explain here. The PELT method detects multiple change points by minimizing a penalized cost function over their number and locations. Consider an ordered univariate sequence $X_{1:n} = \{X_1, \ldots, X_n\}$ with z unknown change points at $1 \leq \tau_1 < \cdots < \tau_z \leq n-1$. The PELT algorithm solves

$$\min_{\{\tau_1, \ldots, \tau_z\}} \left\{ \sum_{j=1}^{z+1} \text{Cost}\big(X_{(\tau_{j-1}+1):\tau_j}\big) + \beta f(z) \right\},$$

where $\text{Cost}(\cdot)$ is a cost function and $\beta f(z)$ is a penalty term (for example, linear with $f(z) = z$) to control overfitting. Here we set $\tau_0 = 0$ and $\tau_{z+1} = n$ for mathematical

convention. In Killick et al. (2012), the cost corresponds to the negative maximum log-likelihood for univariate data. It has been shown that PELT is more computationally efficient than RBS in the univariate setting because it uses a pruning step in the search procedure.

To extend PELT for high dimensional change point detection, our DCCP framework can be incorporated. For p-dimensional observations $\boldsymbol{X}_{1:n} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ with $n \ll p$, the distance-based CUSUM cost can be defined as follows

$$\text{Cost}\big(\boldsymbol{X}_{(\tau_{j-1}+1):\tau_j}\big) = -\max_{\tau_{j-1} \le k \le \tau_j - 1} \frac{1}{\tau_j - \tau_{j-1}} \sum_{i=\tau_{j-1}+1}^{\tau_j} C_i^2\big(k, \tau_{j-1}+1, \tau_j\big),$$

where $C_i(k, \tau_{j-1}+1, \tau_j)$ is the distance-based CUSUM computed within the segment $[\boldsymbol{X}_{\tau_{j-1}}, \ldots, \boldsymbol{X}_{\tau_j}]$, see formula in (5.2). This idea provides a nonparametric way to integrate PELT into high dimensional change point detection without assuming a known likelihood function or distribution. Exploring this extension is a topic for further research as it requires appropriate mathematical and computational developments.

5.6 Proofs

Proof of Theorem 5.2.1

First, for the simplicity of mathematical presentation, let us define the notation

$$\overline{C}_{s_m,e_m}(k) := \frac{1}{e_m - s_m + 1} \sum_{i=s_m}^{e_m} {C_i}^2\left(k\right), \quad k = s_m, s_m+1, \ldots, e_m - 1.$$

Using this notation, we can rewrite the change point estimate used in Algorithm 2 as follows

$$\hat{\gamma}_m = \underset{s_m \le k \le e_m - 1}{\arg\max} \big\{\overline{C}_{s_m,e_m}(k)\big\}.$$

For the sake of clarity, we begin by considering the simplest case when there are two true change points τ_1^0 and τ_2^0, $1 \le \tau_1^0 < \tau_2^0 \le n-1$. From Theorem 3.4.4, the

asymptotic limit of $\overline{C}_{s_m,e_m}(k)$ as $p > n \to \infty$ can be written as follows

$$\overline{C}_{s_m,e_m}(k) = \phi(k)_{s_m,e_m} + o_P(1),$$

where $\phi(k)_{s_m,e_m}$ denotes the value of $\phi(k)$ for the segment $[\boldsymbol{X}_{s_m}, \boldsymbol{X}_{s_m+1}, \ldots, \boldsymbol{X}_{e_m}]$. Similar to the result (3.17) in the proof of Theorem 3.4.4, it is straightforward to show that

$$\begin{aligned}
&\max_{s_m \le k \le e_m - 1} \big\{\phi(k)_{s_m,e_m}\big\} \qquad (5.4)\\
&= \max_{s_m \le k \le e_m - 1} \big\{\phi(k)_{s_m,e_m} \mathbb{1}(k \neq \tau_1^0, \tau_2^0) + \phi(k)_{s_m,e_m} \mathbb{1}(k = \tau_1^0) + \phi(k)_{s_m,e_m} \mathbb{1}(k = \tau_2^0)\big\}\\
&= \max_{s_m \le k \le e_m - 1} \big\{\phi(k)_{s_m,e_m} \mathbb{1}(k = \tau_1^0) + \phi(k)_{s_m,e_m} \mathbb{1}(k = \tau_2^0)\big\},
\end{aligned}$$

see also Figure 5.1 in the main thesis. The binary segmentation in this case starts with $m = 1, s_m = 1$ and $e_m = n$, so

$$\hat{\gamma}_1 = \arg\max_{1 \le k \le n-1} \big\{\overline{C}_{1,n}(k)\big\}.$$

Using (5.4), we find as $p > n \to \infty$

$$\begin{aligned}
\hat{\gamma}_1 &= \arg\max_{1 \le k \le n-1} \big\{\overline{C}_{1,n}(k)\big\}\\
&= \arg\max_{1 \le k \le n-1} \big\{\overline{C}_{1,n}(k) \mathbb{1}(k \neq \tau_1^0, \tau_2^0) + \overline{C}_{1,n}(k) \mathbb{1}(k = \tau_1^0) + \overline{C}_{1,n}(k) \mathbb{1}(k = \tau_2^0)\big\}\\
&= \arg\max_{1 \le k \le n-1} \big\{\phi(k)_{1,n} \mathbb{1}(k \neq \tau_1^0, \tau_2^0) + \phi(k)_{1,n} \mathbb{1}(k = \tau_1^0) + \phi(k)_{1,n} \mathbb{1}(k = \tau_2^0) + o_P(1)\big\}\\
&\xrightarrow{P} \arg\max_{1 \le k \le n-1} \big\{\phi(k)_{1,n} \mathbb{1}(k \neq \tau_1^0, \tau_2^0) + \phi(k)_{1,n} \mathbb{1}(k = \tau_1^0) + \phi(k)_{1,n} \mathbb{1}(k = \tau_2^0)\big\}\\
&= \arg\max_{1 \le k \le n-1} \big\{\phi(k)_{1,n}\big\}.
\end{aligned}$$

Using (5.4), we have either $\hat{\gamma}_1 \xrightarrow{P} \tau_1^0$ or $\hat{\gamma}_1 \xrightarrow{P} \tau_2^0$, depending on which one of $\phi(k)_{1,n} \mathbb{1}(k = \tau_1^0)$ and $\phi(k)_{1,n} \mathbb{1}(k = \tau_2^0)$ is larger. For the first case, if $\phi(k)_{1,n} \mathbb{1}(k = \tau_1^0) > \phi(k)_{1,n} \mathbb{1}(k = \tau_2^0)$, we obtain $\hat{\gamma}_1 \xrightarrow{P} \tau_1^0$, and then split the data sequence into two sub-sequences: one before and including $\hat{\gamma}_1$, and one after $\hat{\gamma}_1$. Specifically, the first sub-sequence has the starting and ending indices $s_2 = 1$ and $e_2 = \hat{\gamma}_1$, while the

second sub-sequence has the starting and ending indices $s_2 = \hat{\gamma}_1 + 1$ and $e_2 = n$. As shown in our discussion after Proposition 3.4.3, it is clear that under the null hypothesis of no change point, we have $C_i(k) = o_P(1)$ as $p > n \to \infty$. Then, since $\hat{\gamma}_1 \xrightarrow{P} \tau_1^0 < \tau_2^0$, we have either

$$\hat{\gamma}_2 = \underset{1 \le k \le \hat{\gamma}_1 - 1}{\arg\max} \left\{ \overline{C}_{1,\hat{\gamma}_1}(k) \right\} = \underset{1 \le k \le \hat{\gamma}_1 - 1}{\arg\max} \left\{ o_P(1) \right\} \xrightarrow{P} \underset{1 \le k \le \hat{\gamma}_1 - 1}{\arg\max} \{0\} = \emptyset, \tag{5.5}$$

where $\emptyset$ indicates no change point found (as previously defined in the main thesis), or

$$\begin{aligned}
\hat{\gamma}_2 &= \underset{\hat{\gamma}_1 + 1 \le k \le n-1}{\arg\max} \left\{ \overline{C}_{\hat{\gamma}_1+1,n}(k) \right\} \\
&= \underset{\hat{\gamma}_1 + 1 \le k \le n-1}{\arg\max} \left\{ \overline{C}_{\hat{\gamma}_1+1,n}(k) \mathbb{1}(k \neq \tau_2^0) + \overline{C}_{\hat{\gamma}_1+1,n}(k) \mathbb{1}(k = \tau_2^0) \right\} \\
&= \underset{\hat{\gamma}_1 + 1 \le k \le n-1}{\arg\max} \left\{ \phi(k)_{\hat{\gamma}_1+1,n} \mathbb{1}(k \neq \tau_2^0) + \phi(k)_{\hat{\gamma}_1+1,n} \mathbb{1}(k = \tau_2^0) + o_P(1) \right\} \\
&\xrightarrow{P} \underset{\hat{\gamma}_1 + 1 \le k \le n-1}{\arg\max} \left\{ \phi(k)_{\hat{\gamma}_1+1,n} \mathbb{1}(k \neq \tau_2^0) + \phi(k)_{\hat{\gamma}_1+1,n} \mathbb{1}(k = \tau_2^0) \right\} \\
&= \underset{\hat{\gamma}_1 + 1 \le k \le n-1}{\arg\max} \left\{ \phi(k)_{\hat{\gamma}_1+1,n} \right\} \\
&= \tau_2^0.
\end{aligned}$$

Thus, we must have $\hat{\gamma}_2 \xrightarrow{P} \tau_2^0$. For the second case, if $\phi(k)_{1,n} \mathbb{1}(k = \tau_1^0) < \phi(k)_{1,n} \mathbb{1}(k = \tau_2^0)$, then similarly to the first case we can show that $\hat{\gamma}_1 \xrightarrow{P} \tau_2^0$ and $\hat{\gamma}_2 \xrightarrow{P} \tau_1^0$. Thus, denoting $(\hat{\tau}_1, \hat{\tau}_2) = \text{sort}\,(\hat{\gamma}_1, \hat{\gamma}_2)$, we have $\left\| (\hat{\tau}_1, \hat{\tau}_2) - (\tau_1^0, \tau_2^0) \right\|_\infty = o_P(1)$ as $p > n \to \infty$ for the simple case of two true change points.

Now suppose that there are z true change points $\tau_1^0, \tau_2^0, \ldots, \tau_z^0$, $1 \le \tau_1^0 < \tau_2^0 < \ldots < \tau_z^0 \le n - 1$. Similar to the previous results in (3.17) and (5.4), we find as $p > n \to \infty$ that

$$\begin{aligned}
&\max_{s_m \le k \le e_m - 1} \left\{ \phi(k)_{s_m,e_m} \right\} \\
&= \max_{s_m \le k \le e_m - 1} \left\{ \phi(k)_{s_m,e_m} \mathbb{1}(k \neq \tau_1^0, \ldots, \tau_z^0) + \sum_{i=1}^{z} \phi(k)_{s_m,e_m} \mathbb{1}(k = \tau_i^0) \right\} \\
&= \max_{s_m \le k \le e_m - 1} \left\{ \sum_{i=1}^{z} \phi(k)_{s_m,e_m} \mathbb{1}(k = \tau_i^0) \right\}.
\end{aligned} \tag{5.6}$$

The binary segmentation again starts with $m = 1, s_m = 1$ and $e_m = n$, so

$$\hat{\gamma}_1 = \underset{1 \le k \le n-1}{\arg\max} \big\{ \overline{C}_{1,n}(k) \big\}.$$

Using (5.6), we get as $p > n \to \infty$

$$\begin{aligned}
\hat{\gamma}_1 &= \underset{1 \le k \le n-1}{\arg\max} \big\{ \overline{C}_{1,n}(k) \big\} \\
&= \underset{1 \le k \le n-1}{\arg\max} \Big\{ \overline{C}_{1,n}(k) \mathbb{1}(k \neq \tau_1^0, \ldots, \tau_z^0) + \sum_{i=1}^{z} \overline{C}_{1,n}(k) \mathbb{1}(k = \tau_i^0) \Big\} \\
&= \underset{1 \le k \le n-1}{\arg\max} \Big\{ \phi(k)_{1,n} \mathbb{1}(k \neq \tau_1^0, \ldots, \tau_z^0) + \sum_{i=1}^{z} \phi(k)_{1,n} \mathbb{1}(k = \tau_i^0) + o_P(1) \Big\} \\
&\xrightarrow{P} \underset{1 \le k \le n-1}{\arg\max} \Big\{ \phi(k)_{1,n} \mathbb{1}(k \neq \tau_1^0, \ldots, \tau_z^0) + \sum_{i=1}^{z} \phi(k)_{1,n} \mathbb{1}(k = \tau_i^0) \Big\} \\
&= \underset{1 \le k \le n-1}{\arg\max} \Big\{ \sum_{i=1}^{z} \phi(k)_{1,n} \Big\}.
\end{aligned}$$

Hence, using (5.6), we have $\hat{\gamma}_1 \xrightarrow{P} \tau_{m_1}^0$ for a $\tau_{m_1}^0$, $m_1 \in \{1, \ldots, z\}$, for which $\phi(k)_{1,n} \mathbb{1}(k = \tau_{m_1}^0) = \max\{\phi(k)_{1,n} \mathbb{1}(k = \tau_1^0), \ldots, \phi(k)_{1,n} \mathbb{1}(k = \tau_z^0)\}$. We then split the data sequence into two sub-sequences: one before and including $\hat{\gamma}_1$, and one after $\hat{\gamma}_1$. Specifically, the first sub-sequence has the starting and ending indices $s_2 = 1$ and $e_2 = \hat{\gamma}_1$, while the second sub-sequence has the starting and ending indices $s_2 = \hat{\gamma}_1 + 1$ and $e_2 = n$. Applying the same process as above for the two estimates $\hat{\gamma}_2$ and $\hat{\gamma}_3$ that $\hat{\gamma}_2 \xrightarrow{P} \tau_{m_2}^0$ and $\hat{\gamma}_3 \xrightarrow{P} \tau_{m_3}^0$, where we have either $\tau_{m_2}^0 < \tau_{m_1}^0$ and $\tau_{m_3}^0 > \tau_{m_1}^0$ or $\tau_{m_2}^0 > \tau_{m_1}^0$ and $\tau_{m_3}^0 < \tau_{m_1}^0$. Continuing this process, we find $\|(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_z) - (\tau_1^0, \tau_2^0, \ldots, \tau_z^0)\|_\infty = o_P(1)$ as $p > n \to \infty$. We note that for the last two sub-sequences of this segmentation process, no additional change points will converge according to the calculation in (5.5) for the case of two true change points.

CHAPTER 6

Online change point detection for high dimensional data streams

In this chapter, we introduce distance-based CUSUM for online change points (DC-OCP), a nonparametric method for detecting high dimensional online change points (Zhang et al., 2025). The online change point problem differs from the offline change point problem. Offline change point analysis estimates both the number and the locations of change points, and consistency is the main goal in terms of the theory. But online change point analysis deals with sequential data streams, where the main task is fast detection while controlling the overall type I error. This is often a harder problem as the tests are multiple and sequential. In Section 6.1, we present the online change point problem setting with a sliding window technique. In Section 6.2, we introduce the DC-OCP method and propose a stopping rule based on permutation-based thresholds to terminate the sequential testing algorithm when the true change point arrives. In Section 6.3, we provide theoretical results to justify the proposed method. In Section 6.4, we assess the numerical performance of DC-OCP, validate the accuracy of derived theoretical properties, and compare it with some of the methods in the literature. In Section 6.5, we apply DC-OCP to human activity recognition data in a high dimensional setting and demonstrate its capability to detect multiple online change points. In Section 6.6, we provide the technical proofs

for all theoretical results developed in this chapter.

6.1 Online change point problem setting

Suppose a p-dimensional data stream is observed sequentially as follows

$$\underbrace{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n}_{\text{Historical data}}, \underbrace{\boldsymbol{X}_{n+1}, \boldsymbol{X}_{n+2}, \ldots, \boldsymbol{X}_{n+m}, \ldots}_{\text{Arriving data}} \qquad m = 1, 2, \ldots,$$

where the historical data is of size n and each observation $\boldsymbol{X}_i$ is p-dimensional as $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ for $i = 1, 2, \ldots, n, n+1, \ldots, n+m, \ldots$, with $n + m \ll p$. We note that the p variables are potentially correlated. We here allow historical data and note that some arriving data over time can be regarded as historical data if a change point occurs. In this framework, if there are any significant change points, we only consider the data sequence after the last detected change point as historical data. Let $F_1, F_2, \ldots, F_{n+m}, \ldots$ denote the corresponding unknown probability distributions, respectively. In line with the above setting, we assume that the historical data do not contain a change point, so the distribution of the historical data is identical, although unknown. That is, $F_1 = F_2 = \cdots = F_n$.

Let us consider the first m arriving observations. We use the index t to denote the time of these arriving observations with $t = 1, \ldots, m$. We define a data sequence of size w as most recent w observations in the form of $[\boldsymbol{X}_{n+t-w+1}, \boldsymbol{X}_{n+t-w+2}, \ldots, \boldsymbol{X}_{n+t}]$ for each time $t = 1, \ldots, m$. We aim to sequentially test if there is a change point in these m arriving observations as follows

$$\begin{cases} H_0^1 : F_{n-w+2} = \cdots = F_{n+1}, \\ H_1^1 : F_{n-w+2} = \cdots = F_{n+\tau} \neq F_{n+\tau+1} = \cdots = F_{n+1}, \end{cases}$$
$$\vdots$$
$$\begin{cases} H_0^m : F_{n+m-w+1} = \cdots = F_{n+m}, \\ H_1^m : F_{n+m-w+1} = \cdots = F_{n+\tau} \neq F_{n+\tau+1} = \cdots = F_{n+m}, \end{cases}$$

or equivalently

$$\begin{cases} H_0^t : F_{n+t-w+1} = \cdots = F_{n+t}, \\ H_1^t : F_{n+t-w+1} = \cdots = F_{n+\tau} \neq F_{n+\tau+1} = \cdots = F_{n+t}, \end{cases} \quad \text{for } t = 1, \ldots, m, \quad (6.1)$$

where $\tau \in \{0, \ldots, m-1\}$ points to a (unknown) change point location in the observations that arrive. For mathematical convenience, if there is no change point, we set $\tau = \infty$. If H_0^t is rejected at some time t, we infer that there is a change point in the arriving data and stop the detection procedure at time t. Otherwise, we continue the detection procedure until all arriving observations are tested. This procedure is for single change point detection. We will discuss the case of multiple change points later in the next section.

