High-dimensional Change-point Estimation Under Structural Assumptions

Hanqing Cai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Statistical Science
University College London

April 2, 2024

Declaration

I, Hanqing Cai, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Change-point analysis has been successfully applied to detect changes in multivariate and high-dimensional data streams over time. However, many existing methods did not consider the additional structures that data may possess. In this thesis, we study the problem of high-dimensional change-point estimation under structural assumptions. We mainly study two structures: group sparsity structure and network structure. For group sparsity structure, we assume that coordinates in mean vectors are naturally divided into groups and changes only occur in a small subset of groups. We propose `groupInspect` which uses the group information to estimate a projection direction so as to aggregate information across the component series to estimate the change-point in the mean under this structure. For network structure, we assume that coordinates are connected into a network, and changes start from a source coordinate and then spread out to the neighbouring coordinates. We propose `SpreadDetect` to estimate the initial time of change as well as the location of the source coordinate of change. For both algorithms, we provide theoretical guarantees on our proposed estimators. We also demonstrate the performance of the two algorithms using simulation studies and real-data examples.

Impact Statement

Streaming data has become an increasingly important data type since the advent of the Internet of Things. In many applications, one is interested in estimating changes in the data distribution in the data stream, for example, the amount of greenhouse gases in the atmosphere, fMRI imaging data and stock prices. Many methods have been proposed to estimate the changes in distributions for high-dimensional data. A common assumption is sparsity where the changes only occur in a small subset of coordinates. However, this assumption often does not capture the full structure of the data. In this thesis, we propose change-point estimation algorithms that exploit two additional structures.

The first structure is the group sparsity structure (in Chapter 3) which assumes that the coordinates are clustered into groups and only a small subset of the groups may experience changes. We propose a new algorithm to estimate the change in the mean vector under this structure which first seeks an optimal projection direction to project the data into a one-dimensional series and then locate the change. The algorithm can be combined with existing top-down methods to estimate multiple change-points recursively. We provide theoretical guarantees on the change-point location estimator under both single and multiple change-point cases. We also extend the theory to settings where data follows sub-Gaussian distributions or has temporal dependence. The simulation studies also demonstrate the good performance over other existing methods under this group sparsity assumption. In addition, this algorithm can be used to solve real-world problems when data exhibit group sparsity structure. This includes financial data streams where changes are often grouped by industry sectors, and functional magnetic resonance imaging data where temporal changes are clustered by voxel locations within the brain. In this thesis, we also present a real-world example with an S&P 500 stock price dataset.

The second structure we consider in this thesis is the network structure (Chapter 4). Although there are many existing methods for change-point analysis with network structures, we consider a different setting of spreading change in this work. To be more specific, the coordinates in the mean vector are connected into a network and initially, the change occurs in a single coordinate (source coordinate) and then spreads out to the neighbour-

ing coordinates. As the change is quite sparse and the signal is weak initially, existing methods may not be able to detect it. We proposed a new algorithm, `SpreadDetect`, which can consistently estimate both the initial time of change and the location of the source coordinate. We also provide the theoretical guarantees for the estimators and perform simulation studies to show the good performance of our method. This spreading assumption is also of practical use, for example, locating the initial time point and the individual (source coordinate) in the spread of infectious disease between individuals over time. We also provide an example by applying our method to a US COVID-19 weekly excess death data.

Publications from this work:

- Chapter 3 has been published as: Cai, H. and Wang, T. (2023) Estimation of high-dimensional change-points under a group sparsity structure. *Electronic Journal of Statistics*, **17**, 858–894.

- Chapter 4 has been uploaded to arXiv as: Cai, H. and Wang, T. (2023) Spread-Detect: Detection of spreading change in a network over time. *arXiv preprint*, arXiv:2306.10475.

Acknowledgement

First, I would like to express my sincere gratitude to my primary supervisor Dr Tengyao Wang. I am thankful for his patient guidance and assistance throughout my PhD. I benefit a lot from our regular meetings.

I sincerely thank my secondary supervisor, Dr Yvo Pokern, for his support during my PhD.

I sincerely thank the Department for offering me the Teaching Assistantship and their assistance and support over the last four years.

Finally, I would like to thank my family and friends for their love and support.

Contents

Chapter 1

Introduction

Modern applications routinely generate time-ordered high-dimensional datasets, where many covariates are simultaneously measured over time. Examples include climate data that tracks the amount of greenhouse gases in the atmosphere (Reeves et al., 2007); (Itoh and Kurths, 2010), wearable technologies recording the health state of individuals from multi-sensor feedbacks (Hanlon and Anderson, 2009), internet traffic data collected by tens of thousands of routers (Peng, Leckie and Ramamohanarao, 2004) and functional Magnetic Resonance Imaging (fMRI) scans that record the time evolution of blood oxygen level dependent (BOLD) chemical contrast in different areas of the brain (Aston and Kirch, 2012). The explosion in the number of such high-dimensional data streams calls for methodological advances for their analysis.

Change-point analysis is an essential statistical technique used in identifying abrupt changes in a time series. Time points at which such abrupt change occurs are called 'change-points'. By estimating the location of change-points, we can divide the time series into shorter segments that can be analysed using methods designed for stationary time series. Moreover, in many applications, the estimated change-points indicate specific events that are themselves of great interest. In the examples mentioned in the previous paragraph, they can be used to raise alarms about certain climate changes, abnormal health events, detect distributed denial of service attacks on the network and pinpoint the onset of certain brain activities.

Classical change-point analysis focuses on univariate time series. The current state-of-art methods including Killick, Fearnhead and Eckley (2012); Frick, Munk and Sieling (2014); Fryzlewicz (2014). However, classical univariate change-point methods are often inadequate for high-dimensional datasets that are routinely encountered in modern applications. When applied componentwise, they are often sub-optimal as signals can spread over many components. Recently, several methodologies have been proposed to test and estimate change-points in high-dimensional settings by borrowing strength across multiple coordinates to detect and localise change-points at a higher accuracy than would otherwise be possible using univariate change-point algorithms alone. These methods include ℓ_2 or ℓ_∞ aggregation of the cumulative sums (CUSUMs) test statistics across different components proposed by Horváth and Hušková (2012); Jirak (2015), the Sparsified Binary Segmentation algorithm by Cho and Fryzlewicz (2015), the double CUSUM algorithm of Cho (2016) and a projection-based approach by Wang and Samworth (2018).

However, in order to handle the high-dimensional nature of the problem, the multivariate or high-dimensional methods mentioned in the previous paragraph often make simplifying assumptions such as all coordinates are exchangeable or that changes are located in a sparse subset of coordinates. In reality, in many applications, there are additional structures in the change-points that one can exploit to improve the estimation accuracy. Examples include group structures where coordinates form natural groups and changes tend to occur within the same group (Wang et al., 2021), and community structures where nodes belong to different (unknown) communities and may switch community at the change-point (Wang, Yu and Rinaldo, 2021).

In this thesis, we focus on the high-dimensional change-point estimation problem under structural assumptions. In Chapter 2, we first review some relevant literature on change-point analysis, including classic offline change-point estimation procedures, recent developments on high-dimensional change-point analysis and online change-point detection problems, as well as high-dimensional problems with structural assumptions. In Chapter 3, we study the group sparsity structure that coordinates are naturally divided into groups and only a small subset will undergo changes. We propose a new change-point estimation procedure, named `groupInspect` which uses the pre-specified group information to estimate a projection direction and then locate the change-point by applying a univariate change-point estimation method to the projected series. The algorithm can be combined with a top-down method to identify multiple change-points. In Chapter 4, we consider the structure where the coordinates represent nodes of a graph/network and the change initially appears in one coordinate (the *source coordinate* of change) and then spreads across the network gradually over time. We propose a method called `SpreadDetect`, that can estimate both the source coordinate and the initial change-point time. The idea is to aggregate the CUSUM statistics across multiple coordinates with suitable time lags according to the given network structure information and we propose both quadratic and linear test statistics.

1.1 Notations

We close this chapter by introducing the notations used in this thesis. For $n \in \mathbb{N}$, we write $[n] = \{1, \ldots, n\}$. For a vector $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$, we define $\|v\|_0 = \sum_{i=1}^n \mathbb{1}_{\{v_i \neq 0\}}$, $\|v\|_\infty = \max_{i \in [n]} |v_i|$ and $\|v\|_q = \left\{\sum_{i=1}^n (v_i)^q\right\}^{1/q}$ for any positive integer q, and let $\mathbb{S}^{n-1} = \{v \in \mathbb{R}^n : \|v\|_2 = 1\}$. For a matrix $A \in \mathbb{R}^{p \times n}$, we write $\|A\|_* = \sum_{i=1}^{\min(p,n)} \sigma_i(A)$ for its nuclear norm, $\|A\|_{\mathrm{op}} = \max_i \sigma_i(A)$ for its operator norm, where $\sigma_1, \ldots, \sigma_{\min(p,n)}(A)$ are its singular values. We write $\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^n \sum_{j=1}^p A_{ij}^2}$ for its Frobenius norm.

For any $S \subseteq [n]$, we write v_S for the $|S|$-dimensional vector obtained by extracting coordinates of v in S. For a matrix $A \in \mathbb{R}^{p \times n}$, $J \in [p]$ and $S \in [n]$, we write $A_{J,S}$ for the submatrix obtained by extracting rows and columns of A indexed by J and S respectively. When $S = [n]$, we abbreviate $A_{J,[n]}$ by A_J. When $S = \{t\}$ is a single element set, we slightly abuse notation and write $A_{J,t}$ instead of $A_{J,\{t\}}$.

We use $\circ$ to denote the Hadamard product. Given two sequences $(a_n)_{n \in \mathbb{N}}$ and $(b_n)_{n \in \mathbb{N}}$ such that $a_n, b_n > 0$ for all n, we write $a_n \lesssim b_n$ (or equivalently $b_n \gtrsim a_n$) if $a_n \leqslant C b_n$ for some universal constant C. We denote $j = \lceil p \rceil$ if j is the smallest integer such that $j \geqslant p$ and denote $j = \lfloor p \rfloor$ if j is the largest integer such that $j \leqslant p$. We write $a_n \asymp b_n$ if $0 < \liminf_{n \to \infty} |a_n / b_n| \leqslant \limsup_{n \to \infty} |a_n / b_n| < \infty$.

Chapter 2

Literature review

In this chapter, we review relevant literature in change-point analysis, as well as other high-dimensional problems where structural assumptions can also be exploited to improve statistical inference. In change-point analysis, we will mainly emphasise on the literature on offline change-point estimation as this is also the key focus of this thesis. We start by revisiting some classic one-dimensional estimation procedures and then review some more recent developments in high-dimensional change-point settings. In addition, we will also give a brief overview of the literature on the online change-point problems. Finally, we conclude this chapter by discussing some other high-dimensional problems where structural assumptions are used to improve inference.

2.1 Univariate offline change-point estimation problem

In the offline change-point analysis, we usually have access to the entire dataset prior to performing statistical analysis. In the general setup of the problem, we are presented with a data sequence $X = (X_1, \ldots, X_n)$, such that for some (unknown) time points $1 \leqslant z_1 < z_2 < \cdots < z_\nu \leqslant n-1$, the marginal distributions of the elements in the sequence satisfy

$$X_t \sim F_i, \quad \text{for } z_i + 1 \leqslant t \leqslant z_{i+1}.$$

for $i \in \{0, \dots, \nu\}$ (by convention, we set $z_0 = 0$ and $z_{\nu+1} = n$). We denote f_i as the density of F_i. The goal is to estimate $z_1, \dots, z_\nu$ given data X.

2.1.1 CUSUM-based approaches

In this subsection, we assume that $p = 1$. In general, there are two types of methods for univariate change-point estimation. The first is based on cumulative sum (CUSUM) statistic. For a data sequence $X = (X_t)_{t \in [n]}$, its CUSUM transformation at time t between the segment $(s, e]$ (for $s < t < e$) is defined as:

$$\mathcal{T}_t^{s,e}(X) = \sqrt{\frac{(t-s)(e-t)}{e-s}} \left(\frac{1}{e-t} \sum_{r=t+1}^{e} X_t - \frac{1}{t-s} \sum_{r=s+1}^{t} X_t \right) \tag{2.1}$$

$$= \sqrt{\frac{e-s}{(t-s)(e-t)}} \left(\frac{t-s}{e-t} \sum_{r=s+1}^{e} X_t - \sum_{s=1}^{t} X_t \right) \tag{2.2}$$

Equation (2.1) and (2.2) are two commonly used equivalent ways of defining the CUSUM transformation. If $X_1, \dots, X_n$ are independent and identically distributed normal random variables, $|\mathcal{T}_t^{0,n}(X)|$ can be viewed as the generalised likelihood ratio statistic for testing the null hypothesis that there is no change against the alternative that there is a change in mean at time t. In the single change-point case, we can detect a change-point if $\max_{1 \leqslant t \leqslant n-1} |\mathcal{T}_t^{0,n}(X)|$ is above a certain threshold and consequently estimate its location by the location of the maximum.

While the above CUSUM-based method is designed for estimating a single change-point, it can be applied in conjunction with a top-down approach to estimate multiple change-points recursively. Binary Segmentation (BS) proposed by Scott and Knott (1974) is one of the most widely used top-down method for locating multiple change-points. It starts by applying a single change-point procedure to the entire data sequence $(X_t)_{t \in (0,n]}$ to test the existence of a change-point and, if present, estimate its location $\hat{z}$. It then splits the original data into two subsequences $(X_t)_{t \in (0,\hat{z}]}$ and $(X_t)_{t \in (\hat{z},n]}$ and repeats the same process, until no more change-points can be detected in all the subsegments (see Algorithm 1 for a pseudocode).

However, it should be noted that when used in conjunction with the BS approach to

Algorithm 1: Pseudocode for binary segmentation

```
Input: data sequence $X = (X_t)_{t \in [n]}$, a single change-point test $\psi : \mathbb{R}^* \to \{0,1\}$
       and a single change-point estimator $\eta : \mathbb{R}^* \to \mathbb{N}$
Set $\hat{Z} \leftarrow \emptyset$
Function BS($s$, $e$):
    if $\psi((X_t)_{t \in (s,e]}) = 1$ then
        $\hat{z} \leftarrow \eta((X_t)_{t \in (s,e]}) + s$
        $\hat{Z} \leftarrow \hat{Z} \cup \{\hat{z}\}$
        Run recursively BS($s$, $\hat{z}$) and BS($\hat{z}$, $e$).
Run BS($0$, $n$)
Output: $\hat{Z}$
```

estimate multiple change-points, the CUSUM estimator is applied in a misspecified way. This is because the mean of the CUSUM statistics $(\mathcal{T}_t^{s,e})_{t \in (s,e)}$ is unimodal peaking at the true change-point when there is only a single change-point present in the segment $(s, e]$, but when more than one change-points are present, neighbouring change-points may offset each other in CUSUM calculation and the series may have a much less well-defined peak at each of the true change-point locations. Such misspecification can lead to the sub-optimality of BS in some scenarios. Figure 2.1 gives an illustration of this situation. We randomly generated 300 independent random variables from a normal distribution with a change in mean at $t = 120, 150, 180$. From the dotted curve, it can be seen that none of the true change-points can be identified if we compute CUSUM statistics on the entire series. However, if we look at the CUSUM curve calculated from a short interval around 150, the CUSUM statistics do appear to have an obvious peak at the true change-point $t = 150$.

To remedy the problem of BS mentioned above, Fryzlewicz (2014) proposed a multiple change-point estimation procedure, named Wild Binary Segmentation (WBS). The main idea is that instead of computing global CUSUM statistics from the entire data sequence, we first randomly draw a large number of intervals. Then for each interval, we apply

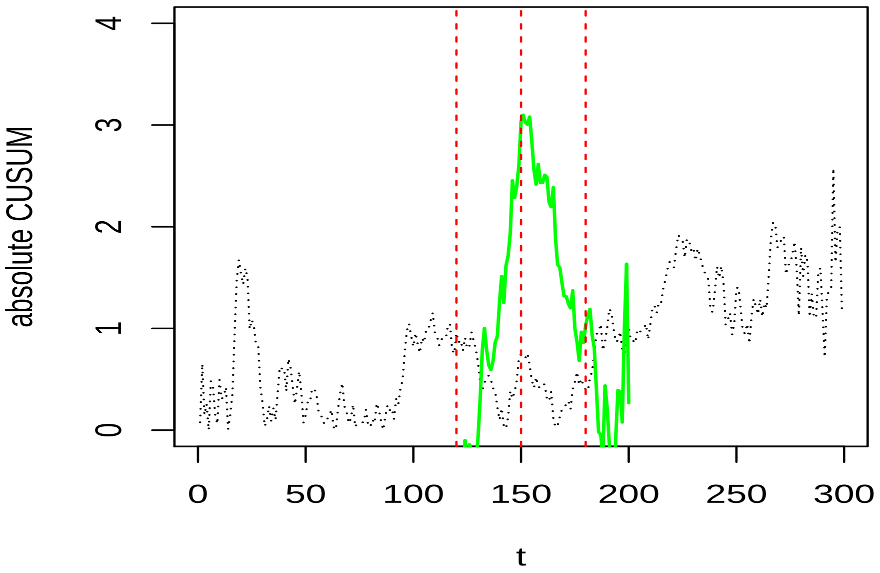

Figure 2.1: Comparisons of CUSUM statistics calculated from the entire data series (the dotted line) and a short window around $t = 150$ (the green curve). The dotted line represents the absolute CUSUM statistics computed on the entire data series used for binary segmentation. The true change points at 120, 150 and 180 are shown with vertical dashed lines.

a single change-point algorithm to find a candidate change-point, which for a CUSUM-based approach is the time point at which the absolute CUSUM statistic is maximise. We then pick the best candidate change-point to be the one where the associated test statistic is maximised. Provided that the test statistic associated with the best candidate change-point is above a certain threshold, we admit it as an estimated change-point. We then split the data from this point and repeat the process within each segment until no more change-point can be detected (see Algorithm 2 for a pseudocode). By choosing M, the number of random intervals, sufficiently large, it is guaranteed that with a high probability, there exists an interval capturing each of the true change-points well inside its interior. Theoretically, a choice of $M \asymp 1/\tau^2$ is sufficient, where $\tau := \min_{i\in\{0,\ldots,\nu\}} n^{-1}(z_{i+1} - z_i)$

(cf. proof of Theorem 3.5).

Algorithm 2: Pseudocode for wild binary segmentation

```
Input: data sequence $X = (X_t)_{t\in[n]}$, number of intervals $M \in \mathbb{N}$, a single
       change-point test statistic $\psi : \mathbb{R}^* \to \mathbb{R}$ with a threshold $\lambda \in \mathbb{R}$, and a
       single change-point estimator $\eta : \mathbb{R}^* \to \mathbb{N}$
Set $\hat{Z} \leftarrow \emptyset$
Draw $M$ pairs of integers $(s_1, e_1), \ldots, (s_M, e_M)$ uniformly at random from the set
  $\{(\ell, r) \in \mathbb{N}^2 : 0 \leqslant \ell < r \leqslant n\}$.
Function WBS($s$, $e$):
    Set $\mathcal{M}_{s,e} = \{m \in [M] : s \leqslant s_m < e_m \leqslant e\}$
    Compute $R_m \leftarrow \psi((X_t)_{t\in(s_m,e_m]})$ for each $m \in \mathcal{M}_{s,e}$
    if $\max_{m\in\mathcal{M}_{s,e}} R_m > \lambda$ then
        $\hat{m} \leftarrow \arg\max_{m\in\mathcal{M}_{s,e}} R_m$
        $\hat{z} \leftarrow \eta((X_t)_{t\in(s_{\hat{m}},e_{\hat{m}}]}) + s_{\hat{m}}$
        $\hat{Z} \leftarrow \hat{Z} \cup \{\hat{z}\}$
        Run recursively WBS($s$, $\hat{z}$) and WBS($\hat{z}$, $e$).
Run WBS($0$, $n$)
Output: $\hat{Z}$
```

However, we remark that when applied with a CUSUM-based estimator, WBS is choosing the interval with the maximum CUSUM statistic, and is not guaranteed that the best candidate change-point belongs to an interval with a single true change-point. If the interval contains more than one true change-points, it is still possible that the estimated change-point may be away from any of the true change-points therein. To guarantee that the chosen interval contains exactly one change-point, Baranowski et al. (2019) further extended the idea of WBS and proposed the Narrowest-Over-Threshold (NOT) algorithm. NOT also starts by drawing a large number of intervals and finding the point with maximum CUSUM statistic within each interval. Unlike WBS, when combined with the CUSUM-based estimator, it then searches for all the intervals which have maximum absolute CUSUM statistics above a certain level and then picks up the narrowest one to estimate the change-point. As NOT focuses on the narrowest intervals in each step, it is guaranteed with a high probability that the interval we finally choose for estimation at each recursion contains exactly one change-point (cf. proof of 3.5, proof of Baranowski et al. (2019, Theorem 1)) . The generic form of NOT is stated in Algorithm 3.

Algorithm 3: Pseudocode for narrowest-over-threshold algorithm

```
Input: data sequence $X = (X_t)_{t \in [n]}$, number of intervals $M \in \mathbb{N}$, a single
       change-point test statistic $\psi : \mathbb{R}^* \to \mathbb{R}$ with a threshold $\lambda \in \mathbb{R}$, and a
       single change-point estimator $\eta : \mathbb{R}^* \to \mathbb{N}$
Set $\hat{Z} \leftarrow \emptyset$
Draw $M$ pairs of integers $(s_1, e_1), \ldots, (s_M, e_M)$ uniformly at random from the set
  $\{(\ell, r) \in \mathbb{N}^2 : 0 \leqslant \ell < r \leqslant n\}$.
Function NOT($s$, $e$):
    Set $\mathcal{M}_{s,e} = \{m \in [M] : s \leqslant s_m < e_m \leqslant e\}$
    Compute $R_m \leftarrow \psi((X_t)_{t \in (s_m, e_m]})$ for each $m \in \mathcal{M}_{s,e}$
    Set $\mathcal{R}_{s,e} := \{m \in \mathcal{M}_{s,e} : R_m > \lambda\}$
    if $\mathcal{R}_{s,e} \neq \emptyset$ then
        Find $\hat{m} \in \arg\min_{m \in \mathcal{R}_{s,e}} |e_m - s_m|$
        $\hat{z} \leftarrow \eta((X_t)_{t \in (s_{\hat{m}}, e_{\hat{m}})}) + s_{\hat{m}}$
        $\hat{Z} \leftarrow \hat{Z} \cup \{\hat{z}\}$
        Run recursively NOT($s, \hat{z}$) and NOT($\hat{z}, e$)
Run NOT(0, $n$)
Output: $\hat{Z}$
```

To better understand the difference between the three top-down approaches mentioned above. We now use the following simulation to compare BS, WBS and NOT. We randomly generate a series from a normal distribution with two changes in mean at 100 and 200 respectively. Figure 2.2 gives the interval picked up by three algorithms in the first step. Both BS and WBS fail to pick out the true change-point. However, as NOT is picking the narrowest interval, it avoids the influence from another change-point and estimates the change-point correctly.

2.1.2 Piecewise model fitting with complexity penalties

Apart from CUSUM-based approaches, a separate line of works estimate the change-points by fitting piecewise constant models to the data sequence with appropriate penalties for model complexity. One common way to do this is by minimising the following cost function:

$$\sum_{i=1}^{\nu} \mathcal{C}(X_{z_{i-1}+1:z_i}) + \beta f(\nu), \tag{2.3}$$

where $\mathcal{C}$ is a cost function and $\beta f(\nu)$ is the penalty term to prevent overfitting. One of the commonly used cost functions is the twice of negative log-likelihood (Horváth, 1993);(Chen and Gupta, 2000). For the penalty function, linear function in terms of ν is commonly used, for example, AIC (Akaike, 1974) and BIC (Schwarz, 1978). The search method Optimal partitioning (OP) proposed by Yao (1984) and Jackson et al (2005) was aimed to solve the minimisation problem above with $f(\nu) = \nu$. Firstly, if we denote the minimisation from 2.3 as $F(s)$, then it can be shown that:

$$F(s) = \min_{s'}\{F(s') + \mathcal{C}(X_{(s'+1):s}) + \beta\}, \quad 0 \leqslant s' < s.$$

The OP algorithm iterates from $s = 1$ to n. In each step, it calculates the minimum of $F(s')$ for each $s' \in [0, s)$ and then find out $s^* = \arg\min_{0 \leqslant s' < s}\{F(s') + \mathcal{C}(X_{(s'+1):n}) + \beta\}$ which is the estimated change-point for data $X_{1:s}$. Then, we record it into the estimated change-point set $\mathrm{cp}(s) = (\mathrm{cp}(s^*), s^*)$ and $\mathrm{cp}(n)$ is the set of estimated change-points for the whole data. Although it has improved the computational efficiency to $O(n^2)$ compared to previous work such as SN method (Auger and Lawrence, 1989), it is not competitive to BS procedure which has a computational cost of $O(n \log n)$.

In order to improve the computational cost of the OP algorithm, Killick, Fearnhead and Eckley (2012) proposed a modified algorithm which adds a pruning step (PELT method). To be more precise, it adds an additional step after updating the estimated change-point set $\mathrm{cp}(s)$ each iteration to remove the time points s that can never be a minimal based on the current minimisation. The cost in this case can be reduced to $O(n)$. In the worst case, the cost is $O(n^2)$ when no pruning is performed.

Another method we introduce here is through multiscale testing which can also be seen as a penalisation method. Frick, Munk and Sieling (2014) proposed simultaneous multiscale change-point estimator (SMUCE) for change-point estimation problems in exponential family. Here, we denote $F_i = F_{\vartheta(i/n)}$ and $F_\theta, \theta \in \Theta$ comes from an exponential family with densities f_θ with $\vartheta : [0,1) \to \Theta$ being a right continuous step function with unknown number ν of change-points. Let $\mathcal{S}$ be the space of all right continuous step functions with an arbitrary but finite number of jumps on the interval $[0,1)$ taking values in Θ. The idea is to estimate an unknown step function by minimising the number of change-points subject to a certain multiscale statistic below a chosen threshold. Let $\mathcal{J}(\vartheta)$ be the set of change-points. Initially, we want to solve the following optimisation problem:

$$\inf_{\vartheta \in \mathcal{S}} |\mathcal{J}(\vartheta)|, \text{ subject to } T_n(X, \vartheta) \leqslant q, \tag{2.4}$$

where

$$T_n(X, \vartheta) = \max_{\substack{1 \leqslant i < j \leqslant n \\ \vartheta(t)=\theta \text{ for } t \in [i/n, j/n]}} \left\{ \sqrt{2T_i^j(X, \theta)} - \sqrt{2 \log\left(\frac{en}{j-i+1}\right)} \right\} \tag{2.5}$$

T_i^j here is the local likelihood ratio test for testing $H_0 : \theta = \theta^*$ against $H_1 : \theta \neq \theta^*$. Let $\hat{\nu}(q)$ be the estimated number of change-points. We denote the solution to (2.4) as:

$$\mathcal{C}(q) = \{\vartheta \in \mathcal{S} : |\mathcal{J}(\vartheta)| = \hat{\nu}(q) \text{ and } T_n(X, \vartheta) \leqslant q\} \tag{2.6}$$

This is the constructed confidence band for the true ϑ that we want to estimate. The estimator $\vartheta(q)$ is then the constrained maximum likelihood estimator within the confidence set $\mathcal{C}(q)$ defined above:

$$\hat{\vartheta}(q) = \arg\max_{\vartheta \in \mathcal{C}(q)} \sum_{i=1}^{n} \log(f_{\vartheta(i/n)}(X_i)). \tag{2.7}$$

SMUCE gives a good estimate of the number of change-points by balancing the probabilities of both overestimating and underestimating $|\mathcal{J}(\vartheta)|$. It follows from equation (2.4) that $\mathbb{P}(\hat{\nu}(q) > \nu) \leqslant \mathbb{P}(T_n(X, \vartheta) > q)$. Therefore, we can control the probability of overestimating the number of change-points at level α by choosing q as the $(1-\alpha)$-quantile

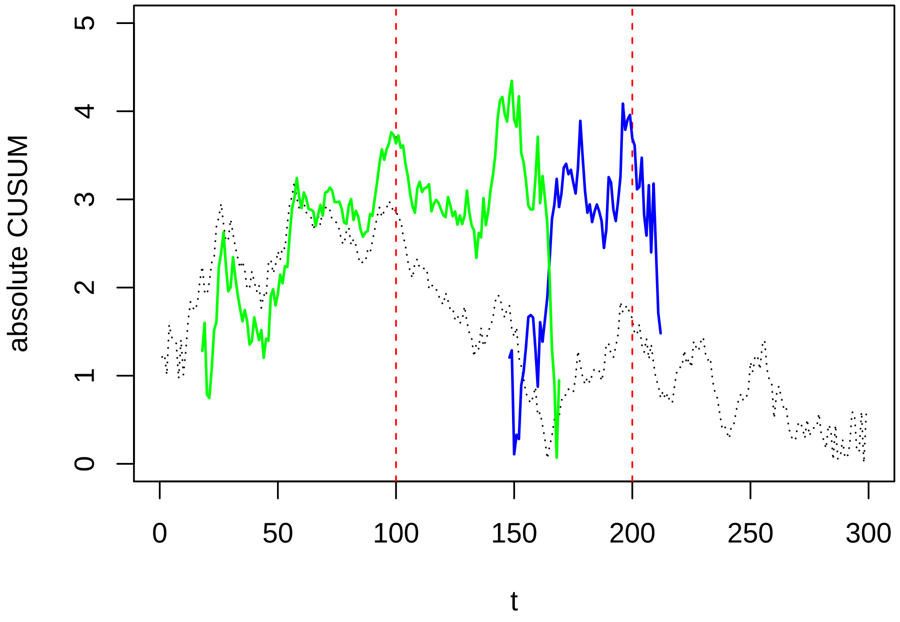

Figure 2.2: Comparisons of BS, WBS, NOT methods on a synthetic univariate data set with 2 change-points. The dotted line represents the absolute CUSUM statistics computed on the entire data series used for binary segmentation. The red curve represents the CUSUM computed in the window with the largest test statistic in wild binary segmentation. The locations of the peak of each of the curves are used as the first estimated change point in each of the procedures respectively. The true change points at 100 and 200 are shown with vertical dashed lines.

of the null distribution of $T_n(X, \vartheta)$. Also, Frick, Munk and Sieling (2014) proved an exponential bound for the probability of underestimating. Combining these two results, the probability that $\hat{\nu} \neq \nu$ tends to 0 for a suitable choice of q.

2.2 High-dimensional change-point estimation

Here, we consider the following setup: $X_1, \dots X_n \sim N(\mu, \sigma^2 I_p)$ and the sequence of mean vectors $(\mu_t)_{t=1}^n$ undergoes changes at times $z_i \in \{1, \dots, n-1\}$ for $i \in \{1, \dots, \nu\}$, in the sense that

$$\mu_{z_i+1} = \cdots = \mu_{z_{i+1}} =: \mu^{(i)}, \quad \forall\, i \in \{0, \dots, \nu\}, \tag{2.8}$$

where we use the convention that $z_0 = 0$ and $z_{\nu+1} = n$.

2.2.1 Methods based on columnwise aggregation of CUSUM matrices

The univariate change-point estimation methods we introduced so far can not be applied directly to high-dimensional data. However, the idea can be adapted to high-dimensional problems. One of the most frequently used quantities in high-dimensional change-point analysis is the CUSUM statistic. We define the CUSUM transformation $\mathcal{T} : \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times(n-1)}$ for a matrix M for $t \in [n]$ as:

$$\mathcal{T}(M)_{j,t} = \sqrt{\frac{t(n-t)}{n}} \left(\frac{1}{n-t} \sum_{r=t+1}^{n} M_{j,r} - \sum_{r=1}^{t} \frac{1}{t} M_{j,r} \right). \tag{2.9}$$

The methods we are going to review are mostly CUSUM-based algorithms. We will focus our discussion on how they estimate a single change-point. For notational simplicity, we write $z = z_1$. Similar to the univariate case, top-down methods, can also be combined to estimate multiple change-points in high-dimensional data.

Horváth and Hušková (2012) proposed the following test statistic based on the ℓ_2 aggregation of CUSUM statistics to test whether there is a change in mean:

$$\Psi_{\ell_2} = \max_{1\leqslant t\leqslant n-1} \sum_{i=1}^{p} (\mathcal{T}(X)_{j,t}^2 - 1) \tag{2.10}$$

under the condition that the coordinates are independent and $p/n^2 \to 0$, which allows the number of coordinates to be larger than the sample size. Jirak (2015) proposed another test statistic based on the ℓ_∞ aggregation:

$$\Psi_{\ell_\infty} = \max_{1 \leqslant t \leqslant n-1} \max_{1 \leqslant j \leqslant p} |\mathcal{T}(X)_{j,t}|. \tag{2.11}$$

This method also allows for large p and small n, where $n/p \to 0$. Also, the construction of the test is non-parametric and can be generalised to many popular models such as ARMA and GARCH.

However, this kind of aggregation may fail in cases such as when the signals are sparse and spread out the coordinates. In order to reduce the impact from the series which do not contain a change, Cho and Fryzlewicz (2015) proposed sparsified binary segmentation which only aggregates the coordinates with the CUSUM values above a certain threshold π_n under the sparsity assumption. Specifically, it uses the following statistic:

$$\Psi_{\text{SBS}} = \max_{1 \leqslant t \leqslant n-1} \sum_{j=1}^{p} |\mathcal{T}(X)_{j,t}| \mathbb{1}_{|\mathcal{T}(X)_{j,t}| > \pi_n} \tag{2.12}$$

Furthermore, Cho (2016) defines the following double CUSUM (DC) statistic to locate the change-point. After performing a first CUSUM transform on the entire series in terms of each time t, it performs a second CUSUM transformation on the coordinates:

$$\begin{aligned}
&\mathcal{D}_s^\phi(\{|\mathcal{T}(X)_{(j),t}|\}_{j=1}^p) \\
&= \left\{ \frac{s(2p-s)}{2p} \right\}^\phi \left(\frac{1}{s} \sum_{j=1}^{s} |\mathcal{T}(X)_{(j),t}| - \frac{1}{2p-s} \sum_{j=s+1}^{p} |\mathcal{T}(X)_{(j),t}| \right) \\
&= \left\{ \frac{s(2p-s)}{2p} \right\}^\phi \times \frac{1}{s} \sum_{j=1}^{s} \left(|\mathcal{T}(X)_{(j),t}| - \frac{1}{2p-s} \sum_{j=s+1}^{p} |\mathcal{T}(X)_{(j),t}| \right),
\end{aligned}$$

where $j \in \{1, \ldots, p\}$, $\phi \in [0, 1]$ and $|\mathcal{T}(X)_{(j),t}|$ is the ordered CUSUM statistic values such that $|\mathcal{T}(X)_{(1),t}| \geqslant |\mathcal{T}(X)_{(2),t}| \geqslant \ldots \geqslant |\mathcal{T}(X)_{(p),t}|$ at each t. Then, the test statistic is defined as:

$$\Psi_{\text{DC}}^\phi = \max_{1 \leqslant t \leqslant n-1} \max_{1 \leqslant s \leqslant p} \mathcal{D}_s^\phi(\{|\mathcal{T}(X)_{(j),t}|\}_{j=1}^p). \tag{2.13}$$

If Ψ^{ϕ}_{DC} is above a certain test criterion, the following time point $\hat{z}$ is identified as a change-point:

$$\hat{z} = \mathop{\arg\max}_{1\leqslant t\leqslant n-1} \max_{1\leqslant s\leqslant p} \mathcal{D}^{\phi}_{s}(\{|\mathcal{T}(X)_{(j),t}|\}_{j=1}^{p}).$$

The innovation here is that in the second aggregation $\mathcal{D}^{\phi}_{s}$, the input is an ordering series. This enables us to not only locate the time point with change but also the coordinates with change. We can achieve this by finding

$$s^{\phi}_{\hat{z}} = \mathop{\arg\max}_{1\leqslant s\leqslant p} \mathcal{D}^{\phi}_{s}(\{|\mathcal{T}(X)_{(j),\hat{z}}|\}_{j=1}^{p},$$

where $\hat{z}$ is the estimated change-point location. The estimated number of coordinates with change is given by $s^{\phi}_{\hat{z}}$ and we can find out the exact coordinates according to the sorted CUSUM statistics in the $\hat{z}$th column. We remark that the double CUSUM statistic can also be viewed as a columnwise aggregation. For each time point t, we input the sorted CUSUM statistics and perform a second CUSUM transformation after combining a zero matrix of $p \times (n-1)$ dimension and then take the maximum. We can rewrite equation (2.13) as:

$$\Psi^{\phi}_{\text{DC}} = \max_{1\leqslant t\leqslant n-1} \|\mathcal{T}(\text{sort}(|\mathcal{T}(X)_t|), 0_p)\|_{\infty}, \tag{2.14}$$

where $\mathcal{T}_t$ denotes the CUSUM transformation of the tth column and $\text{sort}(|\mathcal{T}(X)|_t) = (|\mathcal{T}(X)_{(1),t}|, |\mathcal{T}(X)_{(2),t}|, \ldots, |\mathcal{T}(X)_{(p),t}|)^{\top}$.

