Change-point Detection for Modern Data

By

YI-WEI LIU

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Hao Chen, Chair

Alexander Aue

Jane-Ling Wang

Committee in Charge

2022

To Li Hu, Hsiang Liu and Qin Ding.

Abstract

Change-point detection investigates whether there are abrupt changes in distributions in sequences of observations. The goal is to partition a sequence of observations into homogeneous subsequences, which provides essential screening information for follow-up studies. As we enter the era of big data, it is commonplace to encounter sequences of high-dimensional/non-Euclidean observations. Parametric methods are often limited to those data where the parametric assumptions are reasonable. Nonparametric methods are usually more broadly applicable, but it is often hard to conduct theoretical analysis on them, such as to provide an analytic p-value approximation to facilitate the application to large datasets. The graph-based framework, which utilizes the edge-count information on the similarity graphs constructed on the observations, is the first kind that can be applied to these data with analytic p-value approximates. In this dissertation, we work out three advancements of the graph-based framework to meet the needs for modern data analysis. First, we improve the time efficiency of the algorithms by incorporating the approximate directed k-Nearest Neighbor (k-NN) graphs into the framework. Our new method is many folds faster to run and has power higher than or competitive with state-of-the-art nonparametric methods under various settings. The effectiveness of the new method is illustrated by real applications to fMRI and Neuropixels data sequences. Second, when data are autocorrelated, existing methods that assume independence could result in a higher false discovery rate. Therefore, we use the circular block permutation (CBP) framework that preserves the locally dependent structure among observations. The new framework provides proper controls on the false discovery rate when data have weak serial correlations. Third, we investigate the problem of multiple sequences of high-dimensional/non-Euclidean observations. We propose a new scan statistic that is powerful in detecting changes in all or a subset of the sequences. The new test has much higher power than existing methods and is sensitive to a wide range of alternatives. We illustrate the performance of our new test by applying to the New York Taxi Data over multiple calendar years. For all three new tests, we derive analytic formulas for p-value approximations to make them fast applicable.

Acknowledgments

First of all, I want to express gratitude to my whole family for their support, company and unconditional love - Yin-Hui Wang Liu, my grandmother; Hsiang Liu, my dad; Li Hu, my mom, and Dr. Qin Ding, my fiancée.

The most meaningful and influential person in the journey of my graduate studies must go to Dr. Hao Chen, my Ph.D. advisor in the Department of Statistics at UC Davis. Thank you for providing me with a well-rounded training that makes me a better researcher, educator, and a scientist. I am very grateful for your support and mentorship in all aspects - from conducting research to teaching to professional manners. You are definitely one of the most important persons in both my career and my life.

I also want to say thank you to all the students, staff, and faculty members in the Department of Statistics at UC Davis. Together you has made my Ph.D. journey colorful, fruitful, successful, and unforgettable. Finally, I am aware that there are a lot more people who I owe acknowledgment to. Please indulge me to quote the famous phrase from a Chinese literature - “As there are too many people to thank, I thank the Universe, instead.”

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Change-point detection studies whether there are abrupt changes in distributions in sequences of data observations. Many parametric change-point approaches have been proposed with focuses on various applications. For example, motivated by the problem of detecting recurrent DNA copy number variants in multiple samples, Zhang et al. (2010) studies the changes in mean vector for multivariate Gaussian observations with identity covariance matrix; assuming the changes are in the second-order structure, Barigozzi et al. (2018) used the piecewise stationary time series factor models for multivariate observations; Wang et al. (2018) investigated the change-point detection and localization problem in dynamic networks under the assumption that the entries of the adjacency matrices are from inhomogeneous Bernoulli models. These methods work under certain parametric models. However, as we enter the era of big data, there are many challenging tasks that require change-point analysis for large and complex datasets, such as the authorship debate of *Tirant lo Blanc* (Girón et al., 2005), fMRI sequences analysis (Visconti di Oleggio Castello et al., 2020), the study of brain activities with Neuropixels recordings (Jun et al., 2017), etc.