6.2 DC-OCP method and sequential change point detection algorithm

As noted in Section 3.1, the main challenges in high dimensional change point analysis also apply to the online change point problem, so we avoid repeating them again. The distance-based CUSUM statistic can naturally serve as a nonparametric solution for detecting high dimensional online change points. However, the sequential testing framework is different and more complex than in the offline change point problem. For example, it requires a suitable stopping rule to control the overall type I error across multiple tests. Thus, substantial effort is needed to adapt the distance-based CUSUM statistic to the online framework, which we discuss next.

To detect general types of change points in online settings, we propose distance-based CUSUM statistics to carry out the sequential testing problem in (6.1). For each test at time $t = 1, \ldots, m$, let the data sequence $[\boldsymbol{X}_{n+t-w+1}, \boldsymbol{X}_{n+t-w+2}, \ldots, \boldsymbol{X}_{n+t}]$ denote the test interval of size w. We define the following distance-based CUSUM statistic on this interval, calculated for each $i = n+t-w+1, \ldots, n+t$,

$$C_i(k;t) = \frac{h}{n+t-k} \sum_{j=k+1}^{n+t} d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) - \frac{h}{w-n-t+k} \sum_{j=n+t-w+1}^{k} d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right), \quad (6.2)$$

where $h = \sqrt{(n+t-k)(w-n-t+k)}/w$ is a constant, $k \in \{n+t-w+1, \ldots, n+t-1\}$, and $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ is a suitable pairwise distance function.

Similar to the case of offline change points, the distance-based CUSUM statistic (6.2) measures the average distance differences among all observations before and after each k. It is different than the standard CUSUM statistic, which in this case is given by $\boldsymbol{C}_{\text{standard}}(k;t) = \frac{h^*}{n+t-k}\sum_{j=k+1}^{n+t}\boldsymbol{X}_j - \frac{h^*}{w-n-t+k}\sum_{j=n+t-w+1}^{k}\boldsymbol{X}_j$, where the constant term is $h^* = \sqrt{(n+t-k)(w-n-t+k)/w}$. The proposed distance-based CUSUM statistic (6.2) uses pairwise dissimilarity distances $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ instead of just the observations $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ themselves. This provides a scalar CUSUM value, and because it is defined for each observation i in the test interval, $i = n+t-w+1, \ldots, n+t$, we obtain a sequence of w values of CUSUM for each $k \in \{n+t-w+1, \ldots, n+t-1\}$.

Remark 4. The choice of dissimilarity distance can be crucial in high dimensional regimes. For example, both the squared L_2-norm distance $\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2^2$ and the L_1-norm distance $\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_1$ suffer from convergence issues when dimension p increases (see Hall et al., 2005). To overcome the convergence issue, we can introduce a modifier $1/p$ to the L_1-norm distance, resulting in the following modified L_1-norm distance

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = p^{-1}\sum_{l=1}^{p}|X_{il} - X_{jl}|. \tag{6.3}$$

This is a similar modification to the modified L_2-norm distance $p^{-1/2}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2$ proposed by Hall et al. (2005). We here prefer to use the modified L_1-norm distance as it requires weaker assumptions and is known to show better empirical performance over the modified L_2-norm distance, as shown in Section 4.3.

To proceed with the proposed approach, by applying the distance-based CUSUM statistic (6.2), we obtain a $w \times (w-1)$ squared CUSUM matrix as follows

$$\boldsymbol{C} = \begin{bmatrix} C^2_{n+t-w+1}(n+t-w+1;t) & \cdots & C^2_{n+t-w+1}(n+t-1;t) \\ C^2_{n+t-w+2}(n+t-w+1;t) & \cdots & C^2_{n+t-w+2}(n+t-1;t) \\ \vdots & \ddots & \vdots \\ C^2_{n+t}(n+t-w+1;t) & \cdots & C^2_{n+t}(n+t-1;t) \end{bmatrix}, \tag{6.4}$$

where all the CUSUM values are squared to ensure working with non-negative values.

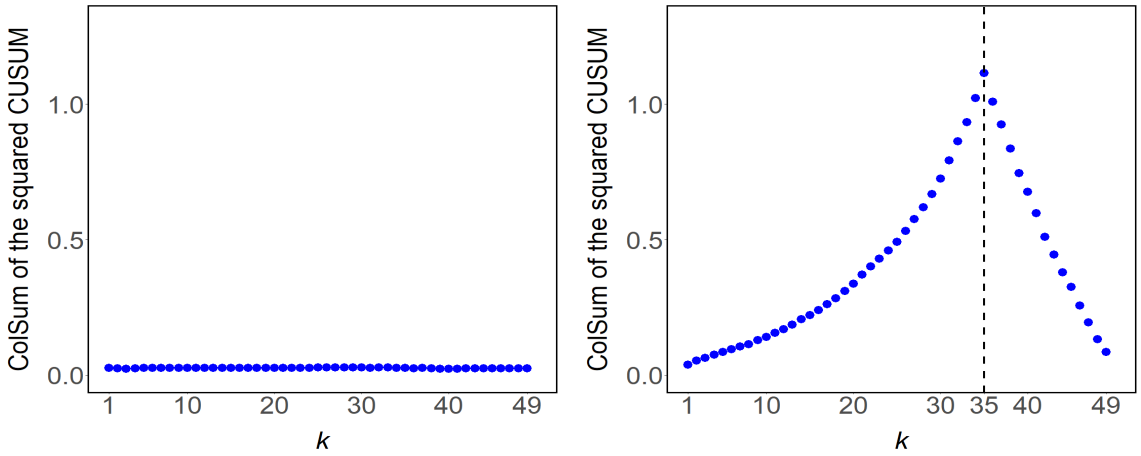
The computational cost for calculating $C_i(k;t)$ for each $i = n+t-w+1, \ldots, n+t$ is $O\big(wp\big)$, so the computational cost for calculating the CUSUM matrix (6.4) is $O\big(w^3p\big)$. Note that when w is much smaller than p as in high dimensional settings, the computational cost becomes $O(p)$. All columns of the squared CUSUM matrix (6.4) may contain information about the potential change point. To measure this information, we suggest the following test statistic using the maximum of column sums of the squared CUSUM matrix $\boldsymbol{C}$:

$$T(t) = \max_{n+t-w+1 \le k \le n+t-1} \left\{ \frac{1}{w} \sum_{i=n+t-w+1}^{n+t} C_i^2(k;t) \right\}. \tag{6.5}$$

As shown in Theorem 6.3.2 below, the test statistic (6.5) converges to 0 under the null hypothesis of no change point. Therefore, large values of the test statistic $T(t)$ suggest rejecting the null hypothesis at time t.

Example 3. To illustrate this using a simple example, Figure 6.1a and Figure 6.1b show the behavior of column sums of the squared CUSUM matrix under the null and alternative hypotheses, respectively. For Figure 6.1a, we randomly simulate 50 observations from a standard multivariate normal distribution with $p = 1000$. For Figure 6.1b, we simulate another 50 observations from the same distribution but with a mean shift of 1 for all variables in the last 15 observations, hence the true change point occurs at location 35. For a simpler visualization here, we apply distance-based CUSUM (6.2) to the entire data sequence. Figure 6.1a shows that all column sums of the squared CUSUM statistics are close to 0 under the null, hence the maximum value is also very small. In contrast, Figure 6.1b presents a different pattern, where the true change point at location 35 has the largest column sum value, and it is much greater than 0. This is due to the different asymptotic behaviors of the distance-based CUSUM statistic under the null and alternative hypotheses, as derived in Section 6.3.

As a nonparametric approach, we use a permutation-based procedure to construct thresholds for the sequential tests in (6.1). The use of a permutation procedure is helpful here because all observations are exchangeable under the null hypothesis of no distributional changes. In each permutation step, we draw w observations



(a) The behavior of column sums of the squared CUSUM matrix $\boldsymbol{C}$ when there is no change point.

(b) The behavior of column sums of the squared CUSUM matrix $\boldsymbol{C}$ when there is a true change point at location 35.

Figure 6.1: Illustrative example on behaviors of column sums of the squared CUSUM matrix $\boldsymbol{C}$ with $p = 1000$. Figure 6.1a shows the results under the null hypothesis. Figure 6.1b shows the results under the alternative hypothesis, where the true change point is highlighted by a vertical dashed line.

from the historical data and randomly permute their indices to obtain a random permutation sample under H_0^t. Using formula (6.5), we obtain a permutation test statistic based on this permutation sample. Repeating this process S times, we find an approximate permutation distribution of the test statistic $T(t)$ under H_0^t as

$$G_{T^S_{\text{perm}}}(u) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\left(T^s_{\text{perm}} \leq u\right) \quad \forall\, u \in \mathbb{R}^+, \tag{6.6}$$

where T^s_{perm} denotes the test statistic calculated for the s-th permutation sample and $\mathbb{1}(\cdot)$ is the indicator function. For the sequential testing problem (6.1), it is important to control the family-wise error rate of the m sequential tests, which can be challenging to achieve. We propose the following threshold to control the FWER using the Bonferroni correction

$$T_r^{\text{BC}} := \inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\text{perm}}}(u) \geq 1 - \alpha/m\right\}, \tag{6.7}$$

where $\alpha \in [0, 1]$ is the nominal significance level. If the value of the test statistic $T(t)$ exceeds the proposed threshold T_r^{BC}, we reject the null hypothesis at time t,

indicating a change point.

The Bonferroni correction is suitable for small or moderate m, but when m is large, the value of T_r^{BC} increases, so rejecting the null hypothesis would be more difficult. To accommodate large-scale data streams, we suggest an alternative threshold based on controlling the average run length (ARL) as follows

$$T_r^{\mathrm{ARL}} := \inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\mathrm{perm}}}(u) \geq 1 - \alpha_{\mathrm{ARL}}\right\}, \tag{6.8}$$

where $\alpha_{\mathrm{ARL}} \in (0, 1)$ is a parameter to control the ARL. The idea of developing a manageable ARL is widely used to control false alarms in online change point detection. Unlike T_r^{BC}, the ARL-based threshold T_r^{ARL} remains fixed with respect to m; however, it is not designed to control FWER (see Table 6.3 in our simulation results). To combine their properties, we propose an additional threshold defined based on the average of these two thresholds as

$$T_r^* = \frac{T_r^{\mathrm{BC}} + T_r^{\mathrm{ARL}}}{2}. \tag{6.9}$$

Using either of the proposed thresholds, we define a stopping rule t_{stop} as the first time t that the test statistic $T(t)$ exceeds a chosen threshold $T_r \in \left\{T_r^{\mathrm{BC}}, T_r^{\mathrm{ARL}}, T_r^*\right\}$ as follows

$$t_{\mathrm{stop}} = \min\left\{1 \leq t \leq m : T(t) > T_r\right\}. \tag{6.10}$$

The choice of threshold T_r should depend on the application and the size of arriving observations to be tested. Based on our simulation results in Section 6.4, we recommend using T_r^{BC} for data streams with small m, T_r^{ARL} for data streams with large m, and T_r^* for data streams with moderate m. We note that the use of empirical thresholds is common in online change point analysis, including Monte Carlo-based thresholds (e.g., Mei, 2010; Chen et al., 2022) and bootstrap calibration schemes (e.g., Avanesov and Buzun, 2018; Gösmann et al., 2022), among others.

Algorithm 3 summarizes our proposed method, which applies the distance-based CUSUM (6.2) with the proposed stopping rule (6.10) to detect an online change point in high dimensional data streams. Unlike offline change point detection which

typically estimates the change point location, Algorithm 3 stops immediately once a significant change point is detected, making it efficient for sequential testing.

Algorithm 3: The DC-OCP method

Input: Historical data $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and m arriving observations $\{\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_{n+m}\}$.

Output: Stopping time t, or "NA" if there is no change point in arriving observations.

Compute threshold T_r using either of the formulas (6.7), (6.8) and (6.9);

$t \leftarrow 1$;

while $t \leq m$ **do**

  $\boldsymbol{X} \leftarrow [\boldsymbol{X}_{n+t-w+1}, \boldsymbol{X}_{n+t-w+2}, \ldots, \boldsymbol{X}_{n+t}]^\top$;

  $\boldsymbol{D} \leftarrow [d(\boldsymbol{X}_i, \boldsymbol{X}_j)]_{i=n+t-w+1}^{n+t} {}_{j=n+t-w+1}^{n+t}$;

  $\boldsymbol{C} \leftarrow [C_i^2(k;t)]_{i=n+t-w+1}^{n+t} {}_{k=n+t-w+1}^{n+t-1}$;

  $T(t) \leftarrow \max_{n+t-w+1 \leq k \leq n+t-1} \left\{ \frac{1}{w} \sum_{i=n+t-w+1}^{n+t} C_i^2(k;t) \right\}$;

  if $T(t) > T_r$ **then**

    return t

  end

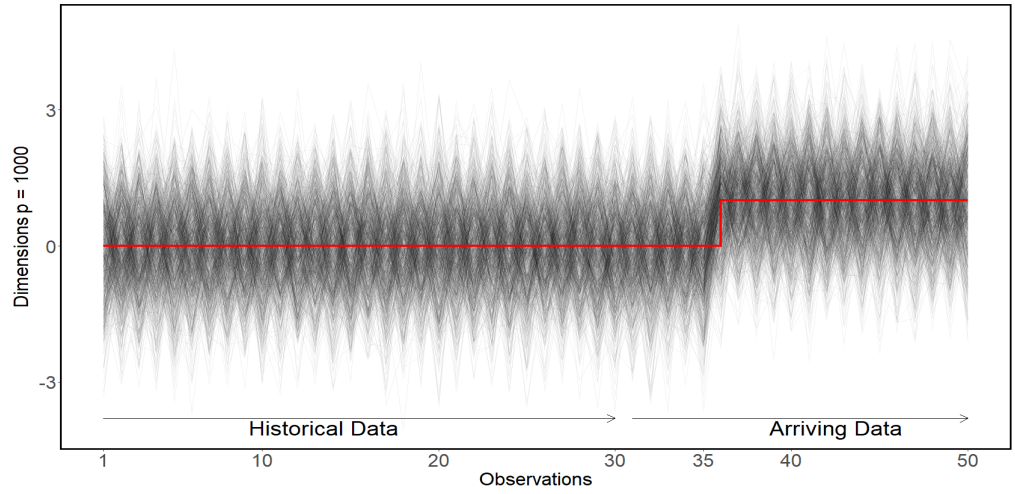
  $t \leftarrow t + 1$;

end

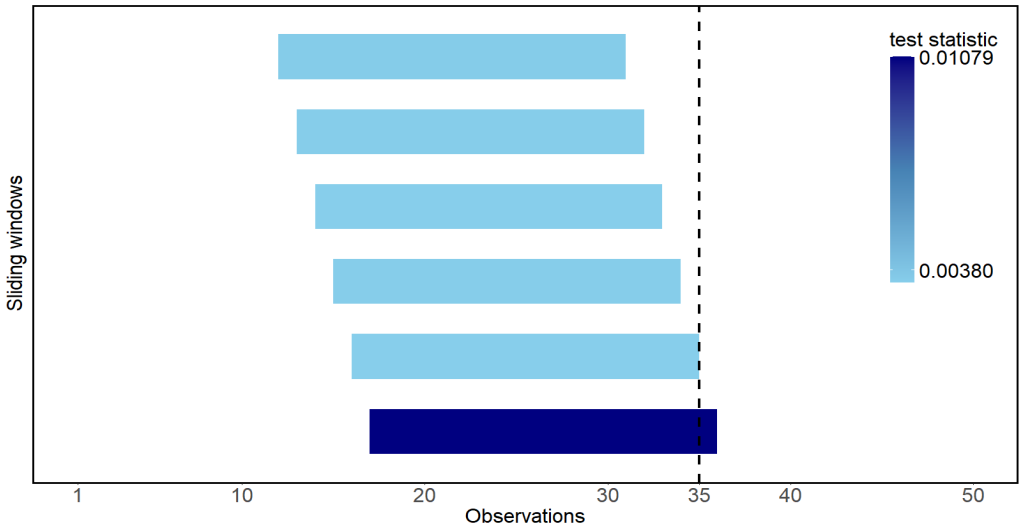
return "NA"

Our Algorithm 3 can be straightforwardly extended to detect multiple online change points. This can be achieved by considering some observations after a new detected change point as historical data and applying the algorithm repeatedly until the data stream is tested in real time (see our real data application in Section 6.5). We note that this requires the assumption that the minimum spacing between consecutive change points is not too small, which is a typical assumption in online change points (see, e.g., Yu et al., 2023).

Example 4. Figure 6.2 illustrates Algorithm 3 for the simulation example in Figure 6.1b with the last 20 observations simulated sequentially. It shows how the algorithm detects a high dimensional online change point ($p = 1000$) using a sliding window approach, and terminates immediately after the time that a true change point occurs (i.e., $n + \tau_0 = 35$).



(a) High dimensional observations containing 30 historical data ($n = 30$), 20 arriving data ($m = 20$), with $p = 1000$. The red lines show the true mean of observations in the simulation.



(b) Algorithm 3 begins at the top with the first test interval $[\boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{31}]$ (here $w = 20$) when the first arriving observation ($\boldsymbol{X}_{31}$) arrives, and it terminates at the bottom with the last test interval $[\boldsymbol{X}_{17}, \ldots, \boldsymbol{X}_{36}]$ where the corresponding test statistic (0.01079) exceeds the threshold (0.0038). Here, the darker blue color indicates larger values of the test statistic.

Figure 6.2: Illustrative example of Algorithm 3 for detecting a true change point with window size $w = 20$ and a pre-computed threshold $T_r^{\mathrm{BC}} = 0.0038$. The time at which the true change point occurs (i.e., $n + \tau_0 = 35$) is highlighted by the vertical dashed line in Figure 6.2b.

6.3 Theoretical results

In this section, we study the asymptotic behaviors of the distance-based CUSUM statistic in the sequential testing framework (6.1) when $p \to \infty$. We also establish theoretical guarantees for the proposed stopping rule under both the null hypothesis and the alternative hypothesis. In particular, we show that the stopping rule controls the FWER and ARL under the null hypothesis of no change point. Under the alternative, we prove that the test is asymptotically consistent as $p \to \infty$, or as both p and m grow large, under some conditions. In the same asymptotic regime, we show that the test also achieves low detection delays.