Enikeeva and Harchaoui (2019) proposed linear and scan statistics to detect high dimensional change-points under both sparse and dense regimes. The linear statistic is identical to Ψ_{ℓ_2} in equation (2.10) and the scan statistic is defined as:

$$\Psi_{\text{scan}} = \max_{1\leqslant t\leqslant n-1} \max_{1\leqslant s\leqslant p} \frac{\sum_{j=1}^{s} \mathcal{T}^{2}_{(j),t} - s}{\sqrt{2s}}, \tag{2.15}$$

where $|\mathcal{T}_{(1),t}| \geqslant |\mathcal{T}_{(2),t}| \geqslant \cdots \geqslant |\mathcal{T}_{(s),t}|$. The decision rule is based on the combination of two test statistics: $\Psi^{*} = \mathbb{1}\{\Psi_{\ell_2} > H\} \bigvee \mathbb{1}\{\Psi_{\text{scan}} > T\}$, where H and T are pre-specified thresholds chosen according to the significance level. We reject the null hypothesis if $\Psi^{*} = 1$. The performance of two tests depend on the sparsity level. Following Enikeeva and Harchaoui (2019, Theorems 1 and 2), in the dense case when the sparsity level is low (s is large), the linear statistic is more effective while in the sparse case (s is small), the scan statistic works better. In addition, we know that the scan statistic can detect a change with a smaller magnitude comparing to linear statistic. The boundary between two sparsity regimes is $s \asymp p^{1/2}$.

2.2.2 The Inspect algorithm

So far, the test statistics we reviewed above are based on columnwise aggregation. We now introduce a projection based estimation procedure: Informative Sparse Projection for Estimation of Change-points algorithm (Inspect) by Wang and Samworth (2018) which is based on the aggregation across columns and rows. The idea is to seek a projection direction v so that the signal-to-noise ratio is maximised. In other words, we would like to maximise $v^\top \theta/\sigma$, where $\theta = \mu^{(1)} - \mu^{(0)}$ is the vector of mean change. Since σ is treated as fixed, the optimal projection direction is $v = \theta/\|\theta\|_2$. By linearity of the CUSUM transformation defined in (2.9), we observe that $\mathcal{T}(X) = \mathcal{T}(\mu) + \mathcal{T}(W)$ and that $\mathcal{T}(\mu)$ has rank 1 with leading left singular vector parallel to θ. Viewing $\mathcal{T}(X)$ as a perturbation of $\mathcal{T}(\mu)$, Inspect estimates v by approximating the sparse leading left singular vector of $\mathcal{T}(X)$ through a convex relaxation scheme. Once the projection direction $\hat{v}$ was obtained, Inspect estimates the location of the change-point by the peak of the univariate projected CUSUM $\hat{v}^\top \mathcal{T}(X)$. The full algorithm for estimating a single change-point is summarised in Algorithm 4.

Algorithm 4: Single change-point estimation using the Inpsect algorithm

```
Input: $X \in \mathbb{R}^{p \times n}$, $(\mathcal{J}_g)_{g \in [G]}$, and $\lambda > 0$
Compute $T \leftarrow \mathcal{T}(X)$ as in (2.9).
Set
    $$\hat{M} = \frac{\mathbf{soft}(T, \lambda)}{\|\mathbf{soft}(T, \lambda)\|_2},$$
  where $\mathbf{soft}(T, \lambda) = \operatorname{sgn} \max\{|T_{i,j}| - \lambda, 0\}$ and $\lambda \geqslant 0$.
Let $\hat{v}$ be the leading left singular vector of $\hat{M}$
Estimate $z$ by $\hat{z} = \arg\max_{1 \leqslant t \leqslant n-1} |\hat{v}^\top T_t|$, where $T_t$ is the $t$th column of $T$.
Output: $\hat{z}$, $\bar{T}_{\max} = |\hat{v}^\top T_{\hat{z}}|$
```

2.2.3 Simulations for comparison of high-dimensional change-point estimation methods

In this subsection, we compare the methods surveyed in the last section. Specifically, ℓ_2 and ℓ_∞ aggregation, SBS, scan statistic from Enikeeva and Harchaoui (2019), Double CUSUM and `Inspect`. Here, we randomly generated a data set with $n = 500$, $p = 1000$ and the change is located at $z = 400$. We compare two settings according to the sparsity: sparse regime, $s = 2$ and dense regime $s = 100$. We take $\|\theta\|_2 = 0.8$ in sparse regime and $\|\theta\|_2 = 1$ in dense regime. For the threshold in SBS, we randomly generate 1000 of $N(0, 1)$ matrices and take the maximum for each of them. The threshold is then the 95th quantile of the maximum.

From Figure 2.3 and Figure 2.4, we see that ℓ_2 aggregation and Doule CUSUM work well for dense signal but not sparse signal. For ℓ_2 aggregation, this is due to the fact that for sparse signals, we add up, for each time point, squares of the CUSUM statistics at both the signal and noise coordinates, which means that the sparse signal could be diluted by the large noise variance when the dimension is high. For Double CUSUM, the problem lies in the fact that for sparse signal, the second CUSUM transformation along the spatial direction has a change-point at s at the population level. The fact that the second CUSUM has a change near 0 means that the CUSUM magnitude is relatively small and easily affected by noise, resulting in inaccurate estimates. On the other hand, both the ℓ_∞ aggregation and SBS work well for sparse signals but behaves poorly in dense signals. For ℓ_∞ aggregation, this is due to the fact that ℓ_∞ aggregation only looks at the maximum CUSUM statistics in each column and does not accumulate evidence across multiple signal coordinates. For SBS, this is likely a result of the hard thresholding operation being too aggressive when the signal is spread across multiple coordinates. Finally, both the scan statistic based approach and the `Inspect` algorithm appear to be robust to the sparsity level. This is because both methods can be viewed as approximating an ℓ_2 aggregation over only the signal coordinates and so should behave like an oracle estimator, where we have removed all the noise coordinates from the data. The `Inspect` algorithm achieves this by first estimating a projection direction $\hat{v}$ that is close to $\theta/\|\theta\|$, which has support

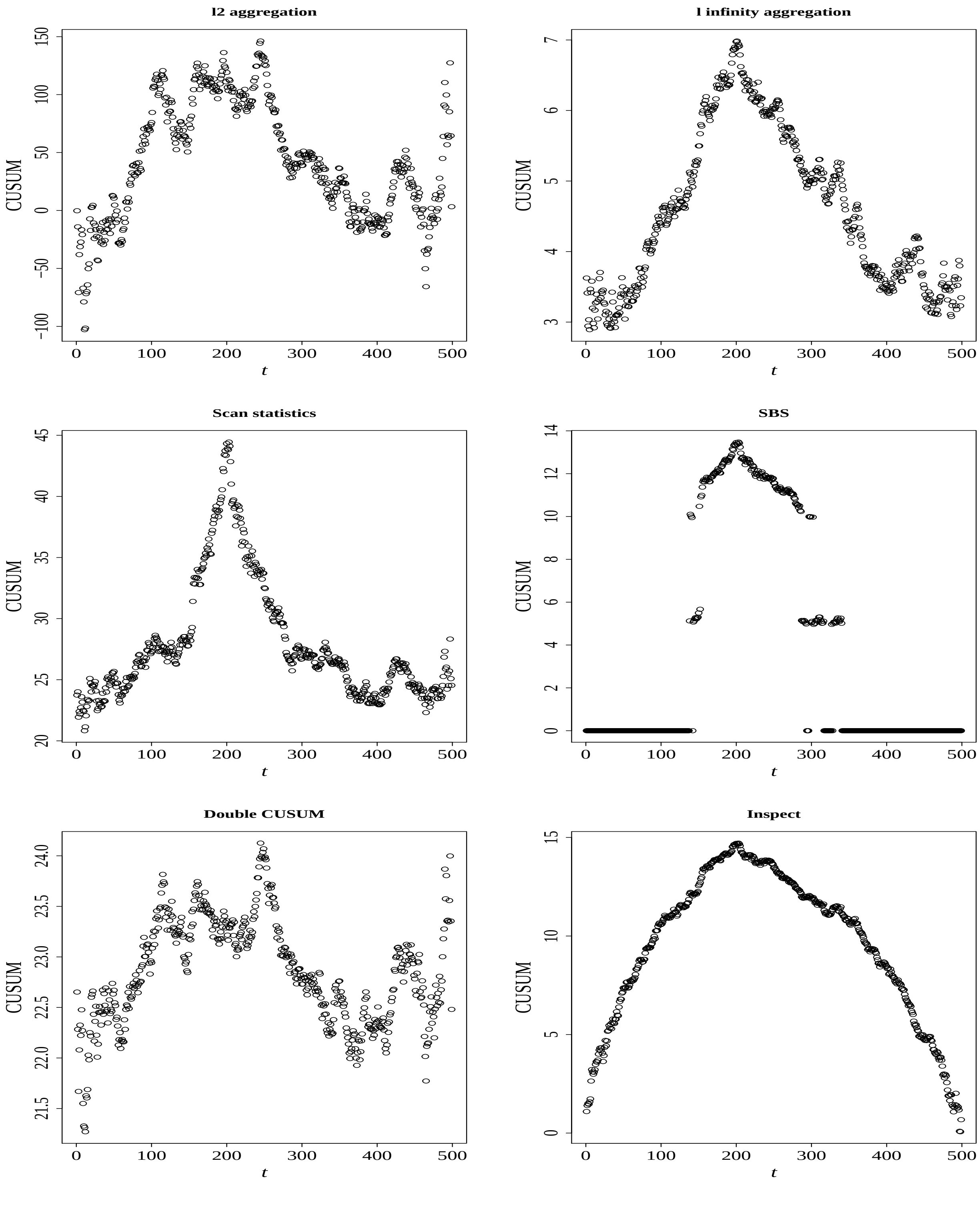

Figure 2.3: Comparisons of aggregated CUSUM statistics of ℓ_2, ℓ_∞, scan statistics, SBS, double CUSUM and `Inspect` under sparse case when $s = 2$. Other parameters used: $p = 500$, $n = 1000$, $\|\theta\|_2 = 0.8$

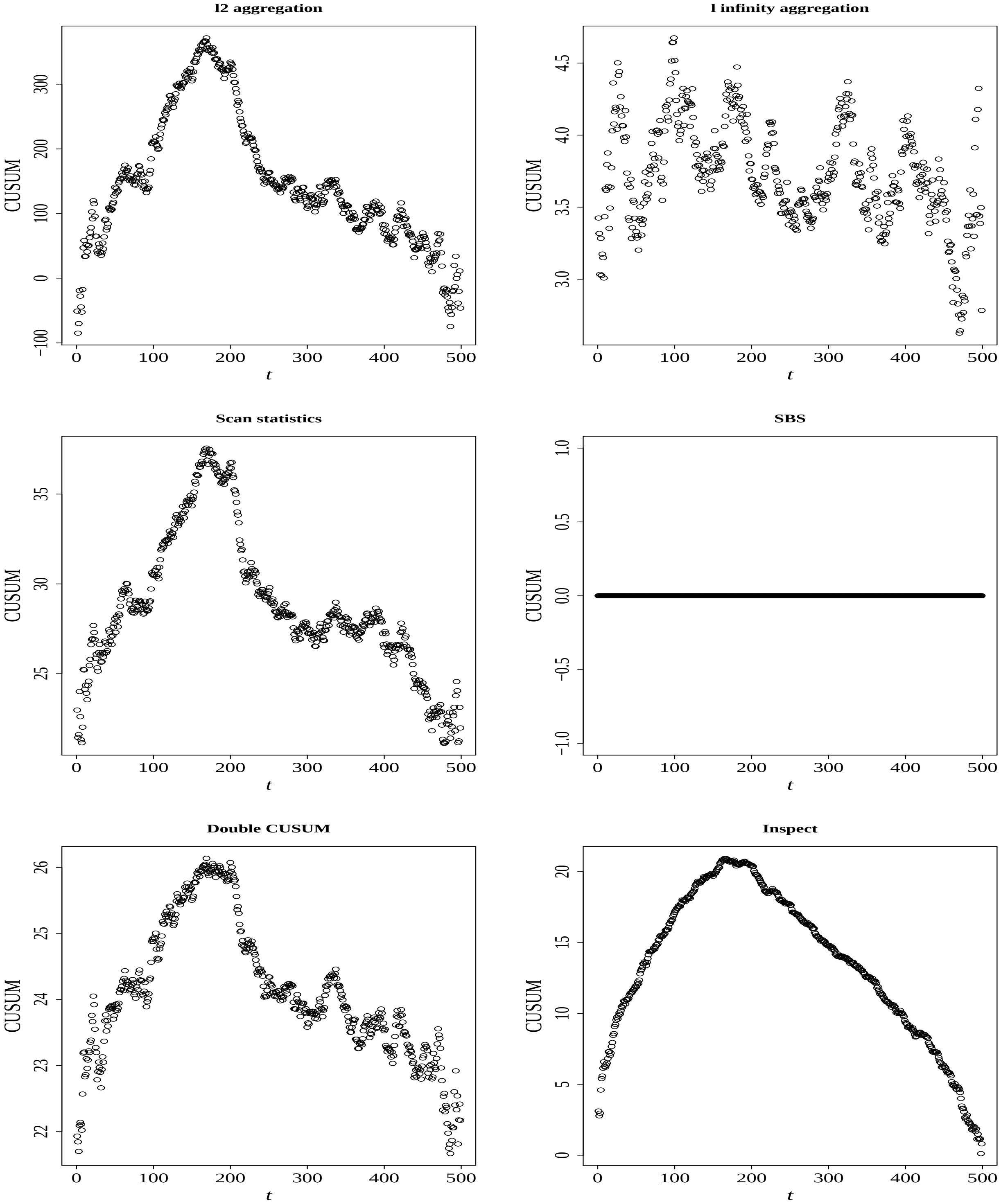

Figure 2.4: Comparisons of aggregated CUSUM statistics of ℓ_2, ℓ_∞, scan statistics, SBS, double CUSUM and `Inspect` under dense case when $s = 100$. Other parameters used: $p = 500$, $n = 1000$, $\|\theta\|_2 = 1$

in the signal coordinates. The scan statistics approach achieves this if the maximum over t in (2.15) is equal to the sparsity level s.

2.3 Online change-point estimation problem

While the main focus of this thesis is on offline change-point estimation with structural constraints, we include this section for review of the online change-point detection problem for completeness.

2.3.1 Classic methods

In this subsection, we review the univariate online change-point estimation problem. Unlike offline change-point problem which obtains the entire dataset over all the time points, online change-point analysis works on real-time series and aims to detect the change as soon as it occurs. To be more specific, we have that $X_1 \dots X_z$ follow distribution F_0 with density f_0 and $X_{z+1}, X_{z+2}, \dots$ follow distribution F_1 with density f_1. We want to find z as soon as it occurs. Online change-point detection can date back to the last century and is often used in quality control in manufacturing. Shewhart (1931) introduced control charts, which is to plot the statistics as well as control limits on the chart, and actions are taken once the points fall outside the control limits. We first introduce some common control charts. Shewhart (1931) used $\bar{X}$ chart which uses $\mu_0 \pm C\sigma_{\bar{X}}$ as the control limit, where μ_0 is the mean of $\bar{X}$ and $\sigma_{\bar{X}}$ is the standard deviation of $\bar{X}$. In many applications, taking $C = 3$ is enough to detect the change without taking actions unnecessarily. However, the detection is slow if the shifts in the mean are small. Therefore, various tests have been used to supplement the Shewhart test so as to increase sensitivity to small changes, for example, taking action if there are a few consecutive points on one side of the central line or using multiple supplementary tests to speed the detection (Roberts, 1966).

The moving average chart plots the following term at time $i \geqslant k$ for a moving average of index k:

$$\bar{X}_i^{(k)} = \frac{\sum_{j=0}^{k-1} \bar{X}_{i-j}}{k},$$

with the limits $\mu_0 \pm C\sigma_{\bar{X}}/\sqrt{k}$. At time i, we plot $\bar{X}_i^{(i)} = \sum_{j=1}^{i} \bar{X}_j/i$ and use $\mu_0 \pm C\sigma_{\bar{X}}/\sqrt{i}$ as limits. Based on the moving average chart, the geometric moving average chart considers plotting the following term:

$$Z_i = (1-r)Z_{i-1} + r\bar{X}_i = (1-r)^i Z_0 + r\sum_{j=0}^{i-1}(1-r)^j \bar{X}_{i-j},$$

with $Z_0 = \mu_0$, $r \in (0, 1]$. The limits in this control chart is $\mu_0 \pm C_r\sigma_{\bar{X}}\sqrt{r/(2-r)}$. Z_i here is a combination of all the points before i, and the closer the point is to i, the higher the weight.

Page (1954) proposed the cumulative sum chart. Let $S_i = \sum_{j=1}^{i} x_j$, where x_j is a score, for example $\bar{X}_j - \mu_0$. The action is taken once $S_n - \min_{0 \leqslant i < n} S_i \geqslant c$. Let $S'_n = \max(S'_{n-1} + x_n, 0)$ with $S'_0 = 0$, Page (1954) also showed that the criteria above is equivalent to $S'_n \geqslant c$. The likelihood ratio score $x_i = \log\{f_1(X_i)/f_0(X_i)\}$ is often used in change-point problem.

The main criteria for online change-point detection are detection delay and false alarm. We denote N as the stopping time which is the estimated time that a change occurs. The false alarm is usually measured by $1/E_\infty(N)$, where $E_\infty(N)$ is the average run length until the false alarm. Page's procedure using the likelihood ratio score is to seek the following stopping time:

$$N = \inf\left\{n : \max_{1 \leqslant k \leqslant n} \sum_{i=k}^{n} \log\left(\frac{f_1(X_i)}{f_0(X_i)}\right) \geqslant c\right\}. \tag{2.16}$$

Lorden (1971) proposed the following minimax type criteria, which corresponds to the worst case detection delay:

$$\bar{E}_1(N) = \sup_{z \geqslant 0} \operatorname{ess\,sup} \mathbb{E}_z[(N-z)^+ | X_1, \ldots, X_z] \tag{2.17}$$

The intuition behind is that, if z is the true change-point, the conditional expectation of $N - z$ given $X_1, \ldots, X_z$ when $N \geqslant z$ should be small. The expectation is maximised over all pre-change observations and all possible change-point locations. Lorden (1971) proved that Page's procedure is asymptotically minimax optimal as $\gamma \to \infty$. Furthermore, Moustakides (1986) proved that Page's procedure is optimal in terms of minimising this worst case detection delay subject to $E_\infty(N) \geqslant \gamma$, where γ is predetermined.

Under the Bayesian framework, the location of change-point is random with some prior distributions. In the special case of geometric prior distribution, the change-point z with value n has the probability: $\mathbb{P}(z = n) = p(1 - p)^{n-1}$, for $n = 1, 2, \ldots$. Shiryaev (1963) proposed the following stopping time :

$$N_p(\gamma) = \inf\left\{n \geqslant 1 : \sum_{k=1}^{n}\prod_{i=k}^{n}\frac{f_1(X_i)}{(1-p)f_0(X_i)} \geqslant \gamma\right\}. \tag{2.18}$$

Roberts (1966) modified the rule as:

$$N(\gamma) = \inf\{n \geqslant 1 : R_n \geqslant \gamma\},$$

where $R_n = \sum_{k=1}^{n}\prod_{i=k}^{n}\frac{f_1(X_i)}{f_0(X_i)}$.

In addition to the worst case detection delay, Pollak (1985) proposed the following supremum conditional average delay:

$$\bar{E}_1(N) = \sup_{z \geqslant 0}\mathbb{E}_z(N - z|N \geqslant z).$$

The stopping rule from Roberts (1966) and Shiryaev (1963) have shown to be asymptotically optimal in the sense of controlling the conditional average delay stated above. Pollak and Tartakovsky (2009) further proved that they are exactly optimal in terms of minimising the following integral average delay subject to $E_\infty(N) \geqslant \gamma$:

$$\bar{E}_1(N) = \sum_{z=1}^{\infty}\mathbb{E}_z(N - z)^+.$$

2.3.2 Multivariate and high-dimensional setting

High-dimensional online change-point detection problem has also been studied where we have a sequence of p dimensional vectors: $X_{1,i}, \ldots, X_{p,i}$ for $i \in [n]$. For example, Mei (2010) proposed a method based on CUSUM statistic by Page (1954). We now briefly go through the idea here. We first define the following local CUSUM statistic of the kth coordinate.

$$W_n^j = \max\left\{0, W_{n-1} + \log\frac{f_1^j(X_{j,n})}{f_0^j(X_{j,n})}\right\},$$

where f_0^j and f_1^j are densities for the jth coordinate before and after change and $W_0^j = 0$. The proposed stopping time is:

$$N = \inf\{n \geqslant 1 : \sum_{j=1}^{p} W_n^j \geqslant c\}$$

The statistic here is a summation over all the coordinates. Here, we assume that we do not know the number of coordinates with a change. Therefore, it is quite natural to use the sum of the CUSUM statistics over all the coordinates and a large value indicates a possible change.

We list here some other existing methodologies for multivariate and high dimensional online change-point estimation: Xie and Siegmund (2013),Chan (2017),Tartakovsky et al. (2006). In addition, there are also some methods proposed for the structural change in covariances, for example: Li and Li (2023).

2.4 Related problems where structural assumptions are used in inference

This thesis primarily concerns with how structural assumptions can be exploited to improve statistical inference in high-dimensional change-point settings. In this section, we survey some existing results in the literature, where similar structural assumptions have been successfully used in inferential tasks.

In high-dimensional problems, one typically assumes sparsity on problem parameters to effectively reduce the complexity of the problem. However, high-dimensional data may possess other structures in addition to sparsity. For example, the coordinates may be naturally clustered into groups or connected in a network. We now review how such additional structures can be exploited in the context of high-dimensional regression. Consider the following setup:

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_p),$$

where X is a $p \times n$ data matrix, Y and β are vectors of length n and p respectively. The well-known lasso (Tibshirani, 1996) regression minimises $\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$, which induces sparsity on the coefficient vector by ℓ_1 penalty to avoid overfitting. Yuan and Lin (2006) generalises lasso to the group lasso, which assumes that the parameters can be divided into groups and the penalty is summed by group index so that it can drive the entire group of parameters to 0 if this group is not significant. To be more specific, for a vector $\eta \in \mathbb{R}^d$ with $d \geqslant 1$, and a symmetric positive definite matrix $K \in \mathbb{R}^{d\times d}$, we first define the following quantity:

$$\|\eta\|_K = (\eta^\top K \eta)^{1/2}.$$

Then, given a sequence of positive definite matrices $K_1, \dots K_G$, where $K_g \in \mathbb{R}^{p_g \times p_g}$, p_g is the group size for group g, the group lasso is to minimise the following quantity:

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_{K_g}, \tag{2.19}$$

where $\lambda \geqslant 0$ is the tuning parameter and β_g is the vector of coefficients belonging to group g. One reasonable choice of K_j is $p_g I_{p_g}$, which was also used by Yuan and Lin (2006) in their implementation. The second term in equation (2.19) is the group lasso penalty, it calculates the sum of ℓ_2 norms of parameters within each group and puts the weights of the square root of group size for each group. There are two special cases for the group lasso: if all groups are of size 1, it then reduces to the original lasso regression. On the other hand, if there is only one group, it becomes ridge regression. Based on the group lasso, Simon et al. (2013) proposed sparse group lasso which not only considers the sparsity between but also the sparsity within the group. It adds an additional ℓ_1 penalty as follow:

$$\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_{K_g} + (1-\lambda)\sum_{j=1}^{p} |\beta_j|, \tag{2.20}$$

where $\lambda \in [0, 1]$. It forms a convex combination between group lasso and lasso.

In addition to the group structure, Tibshirani et al. (2005) proposed fused lasso to exploit the spatial dependence between covariates. The fused lasso minimises the following objective:

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|. \tag{2.21}$$

The first penalty term is to induce the sparsity on the covariates, and the second one is to penalise the difference between two adjacent covariates. Motivated by the gene clustering problem, She (2010) proposed clustered lasso, which is a generalisation of the fused lasso, but does not require the covariates to be ordered:

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{i<j}^{p} |\beta_i - \beta_j|. \tag{2.22}$$

In regression problems mentioned above, we have assumed that parameters have additional structures. Similarly, we can also make assumptions on the coordinates in high-dimensional change-point data. For example, the coordinates are naturally grouped into clusters. In this case, we can aggregate the information from each coordinate by groups which is motivated from the idea of group lasso. We will see the details in the next Chapter. Regarding the network structure, the fused lasso sums over the differences between two consecutive parameters. In fact, we can view the parameters as the coordinates that are connected in a network. $|\beta_j - \beta_{j-1}|$ is the distance between two coordinates that are connected directly with each other, and the summation is an aggregation over the network. This motivates us to aggregate the information along a certain path with suitable time lags in terms of time series data. We will see the details in Chapter 4.

Chapter 3

High-dimensional change-point estimation under group sparsity structure

3.1 Introduction

In this chapter, we provide a new high-dimensional change-point methodology that exploits the group sparsity structure of the changes. As mentioned in Chapter 1, existing high-dimensional change-point methods often assume that the signal of change possesses some form of sparsity such as the difference in mean before and after a change-point is nonzero only in a small subset of coordinates. However, it often does not capture the full extent of the structure in the vector of change available in real data applications. For instance, in many applications, the coordinates of the high-dimensional vectors are naturally clustered into groups and coordinates within the same group tend to change together. At each change-point, only a small number of groups will undergo a change. This is what we mean by group sparsity structure. Such a group sparsity change-point structure is useful in modelling many practical applications. Examples include financial data streams where changes are often grouped by industry sectors and a small number of sectors may experience virtually simultaneous market shocks. Also, in functional magnetic resonance imaging data, voxels belonging to the same brain functional regions tend to change simultaneously over time. We have summarised some problems with similar group sparsity assumptions in Section 2.4. The algorithm we proposed here, named `groupInspect` (standing for **group**-based **in**formative **s**parse **p**rojection **e**stimator of **c**hange-poin**t**s), will use the given pre-specified grouping information of all the coordinates to first estimate a vector of projection that is closely aligned with the true vector of change at each change-point. It will then project the high-dimensional data series along this estimated direction and apply a univariate change-point method on the projected series to identify the location of the change. The above procedure can be combined with the narrowest-over-threshold algorithm of Baranowski et al. (2019) to identify multiple change-points recursively. We show that, in a single change-point setting, the projection direction estimator employed in `groupInspect` has a minimax optimal dependence, up to logarithmic factors, on both the ℓ_0 sparsity parameter and the group-sparsity parameter, representing respectively the number of nonzero elements and the number of nonzero groups in the vector of change. Furthermore, under appropriate conditions, `groupInspect` achieves a minimax optimal $\log\log(n)/(n\vartheta^2)$ rate of convergence for the estimated location of a single change-point and a $\log(n)/(n\vartheta^2)$ rate of convergence for multiple change-points, where ϑ denotes the ℓ_2 norm of the vector of change.

In Section 3.2, we describe the formal setup of our problem. The `groupInspect` methodology is then introduced in Section 3.3, with its theoretical performance guarantees provided in Section 3.4. We illustrate the empirical performance of `groupInspect` via simulations and a real-data example in Section 3.5. The extensions to sub-Gaussian and temporal dependence settings are given in Section 3.6 and Section 3.7. Proofs of all theoretical results are deferred to Section 3.8, and ancillary results and their proofs are given in Section 3.9.

3.2 Problem set up

We now formally describe the data generating mechanism, which is described in Section 3.1. Let $X_1, \ldots, X_n$ be independent random vectors with distribution:

$$X_t \sim N_p(\mu_t, \Sigma), \quad 1 \leqslant t \leqslant n, \quad \text{where } \|\Sigma\|_{\mathrm{op}} \leqslant B \tag{3.1}$$

for some $B \in (0, \infty)$. We remark that the main focus of the current chapter is to understand the effect of group sparsity structure on the change-point estimation accuracy, and as such, to simplify exposition, we have assumed here that observations are independent normal random vectors. All our theoretical results can be extended to the case where the observations are sub-Gaussian or have short-ranged temporal dependence (see Section 3.6 and 3.7 for details). We can combine into a single data matrix $X \in \mathbb{R}^{p\times n}$ and mean vectors undergo changes as described in equation (2.8). We assume that consecutive change-points are sufficiently separated in the sense that

$$\min\{z_{i+1} - z_i : 0 \leqslant i \leqslant \nu\} \geqslant n\tau.$$

Suppose further that each of the p coordinates belongs to (at least) one of the G groups. Specifically, let $\mathcal{J}_g$ denotes the set of indices associated with the gth group for $g \in \{1, \ldots, G\}$, we have that

$$\bigcup_{g=1}^{G} \mathcal{J}_g = [p]. \tag{3.2}$$

We assume that coordinates in the same group will tend to change together. We will consider both the case of overlapping and non-overlapping groups. In the latter scenario, we have $\mathcal{J}_g \cap \mathcal{J}_g' = \emptyset$ so that each coordinate belongs to a unique group and $(\mathcal{J}_g)_{g\in[G]}$ forms a partition of $[p]$.

Our goal is to estimate the locations of change $z_1, \ldots, z_\nu$ from the data matrix X and the pre-specified grouping information $(\mathcal{J}_g)_{g\in[G]}$. Motivated by Wang and Samworth (2018), when the coordinates are independent, the best way to aggregate the component series so as to maximise the signal-to-noise ratio around the ith change-point is to project the data along a direction close to the vector of change $\theta^{(i)} = \mu^{(i)} - \mu^{(i-1)}$. Let $v^{(i)}$ be the unit vector parallel to $\theta^{(i)}$:

$$v^{(i)} = \theta^{(i)}/\|\theta^{(i)}\|_2.$$

In our setting, we would like to maximise the signal-to-noise ratio: $\frac{v^\top \theta}{(v^\top \Sigma v)^{1/2}}$ and the optimiser is $\Sigma^{-1}\theta$. However, as Σ is usually difficult to estimate, we still consider estimating θ. If we use the optimiser $\Sigma^{-1}\theta$, the square of the signal-to-noise ratio is $\|\Sigma^{-1/2}\theta\|_2^2 \leqslant \|\theta\|_2^2/\lambda_{\min}(\Sigma)$. On the other hand, if we use θ instead, the square of signal-to-noise ratio becomes: $\|\theta^4\|/\|\Sigma^{1/2}\theta\|_2^2 \geqslant \|\theta\|_2^2/\lambda_{\max}(\Sigma)$, where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are maximum and minimum eigenvalues of Σ respectively. Therefore, if Σ is well-conditioned in the sense that the maximum and minimum eigenvalues of Σ are bounded away from 0, using θ instead of $\Sigma^{-1}\theta$ incurs a loss of efficiency of at most a factor of $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. If the fraction between the maximum and minimum eigenvalues is of order 1, then the signal-to-noise ratios obtained from actual optimiser Σ^{-1} and θ are of the same order. We measure the quality of any estimated projection direction $\hat{v}$ with the Davis–Kahan sin θ loss (Davis and Kahan, 1970)

$$L(\hat{v}, v^{(i)}) = \sqrt{1 - (\hat{v}^\top v^{(i)})^2}$$

and measure the quality of the subsequent location estimator $\hat{z}_i$ by $\mathbb{E}|\hat{z}_i - z_i|$.

The difficulty of the estimation task depends on both the noise level σ and the vector of change $\theta^{(i)} = \mu^{(i)} - \mu^{(i-1)}$. More precisely, we assume that the change is localised in a small number of the G groups as defined in (3.2). Define $\phi : \mathbb{R}^p \to \mathbb{R}^G$ such that $\phi(x) = (\|x_{\mathcal{J}_1}\|_2, \|x_{\mathcal{J}_2}\|_2, \ldots, \|x_{\mathcal{J}_G}\|_2)^\top$, we assume that

$$\|\phi(\theta^{(i)})\|_0 \leqslant s, \qquad \sum_{g\in[G]:\theta^{(i)}_{\mathcal{J}_g}\neq 0} |\mathcal{J}_g| \leqslant k, \quad \text{and} \quad \|\theta^{(i)}\|_2 \geqslant \vartheta. \tag{3.3}$$

3.3 Methodology

3.3.1 Single change-point estimation

Initially, we will consider the estimation of a single change-point, where $\nu = 1$. This can be extended to estimate multiple change-points in conjunction with top-down approaches such as wild binary segmentation and narrowest-over-threshold approach of Baranowski et al. (2019), which we will discuss in Section 3.3.2.

We compute the CUSUM transformation of X: $T = \mathcal{T}(X)$ as defined in equation (2.9) of Chapter 1. As discussed in Section 3.2, our general strategy is to use the matrix T to estimate a projection direction that is well-aligned with the direction of change, and then project the data along this direction to estimate the change-point location from the univariated projected series. More precisely, we would like to solve for:

$$\hat{v} \in \operatorname*{arg\,max}_{u \in \mathbb{S}^{p-1}, \|\phi(u)\|_0 \leqslant s} \|u^\top T\|_2. \tag{3.4}$$

However, the above optimisation problem is non-convex due to the group-sparsity constraint. Consequently, we perform the following convex relaxation of the above problem. We first note that the set of optimisers of (3.4) is equal to the set of leading left singular vectors of

$$\operatorname*{arg\,max}_{\substack{M \in \mathbb{R}^{p\times(n-1)}: \|M\|_* = 1, \mathrm{rank}(M)=1 \\ \sum_{g\in[G]} \mathbb{1}_{\{\|M_{\mathcal{J}_g}\|_{\mathrm{F}} \neq 0\}} \leqslant s}} \langle M, T\rangle,$$

We relax the above matrix-variate optimisation problem by dropping the combinatorial rank constraint, and replacing the nuclear norm constraint set by the larger Frobenius norm set of $\mathcal{S} = \{M \in \mathbb{R}^{p\times(n-1)} : \|M\|_F \leqslant 1\}$. The constraint that M has at most s groups of non-zero rows can be written as an ℓ_0 constraint on the vector of Frobenius norms of such submatrices, i.e. $\|(\|M_{\mathcal{J}_g}\|_F : g \in \{1, \ldots, G\})\|_0 \leqslant s$. Motivated by the group lasso penalty (Yuan and Lin, 2006), we replace this group sparsity constraint with a *group norm* penalty, where the group norm for a matrix $M \in \mathbb{R}^{p\times(n-1)}$ is defined as

$$\|M\|_{\mathrm{grp}} = \sum_{g=1}^{G} p_g^{1/2} \|M_{\mathcal{J}_g}\|_{2,1}, \tag{3.5}$$

where $\|M_{\mathcal{J}_g}\|_{2,1}$ is the sum of column ℓ_2 norms of the submatrix $M_{\mathcal{J}_g}$ and $p_g = |\mathcal{J}_g|$. Overall, we obtain the following optimisation problem:

$$\hat{M} \in \operatorname*{arg\,max}_{M\in\mathcal{S}} \bigl\{\langle T, M\rangle - \lambda\|M\|_{\mathrm{grp}}\bigr\}, \tag{3.6}$$

where $\lambda \in [0, \infty)$ is a regularisation parameter.

If the groups are non-overlapping, in the sense that $\mathcal{J}_g \cap \mathcal{J}_{g'} = \emptyset$ for all $g \neq g'$, then we see from Proposition 3.12 that (3.6) has a closed form solution

$$\hat{M} = \frac{T - R^*}{\|T - R^*\|_{\mathrm{F}}}, \tag{3.7}$$

where $R^*_{\mathcal{J}_g,t} = T_{\mathcal{J}_g,t} \min\left\{\frac{\lambda p_g^{1/2}}{\|T_{\mathcal{J}_g,t}\|_2}, 1\right\}$. Overall, equation (3.7) reveals a soft-thresholding process between each column ℓ_2 norm in each of the g th group with $\lambda p_g^{1/2}$.

For overlapping groups, (3.6) can be optimised using Frank–Wolfe algorithm (Frank and Wolfe, 1956), as described in Algorithm 5. We first compute the gradient of the objective function which is the step 4 in Algorithm 5. We then project the $\hat{M}$ back onto $\mathcal{S}$.

After solving the optimisation problem, we can obtain the estimated projection direction $\hat{v}$ by computing the leading left singular vector of $\hat{M}$. Then, we project the data along $\hat{v}$ to obtain a univariate series for which existing one-dimensional change-point estimation methods apply. Specifically, we perform the CUSUM transformation over the projected data series, and locate the change-point by the maximum absolute value of the CUSUM vector. The full procedure is described in Algorithm 6.

3.3.2 Multiple change-point estimation

When the data matrix possesses multiple change-points, we may combine Algorithm 6 with a top-down approach (Fryzlewicz, 2014; Baranowski et al., 2019, e.g), to recursively identify all the change-points. Specifically, in Algorithm 7, we adopt the narrowest-over-threshold approach of Baranowski et al. (2019). We start by drawing a large number of random intervals $[s_1, e_1], \dots, [s_Q, e_Q]$ and perform a test in each of these intervals to find windows that contain at least one change-point (Line 5 of Algorithm 7, with justification given by Corollary 3.4 in Section 3.4). We then select the narrowest interval for which the test rejects the null and apply Algorithm 6 to estimate a change-point within that window. We then partition the data into two submatrices to the left and right of this identified change-point and repeat the above procedures until no windows within the segmented submatrices contain any change-point.