The asymptotic behavior of the proposed distance-based CUSUM (6.2) depends on the asymptotic limit of the distance function $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$. To develop the theory, we define the following scalar term

$$\lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} := \sqrt{\mathbb{E}\big(d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)\big)} \qquad \forall\, i \neq j, \tag{6.11}$$

which turns out to appear in the asymptotic limit of $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$, as we will see later. Suppose there is a true change point in the m arriving observations and let τ_0 denote the (unknown) true change point such that $F_1 = \cdots = F_{n+\tau_0} \neq F_{n+\tau_0+1} = \cdots = F_{n+m}$ with $\tau_0 \in \{0, \ldots, m-1\}$. We define $A^-_{\tau_0}$ as a set of indices for pre-change observations and $A^+_{\tau_0}$ as a set of indices for post-change observations as follows

$$A^-_{\tau_0} := \{1, \ldots, n+\tau_0\}, \quad A^+_{\tau_0} := \{n+\tau_0+1, \ldots, n+m\}.$$

We can then write for $i \neq j$ that

$$\begin{aligned}
\lambda_{A^-_{\tau_0} A^-_{\tau_0}} &:= \lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} \quad \forall\, i, j \in A^-_{\tau_0}, \\
\lambda_{A^+_{\tau_0} A^+_{\tau_0}} &:= \lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} \quad \forall\, i, j \in A^+_{\tau_0}, \\
\lambda_{A^-_{\tau_0} A^+_{\tau_0}} &:= \lambda_{\boldsymbol{X}_i \boldsymbol{X}_j} \quad \forall\, i \in A^-_{\tau_0}, j \in A^+_{\tau_0} \text{ or } \forall\, i \in A^+_{\tau_0}, j \in A^-_{\tau_0}.
\end{aligned}$$

Since we focus on the modified L_1-norm distance $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ in (6.3), we recall the following assumptions:

(A1) Assume $\max_{1\le i\le n+m}\max_{1\le l\le p} E(X_{il}^2) < \infty$,

(A2) Assume $\sum_{l=1}^{p}\sum_{l'=1}^{p} \mathrm{Cov}\left(|X_{il} - X_{jl}|, |X_{il'} - X_{jl'}|\right) = o(p^2)$ as $p \to \infty$, where $l \neq l'$.

With these assumptions, it is evident that $\lambda_{\boldsymbol{X}_i\boldsymbol{X}_j} < \infty$ for $1 \le i, j \le n+m$. Assumption (A1) says that the second moment of random variables X_{il} is uniformly bounded. Assumption (A2) imposes the weak dependence among the random variables, which would be trivial in case the variables are independent. Note that under assumptions (A1) and (A2), the WLLN holds on the sequence $\{|X_{il} - X_{jl}| : 1 \le l \le p\}$. Thus, we have $p^{-1}\big(\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_1 - E\left(\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_1\right)\big) \xrightarrow{P} 0$ as $p \to \infty$. Using the results in Theorem 3.4.2, we have the following lemma shows the asymptotic convergence of the modified L_1-norm distance $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ as $p \to \infty$.

Lemma 6.3.1. *Suppose there is a true change point τ_0 in m arriving observations such that $F_1 = \cdots = F_{n+\tau_0} \neq F_{n+\tau_0+1} = \cdots = F_{n+m}$ with $\tau_0 \in \{0, \ldots, m-1\}$. Under assumptions (A1) - (A2), we have as $p \to \infty$*

$$d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = \begin{cases} \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1) & \forall\, i, j \in A^-_{\tau_0}, \\ \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} + o_P(1) & \forall\, i, j \in A^+_{\tau_0}, \\ \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} + o_P(1) & \forall\, i \in A^-_{\tau_0}, j \in A^+_{\tau_0} \text{ or } \forall\, i \in A^+_{\tau_0}, j \in A^-_{\tau_0}. \end{cases}$$

The proof follows the same argument as Theorem 3.4.2 and therefore is omitted. We derive the asymptotic limit of the proposed test statistic $T(t)$ in (6.5) when $p \to \infty$ under the sequential testing framework (6.1).

Theorem 6.3.2. *Suppose that there are m arriving p-dimensional observations. For each sequential test at time $t = 1, \ldots, m$, consider the test statistic $T(t)$ in (6.5) which is computed based on the test interval $[\boldsymbol{X}_{n+t-w+1}, \boldsymbol{X}_{n+t-w+2}, \ldots, \boldsymbol{X}_{n+t}]$. Under assumptions (A1) - (A2), the following results hold as $p \to \infty$*

(a) If there is a true change point τ_0 in m arriving observations such that $F_1 =$

$\cdots = F_{n+\tau_0} \neq F_{n+\tau_0+1} = \cdots = F_{n+m}$ *with* $\tau_0 \in \{0, \ldots, m-1\}$, *then*

$$\begin{cases} T(t)|_{H_1} = \Psi(t) + o_P(1) & \text{for } t > \tau_0 \\ T(t)|_{H_1} = o_P(1) & \text{for } t \leq \tau_0, \end{cases}$$

in which $\Psi(t) = \frac{(t-\tau_0)(w-(t-\tau_0))^2}{w^3}\left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\right)^2 + \frac{(t-\tau_0)^2(w-(t-\tau_0))}{w^3}\left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\right)^2$.

(b) If there is no change point in m *arriving observations, then* $T(t)|_{H_0} = o_P(1)$ *for* $t = 1, \ldots, m$.

Let us use the notations $\mathbb{P}_\infty(\cdot)$ and $\mathbb{E}_\infty(\cdot)$ for the probability and expectation under the null hypothesis of no change point (i.e., $\tau = \infty$), respectively. We first study the theoretical properties of the proposed stopping rule under the null hypothesis. We begin with the following lemma showing that the permutation distribution of the test statistic $T(t)$ is asymptotically equivalent to the (unknown) true distribution under the null hypothesis.

Lemma 6.3.3. *Suppose that there are* m *arriving observations with no true change point. Let* $G_{T_{H_0}}(u)$ *denote the distribution of the test statistic* $T(t)$ *under the null hypothesis. We have as* $S \to \infty$ *that*

(a) $G_{T^S_{perm}}(u) \xrightarrow{P} G_{T_{H_0}}(u) \quad \forall\, u \in \mathbb{R}^+$,

(b) $\mathbb{P}_\infty\left(T(t) > T_r^{BC}\right) \xrightarrow{P} \frac{\alpha}{m} \quad \forall\, \alpha \in [0,1]$, for $t = 1, \ldots, m$.

Using the results in Lemma 6.3.3, we establish the following theorem which shows that the threshold T_r^{BC} in (6.7) controls the FWER of sequential tests.

Theorem 6.3.4. *Under the null hypothesis of no change point, consider the threshold* T_r^{BC} *in* (6.7). *The probability of the first false detection among the* m *arriving observations is bounded above by a nominal level* $\alpha \in [0,1]$ *as follows*

$$\mathbb{P}_\infty\left(\min\left\{1 \leq t \leq m : T(t) > T_r^{BC}\right\} \leq m\right) \leq \alpha.$$

We next derive a theoretical lower bound for ARL using the threshold T_r^{ARL} in (6.8).

Theorem 6.3.5. *Consider the conditions in Theorem 6.3.4 and the threshold* T_r^{ARL} *in* (6.8). *The expectation of the stopping time using this threshold attains the following lower bound*

$$\mathbb{E}_\infty \left(\min\left\{t \in \mathbb{N} : T(t) > T_r^{ARL}\right\}\right) \geq \frac{1}{\alpha_{ARL}}.$$

We now study the theoretical properties under the alternative hypothesis, where a change point τ_0 occurs in the m arriving observations. Let $\mathbb{P}_{\tau_0}(\cdot)$ and $\mathbb{E}_{\tau_0}(\cdot)$ denote the probability and expectation under the alternative hypothesis, respectively. We make the following additional assumption.

(A3) Consider the conditions in part (a) of Theorem 6.3.2 where $t > \tau_0$. Define $\Delta_{\min} := \left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\right)^2 \wedge \left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\right)^2$. Assume $\Delta_{\min} > 0$ as $p \to \infty$.

The detection power is defined as the probability that a change point is detected after it occurs (Tartakovsky et al., 2014). The following theorem establishes the consistency of our proposed test as $p \to \infty$, under the above conditions.

Theorem 6.3.6. *Consider the alternative hypothesis that there is a true change point* $\tau_0 \in \{0, \ldots, m-1\}$ *in the* m *arriving observations such that* $F_1 = \cdots = F_{n+\tau_0} \neq F_{n+\tau_0+1} = F_{n+m}$. *Under assumptions (A1) - (A3), we have as* $p \to \infty$

$$\sup_{0 \leq \tau_0 < m} \left|\mathbb{P}_{\tau_0}\left(T(t) > T_r\right) - 1\right| \to 0 \quad \text{for } t > \tau_0.$$

Under the alternative hypothesis that there is a true unknown change point $\tau_0 \in \{0, \ldots, m-1\}$, the expected detection delay is defined as

$$\mathrm{EDD}(\tau_0) = \mathbb{E}_{\tau_0}\left(\min\{1 \leq t \leq m : T(t) > T_r\} - \tau_0 \,|\, \min\{1 \leq t \leq m : T(t) > T_r\} > \tau_0\right).$$

The literature mostly considers the expected detection delay at the so-called immediate change point, since the supremum over all expected detection delays is often achieved at this point (see also Siegmund and Venkatraman, 1995; Xie and Siegmund, 2013). In our setting, the immediate change point refers to a change that occurs immediately after historical observations, that is, at $\tau_0 = 0$. We therefore expect that

the supremum of EDD is given by $\mathbb{E}_{\tau_0=0}\left(\min\{1 \leq t \leq m : T(t) > T_r\}\right)$, as shown in the following theorem, which establishes the convergence of the expected detection delay as $p \to \infty$.

Theorem 6.3.7. *Consider the conditions in Theorem 6.3.6 with a true change point* $\tau_0 \in \{0, \ldots, m-1\}$. *We have*

(a) The supremum of all expected detection delays is achieved at the immediate true change point $\tau_0 = 0$ *as*

$$\sup_{0 \leq \tau_0 < m} \mathrm{EDD}(\tau_0) = \mathbb{E}_{\tau_0=0}(\min\{1 \leq t \leq m : T(t) > T_r\}).$$

(b) Under assumptions (A1) - (A3), it holds as $p \to \infty$ *that*

$$\mathbb{E}_{\tau_0=0}(\min\{1 \leq t \leq m : T(t) > T_r\}) \to 1.$$

Theorem 6.3.7 shows that on average one post-change observation is required asymptotically as $p \to \infty$ to terminate the detection procedure after a change point occurs. The above two theorems hold for finite m, as long as $p \to \infty$. Do these results also hold when m grows with p? The following result answers this question.

Theorem 6.3.8. *Consider the conditions in Theorem 6.3.6. Under assumptions (A1) - (A3), we have as* $p \to \infty$ *and* $m \to \infty$ *that*

(a) $\sup_{0 \leq \tau_0 < m} \left|\mathbb{P}_{\tau_0}\left(T(t) > T_r\right) - 1\right| \to 0 \quad \text{for } t > \tau_0.$

(b) The expected detection delay satisfies

$$\sup_{0 \leq \tau_0 < m} \mathrm{EDD}(\tau_0) = \mathbb{E}_{\tau_0=0}(\min\{t \in \mathbb{N} : T(t) > T_r\}),$$

and furthermore $\mathbb{E}_{\tau_0=0}(\min\{t \in \mathbb{N} : T(t) > T_r\}) \to 1.$

This theorem proves the asymptotic optimality of our method when both $m \to \infty$ and $p \to \infty$.

6.4 Numerical studies for online change point scenario

In this section, we conduct some numerical studies of DC-OCP in the online change point setting. The advantages of the distance-based CUSUM statistic, as shown in Chapter 4, also apply to online change point detection. So we avoid repeating them. In Subsections 6.4.1 and 6.4.2, we assess the accuracy of the theoretical results by evaluating the empirical EDD and detection power under the alternative, and the FWER and ARL under the null, respectively. In Subsection 6.4.3, we compare DC-OCP with some recent methods.

6.4.1 Detection delay and empirical power

We first evaluate the detection delay and the detection power of our proposed method DC-OCP under some online high dimensional scenarios. Here, we consider two types of distributional change: a change in the mean of observations and a change in the variance of observations. Our simulations include historical data of size $n = 100$, arriving data of sizes $m \in \{100, 300\}$, dimensions $p \in \{500, 1000, 2000\}$ and window sizes $w \in \{50, 100\}$. For historical data, we generate n random observations from a p-variate normal distribution $N(\mathbf{0}_p, \boldsymbol{V}_p)$, where $\mathbf{0}_p$ is a p-dimensional vectors of zeros and $\boldsymbol{V}_p$ is a $p \times p$ autoregressive covariance matrix defined as $\boldsymbol{V}_p = \left[0.5^{|i-j|}_{ij}\right]^{p \quad p}_{i=1 j=1}$. To impose a change point in the arriving observations, we generate the first $m/5$ random observations from the same distribution as historical data whilst the last $4m/5$ random observations are generated from $N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, allowing different means and variances as $\boldsymbol{\mu}_p \in \left\{\mathbf{0}_p, \left(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\right), \left(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\right)\right\}$ and $\boldsymbol{\Sigma}_p \in \{\boldsymbol{V}_p, 1.1\boldsymbol{V}_p, 1.2\boldsymbol{V}_p\}$. All m arriving observations arrive sequentially, and the true change point occurs at location $n + m/5$, except for the case of no change point when $\boldsymbol{\mu}_p = \mathbf{0}_p$ and $\boldsymbol{\Sigma}_p = \boldsymbol{V}_p$. Just to clarify that for the scenarios of a change in the mean of observations, we have $\boldsymbol{\mu}_p \in \{(0.2 \times \mathbf{1}_{3p/4}, \mathbf{0}_{p/4}), (0.3 \times \mathbf{1}_{3p/4}, \mathbf{0}_{p/4})\}$ and $\boldsymbol{\Sigma}_p = \boldsymbol{V}_p$.

We compute the three proposed thresholds using $S = 500$ permutations for each of the 200 replications to train the threshold, with $\alpha = 0.05$ for T_r^{BC} and $\alpha_{\mathrm{ARL}} = 1/3000$ for T_r^{ARL}. Table 6.1 reports the detection power of our method over 200 replications.

Table 6.1: Detection power of our method, DC-OCP, over 200 replications for detecting a true change in the mean of observations, using three thresholds: T_r^{BC}, T_r^{ARL}, and T_r^*.

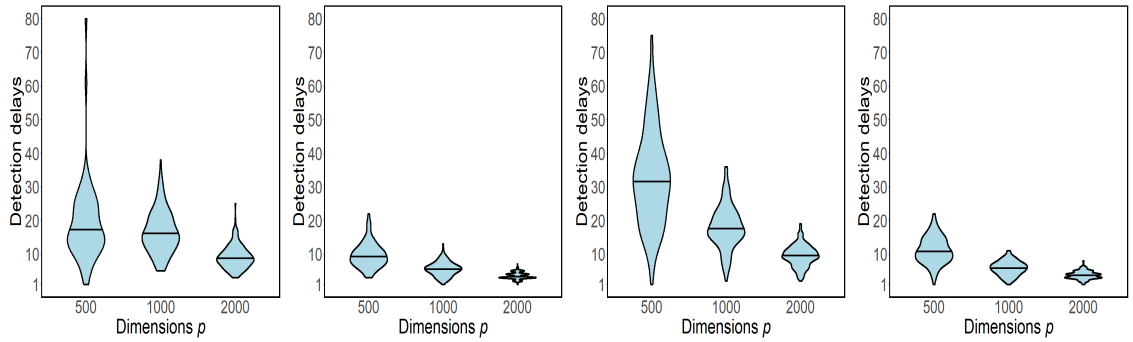
		$\boldsymbol{\mu}_p = \big(0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_p = \boldsymbol{V}_p$						$\boldsymbol{\mu}_p = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_p = \boldsymbol{V}_p$					
		$w = 50$			$w = 100$			$w = 50$			$w = 100$		
m	p	T_r^{BC}	T_r^{ARL}	T_r^*	T_r^{BC}	T_r^{ARL}	T_r^*	T_r^{BC}	T_r^{ARL}	T_r^*	T_r^{BC}	T_r^{ARL}	T_r^*
100	500	0.44	0.38	0.42	0.81	0.72	0.79	0.99	1.00	1.00	1.00	1.00	1.00
100	1000	0.80	0.75	0.77	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99
100	2000	1.00	1.00	1.00	0.99	0.99	0.99	0.99	1.00	0.99	1.00	1.00	1.00
300	500	0.29	0.36	0.32	0.56	0.68	0.63	0.99	0.98	0.99	1.00	0.98	1.00
300	1000	0.73	0.80	0.76	0.99	0.98	0.99	1.00	1.00	1.00	0.99	0.98	0.99
300	2000	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00

The results show that the power is considerably high, especially when the dimension p, window size w, or signal strength increases, which is in line with Theorem 6.3.6. We note that the power with T_r^{BC} slightly drops when m increases as expected, while T_r^{ARL} and T_r^* show a stable power. We note that a larger window size w improves the power of T_r^{BC} for large m. Overall, our method performs reasonably well in detecting a change in the mean of observations in these high dimensional settings with all the proposed thresholds.

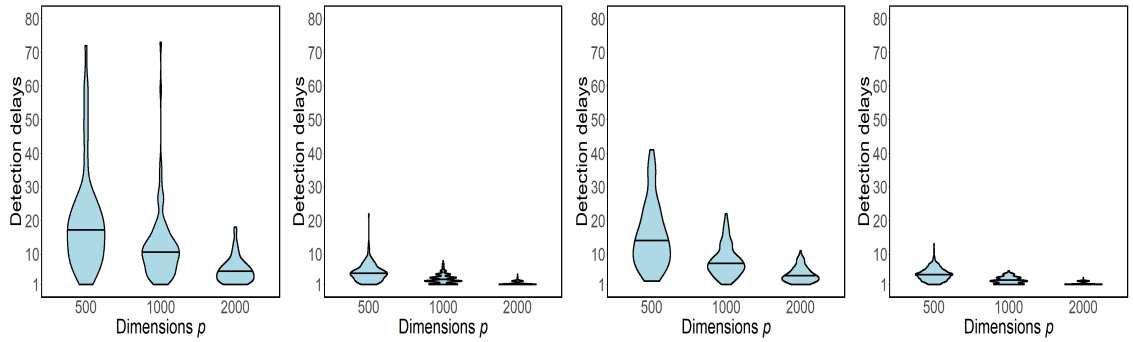
The violin plots in Figures 6.3a -6.3d show the detection delays including the average values, to detect a change in the mean of observations with $m = 100$ using the threshold T_r^{BC}. It can be seen that the average detection delay approaches 1 as the dimension p or the signal increases, which is consistent with Theorem 6.3.7. For scenarios of a change in the variance of observations, we recall $\boldsymbol{\Sigma}_p \in \{1.1\boldsymbol{V}_p, 1.2\boldsymbol{V}_p\}$ and we fix $\boldsymbol{\mu}_p = \mathbf{0}_p$. The empirical power and detection delay for detecting a change in the variance of observations are presented in Table 6.2 and Figures 6.3e -6.3h, respectively. The results indicate that our method has good performance in detecting a change in the variance of observations. We report the results on the detection delay for detecting a change in the mean of observations and a change in the variance of observations with the arriving data size of $m = 300$ in Figure 6.4. We can see that the presented results are similar to those with $m = 100$ in Figure 6.3, with the average detection delay approaching 1 as the dimension p or the signal increases. These results are consistent with Theorem 6.3.8 when m also grows with p.

Table 6.2: Detection power of our method, DC-OCP, over 200 replications for detecting a true change in the variance of observations, using three thresholds: T_r^{BC}, T_r^{ARL}, and T_r^*.

		$\boldsymbol{\mu}_p = \mathbf{0}_p$, $\boldsymbol{\Sigma}_p = 1.1\boldsymbol{V}_p$						$\boldsymbol{\mu}_p = \mathbf{0}_p$, $\boldsymbol{\Sigma}_p = 1.2\boldsymbol{V}_p$					
		$w = 50$			$w = 100$			$w = 50$			$w = 100$		
m	p	T_r^{BC}	T_r^{ARL}	T_r^*	T_r^{BC}	T_r^{ARL}	T_r^*	T_r^{BC}	T_r^{ARL}	T_r^*	T_r^{BC}	T_r^{ARL}	T_r^*
100	500	0.60	0.55	0.57	0.99	0.99	0.99	0.98	1.00	0.99	0.98	0.98	0.98
100	1000	0.90	0.86	0.88	0.99	0.99	0.99	1.00	1.00	1.00	0.99	0.99	0.99
100	2000	0.98	0.99	0.99	0.99	0.99	0.99	0.98	1.00	1.00	0.99	0.99	0.99
300	500	0.57	0.64	0.63	0.96	0.97	0.96	0.98	0.97	0.98	0.99	0.98	0.99
300	1000	0.90	0.96	0.95	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98
300	2000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	0.98	0.99

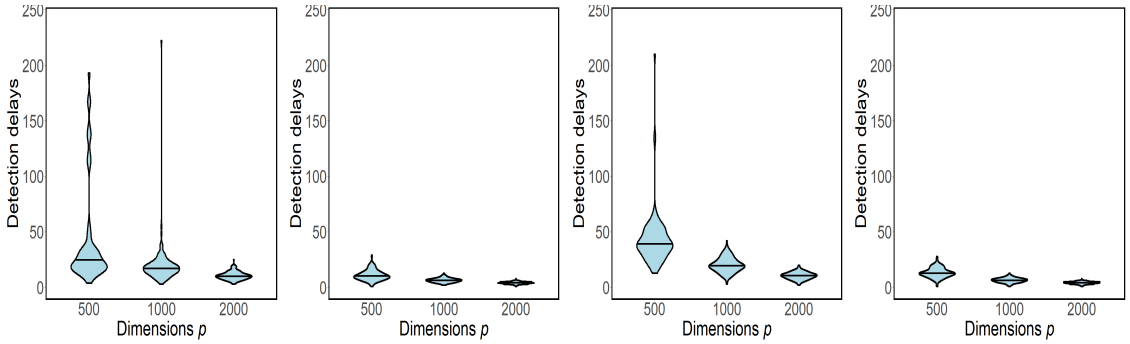


(a) $\boldsymbol{\mu}_p = (0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $w = 50$ (b) $\boldsymbol{\mu}_p = (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $w = 50$ (c) $\boldsymbol{\mu}_p = (0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $w = 100$ (d) $\boldsymbol{\mu}_p = (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $w = 100$

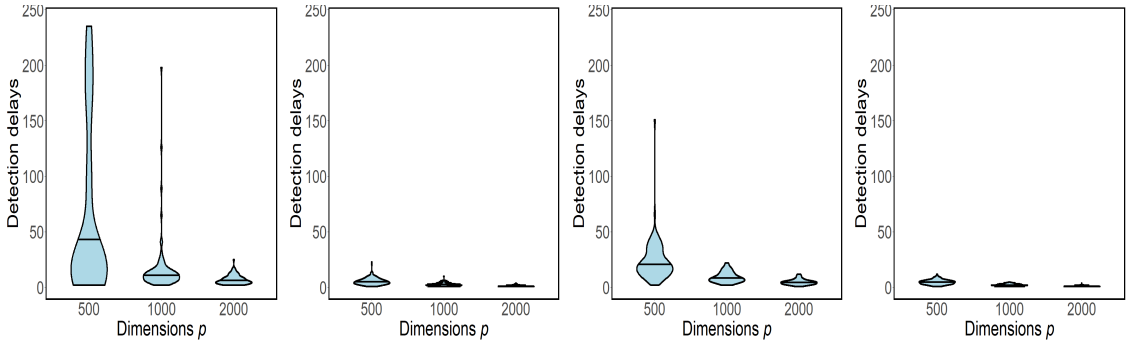


(e) $\boldsymbol{\Sigma}_p = 1.1\boldsymbol{V}_p$, $w = 50$ (f) $\boldsymbol{\Sigma}_p = 1.2\boldsymbol{V}_p$, $w = 50$ (g) $\boldsymbol{\Sigma}_p = 1.1\boldsymbol{V}_p$, $w = 100$ (h) $\boldsymbol{\Sigma}_p = 1.2\boldsymbol{V}_p$, $w = 100$

Figure 6.3: Detection delay of our method, DC-OCP, over 200 replications with $m = 100$ for detecting a true change in the mean of observations (figures (a) - (d)), and for detecting a true change in the variance of observations (figures (e) - (h)). The horizontal line in the middle of each violin plot highlights the average detection delay.



(a) $\boldsymbol{\mu}_p = (0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $w = 50$ (b) $\boldsymbol{\mu}_p = (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $w = 50$ (c) $\boldsymbol{\mu}_p = (0.2 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $w = 100$ (d) $\boldsymbol{\mu}_p = (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})$, $w = 100$



(e) $\boldsymbol{\Sigma}_p = 1.2\boldsymbol{V}_p$, $w = 50$ (f) $\boldsymbol{\Sigma}_p = 1.3\boldsymbol{V}_p$, $w = 50$ (g) $\boldsymbol{\Sigma}_p = 1.2\boldsymbol{V}_p$, $w = 100$ (h) $\boldsymbol{\Sigma}_p = 1.3\boldsymbol{V}_p$, $w = 100$

Figure 6.4: Detection delay of our method, DC-OCP, over 200 replications with $m = 300$ for detecting a true change in the mean of observations (figures (a) - (d)), and for detecting a true change in the variance of observations (figures (e) - (h)). The horizontal line in the middle of each violin plot highlights the average detection delay.

Table 6.3: False alarm rate (FAR) of our method, DC-OCP, over 200 replications for the case of no true change point using three thresholds: T_r^{BC} with $\alpha = 0.05$, T_r^{ARL} with $\alpha_{\mathrm{ARL}} = 1/3000$, and T_r^*.

		$w = 50$			$w = 100$		
m	p	T_r^{BC}	T_r^{ARL}	T_r^*	T_r^{BC}	T_r^{ARL}	T_r^*
100	500	0.03	0.03	0.03	0.01	0.01	0.01
100	1000	0.03	0.02	0.02	0.03	0.02	0.02
100	2000	0.04	0.01	0.02	0.03	0.01	0.02
300	500	0.03	0.08	0.04	0.03	0.06	0.04
300	1000	0.01	0.05	0.03	0.03	0.05	0.04
300	2000	0.01	0.01	0.01	0.02	0.03	0.02

Table 6.4: Average run length (ARL) of our method, DC-OCP, over 200 replications for the case of no true change point using threshold T_r^{ARL} with $\alpha_{\mathrm{ARL}} = 1/3000$. Here, the theoretical lower bound of ARL is 3000.

p	$w = 50$	$w = 100$
500	2989.25	3589.94
1000	3489.99	3300.01
2000	3624.72	3472.16

6.4.2 False alarm rate and average run length

We next assess the False Alarm Rate (FAR) and the ARL of our method under the null hypothesis of no change point. For this, using the previous simulation setting, we set $\boldsymbol{\mu}_p = \mathbf{0}_p$ and $\boldsymbol{\Sigma}_p = \boldsymbol{V}_p$ so that all simulated observations are generated from a p-variate normal distribution $N(\mathbf{0}_p, \boldsymbol{V}_p)$. We run simulations over 200 replications. Table 6.3 reports the FAR of our method using the three proposed thresholds, with different numbers of arriving observations m and window sizes w. We can see that all three thresholds control the FAR reasonably well. In particular, the test with threshold T_r^{BC} controls FAR below the nominal level $\alpha = 0.05$, which is in line with Theorem 6.3.4. Since the threshold T_r^{ARL} is designed to control the ARL of the test, we also calculate the empirical ARL using the threshold T_r^{ARL}. For this simulation scenario, we use the previous setting but set $m = 5000$. We let $\alpha_{\mathrm{ARL}} = 1/3000$ such that the theoretical lower bound of the ARL is 3000. The results presented in Table 6.4 show that the empirical ARL is reasonably close to the theoretical lower bound with different dimensions p and window sizes w, which supports our theoretical result in Theorem 6.3.5.

6.4.3 Comparison with other methods

We now compare the performance of our proposed method DC-OCP with some recent methods for online change points. They are the likelihood ratio test for multiscale high dimensional change point proposed by Chen et al. (2022), here called OCD, the multi-sensor mixture procedure by Xie and Siegmund (2013), here called XS, and the method for detecting changes in high dimensional covariance structure proposed by Li and Li (2023), here called OnlineCOV. In the simulations, we use $w = 100$ and $m = 100$, and set ARL = 3000 for all the methods. We investigate both normal and non-normal data scenarios. For normal data, we use the previous setting. For non-normal data, we first generate n observations from a p-variate Student's-t distribution with standardized covariance matrix $t_{v_1}(\mathbf{0}_p, \boldsymbol{I}_p)$ as historical data, where v_1 is the degrees of freedom. We then generate m arriving observations where the first $m/5$ observations are drawn from the same distribution as the historical data, and the last $4m/5$ observations are drawn from $t_{v_2}(\boldsymbol{\mu}_p, \boldsymbol{I}_p)$. Here, we set $\boldsymbol{\mu}_p \in \{\mathbf{0}_p, (0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4})\}$.

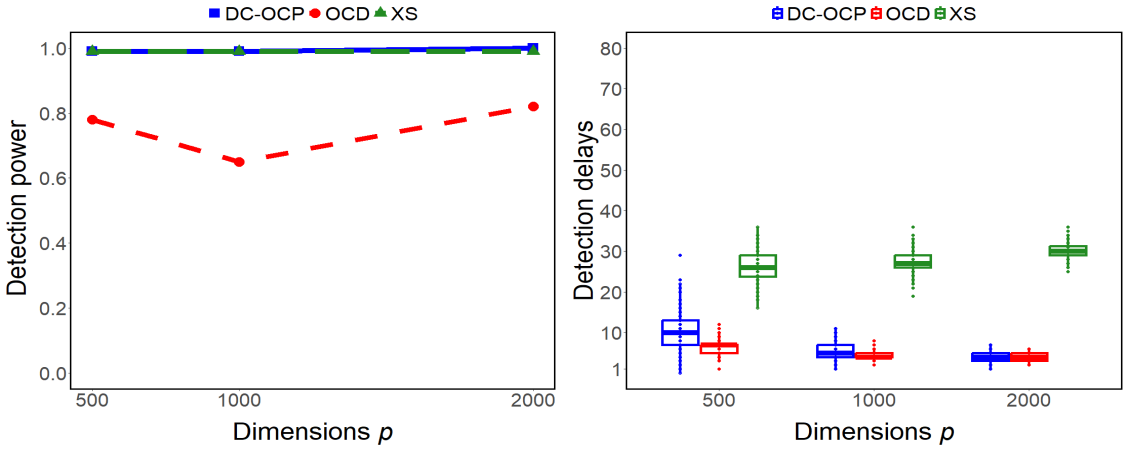
First, we compare our method DC-OCP with the OCD and the XS for detecting a change in the mean of normal observations. The results in Figure 6.5a show that our method DC-OCP has a higher detection power than the OCD, and the results in Figure 6.5b indicate that our method DC-OCP has lower average detection delays compared to the XS, also showing comparable detection delays with respect to the OCD as dimension p increases. This is because both OCD and XS require the assumption of normal observations with an uncorrelated covariance structure. To simulate a change in the mean with Student's-t observations, we use $\boldsymbol{\mu}_p = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$ and set $v_1 = v_2 = 30$. The reason we use a high degree of freedom $v_1 = v_2 = 30$ is to have a distribution still close to normal for a fair comparison with the OCD and the XS. The results in Figure 6.5c suggest that our method DC-OCP has better detection power than OCD. Also, Figure 6.5d shows that our method outperforms OCD and XS in terms of empirical detection delays. So our method outperforms both OCD and XS when the normality assumption does not hold.

Then, we compare our method DC-OCP with the OnlineCOV for detecting a

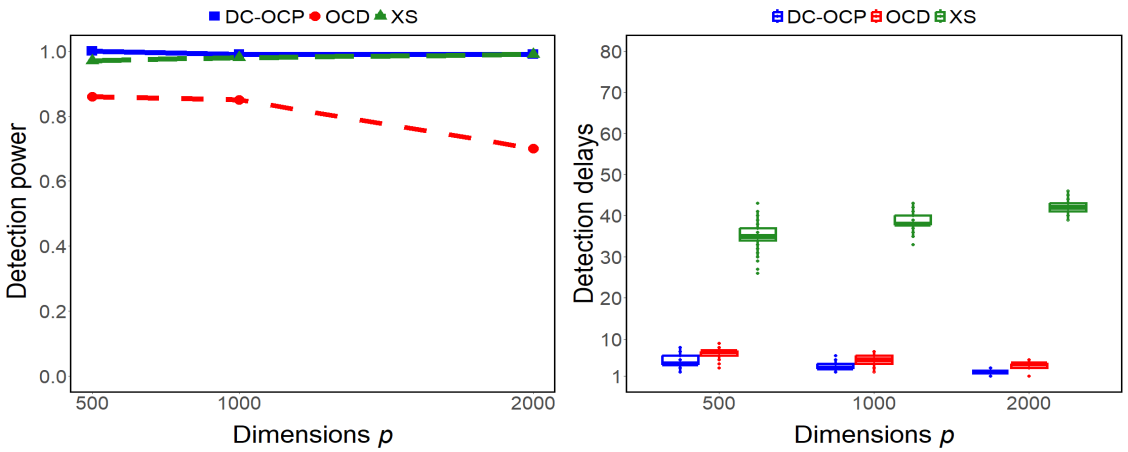
change in the variance of p-variate normal and p-variate Student's-t observations. For normal data, we use the same simulation setting as before, but for Student's-t data, we here set $v_1 = 5$ and $v_2 = 10$, and fix $\boldsymbol{\mu}_p = \mathbf{0}_p$ to keep the mean of observations the same in this comparison. The results in Figure 6.6a and Figure 6.6c show that our method DC-OCP has a higher power than the OnlineCOV under both normal and Student's-t data, and the power of OnlineCOV gets closer to ours only when p is very large. Furthermore, Figure 6.6b and Figure 6.6d indicate that our method DC-OCP has much lower empirical detection delays compared to OnlineCOV under both normal and Student's-t data. We note that the OnlineCOV method of Li and Li (2023) is mainly effective in detecting changes in the non-diagonal elements of the covariance matrix. In fact, in their simulations, they changed the covariance matrix of p-dimensional observations from the identity matrix to a matrix with 1 on the diagonal and 0.6 on the non-diagonal elements. In contrast, our settings focus mostly on changes in the diagonal entries of the covariance matrix. As a result, the OnlineCOV method is not very competitive in this case.

6.5 Real data application: human activity recognition data

We apply our online change point method to a real dataset called the Human Activity Recognition (HAR) data, which is obtained from the UCI Machine Learning Repository (see Anguita et al., 2013; Reyes-Ortiz et al., 2013). The dataset contains time-series data of daily human activities collected from experiments in which volunteers perform six activities (walking, walking upstairs, walking downstairs, sitting, standing, and laying) while carrying a smartphone with embedded sensors. There are 561 features in this database that come from the accelerometer and gyroscope 3-axial raw signals. This dataset was previously analyzed by Yilmaz (2017) for online change point detection in a low dimensional setup, since they considered only 5 features. But, in our analysis, we study the high dimensional case by analyzing all the 561 features. We focus on a time period during which a participant was instructed to perform three activities: walking downstairs, walking upstairs, and

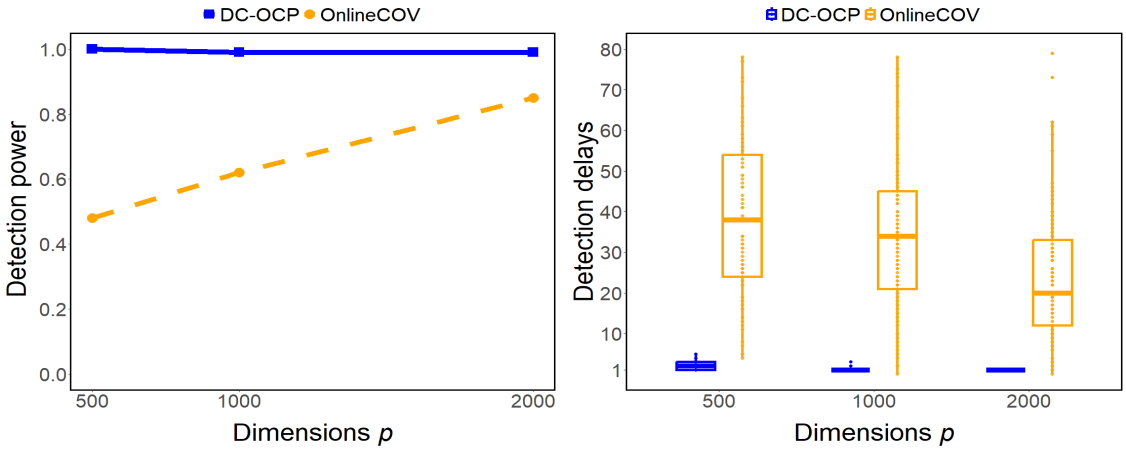


(a) $\boldsymbol{\mu}_p = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_p = \boldsymbol{V}_p$ (b) $\boldsymbol{\mu}_p = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\Sigma}_p = \boldsymbol{V}_p$

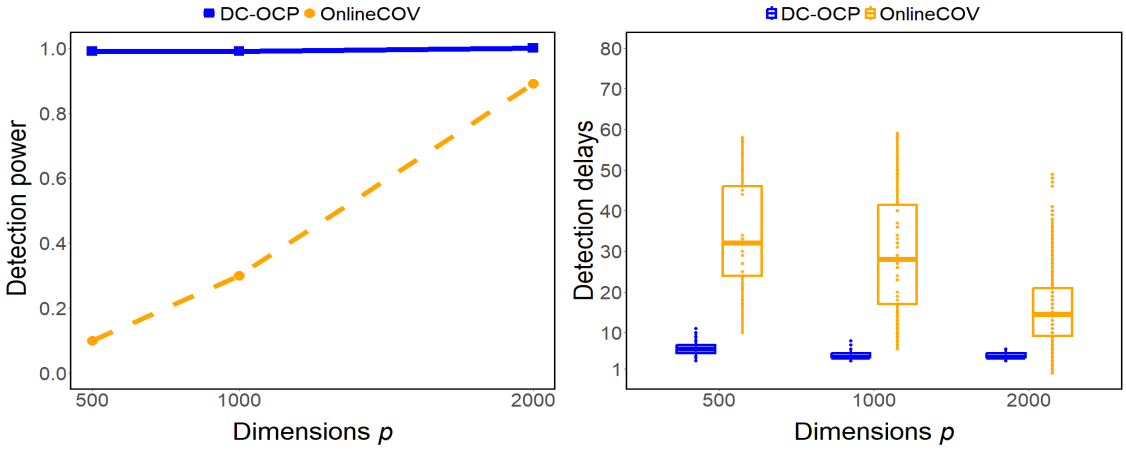


(c) $\boldsymbol{\mu}_p = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $v_1 = v_2 = 30$ (d) $\boldsymbol{\mu}_p = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $v_1 = v_2 = 30$

Figure 6.5: Detection power and detection delay of our method, DC-OCP, in comparison with OCD and XS over 200 replications for detecting a true change in the mean of observations: figures (a) and (b) show the results with p-variate normal observations, and figures (c) and (d) show the results with non-normal observations from a p-variate Student's-t distribution.