Algorithm 5: Frank–Wolfe algorithm for optimising (3.6)

```
Input: $T \in \mathbb{R}^{p\times(n-1)}$, grouping $(\mathcal{J}_g)_{g\in[G]}$, $\lambda > 0$ and $\epsilon > 0$.
Initialise $\hat{M}^{[0]} = T/\|T\|_{\mathrm{F}}$ and $i = 0$.
repeat
    $i \leftarrow i + 1$
    Compute $G^{[i]} = (G_1^{[i]}, \ldots, G_p^{[i]})^\top \in \mathbb{R}^{p\times(n-1)}$ such that
        $$G_{j,t}^{[i]} \leftarrow T_{j,t} - \sum_{g: j\in\mathcal{J}_g} \lambda_g \frac{M_{j,t}^{[i-1]}}{\|M_{\mathcal{J}_g,t}^{[i-1]}\|_{\mathrm{F}}},$$
      where $\lambda_g = p_g^{1/2}\lambda$
    if $G^{[i]} = 0$ then break
    Compute
        $$\tilde{M}^{[i]} = \frac{i}{i+2} M^{[i-1]} + \frac{2}{i+2}\frac{G^{[i]}}{\|G^{[i]}\|_{\mathrm{F}}},$$
    Normalise $\hat{M}^{[i]} \leftarrow \tilde{M}^{[i]}/\|\tilde{M}^{[i]}\|_{\mathrm{F}}$
until $\|\hat{M}^{[i+1]} - \hat{M}^{[i]}\|_{\mathrm{F}} \leqslant \epsilon$;
Output: $\hat{M}^{[i]}$
```

Algorithm 6: Single change-point estimation procedure for data with group structure

```
Input: $X \in \mathbb{R}^{p\times n}$, $(\mathcal{J}_g)_{g\in[G]}$, and $\lambda > 0$
Compute $T \leftarrow \mathcal{T}(X)$ as in (2.9).
Solve
    $$\hat{M} \in \operatorname*{arg\,max}_{M\in\mathcal{S}} \bigl\{\langle T, M\rangle - \lambda\|M\|_{\mathrm{grp}}\bigr\}$$
  using either the closed-form solution in (3.7) if groups are non-overlapping, or
  Algorithm 5.
Let $\hat{v}$ be the leading left singular vector of $\hat{M}$.
Estimate $z$ by $\hat{z} = \arg\max_{1\leqslant t\leqslant n-1} |\hat{v}^\top T_t|$, where $T_t$ is the $t$th column of $T$.
Output: $\hat{z}$, $\bar{T}_{\max} = |\hat{v}^\top T_{\hat{z}}|$
```

Algorithm 7: Multiple change-point estimation procedure

```
Input: $X \in \mathbb{R}^{p \times n}$, $(\mathcal{J}_g)_{g \in [G]}$, $\lambda > 0$, $\beta$, $M \in \mathbb{N}$
Set $\hat{Z} \leftarrow \emptyset$
Draw $M$ pairs of integers $(s_1, e_1), \ldots, (s_M, e_M)$ uniformly at random from the set
  $\{(\ell, r) \in \mathbb{Z}^2 : 0 \leqslant \ell < r \leqslant n\}$
Function NOT($s$, $e$)
    Set $\mathcal{M}_{s,e} = \{m \in [M] : s \leqslant s_m < e_m \leqslant e\}$
    Set $\mathcal{R}_{s,e} := \{m \in \mathcal{M}_{s,e} : \|\mathcal{T}(X^{(s_m+\beta, e_m-\beta]})\|_{\mathrm{grp}*} > \lambda\}$, where $X^{(a,b]}$ is the
      submatrix of $X$ obtained using columns indexed in $(a, b]$
    if $\mathcal{R}_{s,e} \neq \emptyset$ then
        Find $m^* \in \arg\min_{m \in \mathcal{R}_{s,e}} |e_m - s_m|$
        Set $\hat{z}^{[m^*]}$ as the output from Algorithm 6 with inputs $X^{(s_{m^*}, e_{m^*}]}$ and $\lambda$
        $b \leftarrow \hat{z}^{[m^*]} + s_{m^*}$
        $\hat{Z} \leftarrow \hat{Z} \cup \{b\}$
        Run recursively NOT($s, b$) and NOT($b, e$)
Output: $\hat{Z}$
```

3.4 Theoretical results

In this section, we provide theoretical guarantees to the performance of the `groupInspect` algorithm. As we have noted in Section 3.2, a key to the successful change-point estimation in the current problem is a good estimator of the oracle projection direction $v = \theta/\|\theta\|_2$.

The following theorem controls the sine angle risk of the estimated projection direction $\hat{v}$ in Step 3 of Algorithm 6 when data has a single change. We define the following set to be the set of data distribution that satisfying our data generating mechanism:

Definition 1. *Suppose that data $X = (X_1, \ldots, X_n)$ is generated from a probability distribution P. We say that $P \in \mathcal{P}^{(\nu)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$ if it satisfies (3.1), (2.8), (3.2) and (3.3). For any $P \in \mathcal{P}^{(\nu)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$, we write $v(P) = \theta/\|\theta\|_2$ where θ is the difference between post-change and pre-change means.*

Theorem 3.1. *For a given grouping $(\mathcal{J}_g)_{g\in[G]}$, let $p_* = \min_{g\in[G]} |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \leqslant C_1$. Let $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$ be a $p \times n$ data matrix, let θ be the vector of change and let $\hat{v}$ be as in Step 3 of Algorithm 6 with input X, $(\mathcal{J}_g)_{g\in[G]}$ and $\lambda \geqslant B^{1/2}(1 + \sqrt{8\log(nG)/p_*})$. Then there exists $C > 0$, depending only on C_1, such that*

$$\sup_{P\in\mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{P}_P\left\{ \sin\angle(\hat{v}, v) > \frac{C\lambda k^{1/2}}{n^{1/2}\tau\vartheta} \right\} \leqslant \frac{1}{(nG)^3}. \tag{3.8}$$

We remark that the condition $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \leqslant C_1$ is to control the extent of overlapping between different groups. Specifically, it requires that each coordinate can belong to at most C_1 groups. In the special case when all groups $\mathcal{J}_g$ are disjoint, which is often true in practical applications, then it suffices to take $C_1 = 1$.

To understand the probabilistic bound in (3.8), we consider the optimal tuning parameter $\lambda = B^{1/2}(1 + \sqrt{8\log(nG)/p_*})$, for which the upper bound on the sine angle loss is of order $\sqrt{Bk(1 + \log(nG)/p_*)/(n\tau^2\vartheta^2)}$. The upper bound reveals an interesting interaction between the ℓ_0-sparsity k and the group sparsity s when all groups are of comparable size. Specifically, assuming that $\max_{g\in[G]} p_g \lesssim p_*$, from (3.8) we can derive for this particular choice of λ that if $\vartheta \lesssim n^{5/2}$, then

$$\mathbb{E}\{\sin \angle(\hat{v}, v)\} \leqslant \frac{C\lambda k^{1/2}}{n^{1/2}\tau\theta} + \frac{1}{(nG)^3} \lesssim \sqrt{\frac{B\{k + s\log(nG)\}}{n\tau^2\vartheta^2}}.$$

In other words, when the number of coordinates per group is at least of order $\log(nG)$, the risk upper bound is of order $\sqrt{\frac{Bs\log(nG)}{n\tau^2\vartheta^2}}$. On the other hand, when number of coordinates per group is smaller than this order, the risk upper bound is of order $\sqrt{\frac{Bk}{n\tau^2\vartheta^2}}$. Similar phase transitions have been previously observed in the context of high-dimensional linear model where the regression coefficients satisfy a group sparsity assumption (see, e.g. Cai et al., 2019, Theorem 3). We also note that when $G = p$, each group only has one element, and the modelling assumption is identical to that in Wang and Samworth (2018) and $s = k$ in this case. The upper bound of the sine angle loss has the order of $\sqrt{\frac{Bk\log(pn)}{n\tau^2\vartheta^2}}$, which could be slightly worse comparing to the $\sqrt{\frac{Bk\log(p\log n)}{n\tau^2\vartheta^2}}$ bound achieved in (Wang and Samworth, 2018, Proposition 1) if $\log n \gg \log p$. However, if the group structure is nontrivial in the sense that each group has at least $\log(nG)$ elements, then the projection direction estimator in `groupInspect` is closer to the truth compared to that from the `Inspect` algorithm.

We now turn our attention to a minimax lower bound of the estimation risk of the oracle projection direction. Theorem 3.2 below shows that the phase transition observed in Theorem 3.1 is not due to the specific proof techniques employed but rather an intrinsic feature of the problem.

Theorem 3.2. *Suppose $s > 0$, $k > 0$ and a grouping $(\mathcal{J}_g)_{g\in[G]}$ satisfy that $\mathcal{J}_g \cap \mathcal{J}_{g'} = \emptyset$ for all $g \neq g'$, $\min\{k, (s-1)\log(G/s)\} \geqslant 20$, and $\sum_{r=1}^{s} p_{(G-r+1)} \geqslant k/2$, where $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(G)}$ are order statistics of $p_1, \ldots, p_G$. Let $\Sigma = BI_p$. Then for some universal constant $c > 0$, we have*

$$\inf_{\tilde{v}} \sup_{P \in \mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{E}_P L(\tilde{v}(X), v(P)) \geqslant c\sqrt{\frac{B\{k + s\log(G/s)\}}{n\tau\vartheta^2}},$$

where the infimum is taken over the set of all measurable functions $\tilde{v}$ of the data X.

The condition that $\sum_{r=1}^{s} p_{(G-r+1)} \geqslant k/2$ is to ensure that the upper bound k on the ℓ_0-sparsity is not too loose in the sense that k is not too much larger than the cardinality of the union of the largest s groups. If we assume that $\log(G/s) \asymp \log(n)$, $\tau \asymp 1$ and $\max_{g\in[G]} p_g \lesssim p_*$, then the lower bound in Theorem 3.2 matches the upper bound of Theorem 3.1 up to universal constants, when all groups are non-overlapping. We remark that the upper and lower bounds in Theorems 3.1 and 3.2 do not match in their dependence on the parameter τ. As Proposition 3.15 shows, this suboptimality is unlikely due to the convex relaxation carried out in (3.6) since the same τ dependence appears in the risk upper bound of the (computationally infeasible) optimiser of (3.4). We further remark that if we derive this lower bound using Σ instead of BI_p, ϑ^2 in the denominator would become $\|\Sigma^{-1}\theta\|_2^2 \leqslant \lambda_{\min}^2/\vartheta^2$. This implies our derived bound in Theorem 3.1 may be sub-optimal in the generic setting.

After obtaining guarantees on the quality of the projection direction estimator, we now provide theoretical guarantees of the overall change-point procedure. We note that the projection direction estimator $\hat{v}$ is dependent on the CUSUM panel T. While this dependence is observed to be very weak in practice, it creates difficulties in analysing the projected CUSUM series $\hat{v}^\top T$ in Step 4 of Algorithm 6. As such, for theoretical convenience, we will instead analyse a sample-splitting version of the algorithm. Specifically, we split the data into $X^{(1)}$ and $X^{(2)}$, consisting of odd and even time points respectively, as described in Algorithm 8. We use $X^{(1)}$ to estimate the projected direction $\hat{v}^{(1)}$ and then project $X^{(2)}$ along this direction to locate the change-point via a univariate CUSUM procedure (Step 4 of Algorithm 6). Theorem 3.3 below provides a performance guarantee for the estimated location of the change-point of this sample-splitting version of our procedure.

Theorem 3.3. *Given data matrix $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})$, let $\hat{z}$ be the output from the Algorithm 8 with input X and $\lambda \geqslant B^{1/2}(1+\sqrt{p_*^{-1}8\log(nG)})$. There exist universal constants C, $C' > 0$ such that, if $n \geqslant 12$ is even, z is even, and*

$$\frac{C\sqrt{k}\lambda}{\vartheta\tau\sqrt{n}} \leqslant 1, \tag{3.9}$$

then for any $\lambda_1 > \sqrt{B}$, we have

$$\mathbb{P}\left\{\frac{1}{n}|\hat{z}-z| \leqslant \frac{C'\lambda_1^2}{n\vartheta^2}\right\} \geqslant 1 - \frac{8}{n^3} - (3\lambda_1+1)e^{-\lambda_1^2/(4B)}\log n.$$

Algorithm 8: Change-point estimation procedure: sample splitting version

```
Input: $X \in \mathbb{R}^{p\times n}$ and $\lambda > 0$
Define $X^{(1)}$ as $X^{(1)}_{j,t} = X_{j,2t-1}$ and $X^{(2)}$ as $X^{(2)}_{j,t} = X_{j,2t}$.
Compute $T^{(1)} \leftarrow \mathcal{T}(X^{(1)})$ and $T^{(2)} \leftarrow \mathcal{T}(X^{(2)})$ as in (2.9).
Solve
    $$\hat{M}^{(1)} \in \arg\max_{M\in\mathcal{S}} \bigl\{\langle T^{(1)}, M\rangle - \lambda\|M\|_{\mathrm{grp}}\bigr\}$$
    using either the closed-form solution in (3.7) if groups are non-overlapping, or Algorithm 5.
Let $\hat{v}$ be the leading left singular vector of $\hat{M}^{(1)}$.
Estimate $z$ by $\hat{z} = 2\arg\max_{1\leqslant t\leqslant n_1-1} |(\hat{v}^{(1)})^\top T^{(2)}_t|$, where $T^{(2)}_t$ is the $t$th column of $T^{(2)}$.
Output: $\hat{z}$
```

If we choose $\lambda_1 = C\sqrt{B\log\log n}$ for a sufficiently large absolute constant $C > 0$, then Theorem 3.3 shows that the location estimator $\hat{z}/n$ converges to z/n at a rate of $\frac{B\log\log n}{n\vartheta^2}$ in probability. This rate is minimax optimal even for the problem of estimating a single change in mean in a univariate series; see Proposition 3.10. While Theorem 3.3 concerns primarily with the estimation task, we remark that the argument used in its proof can be easily adapted to derive a testing procedure with good theoretical guarantees. Specifically, given data matrix $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$, we are interested to test the null hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta \neq 0$. We can construct a test based on the dual norm to the $\|\cdot\|_{\mathrm{grp}}$ norm defined in (3.5). More precisely, for any $R \in \mathbb{R}^{p\times n}$ and a grouping $(\mathcal{J}_g)_{g\in[G]}$ of $[p]$, we define

$$\|R\|_{\mathrm{grp}*} = \max_{g\in[G]}\max_{t\in[n]} p_g^{-1/2}\|R_{\mathcal{J}_g,t}\|_2. \tag{3.10}$$

It can be seen from Lemma 3.11 that $\|\cdot\|_{\mathrm{grp}*}$ is indeed dual to $\|\cdot\|_{\mathrm{grp}}$. For any $\lambda > 0$, we define a test ψ_λ such that

$$\psi_\lambda(X) = \mathbb{1}_{\{\|\mathcal{T}(X)\|_{\mathrm{grp}*}\geqslant\lambda\}}.$$

The following Corollary shows that with an appropriately chosen testing threshold λ, the test ψ_λ defined above has good size and power controls.

Corollary 3.4. *Given data matrix $X \sim P \in \mathcal{P}_{n,p}^{(1)}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$. Let k be the total number of coordinates with change and $\|\theta\|_2$ be the magnitude of the change. Fix $\lambda \geqslant B^{1/2}\big(1 + \sqrt{4p_*^{-1}\log(nG)}\big)$.*

- *If $s = 0$, then $\mathbb{P}_P(\psi_\lambda(X) = 1) \leqslant 1/(nG)$.*
- *If $\vartheta \geqslant \frac{\sqrt{8k}\lambda}{\sqrt{n\tau}}$, then $\mathbb{P}_P(\psi = 1) \geqslant 1 - 1/(nG)$.*

Our single change-point theory can be applied iteratively to show that the `groupInspect` algorithm in in Algorithm 7 can consistently estimate both the number and the locations of the true change-points. In line with Theorem 3.3, we consider a sample-splitting version of Algorithm 7, which we call Algorithm Algorithm 7′, where we use Algorithm 8 in place of Algorithm 6 in line 8 of Algorithm 7.

Theorem 3.5. *Given data matrix $X \sim P \in \mathcal{P}_{n,p}^{(\nu)}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$. Let $\hat{Z}$ be the output from the Algorithm 7′ with input X and $\lambda = B^{1/2}(1 + \sqrt{8p_*^{-1}\log(nG)})$, Q and $\beta = n\tau/10$. Let $\tau\sqrt{n} \geqslant C'B\log n/\vartheta^2$. There exist universal constants C, $C' > 0$ such that, if $n \geqslant 12$ is even, z is even, and*

$$\frac{C\sqrt{Bk}}{\vartheta\tau\sqrt{n\tau}}\left(1 + \sqrt{\frac{8\log(nG)}{p_*}}\right) \leqslant 1, \tag{3.11}$$

then,

$$\mathbb{P}\left(\hat{\nu} = \nu \text{ and } |\hat{z}_i - z_i| \leqslant \frac{C'B\log n}{\vartheta^2}\ \forall\, i \in [\nu]\right) \geqslant 1 - \nu e^{-\tau^2 M/36} - \frac{1}{nG^3} - \frac{7}{n\tau^3}.$$

Theorem 3.5 shows that under suitable assumptions about the spacings between consecutive change-points, our multiple change-point estimator $\hat{z}_i/n$ converges to z_i/n with a rate of convergence $B\log n/(n\vartheta^2)$. Up to a logarithmic factor, this rate is essentially the same as the rate for single change-point estimation as proved in Theorem 3.3.

3.5 Numerical studies

In this section, we provide some simulation results to demonstrate the empirical performance of the `groupInspect` method. In all our numerical studies, unless otherwise specified, we will assume that data are generated according to (3.1), (2.8), (3.2) and (3.3).

In all simulations, we do not assume that the covariance matrix Σ is known. Instead, we estimate the variance in each row using the mean absolute deviation of successive differences of the observations. We then standardise the data by the estimated row standard deviation. The `groupInspect` procedure is then applied to the standardised data assuming that Σ is a well-conditioned matrix with all diagonal entries equal to 1.

3.5.1 Theory validation

We first show that the practical performance of the `groupInspect` procedure is well captured by the theoretical results in Theorems 3.1 and 3.2. There are two related measures of the signal sparsity in our problem, which are the total number of coordinates of change k and the total number of groups with a change s. We conduct two sets of simulation experiments fixing one of these sparsity measures and varying the other. Specifically, for $n = 1000$, $p \in \{600, 1200, 2400\}$ and $\vartheta \in \{1, 2, 4, 8, 16\}$ and $\Sigma = I_p$, we split the p coordinates into disjoint groups of p_* coordinates per group, where p_* is allowed to vary over all divisors of 60. In the first set of experiments, we fix $k = 60$ so that $s = k/p_*$ varies with p_*, whereas in the second set of experiments, we fix $s = 3$ so that $k = sp_*$ varies with p_*. The vector of change is constructed so that the magnitude of change is equal across all coordinates of change. We will use the theoretical choice of tuning parameter λ for both sets of experiments here. Figure 3.1 shows how the $\sin\theta$ loss, averaged over 100 Monte Carlo repetitions, varies with p_*, for different choices of p and ϑ in both settings.

In the left panel of Figure 3.1, where the number of signal coordinates k is fixed, we see that the average loss decreases as p_* increases. Furthermore, at a log-log scale, and for relatively large signal sizes of $\vartheta \in \{4, 8, 16\}$, we see the loss curves follow an initial linear decreasing trend as p_* increases before plateauing eventually. This is in agreement with the two terms contributing to the loss described in Theorem 3.1. Specifically, for small p_*, we expect the second term of (3.8) to dominate and the loss decreases at a rate approximately proportional to $1/\sqrt{p_*}$ initially. For large p_*, we expect the first term of (3.8) to dominate and the loss will have minimal dependence on p_*. In the right panel of Figure 3.1, where the number of signal groups s is fixed, the average loss increases

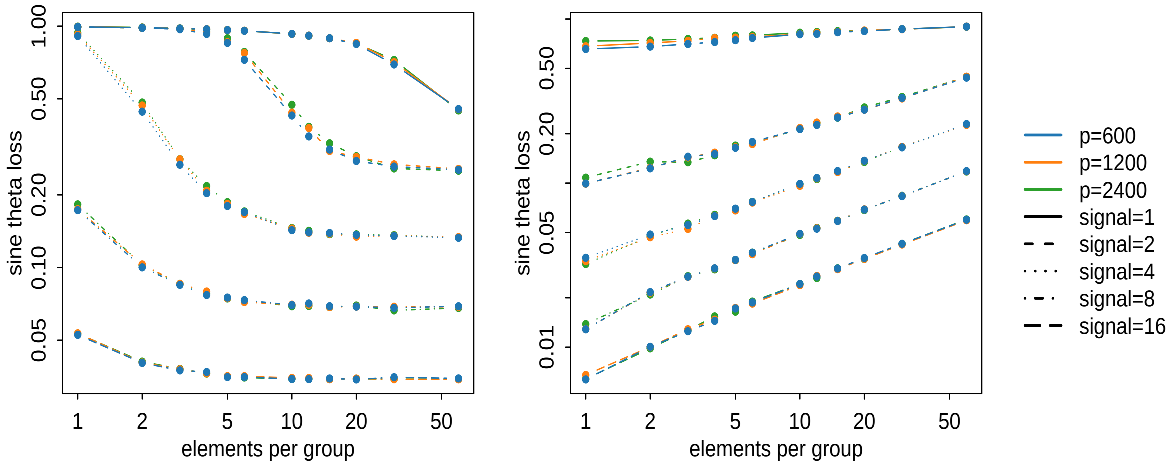

Figure 3.1: Average loss (over 100 repetitions) of `groupInspect` for varying elements per group p_*, plotted on a log-log scale. Left panel: $k = 60$ and $s = k/p_*$. Right panel: $s = 3$ and $k = sp_*$. Other parameter: $n = 1000$.

with p_*, as expected from our theory. It appears that for $s = 3$ studied here, the first term of (3.8) is dominant and the average loss increases linearly at the log-log scale with respect to p_*.

We further remark that in both panels of Figure 3.1, the average loss for large p_* shows equally spaced separation for the signal size ϑ in the dyadic grid $\{1, 2, 4, 8, 16\}$. This is in good agreement with the $1/\theta$ dependence of expected loss given in Theorem 3.1. Finally, we note that the ambient dimension p has minimal effect on the loss curves, for all signal strengths studied here. Again, this is predicted by our theory as the dimension p enters the mean loss in (3.8) only through the $\log(nG) = \log(pn/p_*)$ expression in the second term.

3.5.2 Practical choice of tuning parameter

The theoretical choice of λ turns out to be conservative in practical use. In this subsection, we will perform numerical simulations to suggest a suitable practical tuning parameter choice. We fix $n = 1000$, $z = 400$, $s = 3$, $G \in \{10, 25\}$ and assume $\Sigma = I_p$. The signal size ϑ is varied in $\{1, 2, 4, 8, 16\}$ and p is chosen from $\{500, 1000\}$. All groups are set to have equal size. We run the `groupInspect` algorithm for tuning parameters $\lambda = a(1 + \sqrt{4p_*^{-1} \log(nG)})$, where a is chosen from a logarithmic sequence of values between 0.1 and 3.

We plot $\sin\theta$ loss against a in Figure 3.2. In most cases, the loss is minimised when $a \approx 1/2$, i.e. tuning parameter value is half of the theoretical value. This suggests that when $\Sigma = I_p$, the choice $\lambda = 2^{-1}(1 + \sqrt{4p_*^{-1} \log(nG)})$ leads to more accurate estimation in practice. Theorems 3.1 and 3.3 suggest that for non-identity covariance structure, the tuning parameter choice should scale proportional to the square root of the operator norm of Σ. It is in general a challenging statistical problem to estimate the operator norm of the covariance matrix in a high-dimensional setting. One can in principal use the estimator proposed by Liu, Gao and Samworth (2021), though we observe that this estimator typically incurs a large upward bias when the dimension is high in comparison to the sample size. Moreover, an inspection of our proof reveals that the presence of the additional factor B is used to capture some worst-case large deviation bound, which is often too conservative for a generic covariance Σ. In view of the above, we recommend that practitioners use the same $\lambda = 2^{-1}(1 + \sqrt{4p_*^{-1} \log(nG)})$ when Σ is unknown.

3.5.3 Comparison between different methods

Now, we would like to compare our method with other existing change-point estimation procedures. As `groupInspect` is a two-stage procedure that first estimates a projection direction before localising the change-point on the projected series, we will investigate its performance both in terms of its accuracy in estimating the projection direction and the quality of the final change-point location estimator. For the former, we compare the estimated projection direction from `groupInspect` with that from the `Inspect` algorithm. We measure the accuracy in terms of the sine angle loss introduced in Section 3.2. We use the recommended values for tuning parameters in both methods, i.e., $\sqrt{2^{-1} \log\{p \log n\}}$ in `Inspect` as in Wang and Samworth (2018) and $2^{-1}(1+\sqrt{4p_*^{-1} \log(nG)})$ for `groupInspect` as suggested in Section 3.5.2.

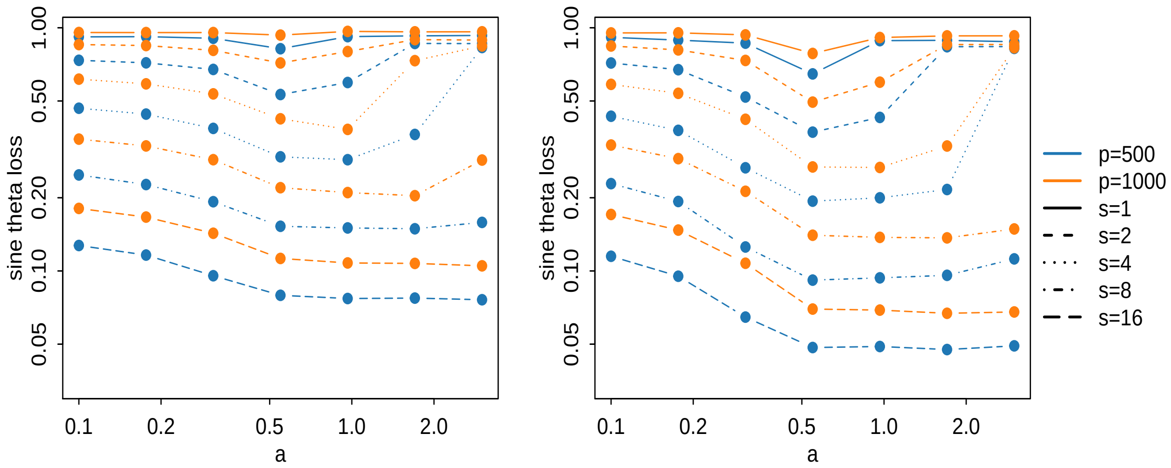

Figure 3.2: Average loss (over 100 repetitions) of `groupInspect` for tuning parameter $\lambda = a(1 + \sqrt{4p_*^{-1}\log(nG)})$ with varying choice of a. Left panel: $G = 10$. Right panel: $G = 25$. Other parameter: $n = 1000$, $s = 3$.

We fix $n = 1000$, $p = 1000$, vary ϑ in $\{1, 2, 4, 8, 16\}$ and set the covariance matrix to be $\Sigma = I_p$. We consider settings with both non-overlapping groups and overlapping groups. For the non-overlapping setting, we have $G = 10$ groups of equal size $p_* = 100$, whereas for the overlapping setting, we have $G = 19$ groups of size 100 each, where neighbouring groups overlap in exactly 50 coordinates. Both methods have access to exactly the same data sets and the performance is averaged over 100 Monte Carlo repetitions.

Figure 3.3 shows the comparison of the average sine angle loss between `Inspect` and `groupInspect` over all signal sizes on a logarithmic scale, in both the non-overlapping and overlapping settings. In both cases, `groupInspect` outperforms the `Inspect` algorithm. From the left panel, we can see that the estimation accuracy of the projection direction using `groupInspect` is substantially better even when the signal is small.

We now turn our attention to the overall change-point localisation accuracy of the `groupInspect` procedure. To this end, we compare the mean absolute deviation of various high-dimensional change-point procedures over 300 Monte Carlo repetitions using the same data sets. In addition to `Inspect`, we also compare against the ℓ_2 aggrega-

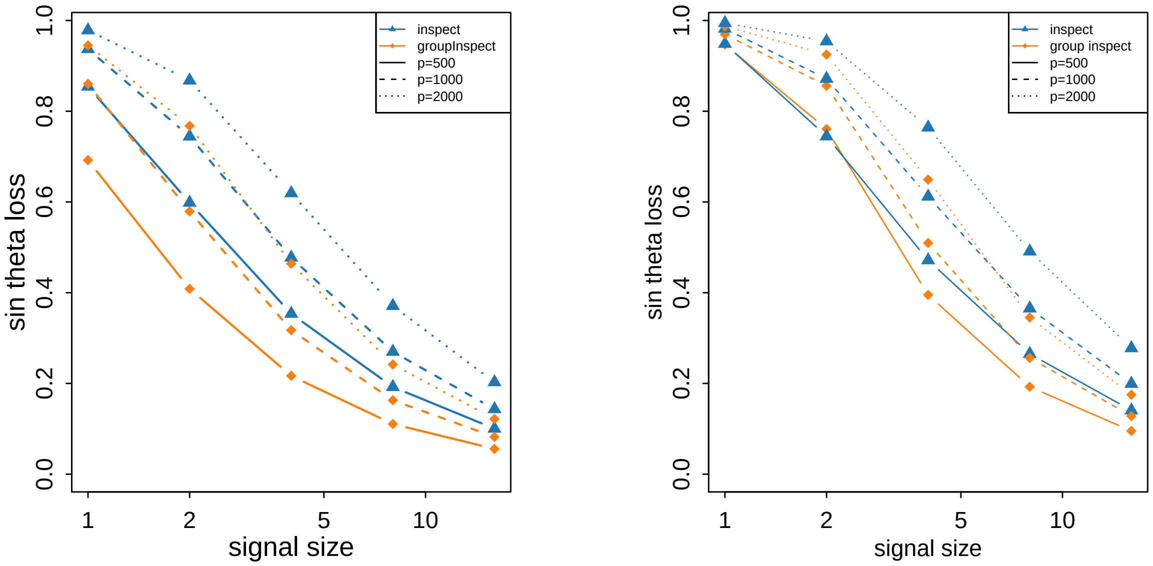

Figure 3.3: Average loss (over 100 repetitions) comparison between `groupInspect` and `Inspect`. Left panel: non-overlap setting. Right panel: overlap setting

tion procedures of Horváth and Hušková (2012) (ℓ_2-agg), the ℓ_∞ aggregation procedure of Jirak (2015) (ℓ_∞-agg), the double CUSUM procedure of Cho (2016) (DC) and a multiscale testing procedure Pilliat et al. (2020). We set $n = 1000$, $p \in \{500, 1000, 2000\}$, $\vartheta \in \{0.25, 0.5, 1, 2, 4\}$ and $\Sigma = (2^{-|j-k|})_{j,k\in[p]}$. The simulation results are presented in Table 3.1. For simplicity, we have only shown the results for 10 equal-sized non-overlapping groups here, but qualitatively similar results were obtained in other settings as well. We see that `groupInspect` is very competitive over a wide range of dimensions and signal-to-noise ratio settings, and `groupInspect` dominates the `Inspect` procedure in all simulation settings by successfully exploiting the group-sparsity structure.

3.5.4 Multiple change-points simulation

The numerical studies so far have focused mainly on the single change-point estimation problem. In this subsection, we investigate the empirical performance of `groupInspect` in multiple change-point estimation tasks. We will compare its performance as implemented in Algorithm 7 to that of the `Inspect` algorithms for estimating multiple change-points

p	ϑ	`groupInspect`	`Inspect`	ℓ_2-agg	ℓ_∞-agg	DC	pilliat
500	0.25	151	158	370	368	364	**113**
500	0.5	**89.6**	98.6	271	332	298	102.6
500	1	**8.7**	14.8	18.5	108	66.8	56.82
500	2	**0.95**	1.30	1.64	15.9	5.42	19.53
500	4	**0.057**	0.063	0.080	3.11	0.51	15
1000	0.25	116	147	368	344	385	**115**
1000	0.5	**85**	120	309	316	335	102
1000	1	**23.4**	32.6	41.0	194	110	67.2
1000	2	**1.31**	1.67	2.04	32.2	7.47	24.36
1000	4	**0.09**	0.14	0.123	6.29	0.850	15
2000	0.25	**106**	128	356	356	374	131
2000	0.5	**89.6**	118	321	344	341	119
2000	1	**47.61**	55.56	106	283	177	92.91
2000	2	**2.91**	3.23	3.39	63.3	10.4	39.141
2000	4	**0.11**	0.160	0.17	9.94	1.32	30.75

Table 3.1: Average mean absolute deviation (over 300 repetitions) comparison between different methods. Other parameters used: $n = 1000$ with $G = 10$

under different settings. We choose $n = 1200$, $p \in \{500, 1000\}$, $s \in \{3, 10\}$, $G \in \{50, 100\}$ and $\Sigma = I_p$. Each data series contains three true change-points located at 300, 600 and 900 with the ℓ_2 norm of the change equal to ϑ, 1.5ϑ and 2ϑ respectively. We vary ϑ in $\{0.6, 0.8, 1, 1.2, 1.4\}$. For simplicity, we further assume that the same s coordinates undergo change in all three change-points and that all groups have 10 elements. The total number of coordinates with change k is calculated as $10s$. We use the λ tuning parameter choice suggested in Section 3.5.2 for the `groupInspect` method and that suggested in Wang and Samworth (2018) for the `Inspect` algorithm. For the thresholding parameter ξ of the wild binary segmentation recursion used in both `groupInspect` and `Inspect`, we choose via Monte Carlo simulation. More precisely, we randomly generate 1000 data sets from the null model with no change-points and take the maximum absolute CUSUM statistic from Algorithm 7 and Wang and Samworth (2018, Algorithm 4) as ξ_g and ξ_i respectively. We compare the performance of two algorithms using the Adjusted Rand index (ARI) of the estimated segmentation against the truth (Rand, 1971; Hubert and Arabie, 1985).

From Figure 3.4, we see that the `groupInspect` algorithm generally performs much better than the `Inspect` algorithm in the multiple change-point localisation tasks. The advantage of `groupInspect` is more pronounced when the signal is sparser and when the dimension of the data is higher.

To further visualise the output of the two procedures, we plot the estimated change-point locations for one specific setting ($s = 3$ and $\vartheta = 1$) of each of the two panels in Figure 3.4. The resulting histograms in Figure 3.5 show that when $p = 500$, `groupInspect` was better at picking out all three change-points with higher accuracy. When $p = 1000$, `Inspect` was only able to pick out the change at $t = 600$ in most of the trials, whereas `groupInspect` was still able to identify even the weakest change signal at $t = 300$ in a substantial fraction of all trials.

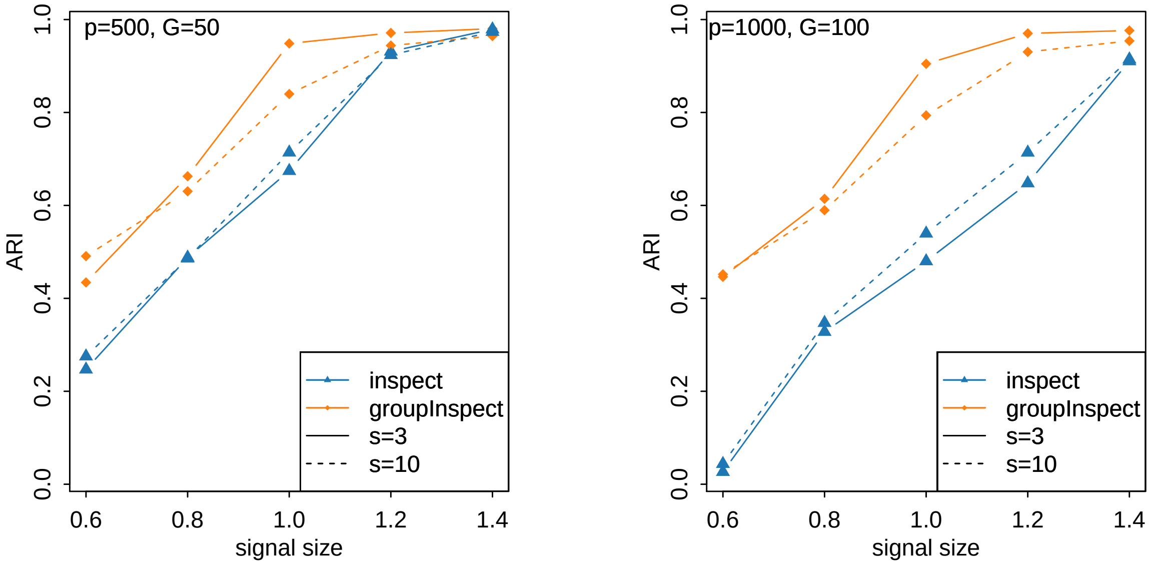

Figure 3.4: Average ARI comparsion between `groupInspect` and `Inspect`. Left panel: $p = 500, G = 50$. Right panel: $p = 1000, G = 100$.

3.5.5 Real data analysis

In this section, we apply `groupInspect` to an S&P 500 daily stock return dataset. The data consists of the logarithmic daily returns (computed from the adjusted closing prices) of S&P 500 stocks traded during the period of 1 January 2007 to 31 December 2011. We only included the 257 stocks which have continuously traded throughout this period to construct a multivariate time series of dimension $p = 257$ and length $n = 1259$. We divided the 257 companies into $G = 11$ non-overlapping groups according to their Global Industry Classification Standard sector memberships. For each stock logarithmic returns, we fitted an AR(1) model, and then rescaled the residuals by their estimated standard deviation according to the method described in Section 3.5.