(a) $\boldsymbol{\mu}_p = \mathbf{0}_p$, $\boldsymbol{\Sigma}_p = 1.2\boldsymbol{V}_p$ (b) $\boldsymbol{\mu}_p = \mathbf{0}_p$, $\boldsymbol{\Sigma}_p = 1.2\boldsymbol{V}_p$



(c) $\boldsymbol{\mu}_p = \mathbf{0}_p$, $v_1 = 5$, $v_2 = 10$ (d) $\boldsymbol{\mu}_p = \mathbf{0}_p$, $v_1 = 5$, $v_2 = 10$

Figure 6.6: Detection power and detection delay of our method, DC-OCP, in comparison with OnlineCOV over 200 replications for detecting a change in the variance of observations: figures (a) and (b) show the results with p-variate normal observations, and figures (c) and (d) show the results with non-normal observations from a p-variate Student's-t distribution.

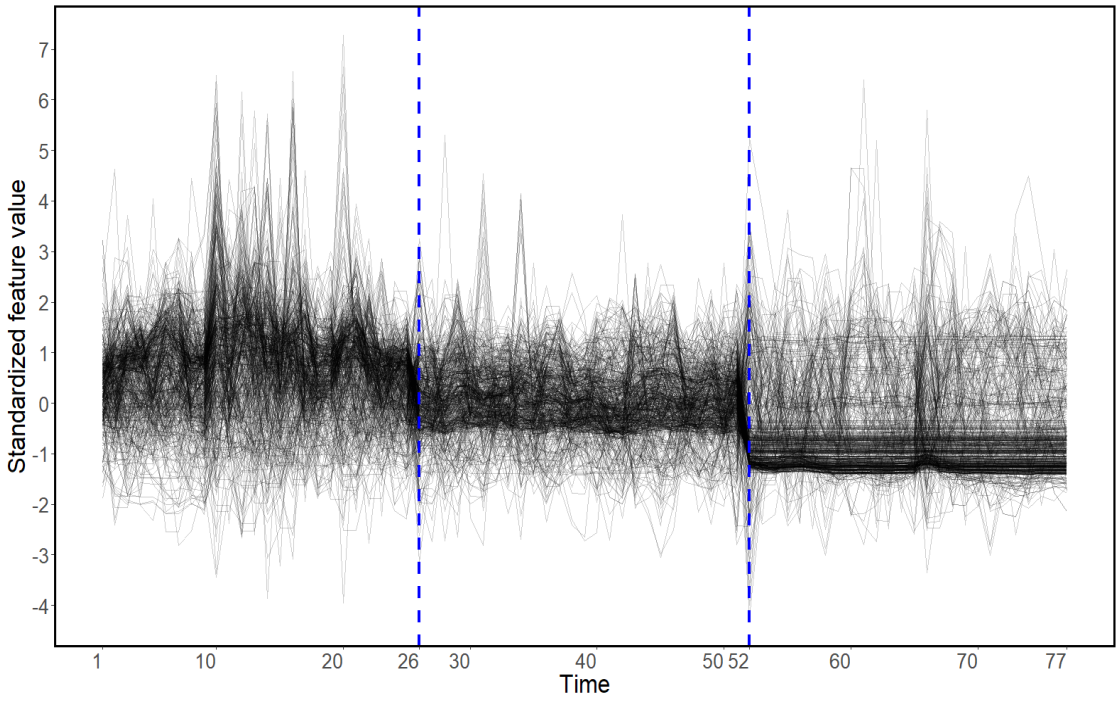


Figure 6.7: Human Activity Recognition (HAR) data with $p = 561$ features. The blue dashed lines mark the detected change points by our method. Note that all features are standardized here for the purpose of visualization.

standing. This period contains 77 time points (each representing 2.56 seconds), and it results in a high dimensional 77×561 data matrix. While the observations arrive sequentially, we use the first 20 observations as historical data, where no change point is found, to train thresholds T_r^{BC} with $\alpha = 0.05$ and T_r^{ARL} with $\alpha_{\mathrm{ARL}} = 1/3000$. We apply Algorithm 3 with two different window sizes $w \in \{15, 20\}$ to detect online change points in the arriving data. Unlike simulations, we here use $w = 15$ or $w = 20$ because the sample size is small. As discussed in Section 6.2, if Algorithm 3 detects a significant change point, we can continue detecting more change points as data still arrives. As shown in Figure 6.7, our method detects two significant change points at time points 26 and 52. To make sense of these findings, we notice that in the testing data, the participant walked downstairs from time 1 to 25, switched to walking upstairs from 26 to 51, and remained standing for the rest of the time. Therefore, two potential change points could be at times 25 and 51. These are very close to the expected change point locations. One reason for the effectiveness of our method is that both the mean and variance of the observations change when the

participant switches activities.

Furthermore, no additional change point is found after time 52 where the participant remains standing. Note that the same two change points are obtained when we use either of $w = 15$ and $w = 20$, and thresholds T_r^{BC} and T_r^{ARL}. These results highlight the effectiveness of our method in quickly detecting motion transitions in HAR data. Also, this example explains how Algorithm 3 can be naturally extended to detect multiple online change points in real-world data streams.

6.6 Proofs

Proof of Theorem 6.3.2

(a) Under the alternative hypothesis, since there is a single change point τ_0 in the m arriving observations, we have $F_1 = \cdots = F_{n+\tau_0} \neq F_{n+\tau_0+1} = \cdots = F_{n+m}$ with $\tau_0 \in \{0, \ldots, m-1\}$. For time $t > \tau_0$, it is straightforward to see that all observations in the test interval $[\boldsymbol{X}_{n+t-w+1}, \boldsymbol{X}_{n+t-w+2}, \ldots, \boldsymbol{X}_{n+t}]$ satisfy $F_{n+t-w+1} = \cdots = F_{n+\tau_0} \neq F_{n+\tau_0+1} = \cdots = F_{n+t}$ for $w > t - \tau_0$. Note that we must have $w > t - \tau_0$ because otherwise if, for instance, $w = t - \tau_0 - 1$ then the corresponding interval $[\boldsymbol{X}_{n+\tau_0+2}, \boldsymbol{X}_{n+\tau_0+3}, \ldots, \boldsymbol{X}_{n+t}]$ would not include the true change point τ_0. Similarly to the results of Theorem 3.4.4, we have as $p \to \infty$ that

$$\max_{n+t-w+1 \leq k \leq n+t-1} \left\{ \frac{1}{w} \sum_{i=n+t-w+1}^{n+t} C_i^2(k;t)|_{H_1} \right\} = \psi(k;t)\mathbf{1}(k = n+\tau_0) + o_P(1), \tag{6.12}$$

where $\psi(k;t)$ is the asymptotic limit of $\frac{1}{w}\sum_{i=n+t-w+1}^{n+t} C_i^2(k;t)|_{H_1}$ as $p \to \infty$ which is given by

$$\psi(k;t) = \begin{cases} \frac{(w-(t-\tau_0))^2(t-\tau_0)}{w^3}\left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\right)^2 & \\ + \frac{(w-(t-\tau_0))(t-\tau_0)^2}{w^3}\left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\right)^2 & \text{for } k = n+\tau_0, \\ \frac{(w-(t-\tau_0))(t-\tau_0)^2(w-n-t+k)}{w^3(n+t-k)}\left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\right)^2 & \\ + \frac{(t-\tau_0)^3(w-n-t+k)}{w^3(n+t-k)}\left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\right)^2 & \forall\, k \in \{n+t-w+1, \ldots, n+\tau_0-1\}, \\ \frac{(w-(t-\tau_0))^3(n+t-k)}{w^3(w-n-t+k)}\left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\right)^2 & \\ + \frac{(w-(t-\tau_0))^2(t-\tau_0)(n+t-k)}{w^3(w-n-t+k)}\left(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\right)^2 & \forall\, k \in \{n+\tau_0+1, \ldots, n+t-1\}. \end{cases} \tag{6.13}$$

The result in (6.12) indicates that the value of $\frac{1}{w}\sum_{i=n+t-w+1}^{n+t} C_i^2(k;t)$ reaches its maximum at $k = n + \tau_0$ when $p \to \infty$. Using the result in (6.13), this property can be justified as

$$
\begin{aligned}
&\psi(k;t)\mathbf{1}(k = n+\tau_0) - \psi(k;t)\mathbf{1}(k < n+\tau_0) \\
&= \frac{(w-(t-\tau_0))^2(t-\tau_0)}{w^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 + \frac{(w-(t-\tau_0))(t-\tau_0)^2}{w^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 \\
&\quad - \frac{(w-(t-\tau_0))(t-\tau_0)^2(w-n-t+k)}{w^3(n+t-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 \\
&\quad - \frac{(t-\tau_0)^3(w-n-t+k)}{w^3(n+t-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 \\
&= \frac{(w-(t-\tau_0))(t-\tau_0)w(n+\tau_0-k)}{w^3(n+t-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 \\
&\quad + \frac{(t-\tau_0)^2 w(n+\tau_0-k)}{w^3(n+t-k)}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 \\
&\geq 0
\end{aligned}
$$

Similarly, we use (6.13) to obtain that $\psi(k;t)\mathbf{1}(k = n+\tau_0) - \psi(k;t)\mathbf{1}(k > n+\tau_0) \geq 0$. Putting these two results together, we find for all $k \in \{n+t-w+1, \ldots, n+t-1\}$ that as $p \to \infty$

$$
\begin{aligned}
&\max_{n+t-w+1 \leq k \leq n+t-1}\{\psi(k;t)\} \\
&= \max_{n+t-w+1 \leq k \leq n+t-1}\{\psi(k;t)\mathbf{1}(k = n+\tau_0) + \psi(k;t)\mathbf{1}(k \neq n+\tau_0)\} \\
&= \psi(k;t)\mathbf{1}(k = n+\tau_0) \qquad (6.14)
\end{aligned}
$$

Note that Figure 6.1b in the main thesis shows an illustrative example of this property. We provide the calculation of $\psi(k;t)\mathbf{1}(k = n+\tau_0)$ given in (6.13). From the results of Lemma 6.3.1, the asymptotic limit of the proposed CUSUM statistic $C_i(k;t)$ in (3.4) depends on the values of i, k, and $n+\tau_0$. We recall that $i \in \{n+t-w+1, \ldots, n+t\}$ and $k \in \{n+t-w+1, \ldots, n+t-1\}$. We here use the indicator function $\mathbf{1}(\cdot)$ to help clarify the different cases of this in our proof. Obviously, $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if $i = j$. First, for all $i \in \{n+\tau_0+1, \ldots, n+t\}$ and $k = n+\tau_0$, we obtain as $p \to \infty$

$$
\begin{aligned}
&C_i(k = n+\tau_0; t)\,\mathbf{1}(n+t-w+1 \leq i \leq n+\tau_0) \\
&= h\Big(\frac{1}{n+t-k}\sum_{j=k+1}^{n+t} d(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{w-n-t+k}\sum_{j=n+t-w+1}^{k} d(\boldsymbol{X}_i, \boldsymbol{X}_j)\Big)
\end{aligned}
$$

$$= \frac{\sqrt{(t-\tau_0)(w-(t-\tau_0))}}{w}\Big(\frac{1}{t-\tau_0}\sum_{j=n+\tau_0+1}^{n+t} d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) - \frac{1}{w-(t-\tau_0)}\sum_{j=n+t-w+1}^{n+\tau_0} d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)\Big)$$

$$= \frac{\sqrt{(t-\tau_0)(w-(t-\tau_0))}}{w}\Big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} + o_P(1) - \frac{w-(t-\tau_0)-1}{w-(t-\tau_0)}\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)\Big)$$

$$= \frac{\sqrt{(t-\tau_0)(w-(t-\tau_0))}}{w}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) + \sqrt{\frac{t-\tau_0}{w^2(w-(t-\tau_0))}}\,\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} + o_P(1)$$

$$= \frac{\sqrt{(t-\tau_0)(w-(t-\tau_0))}}{w}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big) + o_P(1), \tag{6.15}$$

where we note $\sqrt{\frac{t-\tau_0}{w^2(w-(t-\tau_0))}}\,\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} < \sqrt{\frac{w}{w^2(w-(t-\tau_0))}}\,\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}} = \sqrt{\frac{1}{w(w-(t-\tau_0))}}\,\lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}$ is negligible for sufficiently large w. For all $i \in \{n+\tau_0+1, \ldots, n+t\}$ and $k = n+\tau_0$, we similarly obtain as $p \to \infty$

$$C_i\left(k = n+\tau_0; t\right)\mathbf{1}(n+\tau_0+1 \le i \le n+t)$$

$$= h\Big(\frac{1}{n+t-k}\sum_{j=k+1}^{n+t} d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) - \frac{1}{w-n-t+k}\sum_{j=n+t-w+1}^{k} d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)\Big)$$

$$= \frac{\sqrt{(t-\tau_0)(w-(t-\tau_0))}}{w}\Big(\frac{1}{t-\tau_0}\sum_{j=n+\tau_0+1}^{n+t} d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) - \frac{1}{w-(t-\tau_0)}\sum_{j=n+t-w+1}^{n+\tau_0} d\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)\Big)$$

$$= \frac{\sqrt{(t-\tau_0)(w-(t-\tau_0))}}{w}\Big(\frac{t-\tau_0-1}{t-\tau_0}\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} + o_P(1) - \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} + o_P(1)\Big)$$

$$= \frac{\sqrt{(t-\tau_0)(w-(t-\tau_0))}}{w}\big(\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}\big) - \sqrt{\frac{w-(t-\tau_0)}{w^2(t-\tau_0)}}\,\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} + o_P(1)$$

$$= \frac{\sqrt{(t-\tau_0)(w-(t-\tau_0))}}{w}\big(\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^+_{\tau_0}}\big) + o_P(1), \tag{6.16}$$

where $\sqrt{\frac{w-(t-\tau_0)}{w^2(t-\tau_0)}}\,\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} < \sqrt{\frac{w}{w^2(t-\tau_0)}}\,\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}} = \sqrt{\frac{1}{w(t-\tau_0)}}\,\lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}$ is negligible for sufficiently large w. Hence, using the results in (6.15) and (6.16), we have as $p \to \infty$

$$\begin{aligned}
&\frac{1}{w}\sum_{i=n+t-w+1}^{n+t} C_i^2(k = n+\tau_0; t) \\
&= \frac{1}{w}\Big(\sum_{i=n+t-w+1}^{n+\tau_0} C_i^2(k = n+\tau_0; t)\mathbf{1}(n+t-w+1 \le i \le n+\tau_0) \\
&\quad + \sum_{i=n+\tau_0+1}^{n+t} C_i^2(k = n+\tau_0; t)\mathbf{1}(n+\tau_0+1 \le i \le n+t)\Big) \\
&= \frac{1}{w}\Big((w-(t-\tau_0))\frac{(t-\tau_0)(w-(t-\tau_0))}{w^2}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2 \\
&\quad + (t-\tau_0)\frac{(t-\tau_0)(w-(t-\tau_0))}{w^2}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2\Big) + o_P(1)
\end{aligned}$$

$$= \frac{(t-\tau_0)(w-(t-\tau_0))^2}{w^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2$$
$$+ \frac{(t-\tau_0)^2(w-(t-\tau_0))}{w^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2 + o_P(1)$$
$$= \psi(k;t)\mathbf{1}(k = n+\tau_0) + o_P(1).$$

Repeating the same calculations as above for all $i \in \{n+t-w+1,\ldots,n+t\}$ and $k \in \{n+t-w+1,\ldots,n+t-1\}$, we get the result for $\psi(k;t)$ given in (6.13) when $p \to \infty$. Now, let $\Psi(t)$ denote the value of $\psi(k;t)\mathbf{1}(k = n+\tau_0)$ as follows

$$\Psi(t) = \frac{(t-\tau_0)\,(w-(t-\tau_0))^2}{w^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^-_{\tau_0}A^-_{\tau_0}}\big)^2$$
$$+ \frac{(t-\tau_0)^2\,(w-(t-\tau_0))}{w^3}\big(\lambda^2_{A^-_{\tau_0}A^+_{\tau_0}} - \lambda^2_{A^+_{\tau_0}A^+_{\tau_0}}\big)^2. \tag{6.17}$$

Using (6.14) and (6.17), we obtain as $p \to \infty$ that

$$\begin{aligned}
T(t)|_{H_1} &= \max_{n+t-w+1\le k\le n+t-1}\Big\{\frac{1}{w}\sum_{i=n+t-w+1}^{n+t} C_i^2(k;t)\Big\} \\
&= \max_{n+t-w+1\le k\le n+t-1}\big\{\psi(k;t) + o_P(1)\big\} \\
&= \max_{n+t-w+1\le k\le n+t-1}\big\{\psi(k;t)\mathbf{1}(k = n+\tau_0) + \psi(k;t)\mathbf{1}(k \ne n+\tau_0)\big\} + o_P(1) \\
&= \psi(k;t)\mathbf{1}(k = n+\tau_0) + o_P(1) \\
&= \Psi(t) + o_P(1), \quad \text{for } t > \tau_0.
\end{aligned} \tag{6.18}$$

For time $t \le \tau_0$, all observations in the test interval $[\boldsymbol{X}_{n+t-w+1}, \boldsymbol{X}_{n+t-w+2},\ldots,\boldsymbol{X}_{n+t}]$ do not include the true change point. Therefore, observations are identically distributed and we have $\lambda_{A^-_{\tau_0}A^-_{\tau_0}} = \lambda_{A^+_{\tau_0}A^+_{\tau_0}} = \lambda_{A^-_{\tau_0}A^+_{\tau_0}}$. Hence, we obtain $\psi(k;t) = 0$ in (6.13) for all $i \in \{n+t-w+1,\ldots,n+t\}$ and $k \in \{n+t-w+1,\ldots,n+t-1\}$ as $p \to \infty$. Therefore, we have as $p \to \infty$

$$\begin{aligned}
T(t)|_{H_1} &= \max_{n+t-w+1\le k\le n+t-1}\left\{\frac{1}{w}\sum_{i=n+t-w+1}^{n+t} C_i^2(k;t)|_{H_1}\right\} \\
&= \max_{n+t-w+1\le k\le n+t-1}\{\psi(k;t) + o_P(1)\} \\
&= \max_{n+t-w+1\le k\le n+t-1}\{o_P(1)\}
\end{aligned}$$

$$= o_P(1) \quad \text{for } t \leq \tau_0. \tag{6.19}$$

This completes the proof of part (a).

(b) Under the null hypothesis of no change point, consider the aforementioned test interval $[\boldsymbol{X}_{n+t-w+1}, \boldsymbol{X}_{n+t-w+2}, \ldots, \boldsymbol{X}_{n+t}]$ for time $t = 1, \ldots, m$. Similarly to the proof in part (a) where time $t \leq \tau_0$, it holds trivially that, as $p \to \infty$,

$$T(t)|_{H_0} = o_P(1) \quad \text{for } t = 1, \ldots, m. \tag{6.20}$$

Figure 6.1a in the main thesis gives an illustrative example of this property. This completes the proof of part (b).

Proof of Lemma 6.3.3

(a) Using the Weak Law of Large Numbers (WLLN), we obtain as $S \to \infty$

$$\begin{aligned} G_{T^S_{\text{perm}}}(u) &= \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\left(T^s_{\text{perm}} \leq u\right) \\ &\xrightarrow{P} \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}\left(\mathbf{1}\left(T^s_{\text{perm}} \leq u\right)\right) \\ &= \frac{1}{S} \sum_{s=1}^{S} \mathbb{P}_\infty\left(T^s_{\text{perm}} \leq u\right), \end{aligned} \tag{6.21}$$

where $u \in \mathbb{R}^+$. Let $G_{T_{H_0}}(u)$ denote the distribution of $T(t)$ under the null hypothesis as follows

$$G_{T_{H_0}}(u) = \mathbb{P}_\infty\left(T(t) \leq u\right),$$

Under the null hypothesis of no change point, all observations have the same distribution. Hence, we must have

$$\frac{1}{S} \sum_{s=1}^{S} \mathbb{P}_\infty\left(T^s_{\text{perm}} \leq u\right) = G_{T_{H_0}}(u), \tag{6.22}$$

where $u \in \mathbb{R}^+$. From (6.21) and (6.22), it follows that as $S \to \infty$

$$G_{T^S_{\text{perm}}}(u) \xrightarrow{P} G_{T_{H_0}}(u) \quad \forall\, u \in \mathbb{R}^+. \tag{6.23}$$

(b) For the threshold T_r^{BC} in (6.7), it is straightforward to obtain using the result in (6.23) that

$$T_r^{\text{BC}} \xrightarrow{P} q_{1-\alpha/m}, \tag{6.24}$$

where $q_{1-\alpha/m}$ is defined as $G_{T_{H_0}}(q_{1-\alpha/m}) = 1 - \alpha/m$ with $\alpha \in [0, 1]$. Using (6.24) and applying the continuous mapping theorem to the CDF of continuous random variables $T(t)$, we get as $S \to \infty$

$$\begin{aligned}
\mathbb{P}_\infty\left(T(t) > T_r^{\text{BC}}\right) &= 1 - \mathbb{P}_\infty\left(T(t) \leq T_r^{\text{BC}}\right) \\
&\xrightarrow{P} 1 - \mathbb{P}_\infty(T(t) \leq q_{1-\alpha/m}) \\
&= 1 - G_{T_{H_0}}(q_{1-\alpha/m}) \\
&= 1 - (1 - \frac{\alpha}{m}) \\
&= \frac{\alpha}{m},
\end{aligned}$$

where $\alpha \in [0, 1]$.

Proof of Theorem 6.3.4

Applying Boole's inequality and Lemma 6.3.3, we get as $S \to \infty$ that

$$\begin{aligned}
\mathbb{P}_\infty\left(\min\left\{1 \leq t \leq m : T(t) > T_r^{\text{BC}}\right\} \leq m\right) &= \mathbb{P}_\infty\Big(\bigcup_{t=1}^{m}\left\{T(t) > T_r^{\text{BC}}\right\}\Big) \\
&\leq \sum_{t=1}^{m} \mathbb{P}_\infty\left(T(t) > T_r^{\text{BC}}\right) \\
&\xrightarrow{P} m \cdot \frac{\alpha}{m} \\
&= \alpha.
\end{aligned}$$

Proof of Theorem 6.3.5

Let $t \in \mathbb{N}$ denote discrete time. For clarity of exposition, we define notation $t_{\text{stop}}^{\text{ARL}}$ to denote the stopping time using threshold T_r^{ARL} as follows

$$t_{\text{stop}}^{\text{ARL}} = \min\left\{t \in \mathbb{N} : T(t) > T_r^{\text{ARL}}\right\}.$$

The expectation of stopping time $t_{\text{stop}}^{\text{ARL}}$ under the null hypothesis can be written as

$$\begin{aligned}
\mathbb{E}_\infty\left(t_{\text{stop}}^{\text{ARL}}\right) &= \sum_{i=1}^{\infty} i \cdot \mathbb{P}_\infty(t_{\text{stop}}^{\text{ARL}} = i) \\
&= \sum_{i=1}^{\infty} \Big(\sum_{j=1}^{i} 1\Big) \cdot \mathbb{P}_\infty(t_{\text{stop}}^{\text{ARL}} = i) \\
&= \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} \mathbb{P}_\infty(t_{\text{stop}}^{\text{ARL}} = i) \\
&= \sum_{j=1}^{\infty} \mathbb{P}_\infty(t_{\text{stop}}^{\text{ARL}} \geq j), \qquad (6.25)
\end{aligned}$$

where the order of summations is exchanged using Tonelli's theorem because all terms are non-negative. We define $\mathcal{A}_t$ as the event that the test statistic is less than or equal to the threshold at time t as follows

$$\mathcal{A}_t := \left\{T(t) \leq T_r^{\text{ARL}}\right\}. \qquad (6.26)$$

Then, using the chain rule in probability, we can write the probability $\mathbb{P}_\infty(t_{\text{stop}}^{\text{ARL}} \geq j)$ in (6.25) as

$$\begin{aligned}
\mathbb{P}_\infty\left(t_{\text{stop}}^{\text{ARL}} \geq j\right) &= \mathbb{P}_\infty\Big(\bigcap_{t=1}^{j-1} \mathcal{A}_t\Big) \\
&= \prod_{t=1}^{j-1} \mathbb{P}_\infty\Big(\mathcal{A}_t \mid \bigcap_{k=1}^{t-1} \mathcal{A}_k\Big) \\
&\geq \prod_{t=1}^{j-1} \mathbb{P}_\infty(\mathcal{A}_t) \qquad (6.27)
\end{aligned}$$

where in the last inequality we used $\mathbb{P}_\infty\big(\mathcal{A}_t \mid \bigcap_{k=1}^{t-1} \mathcal{A}_k\big) \geq \mathbb{P}_\infty(\mathcal{A}_t)$ which follows from the fact that the probability of no rejection at time t given no previous rejection is at least as large as its unconditional probability under H_0. This is because each test statistic is computed from overlapped sliding window data. Similarly to the result in part (b) of Lemma 6.3.3, it is straightforward to show that as $S \to \infty$

$$\mathbb{P}_\infty(\mathcal{A}_t) = 1 - \mathbb{P}_\infty(T(t) > T_r^{\text{ARL}}) \xrightarrow{P} 1 - \alpha_{\text{ARL}}, \tag{6.28}$$

where $\alpha_{\text{ARL}} \in (0, 1)$. Using the results in (6.27) and (6.28), equation (6.25) can be written as

$$\begin{aligned}
\mathbb{E}_\infty\left(t_{\text{stop}}^{\text{ARL}}\right) &= \sum_{j=1}^{\infty} \mathbb{P}_\infty(t_{\text{stop}}^{\text{ARL}} \geq j) \\
&= \sum_{j=1}^{\infty} \left(\prod_{t=1}^{j-1} \mathbb{P}_\infty\Big(\mathcal{A}_t \mid \bigcap_{k=1}^{t-1} \mathcal{A}_k\Big) \right) \\
&\geq \sum_{j=1}^{\infty} \prod_{t=1}^{j-1} \mathbb{P}_\infty(\mathcal{A}_t) \\
&\geq \sum_{j=1}^{\infty} (1 - \alpha_{\text{ARL}})^{j-1} \\
&= \frac{1}{\alpha_{\text{ARL}}},
\end{aligned}$$

where the last step is followed using the formula for the sum of an infinite geometric series. This completes the proof.

Proof of Theorem 6.3.6

First, recall the result in part (b) of Theorem 6.3.2 saying that as $p \to \infty$

$$T(t) \xrightarrow{P} \Psi(t) \quad \text{for } t > \tau_0. \tag{6.29}$$

We notice that

$$\Psi(t) > \Big(\frac{(t - \tau_0)\,(w - (t - \tau_0))^2}{w^3} + \frac{(t - \tau_0)^2\,(w - (t - \tau_0))}{w^3} \Big) \Delta_{\min}$$

$$= \frac{(t-\tau_0)(w-(t-\tau_0))}{w^2}\Delta_{\min}.$$

Under assumption (A3), it follows that $\Psi(t) > 0$. Let $a = \frac{(t-\tau_0)(w-(t-\tau_0))}{w^2}\Delta_{\min} > 0$. By the convergence in probability of test statistic $T(t)$ in (6.29) when $p \to \infty$, we have for a given $\epsilon_1 = \frac{1}{2}a > 0$ that

$$\lim_{p\to\infty} \mathbb{P}_{\tau_0}\left(|T(t) - \Psi(t)| \geq \epsilon_1\right) = 0. \tag{6.30}$$

Using the limit for a sequence to the result in (6.30) when $p \to \infty$, for all $\delta_1 > 0$, there exists a $p_1 > 0$ such that, for every $p > p_1$, we have

$$\begin{aligned} & \mathbb{P}_{\tau_0}\left(|T(t) - \Psi(t)| \geq \epsilon_1\right) < \delta_1 \\ \Rightarrow \quad & \mathbb{P}_{\tau_0}\left(T(t) - \Psi(t) \geq \epsilon_1\right) + \mathbb{P}_{\tau_0}\left(T(t) - \Psi(t) \leq -\epsilon_1\right) < \delta_1 \\ \Rightarrow \quad & \mathbb{P}_{\tau_0}\left(T(t) \leq \Psi(t) - \epsilon_1\right) < \delta_1. \end{aligned} \tag{6.31}$$

Since the threshold T_r is obtained as the upper quantile of the permutation distribution of the test statistic under the null hypothesis, we have used part (b) of Theorem 6.3.2 that as $p \to \infty$

$$T_r \xrightarrow{P} 0. \tag{6.32}$$

Similarly to the above, using the convergence in probability and limit for a sequence to (6.32), for a given $\epsilon_2 = \frac{1}{2}a > 0$, for all $\delta_2 > 0$, there exists a $p_2 > 0$ such that, for every $p > p_2$, we have

$$\mathbb{P}_{\tau_0}(T_r \geq \epsilon_2) < \delta_2. \tag{6.33}$$

Since $\Psi(t) - \epsilon_1 > \frac{1}{2}a = \epsilon_2$, using Boole's inequality and the results in (6.31) and (6.33), for every $p > \max\{p_1, p_2\}$, we can write

$$\begin{aligned} \mathbb{P}_{\tau_0}(T(t) > T_r) &\geq \mathbb{P}_{\tau_0}\big(\{T(t) > \Psi(t) - \epsilon_1\} \cap \{T_r < \epsilon_2\}\big) \\ &= 1 - \mathbb{P}_{\tau_0}\big(\{T(t) \leq \Psi(t) - \epsilon_1\} \cup \{T_r \geq \epsilon_2\}\big) \\ &\geq 1 - \mathbb{P}_{\tau_0}(T(t) \leq \Psi(t) - \epsilon_1) - \mathbb{P}_{\tau_0}(T_r \geq \epsilon_2) \\ &> 1 - \delta_1 - \delta_2. \end{aligned} \tag{6.34}$$

Since $\delta_1, \delta_2 > 0$ in (6.34) are arbitrary, we have as $p \to \infty$ that

$$\mathbb{P}_{\tau_0}(T(t) > T_r) > 1-\delta_1-\delta_2 \Rightarrow |\mathbb{P}_{\tau_0}(T(t) > T_r)-1| < \delta_1+\delta_2 \Rightarrow |\mathbb{P}_{\tau_0}(T(t) > T_r)-1| \to 0, \tag{6.35}$$

where $t > \tau_0$. We note that the result $\Psi(t) > \frac{(t-\tau_0)(w-(t-\tau_0))}{w^2}\Delta_{\min} > 0$ holds for any $\tau_0 \in \{0, \ldots, m-1\}$ with $t > \tau_0$. Repetition of the above process produces the result in (6.35) for any $\tau_0 \in \{0, \ldots, m-1\}$. Hence, we have as $p \to \infty$ that

$$\sup_{0 \le \tau_0 < m} |\mathbb{P}_{\tau_0}(T(t) > T_r) - 1| \to 0 \quad \text{for } t > \tau_0,$$

which completes the proof.

Proof of Theorem 6.3.7

(a) Under the alternative hypothesis, there is a single change point τ_0 in the m arriving observations such that $F_1 = \cdots = F_{n+\tau_0} \neq F_{n+\tau_0+1} = \cdots = F_{n+m}$ with $\tau_0 \in \{0, \ldots, m-1\}$. We recall that the stopping time is defined as

$$t_{\text{stop}} = \min\{1 \le t \le m : T(t) > T_r\}. \tag{6.36}$$

Using the tail sum formula of expectation, we can write the worst-case EDD as

$$\begin{aligned}
\sup_{0 \le \tau_0 < m} \text{EDD}(\tau_0) &= \sup_{0 \le \tau_0 < m} \mathbb{E}_{\tau_0}(t_{\text{stop}} - \tau_0 \mid t_{\text{stop}} > \tau_0) \\
&= \sup_{0 \le \tau_0 < m} \sum_{i=1}^{\infty} \mathbb{P}_{\tau_0}(t_{\text{stop}} \ge \tau_0 + i \mid t_{\text{stop}} > \tau_0) \\
&= \sup_{0 \le \tau_0 < m} \sum_{i=1}^{\infty} \big(1 - \mathbb{P}_{\tau_0}(t_{\text{stop}} < \tau_0 + i \mid t_{\text{stop}} > \tau_0)\big).
\end{aligned} \tag{6.37}$$

From the definition of t_{stop} in (6.36), the stopping time is strictly positive. Therefore, the EDD for an immediate true change point $\tau_0 = 0$ can be written as

$$\begin{aligned}
\mathbb{E}_{\tau_0=0}(t_{\text{stop}} - 0 \mid t_{\text{stop}} > 0) &= \mathbb{E}_{\tau_0=0}(t_{\text{stop}}) \\
&= \sum_{i=1}^{\infty} \mathbb{P}_{\tau_0=0}(t_{\text{stop}} \ge i)
\end{aligned}$$

$$= \sum_{i=1}^{\infty} \big(1 - \mathbb{P}_{\tau_0=0}(t_{\text{stop}} < i)\big).$$

Now, to show that the supremum of EDD is reached at the immediate change point $\tau_0 = 0$, it suffices to prove that $\mathbb{P}_{\tau_0}(t_{\text{stop}} < \tau_0 + i \mid t_{\text{stop}} > \tau_0) \geq \mathbb{P}_{\tau_0=0}(t_{\text{stop}} < i)$. For this, we can write

$$\begin{aligned}\mathbb{P}_{\tau_0}(t_{\text{stop}} < \tau_0 + i \mid t_{\text{stop}} > \tau_0) &= \frac{\mathbb{P}_{\tau_0}(\tau_0 < t_{\text{stop}} < \tau_0 + i)}{\mathbb{P}_{\tau_0}(t_{\text{stop}} > \tau_0)} \\ &= \frac{\mathbb{P}_{\tau_0}(t_{\text{stop}} < \tau_0 + i) - \mathbb{P}_{\tau_0}(t_{\text{stop}} \leq \tau_0)}{1 - \mathbb{P}_{\tau_0}(t_{\text{stop}} \leq \tau_0)}, \end{aligned} \tag{6.38}$$

where the denominator $\mathbb{P}_{\tau_0}(t_{\text{stop}} > \tau_0)$ must be nonzero under the alternative hypothesis. For the numerator, we can write

$$\mathbb{P}_{\tau_0}(t_{\text{stop}} < \tau_0 + i) = \mathbb{P}_{\tau_0}\big(\min\{1 \leq t \leq m : T(t) > T_r\} < \tau_0 + i\big),$$

and

$$\mathbb{P}_{\tau_0}(t_{\text{stop}} \leq \tau_0) = \mathbb{P}_{\tau_0}\big(\min\{1 \leq t \leq m : T(t) > T_r\} \leq \tau_0\big).$$