Figure 3.6 displays the ten most significant change-points identified by our `groupInspect` algorithm. For each change-point, we derived a sector-weighting vector from the estimated projection direction by `groupInspect`. Specifically, given the projection direction $\hat{v} \in \mathbb{S}^{p-1}$ for each estimated change-point, and the grouping $(\mathcal{J}_g)_{g\in[G]}$, we computed a weight vector $\hat{w} := (\|\hat{v}_{\mathcal{J}_g}\|)_{g\in[G]}$. This vector gives us information about which sectors

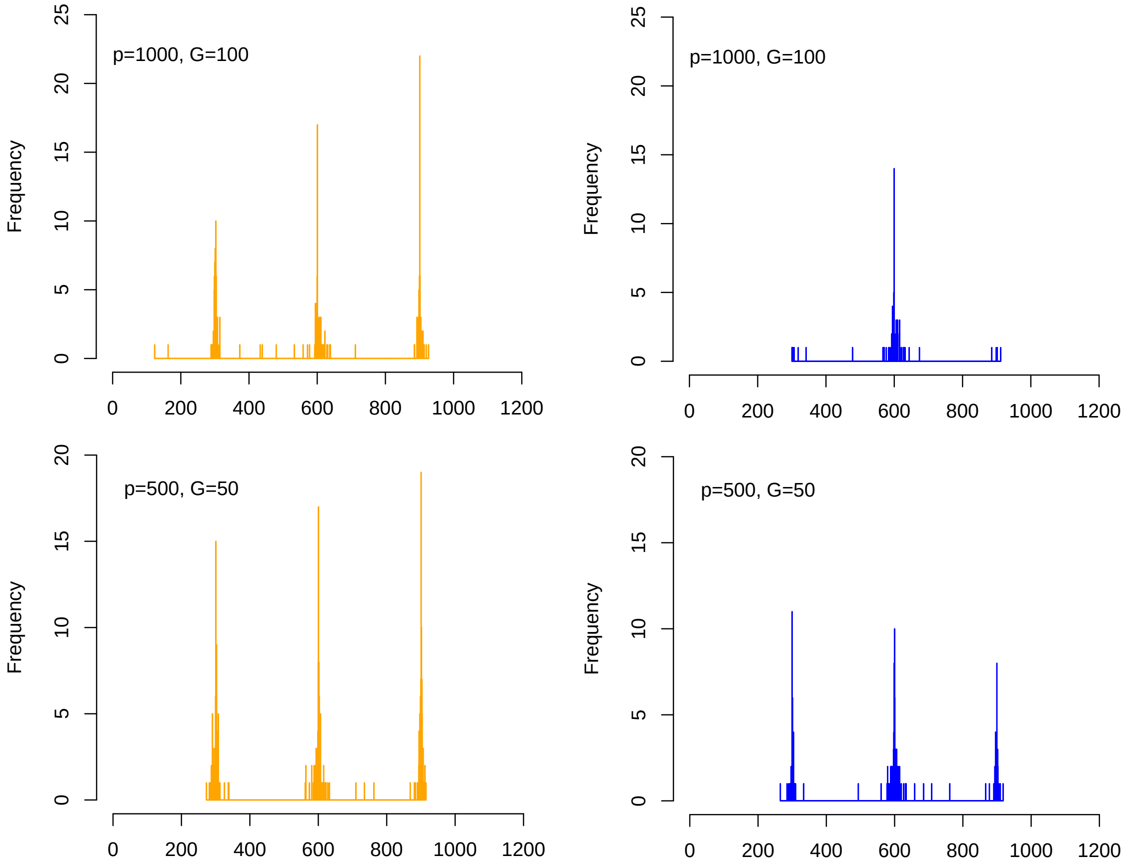

Figure 3.5: Histograms of estimated locations by `groupInspect` and `Inspect` under two settings when $P = 500, G = 50$ and $p = 1000, G = 100$. Other parameter used: $s = 3$, $\vartheta = 1$ are fixed in both settings.

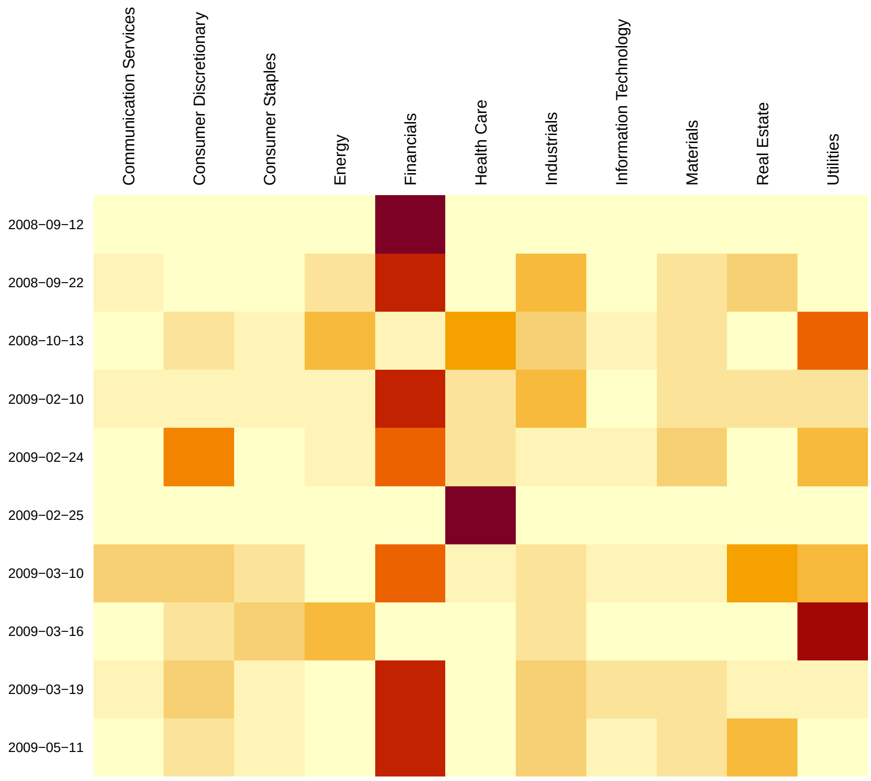

Figure 3.6: Estimated change point locations (red dashed lines) by `groupInspect` applied to the stock return data. For ease of illustration, we have plotted the ℓ_2 norm of the returns of all stocks within each of the 11 groups over time.

had driven the change for each change-point estimated. For instance, we see from Figure 3.6 that the change-point at 12 September 2008 was predominantly driven by price fluctuations in financial stocks, which coincides with the Federal takeover of Fannie Mae and Freddie Mac on 7 September 2008 and the bankruptcy of Lehman Brothers on 15 September 2008. The change-point identified at 10 February 2009, though still heavily weighted on financial stocks, showed a broader impact across other sectors. This is consistent with the passing of the American Recovery and Reinvestment Act of 2009 on 13 February 2009 sending a general positive signal to the entire economy.

3.6 Extensions to sub-Gaussian distributions

In the previous sections, we assumed that $X_i = \mu_i + W_i$, for $W_1, \ldots, W_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$. In this session, we discuss how the previous results can be generalised to settings where $W_1, \ldots, W_n$ are independent sub-Gaussian random vectors. Adpoting notation from Zhu, Wang and Samworth (2022), for any random vector U in $\mathbb{R}^p$, we write

$$\|U\|_{\psi_2} := \sup_{w \in \mathcal{S}^{p-1}} \sup_{q \in \mathbb{N}} \frac{\mathbb{E}(|w^\top U|^q)^{1/q}}{\sqrt{q}},$$

$$\|U\|_{\psi_2^*} := \sup_{w \in \mathcal{S}^{p-1}} \frac{\|w^\top U\|_{\psi_2}}{(w^\top \mathrm{Var}(U) w)^{1/2}} = \|\mathrm{Var}^{-1/2}(U) U\|_{\psi_2}.$$

We say that a U is a p dimensional sub-Gaussian random vector if $\|U\|_{\psi_2^*} < \infty$.

For sub-Gaussian data, Lemma 3.21 can be used in place of Lemma 3.18 to derive the equivalent result of Theorem 3.1 for the sub-Gaussian data. In this section we denote $\mathcal{P}^{(\nu)}_{n,p}(s, k, \tau, \vartheta, L, B, (\mathcal{J}_g)_{g \in [G]})$ as the data generating mechanism, where the data is generated in the same way as before except the noise W_i are drawn from sub-Gaussian distributions with $\|W_t\|_{\psi_2^*} \leqslant L$. Theorem 3.6 gives the result for the projection direction and Theorem 3.7 gives the result for the estimation accuracy of change-point location.

Theorem 3.6. *For a given grouping $(\mathcal{J}_g)_{g \in [G]}$, let $p_* = \min_{g \in [G]} |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j \in [p]} |\{g : j \in \mathcal{J}_g\}| \leqslant C_1$. Let $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, L, B, (\mathcal{J}_g)_{g \in [G]})$ be a $p \times n$ data matrix, let θ be the vector of change and let $\hat{v}$ be as in Step 3 of Algorithm 6 with input X, $(\mathcal{J}_g)_{g \in [G]}$ and $\lambda \geqslant C_2 L B^{1/2}(1 + \sqrt{\log(nG)/p_*})$. Then there exists $C > 0$, depending only on C_1, such that*

$$\sup_{P \in \mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,L,B(\mathcal{J}_g)_{g \in [G]})} \mathbb{P}_P\biggl\{ \sin \angle(\hat{v}, v) > \frac{C \lambda k^{1/2}}{n^{1/2} \tau \vartheta} \biggr\} \leqslant \frac{1}{nG}. \tag{3.12}$$

The proof follows the same argument as in that of Theorem 3.1, but using Lemma 3.21 in place of Lemma 3.18.

Theorem 3.7. *Given data matrix $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, L, B, (\mathcal{J}_g)_{g \in [G]})$, let $\hat{z}$ be the output from the Algorithm 8 with input X and $\lambda \geqslant C_2 L B^{1/2}(1 + \sqrt{\log(nG)/p_*})$. There exist universal constants C, $C' > 0$ such that, if $n \geqslant 12$ is even, z is even, and*

$$\frac{C\sqrt{k}\lambda}{\vartheta \tau \sqrt{n}} \leqslant 1, \tag{3.13}$$

then for any $\lambda_1 > L\sqrt{CB\log n}$*, we have*

$$\mathbb{P}\left\{\frac{1}{n}|\hat{z}-z| \leqslant \frac{C'\lambda_1^2}{n\vartheta^2}\right\} \geqslant 1 - \frac{8}{n^3} - \frac{16\log n}{n}.$$

The proof of this theorem follows from Lemma 3.23.

Theorem 3.6 and Theorem 3.7 show that we can obtain the similar theoretical guarantee results in terms of both projection direction and location estimation accuracy if the data is consisting of sub-Gaussian random vectors. We remark that the main challenge in the derivation of the theorems above here is how to bound the error terms, i.e, $\|E\|_{\mathrm{grp}^*}$, as the usual concentration bounds designed for normal distribution could not be applied here. We instead considered Hoeffding-type inequality. Also, we define the sub-Gaussian random variable in a way that it is invariant for linear transformation due to projection.

3.7 Extensions to temporal dependence

In this section, we consider the case when the columns of X are not independent across time. We assume that $W_1, \ldots, W_n$ are stationary and let $K(u) = \mathrm{Cov}(W_t, W_{t+u})$. We further assume that the dependence is short-ranged in the sense that:

$$\left\|\sum_{u=0}^{n-1} K(u)\right\|_{\mathrm{op}} \leqslant B^*. \tag{3.14}$$

The oracle projection direction maximises the signal-to-noise ratio after projection, which does not change in the presence of temporal dependence. The main difference, however, is that when the noise $W_1, \ldots, W_n$ are dependent across time, the CUSUM matrix E will have columns whose covariance will be inflated as we are taking a weighted average of correlated noise. Condition (3.14) ensures that the temporal correlation is short-ranged and that the covariance of the noise CUSUM is inflated by at most a constant factor, depending on B^*. The following theorems give the theoretical results on projection direction (Theorem 3.8) and estimation accuracy (Theorem 3.9) in this case. In this section we denote $\mathcal{P}_{n,p}^{(\nu)}(s, k, \tau, \vartheta, B^*, (\mathcal{J}_g)_{g\in[G]})$ as the data generating mechanism, where the data is generated in the same way as before except the noise W_i has the temporal dependence structure as in 3.14.

Theorem 3.8. *For a given grouping $(\mathcal{J}_g)_{g\in[G]}$, let $p_* = \min_{g\in[G]} |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \leqslant C_1$. Let $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B^*, (\mathcal{J}_g)_{g\in[G]})$ be a $p \times n$ data matrix, let θ be the vector of change and let $\hat{v}$ be as in Step 3 of Algorithm 6 with input X, $(\mathcal{J}_g)_{g\in[G]}$ and $\lambda \geqslant \sqrt{2B^*}(1 + \sqrt{8\log(nG)/p_*})$. Then there exists $C > 0$, depending only on C_1, such that*

$$\sup_{P\in\mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B^*,(\mathcal{J}_g)_{g\in[G]})} \mathbb{P}_P\left\{ \sin\angle(\hat{v}, v) > \frac{C\lambda k^{1/2}}{n^{1/2}\tau\vartheta} \right\} \leqslant \frac{1}{(nG)^3}. \tag{3.15}$$

Theorem 3.9. *Given data matrix $X \sim P \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B^*, (\mathcal{J}_g)_{g\in[G]})$, let $\hat{z}$ be the output from the Algorithm 8 with input X and $\lambda \geqslant \sqrt{2B^*}(1 + \sqrt{p_*^{-1} 8\log(nG)})$. There exist universal constants C, $C' > 0$ such that, if $n \geqslant 12$ is even, z is even, and*

$$\frac{C\sqrt{k}\lambda}{\vartheta\tau\sqrt{n}} \leqslant 1, \tag{3.16}$$

then for any $\lambda_1 > \sqrt{2B^*}$, we have

$$\mathbb{P}\left\{ \frac{1}{n}|\hat{z} - z| \leqslant \frac{C'\lambda_1^2}{n\vartheta^2} \right\} \geqslant 1 - \frac{8}{n^3} - (3\lambda_1 + 1)e^{-\lambda_1^2/(8B^*)} \log n.$$

The proofs of two theorems above follow from Lemma 3.24.

3.8 Proof of main results

In this section, we will give the proof of our results in Chapter 3.

3.8.1 Proof of Theorem 3.1

Proof. From the definition of the CUSUM transformation in (2.9), we can explicitly write the matrix $A := \mathbb{E}(T) = (A_{j,t})_{j\in[p],t\in[n-1]}$ as

$$A_{j,t} = \begin{cases} \sqrt{\frac{t}{n(n-t)}}(n - z)\theta_j & \text{if } 1 \leqslant t \leqslant z, \\ \sqrt{\frac{n-t}{nt}} z\theta_j & \text{if } z < t \leqslant n - 1. \end{cases}$$

In particular, we have that A is a rank 1 matrix of the form

$$A = \theta\gamma^\top, \tag{3.17}$$

with

$$\gamma = \frac{1}{\sqrt{n}}\Big(\sqrt{\frac{1}{n-1}}(n-z), \sqrt{\frac{2}{n-2}}(n-z), \cdots, \\ \sqrt{z(n-z)}, \sqrt{\frac{n-z-1}{z+1}}z, \cdots, \sqrt{\frac{1}{n-1}}z\Big)^\top.$$

By Wang and Samworth (2018, Lemma 3), we have $\|\gamma\|_2 \geqslant n\tau/4$, so $\|A\|_{\mathrm{op}} = \|\theta\|_2\|\gamma\| \geqslant n\tau\vartheta/4$. By Lemma 3.18 with $\delta = (nG)^{-4}$, we have

$$\mathbb{P}(\|T - A\|_{\mathrm{grp}*} > \lambda) < \frac{1}{(nG)^3}.$$

By Proposition 3.16, on the event $\{\|T - A\|_{\mathrm{grp}_*} \leqslant \lambda\}$, we have

$$\max\big\{\sin\angle(v,\hat{v}), \sin\angle(u,\hat{u})\big\} \leqslant \frac{32\lambda(C_1k)^{1/2}}{n^{1/2}\tau\vartheta},$$

as desired. □

3.8.2 Proof of Theorem 3.2

Proof. We will use two different constructions to derive separate lower bounds of order $\sqrt{Bs\log(G/s)/(n\tau\vartheta^2)}$ and $\sqrt{Bk/(n\tau\vartheta^2)}$ respectively. Without loss of generality, we may assume that $z < n/2$.

For the first bound, let $s_0 = s-1$, $G_0 = G-1$. By the Gilbert–Varshamov lemma as stated in Massart (2007, Lemma 4.10) (applied with $\alpha = 3/4$ and $\beta = 1/3$), we can construct a set $\mathcal{U}_0$ of s_0-sparse vectors in $\{0,1\}^{G_0}$, with cardinality at least $(G_0/s_0)^{s_0/5}$, such that the pairwise Hamming distance between any pair of vectors in $\mathcal{U}_0$ is at least $s_0/2$. Let $\epsilon \in (0,1)$ to be chosen later, we can define a set

$$\mathcal{U} = \left\{\begin{pmatrix}\sqrt{1-\epsilon^2}\\ s_0^{-1/2}\epsilon u_0\end{pmatrix} : u_0 \in \mathcal{U}_0\right\} \subseteq \mathbb{S}^{G-1}.$$

We remark that for any pair of distinct $u, u' \in \mathcal{U}$, we have by construction that $\epsilon/\sqrt{2} \leqslant \|u' - u\|_2 \leqslant \epsilon$. We then define a map $\psi : \mathbb{R}^G \to \mathbb{R}^p$ such that for any $u \in \mathcal{U}$ and $j \in \mathcal{J}_g$, we have $\psi(u)_j = u_g p_g^{-1/2}$. Finally, let $\mathcal{V} = \{\psi(u) : u \in \mathcal{U}\}$. We note that $\|\psi(u') - \psi(u)\|_2 = \|u' - u\|_2$. Therefore, for distinct $v, v' \in \mathcal{V}$, we have

$$L(v', v) = \sqrt{1 - (v^\top v')^2} = \frac{\|v' - v\|_2}{\sqrt{2}} \geqslant \frac{\epsilon}{2}. \tag{3.18}$$

Now, for each $v \in \mathcal{V}$, we define a distribution $P_v \in \mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$, such that the pre-change mean is $-\vartheta v$ and the post-change mean is 0 (we check that P_v indeed satisfies the conditions of $\mathcal{P}^{(1)}_{n,p}(s, k, \tau, \vartheta, B, (\mathcal{J}_g)_{g\in[G]})$). Then for any distinct $v, v' \in \mathcal{V}$, we have

$$\begin{aligned} D(P_v \| P_{v'}) = zD(N_p(-v\vartheta, B)\|N_p(-v'\vartheta, B)) &\leqslant \frac{z\vartheta^2}{2B}\|v - v'\|_2^2 \\ &\leqslant \frac{z\vartheta^2\epsilon^2}{2B}. \end{aligned} \tag{3.19}$$

By (3.18) and (3.19), we can apply Fano's lemma (Yu, 1997, Lemma 3) to obtain that

$$\begin{aligned} \inf_{\tilde{v}} \sup_{P\in\mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{E}_P L(\tilde{v}(X), v(P)) &\geqslant \inf_{\tilde{v}} \sup_{v\in\mathcal{V}} \mathbb{E}_{P_v} L(\tilde{v}(X), v) \\ &\geqslant \frac{\epsilon}{4}\left\{1 - \frac{z\vartheta^2\epsilon^2/(2B) + \log 2}{(s_0/5)\log(G_0/s_0)}\right\}. \end{aligned}$$

By the condition $(s-1)\log(G/s) \geqslant 20$, we have $(s_0/5)\log(G_0/s_0) \geqslant 2\log 2$. Moreover, the choice of

$$\epsilon = \sqrt{\frac{Bs_0\log(G_0/s_0)}{10z\vartheta^2}}$$

ensures that $(s_0/5)\log(G_0/s_0) \geqslant 2z\vartheta^2\epsilon^2/B$. Therefore,

$$\inf_{\tilde{v}} \sup_{P\in\mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})} \mathbb{E}_P L(\tilde{v}(X), v(P)) \geqslant \frac{\epsilon}{16} \geqslant \frac{1}{72}\sqrt{\frac{Bs\log(G/s)}{z\vartheta^2}}. \tag{3.20}$$

For the second lower bound, let $g_1, \ldots, g_s$ be the indices of the s groups with largest cardinalities. By the given condition of the Theorem, we have that $\tilde{k} = \sum_{r=1}^s p_{g_r} = \sum_{r=1}^s p_{(G-r+1)} \geqslant k/2$. Let $S = \cup_{r=1}^s \mathcal{J}_{g_r}$, so $|S| = \tilde{k}$. By Massart (2007, Lemma 4.7), we can construct a subset $\mathcal{V}_0$ of $\{-1, 1\}^{\tilde{k}_0}$ of cardinality at least $e^{\tilde{k}/8}$, such that any two points in the set are separated in Hamming distance by at least $\tilde{k}/4$. Construct

$$\mathcal{V} = \left\{v : v_S = \begin{pmatrix} \sqrt{1-\epsilon^2} \\ \tilde{k}_0^{-1/2}\epsilon v_0 \end{pmatrix} \text{ for some } v_0 \in \mathcal{V}_0 \text{ and } v_{S^c} = 0\right\}.$$

Therefore, for distinct $v, v' \in \mathcal{V}$, we have $\epsilon \leqslant \|v' - v\|_2 \leqslant 2\epsilon$,then,

$$L(v', v) = \sqrt{1 - (v^\top v')^2} = \frac{\|v' - v\|_2}{\sqrt{2}} \geqslant \frac{\epsilon}{\sqrt{2}}.$$

Following the same derivation as in (3.19), we have that

$$\begin{aligned}D(P_v\|P_{v'}) &= zD(N_p(-v\vartheta,\Sigma)\|N_p(-v'\vartheta,\Sigma))\\ &\leqslant \frac{z\vartheta^2}{2B}\|v-v'\|_2^2 \leqslant \frac{2z\vartheta^2\epsilon^2}{B}.\end{aligned}$$

Again, we can use Fano's lemma (Yu, 1997, Lemma 3) to obtain that

$$\begin{aligned}\inf_{\tilde{v}}\sup_{v\in\mathcal{V}}\mathbb{E}_{P_v}L(\tilde{v}(X),v) &\geqslant \frac{\epsilon}{\sqrt{2}}\left\{1-\frac{2z\vartheta^2\epsilon^2/B+\log 2}{\tilde{k}/8}\right\}\\ &\geqslant \frac{\epsilon}{\sqrt{2}}\left\{1-\frac{2z\vartheta^2\epsilon^2/B+\log 2}{k/16}\right\}.\end{aligned}$$

Now, choose $\epsilon=(kB)^{1/2}z^{-1/2}\vartheta^{-1}/4\sqrt{6}$. Since $k\geqslant 20$, we have $k/16\geqslant 9\log(2)/5$, so that

$$\begin{aligned}\inf_{\tilde{v}}\sup_{P\in\mathcal{P}^{(1)}_{n,p}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})}\mathbb{E}_P L(\tilde{v}(X),v(P)) &\geqslant \inf_{\tilde{v}}\sup_{v\in\mathcal{V}}\mathbb{E}_{P_v}L(\tilde{v}(X),v)\\ &\geqslant \frac{\epsilon}{9\sqrt{2}} \geqslant \frac{1}{72\sqrt{3}}\sqrt{\frac{kB}{z\theta^2}}. \end{aligned} \tag{3.21}$$

The desired result follows by combining (3.20) with (3.21), and noting that $z\geqslant n\tau$. □

3.8.3 Proof of Theorem 3.3

Proof. Recall the definition of $X^{(2)}$ and let $T^{(2)}=\mathcal{T}(X^{(2)})$. Define similarly $\mu^{(2)}=(\mu_1^{(2)},\ldots,\mu_{n_1}^{(2)})\in\mathbb{R}^{p\times n_1}$ and a random $W^{(2)}=(W_1^{(2)},\ldots,W_{n_1}^{(2)})$ taking values in $\mathbb{R}^{p\times n_1}$ by $\mu_t^{(2)}=\mu_{2t}$ and $W_t^{(2)}=W_{2t}$. Now, let $A^{(2)}=\mathcal{T}(\mu^{(2)})$ and $E^{(2)}=\mathcal{T}(W^{(2)})$. We also write $\bar{X}=(\hat{v}^{(1)})^\top X^{(2)}$, $\bar{\mu}=(\hat{v}^{(1)})^\top\mu^{(2)}$, $\bar{W}=(\hat{v}^{(1)})^\top W^{(2)}$, $\bar{A}=(\hat{v}^{(1)})^\top A^{(2)}$, $\bar{E}=(\hat{v}^{(1)})^\top E^{(2)}$ and $\bar{T}=(\hat{v}^{(1)})^\top T^{(2)}$ for the one-dimensional projected images. Note that by linearity, we have $\bar{T}=\mathcal{T}(\bar{X})$, $\bar{A}=\mathcal{T}(\bar{\mu})$ and $\bar{E}=\mathcal{T}(\bar{W})$.

Now, conditional on $\hat{v}^{(1)}$, the random variables $\bar{X}_1,\ldots,\bar{X}_{n_1}$ are independent with

$$\bar{X}_t \mid \hat{v}^{(1)} \sim N(\bar{\mu}_t,\sigma^2)$$

and the row vector $\bar{\mu}$ undergoes a single change at $z^{(2)}=z/2$ with magnitude of change

$$\bar{\theta}=\bar{\mu}_{z^{(2)}+1}-\bar{\mu}_{z^{(2)}}=\hat{v}^{(1)\top}\theta.$$

Finally, let $\hat{z}^{(2)} \in \arg\max_{1\leqslant t\leqslant n_1-1} |\bar{T}_t|$, so the first component of the output of the algorithm is $\hat{z} = 2\hat{z}^{(2)}$. Consider the set

$$\Upsilon = \{u \in \mathbb{S}^{p-1} : \sin\angle(u, v) \leqslant 1/2\}.$$

By Condition (3.9) and Theorem 3.1, we have that

$$\mathbb{P}(\hat{v}^{(1)} \in \Upsilon) \geqslant 1 - \frac{1}{(n_1 G)^3}. \tag{3.22}$$

Moreover, on the event $\{\hat{v}^{(1)} \in \Upsilon\}$, we have that $|\bar{\theta}| \geqslant \sqrt{3}\vartheta/2$. Noting that we have $\bar{E}_t \mid \hat{v}^{(1)} \sim N(0, \hat{v}^{(1)\top}\Sigma\hat{v}^{(1)})$, we have by Wang and Samworth (2018, Lemma 4) for any $\lambda_1 \geqslant \sqrt{B}$ that

$$\mathbb{P}(\|\bar{E}\|_\infty \geqslant \lambda_1) \leqslant \sqrt{\frac{2}{\pi}}\lceil \log n_1 \rceil \biggl(\frac{\lambda_1}{\sqrt{B}} + 2\biggr) e^{-\lambda_1^2/B} \leqslant 3\lambda_1 e^{-\lambda_1^2/B} \log n. \tag{3.23}$$

Define $\Omega_0 := \{\hat{v}_1 \in \Upsilon, \|\bar{E}\|_\infty \leqslant \lambda_1\}$. From (3.22) and (3.23), we have $\mathbb{P}(\Omega_0) \geqslant 1 - n_1^{-3} - 3\lambda_1 e^{-\lambda_1^2/B} \log n$.

Notice that the procedure produces the same output if we replace $\hat{v}^{(1)}$ by $-\hat{v}^{(1)}$, hence we may assume without loss of generality that $\bar{\theta} \geqslant 0$, which implies that $\bar{A}_t \geqslant 0$ for all $t \in [n_1 - 1]$. Condition (3.9) implies that

$$\sqrt{n}\tau\vartheta \geqslant C\lambda_1, \tag{3.24}$$

for sufficient large C. Therefore, by Lemma 3.20 and (3.24), if we choose $C \geqslant 8/\sqrt{3}$, then for t satisfying $|z^{(2)} - t| \geqslant n_1\tau/2$, we have

$$A_{z^{(2)}} = \sqrt{\frac{z^{(2)}(n_1 - z^{(2)})}{n_1}}\bar{\theta} \geqslant \sqrt{\frac{n_1\tau}{2}}\bar{\theta} \geqslant \frac{\sqrt{3}}{4}\sqrt{n\tau}\vartheta \geqslant 2\lambda_1.$$

In particular, we must have on Ω_0 that $T_{\hat{z}^{(2)}} \geqslant T_{z^{(2)}} \geqslant A_{z^{(2)}} - \lambda_1 \geqslant -A_t + \lambda_1 \geqslant -T_t$ for any $t \in [n-1]$. Hence, $\arg\max_{t\in[n-1]} |\bar{T}_t| = \arg\max_{t\in[n-1]} \bar{T}_t$.

Since $\bar{T} = \bar{A} + \bar{E}$ and $(\bar{A}_t)_t$ and $(\bar{T}_t)_t$ are respectively maximised at $t = z^{(2)}$ and $t = \hat{z}^{(2)}$. We have on the event Ω_0 that

$$\bar{A}_{z^{(2)}} - \bar{A}_{\hat{z}^{(2)}} = (\bar{A}_{z^{(2)}} - \bar{T}_{\hat{z}^{(2)}}) + (\bar{T}_{z^{(2)}} - \bar{T}_{\hat{z}^{(2)}}) + (\bar{T}_{\hat{z}^{(2)}} - \bar{A}_{\hat{z}^{(2)}}) \leqslant \bar{E}_{\hat{z}^{(2)}} - \bar{E}_{z^{(2)}}. \tag{3.25}$$

Note that on Ω_0, the right-hand side of (3.25) is bounded by $2\lambda_1$. Hence, applying Lemma 3.20 to the left-hand side of (3.25), and using the unimodality of $\bar{A}$, if $C \geqslant 24$, on the event Ω_0, we have that

$$\frac{|\hat{z}^{(2)} - z^{(2)}|}{n_1\tau} \leqslant \frac{3\sqrt{6}\lambda_1}{\bar{\theta}\sqrt{n_1\tau}} \leqslant \frac{12\lambda_1}{\vartheta\sqrt{n\tau}} \leqslant \frac{1}{2}.$$

By Lemma 3.19, there exists an event Ω_1 with probability at least $1 - e^{-\lambda_1^2/(2B)}\log n$ on which

$$|\bar{E}_{z^{(2)}} - \bar{E}_{\hat{z}^{(2)}}| \leqslant 4\lambda_1\sqrt{\frac{|z^{(2)} - \hat{z}^{(2)}|}{n\tau}} + 16\lambda_1\frac{|z^{(2)} - \hat{z}^{(2)}|}{n\tau}. \tag{3.26}$$

Substituting the improved bound of (3.26) into the right-hand side of (3.25), and again applying Lemma 3.20 to the left-hand side of (3.25), we have on $\Omega_0 \cap \Omega_1$ that

$$\frac{\vartheta}{3}\frac{|z^{(2)} - \hat{z}^{(2)}|}{\sqrt{n\tau}} \leqslant 4\lambda_1\sqrt{\frac{|z^{(2)} - \hat{z}^{(2)}|}{n\tau}} + 16\lambda_1\frac{|z^{(2)} - \hat{z}^{(2)}|}{n\tau}.$$

When $C \geqslant 96$, from (3.9), we have $16\lambda_1\frac{|z^{(2)}-\hat{z}^{(2)}|}{n\tau} \leqslant \frac{\vartheta}{6}\frac{|z^{(2)}-\hat{z}^{(2)}|}{\sqrt{n\tau}}$. Consequently, on $\Omega_0 \cap \Omega_1$, we have

$$|\hat{z} - z| \leqslant \frac{C'\lambda_1^2}{\vartheta^2},$$

as desired. Finally, we compute that the desired event occurs with probability

$$\mathbb{P}(\Omega_0 \cap \Omega_1) \geqslant 1 - \frac{1}{n_1^3} - (3\lambda_1 + 1)e^{-\lambda_1^2/(2B)}\log n.$$

as desired. □

3.8.4 Proof of Corollary 3.4

Proof. Define $A := \mathbb{E}(T)$ and $E := T - A$. Under null hypothesis where there is no change in the segment, by Lemma 3.18, we have that $\mathbb{P}(\|T\|_{\mathrm{grp}^*} \geqslant \lambda) = \mathbb{P}(\|E\|_{\mathrm{grp}^*} \geqslant \lambda) < 1/(nG)$.

Under the alternative, we have:

$$\|T\|_{\mathrm{grp}^*} = \|A + E\|_{\mathrm{grp}^*} \geqslant \|A\|_{\mathrm{grp}^*} - \|E\|_{\mathrm{grp}^*}.$$

By (3.17), we have

$$\|A\|_{\mathrm{grp}^*} = \|\theta\gamma^\top\|_{\mathrm{grp}^*} = \|\gamma\|_\infty \max_{g\in[G]} p_g^{-1/2}\|\theta_{\mathcal{J}_g}\|_2 \geqslant \frac{\|\gamma\|_\infty\|\theta\|_2}{\sqrt{k}}.$$

Also, by definition of γ, we have that $\|\gamma\|_\infty = \sqrt{\frac{z(n-z)}{n}} \geqslant \sqrt{n\tau/2}$. Therefore, for $\|\theta\|_2 \geqslant \frac{2\sqrt{2k}\lambda}{\sqrt{n\tau}}$, combining with Lemma 3.17, we have that with probability at least $1 - 1/(nG)$ that $\|T\|_{\mathrm{grp}^*} \geqslant 2\lambda - \lambda = \lambda$. □

3.8.5 Proof of Theorem 3.5

Proof. Let $\{z_1, \ldots, z_\nu\}$ be the set of true change points, such that $0 =: z_0 < z_1 < \cdots < z_\nu < n =: z_{\nu+1}$. For each $i \in [\nu]$, define intervals

$$\mathcal{I}_i^L = \big(z_i - n\tau/3, z_i - n\tau/6\big) \quad \text{and} \quad \mathcal{I}_i^R = \big(z_i + n\tau/6, z_i + n\tau/3\big).$$

These intervals contain at least one integer for $n\tau \geqslant 6$. For simplicity of exposition, we have ignored various rounding issues in this proof. Now, define the following event:

$$\Omega_0 := \{\forall i \in [\nu],\, \exists m \in [M],\, \text{ s.t. } (s_m, e_m) \in \mathcal{I}_i^L \times \mathcal{I}_i^R\}.$$

Then, we have

$$\mathbb{P}(\Omega_0^{\mathrm{c}}) \leqslant \sum_{i=1}^{\nu} \prod_{m=1}^{M} \Big(1 - \mathbb{P}((s_m, e_m) \in \mathcal{I}_i^L \times \mathcal{I}_i^R)\Big) \leqslant \nu \Big(1 - \frac{\tau^2}{36}\Big)^M \leqslant \nu e^{-\tau^2 M/36}.$$

On Ω_0, for each change point z_i, we can find an interval $(s_m, e_m]$ which only captures one change-point, which is at least $n\tau/6$ away from the endpoints s_m and e_m of the interval.

We write $X^{(s,e]}$ for the submatrix of X obtained by extracting columns indexed in $(s, e]$. Let $T^{(s,e]} := \mathcal{T}(X^{(s,e]})$, $A^{(s,e]} := \mathbb{E}T^{(s,e]}$ and $E^{(s,e]} := T^{(s,e]} - A^{(s,e]}$. Set

$$\Omega_1 := \Big\{\max_{1 \leqslant s < e \leqslant n} \|E^{(s,e]}\|_{\mathrm{grp}*} < \lambda\Big\}.$$

By Lemma 3.18 and a union bound, we have that

$$\mathbb{P}(\Omega_1^{\mathrm{c}}) \leqslant n^2 \frac{(n-1)G}{(nG)^4} \leqslant \frac{1}{nG^3}.$$

Now, for any interval $(s, e]$, we write $\hat{z}^{(s,e]}$ to be the change-point estimate of Algorithm 8 applied to data $X^{(s,e]}$. We define the following set: $O := \{(s, e) : 0 \leqslant s < e \leqslant n,\ z_{i-1} \leqslant s < z_i < e < z_{i+1}$ for some $i \in [\nu]\}$ and $\min\{z_i - s, e - z_i\} \geqslant n\tau/10\}$ to be the set of intervals $(s, e]$ that captures exactly one true change-point, which is at least $n\tau/10$ away from the boundaries. We then define the event

$$\Omega_2 := \left\{ |\hat{z}^{(s,e]} + s - z_i| \leqslant \frac{C'B\log n}{\vartheta^2} \text{ for all } (s,e] \in O \right\}.$$

For a sufficiently large C and C', by Condition (3.3) and Theorem 3.3 applied with $\lambda_1 = \sqrt{16B\log(n\tau B)}$, together with union bound, we have that

$$\mathbb{P}(\Omega_2^{\mathrm{c}}) \leqslant \frac{7}{n\tau^3}.$$

We will henceforth work on $\Omega_0 \cap \Omega_1 \cap \Omega_2$.