Hence, we have

$$\begin{aligned}\mathbb{P}_{\tau_0}(t_{\text{stop}} < \tau_0 + i) - \mathbb{P}_{\tau_0}(t_{\text{stop}} \leq \tau_0) &= \mathbb{P}_{\tau_0}\big(\tau_0 < \min\{1 \leq t \leq m : T(t) > T_r\} < \tau_0 + i\big) \\ &= \mathbb{P}_{\tau_0=0}\big(0 < \min\{1 \leq t \leq m : T(t) > T_r\} < i\big), \end{aligned} \tag{6.39}$$

where the above result holds because both probabilities $\mathbb{P}_{\tau_0}\big(\tau_0 < \min\{1 \leq t \leq m : T(t) > T_r\} < \tau_0 + i\big)$ and $\mathbb{P}_{\tau_0=0}\big(0 < \min\{1 \leq t \leq m : T(t) > T_r\} < i\big)$ are based on observations after the true change point $\tau_0 \in \{0, \ldots, m-1\}$ and $\tau_0 = 0$ respectively, and post-change observations have the same distribution. Then, using (6.39), we can compare the two probabilities $\mathbb{P}_{\tau_0}(t_{\text{stop}} < \tau_0 + i \mid t_{\text{stop}} > \tau_0)$ and $\mathbb{P}_{\tau_0=0}(t_{\text{stop}} < i)$ as follows

$$\begin{aligned}&\mathbb{P}_{\tau_0}(t_{\text{stop}} < \tau_0 + i \mid t_{\text{stop}} > \tau_0) - \mathbb{P}_{\tau_0=0}(t_{\text{stop}} < i) \\ &= \frac{\mathbb{P}_{\tau_0}(\tau_0 < t_{\text{stop}} < \tau_0 + i)}{\mathbb{P}_{\tau_0}(t_{\text{stop}} > \tau_0)} - \mathbb{P}_{\tau_0=0}(t_{\text{stop}} < i)\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathbb{P}_{\tau_0=0}(t_{\text{stop}} < i)}{\mathbb{P}_{\tau_0}(t_{\text{stop}} > \tau_0)} - \mathbb{P}_{\tau_0=0}(t_{\text{stop}} < i) \\
&= \mathbb{P}_{\tau_0=0}(t_{\text{stop}} < i)\Big(\frac{1 - \mathbb{P}_{\tau_0}(t_{\text{stop}} > \tau_0)}{\mathbb{P}_{\tau_0}(t_{\text{stop}} > \tau_0)}\Big) \\
&\geq 0.
\end{aligned}
$$

We therefore conclude that

$$
\sup_{0 \leq \tau_0 < m} \mathbb{E}_{\tau_0}(t_{\text{stop}} - \tau_0 \mid t_{\text{stop}} > \tau_0) = \mathbb{E}_{\tau_0=0}(t_{\text{stop}}). \tag{6.40}
$$

(b) We can write

$$
\begin{aligned}
\mathbb{E}_{\tau_0=0}(t_{\text{stop}}) &= \sum_{i=1}^{\infty} i \cdot \mathbb{P}_{\tau_0=0}(t_{\text{stop}} = i) \\
&= \mathbb{P}_{\tau_0=0}(t_{\text{stop}} = 1) + \sum_{i=2}^{\infty} i \cdot \mathbb{P}_{\tau_0=0}(t_{\text{stop}} = i) \\
&\geq \mathbb{P}_{\tau_0=0}(t_{\text{stop}} = 1).
\end{aligned} \tag{6.41}
$$

Using the result in Theorem 6.3.6 obtained under assumption (A1) - (A3), for all $\epsilon > 0$ there exists a $p_0 > 0$ such that, for every $p > p_0$, we have

$$
\mid \mathbb{P}_{\tau_0=0}(T(1) > T_r) - 1 \mid < \epsilon \Rightarrow \mathbb{P}_{\tau_0=0}(T(1) > T_r) > 1 - \epsilon \Rightarrow \mathbb{P}_{\tau_0=0}(t_{\text{stop}} = 1) > 1 - \epsilon.
$$

Since $\epsilon > 0$ is arbitrary, we must have as $p \to \infty$

$$
\mathbb{E}_{\tau_0=0}(t_{\text{stop}}) \to 1, \tag{6.42}
$$

which completes the proof.

Proof of Theorem 6.3.8

When $m \to \infty$, we have

$$
\begin{aligned}
T_r^{\text{BC}} &= \inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\text{perm}}}(u) \geq 1 - \alpha/m\right\} \\
&\to \inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\text{perm}}}(u) = 1\right\}
\end{aligned}
$$

$$= \inf\left\{u \in \mathbb{R}^+ : \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\left(T^s_{\text{perm}} \leq u\right) = 1\right\}$$
$$= \max_{1\leq s\leq S}\{T^s_{\text{perm}}\},$$

where $\alpha \in [0, 1]$. Also, using the results in part (b) of Theorem 6.3.2 and part (a) of Lemma 6.3.3, we have when $p \to \infty$ that

$$T^s_{\text{perm}} \xrightarrow{P} 0.$$

Therefore, we must have, as $m \to \infty$ and $p \to \infty$, that

$$T^{\text{BC}}_r \xrightarrow{P} 0. \tag{6.43}$$

Similarly, for $\alpha_{\text{ARL}} \in (0, 1)$, it is straightforward to obtain that as $m \to \infty$ and $p \to \infty$

$$\begin{aligned}
T^*_r &= \left(T^{\text{BC}}_r + T^{\text{ARL}}_r\right)/2 \\
&= \left(\inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\text{perm}}}(u) \geq 1 - \alpha/m\right\} + \inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\text{perm}}}(u) \geq 1 - \alpha_{\text{ARL}}\right\}\right)/2 \\
&\to \left(\inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\text{perm}}}(u) = 1\right\} + \inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\text{perm}}}(u) \geq 1 - \alpha_{\text{ARL}}\right\}\right)/2 \\
&\leq \inf\left\{u \in \mathbb{R}^+ : G_{T^S_{\text{perm}}}(u) = 1\right\} \\
&= o_P(1).
\end{aligned} \tag{6.44}$$

Using the results in (6.43) and (6.44) and noting that T^{ARL}_r is not affected when m grows, we obtain as $m \to \infty$ and $p \to \infty$ that

$$T_r \xrightarrow{P} 0. \tag{6.45}$$

The result in (6.45) is exactly the same as that in (6.32). Therefore, the previous results in (6.34) and (6.35) also hold here as $p \to \infty$ and $m \to \infty$. Note that the proof of part (b) can be carried out in the same way as in Theorem 6.3.7, since Theorem 6.3.6 holds under the same asymptotic regime. This completes the proof.

CHAPTER 7

Conclusions and future work

High dimensional change point detection is less studied compared to low dimensional change points, and classical methods are often inapplicable to high dimensional data. This thesis introduces two nonparametric approaches for such data: DCCP for offline change point detection and DC-OCP for online change point detection. These contributions advance both the methodology and theoretical understanding of high dimensional change point problems from a nonparametric perspective.

To conclude our discussion about the DCCP method, we have proposed distance-based CUSUM statistics for detecting single and multiple change points in high dimensional data sequences, which do not require normality or any other distribution for data. The distance-based CUSUM statistic can be regarded as a general purposeful alternative to the standard CUSUM statistic for high dimensional observations. A main advantage of the distance-based CUSUM statistics is their capability in detecting more general types of changes, including linear and non-linear changes such as a change in the mean, variance, correlation, and other distributional changes. Also, due to the nature of the distance functions used, our distance-based CUSUM method is particularly suitable for HDLSS data when the sample size is very small compared to the dimension. This is currently an understudied problem in the

literature of high dimensional change points. We have proved that our change point estimates are consistent under some conditions, as shown in Theorems 3.4.4 and 5.2.1. We have shown that our method can effectively detect non-sparse high dimensional change points. Moreover, the numerical results support its capability to detect more general types of changes in the underlying distribution. We have developed an R package called `distCUSUM` for the implementation of our distance-based CUSUM method, which will be made available online on GitHub at `https://github.com/lupengzhang/distCUSUM`.

For the DC-OCP method, we have established several theoretical results demonstrating the effectiveness of our method in finding an online change point without requiring knowledge of likelihood information or data distribution. This is technically demanding, as rigorous theoretical justifications in existing nonparametric work remain understudied. Under the null hypothesis of no change point, we have shown that our method controls the FWER and ARL at a pre-specified level. Under the alternative hypothesis that there is a change point in the arriving data stream, we have proved the consistency of our test and provided convergence results of EDD under some conditions. Another advantage of our method for online change points is its capability in detecting general types of changes, including linear and non-linear changes in data distribution, such as a change in the mean, variance, or correlation. In the simulation studies, we have validated the accuracy of the theoretical results, including FWER, ARL, detection power, and EDD, through empirical evidence. We also compared our DC-OCP method with several methods in the literature. The results have shown that our method performs particularly well for detecting online change points in non-normal observations. In our real data application, we have demonstrated how the proposed method can be naturally extended to detect multiple change points in an online setting. Also, we have developed an R package, called `DC-OCP`, for the implementation of our method, which will be made available online on GitHub at `https://github.com/lupengzhang/DC-OCP`.

Finally, some extensions merit further study. For the DCCP method, alternative techniques for computing multiple change points could be explored. In particular, the PELT method (Killick et al., 2012) seems an attractive approach for its computational

efficiency, though it has been developed primarily for univariate data. Our distance-based CUSUM offers a nonparametric way to adapt PELT to high dimensional multiple change points, as shown in Section 5.5. Further research is required to incorporate PELT into the proposed distance-based CUSUM statistics. For both DCCP and DC-OCP methods, it is worth mentioning that the distance-based CUSUM approach can be applied with any suitable dissimilarity measure for high dimensional data. For example, one can use the modified L_q-norm distances in (3.2) with $q > 2$. It is a topic for further research to study and discover more properties of distance-based CUSUM statistics using other dissimilarity distances. Moreover, we have shown that our methods perform well across diverse data settings, including spatially correlated variables, temporally dependent observations, and non-normal data. Future work could extend them to more complex data structures common in modern applications, such as datasets with missing values (e.g., Xie et al., 2012; Follain et al., 2022) or contaminated observations (e.g., Li and Yu, 2021).

APPENDIX A

R code for the offline change point method DCCP

We here provide the core R code for DCCP covering single and multiple change point detection with some illustrative examples. The complete version will be available at `https://github.com/lupengzhang/distCUSUM`. The function below computes the proposed distance-based CUSUM statistics (3.4) and returns the distance matrix (3.3), the CUSUM matrix (3.5), the offline change point estimate (3.6), and the test statistic (3.7). The default distance function is the modified L_1 norm in (3.2), implemented as `Fun_DCCP = M.L1norm`. To use the modified L_2 norm, we set `Fun_DCCP = dist`, and the code rescales the distance matrix by $\sqrt{p}$ accordingly. Users can also specify any other distance function by changing the argument `Fun_DCCP`.

```
M.L1norm <- function(data)##modified L1norm##
{
  p = ncol(data)
  L1norm_matrix =(1/p)*as.matrix(dist(data,method = "manhattan"))
  return(L1norm_matrix)
}


DCCP <- function(data, Fun_DCCP = M.L1norm)##DCCP without permutation##
```

```
{
  n = nrow(data)
  p = ncol(data)
  D = as.matrix(Fun_DCCP(data))
  if(identical(Fun_DCCP, dist))
  {
    D <- D/sqrt(p)
  }
  C = matrix(0, nrow = n, ncol = n-1)
  for (i in 1:n)
  {
    R = D[i, ]
    c = numeric(n-1)
    for (k in 1: (n-1))
    {
      c[k] = sqrt(length(R[1: k])*length(R[-(1: k)]))/n*(mean(R[-(1: k)]) -
    mean(R[1: k]))
    }
    C[i, ] = c
  }
  changepoint <- which.max(colSums(C^2))
  test.statistic <- (1/n)*colSums(C^2)[changepoint]
  list_all <- list("changepoint" = changepoint, "test.statistic" =
    test.statistic, "Distance matrix" = D, "CUSUM matrix" = C)
  return(list_all)
}
```

The following code implements Algorithm 1, which is the DCCP for single change point detection. It requires an $n \times p$ multivariate data matrix with $p > 1$ as input and outputs a significant change point estimate together with its corresponding p-value, or "NA" if no significant change point is found.

```
DCCP_single_changepoint <- function(data, FUN_single_changepoint = M.L1norm,
    nperm = 200, sig.lvl = 0.05) {
  if (!is.matrix(data) && !is.data.frame(data)) {
    stop("Input must be a matrix or data frame for multivariate data.")
```

```
}
if (ncol(data) <= 1) {
  stop("Data must have more than one column for multivariate change point
  detection.")
}
##DCCP
n <- nrow(data)
p <- ncol(data)
D <- as.matrix(FUN_single_changepoint(data))
if (identical(FUN_single_changepoint, dist)) {
  D <- D / sqrt(p)
}
C <- matrix(0, nrow = n, ncol = n - 1)
for (i in 1:n) {
  R <- D[i, ]
  c <- numeric(n - 1)
  for (k in 1:(n - 1)) {
    c[k] <- sqrt(length(R[1:k]) * length(R[-(1:k)])) / n * (mean(R[-(1:k)])
  - mean(R[1:k]))
  }
  C[i, ] <- c
}
changepoint <- which.max(colSums(C^2))
test.statistic <- (1/n) * colSums(C^2)[changepoint]

##permutation
permuted_test.statistics <- numeric(nperm)
for (i in 1:nperm) {
  index <- 1:n
  perm_before <- index[1:(changepoint - 1)]
  perm_after  <- index[(changepoint + 1):n]
  index <- c(perm_before, perm_after)
  permuted_index <- sample(index)
  permuted_data <- data[permuted_index, ]
  permuted_test.statistics[i] <- DCCP(data = permuted_data, Fun_DCCP =
  FUN_single_changepoint)$test.statistic
}
```

```
  pvalue_numeric <- sum(permuted_test.statistics >= test.statistic) / nperm
  pvalue_report <- if (pvalue_numeric < 0.01) "<0.01" else pvalue_numeric
  significance <- ifelse(pvalue_numeric < sig.lvl, "significant",
    "non-significant")
  ##output
  if (significance == "non-significant") {
    return(list(
      changepoint = NA,
      pvalue = pvalue_report
    ))
  } else {
    return(list(
      changepoint = changepoint,
      pvalue = pvalue_report
    ))
  }
}
```

We run two examples to illustrate how `DCCP_single_changepoint` works. The first example considers a change in the mean of observations (with correlated variables) in Section 4.5, where $n = 100$, $p = 500$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4},\, 0 \times \mathbf{1}_{p/4}\big)$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$. Here, the true change point is 60. It can be seen that the function `DCCP_single_changepoint` returns a significant change point estimate of 60, which coincides with the true change point.

```
> n <- 100
> p <- 500
> n1 <- 0.6*n
> n2 <- 0.4*n
> mu <- 0.3
> s <- 0.75
> Deltamu <- sample(c(rep(0,(1-s)*p),rep(mu,s*p)))
> rho <- 0.5
> cov_matrix <- matrix(0, nrow = p, ncol = p)
> for (i in 1:p)
+ {
```

```
+   for (j in 1:p)
+   {
+     cov_matrix[i, j] <- rho^abs(i - j)
+   }
+ }
> Obs_before <- mvrnorm(n1, mu = rep(0, p), Sigma = cov_matrix)
> Obs_after <- mvrnorm(n2, mu = Deltamu, Sigma = cov_matrix)
> Obs <- rbind(Obs_before, Obs_after)
> DCCP_single_changepoint(Obs)
$changepoint
[1] 60

$pvalue
[1] "<0.01"
```

The second example considers the case with no change point, where $n = 100$, $p = 500$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$. The function `DCCP_single_changepoint` returns `NA` with a non-significant p-value (p-value$>$ 0.05).

```
> n <- 100
> p <- 500
> n1 <- 0.6*n
> n2 <- 0.4*n
> mu <- 0
> s <- 0.75
> Deltamu <- sample(c(rep(0,(1-s)*p),rep(mu,s*p)))
> rho <- 0.5
> cov_matrix <- matrix(0, nrow = p, ncol = p)
> for (i in 1:p)
+ {
+   for (j in 1:p)
+   {
+     cov_matrix[i, j] <- rho^abs(i - j)
+   }
+ }
> Obs_before <- mvrnorm(n1, mu = rep(0, p), Sigma = cov_matrix)
> Obs_after <- mvrnorm(n2, mu = Deltamu, Sigma = cov_matrix)
```

```
> Obs <- rbind(Obs_before, Obs_after)
> DCCP_single_changepoint(Obs)
$changepoint
[1] NA

$pvalue
[1] 0.685
```

The following code implements Algorithm 2, which is DCCP for multiple change point detection with recursive binary segmentation. It takes an $n \times p$ multivariate data matrix with $p > 1$ and outputs an ordered list of significant change point estimates, or "NA" if none are found. Here we set the minimum segment length to 10 and the number of permutations to 200. If users believe there is only a single change point to detect, we recommend using the function `DCCP_single_changepoint`. Otherwise, we recommend the function `DCCP_multiple_changepoints` as the first option.

```
DCCP_multiple_changepoints <-function (data, FUN_multiple_changepoints =
    M.L1norm, minsegment = 10, nperm = 200, sig.lvl = 0.05)
{
  if (!is.matrix(data) && !is.data.frame(data)) {
    stop("Input must be a matrix or data frame for multivariate data.")
  }

  if (ncol(data) <= 1) {
    stop("Data must have more than one column for multivariate change point
    detection.")
  }
  changepoint_all <- NA
  length_data <- nrow(as.matrix(rbind(data, data)))/2
  while (length_data > minsegment)
  {
    changepoint_test <- DCCP_single_changepoint(data = data,
    FUN_single_changepoint = FUN_multiple_changepoints, nperm = nperm,
    sig.lvl = sig.lvl)
    changepoint <- changepoint_test$changepoint
    changepoint_pvalue <- changepoint_test$pvalue
```

```
    if (is.character(changepoint_pvalue)) {
      if (grepl("^<", changepoint_pvalue)) {
        changepoint_pvalue <- as.numeric(sub("^<", "", changepoint_pvalue)) *
    0.999
      } else {
        changepoint_pvalue <- as.numeric(changepoint_pvalue)
      }
    }
    end <- nrow(data)
    if (!is.na(changepoint) && changepoint_pvalue <= sig.lvl & changepoint >
    2 & changepoint <
        end - 2) {
      data_before <- data[1:changepoint, ]
      data_after <- data[(changepoint + 1):end, ]
      changepoint_before <- DCCP_multiple_changepoints(data = data_before,
    FUN_multiple_changepoints = FUN_multiple_changepoints, minsegment =
    minsegment, nperm = nperm, sig.lvl = sig.lvl)
      changepoint_after <- DCCP_multiple_changepoints(data = data_after,
    FUN_multiple_changepoints = FUN_multiple_changepoints, minsegment =
    minsegment, nperm = nperm, sig.lvl = sig.lvl) + changepoint
      changepoint_all <- c(changepoint, changepoint_before, changepoint_after)
      return(sort(changepoint_all[!is.na(changepoint_all)]))
    }
    else {
      return(NA)
    }
  }
  return(sort(changepoint_all[!is.na(changepoint_all)]))
}
```