For any interval $(s, e] \subseteq (0, n]$, we define $\mathcal{Z}^{(s,e]} := \{z_i : i \in [\nu] \text{ and } z_i \in (s,e]\}$ and the following subsets of $\mathcal{Z}^{(s,e]}$:

$$\mathcal{Z}^{(s,e]}_{\mathrm{good}} := \{z \in \mathcal{Z}^{(s,e]} : \min\{z - s, e - z\} \geqslant n\tau/3\},$$
$$\mathcal{Z}^{(s,e]}_{\mathrm{bad}} := \left\{ z \in \mathcal{Z}^{(s,e]} : \min\{z - s, e - z\} \leqslant \frac{C'B\log n}{\vartheta^2} \right\},$$

where C' is chosen to be the same constant as in the definition of Ω_2. We note that $\mathcal{Z}^{(s,e]}_{\mathrm{good}}$ and $\mathcal{Z}^{(s,e]}_{\mathrm{bad}}$ respectively contain change-points within $(s, e]$ that are well-separated from the boundary and close to the boundary. We will informally refer to these change-points as "good" and "bad" change-points in $(s, e]$. On Ω_0, for every $i \in [\nu]$, we can associate it with an $m_i \in [M]$ such that $s_{m_i} \in \mathcal{I}_i^L$ and $e_{m_i} \in \mathcal{I}_i^R$. We claim that

$$\{m_i : z_i \in \mathcal{Z}^{(s,e]}_{\mathrm{good}}\} \subseteq \mathcal{R}_{s,e}. \tag{3.27}$$

To see this, we first note that from the definition of $\mathcal{I}_i^L$ and $\mathcal{I}_i^R$, and the condition $\min\{z_i - s, e - z_i\} \geqslant n\tau/3$ that for every i with $z_i \in \mathcal{Z}^{(s,e]}_{\mathrm{good}}$ we have $(s_{m_i}, e_{m_i}] \subseteq (s, e]$. On Ω_1, by Condition (3.11) with a sufficiently large choice of $C > 0$ and the proof of Corollary 3.4 we have

$$\|T^{(s_{m_i}+\beta, e_{m_i}-\beta]}\|_{\mathrm{grp}*} \geqslant \lambda.$$

Hence $m_i \in \mathcal{R}_{s,e}$, establishing the claim. On the other hand, under Condition (3.11) for sufficiently large C, we have $\frac{C'B\log n}{\vartheta^2} < n\tau/10 = \beta$. Hence on Ω_1, for any $(s_0, e_0] \subseteq (s, e]$ containing only "bad" change-points, i.e. $(s_0, e_0] \cap \mathcal{Z}^{(s,e]} \subseteq \mathcal{Z}^{(s,e]}_{\mathrm{bad}}$, we get:

$$\|T^{(s_0+\beta, e_0-\beta]}\|_{\mathrm{grp}*} < \lambda,$$

as there are no change points within the interval $(s_0+\beta, e_0-\beta]$. Thus,

$$\left\{m \in \mathcal{M}_{s,e} : (s_m, e_m] \cap \mathcal{Z}^{(s,e]} \subseteq \mathcal{Z}^{(s,e]}_{\text{bad}}\right\} \cap \mathcal{R}_{s,e} = \emptyset \tag{3.28}$$

Given a set $\hat{Z}$ of estimated change-points, we can partition $(0,n]$ into $|\hat{Z}|+1$ segments. We call these the segments induced by $\hat{Z}$. We now prove by induction that throughout the recursion of NOT, the following statement holds:

$$\text{For any } (s,e] \text{ induced by } \hat{Z},\ \mathcal{Z}^{(s,e]} = \mathcal{Z}^{(s,e]}_{\text{good}} \cup \mathcal{Z}^{(s,e]}_{\text{bad}}. \tag{P}$$

For the base case, at the beginning of the algorithm, we have $\hat{Z} = \emptyset$, so the only induced segment by $\hat{Z}$ is $(0,n]$. The statement (P) is true since the cloest change-point from the boundary is at least $n\tau$ away. Now assuming that (P) is true at some stage of the recursion when $\hat{Z}$ is the set of estimated change-points so far, we need to show that (P) still holds when a new change-point is estimated by NOT. This new change-point must be identified from running NOT on some $(s,e]$ where $(s,e]$ is one of the induced segments by $\hat{Z}$. From the inductive hypothesis, we know that $\mathcal{Z}^{(s,e]} = \mathcal{Z}^{(s,e]}_{\text{good}} \cup \mathcal{Z}^{(s,e]}_{\text{bad}}$. We note that $\mathcal{Z}^{(s,e]}_{\text{good}}$ is necessarily nonempty for otherwise by (3.28) we have $\mathcal{M}_{s,e} \cap \mathcal{R}_{s,e} = \emptyset$ and hence $\mathcal{R}_{s,e} = \emptyset$, so no new change-point will be identified in $(s,e]$. Thus, there exists some i' with $z_{i'} \in \mathcal{Z}^{(s,e]}_{\text{good}}$ and by (3.27), $m_{i'} \in \mathcal{R}_{s,e}$ and hence $e_{m^*} - s_{m^*} \leqslant e_{m_{i'}} - s_{m_{i'}} \leqslant n\tau/3$. In particular, we have that $(s_{m^*}, e_{m^*}]$ must capture exactly one change-point (it has to capture at least one change-point by (3.28) and cannot capture more than one since two consecutive change-points are spaced at least $n\tau$ away), say z_{i^*}. On the event Ω_2, we know that the change-point output $\hat{z}$ of Algorithm 8 on $X^{(s_{m^*}, e_{m^*}]}$ satisfies

$$|\hat{z} + s_{m^*} - z_{i^*}| \leqslant \frac{C' B \log n}{\vartheta^2}. \tag{3.29}$$

We now check that the two new segments induced by $\hat{Z} \cup \{\hat{z} + s_{m^*}\}$ still satisfy (P). For this, it suffices to check that z_{i^*-1}, z_{i^*} and z_{i^*+1} are either within $\frac{C'B\log n}{\vartheta^2}$ of $\hat{z} + s_{m^*}$ or at least $n\tau/3$ away from it. This can be seen by combining (3.29) with the fact that $\min\{z_{i^*} - z_{i^*-1}, z_{i^*+1} - z_{i^*}\} \geqslant n\tau$. This completes the induction.

We remark that as a side product of the above inductive argument, we have shown that if $(s,e] \cap \mathcal{Z}^{(s,e]}_{\text{good}} \neq \emptyset$, then $\mathcal{R}_{s,e}$ is non-empty and NOT will estimate a new change-point. Hence, at the end of the recursion, we must have that all segments induced by

$\hat{Z}$ contains no change-point at least $n\tau/3$ away from the boundaries. In other words, all change-points $z_1, \ldots, z_\nu$ must be at most $n\tau/10$ away from the endpoints of one of the induced segments. This, together with the fact that consecutive change-points (including z_0 and z_{n+1}) are spaced at least $n\tau$ away, means that there must be exactly ν estimated change-points in $\hat{Z}$ at the end of the algorithm. Let $\hat{z}_1 < \hat{z}_2 < \cdots < \hat{z}_\nu$ be elements of $\hat{Z}$ arranged in an increasing order. Then, since all change-points are "bad" at the end of the NOT recursion, we must have

$$\max_{i \in [\nu]} |\hat{z}_i - z_i| \leqslant \frac{C' B \log n}{\vartheta^2}$$

as desired. □

3.9 Ancillary results

We collect in this section all ancillary propositions and lemmas used in Chapter 3. For all results in this section, we assume that we are given a grouping $(\mathcal{J}_g)_{g \in [G]}$ of $[p]$ and the associated group norm $\|\cdot\|_{\mathrm{grp}}$.

Proposition 3.10. *Fix $n \in \mathbb{N}$. Let $P_{z,\mu_{\mathrm{L}},\mu_{\mathrm{R}}}$ denote the joint distribution of $(X_i)_{i \in [n]}$ such that $X_i \sim N(\mu_i, \sigma^2)$ are independent random variables with $\mu_i = \mu_{\mathrm{L}} \mathbb{1}_{\{i \leqslant z\}} + \mu_{\mathrm{R}} \mathbb{1}_{\{i > z\}}$. Then*

$$\inf_{\hat{z}} \sup_{(z,\mu_{\mathrm{L}},\mu_{\mathrm{R}}) \in [n-1] \times \mathbb{R}^2} \mathbb{E}_{P_{z,\mu_{\mathrm{L}},\mu_{\mathrm{R}}}} |\hat{z} - z| (\mu_{\mathrm{L}} - \mu_{\mathrm{R}})^2 \geqslant c\sigma^2 \log \log n.$$

Proof. Suppose $n = 2^L$ for some $L \in \mathbb{N}$. For $\ell \in [L]$, we define $\boldsymbol{\mu}^{(\ell)} \in \mathbb{R}^{2n}$ to be the vector whose last 2^ℓ entries are equal to $\sqrt{\sigma^2 2^{-\ell} \log \log_2(2n)/60}$ and the remaining entries are 0. Gao et al. (2020, Theroem 2.2 and the argument immediately above its statement) shows that for some universal constant $c_1 > 0$, we have

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\ell \in [L]} \mathbb{E}_{\boldsymbol{\mu}^{(\ell)}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(\ell)}\|_2^2 \geqslant c_1 \sigma^2 \log \log(16n). \tag{3.30}$$

Let $c > 0$ be a constant to be chosen later. We assume that the conclusion of the proposition does not hold, which means that there exists an estimator $\hat{z}$ such that for all

$z \in [n-1]$ and $\mu_{\mathrm{L}}, \mu_{\mathrm{R}} \in \mathbb{R}$, we have

$$\mathbb{E}_{P_{z,\mu_{\mathrm{L}},\mu_{\mathrm{R}}}} |\hat{z} - z| < \frac{c\sigma^2 \log\log n}{(\mu_{\mathrm{L}} - \mu_{\mathrm{R}})^2}. \tag{3.31}$$

Let $(Z_i)_{i\in[2n]}$ be a sequence of $2n$ independent random variables such that $Z_i \sim N(\mu_{\mathrm{L}}\mathbb{1}_{\{i\leqslant 2z\}} + \mu_{\mathrm{R}}\mathbb{1}_{\{i>2z\}}, \sigma^2)$. We can apply the estimator $\hat{z}$ on data $\mathcal{Z}_{\mathrm{odd}} := (Z_1, Z_3, \ldots, Z_{2n-1})$ of length n to obtain a changepoint location estimate $\hat{z}(\mathcal{Z}_{\mathrm{odd}})$, which for notational simplicity, we will denote also as $\hat{z}$ henceforth. Now, define

$$\hat{\mu}_{\mathrm{L}} := \frac{1}{\hat{z}} \sum_{i=1}^{\hat{z}} Z_{2i} \quad \text{and} \quad \hat{\mu}_{\mathrm{R}} := \frac{1}{n - \hat{z}} \sum_{i=\hat{z}+1}^{n} Z_{2i}.$$

Then the vector $\hat{\boldsymbol{\mu}} := (\hat{\mu}_{\mathrm{L}}\mathbb{1}_{\{i\leqslant 2\hat{z}\}} + \hat{\mu}_{\mathrm{R}}\mathbb{1}_{\{i>2\hat{z}\}})_{i\in[2n]}$ is an estimator of $\boldsymbol{\mu} := (\mathbb{E}Z_i)_{i\in[2n]}$. Without loss of generality, we may assume that $\hat{z} \geqslant z$; the opposite case can be handled symmetrically. This means that

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 = 2z(\hat{\mu}_{\mathrm{L}} - \mu_{\mathrm{L}})^2 + 2(\hat{z} - z)(\hat{\mu}_{\mathrm{L}} - \mu_{\mathrm{R}})^2 + 2(n - \hat{z})(\hat{\mu}_{\mathrm{R}} - \mu_{\mathrm{R}})^2 \tag{3.32}$$

Using independence between $\hat{z}$ and $(Z_{2i})_{i\in[n]}$, we have $\hat{\mu}_{\mathrm{L}} \mid \mathcal{Z}_{\mathrm{odd}} \sim N\bigl(\frac{z}{\hat{z}}\mu_{\mathrm{L}} + \frac{\hat{z}-z}{\hat{z}}\mu_{\mathrm{R}}, \sigma^2/\hat{z}\bigr)$ and $\hat{\mu}_{\mathrm{R}} \mid \mathcal{Z}_{\mathrm{odd}} \sim N\bigl(\mu_{\mathrm{R}}, \sigma^2/(n-\hat{z})\bigr)$. Hence, from (3.32), we have

$$\begin{aligned}\mathbb{E}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 \mid \mathcal{Z}_{\mathrm{odd}}) &= 4\sigma^2 + (\mu_{\mathrm{L}} - \mu_{\mathrm{R}})^2 \biggl\{\frac{2z(\hat{z} - z)^2}{\hat{z}^2} + \frac{2z^2(\hat{z} - z)}{\hat{z}^2}\biggr\} \\ &\leqslant 4\sigma^2 + 4(\mu_{\mathrm{L}} - \mu_{\mathrm{R}})^2(\hat{z} - z).\end{aligned}$$

Then, since $\mathbb{E}_{P_{z,\mu_{\mathrm{L}},\mu_{\mathrm{R}}}} |\hat{z} - z| < \frac{c\sigma^2 \log\log n}{(\mu_{\mathrm{L}} - \mu_{\mathrm{R}})^2}$, we have

$$\begin{aligned}\mathbb{E}_{P_{z,\mu_{\mathrm{L}},\mu_{\mathrm{R}}}}(\|\hat{\mu} - \mu\|_2^2) &\leqslant 4\sigma^2 + 4(\mu_L - \mu_R)^2 \mathbb{E}_{P_{z,\mu_{\mathrm{L}},\mu_{\mathrm{R}}}}(\hat{z} - z) \\ &< 4\sigma^2 + c\sigma^2 \log\log n.\end{aligned}$$

Now, choosing $c = c_1/2$, then for sufficiently large n, the above inequality contradicts (3.30), which means that (3.31) cannot hold, thus establishing the desired conclusion. □

Lemma 3.11. *The norm $\|\cdot\|_{\mathrm{grp}*}$ is a dual to $\|\cdot\|_{\mathrm{grp}}$ with respect to the inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^{p\times n}$.*

Proof. To prove the lemma, it suffices to show that $\|M\|_{\mathrm{grp}} = \sup_{\|R\|_{\mathrm{grp}*} \leqslant 1} \langle R, M\rangle$ for all $M \in \mathbb{R}^{p\times(n-1)}$. First, for any $M \in \mathbb{R}^{p\times(n-1)}$, let $M_{\mathcal{J}_g,t}$ be the tth column of $M_{\mathcal{J}_g}$. Define $\tilde{R} = \tilde{R}(M)$ such that

$$\tilde{R}_{\mathcal{J}_g,t} = \frac{p_g^{1/2} M_{\mathcal{J}_g,t}}{\|M_{\mathcal{J}_g,t}\|_2}.$$

By convention, we set $\tilde{R}_{\mathcal{J}_g,t} = 0$ if $\|M_{\mathcal{J}_g,t}\|_2 = 0$ Then,

$$\|\tilde{R}\|_{\mathrm{grp}*} = \max_{\substack{g\in[G],t\in[n-1],\\ \|M_{\mathcal{J}_g,t}\|_2\neq 0}} \frac{p_g^{-1/2}\|M_{\mathcal{J},t}\|_2}{\|M_{\mathcal{J}_g,t}\|_2} \leqslant \max_{\substack{g\in[G],t\in[n-1],\\ \|M_{\mathcal{J}_g,t}\|_2\neq 0}} p_g^{-1/2} p_g^{1/2} \frac{\|M_{\mathcal{J}_g,t}\|_2}{\|M_{\mathcal{J}_g,t}\|_2} = 1.$$

Hence,

$$\begin{aligned}\sup_{\|R\|_{\mathrm{grp}*}\leqslant 1} \langle R, M\rangle \geqslant \langle \tilde{R}, M\rangle &= \sum_{\substack{g\in[G],t\in[n-1],\\ \|M_{\mathcal{J}_g,t}\|_2\neq 0}} p_g^{1/2} \frac{\langle M_{\mathcal{J}_g,t}, M_{\mathcal{J}_g,t}\rangle}{\|M_{\mathcal{J}_g,t}\|_2} \\ &= \sum_{g=1}^{G}\sum_{t=1}^{n-1} p_g^{1/2}\|M_{\mathcal{J}_g,t}\|_2 = \|M\|_{\mathrm{grp}}.\end{aligned}$$

On the other hand, for any R such that $\|R\|_{\mathrm{grp}*} \leqslant 1$, we have $\|R_{\mathcal{J}_g,t}\|_2 \leqslant p_g^{1/2}$ for all g and t. Consequently, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\langle R, M\rangle = \sum_{g\in[G]}\sum_{t\in[n-1]} \langle R_{\mathcal{J}_g,t}, M_{\mathcal{J}_g,t}\rangle &\leqslant \sum_{g\in[G]}\sum_{t\in[n-1]} \|R_{\mathcal{J}_g,t}\|_2\|M_{\mathcal{J}_g,t}\|_2 \\ &\leqslant \sum_{g\in[G]}\sum_{t\in[n-1]} p_g^{1/2}\|M_{\mathcal{J}_g,t}\|_2 = \|M\|_{\mathrm{grp}},\end{aligned}$$

thus establishing the result. □

Proposition 3.12. *Let $\mathcal{S} = \{M \in \mathbb{R}^{p\times(n-1)} : \|M\|_{\mathrm{F}} \leqslant 1\}$. For $T \in \mathbb{R}^{p\times(n-1)}$, $\lambda > 0$, we have*

$$\operatorname*{arg\,max}_{M\in\mathcal{S}} \Big\{\langle T, M\rangle - \lambda\|M\|_{\mathrm{grp}}\Big\} = \frac{T - R^*}{\|T - R^*\|_{\mathrm{F}}},$$

where R^* satisfies $R^*_{\mathcal{J}_g,t} = T_{\mathcal{J}_g,t}\min\big\{\frac{\lambda p_g^{1/2}}{\|T_{\mathcal{J}_g,t}\|_{\mathrm{F}}}, 1\big\}$.

Proof. Define functions $h : \mathbb{R}^{p\times(n-1)} \times \mathbb{R}^{p\times(n-1)} \to \mathbb{R}$ and $f, g : \mathbb{R}^{p\times(n-1)} \to \mathbb{R}$ such that for $M, R \in \mathbb{R}^{p\times(n-1)}$, $h(M, R) = \langle T - \lambda R, M\rangle$ and $f(M) = \inf_{\|R\|_{\mathrm{grp}*}\leqslant 1} h(M, R)$ and

$g(R) = \sup_{M \in \mathcal{S}} h(M, R)$. By (3.10) and Lemma 3.11, we have that

$$\begin{aligned}\langle T, M\rangle - \lambda \|M\|_{\mathrm{grp}} &= \langle T, M\rangle - \lambda \sup_{\|R\|_{\mathrm{grp}*} \leqslant 1} \langle R, M\rangle \\ &= \inf_{\|R\|_{\mathrm{grp}*} \leqslant 1} \langle T - \lambda R, M\rangle = f(M).\end{aligned}$$

By the minimax equality theorem (Fan, 1953, Theorem 1), we obtain that

$$\sup_{M \in \mathcal{S}} f(M) = \sup_{M \in \mathcal{S}} \inf_{\|R\|_{\mathrm{grp}*} \leqslant 1} h(M, R) = \inf_{\|R\|_{\mathrm{grp}*} \leqslant 1} \sup_{M \in \mathcal{S}} h(M, R) = \inf_{\|R\|_{\mathrm{grp}*} \leqslant 1} g(R).$$

Observe that $g(R) = \|T - \lambda R\|_{\mathrm{F}}$. To find the $R^* \in \arg\min_{\|R\|_{\mathrm{grp}*} \leqslant 1} \|T - \lambda R\|_{\mathrm{F}}$, we consider the G groups individually. For each group g, and in the tth column, if $\|T_{\mathcal{J}_g,t}\|_2 \leqslant \lambda p_g^{1/2}$, then $R^*_{\mathcal{J}_g,t} = T_{\mathcal{J}_g,t}/\lambda$; and if $\|T_{\mathcal{J}_g,t}\|_2 > \lambda p_g^{1/2}$, then $R^*_{\mathcal{J}_g,t} = p_g^{1/2} T_{\mathcal{J}_g,t}/\|T_{\mathcal{J}_g,t}\|_2$. Since the minimiser of $g(R)$ is unique, we have that

$$\arg\max_{M \in \mathcal{S}} f(M) = \arg\max_{M \in \mathcal{S}} h(M, R^*) = \frac{T - \lambda R^*}{\|T - \lambda R^*\|_{\mathrm{F}}},$$

as desired. □

Lemma 3.13. *For any $A, B \in \mathbb{R}^{p \times n}$, we have $\langle A, B\rangle \leqslant \|A\|_{\mathrm{grp}} \|B\|_{\mathrm{grp}*}$.*

Proof. By Cauchy–Schwarz inequality, we have that

$$\begin{aligned}\langle A, B\rangle &= \sum_{g,t} \langle A_{\mathcal{J}_g,t}, B_{\mathcal{J}_g,t}\rangle \leqslant \sum_{g \in [G], t \in [n]} \|A_{\mathcal{J}_g,t}\|_{\mathrm{F}} \|B_{\mathcal{J}_g,t}\|_{\mathrm{F}} \\ &\leqslant \biggl(\sum_{g \in [G], t \in [n]} p_g^{1/2} \|A_{\mathcal{J}_g,t}\|_{\mathrm{F}}\biggr) \Bigl(\max_{g \in [G], t \in [n]} p_g^{-1/2} \|B_{\mathcal{J}_g,t}\|_{\mathrm{F}}\Bigr) = \|A\|_{\mathrm{grp}} \|B\|_{\mathrm{grp}*}.\end{aligned}$$

as desired. □

Lemma 3.14. *Let $p_g = |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j \in [p]} |\{g : j \in \mathcal{J}_g\}| \leqslant C_1$. Then, for any $M \in \mathbb{R}^{p \times n}$, we have $\|M\|_{\mathrm{grp}} \leqslant (C_1 n \sum_g p_g)^{1/2} \|M\|_{\mathrm{F}}$.*

Proof. Define m with $m_{\mathcal{J}_g,t} = \|M_{\mathcal{J}_g,t}\|_{\mathrm{F}}$. Then by applying the Cauchy–Schwarz inequality twice, we have

$$\begin{aligned}\|M\|_{\mathrm{grp}} &= \sum_{g\in[G]} p_g^{1/2}\sum_{t=1}^{n}\|M_{\mathcal{J}_g,t}\|_2 \leqslant \sum_{g\in[G]}(np_g)^{1/2}\|M_{\mathcal{J}_g}\|_{\mathrm{F}}\\ &\leqslant \sqrt{n}\biggl(\sum_{g\in[G]}p_g\biggr)^{1/2}\biggl(\sum_{g\in[G]}\|M_{\mathcal{J}_g}\|_{\mathrm{F}}^2\biggr)^{1/2} \leqslant \biggl(C_1 n\sum_{g\in[G]}p_g\biggr)^{1/2}\|M\|_{\mathrm{F}},\end{aligned}$$

as desired. □

The following proposition establishes a sine angle loss upper bound for the (computationally infeasible) optimiser of (3.4). We see that the risk bound has essentially the same form as that given in Theorem 3.1.

Proposition 3.15. *For a given grouping $(\mathcal{J}_g)_{g\in[G]}$, let $p_* = \min_{g\in[G]}|\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j\in[p]}|\{g : j\in\mathcal{J}_g\}| \leqslant C_1$. Let $X\sim P\in\mathcal{P}_{n,p}^{(1)}(s,k,\tau,\vartheta,B,(\mathcal{J}_g)_{g\in[G]})$ be a $p\times n$ data matrix, let θ be the vector of change and let $\hat{v}\in\arg\max_{\tilde{v}\in\mathbb{S}^{p-1},\|\phi(\tilde{v})\|_0\leqslant s}\|\tilde{v}^\top T\|_2$. Let $\lambda\geqslant B^{1/2}(1+\sqrt{4\log(nG)/p_*})$. Then, with probability at least $1-\frac{1}{nG}$ we have that*

$$\sin\angle(v,\hat{v}) \leqslant \frac{8\sqrt{2C_1}\lambda k^{1/2}}{n^{1/2}\tau\vartheta} \tag{3.33}$$

Proof. Let A,γ be defined as in the proof of Theorem 3.1. Let $u := \gamma/\|\gamma\|_2$ and $\hat{u} := T^\top\hat{v}/\|T^\top\hat{v}\|_2$. Then, by the basic inequality, we have that:

$$\langle\hat{v}\hat{u}^\top, T\rangle = \|T^\top\hat{v}\|_2 \geqslant \|T^\top v\|_2 \geqslant v^\top Tu = \langle vu^\top, T\rangle.$$

Combining with Wang and Samworth (2018, Lemma 2), we have:

$$\begin{aligned}\|vu^\top - \hat{v}\hat{u}^\top\|_{\mathrm{F}}^2 &= \frac{2}{\|\theta\|_2\|\gamma\|_2}(\langle A-T, vu^\top-\hat{v}\hat{u}^\top\rangle + \langle T, vu^\top-\hat{v}\hat{u}^\top\rangle)\\ &\leqslant \frac{2}{\|\theta\|_2\|\gamma\|_2}\langle A-T, vu^\top-\hat{v}\hat{u}^\top\rangle\\ &\leqslant \frac{2}{\|\theta\|_2\|\gamma\|_2}\|A-T\|_{\mathrm{grp}*}\|vu^\top-\hat{v}\hat{u}^\top\|_{\mathrm{grp}}\end{aligned}$$

Since $vu^\top - \hat{v}\hat{u}^\top$ has at most $2k$ rows with non-zero entries, By Lemmas 3.14 and 3.18, for the choice of λ in the proposition, we have with probability at least $1 - 1/(nG)$ that

$$\|uv^\top - \hat{v}\hat{u}^\top\|_{\mathrm{F}}^2 \leqslant \frac{2\sqrt{2}\lambda(C_1 nk)^{1/2}}{\|\theta\|_2\|\gamma\|_2}\|uv^\top - \hat{v}\hat{u}^\top\|_{\mathrm{F}}.$$

Consequently, by the same argument as in the proof of Proposition 3.16 we have

$$\sin\angle(v,\hat{v}) \leqslant \|vu^\top - \hat{v}\hat{u}^\top\|_{\mathrm{F}} \leqslant \frac{2\sqrt{2}\lambda(C_1 nk)^{1/2}}{\|\theta\|_2\|\gamma\|_2} \leqslant \frac{8\sqrt{2}\lambda(C_1 k)^{1/2}}{n^{1/2}\tau\vartheta},$$

as required. □

Proposition 3.16. *Let $p_g = |\mathcal{J}_g|$ and suppose further that there exists a universal constant $C_1 > 0$, such that $\max_{j\in[p]} |\{g : j \in \mathcal{J}_g\}| \leqslant C_1$. Let A be a rank one matrix with $A = \delta vu^\top$ for $\delta > 0$, $\|v\|_2 = \|u\|_2 = 1$ and $\sum_{g:v_{\mathcal{J}_g}\neq 0} p_g \leqslant k$. Suppose $T \in \mathbb{R}^{p\times(n-1)}$ satisfies $\|T - A\|_{\mathrm{grp}*} \leqslant \lambda$ for some $\lambda > 0$, and let $\mathcal{S} = \{M \in \mathbb{R}^{p\times(n-1)} : \|M\|_{\mathrm{F}} \leqslant 1\}$. Then, for any*

$$\hat{M} \in \operatorname*{arg\,max}_{M\in\mathcal{S}} \bigl\{\langle T, M\rangle - \lambda\|M\|_{\mathrm{grp}}\bigr\},$$

we have

$$\|vu^\top - \hat{M}\|_{\mathrm{F}} \leqslant \frac{4\lambda(C_1 nk)^{1/2}}{\delta},$$

and

$$\max\{\sin\angle(v,\hat{v}), \sin\angle(u,\hat{u})\} \leqslant \frac{8\lambda(C_1 nk)^{1/2}}{\delta}.$$

Proof. Define $\mathcal{G}_0 = \{g : v_{\mathcal{J}_g} \neq 0\}$. Since $vu^\top \in \mathcal{S}$, from the basic inequality, we have

$$\langle T, vu^\top\rangle - \lambda\|vu^\top\|_{\mathrm{grp}} \leqslant \langle T, \hat{M}\rangle - \lambda\|\hat{M}\|_{\mathrm{grp}}. \tag{3.34}$$

When $\|A - T\|_{\mathrm{grp}*} \leqslant \lambda$, or equivalently, $p_g^{-1/2}\|A_{\mathcal{J}_g,t} - T_{\mathcal{J}_g,t}\|_2 \leqslant \lambda$ for all $g \in [G]$ and $t \in [n-1]$, we have by Wang and Samworth (2018, Lemma 2) and (3.34) that

$$\begin{aligned}\|vu^\top - \hat{M}\|_{\mathrm{F}}^2 &\leqslant \frac{2}{\delta}\langle A, vu^\top - \hat{M}\rangle \leqslant \frac{2}{\delta}\bigl(\langle T, vu^\top - \hat{M}\rangle + \langle A - T, vu^\top - \hat{M}\rangle\bigr)\\ &\leqslant \frac{2\lambda}{\delta}\bigl(\|vu^\top\|_{\mathrm{grp}} - \|\hat{M}\|_{\mathrm{grp}} + \|vu^\top - \hat{M}\|_{\mathrm{grp}}\bigr)\\ &= \frac{4\lambda}{\delta}\sum_{g\in\mathcal{G}_0}\sum_{t\in[n-1]} \|(vu^\top - \hat{M})_{\mathcal{J}_g,t}\|_2 \leqslant \frac{4\lambda(C_1 nk)^{1/2}}{\delta}\|vu^\top - \hat{M}\|_{\mathrm{F}},\end{aligned}$$

where we used Lemma 3.13 in the penultimate inequality and Lemma 3.14 in the final bound. This proves the first claim of the proposition, and the second claim follows from the first by the same argument as used in Wang and Samworth (2018, online supplement (18) and (19)). □

Lemma 3.17. *Suppose $\Sigma \in \mathbb{R}^{d\times d}$ is a symmetric positive semidefinite matrix and let $E \sim N(0, \Sigma)$. Then we have for any $\delta > 0$ that*

$$\mathbb{P}\big(\|E\|^2 > \mathrm{tr}(\Sigma) + 2\|\Sigma\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + 2\|\Sigma\|_{\mathrm{op}}\log(1/\delta)\big) \leqslant \delta.$$

Proof. Let $\Sigma = U^\top \Lambda U$ be the eigendecomposition of Σ, such that $U \in \mathbb{R}^{d\times d}$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1(\Sigma), \ldots, \lambda_d(\Sigma))$ is a diagonal matrix with eigenvalues of E on its diagonal. Hence, there exist $Z_1, \ldots, Z_d \overset{\text{iid}}{\sim} N(0,1)$ such that $\|E\|_2^2 = \|UE\|_2^2 = \sum_{j=1}^d \lambda_j(\Sigma) Z_j^2$. Applying Laurent and Massart (2000, Lemma 1), we have with probability at least $1-\delta$ that

$$\begin{aligned}\|E\|_2^2 &\leqslant \sum_{j=1}^d \lambda_j(\Sigma) + 2\Big(\sum_{j=1}^d \lambda_j^2(\Sigma)\Big)^{1/2}\sqrt{\log(1/\delta)} + 2\max_{j=1}^d \lambda_j(\Sigma)\log(1/\delta) \\ &\leqslant \mathrm{tr}(\Sigma) + 2\|\Sigma\|_{\mathrm{F}}\sqrt{\log(1/\delta)} + 2\|\Sigma\|_{\mathrm{op}}\log(1/\delta)\end{aligned}$$

as desired. □

Lemma 3.18. *Suppose $\Sigma \in \mathbb{R}^{p\times p}$ is a symmetric positive semidefinite matrix with $\|\Sigma\|_{\mathrm{op}} \leqslant B$. Let $W = (W_1, \ldots, W_n)$ be an $p \times n$ random matrix with independent columns $W_t \sim N_p(0, \Sigma)$. Define $E := \mathcal{T}(W)$. Let $p_g = |\mathcal{J}_g|$ with $p_* = \min_{g\in[G]} p_g$. Then for any $\delta \in (0,1)$ and $\lambda = B^{1/2}\big(1 + \sqrt{2p_*^{-1}\log(1/\delta)}\big)$, we have that*

$$\mathbb{P}(\|E\|_{\mathrm{grp}*} > \lambda) \leqslant (n-1)G\delta.$$

Proof. By the definition of the CUSUM transformation $\mathcal{T}$ in (2.9), we have that $E_{\mathcal{J}_g,t} \sim$

$N(0, \Sigma_{\mathcal{J}_g,\mathcal{J}_g})$. By a union bound, we have

$$\begin{aligned}
\mathbb{P}(\|E\|_{\mathrm{grp}*} > \lambda) &\leqslant \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}(\|E_{\mathcal{J}_g,t}\|_2^2 > p_g\lambda^2) \\
&\leqslant \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}\biggl(\|E_{\mathcal{J}_g,t}\|_2^2 > Bp_g\biggl(1+\sqrt{\frac{2\log(1/\delta)}{p_g}}\biggr)^2\biggr) \\
&\leqslant \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}\biggl(\|E_{\mathcal{J}_g,t}\|_2^2 > B\Bigl(p_g + 2\sqrt{p_g\log(1/\delta)} + 2\log(1/\delta)\Bigr)\biggr) \\
&\leqslant \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}\bigl(\|E_{\mathcal{J}_g,t}\|_2^2 > \mathrm{tr}(\Sigma_{\mathcal{J}_g,\mathcal{J}_g}) + 2\|\Sigma_{\mathcal{J}_g,\mathcal{J}_g}\|_{\mathrm{F}}\sqrt{\log(1/\delta)} \\
&\qquad\qquad + 2\|\Sigma_{\mathcal{J}_g,\mathcal{J}_g}\|_{\mathrm{op}}\log(1/\delta)\bigr) \\
&\leqslant (n-1)G\delta.
\end{aligned}$$

as desired, where we used the fact that $\|\Sigma_{\mathcal{J}_g,\mathcal{J}_g}\|_{\mathrm{op}} \leqslant \|\Sigma\|_{\mathrm{op}} \leqslant B$ in the penultimate inequality and Lemma 3.17 in the final bound. □

Lemma 3.19. *Let $W = (W_1, \ldots, W_n)$ be a $p \times n$ random matrix with $W_i \overset{\mathrm{iid}}{\sim} N_p(0, \Sigma)$ and $E = \mathcal{T}(W) = (E_1, \ldots, E_{n-1})$. Suppose $\|\Sigma\|_{\mathrm{op}} \leqslant B$ and that $\min(z, n-z) \geqslant n\tau$ and $|z-t| \leqslant n\tau/2$. For a deterministic vector $v \in \mathbb{R}^p$ and any $\lambda_1 > 0$, there exists an event Ω_1 with probability at least $1 - 16e^{-\lambda_1^2/(4B)}\log n$ such that on this event, we have*

$$|v^\top E_z - v^\top E_t| \leqslant 2\sqrt{2}\lambda_1\sqrt{\frac{z-t}{n\tau}} + 8\lambda_1\frac{z-t}{n\tau}.$$

Proof. Define event

$$\Omega_1 := \biggl\{\biggl|\sum_{r=1}^{s} v^\top W_r - \sum_{r=1}^{t} v^\top W_r\biggr| \leqslant \lambda_1\sqrt{|s-t|}, \text{ for } 0 \leqslant t \leqslant n \text{ and } s \in \{0, z, n\}\biggr\}.$$

Since $v^\top W_1, \ldots, v^\top W_n \overset{\mathrm{iid}}{\sim} N(0, v^\top\Sigma v)$, with $v^\top\Sigma v \leqslant B$, by Wang and Samworth (2018, Lemma 5), for any $u \geqslant 0$, and $m \in \mathbb{N}$, we have

$$\mathbb{P}\biggl(\max_{1\leqslant t\leqslant m}\biggl|\frac{1}{\sqrt{t}}\sum_{r=1}^{t} v^\top W_r\biggr| \geqslant uB^{1/2}\biggr) \leqslant 4e^{-u^2/4}\log m. \tag{3.35}$$

Applying the above bound four times, we have

$$\mathbb{P}(\Omega_1^{\mathrm{c}}) \leqslant 4e^{-\lambda_1^2/(4B)}\{2\log n + \log z + \log(n-z)\} \leqslant 16e^{-\lambda_1^2/(4B)}\log n.$$

It hence suffices to show that on Ω_1, the desired inequality holds. By symmetry, we may assume without loss of generality that $t < z$. From the definition of the CUSUM transformation in (2.9), we have

$$\begin{aligned} v^\top E_z - v^\top E_t &= \sqrt{\frac{n}{z(n-z)}}\left(\frac{z}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{z} v^\top W_r\right) \\ &\qquad - \sqrt{\frac{n}{t(n-t)}}\left(\frac{t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r\right) \\ &= \sqrt{\frac{n}{z(n-z)}}\left(\frac{z-t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=t+1}^{z} v^\top W_r\right) \\ &\qquad + \left(\sqrt{\frac{n}{z(n-z)}} - \sqrt{\frac{n}{t(n-t)}}\right)\left(\frac{t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r\right). \end{aligned} \tag{3.36}$$