We run the following example to illustrate how `DCCP_multiple_changepoints` works. It considers changes in the mean of observations in Subsection 5.3.1, where $n = 100$, $p = 500$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4}, 0 \times \mathbf{1}_{p/4}\big)$, $\boldsymbol{\mu}_3 = 2\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = 3\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4 = \boldsymbol{V}_p$. The true change points are $\tau_1 = 20$, $\tau_2 = 60$, and $\tau_3 = 80$, which are exactly detected by the following code.

```
> n <- 100
> p <- 500
> n1 <- 0.2*n
> n2 <- 0.4*n
> n3 <- 0.2*n
> n4 <- 0.2*n
> mu <- 0.3
> s <- 0.75
> Deltamu <- sample(c(rep(0,(1-s)*p),rep(mu,s*p)))
> rho <- 0.5
> cov_matrix <- matrix(0, nrow = p, ncol = p)
> for (i in 1:p)
+ {
+   for (j in 1:p)
+   {
+     cov_matrix[i, j] <- rho^abs(i - j)
+   }
+ }
> Obs_1 <- mvrnorm(n1, mu = rep(0, p), Sigma = cov_matrix)
> Obs_2 <- mvrnorm(n2, mu = Deltamu, Sigma = cov_matrix)
> Obs_3 <- mvrnorm(n3, mu = 2*Deltamu, Sigma = cov_matrix)
> Obs_4 <- mvrnorm(n4, mu = 3*Deltamu, Sigma = cov_matrix)
> Obs <- rbind(Obs_1, Obs_2, Obs_3, Obs_4)
> DCCP_multiple_changepoints(Obs)
[1] 20 60 80
```

APPENDIX B

R code for the online change point method DC-OCP

We here provide the core R code for DC-OCP for online change point detection. The complete version will be available at `https://github.com/lupengzhang/DC-OCP`. Building on the distance-based CUSUM code `DCCP` in Appendix A, `DCOCP` performs sequential testing and returns the stopping time if a change occurs. Otherwise, it returns "NA" if no significant change point is found in the arriving data. The idea of the function `DCOCP` is given in Algorithm 3 in Section 6.2. The procedure takes an $n \times p$ historical data matrix and an $m \times p$ arriving data matrix. We do not assume a known distribution for the historical data, but we require that it contains no change points. This can be checked using our offline code in Appendix A or other existing methods such as E.divisive, Inspect, or HDcp.

```
DCOCP <- function(historical_data, arriving_data, Fun_DCOCP = M.L1norm, w =
    50, nperm = 10000, threshold_choice = c("BC","ARL","star"), alpha_BC =
    0.05, alpha_ARL = 0.001, return_details = FALSE)
{
  ## data check
  threshold_choice <- match.arg(threshold_choice)
  if (!is.matrix(historical_data) && !is.data.frame(historical_data))
    stop("historical_data must be a matrix or data frame.")
```

```
if (!is.matrix(arriving_data) && !is.data.frame(arriving_data))
  stop("arriving_data must be a matrix or data frame.")
if (ncol(historical_data) <= 1)
  stop("Data must have >1 column for multivariate detection.")
if (ncol(historical_data) != ncol(arriving_data))
  stop("historical_data and arriving_data must have the same number of
  columns.")

n <- nrow(historical_data)
m <- nrow(arriving_data)
if (n < w) stop("historical_data must contain at least w rows.")

## threshold (permute last w rows of historical data)
recent_data <- historical_data[(n - w + 1):n, , drop = FALSE]
permuted_stats <- numeric(nperm)
for (r in seq_len(nperm)) {
  permuted_stats[r] <- DCCP(recent_data[sample.int(w), , drop = FALSE],
  Fun_DCCP = Fun_DCOCP)$test.statistic
}
threshold_BC   <- as.numeric(quantile(permuted_stats, 1 - alpha_BC / m))
threshold_ARL  <- as.numeric(quantile(permuted_stats, 1 - alpha_ARL))
threshold_star <- (threshold_BC + threshold_ARL) / 2
threshold <- switch(threshold_choice, "BC"   = threshold_BC, "ARL"  =
  threshold_ARL, "star" = threshold_star)

## detection
detection_flag  <- FALSE
detection_times <- n + m
i <- 1
while (!detection_flag && i <= m) {
  end_idx   <- n + i
  start_idx <- end_idx - w + 1
  if (start_idx < 1) { i <- i + 1; next }
  testing_data <- rbind(historical_data, arriving_data[seq_len(i), , drop =
  FALSE])[start_idx:end_idx, , drop = FALSE]
  observed_stat <- DCCP(testing_data, Fun_DCCP = Fun_DCOCP)$test.statistic
  if (observed_stat > threshold) {
```

```
      detection_flag  <- TRUE
      detection_times <- end_idx
      cat("Change point detected at time:", detection_times,
    "threshold_choice:", threshold_choice, "\n")
      break
    }
    i <- i + 1
  }
  if (!detection_flag) {
    cat("No change point detected in the arriving data.", "\n")
    detection_times <- NA
  }

  if (!return_details) return(detection_times)
  return(list(
    detection_time  = detection_times,
    threshold_choice= threshold_choice,
    thresholds      = list(threshold_BC = threshold_BC, threshold_ARL =
    threshold_ARL, threshold_star = threshold_star),
    permuted_stats  = permuted_stats
  ))
}
```

We run two examples to illustrate how the function `DCOCP` works. The first example considers a change in the mean of observations with correlated variables in Subsection 6.4.1. Here we set $n = 100$, $m = 300$, $p = 500$, $w = 50$, $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \big(0.3 \times \mathbf{1}_{3p/4},\, 0 \times \mathbf{1}_{p/4}\big)$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$. We use the threshold in (6.7) with nominal level $\alpha = 0.05$ to terminate the sequential testing. Here, the true change point arrives at time 160, since $n + 0.2m = 160$. The function `DCOCP` returns a stopping time $t = 164$, which gives a detection delay of 4 in this run.

```
> n <- 100
> m <- 300
> p <- 500
> w <- 50
> mu <- 0.3
```

```
> s <- 0.75
> Deltamu <- sample(c(rep(0,(1-s)*p),rep(mu,s*p)))
> rho <- 0.5
> cov_matrix <- matrix(0, nrow = p, ncol = p)
> for (i in 1:p)
+ {
+   for (j in 1:p)
+   {
+     cov_matrix[i, j] <- rho^abs(i - j)
+   }
+ }
> Obs_historical <- mvrnorm(n, mu = rep(0, p), Sigma = cov_matrix)
> Obs_before <- mvrnorm(0.2 * m, mu = rep(0, p), Sigma = cov_matrix)
> Obs_after <- mvrnorm(0.8 * m, mu = Deltamu, Sigma = cov_matrix)
> Obs_arrive <- rbind(Obs_before, Obs_after)
> DCOCP(Obs_historical, Obs_arrive)
Change point detected at time: 164 threshold_choice: BC
[1] 164
```

The second example is the case with no change point in Subsection 6.4.2, where $n = 100$, $m = 300$, $p = 500$, $w = 50$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}_p$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{V}_p$. The function `DCOCP` returns `NA`.

```
> n <- 100
> m <- 300
> p <- 500
> w <- 50
> mu <- 0
> s <- 0.75
> Deltamu <- sample(c(rep(0,(1-s)*p),rep(mu,s*p)))
> rho <- 0.5
> cov_matrix <- matrix(0, nrow = p, ncol = p)
> for (i in 1:p)
+ {
+   for (j in 1:p)
+   {
+     cov_matrix[i, j] <- rho^abs(i - j)
```

```
+   }
+ }
> Obs_historical <- mvrnorm(n, mu = rep(0, p), Sigma = cov_matrix)
> Obs_before <- mvrnorm(0.2 * m, mu = rep(0, p), Sigma = cov_matrix)
> Obs_after <- mvrnorm(0.8 * m, mu = Deltamu, Sigma = cov_matrix)
> Obs_arrive <- rbind(Obs_before, Obs_after)
> DCOCP(Obs_historical, Obs_arrive)
No change point detected in the arriving data.
[1] NA
```

Bibliography

[1] Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

[2] Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J. L., et al. (2013). A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, pages 3–4.

[3] Apostol, T. M. (1967). *Calculus*. Wiley, New York.

[4] Aue, A. and Kirch, C. (2024). The state of cumulative sum sequential changepoint testing 70 years after Page. *Biometrika*, 111(2):367–391.

[5] Austin, E., Romano, G., Eckley, I. A., and Fearnhead, P. (2023). Online non-parametric changepoint detection with application to monitoring operational performance of network devices. *Computational Statistics & Data Analysis*, 177:107551.

[6] Avanesov, V. and Buzun, N. (2018). Change-point detection in high-dimensional covariance structure. *Electronic Journal of Statistics*, 12(2):3254 – 3294.

[7] Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):649–672.

[8] Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice Hall.

[9] Biswas, M., Mukhopadhyay, M., and Ghosh, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, 101(4):913–926.

[10] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

[11] Cai, T. T., Ren, Z., and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1 – 59.

[12] Carlstein, E. (1988). Nonparametric change-point estimation. *The Annals of Statistics*, 16(1):188–197.

[13] Chakraborty, S. and Zhang, X. (2021). High-dimensional change-point detection using generalized homogeneity metrics. *arXiv preprint arXiv:2105.08976*.

[14] Chan, H. P. (2017). Optimal sequential detection in multi-stream data. *The Annals of Statistics*, 45(6):2736 – 2763.

[15] Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, 43(1):139 – 176.

[16] Chen, J. and Gupta, A. (2013). *Parametric Statistical Change Point Analysis*. Springer Science & Business Media.

[17] Chen, Y., Wang, T., and Samworth, R. J. (2022). High-dimensional, multiscale online changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):234–266.

[18] Chen, Z. and Tian, Z. (2010). Modified procedures for change point monitoring in linear models. *Mathematics and Computers in Simulation*, 81(1):62–75.

[19] Chu, L. and Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *The Annals of Statistics*, 47(1):382 – 414.

[20] Csörgö, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics.

[21] Dette, H., Pan, G., and Yang, Q. (2022). Estimating a change point in a sequence of very high-dimensional covariance matrices. *Journal of the American Statistical Association*, 117(537):444–454.

[22] Dörnemann, N. and Dette, H. (2024). Detecting change points of covariance matrices in high dimensions. *arXiv preprint arXiv:2409.15588*.

[23] Drikvandi, R. and Modarres, R. (2025). A distribution-free method for change point detection in non-sparse high dimensional data. *Journal of Computational and Graphical Statistics*, 34(1):290–305.

[24] Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *The Annals of Statistics*, 19(3):1471–1495.

[25] Eagle, N. and Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10:255–268.

[26] Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051 – 2079.

[27] Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):589–605.

[28] Follain, B., Wang, T., and Samworth, R. J. (2022). High-dimensional changepoint estimation with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):1023–1055.

[29] Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

[30] Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*, 49(4):1027–1070.

[31] Garreau, D. and Arlot, S. (2018). Consistent change-point detection with kernels. *Electronic Journal of Statistics*, 12(2):4440 – 4486.

[32] Giraud, C. (2021). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

[33] Gombay, E. and Horvath, L. (1990). Asymptotic distributions of maximum likelihood tests for change in the mean. *Biometrika*, 77(2):411–414.

[34] Gösmann, J., Stoehr, C., Heiny, J., and Dette, H. (2022). Sequential change point detection in high dimensional time series. *Electronic Journal of Statistics*, 16(1):3608–3671.

[35] Grundy, T., Killick, R., and Mihaylov, G. (2020). High-dimensional change-point detection via a geometrically inspired mapping. *Statistics and Computing*, 30(4):1155–1166.

[36] Hahn, G., Fearnhead, P., and Eckley, I. A. (2020). Bayesproject: Fast computation of a projection direction for multivariate changepoint detection. *Statistics and Computing*, 30:1691–1705.

[37] Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3):427–444.

[38] Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17.

[39] Horváth, L., Kühn, M., and Steinebach, J. (2008). On the performance of the fluctuation test for structural change. *Sequential Analysis*, 27(2):126–140.

[40] Jewell, S., Fearnhead, P., and Witten, D. (2022). Testing for a change in mean after changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1082–1104.

[41] Jirak, J. M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, 43(6):2451–2483.

[42] Kiefer, J. (1959). K-Sample Analogues of the Kolmogorov-Smirnov and Cramér-V. Mises Tests. *The Annals of Mathematical Statistics*, 30(2):420–447.

[43] Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

[44] Kirch, C. and Weber, S. (2018). Modified sequential change point procedures based on estimating functions. *Electronic Journal of Statistics*, 12(1):1579 – 1613.

[45] Kovács, S., Bühlmann, P., Li, H., and Munk, A. (2023). Seeded binary segmentation: a general methodology for fast and optimal changepoint detection. *Biometrika*, 110(1):249–256.

[46] Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information theory*, 44(7):2917–2929.

[47] Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika*, 105(3):529–546.

[48] Li, J. (2020). Asymptotic distribution-free change-point detection based on interpoint distances for high-dimensional data. *Journal of Nonparametric Statistics*, 32(1):157–184.

[49] Li, J., Xu, M., Zhong, P.-S., and Li, L. (2019a). Change point detection in the mean of high-dimensional time series data under dependence. *arXiv preprint arXiv:1903.07006*.

[50] Li, L. and Li, J. (2023). Online change-point detection in high-dimensional covariance structure with application to dynamic networks. *Journal of Machine Learning Research*, 24(51):1–44.

[51] Li, M. and Yu, Y. (2021). Adversarially robust change point detection. *Advances in Neural Information Processing Systems*, 34:22955–22967.

[52] Li, S., Xie, Y., Dai, H., and Song, L. (2019b). Scan b-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544.

[53] Li, Z. and Gao, J. (2024). Efficient change point detection and estimation in high-dimensional correlation matrices. *Electronic Journal of Statistics*, 18(1):942–979.

[54] Liu, B., Zhang, X., and Liu, Y. (2022). High dimensional change point inference: Recent developments and extensions. *Journal of Multivariate Analysis*, 188:104833.

[55] Lorden, G. (1971). Procedures for reacting to a change in distribution. *The annals of mathematical statistics*, pages 1897–1908.

[56] Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, 3(2):637–662.

[57] Maa, J.-F., Pearl, D. K., and Bartoszyński, R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *The Annals of Statistics*, 24(3):1069–1074.

[58] Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and computing*, 27(2):519–533.

[59] 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(505):334–345.

[60] Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, 97(2):419–433.

[61] Montgomery, D. C. (2007). *Introduction to statistical quality control*. John Wiley & Sons.

[62] Padilla, O. H. M., Yu, Y., Wang, D., and Rinaldo, A. (2021). Optimal nonparametric change point analysis. *Electronic Journal of Statistics*, 15(1):1154 – 1201.

[63] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

[64] Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 3571–3577. AAAI Press.

[65] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

[66] Reddi, S., Ramdas, A., Póczos, B., Singh, A., and Wasserman, L. (2015). On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial intelligence and statistics*, pages 772–780. PMLR.

[67] Reyes-Ortiz, J. L., Anguita, D., Ghio, A., Oneto, L., and Parra, X. (2013). Human activity recognition using smartphones. UCI Machine Learning Repository. `https://doi.org/10.24432/C54S4K`.

[68] Romano, G., Eckley, I. A., Fearnhead, P., and Rigaill, G. (2023). Fast online changepoint detection via functional pruning cusum statistics. *Journal of Machine Learning Research*, 24(81):1–36.

[69] Safikhani, A. and Shojaie, A. (2022). Joint structural break detection and parameter estimation in high-dimensional nonstationary var models. *Journal of the American Statistical Association*, 117(537):251–264.

[70] Siegmund, D. and Venkatraman, E. S. (1995). Using the Generalized Likelihood Ratio Statistic for Sequential Detection of a Change-Point. *The Annals of Statistics*, 23(1):255 – 271.

[71] 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(393):199–204.

[72] Statista Research Department (2022). Weekly Development of the S&P 500 Index from January 2020 to September 2022. `https://www.statista.com/statistics/1104270/weekly-sandp-500-index-performance/`.

[73] Szekely, G. J., Rizzo, M. L., et al. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 22(2):151–184.

[74] Tartakovsky, A., Nikiforov, I., and Basseville, M. (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press.

[75] Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

[76] Vostrikova, L. Y. (1981). Detecting "disorder" in multidimensional random processes. In *Doklady Akademii Nauk*, volume 259, pages 270–274. Russian Academy of Sciences.

[77] Waldmann, K.-H. (1996). Design of double cusum quality control schemes. *European Journal of Operational Research*, 95(3):641–648.

[78] Wang, H. and Xie, Y. (2024). Sequential change-point detection: Computation versus statistical performance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(1):e1628.

[79] Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):57–83.

[80] Wei, S. and Xie, Y. (2026). Online kernel cusum for change-point detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkag020.

[81] Xiao, W., Huang, X., He, F., Silva, J., Emrani, S., and Chaudhuri, A. (2019). Online robust principal component analysis with change point detection. *IEEE Transactions on Multimedia*, 22(1):59–68.

[82] Xie, L., Moustakides, G. V., and Xie, Y. (2023). Window-limited CUSUM for sequential change detection. *IEEE Transactions on Information Theory*, 69(9):5990–6005.

[83] Xie, Y., Huang, J., and Willett, R. (2012). Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27.

[84] Xie, Y. and Siegmund, D. (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics*, 41(2):670 – 692.

[85] Yao, Y.-C. (1988). Estimating the number of change-points via schwarz' criterion. *Statistics & Probability Letters*, 6(3):181–189.

[86] Yilmaz, Y. (2017). Online nonparametric anomaly detection based on geometric entropy minimization. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3010–3014. IEEE.

[87] Yu, Y., Madrid Padilla, O. H., Wang, D., and Rinaldo, A. (2023). A note on online change point detection. *Sequential Analysis*, 42(4):438–471.

[88] Zeileis, A., Kleiber, C., Krämer, W., and Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(1-2):109–123.

[89] Zhang, L. and Drikvandi, R. (2023). High Dimensional Change Points: Challenges and Some Proposals. In *Proceedings of the 5th International Conference on Statistics: Theory and Applications*. Paper No. 142.

[90] Zhang, L. and Drikvandi, R. (2025). Distance-based CUSUM statistics for high dimensional change points. *Statistics and Computing*, 35(6):1–19.

[91] Zhang, L., Drikvandi, R., and Chen, Y. (2025). Nonparametric online change point detection in high dimensional data streams. Submitted.

[92] Zhong, P.-S., Li, J., and Kokoszka, P. (2021). Multivariate analysis of variance and change points estimation for high-dimensional longitudinal data. *Scandinavian Journal of Statistics*, 48(2):375–405.

[93] Zhu, C. and Shao, X. (2021). Interpoint distance based two sample tests in high dimension. *Bernoulli*, 27(2):1189 – 1211.

[94] Zhu, C., Zhang, X., Yao, S., and Shao, X. (2020). Distance-based and rkhs-based dependence metrics in high dimension. *The Annals of Statistics*, 48(6):3366–3394.

[95] Zou, C., Wang, Z., Zi, X., and Jiang, W. (2015). An efficient online monitoring method for high-dimensional data streams. *Technometrics*, 57(3):374–387.