On the event Ω_1,

$$\begin{aligned} \left|\frac{z-t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=t+1}^{z} v^\top W_r\right| &\leqslant \frac{z-t}{n}\left|\sum_{r=1}^{n} v^\top W_r\right| + \left|\sum_{r=t+1}^{z} v^\top W_r\right| \\ &\leqslant \frac{z-t}{n}\lambda_1\sqrt{n} + \lambda_1\sqrt{z-t} \leqslant 2\lambda_1\sqrt{z-t} \end{aligned} \tag{3.37}$$

Similarly, we have on Ω_1 that

$$\left|\frac{t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r\right| \leqslant \frac{\lambda_1 t}{\sqrt{n}} + \lambda_1\sqrt{t} \leqslant 2\lambda_1\sqrt{t}.$$

Noticing that $\frac{t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r = \frac{n-t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=t+1}^{n} v^\top W_r$, we can similarly bound the left-hand side above by $2\lambda_1\sqrt{n-t}$. Therefore, on Ω_1, we have

$$\begin{aligned} \left|\frac{t}{n}\sum_{r=1}^{n} v^\top W_r - \sum_{r=1}^{t} v^\top W_r\right| &\leqslant 2\lambda_1 \min\{\sqrt{t}, \sqrt{n-t}\} \\ &\leqslant 2\lambda_1 \min\left\{\sqrt{z}, \sqrt{n-z+\frac{n\tau}{2}}\right\}. \end{aligned} \tag{3.38}$$

By the mean value theorem, there exists $\xi \in [t, z]$ such that

$$\begin{aligned} \left|\sqrt{\frac{n}{z(n-z)}} - \sqrt{\frac{n}{t(n-t)}}\right| &\leqslant \frac{z-t}{2}\left(\frac{n}{\xi(n-\xi)}\right)^{3/2} \\ &\leqslant \frac{\sqrt{2}(z-t)}{\min\{(z-n\tau/2)^{3/2}, (n-z)^{3/2}\}}. \end{aligned} \tag{3.39}$$

Combining (3.36), (3.37), (3.38) and (3.39), we have on Ω_1 that

$$\begin{aligned}\left|v^\top E_z - v^\top E_t\right| &\leqslant 2\lambda_1\sqrt{\frac{n(z-t)}{z(n-z)}} + \frac{2^{3/2}\lambda_1(z-t)\min\left\{z^{1/2},(n-z+n\tau/2)^{1/2}\right\}}{\min\{(z-n\tau/2)^{3/2},(n-z)^{3/2}\}}\\ &\leqslant 2\sqrt{2}\lambda_1\sqrt{\frac{z-t}{n\tau}} + 8\lambda_1\frac{z-t}{n\tau},\end{aligned}$$

as desired. □

Lemma 3.20. *Suppose $\mu = (\mu_1,\ldots,\mu_n)$ has a single change point at z, in the sense that $\mu_1 = \cdots = \mu_z = \mu^{(1)}$ and $\mu_{z+1} = \cdots = \mu_n = \mu^{(2)}$. Let $A = \mathcal{T}(\mu) = (A_1,\ldots,A_n)$. Define $\theta = \mu^{(1)} - \mu^{(2)}$. Then for any $v \in \mathbb{R}^p$, and $|z-t| \leqslant n\tau/2$, we have*

$$\left|v^\top A_z - v^\top A_t\right| \geqslant \frac{2}{3\sqrt{6}}\frac{|z-t|}{\sqrt{n\tau}}(v^\top\theta).$$

Proof. Observe that A is a rank one matrix given by (3.17). Hence, $v^\top A = (v^\top\theta)\gamma^\top$. The desired result is then a consequence of Wang and Samworth (2018, Lemma 7). □

Lemma 3.21. *Let $W = (W_1,\ldots,W_n)$ be a $p\times n$ random matrix with independent columns W_t satisfying $\|W_t\|_{\psi_2^*} \leqslant L$ and $\|\mathrm{Var}(W_t)\|_{\mathrm{op}} \leqslant B$ for $t \in [n-1]$. Define $E := \mathcal{T}(W)$. Let $p_g = |\mathcal{J}_g|$ with $p_* = \min_{g\in[G]} p_g$. There exists a universal constant $C > 0$ such that for any $\delta \in (0,1)$, we have*

$$\mathbb{P}\left\{\|E\|_{\mathrm{grp}*} > CLB^{1/2}\left(1 + \sqrt{\frac{\log(nG/\delta)}{p_*}}\right)\right\} \leqslant \delta.$$

Proof. By the definition of the CUSUM transformation $\mathcal{T}$ in (2.9), we can write E_t as $E_t = \sum_{s\in[n]} a_s W_s$ for a contrast vector $a = (a_1,\ldots,a_n)^\top$ such that $\|a\|_2 = 1$. For each $t \in [n]$, Since $\|W_t\|_{\psi_2^*} \leqslant L$, we have for any $v \in \mathcal{S}^{p-1}$ that $\|v^\top W_s/\{v^\top\mathrm{Var}(W_t)v\}^{1/2}\|_{\psi_2} \leqslant L$. Therefore, by Vershynin (2012, Proposition 5.10), there exists a constant $C_1 > 0$ such that for every $t \in [n-1]$ we have

$$\|E_t\|_{\psi_2^*} = \sup_{v\in\mathcal{S}^{p-1}} \frac{\|v^\top E_t\|_{\psi_2}}{(v^\top\mathrm{Var}(W_t)v)^{1/2}} = \sup_{v\in\mathcal{S}^{p-1}} \left\|\frac{\sum_{s=1}^n a_s v^\top W_s}{(v^\top\mathrm{Var}(W_t)v)^{1/2}}\right\|_{\psi_2} \leqslant C_1 L.$$

Then, we can bound $\|E_t\|_{\psi_2}$ by:

$$\|E_t\|_{\psi_2} \leqslant \|E_t\|_{\psi_2^*}\|\Sigma\|_{\mathrm{op}}^{1/2} \leqslant C_1 LB^{1/2}.$$

Define $\mathcal{S}_g^{p-1} := \{v \in \mathcal{S}^{p-1} : \mathrm{supp}(v) \subseteq \mathcal{J}_g\}$ and let $\mathcal{N}_g \subseteq \mathcal{S}_g^{p-1}$ be a 1/2-net of the set $\mathcal{S}_g^{p-1}$. By Vershynin (2012, Lemma 5.2), we can choose $\mathcal{N}_g$ such that $|\mathcal{N}_g| \leqslant 5^{p_g}$. Obseve that

$$\begin{aligned}\|E_{\mathcal{J}_g,t}\|_2 &= \sup_{v\in\mathcal{S}_g^{p-1}} v^\top E_t \leqslant \sup_{v\in\mathcal{N}_g} v^\top E_t + \sup_{u:\|u\|_2\leqslant 1/2,\mathrm{supp}(u)\subseteq\mathcal{J}_g} |u^\top E_t| \\ &= \sup_{v\in\mathcal{N}_g} v^\top E_t + \frac{1}{2}\|E_{\mathcal{J}_g,t}\|_2 \leqslant 2\sup_{v\in\mathcal{N}_g} v^\top E_t.\end{aligned}$$

Hence, by a union bound and a tail bound of sub-Gaussian random variables, we have we have for some universal constant $C_2 > 0$ that

$$\mathbb{P}(\|E_{\mathcal{J}_g,t}\|_2 \geqslant x) \leqslant \mathbb{P}\biggl(\sup_{v\in\mathcal{N}_g} v^\top E_t \geqslant \frac{x}{2}\biggr) \leqslant 5^{p_g} e^{-x^2/(C_2^2 L^2 B)}.$$

By another union bound, we have

$$\begin{aligned}&\mathbb{P}\biggl\{\|E\|_{\mathrm{grp}*} > 2C_2 L B^{1/2}\biggl(1 + \sqrt{\frac{\log(nG/\delta)}{p_*}}\biggr)\biggr\} \\ &\leqslant \sum_{g\in[G]}\sum_{t\in[n-1]} \mathbb{P}\Bigl(\|E_{\mathcal{J}_g,t}\|_2 > C_2 L B^{1/2}\sqrt{2p_g + \log(nG/\delta)}\Bigr) \\ &\leqslant \sum_{g\in[G]} (n-1) 5^{p_g} e^{-2p_g - \log(nG/\delta)} \leqslant \delta,\end{aligned}$$

as desired. □

Lemma 3.22. *Let $W_1, \ldots, W_n$ be independent centered sub-Gaussian random variables with $\max_t \|W_t\|_{\psi_2} \leqslant K$ for $t \in [n]$. Define $Z_t := t^{-1/2}\sum_{r=1}^t W_r$. Then for $n \geqslant 5$ and $u \geqslant 0$, we have for some universal constant $C > 0$ that*

$$\mathbb{P}(\max_{1\leqslant t\leqslant n} Z_t \geqslant u) \leqslant 2e^{-u^2/(CK^2)}\log n.$$

Proof. Define $S_t := \sum_{r=1}^t W_r$. Then, $(S_t)_t$ is a martingle and $(e^{S_t})_t$ is a non-negative sub-martingle. Then, by a union bound, we have

$$\mathbb{P}\Bigl(\max_{1\leqslant t\leqslant n} Z_t \geqslant u\Bigr) \leqslant \sum_{j=1}^{\lceil \log_2(n+1)\rceil} \mathbb{P}\Bigl(\max_{2^{j-1}\leqslant t < 2^j} Z_t \geqslant u\Bigr)$$

Then by Doob's martingle inequality and Vershynin (2012, Lemma 5.9), we have for some universal constant $C_1 > 0$ that

$$\begin{aligned}
\mathbb{P}\Bigl(\max_{2^{j-1}\leqslant t<2^j} Z_t \geqslant u\Bigr) &\leqslant \sum_{j=1}^{\lceil \log_2(n+1)\rceil} \inf_{\lambda>0} \mathbb{P}\Bigl(\max_{2^{j-1}\leqslant t<2^j} e^{\lambda S_t} \geqslant e^{2^{(j-1)/2}\lambda u}\Bigr) \\
&\leqslant \sum_{j=1}^{\lceil \log_2(n+1)\rceil} \inf_{\lambda>0} \mathbb{E}e^{\lambda S_{2^j}} e^{-2^{(j-1)/2}\lambda u} \\
&\leqslant \sum_{j=1}^{\lceil \log_2(n+1)\rceil} \inf_{\lambda>0} e^{C_1\lambda^2 2^{j-1}K^2} e^{-2^{(j-1)/2}\lambda u} \\
&= \sum_{j=1}^{\lceil \log_2(n+1)\rceil} e^{-u^2/(4CK^2)} \leqslant 2e^{-u^2/(4C_1K^2)} \log n,
\end{aligned}$$

where in the final step, we used the fact that $n \geqslant 5$. The desired result follows by taking $C = 4C_1$. □

Lemma 3.23. *Let $W = (W_1, \ldots, W_n)$ be a $p \times n$ random matrix with columns satisfying $\max_t \|W_t\|_{\psi_2^*} \leqslant L$ and $E = \mathcal{T}(W) = (E_1, \ldots, E_{n-1})$. Suppose $\min(z, n-z) \geqslant n\tau$ and $|z - t| \leqslant n\tau/2$. For a deterministic vector v and $\lambda_1 = L\sqrt{CB\log n}$, we have with probability at least $1 - \frac{16\log n}{n}$ that*

$$|v^\top E_z - v^\top E_t| \leqslant 2\sqrt{2}\lambda_1\sqrt{\frac{z-t}{n\tau}} + 8\lambda_1\frac{z-t}{n\tau}$$

Proof. By a similar argument as in the proof of Lemma 3.21, we have for all $r \in [n]$ and $v \in \mathcal{S}^{p-1}$ that $\|v^\top W_r\|_{\psi_2} \leqslant LB^{1/2}$. Define event

$$\Omega_1 := \Biggl\{\Biggl|\sum_{r=1}^{s} v^\top W_r - \sum_{r=1}^{t} v^\top W_r\Biggr| \leqslant \lambda_1\sqrt{|s-t|},\ \text{for } 0 \leqslant t \leqslant n \text{ and } s \in \{0, z, n\}\Biggr\}.$$

Then, by Lemma 3.22, for any $u \geqslant 0$, and $m \in \mathbb{N}$, we have

$$\mathbb{P}\Biggl(\max_{1\leqslant t\leqslant m}\Biggl|\frac{1}{\sqrt{t}}\sum_{r=1}^{t} v^\top W_r\Biggr| \geqslant \lambda_1\Biggr) \leqslant 4e^{-\lambda_1^2/(CL^2B)}\log m.$$

Applying the above bound four times, we have

$$\begin{aligned}
\mathbb{P}(\Omega_1^{\mathrm{c}}) &\leqslant 4e^{-\lambda_1^2/(CL^2B)}\{2\log n + \log z + \log(n-z)\} \\
&\leqslant 16e^{-\lambda_1^2/(CL^2B)}\log n \leqslant \frac{16\log n}{n}.
\end{aligned}$$

It hence suffices to show that on Ω_1, the desired inequality holds. This deterministic calculation follows verbatim from the proof of Lemma 3.19. $\square$

Lemma 3.24. *Suppose $\Sigma \in \mathbb{R}^{p\times p}$ is a symmetric positive semidefinite matrix. Let $W = (W_1,\dots,W_n)$ be an $p\times n$ random matrix with dependent columns $W_t \sim N_p(0,\Sigma)$ satisfying equation (3.14). Define $E := \mathcal{T}(W)$. Let $p_g = |\mathcal{J}_g|$ with $p_* = \min_{g\in[G]} p_g$. Then for any $\delta \in (0,1)$ and $\lambda = \sqrt{2B^*}\big(1+\sqrt{2p_*^{-1}\log(1/\delta)}\big)$, we have that*

$$\mathbb{P}(\|E\|_{\mathrm{grp}*} > \lambda) \leqslant (n-1)G\delta.$$

Proof. Fix $t \in [n-1]$ and define $\kappa = (\kappa_1,\dots,\kappa_n)^\top \in \mathbb{R}^n$ by $\kappa_r = -\sqrt{\frac{n-t}{nt}}\mathbb{1}_{\{r\leqslant t\}} + \sqrt{\frac{t}{n(n-t)}}\mathbb{1}_{\{r>t\}}$ (for simplicity, we have suppressed the t dependence in the definition of κ). Then we have $E_t = \sum_{r=1}^n \kappa_r W_r \sim N(0,\Sigma^*)$ for some positive semidefinite matrix $\Sigma^* \in \mathbb{R}^{p\times p}$. For any $v \in \mathcal{S}^{p-1}$, we have

$$\begin{aligned}
v^\top\Sigma^* v = \mathrm{Var}(v^\top E_t) &= \sum_{r_1=1}^n\sum_{r_2=1}^n \kappa_{r_1}\kappa_{r_2} v^\top K(|r_2-r_1|)v \\
&\leqslant 2\sum_{u=0}^{n-1} v^\top K(u)v \sum_{r=1}^{n-u}\kappa_r\kappa_{r+u} \\
&\leqslant 2\sum_{u=0}^{n-1} v^\top K(u)v\bigg\{\frac{(n-t)(t-u)_+}{nt} + \frac{t(n-t-u)_+}{n(n-t)}\bigg\} \leqslant 2B^*.
\end{aligned}$$

Consequently, we have $\|\Sigma^*\|_{\mathrm{op}} \leqslant 2B^*$. Then, following the proof of Lemma 3.18 and B with $2B^*$, we can obtaine the desired result. $\square$

Chapter 4

High dimensional change-point estimation under network structure

4.1 Introduction

In this chapter, we proposed a method, called `SpreadDetect`, to estimate the change-points under network structure. Existing work on change-point analysis with network structure includes Chen et al. (2022), Wang et al, (2017) which detect changes in the network, Dette et al. (2022) which detects changes in covariance structure. Here, we consider a different setting where the coordinates represent nodes of a graph/network and the change, instead of occurring simultaneously in all coordinates of interest, may initially appear in one coordinate (the *source coordinate* of change), and then spread across the network gradually over time. Such a statistical model is useful to represent, for instance, the spread of infectious disease between individuals over time. We are interested in estimating both the source coordinate and the time point where the change occurs at the source coordinate. Note that different coordinates will have a change occurring at a different time point. To avoid ambiguity, we refer to the time of the change in the source coordinate as the *initial change-point*, or simply the *change-point* of the model, and the time point of change in any given coordinate as the *time of spread* to that coordinate, which is typically later than the change-point. In such a setting, the change signal may be very small and sparse when it first appears, and increases as the change is spread across the network. Thus, a naive application of a multivariate change-point procedure may miss the initial part of the change and likely estimate a change-point with a positive bias. Moreover, in many applications, the coordinate(s) where the change first appears may be of separate interest. The task is to estimate both the source coordinate and the initial change-point time in a statistical model where the change is spread across the network via adjacent nodes.

The key idea here is to aggregate evidence of change, measured in terms of coordinatewise CUSUM statistics, across multiple coordinates with suitable time lags. We then centre these aggregated CUSUM statistics so that under the null distribution, candidate change-points near and far away from the boundary of the time window considered are treated on equal footings. The method is explained in detail in Section 4.2. Depending on whether the signs of the change in different coordinates are equal, we propose quadratic and linear test statistics respectively, indexed both in time and over the coordinates. The final estimator for the time and coordinate of initial change is obtained by maximising these aggregated statistics.

In Section 4.3, we derive theoretical guarantees of our proposed `SpreadDetect` method. For simplicity, we focus on the case where the change is spreading across the network at a deterministic rate. We assume that if the change-point and source coordinate pair varies from (z^*, j^*) to (t^*, k^*), at least m nodes in the network will witness a difference in their time of spread at least proportional to the sum of the time difference between z^* and t^* and the graph distance between j^* and k^*. We first derive a key result in Theorem 4.1, saying that assuming that the change is bounded away from the endpoint, and provided the magnitude of change is up to logarithmic factors above $\sqrt{p}/(nm) + p/(nm^2)$, then both the source coordinate and initial change-point time can be accurately estimated. Theorem 4.4 then shows that our estimation procedure can be turned into a test with good size and power controls for testing the existence of a change-point of the above signal size. Theorem 4.5 shows that when $m \asymp p$ (a condition that can be verified in many common graphs), the signal size required in Theorem 4.1 is in fact minimax optimal. In addition, we derive in Theorem 4.6 the result for the special case when we know the sign of the signal so that the linear statistics in Algorithm 8 is used. In this case, the estimation accuracy is guaranteed if the magnitude of change is above $1/\sqrt{mn\tau^2}$ up to logarithmic factor.

In Section 4.4, we evaluate the empirical performance of the method through simulated data and a COVID-19 real data example. We evaluate our method under two settings when the signal spreads to the nearby coordinates in a fixed or random way using the simulated data. Proofs of all theoretical results are deferred to Section 4.5, and ancillary results and their proofs are given in Section 4.6.

4.2 Problem setup and methodology

Given a network represented by a connected graph G, with vertices $V(G) := [p]$ and edges $E(G) \subseteq [p] \times [p]$, let $j^* \in V(G)$ be the source coordinate and $z^* \in [n]$ the change-point and write $S_t \subseteq [p]$ for the set of "infected nodes", i.e., coordinates that have undergone a change at or before time t. We have $S_t = \emptyset$ for $t < z^*$, $S_{z^*} = \{j^*\}$ and we assume that the change spreads from infected nodes to their neighbours at a constant rate in the sense that at any time $t > z^*$, $S_t := \{j : (j,k) \in E(G) \text{ for some } k \in S_{t-1}\}$. Suppose the data $X_1, \ldots, X_n \in \mathbb{R}^{V(G)} \cong \mathbb{R}^p$ follow multivariate normal distribution with an identity covariance such that

$$\mathbb{E}(X_t) = \mu^0 \circ \mathbf{1}_{S_t^{\mathrm{c}}} + \mu^1 \circ \mathbf{1}_{S_t}, \quad \text{for } t \in [n],$$

where μ^0 and μ^1 are respectively vectors of means pre- and post-change, and $\mathbf{1}_A := (\mathbb{1}_{j\in A})_{j\in[p]}$ for any $A \subseteq [p]$.

Let $d_G(j,k)$ be the graph distance between nodes j and k, i.e., the length of the shortest path from j to k on graph G. Then, the data consists of independent random variables:

$$X_{j,t} \sim \begin{cases} N(\mu_j^0, 1) & \text{if } t \leqslant z^* + d_G(j,j^*) \\ N(\mu_j^1, 1) & \text{if } t > z^* + d_G(j,j^*), \end{cases} \quad \text{for } j \in [p] \text{ and } t \in [n].$$

We define P_{j^*,z^*,μ^0,μ^1} to be the distribution of the data matrix $X = (X_1, \ldots, X_n) \in \mathbb{R}^{p\times n}$ given parameters $(j^*, z^*, \mu^0, \mu^1) \in \Theta := [p] \times [n] \times \mathbb{R}^p \times \mathbb{R}^p$. Our task is to estimate j^* and z^* given data $X \sim P_{j^*,z^*,\mu^0,\mu^1}$. Define $\theta = (\theta_1, \ldots, \theta_p)^\top := \mu^1 - \mu^0$. We assume that $|\theta_j| \geqslant a$ for some $a > 0$ for all $j \in [p]$.

Writing $\mu = \mathbb{E}X \in \mathbb{R}^{p\times n}$, we have the decomposition $X = \mu + W$, where W is a $p \times n$ random matrix with independent $N(0, 1)$ entries. we form $T = \mathcal{T}(\mathbf{X})$ to be the CUSUM transform of matrix X as defined in equation (2.9). The CUSUM transformation is the normalised difference before and after the change for a single entry in the data matrix. Motivated by this meaning, and that in each coordinate, the CUSUM statistic is maximised at the time of spread, we propose to aggregate these CUSUM statistics in different coordinates at appropriate lags. Specifically, given any candidate source coordinate and change-point pair (j, t), we compute the time of spread to each coordinate k as $t+d_G(j, k)$ and aggregate $T_{k,t+d_G(j,k)}$ over $k \in [p]$ provided that $t + d_G(j, k) \leqslant n$. For each $t \in [n - 1]$ and $k \in [p]$, we define $\mathcal{J}_{j,t} := \{k \in [p] : t + d_G(j, k) < n\}$. If we do not know the sign of the signal, We form the following quadratic statistic

$$Q_{j,t} := \sum_{k\in\mathcal{J}_{j,t}} (T^2_{k,t+d_G(j,k)} - 1). \tag{4.1}$$

Here, we subtract 1 from the summands to make them mean-centred, so that candidate change-points near the right boundary will not be disfavoured due to the set $\mathcal{J}_{j,t}$ being smaller. We then estimate the location of the change-point z^* and the source coordinate of the spread via

$$(\hat{j}, \hat{z}) = \arg\max_{j,t} Q_{j,t}. \tag{4.2}$$

Practically, the $p\times p$ distance matrix $(d_G(j, k) : j, k \in [p])$ between every pair of vertices can be pre-calculated from the adjacency matrix in $O(p^3)$ time using the Floyd–Warshall algorithm (Floyd, 1962). The entire estimation procedure is summarised in Algorithm 9.

Figure 4.1 illustrates the working of Algorithm 9 in action. Here, we have a data matrix $X \in \mathbb{R}^{200\times 200}$, which contains a change spreading from the source coordinate $j^* = 50$ from the change-point time $z^* = 50$. The right panel displays the matrix $(Q_{j,t})_{j\in[p],t\in[n-1]}$ of aggregated squared CUSUM statistics from equation 4.1. The darker colour indicates

Algorithm 9: Spreading change estimation procedure

```
Input: $X \in \mathbb{R}^{p \times n}$, graph $G$
Compute $T \leftarrow \mathcal{T}(X)$ as in (2.9).
Compute $Q_{j,t}$ for $j \in [p]$ and $t \in [n-1]$ via Equation (4.1).
Estimate $(\hat{j}, \hat{z}) = \arg\max_{j,t} Q_{j,t}$.
Output: $(\hat{j}, \hat{z})$
```

larger values of the $Q_{j,t}$ statistics. We can see that the aggregation proposed by 4.1 indeed helps us locate both the source coordinate and the true time of change-point.

In some practical applications, it is reasonable to assume additionally that signs of the changes are the same across all coordinates. In such settings, we can modify the quadratic aggregation proposed in (4.1) by using the following linear statistic instead:

$$L_{j,t} := \left| \sum_{k \in \mathcal{J}_{j,t}} T_{k,t+d_G(j,k)} \right|. \tag{4.3}$$

The source coordinate and the change-point are then correspondingly estimated via

$$(\hat{j}, \hat{z}) = \arg\max_{j,t} L_{j,t}. \tag{4.4}$$

4.3 Theoretical results

In this section, we derive theoretical guarantees of the change-point estimation procedure proposed in Algorithm 9.

The main challenge here is that when the change first occurs, it is localised in a small number of coordinates in a neighbourhood of the source node, making the change signal relatively weak. Hence, to accurately identify the time that the change first started and the associated source node, we need to use information further down the line to infer back the origin of the change. However, this is not always possible, as can be seen from the following example. Suppose that we have $G = (V, E)$ is a path graph, where $V = \{1, 2, \ldots, p\}$ and $E = \{(1,2), (2,3), \ldots, (p-1, p)\}$. Then the data generated from P_{1,z^*,μ^0,μ^1} and P_{2,z^*+1,μ^0,μ^1} have exactly the same distribution in all times t except $t = z^*$.

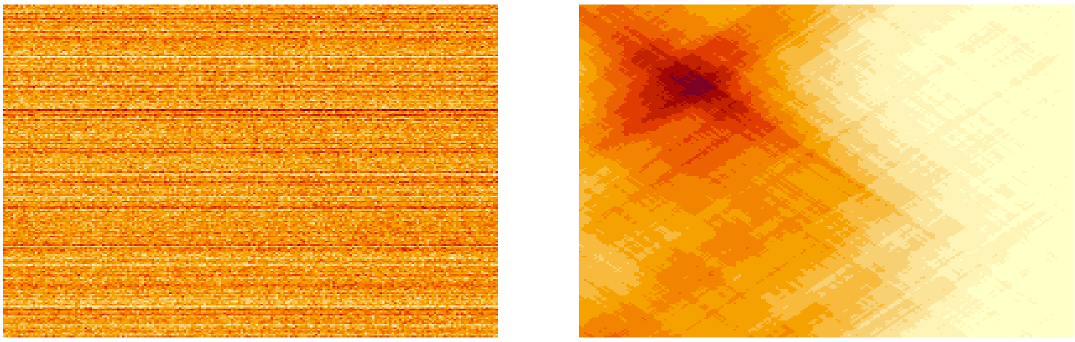

Figure 4.1: Illustration of the `SpreadDetect` algorithm. The heatmap of the original data matrix X is shown on the left panel, where data consist of $p = 200$ nodes in a cycle graph measured over a period of $n = 200$ time points. A true change occurs at $z = 50$ from coordinate 50 and spread across the graph following the model described in Section 4.4. The right panel depicts the heatmap of the aggregated CUSUM statistics generated in the `SpreadDetect` algorithm. The estimated time of change $\hat{z} = 52$ and the estimated origin of change $\hat{j} = 46$ is where the matrix of the aggregated CUSUM statistics achieves its maximum value.

In other words, knowledge of the later stage spread pattern is not helpful in testing apart whether the source of the change is at $j^* = 1$ or $j^* = 2$. A closer inspection of the above simple example reveals that an essential condition for consistent for any fixed t^*, k^*, we define the following set:

$$\mathcal{J}_{t^*,k^*}(C_1) = \Big\{ j \in V(G) : |z^* + d_G(j, j^*) - (t^* + d_G(j, k^*))| \geqslant C_1(|z^* - t^*| + d_G(j^*, k^*)) \Big\}. \tag{4.5}$$

This set counts the number of nodes that will witness a difference in their time of spread at least proportional to the sum of the time difference between z^* and t^* and the distance between j^* and k^* given that the change-point and source coordinate pair varies from (z^*, j^*) to (t^*, k^*). We remark that $\mathcal{J}_{t^*,k^*}(C_1)$ also depends on z^* and j^*, though we will suppress this dependence in the notation since in what follows, we will mostly treat z^* and j^* as fixed or can be inferred from the context. For consistent estimation to be possible, we would require $\min_{t^*,k^*} |\mathcal{J}_{t^*,k^*}(C_1)|$ to be sufficiently large, as demonstrated in the theorem below.

Theorem 4.1. *Suppose $n\tau \geqslant 2p$ and $X \sim P_{j^*,z^*,\mu_0,\mu_1}$ with $\mu_0 - \mu_1 \in \{-a, a\}^p$. Define $m = m_G(C_1) := \min_{t^*,k^*} |\mathcal{J}_{t^*,k^*}(C_1)|$. There exists a universal constant $c > 0$ such that if*

$$a^2 \geqslant c\biggl\{\frac{\sqrt{p} + \log(2pn)}{n\tau m} + \frac{p\log(2pn)}{n\tau^2 m^2}\biggr\}. \tag{4.6}$$

then, the estimator $(\hat{j}, \hat{z})$ from (4.2) satisfies with probability at least $1 - 1/(2pn)$ that

$$|\hat{z} - z^*| + d_G(\hat{j}, j^*) \leqslant \frac{12\sqrt{6}}{C_1 m}\biggl\{\frac{\sqrt{p} + \log(2pn)}{a^2} + \frac{\sqrt{pn\log(2pn)}}{a}\biggr\}.$$

The $n\tau \geqslant 2p$ condition is placed to ensure that the change happens early in the time series to allow sufficient time to spread to all nodes in the network. This helps simplifying our analysis and presentation. However, we note that a similar result can be derived without this assumption; see Theorem 4.8. We remark also that Condition (4.6) is mild in view of the conclusion of Theorem 4.1. Indeed, for the right-hand side of the loss bound to be nontrivial (i.e. less than $n + p$), we would at least need $a^2 \gtrsim \{\sqrt{p} + \log(2pn)\}/(nm) + p\log(2pn)/(nm^2)$. Thus, (4.6) only requires a^2 to be larger than a factor of at most τ^{-2} than minimally what is required in Theorem 4.1. The final loss bound is inversely proportional to $C_1 m_G(C_1)$. In general, $m_G(C_1)$ is a decreasing function of C_1 and by the triangle inequality, $m_G(C_1) = 0$ for all $C_1 \geqslant 1$. Hence, the optimal loss bound we can obtain involves a carefully chosen trade-off between C_1 and $m_G(C_1)$ in the denominator of the final bound. In practice, in many applications, we have $m_G(C_1) \asymp p$ for some $C_1 \asymp 1$. Under such assumptions, and if in addition $\log(n) = O(\sqrt{p})$, the conclusion of Theorem 4.1 simplifies to

$$\frac{|\hat{z} - z^*|}{n} + \frac{d_G(\hat{j}, j^*)}{p} = O\biggl(\frac{1}{p^{1/2}\|\theta\|_2^2} + \frac{\sqrt{n\log(2pn)}}{p\|\theta\|_2}\biggr),$$

showing that both the location of change and the origin of change estimators are consistent when $\|\theta\|_2 \gg \max\{p^{-1/4}, p^{-1}\sqrt{n\log(2pn)}\}$.

As mentioned above, the quantity $m_G(C_1)$ plays an important role in our theoretical control of the loss of change-point location and origin estimation. To get a sense of the magnitude of this quantity, we compute $m_G(1/4)$ for grid graphs, binary trees and random Erdős–Rényi graphs. Figure 4.2 shows that we have $m_G(1/4) \geqslant cp$ for some constant $c > 0$ in all these simulation settings. Moreover, for each specific type of graph,

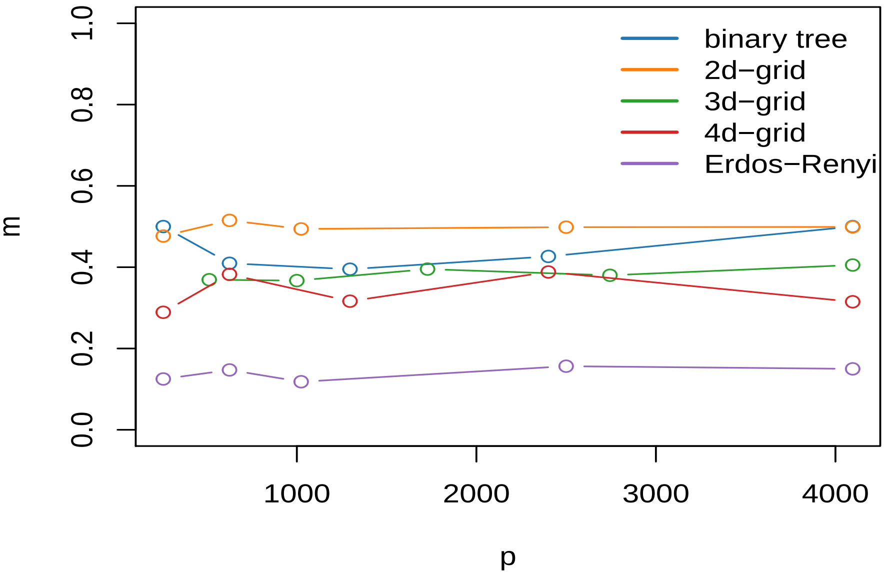

Figure 4.2: $m_G(1/4)/p$ for different graphs.

$m_G(1/4)/p$ tends to be relatively stable when p is large. Theoretically, $m(C_1)$ needs to be controlled in a case-specific manner. Below, we illustrate how this can be done in the setting of a d-dimensional grid graph. For simplicity of exposition, we introduce additional symmetry to require that the grid is 'wrapped around the edges', in the sense that $G = \prod_{r=1}^{d} G_r$, where each G_r is a p_1-cycle C_{p_1} with $p_1^d = p$. Working with the product of cycles instead of paths makes all vertices of G equivalent. The following proposition controls $m_G(1/(4d))$ of such a graph G.

Proposition 4.2. *Suppose $G = \prod_{r=1}^{d} G_r$ with $G_r \cong C_{p_1}$ for all $r \in [d]$ and $p = p_1^d$. Assume further that $n\tau \geqslant 2p_1$. Then we have $m_G(1/(4d)) \geqslant p/8^d$.*

Treating the dimension of the grid as fixed, we have the desired bound that $m_G(C_1) \asymp p$ for some $C_1 \asymp 1$. The following result is an immediate consequence of Theorem 4.1 together with Proposition 4.2.

Corollary 4.3. *Under the same assumption as in Proposition 4.2. Suppose $X \sim P_{j^*,z^*,\mu_0,\mu_1}$*

with $\mu_0 - \mu_1 \in \{-a, a\}^p$*. There exist* $c, C > 0$*, depending only on* d*, such that if*

$$a^2 \geqslant c\left\{\frac{\sqrt{p} + \log(2pn)}{n\tau p} + \frac{pn\log(2pn)}{n^2\tau^2p^2}\right\}, \tag{4.7}$$

then with probability at least $1 - 1/(2pn)$*, the estimator* $(\hat{j}, \hat{z})$ *defined in* (4.2) *satisfies that*

$$|\hat{z} - z| + d_G(\hat{j}, j) \leqslant C\left\{\frac{\sqrt{p} + \log(2pn)}{a^2p} + \frac{\sqrt{pn\log(2pn)}}{ap}\right\}.$$

While the focus of our discussion so far has been the estimation of a change-point (both in terms of the time of change and location of the source of change), our method can be easily modified for the related testing problem. More precisely, given the data X described in Section 4.2, we are interest in testing $H_0 : \theta = 0$ against the alternative: $H_1 : \theta \neq 0$. We can construct a test based on the quadratic statistics computed according to Algorithm 8 as follows:

$$\psi_\lambda(X) = \mathbb{1}_{\{\max_{j\in[p],t\in[n-1]} Q_{j,t} \geqslant \lambda\}}. \tag{4.8}$$

The following theorem shows that for an appropriate choice of λ, the test ψ_λ defined above has small Type I and Type II errors.

Theorem 4.4. *Given* $X \sim P = P_{j^*,z^*,\mu_0,\mu_1}$*. For any* $\delta \in (0, 1)$ *and* $\lambda \geqslant 2\sqrt{p\log(pn/\delta)} + 2\log(pn/\delta)$*, the test* ψ_λ *defined in* (4.8) *has the following properties.*

(a) If $\theta = 0$*, then*

$$\mathbb{P}_P(\psi_\lambda(X) = 1) \leqslant \delta.$$

(b) There exists a universal constant $C > 0$ *such that if* $a^2 \geqslant \frac{C\lambda}{n\tau\min\{2p,n\tau\}}$*, then*

$$\mathbb{P}_P(\psi_\lambda(X) = 1) \geqslant 1 - \delta.$$

From Theorem 4.4 above, if $p = O(n\tau)$ and taking $\delta = 1/(pn)$, then the test ψ_λ defined in (4.8) is able to detect a change when $a^2 \geqslant \frac{C\sqrt{\log(pn)}}{\sqrt{pn\tau}}$. Note that when $m_G(C_1)$ is of order p and $\sqrt{p}\tau \gtrsim \log(2pn)$, then the signal-size condition in (4.6) is equivalent to

$$a^2 \gtrsim \frac{1}{n\tau\sqrt{p}} + \frac{\log(2pn)}{n\tau^2p} \gtrsim \frac{1}{n\tau\sqrt{p}}.$$

Hence, the signal strength needed here for testing is consistent, up to logarithmic factors, with (4.6) in Theorem 4.1 in such a setting. However, the estimation problem is harder comparing to the testing problem, when $C_1 m_G(C_1)$ is much smaller than p for all choices of $C_1 \in (0,1)$. This can happen, for instance, in the case when there exists a t^* close to z^* such that the signal from z^* needs to pass from t^* to spread over the rest of the coordinate, it is hard to tell which time point does the signal start. However, for the testing problem, we only need to know whether there is a change regardless of the location.

To understand the optimality of the signal-size condition in (4.6), we derive a minimax lower bound for testing the existence of a change-point. Let

$$\Theta_0 := \{(j^*, z^*, \mu_0, \mu_1) \in \Theta : \mu_0 = \mu_1, \min(z^*, n - z^*) \geqslant n\tau\}$$
$$\Theta_{1,a} := \{(j^*, z^*, \mu_0, \mu_1) \in \Theta : \mu_0 - \mu_1 \in \{-a, a\}^p, \min(z^*, n - z^*) \geqslant n\tau\}$$

be two subspaces in the parameter space Θ. We consider the problem of testing the null hypothesis $(j^*, z^*, \mu_0, \mu_1) \in \Theta_0$ against the alternative $(j^*, z^*, \mu_0, \mu_1) \in \Theta_1$ using data X.

Theorem 4.5. *If $n\tau \geqslant 1$, then for $a^2 \leqslant \frac{\sqrt{\log 2}}{\sqrt{2pn\tau}}$, we have that*

$$\inf_{\psi}\Big\{ \sup_{(j^*,z^*,\mu_0,\mu_1)\in\Theta_0} P_{j^*,z^*,\mu_0,\mu_1}(\psi = 1) + \sup_{(j^*,z^*,\mu_0,\mu_1)\in\Theta_{1,a}} P_{j^*,z^*,\mu_0,\mu_1}(\psi = 0)\Big\} \geqslant 1/2,$$

where the infimum is taken over all measurable test functions $\psi : \mathbb{R}^{p\times n} \to \{0,1\}$.

In the setting described after Theorem 4.4, Theorem 4.5 shows that Condition (4.6) in our estimation result is necessary for the even simpler task of testing the existence of a change-point.

We then consider the special case when we know the sign of the changes in each coordinate. Without loss of generality, we may assume that all changes are positive. In this case, we can use the linear statistic defined in equation (4.3) and the following theorem shows that this linear statistic achieves good performance in terms of the estimation consistency:

Theorem 4.6. *Suppose $n\tau \geqslant 2p$ and $X \sim P_{j^*,z^*,\mu_0,\mu_1}$ with $\mu_0 - \mu_1 \in \{-a\}^p \cup \{a\}^p$. Define $m = m_G(C_1) := \min_{t^*,k^*} |\mathcal{J}_{t^*,k^*}(C_1)|$. There exists a universal constant c such that* *if $a \geqslant c\sqrt{\log(pn)/(mn\tau^2)}$, then the estimator $(\hat{j}, \hat{z})$ from (4.4) satisfies with probability at least $1 - 1/(2pn)$ that*

$$|\hat{z} - z^*| + d_G(\hat{j}, j^*) \leqslant \frac{C^* \log(pn)}{a^2 m}.$$

From this result, we can see that the estimation accuracy of the estimator from the linear statistic achieves the convergence rate of $\log(pn)/(a^2p)$ and a^2p in the denominator is the ℓ_2 norm of θ. Similar rates have also been observed in many change-point results (Csörgö and Horváth , 1997). This condition is also the same as the second term in equation 4.6 in Theorem 4.1. The following result is an immediate consequence of Theorem 4.6 together with Proposition 4.2.

Corollary 4.7. *Under the same assumption as in Proposition 4.2. Suppose $X \sim P_{j^*,z^*,\mu_0,\mu_1}$ with $\mu_0 - \mu_1 \in \{-a\}^p \cup \{a\}^p$. There exist $c, C > 0$, depending only on d, such that if $a \geqslant c\sqrt{\log(pn)/(pn\tau^2)}$, then with probability at least $1 - 1/(2pn)$, the estimator $(\hat{j}, \hat{z})$ defined in (4.4) satisfies that*

$$|\hat{z} - z^*| + d_G(\hat{j}, j^*) \leqslant \frac{C \log(pn)}{a^2 p}.$$

We also present here a general result of Theorem 4.1 without the condition $n\tau \geqslant 2p$:

Theorem 4.8. *Suppose $X \sim P_{j^*,z^*,\mu_0,\mu_1}$ with $\mu_0 - \mu_1 \in \{-a, a\}^p$. Assuming*

$$a^2 \geqslant c\left\{\frac{\sqrt{p} + \log(2pn)}{n\tau \min(p, n\tau)} + \frac{pn \log(2pn)}{n^2\tau^2 \min(p, n\tau)^2}\right\}. \tag{4.9}$$

Suppose we are using the quadratic statistics defined in equation (4.1), we have with probability at least $1 - 1/(2pn)$ that

$$|\hat{z} - z| + d_G(\hat{j}, j) \leqslant C\left\{\frac{\sqrt{p} + \log(2pn)}{a^2 \min(p, n\tau)} + \frac{\sqrt{pn \log(2pn)}}{a \min(p, n\tau)}\right\}.$$

4.4 Numerical studies

4.4.1 Deterministic spreading model

In this subsection, we compare our method with other possible ways to locate the change. The first possible way is for each row of the data, we perform a one-dimensional change-point testing, that is, pick out the time point with the largest absolute value of the CUSUM statistics for each coordinate. The earliest time and the coordinate corresponding to that time is the desired change-point location. We try two different kinds of change-point locations: in the middle and near the end of the boundary. For the first case, we set $n = 200$ and vary $p \in \{100, 200, 500\}$ and signal size $\mu_j^1 - \mu_j^0 \in \{0.1, 0.2, 0.5\}$. For the second case, we set $n = 500$ and vary $p \in \{500, 800, 1000\}$ and signal size $\mu_j^1 - \mu_j^0 \in \{0.2, 0.3, 0.4, 0.5\}$. We compare the mean absolute deviation between the estimated and true location of z^* and j^* respectively. Columns $\hat{z}^*_{SD}$ and $\hat{j}^*_{SD}$ are mean absolute deviation for z^* and j^* from Algorithm 9 respectively while columns $\hat{z}^*_{coordwise}$and $\hat{j}^*_{coordwise}$ are results from testing procedure stated above. Table 4.1 shows that our method can locate the change-point accurately especially when μ_j^0, μ_j^0 grows above 0.2 in both change-point settings.

4.4.2 Stochastic spreading model

In this subsection, we consider the case when the spread of the change occur independently with probability q each time from an infected node to each of its neighbours. In this case, we can modify our existing methodology, which monitors for deterministic spreading of the change as follows. If the probability q is known, then we can adjust the distance between coordinates j and k as the expected time that a change spreading from source coordinate j will reach k under this stochastic model. For a line graph $G = C_p$, this would simply be $d_G(j, k)/q$. When q is unknown, we may search over a grid $\mathcal{Q}$ of q values in $[0, 1]$, compute the test statistics $\max_{j,t} Q_{j,t}^{(q)}$ for each q as in (4.2) with this adjusted distance metric and then choose the optimal q by $\hat{q} := \arg\max_{q \in \mathcal{Q}} \max_{j,t} Q_{j,t}^{(q)}$. The final estimator for the source coordinate and the time of change-point is defined as $(\hat{j}, \hat{t}) := \arg\max_{(j,t)} Q_{j,t}^{(\hat{q})}$. In Table 4.2, we compare the performance of the method described above (denoted by rSD) and the vanilla `SpreadDetect` algorithm (denoted by SD), together with the baseline coordinatewise procedure mentioned in Section 4.4.1. We set the true probability of change spread to $q = 0.5$, and search over the grid $\mathcal{Q} = \{0.1, 0.2, \ldots, 0.9\}$ and vary n, p, z^* and j^*. We see that the modified `SpreadDetect` algorithm described in this subsection has the best performance over the wide range of parameter settings considered.

n	p	z^*	signal size	$\hat{z}^*_{SD}$	$\hat{z}^*_{coordwise}$	$\hat{j}^*_{SD}$	$\hat{j}^*_{coordwise}$
200	100	100	0.1	25.1	92.24	20.79	46.08
200	100	100	0.2	2.07	61.44	2.35	33.35
200	100	100	0.5	0.06	28.78	0.07	14.91
200	200	100	0.1	23.34	85.28	33.12	87.6
200	200	100	0.2	1.72	59.36	1.69	62.19
200	200	100	0.5	0.01	29.78	0.01	19.92
200	500	100	0.1	59.84	87.24	77.93	204.56
200	500	100	0.2	4.14	60.92	4.05	110.07
200	500	100	0.5	0	34.61	0	26.35
500	500	400	0.2	10.4	106.24	10.36	101.2
500	500	400	0.3	0.2	98.02	0.16	31.39
500	500	400	0.4	0.04	121.48	0.03	30.65
500	500	400	0.5	0	131.16	0	30.18
500	800	400	0.2	51.59	161.95	51.9	142.05
500	800	400	0.3	0.18	160.9	0.15	97.21
500	800	400	0.4	0	179.76	0	86.56
500	800	400	0.5	0	171.38	0	87.75
500	1000	400	0.2	77.04	160.7	77.18	173.15
500	1000	400	0.3	0.22	158.25	0.2	104.48
500	1000	400	0.4	0.02	166.98	0	98.88
500	1000	400	0.5	0.01	171.3	0	94.17

Table 4.1: Average mean absolute deviation (over 100 repetitions) comparison between different methods. Other parameters used: $j^* = p/2$.

n	p	z^*	j^*	signals	$\hat{z}^*_{SD}$	$\hat{z}^*_{rSD}$	$\hat{z}^*_{coordwise}$	$\hat{j}^*_{SD}$	$\hat{j}^*_{rSD}$	$\hat{j}^*_{coordwise}$
200	100	100	50	0.2	17.29	9.05	97.79	6.64	3.35	48.59
200	100	100	50	0.3	27.86	4.35	83.97	3.41	1.89	41.79
200	100	100	50	0.4	16.67	3.57	41.47	2.53	1.62	21.46
200	200	100	100	0.2	19.23	20.67	96.43	23.07	16.33	96.29
200	200	100	100	0.3	19.10	5.79	88.02	11.24	2.38	85.25
200	200	100	100	0.4	17.07	3.66	52.35	4.6	1.81	46.18
200	500	100	250	0.2	61.09	44.43	98.15	77.36	42.95	246
200	500	100	250	0.3	39.23	7.82	87.6	42.23	5.3	209.06
200	500	100	250	0.4	22.59	4.16	46.55	10.6	1.71	83.5
500	200	250	100	0.2	41.69	6.68	152.61	5	2.61	64.23
500	200	250	100	0.3	41.47	5.77	29.53	3.92	2.36	14.77
500	200	250	100	0.4	41.15	5.39	43.72	3.37	2.39	12.28
500	500	250	250	0.2	43.69	5.76	170.54	7.64	2.46	151.04
500	500	250	250	0.3	42.34	5.09	35.02	6.16	2.38	14.48
500	500	250	250	0.4	43.02	5.19	44.59	4.99	2.41	13.11

Table 4.2: Average mean absolute deviation (over 100 repetitions) comparison between different methods for estimating the time of change-point and source coordinate under a stochastic spreading model described in Section 4.4.2

4.4.3 Real data example

We now apply Algorithm 9 to the data set of weekly deaths between January 2017 and December 2020 in United States. The aim is to find the time of the change in the number of deaths and state where the change first occurs. We exclude two states: Alaska and Hawaii in our analysis as they have no adjacent states. To form the adjacency matrix, if two states are adjacent to each other, then we assign the corresponding entry with 1, otherwise, the entries are 0. Before applying Algorithm 9 to the data, we first remove the seasonal trend from the data. Specifically, we use the data up to 30 June 2019 as the training data and estimate the daily death by averaging the weekly total death and then use a Gaussian Kernel with bin width of 20 to estimate the deaths on each day of a year. As daily death follows Poisson distribution, we stabilise the variance by applying a square root transformation. Then, we calculate the difference between the actual data and the fitted data and standardize it using the mean and standard deviation of the calculated difference.

We apply Algorithm 9 to the pre-processed data set. The resulting time is 7 March 2020, and the state which first started to change is Pennsylvania. The date matches the actual situation, as during that time, death due to COVID-19 began to occur. Figure 4.3 shows the aggregated CUSUM statistics with the states arranged such that Pennsylvania is in the centre and the graph distance increases as we move towards the top and bottom of the plot. The heatmap shown in the figure is consistent with a change spreading from Pennsylvania. However, we remark that the conclusion here should be treated with caution for two reasons. Firstly, this is a weekly recorded data and the frequency of recording is likely to be inadequate to capture the rapid spreading of the disease across multiple states. Secondly, we computed the distance between states by the number of state borders one needs to cross from one to the other. While this is a proxy for the distance between states during the pandemic spread, a better measure would involve, for instance, the number of passengers crossing from one state to another, though the latter data are difficult to obtain.

Furthermore, as the assumptions in our model are quite simple here, we discuss some possible extensions to the model to fit this COVID-19 data. The first one is that we can extend the model to allow spatial dependency, as each state is more likely to have some correlation with other states in reality. Temporal dependency can also be considered as it has also been observed in COVID-19 data (Jiang, 2022). In addition, it is unlikely that all the coordinates undergo changes of the same size. One possible way to extend this is by assuming that the signal for each coordinate is randomly drawn from a uniform distribution.

4.5 Proof of main results

In this section, we give the proofs of main results in Chapter 4.

Proof of Theorem 4.1. Let $A_{j,t}$ be the entries of $A = \mathcal{T}(\boldsymbol{\mu})$ then for $j \in [p]$, we have

$$A_{j,t} = \begin{cases} \sqrt{\frac{t}{n(n-t)}}(n - z^* - d_G(j,j^*))\theta_j, & \text{if } t \leqslant z^* + d_G(j,j^*), \\ \sqrt{\frac{n-t}{nt}}(z^* + d_G(j,j^*))\theta_j, & \text{if } t > z^* + d_G(j,j^*). \end{cases}$$

Since the test statistic is unchanged by flipping signs in any one row of the data, we may assume without loss of generality that $\theta_j > 0$ for all j.

Fix $k^* \in [p]$ and $t^* \in [n-1]$, we define

$$B_{k^*,t^*} := \sum_{j \in \mathcal{J}_{k^*,t^*}} A^2_{j,t^*+d_G(k^*,j)}.$$

For each j such that $t^* + d_G(k^*, j) < n$, we have $T_{j,t^*+d_G(k^*,j)} \sim N(A_{j,t^*+d_G(k^*,j)}, 1)$ and obtain that $Q_{k^*,t^*} + |\mathcal{J}_{k^*,t^*}| \sim \chi^2_{\mathcal{J}_{k^*,t^*}}(B_{k^*,t^*})$. Therefore, by Birgé (2001, Lemma 8.1), for each $j \in [p]$ and $t \in [n-1]$, we have for any $\delta \in (0,1)$ that

$$\mathbb{P}\Big(|Q_{k^*,t^*} - B_{k^*,t^*}| > 2\sqrt{(|\mathcal{J}_{k^*,t^*}| + 2B_{k^*,t^*})\log(2/\delta)} + 2\log(2/\delta)\Big) \leqslant \delta.$$

Taking a union bound over $k^* \in [p]$ and $t^* \in [n-1]$ of the above inequality, we therefore obtain that with probability at least $1 - \delta$,

$$\begin{aligned} B_{j^*,z^*} - 2\sqrt{(|\mathcal{J}_{j^*,z^*}| + 2B_{j^*,z^*})\log(2pn/\delta)} - 2\log(2pn/\delta) \leqslant Q_{j^*,z^*} \leqslant Q_{\hat{j},\hat{z}} \\ \leqslant B_{\hat{j},\hat{z}} + 2\sqrt{(|\mathcal{J}_{\hat{j},\hat{t}}| + 2B_{\hat{j},\hat{t}})\log(2pn/\delta)} + 2\log(2pn/\delta). \end{aligned} \tag{4.10}$$

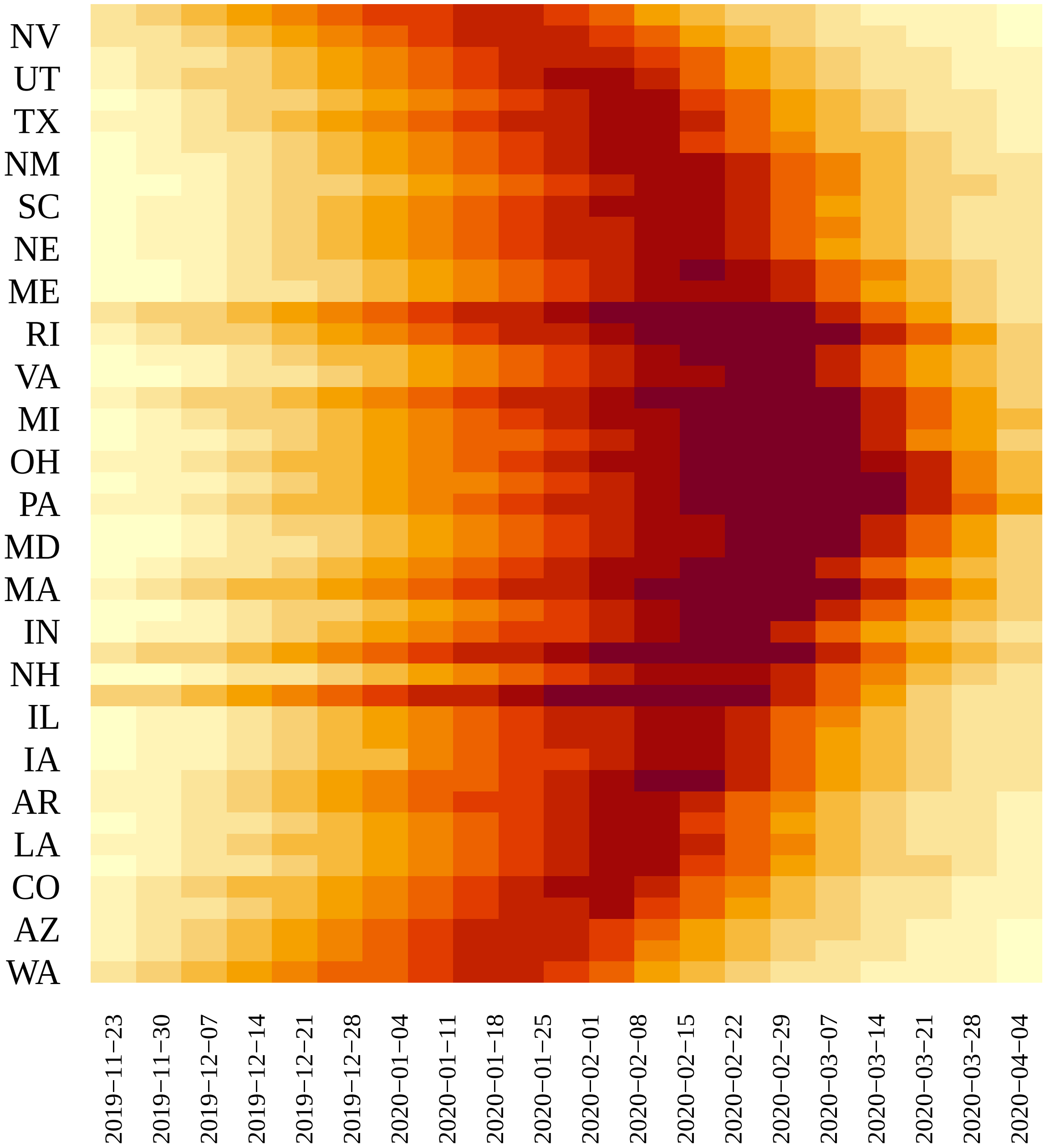

Figure 4.3: Aggregated CUSUM statistics as computed in (4.1) for the weekly COVID-19 excess death data in 46 US states during 2019-11-23 to 2020-04-04. Data are preprocessed as described in Section 4.4.3.

Notice that for every $k^* \in [p]$ and $t^* \in [n-1]$, we have $|\mathcal{J}_{k^*,t^*}| \leqslant p$ and

$$B_{k^*,t^*} \leqslant \sum_{j\in[p]} A^2_{j,z^*+d_G(j,j^*)} \leqslant \sum_j \theta_j^2 \frac{(z^*+d_G(j,j^*))(n-z^*-d_G(j,j^*))}{n} \leqslant \frac{n\|\theta\|_2^2}{2}.$$

Thus, after rearranging (4.10), we have with probability at least $1-\delta$ that

$$B_{j^*,z^*} - B_{\hat{j},\hat{z}} \leqslant 4\sqrt{(p+n\|\theta\|_2^2)\log(2pn/\delta)} + 4\log(2pn/\delta). \tag{4.11}$$

On the other hand, we can obtain a lower bound of $B_{j^*,z^*} - B_{\hat{j},\hat{z}}$ as follows. For each $j \in [p]$, the sequence $(A_{j,t})_t$ is unimodal with a single peak at $z^* + d_G(j,j^*)$. Moreover, since $d_G(j,j^*) \leqslant p \leqslant n\tau/2$, we have

$$\theta_j A_{j,z^*+d_G(j,j^*)} = \theta_j^2 \sqrt{\frac{(z^*+d_G(j,j^*)(n-z^*-d_G(j,j^*))}{n}} \geqslant \theta_j^2\sqrt{\frac{(n\tau/2)(n/2)}{n}} \geqslant \frac{\theta_j^2\sqrt{n\tau}}{2}. \tag{4.12}$$

Therefore, by Wang and Samworth (2018, Lemma 7), we have for each $j \in \mathcal{J}_{\hat{z},\hat{j}}(C_1)$ that

$$\begin{aligned}\theta_j(A_{j,z^*+d_G(j,j^*)} - A_{j,\hat{z}+d_G(j,\hat{j})}) &\geqslant \frac{2\theta_j^2}{3\sqrt{6n\tau}} \min\Big\{|z^*+d_G(j,j^*)-\hat{z}-d_G(j,\hat{j})|, \frac{n\tau}{2}\Big\}\\ &\geqslant \frac{2\theta_j^2}{3\sqrt{6n\tau}} \min\Big\{C_1(|z^*-\hat{z}|+d_G(j^*,\hat{j})), \frac{n\tau}{2}\Big\},\end{aligned} \tag{4.13}$$

where we have used the definition of $\mathcal{J}_{\hat{z},\hat{j}}(C_1)$ from (4.5) in the final bound. Combining (4.13) with (4.12), we obtain that

$$\begin{aligned}B_{j^*,z^*} - B_{\hat{j},\hat{z}} &\geqslant \sum_{j\in\mathcal{J}_{\hat{z},\hat{j}}(C_1)} (A^2_{j,z^*+d_G(j,j^*)} - A^2_{j,\hat{z}+d_G(j,\hat{j})})\\ &\geqslant \sum_{j\in\mathcal{J}_{\hat{z},\hat{j}}(C_1)} A_{j,z^*+d_G(j,j^*)}(A_{j,z^*+d_G(j,j^*)} - A_{j,\hat{z}+d_G(j,\hat{j})})\\ &\geqslant \frac{2a^2m}{3\sqrt{6}} \min\Big\{C_1(|z^*-\hat{z}|+d_G(j^*,\hat{j})), \frac{n\tau}{2}\Big\},\end{aligned} \tag{4.14}$$

where we have used the fact that $|\mathcal{J}_{\hat{z},\hat{j}}(C_1)| \geqslant m$ in the final inequality. Combining (4.11) and (4.14), and choosing $\delta = 1/(2pn)$, we have with probability at least $1-1/(2pn)$ that

$$\frac{2a^2m}{3\sqrt{6}} \min\Big\{C_1(|z^*-\hat{z}|+d_G(j^*,\hat{j})), \frac{n\tau}{2}\Big\} \leqslant 4\sqrt{p+2npa^2\log(2pn)} + 8\log(2pn). \tag{4.15}$$

From condition (4.6), we have

$$a^2 \geqslant \max\Big\{\frac{c(\sqrt{p}+\log(2pn))}{n\tau m}, a\sqrt{\frac{cp\log(2pn)}{n\tau^2m^2}}\Big\} \geqslant \frac{c\sqrt{p}+c\log(2pn)+\sqrt{cnpa^2\log(2pn)}}{2n\tau m},$$

which for sufficiently large c implies that the minimum on the left-hand side of (4.15) is achieved by the first term. Consequently, we derive from (4.15) that with probability at least $1 - 1/(2pn)$,

$$|z^* - \hat{z}| + d_G(j^* - \hat{j}) \leqslant \frac{12\sqrt{6}}{C_1}\left\{\frac{\sqrt{p} + \log(2pn)}{a^2 m} + \frac{\sqrt{pn\log(2pn)}}{am}\right\},$$

as desired. □

Proof of Proposition 4.2. Denote G_r to be the rth copy of C_{p_1} making up G, i.e. $G = \prod_{r=1}^d G_r$. Each vertex $j \in G$ can be represented by a d-tuple of coordinates $(\pi_1(j), \ldots, \pi_d(j))$, where $\pi_r(j) \in V(G_r) = [p_1]$. Define $\ell(j) = \ell_G(j) := (z^* + d_G(j, j^*)) - (t^* + d_G(j, k^*))$, by Proposition 4.10, we have that each of the following set

$$\mathcal{J}_r := \left\{\tilde{j} : \operatorname{sgn}(z^* - t^*)\ell_{G_r}(\tilde{j}) \geqslant \frac{|z^* - t^*| + d_{G_r}(\pi_r(j^*), \pi_r(k^*))}{4}\right\},$$

has cardinality at least $p_1/8$. Then, for all $j \in \mathcal{J} := \prod_{r=1}^d \mathcal{J}_r$, when $z^* \geqslant t^*$, we have:

$$\begin{aligned} z^* - t^* + d_G(j, j^*) - d_G(j, k^*) &= \sum_{r=1}^d \left\{\frac{z^* - t^*}{d} + d_{G_r}(\pi_r(j), \pi_r(j^*)) - d_{G_r}(\pi_r(j), \pi_r(k^*))\right\} \\ &\geqslant \sum_{r=1}^d \frac{\ell_{G_r}(j)}{d} \geqslant \sum_{r=1}^d \frac{z^* - t^* + d_{G_r}(\pi_r(j^*), \pi_r(k^*))}{4d} \\ &\geqslant \frac{z^* - t^*}{4d} + \frac{d_G(j^*, k^*)}{4d} \end{aligned}$$

Similarly, if $t^* > z^*$, we have

$$t^* - z^* + d_G(j, k^*) - d_G(j, j^*) \geqslant \frac{t^* - z^*}{4d} + \frac{d_G(j^*, k^*)}{4d}.$$

Overall, we have for $j \in \mathcal{J}$ that

$$|\ell_G(j)| = |z^* - t^* + d_G(j, j^*) - d_G(j, k^*)| \geqslant \frac{|z^* - t^*|}{4d} + \frac{d_G(j^*, k^*)}{4d}.$$

Hence, $m_G(1/(4d)) \geqslant |\mathcal{J}| = \prod_{r=1}^d |\mathcal{J}_r| \geqslant (p_1/8)^d = p/8^d$ as desired. □

Proof of Theorem 4.4. If $\theta = 0$, then $T_{j,t} \sim N(0, 1)$ for all $t \in [n-1]$ and $j \in [p]$ and hence $Q_{j,t} + |\mathcal{J}_{j,t}| \sim \chi^2_{|\mathcal{J}_{j,t}|}$. By Laurent and Massart (2000, Lemma 1) together with a union bound, we have that

$$\begin{aligned}\mathbb{P}(\max_{j\in[p],t\in[n-1]} Q_{j,t} \geqslant \lambda) &\leqslant \sum_{j=1}^{p}\sum_{t=1}^{n-1}\mathbb{P}(Q_{j,t} \geqslant \lambda)\\ &\leqslant \sum_{j=1}^{p}\sum_{t=1}^{n-1}\mathbb{P}\Bigl\{Q_{j,t} \geqslant 2\sqrt{|\mathcal{J}_{j,t}|\log(pn/\delta)} + 2\log(pn/\delta)\Bigr\} \leqslant \delta.\end{aligned}$$

This establishes part (a). For part (b), let $A_{j,t}$ and $B_{j,t}$ be defined as in the proof of Theorem 4.1. Note that under the alternative hypothesis, $Q_{j,t} + |\mathcal{J}_{j,t}| \sim \chi^2_{|\mathcal{J}_{j,t}|}(B_{j,t})$. Hence, by Birgé (2001, Lemma 8.1), we have

$$\mathbb{P}\Bigl\{Q_{j^*,z^*} \geqslant B_{j^*,z^*} - 2\sqrt{(|\mathcal{J}_{j^*,t^*}| + 2B_{j^*,z^*})\log(1/\delta)}\Bigr\} \geqslant 1-\delta.$$

Under the assumption that $B_{j^*,z^*} \geqslant 8\lambda$, we have

$$\begin{aligned}B_{j^*,z^*} - 2\sqrt{(|\mathcal{J}_{j^*,t^*}| + 2B_{j^*,z^*})\log(1/\delta)} &\geqslant B_{j^*,z^*} - 2\sqrt{p\log(1/\delta)} - 2\sqrt{2B_{j^*,z^*}\log(1/\delta)}\\ &\geqslant B_{j^*,z^*} - \lambda - 2\sqrt{B_{j^*,z^*}\lambda}\\ &= (\sqrt{B_{j^*,z^*}} - \sqrt{\lambda})^2 - 2\lambda \geqslant \lambda.\end{aligned}$$

Since

$$A^2_{j,z^*+d_G(j,j^*)} = \theta_j^2\frac{(z^*+d_G(j,j^*))(n-z^*-d_G(j,j^*))}{n} \geqslant \theta_j^2\frac{(n\tau/2)(n/2)}{n} \geqslant \frac{\theta_j^2 n\tau}{4},$$

there are at least $n\tau/2$ points with $d_G(j,j^*) \leqslant n\tau/2$,

$$B_{j^*,z^*} = \sum_{j\in\mathcal{J}_{j^*,z^*}} A^2_{j,z^*+d_G(j^*,j)} \geqslant \min\bigl(p,\frac{n\tau}{2}\bigr)\frac{a^2 n\tau}{4} = \frac{a^2 n\tau\min(2p,n\tau)}{8}$$

Then, for $a^2 \geqslant 64\lambda/(n\tau\min\{2p,n\tau\})$, we have the desired result. □

Proof of Theorem 4.5. Fix $j^* \in \arg\min_{v\in V(G)}\max_{w\in V(G)} d_G(v,w)$ and $z^* = n - \lceil n\tau\rceil$. Let π be the uniform distribution on $\{-a,a\}^p$. For notational simplicity, define $P_0 := P_{j^*,z^*,0,0}$ and $P_1 := \int P_{j^*,z^*,0,\mu_1}d\pi(\mu_1)$. Then, for any test function ψ, we have

$$\begin{aligned}&\sup_{(j^*,z^*,\mu_0,\mu_1)\in\Theta_0} P_{j^*,z^*,\mu_0,\mu_1}(\psi=1) + \sup_{(j^*,z^*,\mu_0,\mu_1)\in\Theta_1} P_{j^*,z^*,\mu_0,\mu_1}(\psi=0)\\ &\qquad\geqslant P_0(\psi=1) + P_1(\psi=0) \geqslant 1 - d_{\mathrm{TV}}(P_0,P_1)\\ &\qquad= 1 - \frac{1}{2}\int\Bigl|\frac{dP_1}{dP_0} - 1\Bigr|dP_0 \geqslant 1 - \frac{1}{2}\Bigl\{\int\Bigl(\frac{dP_1}{dP_0} - 1\Bigr)^2 dP_0\Bigr\}^{1/2}\\ &\qquad= 1 - \frac{1}{2}\Bigl\{\int\Bigl(\frac{dP_1}{dP_0}\Bigr)^2 dP_0 - 1\Bigr\}^{1/2}. \qquad (4.16)\end{aligned}$$

Let $\boldsymbol{\mu}$ be the conditional mean of X given μ_1 under $P_{j^*,z^*,0,\mu_1}$ and let $\tilde{\boldsymbol{\mu}}$ be an independent copy of $\boldsymbol{\mu}$. By Ingster and Suslina (2012), we have for some independent Rademacher random variables $\xi_1, \ldots, \xi_p$ that

$$\begin{aligned}\int \left(\frac{dP_1}{dP_0}\right)^2 dP_0 = \mathbb{E}\exp\langle \mu, \tilde{\mu}\rangle &= \mathbb{E}\exp\left(\sum_{j=1}^p \max\{n - z^* - d_G(j, j^*), 0\}a^2\xi_j\right) \\ &= \prod_{j=1}^p \Big[\frac{1}{2}e^{\max\{n-z^*-d_G(j,j^*),0\}a^2} + \frac{1}{2}e^{-\max\{n-z^*-d_G(j,j^*),0\}a^2}\Big] \\ &\leqslant \prod_{j=1}^p e^{\max\{n-z^*-d_G(j,j^*),0\}^2 a^4/2} \leqslant e^{2pn^2\tau^2a^4} \leqslant 2,\end{aligned}$$

where the first inequality follows from the fact that $(e^x + e^{-x})/2 \leqslant e^{x^2/2}$ for all $x \in \mathbb{R}$ and the second bound uses the fact that $n - z^* - d_G(j, j^*) \leqslant \lceil n\tau \rceil \leqslant 2n\tau$. Finally, substituting the above inequality into (4.16) we arrive at the desired conclusion. □

Proof of Theorem 4.6. From the definition of $(\hat{j}, \hat{z})$, we have $\sum_{j\in\mathcal{J}_{\hat{z},\hat{j}}(C_1)}(A_{j,z^*+d_G(j,j^*)} + E_{j,z^*+d_G(j,j^*)}) \leqslant \sum_{j\in\mathcal{J}_{\hat{z},\hat{j}}(C_1)}(A_{j,\hat{z}+d_G(j,\hat{j})} + E_{\hat{z}+d_G(j,\hat{j})})$, which can be combined with Proposition 4.11 to obtain that for some universal constant $C_2 > 0$, we have with probability at least $1 - 1/(pn)$ that

$$\sum_{j\in\mathcal{J}_{\hat{z},\hat{j}}(C_1)} (A_{j,z^*+d_G(j,j^*)} - A_{j,\hat{z}+d_G(j,\hat{j})}) \leqslant C_2\left\{|\mathcal{J}_{\hat{z},\hat{j}}(C_1)| \log(pn)\frac{|z^* - \hat{z}| + d_G(j^*, \hat{j})}{n\tau}\right\}^{1/2}. \tag{4.17}$$

On the other hand, by (4.13), we have

$$\sum_{j\in\mathcal{J}_{\hat{z},\hat{j}}(C_1)} (A_{j,z^*+d_G(j,j^*)} - A_{j,\hat{z}+d_G(j,\hat{j})}) \geqslant \frac{2a|\mathcal{J}_{\hat{z},\hat{j}}(C_1)|}{3\sqrt{6n\tau}} \min\left\{C_1(|z^* - \hat{z}| + d_G(j^*, \hat{j})), \frac{n\tau}{2}\right\}. \tag{4.18}$$

Combining (4.17) and (4.18), we have with probability at least $1 - 1/(pn)$ that

$$\frac{2a|\mathcal{J}_{\hat{z},\hat{j}}(C_1)|^{1/2}}{3\sqrt{6\log(pn)}} \min\left\{C_1(|z^* - \hat{z}| + d_G(j^*, \hat{j})), \frac{n\tau}{2}\right\} \leqslant C_2\{|z^* - \hat{z}| + d_G(j^*, \hat{j})\}^{1/2}. \tag{4.19}$$

We claim that when $c \geqslant 6\sqrt{3}C_2$, the minimum on the left-hand side above cannot be achieved at $\frac{n\tau}{2}$. Indeed, from the assumption on a, we have

$$\frac{2a|\mathcal{J}_{\hat{z},\hat{j}}(C_1)|^{1/2}}{3\sqrt{6\log(pn)}}\frac{n\tau}{2} \geqslant \frac{c\sqrt{n}}{3\sqrt{6}} \geqslant C_2\sqrt{2n} > C_2\{|z^* - \hat{z}| + d_G(j^*, \hat{j})\}^{1/2}.$$

Therefore, we have that

$$|z^* - \hat{z}| + d_G(j^*, \hat{j}) \leqslant \frac{27C_2^2 \log(pn)}{2C_1^2 a^2 |\mathcal{J}_{\hat{z},\hat{j}}(C_1)|} \leqslant \frac{C^* \log(pn)}{a^2 m}.$$

□

Proof of Theorem 4.8. Following the proof of Theorem 4.1 and Proposition 4.9, there exists $\mathcal{J} \subset [p]$ such that $|\mathcal{J}| \geqslant \min(p, n\tau)/32$, and for each $j \in \mathcal{J}$, we have

$$|j - j^*| \leqslant \frac{n\tau}{2} \quad \text{and} \quad |z^* + d_G(j, j^*) - (\hat{z} + d_G(j, \hat{j}))| \geqslant \min\left(\frac{n\tau}{16}, \frac{|z^* - \hat{z}| + d_G(j^*, \hat{j})}{4}\right).$$

Combining equation (4.12) and Proposition 4.9, we have that

$$\begin{aligned} A_{j,z^*+d_G(j,j^*)} - A_{j,\hat{z}+d_G(j,\hat{j})} &\geqslant \frac{2\theta_j}{3\sqrt{6n\tau}} \min(|z^* + d_G(j, j^*) - (\hat{z} + d_G(j, \hat{j}))|, n\tau/2) \\ &\geqslant \frac{2\theta_j}{3\sqrt{6n\tau}} \min\left(\frac{|z^* - \hat{z}| + d_G(j^*, \hat{j})}{4}, \frac{n\tau}{16}\right). \end{aligned}$$

Then

$$\begin{aligned} B_{j^*,z^*} - B_{\hat{j},\hat{z}} &\geqslant \sum_{j \in \mathcal{J}} A_{j,z^*+d_G(j,j^*)} (A_{j,z^*+d_G(j,j^*)} - A_{j,\hat{z}+d_G(j,\hat{j})}) \\ &\geqslant \frac{a^2 \min(p, n\tau)}{96\sqrt{6}} \min\left(\frac{|z^* - \hat{z}| + d_G(j^*, \hat{j})}{4}, \frac{n\tau}{16}\right). \end{aligned} \tag{4.20}$$

Combining (4.11) in the proof of Theorem 4.1 and (4.20), and choosing $\delta = 1/(2pn)$, we have with probability at least $1 - 1/(2pn)$ that

$$\frac{a^2 \min(p, n\tau)}{96\sqrt{6}} \min\left(\frac{|z^* - \hat{z}| + d_G(j^*, \hat{j})}{4}, \frac{n\tau}{16}\right) \leqslant 4\sqrt{p + 2npa^2 \log(2pn)} + 8\log(2pn). \tag{4.21}$$

When c is sufficently large, we have from (4.9) that the minimum on the left-hand side of (4.21) is necessarily achieved by the first term. Consequently, we derive from (4.21) that with probability at least $1 - 1/(2pn)$, we have

$$|z^* - \hat{z}| + d_G(j^*, \hat{j}) \leqslant C \left\{ \frac{\sqrt{p} + \log(2pn)}{a^2 \min(p, n\tau)} + \frac{\sqrt{pn \log(2pn)}}{a \min(p, n\tau)} \right\},$$

as desired. □

4.6 Ancillary results

In this section, we collect all ancillary propositions and lemmas used in Chapter 4. We first show that when $G = C_p$, a cycle graph, for a nontrivial fraction of coordinates $j \in [p]$, the difference $\ell(j) = \ell_G(j) := (z^* + d_G(j, j^*)) - (t^* + d_G(j, k^*))$ is large in absolute value.

Proposition 4.9. *Let $G = C_p$ be a p-cycle graph. Let $\tau = \min\{z^*/n, 1-z^*/n\}$. Assuming that $n\tau \geqslant 16$ and $p \geqslant 4$, the following set*

$$\mathcal{J} := \left\{ j : |\ell(j)| \geqslant \min\left(\frac{n\tau}{16}, \frac{|z^* - t^*| + d_G(j^*, k^*)}{4} \right) \text{ and } d_G(j, j^*) \leqslant \frac{n\tau}{2} \right\}$$

has cardinality at least $\min(p, n\tau)/32$.

Proof. Without loss of generality, we may assume by symmetry that $j^* = \lceil p/2 \rceil$ and $k^* \geqslant j^*$. This choice is convenience since $d_G(j, j^*) = |j - j^*|$. With this choice, we can write

$$\ell(j) = \begin{cases} (z^* - t^*) + (k^* - j^*) - 2j & 1 \leqslant j \leqslant k^* - j^* \\ (z^* - t^*) - (k^* - j^*) & k^* - j^* \leqslant j \leqslant j^* \\ (z^* - t^*) + (k^* - j^*) - 2(k^* - j) & j^* \leqslant j \leqslant k^* \\ (z^* - t^*) + (k^* - j^*) & k^* \leqslant j \leqslant p. \end{cases} \tag{4.22}$$

We then prove the result by discussing the following four cases.

Case 1: assume $k^* - t^* \geqslant j^* - z^*$ and $k^* + t^* \leqslant j^* + z^*$. In this case, we have $t^* \leqslant z^*$ and $k^* - j^* \leqslant z^* - t^*$. Hence $\ell(j) \geqslant 0$ for all j. Notice that $\ell(j)$ is an non-decreasing function of j for $j \geqslant j^*$. Then for all j such that

$$j^* + \min\{n\tau/4, (k^* - j^*)/4\} \leqslant j \leqslant j^* + \min\{n\tau/2, p/4\},$$

we have

$$\begin{aligned} \ell(j) &\geqslant \min\bigl\{\ell(j^* + \lceil n\tau/4 \rceil), \ell(j^* + \lceil (k^* - j^*)/4 \rceil)\bigr\} \\ &\geqslant \min\left\{ (z^* - t^*) + \min\left\{ \frac{n\tau}{2} - (k^* - j^*), k^* - j^* \right\}, z^* - t^* - \frac{1}{2}(k^* - j^*) \right\} \\ &\geqslant \min\left\{ \frac{n\tau}{2}, \frac{|z^* - t^*| + |j^* - k^*|}{4} \right\}. \end{aligned}$$

Consequently, in this case, we have

$$|\mathcal{J}| \geqslant \min\{\lfloor n\tau/4 \rfloor, p/8\}.$$

Case 2: assume $k^* - t^* \geqslant j^* - z^*$, $k^* + t^* \geqslant j^* + z^*$ and $z^* \geqslant t^*$. In this case, $k^* - j^* \geqslant z^* - t^* \geqslant 0$. We define h^* to be the point such that $h^* - j^* = \lceil \frac{(k^*-j^*)-(z^*-t^*)}{2} \rceil$. Then $j^* \leqslant h^* \leqslant k^*$ and $\ell(h^*) \in \{0, 1\}$.

We discuss three sub-cases. Case 2a: When $k^* - j^* \leqslant n\tau/4$, let $A = \{j : \frac{k^*+h^*}{2} \leqslant j \leqslant \min(j^* + \frac{n\tau}{2}, p)\}$. Then, for all $j \in A$, we have that

$$\begin{aligned}\ell(j) &\geqslant \ell\left(\left\lceil \frac{h^*+k^*}{2} \right\rceil\right) \geqslant \ell(h^*) + 2\left\lceil \frac{k^*-h^*}{2} \right\rceil \\ &\geqslant \ell(h^*) + \left\lfloor \frac{(z^*-t^*)+(k^*-j^*)}{2} \right\rfloor \geqslant \frac{(z^*-t^*)+(k^*-j^*)}{4}.\end{aligned}$$

and

$$\begin{aligned}|A| = \min\left(p, j^* + \frac{n\tau}{2}\right) - \frac{k^*+h^*}{2} &= \min\left\{p - \frac{k^*+h^*}{2}, \frac{n\tau}{2} - \left(\frac{k^*+h^*}{2} - j^*\right)\right\} \\ &\geqslant \min\left(p - k^* + \frac{k^*-h^*}{2}, n\tau/4\right) \geqslant \min\left(\frac{p-j^*}{8}, n\tau/4\right) \\ &\geqslant \min(p/32, n\tau/4).\end{aligned}$$

Case 2b: When $h^* - j^* \geqslant n\tau/8$, let $A = \{j : j^* \leqslant j \leqslant \frac{j^*+h^*}{2}\}$. Then,

$$|\ell(j)| \geqslant \left|\ell\left(\left\lfloor \frac{j^*+h^*}{2} \right\rfloor\right)\right| \geqslant 2\left\{h^* - \left\lfloor \frac{j^*+h^*}{2} \right\rfloor\right\} - \ell(h^*) \geqslant \frac{n\tau}{16},$$

and $|A| \geqslant \lfloor (h^* - j^*)/2 \rfloor \geqslant \lfloor n\tau/16 \rfloor \geqslant n\tau/32$.

Case 2c: When $k^* - j^* > n\tau/4$ and $h^* - j^* < n\tau/8$, let $A = \{j : h^* + \frac{n\tau}{16} \leqslant j^* \leqslant h^* + \frac{n\tau}{8}\}$. Then,

$$|\ell(j)| \geqslant \ell(h^*) + 2\left\lceil \frac{n\tau}{16} \right\rceil \geqslant \frac{n\tau}{8},$$

and $|A| \geqslant \lfloor n\tau/16 \rfloor \geqslant n\tau/32$.

Combining all subcases and noticing that $A \subseteq \mathcal{J}$, we have the desired result.

Case 3: assume $k^* - t^* \geqslant j^* - z^*$, $k^* + t^* \geqslant j^* + z^*$ and $t^* > z^*$. We define h^* to be the point such that $h^* - j^* = \lfloor \frac{(k^*-j^*)+(t^*-z^*)}{2} \rfloor$. Observe that $\ell(h) \in \{0, -1\}$. In this case,

$k^*-j^* \geqslant t^*-z^* \geqslant 0$. Let $A=\{j : j^*-\min\{2j^*-k^*, n\tau/2\} \leqslant j \leqslant j^*+\min\{\frac{h^*-j^*}{2}, n\tau/2\}$. Since $\ell(j)$ is negative and increasing for $j \in [k^*-j^*, \frac{h^*+j^*}{2}]$, we have for all $j \in A$ that

$$|\ell(j)| \geqslant \left|\ell\left(\left\lfloor\frac{j^*+h^*}{2}\right\rfloor\right)\right| \geqslant 2\left\{h^*-\left\lfloor\frac{j^*+h^*}{2}\right\rfloor\right\}+|\ell(h^*)| \geqslant \lfloor h^*-j^*\rfloor \geqslant \frac{k^*-j^*+t^*-z^*}{4}.$$

Finally, observe by the definition of h^* and the condition $t^* > z^*$ that $h^*-j^* \geqslant \lfloor(k^*-j^*+1)/2\rfloor \geqslant (k^*-j^*)/2$. Hence,

$$\frac{h^*-j^*}{2}+(2j^*-k^*) \geqslant \frac{k^*-j^*+(2j^*-k^*)}{4} \geqslant \frac{j^*}{4} \geqslant \frac{p}{16}.$$

Consequently, we have $|\mathcal{J}| \geqslant |A| \geqslant \min(p/16, n\tau/2)$.

Case 4: assume $k^*-t^* \leqslant j^*-z^*$ and $k^*+t^* \geqslant j^*+z^*$. In this case, we have $t^* \geqslant z^*$. Let $A=\{j : j^*-\min\{2j^*-k^*, n\tau/2\} \leqslant j \leqslant j^*+\min\{\frac{k^*-j^*}{2}, n\tau/2\}\}$. Noticing that $\ell(j)$ is negative and increasing for $j \in [k^*-j^*, k^*]$, we have

$$|\ell(j)| \geqslant \left|\ell\left(\left\lfloor\frac{j+k}{2}\right\rfloor\right)\right| \geqslant t^*-z^* \geqslant \frac{t^*-z^*+k^*-j^*}{2}.$$

We have $(k^*-j^*)/2+(2j^*-k^*) \geqslant j^*/2 \geqslant p/8$. Hence $|\mathcal{J}| \geqslant |A| \geqslant \min\{p/8, n\tau/2\}$. □

With the addition assumption that $n\tau \geqslant 2p$, for $\tau=\min\{z^*/n, 1-z^*/n\}$, we may establish the following improved version of Proposition 4.9 that can be used to prove Theorem 4.1.

Proposition 4.10. *Let $G=C_p$ be a p-cycle graph and $\tau=\min\{z^*/n, 1-z^*/n\}$. Assuming that $n\tau \geqslant 2p$, the following set*

$$\mathcal{J} := \left\{j : \operatorname{sgn}(z^*-t^*)\ell(j) \geqslant \frac{|z^*-t^*|+d_G(j^*, k^*)}{4}\right\}$$

has cardinality at least $p/8$.

Proof. Following the proof of Proposition 4.9, we may assume without loss of genrality that $j^*=\lceil p/2\rceil$ and $k^* \geqslant j^*$, which imlpies that $\ell(j)$ takes the form given in (4.22). We then prove the result by considering four cases as in the proof of Proposition 4.9.

Case 1: assume $k^*-t^* \geqslant j^*-z^*$ and $k^*+t^* \leqslant j^*+z^*$. In this case, we have $t^* \leqslant z^*$ and $k^*-j^* \leqslant z^*-t^*$. Hence $\ell(j) \geqslant 0$ for all j. Notice that $\ell(j)$ is an non-decreasing function of j for $j \geqslant j^*$. Then, for all j such that $j^*+(k^*-j^*)/4 \leqslant j \leqslant p$, we have

$$\ell(j) \geqslant \ell\big(j^*+\lceil(k^*-j^*)/4\rceil\big) \geqslant z^*-t^*+\frac{1}{2}(k^*-j^*) \geqslant \frac{z^*-t^*+k^*-j^*}{4}.$$

Consequently, in this case, $|\mathcal{J}| \geqslant p - j^* - \lceil (k^* - j^*)/4 \rceil + 1 \geqslant p/4$ as required.

Case 2: assume $k^* - t^* \geqslant j^* - z^*$, $k^* + t^* \geqslant j^* + z^*$ and $z^* \geqslant t^*$. In this case, $k^* - j^* \geqslant z^* - t^* \geqslant 0$. We define $h^* := j^* + \lceil \frac{(k^*-j^*)-(z^*-t^*)}{2} \rceil$, and observe that $j^* \leqslant h^* \leqslant k^*$, $k^* - h^* \geqslant (k^* - j^*)/2$ and $\ell(h^*) \in \{0, 1\}$, and that $\ell(j)$ is increasing for $j \in [h^*, p]$. Then, for all j such that $(k^* + h^*)/2 \leqslant j \leqslant p$, we have

$$\begin{aligned}\ell(j) &\geqslant \ell\biggl(\biggl\lceil \frac{h^* + k^*}{2} \biggr\rceil\biggr) = \ell(h^*) + 2\biggl\lceil \frac{k^* - h^*}{2} \biggr\rceil \geqslant \ell(h^*) + (k^* - h^*) \\ &\geqslant \ell(h^*) + \frac{z^* - t^* + k^* - j^* - \ell(h^*)}{2} \geqslant \frac{z^* - t^* + k^* - j^*}{2},\end{aligned}$$

where in the penultimate inequality, we have used the property that $\lceil \frac{(k^*-j^*)-(z^*-t^*)}{2} \rceil = \frac{(k^*-j^*)-(z^*-t^*)+\ell(h^*)}{2}$. Consequently, in this case, $|\mathcal{J}| \geqslant p - \lceil (k^* + h^*)/2 \rceil + 1$. The right-hand side is a decreasing function of k^*. Hence, using the fact that $k^* \leqslant p$ and $h^* \leqslant (j^* + k^*)/2$, we have $|\mathcal{J}| \geqslant p/8$ as desired.

Case 3: assume $k^* - t^* \geqslant j^* - z^*$, $k^* + t^* \geqslant j^* + z^*$ and $t^* > z^*$. We define $h^* := j^* + \lfloor \frac{(k^*-j^*)+(t^*-z^*)}{2} \rfloor$. Observe that $\ell(h^*) \in \{0, -1\}$ and $h^* = \frac{(k^*+j^*)+(t^*-z^*)}{2} - \frac{\ell(h^*)}{2}$. In this case, $k^* - j^* \geqslant t^* - z^* \geqslant 0$. For all j such that $k^* - j^* \leqslant j \leqslant \frac{h^*+j^*}{2}$, $\ell(j)$ is negative and increasing, satisfying

$$\begin{aligned}-\ell(j) &\geqslant -\ell\biggl(\biggl\lfloor \frac{j^* + h^*}{2} \biggr\rfloor\biggr) = 2\biggl\{h^* - \biggl\lfloor \frac{j^* + h^*}{2} \biggr\rfloor\biggr\} - \ell(h^*) \\ &\geqslant h^* - j^* - \ell(h^*) \geqslant \frac{k^* - j^* + t^* - z^*}{2}.\end{aligned}$$

Finally, observe by the definition of h^* and the condition $t^* > z^*$ that $h^* - j^* \geqslant \lfloor (k^* - j^* + 1)/2 \rfloor \geqslant (k^* - j^*)/2$. Hence, we have

$$|\mathcal{J}| \geqslant \frac{h^* + j^*}{2} - (k^* - j^*) \geqslant \frac{h^* - j^*}{2} + (2j^* - k^*) \geqslant \frac{k^* - j^* + (2j^* - k^*)}{4} \geqslant \frac{j^*}{4} \geqslant \frac{p}{8}.$$

Case 4: assume $k^* - t^* \leqslant j^* - z^*$ and $k^* + t^* \geqslant j^* + z^*$. In this case, we have $t^* \geqslant z^*$. For all j such that $k^* - j^* \leqslant j \leqslant \frac{k^*+j^*}{2}$, we note that $\ell(j)$ is negative and increasing, satisfying

$$-\ell(j) \geqslant -\ell\biggl(\biggl\lfloor \frac{j^* + k^*}{2} \biggr\rfloor\biggr) \geqslant t^* - z^* \geqslant \frac{t^* - z^* + k^* - j^*}{2}.$$

Hence, We have $|\mathcal{J}| \geqslant (k^* + j^*)/2 - (k^* - j^*) \geqslant j^*/2 \geqslant p/4$, completing the proof. □

For the case when we are using the linear statistics, we provide the following result of the difference between the sum of $E_{j,z^*+d_G(j,j^*)}$ and $E_{j,t^*+d_G(j,k^*)}$ for coordinates in set $\mathcal{J}_{t^*,k^*}(C_1)$.

Proposition 4.11. *Fix $z^* \in [n-1]$ and $j^* \in [p]$. If $n\tau \geqslant 2p$, then there exists a universal constant $C > 0$ and an event with probability at least $1 - 1/(pn)$, such that on this event, for all $t^* \in [n-1]$, $k^* \in [p]$ and $\mathcal{J}_{t^*,k^*}(C_1) \subseteq [p]$ defined in (4.5), we have*

$$\sum_{j \in \mathcal{J}_{t^*,k^*}(C_1)} (E_{j,z^*+d_G(j,j^*)} - E_{j,t^*+d_G(j,k^*)}) \leqslant C\sqrt{\frac{|\mathcal{J}_{t^*,k^*}(C_1)|(|z^*-t^*| + d_G(k^*,j^*))\log(pn)}{n\tau}}.$$

Proof. First, we claim that if $|z^* - t^*| + d_G(k^*, j^*) \geqslant n\tau/2$, then the conclusion holds trivially. To see this, we note that $\sum_{j\in[p]} E_{j,z^*+d_G(j,j^*)} \sim N(0,p)$ and $\sum_{j\in[p]} E_{j,t^*+d_G(j,k^*)} \sim N(0,p)$. Taking a union bound over t^* and k^*, there is an event with probability at least $1 - 1/(pn)$ such that

$$\max\Bigg\{\sum_{j\in\mathcal{J}_{t^*,k^*}(C_1)} E_{j,z^*+d_G(j,j^*)}, \max_{t^*\in[n-1],k^*\in[p]} \sum_{j\in\mathcal{J}_{t^*,k^*}(C_1)} E_{j,t^*+d_G(j,k^*)}\Bigg\} \leqslant 2\sqrt{p\log(pn)}.$$

So it suffices to take $C = 2\sqrt{2}$ for the desired conclusion to hold. Hence, we may assume without loss of generality that $|z^* - t^*| + d_G(k^*, j^*) < n\tau/2$.

We control $\sum_{j\in\mathcal{J}_{t^*,k^*}(C_1)}(E_{j,z^*+d_G(j,j^*)} - E_{j,t^*+d_G(j,k^*)})$ for fixed $t^* \in [n-1]$, $k^* \in [p]$. For simplicity of notation, we denote $z_j := z^* + d_G(j, j^*)$ and $t_j := t^* + d_G(j, k^*)$. Note that $\sum_{j\in\mathcal{J}_{t^*,k^*}(C_1)}(E_{j,z^*+d_G(j,j^*)} - E_{j,t^*+d_G(j,k^*)})$ is a sum of $|\mathcal{J}_{t^*,k^*}(C_1)|$ independent normal random variables. Hence, we start by controlling the variance of each summand. We consider first the case where $t_j \leqslant z_j$. From the definition of the CUSUM transformation, we can write

$$\begin{aligned}
E_{j,z_j} - E_{j,t_j} &= \sqrt{\frac{n}{z_j(n-z_j)}}\left(\frac{z_j}{n}\sum_{r=1}^{n} W_{j,r} - \sum_{r=1}^{z_j} W_{j,r}\right) \\
&\qquad\qquad - \sqrt{\frac{n}{t_j(n-t_j)}}\left(\frac{t_j}{n}\sum_{r=1}^{n} W_{j,r} - \sum_{r=1}^{t_j} W_{j,r}\right) \\
&= \sqrt{\frac{n}{z_j(n-z_j)}}\left(\frac{z_j - t_j}{n}\sum_{r=1}^{n} W_{j,r} - \sum_{r=t_j+1}^{z_j} W_{j,r}\right) \\
&\qquad + \left(\sqrt{\frac{n}{z_j(n-z_j)}} - \sqrt{\frac{n}{t_j(n-t_j)}}\right)\left(\frac{t_j}{n}\sum_{r=1}^{n} W_{j,r} - \sum_{r=1}^{t_j} W_{j,r}\right).
\end{aligned} \tag{4.23}$$

By the mean value theorem, there exists $\xi \in [t_j, z_j]$, such that

$$\left(\sqrt{\frac{n}{z_j(n-z_j)}} - \sqrt{\frac{n}{t_j(n-t_j)}}\right) \leqslant (z_j - t_j)\left|\frac{\xi}{n} - \frac{1}{2}\right|\left(\frac{n}{\xi(n-\xi)}\right)^{3/2} \leqslant \frac{\sqrt{2}(z_j - t_j)}{\min(\xi, n-\xi)^{3/2}}$$

Also, we observe that:

$$\frac{t_j}{n}\sum_{r=1}^{n} W_{j,r} - \sum_{r=1}^{t_j} W_{j,r} = \sum_{r=t+1}^{n} W_{j,r} - \frac{n-t_j}{n}\sum_{r=1}^{n} W_{j,r}.$$

Since $\sum_{r=1}^{n} W_{j,r}$ and $\sum_{r=t_j+1}^{z_j} W_{j,r}$ are positively corrected with each other, we have

$$\begin{aligned}
\mathbb{V}(E_{j,z_j} - E_{j,t_j}) &\leqslant \frac{2n}{z_j(n-z_j)}\left(\frac{(z_j-t_j)^2}{n} + z_j - t_j\right) \\
&\qquad + \frac{4(z_j-t_j)^2}{\min(\xi, n-\xi)^3}\min\left(\frac{t_j^2}{n} + t_j, n - t_j + \frac{(n-t_j)^2}{n}\right) \\
&\leqslant 4(z_j - t_j)\left(\frac{1}{z_j} + \frac{1}{n-z_j}\right) + \frac{8(z_j-t_j)^2}{\min(t_j, n-z_j)^2}\max\left(1, \frac{n-t_j}{n-z_j}\right)
\end{aligned}$$

Since $n\tau \geqslant 2p$, we have $|z_j - z^*| = d_G(j - j^*) \leqslant p \leqslant n\tau/2$ and consequently $n\tau \leqslant z_j \leqslant n - n\tau/2$. Also, by (4.22), we have $z_j - t_j < |z^* - t^*| + d_G(k^*, j^*) < n\tau/2$, so $n\tau/2 \leqslant t_j \leqslant n - n\tau$. Thus, for some universal constant $C > 0$, we have

$$\begin{aligned}
\mathbb{V}(E_{j,z_j} - E_{j,t_j}) &\leqslant \frac{8(z_j - t_j)}{n\tau} + \frac{4n\tau(z_j - t_j)}{(n\tau/2)^2}\left(1 + \frac{n\tau/2}{n\tau/2}\right) \\
&\leqslant \frac{C(z_j - t_j)}{n\tau} \leqslant \frac{C(|z^* - t^*| + d_G(k^*, j^*))}{n\tau}.
\end{aligned} \tag{4.24}$$

If $z_j > t_j$, a symmetric argument will show that the same variance bound as in (4.24) holds. Therefore,

$$\mathbb{V}\Bigg(\sum_{j \in \mathcal{J}_{t^*,k^*}(C_1)} (E_{j,z_j} - E_{j,t_j})\Bigg) \leqslant \frac{C|\mathcal{J}_{t^*,k^*}(C_1)|(|z^* - t^*| + d_G(k^*, j^*))}{n\tau}$$

Then, for a fixed t^* and k^*, we have that

$$\mathbb{P}\Bigg(\sum_{j \in \mathcal{J}_{t^*,k^*}(C_1)} (E_{j,z^*+d_G(j,j^*)} - E_{j,t^*+d_G(j,k^*)}) \geqslant \sqrt{\frac{4C|\mathcal{J}_{t^*,k^*}(C_1)|(|z^* - t^*| + d_G(k^*, j^*)) \log(pn)}{n\tau}}\Bigg) \leqslant \frac{1}{(pn)^2}.$$

The desired conclusion then follows by taking a union bound over $t^* \in [n-1]$ and $k^* \in [p]$. □

Chapter 5

Discussion

In this thesis, we considered high-dimensional change-point estimation problems under a group sparsity structure and network structure. We proposed the estimation procedures for both cases. We also provided theoretical guarantees for the two algorithms and demonstrated the good performance using simulation studies. We can see that some structural assumptions that appeared in high-dimensional problems, such as regression, can also be adapted to change-point estimation problems. In this thesis, we mainly use the structural information to seek an optimal way to aggregate the data.

In `groupInspect`, we used the grouping information between coordinates to seek an optimal projection direction so as to aggregate the data. In `SpreadDetect`, we aggregate the CUSUM statistics along potential spreading direction according to the network information from coordinates. In both cases, the prior information about structures helps us to find a better estimator for the change-point location. Further work can also be done to extend two algorithms. For the `groupInspect` algorithm we proposed in this thesis, our sparsity assumption is on the groups but not within the group. In reality, there may be cases when there is sparsity within each group. A similar extension appears from group lasso by Yuan and Lin (2006) to sparse group lasso by Simon et al. (2013). In this case, we can use a similar idea to modify equation (3.6) by adding an ℓ_1 penalty for each group to form a convex combination between group norm and ℓ_1 norm. That is, we can solve the following modified optimisation problem:

$$\hat{M} \in \operatorname*{arg\,max}_{M \in \mathcal{S}} \big\{ \langle T, M \rangle - \lambda \|M\|_{\mathrm{grp}} + (1-\lambda)\|M\|_1 \big\}, \tag{5.1}$$

where $\lambda \in [0,1]$. One possible way to solve this optimisation is by coordinate descent.

For `SpreadDetect` algorithm proposed in Chapter 4, we have assumed that the data is of independent normal vectors with deterministic spread of signals. The model assumptions are quite simple here, we discuss some possible extensions. First of all, it is more realistic to assume that there are some dependence structures with the data. We can use a single covariance matrix $\Sigma := \mathrm{Cov}(\mathrm{vec}(X)) \in \mathbb{R}^{np \times np}$ to capture both spatial and temporal dependence in the data matrix X, where $\mathrm{vec}(X)$ denotes the vectorised version of X. For instance, one possibility is to model $\Sigma = \Sigma_{\mathrm{S}} \otimes \Sigma_{\mathrm{T}}$, where Σ_{T} captures the temporal dependence of an individual coordinate series and Σ_{S} captures the cross-sectional dependence over the graph. When $\Sigma_{\mathrm{T}} = I_n$, columns of W consist of n independent $N(0, \Sigma_{\mathrm{S}})$ random vectors. When $\Sigma_{\mathrm{S}} = I_n$, rows of W are independent and generated from $N(0, \Sigma_{\mathrm{T}})$. We can assume that both the temporal and spatial dependence are short-ranged in the sense that $\|\Sigma\|_{\mathrm{op}} \leqslant B$ in order to control the magnitude of the noise series. In addition, it may be more realistic to assume that the change magnitudes of the signal in different coordinates are not fixed to $\pm a$. To handle this, we may model the magnitudes of change as drawn from some random distribution (e.g.uniform distribution over a pre-determined interval). Under the above spatially and temporally correlated noise condition, as well as possible randomly drawn change magnitude size, the current theoretical results will remain qualitatively unchanged, for the following two reasons. Firstly, we expect that with a high probability, there are at least order p number of coordinates with a change signal that is large enough (e.g, $\geqslant a/2$). Therefore, the main results will only change with a constant factor here. Secondly, both temporal and spatial dependence affect only the noise part of the CUSUM series. As long as our covariance matrix is well-conditioned, our main results will differ at most a constant factor.

Furthermore, the algorithm is based on the case when the spread of the signal is deterministic, when the signal will definitely spread to the next coordinate at the next time point. In reality, there is likely some randomness in the spreading of the signal. For example, at each time point, the corresponding coordinate changes with a probability q. In this case, one possible way is to first estimate q and use the expected time of spreading from the source coordinate to a specific coordinate as the distance matrix. We have described the modified algorithm and demonstrated its good performance in Section 4.4.2. Although it works well under simulation study, the theoretical result remains an open question. Although we roughly modified the algorithm to adapt to this case in the simulation study, theoretical guarantees for this modified algorithm still need to be derived.

Finally, it is also possible to extend the algorithm in order to solve epidemic change-point problems. The difference is that, the mean will change for a time period of length l and then go back to the original values. In this case, given t^*, k^*, we need to aggregate the CUSUM statistics along all possible length l and then find the maximum.

Bibliography

Auger, I. E., and Lawrence, C. E. (1989), Algorithms for the Optimal Identification of Segment Neighborhoods, *Bulletin of Mathematical Biology*, **51**,39–54.

Akaike, H. (1974), A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

Aston, J. A. D. and Kirch, C. (2012) Evaluating stationarity via change point alternatives with applications to fMRI data. *Ann. Appl. Stat.*, **6**, 1906–1948.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2019) Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. Roy. Statist. Soc., Ser. B*, **81**, 649–672.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

Birgé, Lucien. (2001) An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, 113–133

Bosc, M. e. a. (2003) Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage*, **20**, 643–656.

Cai, T. T., Zhang, A. and Zhou, Y. (2019) Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *arXiv preprint*, arxiv:1909.09851.

Chan, H. P. (2017) Optimal sequential detection in multi-stream data. *Ann. Statist.*, **45**, 2736–2763.

Chen, C. Y.-H., Okhrin, Y., and Wang, T. (2022). Monitoring network changes in social media. *Journal of Business & Economic Statistics*, 1–16.

Chen, J. and Gupta, A. K. (1997) Testing and Locating Variance Changepoints with Application to Stock Prices. *Journal of the American Statistical Association*, **92**, 739–747

Chen, J., and Gupta, A. K. (2000), *Parametric Statistical Change Point Analysis*, Birkhauser, New York.

Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, **43**, 139–176.

Cho, H. (2016) Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.*, **10**, 2000-2038.

Cho, H. and Fryzlewicz, P. (2015) Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B*, **77**, 475–507.

Csörgö, M. and Horváth , L. (1997), *Limit Theorems in Change-Point Analysis.* John Wiley and Sons, New York.

Davis, C. and Kahan, W. M. (1970) The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, **7**, 1–46.

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**, 444–454.

Enikeeva, F. and Harchaoui, Z. (2019) High-dimensional change-point detection under sparse alternatives. *Ann. Statist.*, **47**, 2051–2079.

Fan, K. (1953) Minimax theorems. *Proc. Natl. Acad. Sci. USA*, **39**, 42–47.

Follain, B., Wang, T. and Samworth, R. J. (2022) High-dimensional changepoint estimation with heterogeneous missingness. *J. Roy. Statist. Soc., Ser. B*, **84**, 1023–1055.

Floyd, R. W. (1962) Algorithm 97: shortest path. *Communications of the ACM*, **5**, 345.

Frank, M. and Wolfe, P. (1956) An algorithm for quadratic programming. *Naval Research Logistics Quarterly.* **3 (1–2)**, 95–110.

Frick, K., Munk, A. and Sieling, H. (2014) Multiscale change point inference. *J. R. Stat. Soc. Ser. B*, **76**, 495–580.

Fryzlewicz, P. (2014) Wild binary segmentation for multiple change point detection. *Ann. Statist.*, **42**, 2243–2281.

Gao, C., Han, F., Zhang, C. H. (2020) On estimation of isotonic piecewise constant signals. *Ann. Statist.*, **48**, 629–654.

Hanlon, M. and Anderson, R. (2009). Real-time gait event detection using wearable sensors. *Gait & Posture*, **30**, 523–527.

Horváth, L. and Hušková, M. (2012) Change-point detection in panel data. *J. Time Ser. Anal.*, **33**, 631–648.

Horváth, L. (1993) The Maximum Likelihood Method of Testing Changes in the Parameters of Normal Observations. *The Annals of Statistics*, **21**, 671–680

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classification*, **2**, 193—218.

Ingster, Y. and Suslina, I. A. (2012) *Nonparametric Goodness-of-Fit testing under Gaussian models*. Springer Science & Business Media.

Itoh, N., and Kurths, J. (2010). Change-point detection of climate time series by nonparametric method. *World Congress on Engineering and Computer Science* **1**, 445–448.

Jackson, B., Sargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005), An Algorithm for Optimal Partitioning of Data on an Interval. *IEEE Signal Processing Letters*, **12**, 105–108.

Jiang, F., Zhao, Z., and Shao, X. (2022) Modelling the covid-19 infection trajectory: A piecewise linear quantile trend model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84**:1589–1607.

Jirak, M. (2015) Uniform change point tests in high dimension. *Ann. Statist.*, **43**, 2451–2483.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012) Optimal detection of changepoints with a linear computational cost. *J. Amer. Stat. Assoc.*, **107**, 1590–1598.

Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection.

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**, 1–44.

Mei, Y. (2010) Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, **97**, 419–433.

Moustakides, G. V. (1986) Optimal stopping times for detecting changes in distributions. *Ann. Statist.*, **14**, 1379–1387.

Li, A. and Baeber, R.F(2019) Multiple Testing with the Structure-Adaptive Benjamini–Hochberg Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*,**81**, 45–74.

Liu, H., Gao, C. and R. J. Samworth.(2021) Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.*, **49**, 1081–1112.

Lorden, G. (1971) Procedures for reacting to a change in distribution. *Ann. Math. Statist.*, **42**, 1897–1908.

Massart, P. (2007) *Concentration Inequalities and Model Selection*, Springer, Berlin.

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

Peng, T., Leckie, C. and Ramamohanarao, K. (2004) Proactively detecting distributed denial ofservice attacks using source IP address monitoring. In Mitrou, N., Kontovasilis, K., Rouskas, G. N., Iliadis, I. and Merakos, L. eds, *Networking 2004*, pp. 771–782. Springer-Verlag, Berlin.

Pilliat, E., Carpentier, A. and Verzelen, N. (2020) Optimal multiple change-point detection for high-dimensional data.

Pollak, M. (1985) Optimal detection of a change in distribution. *Ann. Statist.*, **13**, 206–227.

Pollak, M. and Tartakovsky, A. G. (2009) Optimality properties of the Shiryaev–Roberts procedure. *Statist. Sinica*, **19**, 1729–1739.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.*, **66**, 846—850.

Reeves, J., Chen, J., Wang, X. L., Lund, R. B. and Lu, Q. (2007) A Review and Comparison of Changepoint Detection Techniques for Climate Data. *Journal of Applied Meteorology and Climatology*, **46**, 900–915.

Roberts, S. W. (1966) A comparison of some control chart procedures. *Technometrics*, **8**, 411–430.

Scott, A. J., and Knott, M. (1974), A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, **30**, 507–512.

Schwarz, G. (1978), Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461–464.

She, Yiyuan (2010) Sparse regression with exact clustering. *Electronic Journal of Statistics*, **4**, 1055–1096.

Simon N., Jerome Friedman, J., Hastie, T. and Tibshirani, R. (2013) A Sparse-Group Lasso, *Journal of Computational and Graphical Statistics*, **22**, 231–245, Shewhart, W. A. (1931) *Economic Control of Quality of Manufactured Product*, Van Nostrand, New York.

Shiryaev, A. N. (1963) On optimum methods in quickest detection problems. *Theory Probab. Appl.*, **8**, 22–46.

Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, **64**, 479–498.

Tartakovsky, A. G., Rozovskii, B. L., Blažek, R. B. and Kim, H. (2006) Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, **3**, 252–293.

Tibshirani, Robert (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B* . **58**. 267–288.

Tibshirani Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. (2005) Sparsity and Smoothness via the Fused lasso. *Journal of the Royal Statistical Society. Series B*, **67**, 91–108.

Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.) *Compressed Sensing, Theory and Applications.* Cambridge University Press, Cambridge. 210–268.

Wang, T and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., Ser. B*, **80**, 57–83.

Wang, Y., Chakrabarti, A., Sivakoff, D., and Parthasarathy, S. (2017). Fast change point detection on dynamic social networks. arXiv preprint arXiv:1705.07325.

Wang, D., Zhao, Z., Lin, K. Z. and Willett, R. (2021) Statistically and computationally efficient change point localisation in regression settings. *The Journal of Machine Learning Research*, **22**, 11255–11300.

Wang, D., Yu, Y. and Rinaldo, A. (2021) Optimal change point detection and localisation in sparse dynamic networks. *The Annals of Statistics*, **49**, 203–232.

Xie, Y. and Siegmund, D. (2013) Sequential multi-sensor change-point detection. *Ann. Statist.*, **41**, 670–692.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc., Ser. B*, **68**, 49–67.

Yao, Y. (1984), Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches. *The Annals of Statistics*, **12**, 1434–1447.

B. Yu Assouad, Fano, and Le Cam (1997).*Festschrift for Lucien Le Cam, Springer-Verlag* ,423–435

Zhu, Z., Wang, T. and Samworth, R. J. (2022) High-dimensional principal component analysis with heterogeneous missingness. *J. Roy. Statist. Soc., Ser. B*, to appear.