For the aforementioned and many more modern data analysis problems, the tasks usually involve sequences of high-dimensional/non-Euclidean data observations, and there are no universal parametric models that tackle all these problems. In the realm of nonparametric change-point detection, methods based on various frameworks were proposed, such as kernel methods (Harchaoui and Cappé, 2007; Harchaoui et al., 2009; Arlot et al., 2019), distance-based methods (Matteson and James, 2014), graph-based methods (Chen and Zhang, 2015; Chu and Chen, 2019), etc. Nevertheless, many of the existing nonparametric methods could be very slow to run when either the data dimensionality is high or the sequence is long (Liu and Chen, 2022). In Chapter 2, we improve the time efficiency upon the graph-based methods by utilizing the directed approximate k-Nearest Neighbor information. Our new method is much faster than the fastest

state-of-the-art methods, while having power better than or on par with its competitors under a variety of settings.

Most of the graph-based methods assume that the observations are sampled independently (Chen and Zhang, 2015; Chu and Chen, 2019; Liu and Chen, 2022). However, when the observations are autocorrelated, which is usually the case for many real applications, methods that assume independence could result in a higher false discovery rate (Chen, 2019a). To tackle this problem, the circular block permutation (CBP) framework is proposed in Chen (2019a) to approximate the distribution of the test statistic under the null hypothesis. Chen (2019a) worked out the analytic expression of the original edge-count test statistic under CBP. However, the procedure depends on the edge-count two-sample test, and could lead to biased estimates of the location of the change-point for some types of changes and low efficiency when the change-point is away from the center of the sequence. In Chapter 3, we further extend the three other edge-count statistics (weighted/generalized/max-type) to the CBP framework. In particular, we find that a new optimal weight function should be adopted in the weighted edge-count test statistic under CBP, and the construction of the generalized/max-type edge-count statistics should hence be modified accordingly. The modified tests under CBP have proper type I error control for autocorrelated data, and the weighted edge-count test after modification exhibits power higher than the one in Chu and Chen (2019) under CBP through simulation studies.

In many applications, researchers may also be interested in detecting simultaneous change-points in multiple sequences of observations. A parametric method was proposed in Zhang et al. (2010), with a focus on detecting common shifts in mean in multiple sequences of univariate Gaussian variables. However, there were few literatures that explore this area over the past decade, whereas the demand for change-point analysis in multiple sequences of high-dimensional/non-Euclidean observations has increased. For example, the dataset in Visconti di Oleggio Castello et al. (2020) consists of the fMRI sequences from 25 patients with each of them watching six selected pieces of the movie "The Grand Budapest Hotel" by Wes Anderson. In Chapter 4, we design a new MS-statistic that accumulates signals from each of the sequences. The new test can be applied to high-dimensional/non-Euclidean data, and is powerful in detecting various types of changes. We also derive the analytic formulas to approximate the p-value, making the new test fast-applicable to large datasets.

In this thesis, we discuss in detail the aforementioned three evolutions of the nonparametric graph-based change-point detection methods. The three versions serve as better solutions to the demand of modern change-point analysis for large datasets with complex data structure. The innovations include improving the time efficiency of the algorithms, controlling type I error for locally dependent data, and accomplishing change-point analysis for multiple sequences of high-dimensional/non-Euclidean data.

1.2 Thesis Outline and Contributions

Build upon the discussions in Section 1.1, this thesis is arranged as follows:

In Chapter 2 - **"A Fast and Efficient Change-point Detection Framework based on Approximate k-NN Graphs,"** we propose a new approach making use of the approximate k-nearest neighbor information from the observations. We derive an analytic formula to control the type I error. The time complexity of our proposed method is $O\left(dn(\log n + k \log d) + nk^2\right)$ for an n-length sequence of d-dimensional data. The test statistic we consider incorporates a useful pattern for moderate- to high- dimensional data so that the proposed method could detect various types of changes in the sequence. The new approach is also asymptotic distribution free, facilitating its usage for a broader community. We apply our method to fMRI and Neuropixels data sequences to illustrate its effectiveness.

In Chapter 3 - **"Graph-based Change-point Detection for Locally Dependent Data,"** we study the circular block permutation framework combined with weighted/generalized/max-type edge-count test statistics to resolve the issue of the original edge-count test. To handle the difficulties caused by the circular block permutation, we propose new edge-count test statistics and provide theoretical treatments to study the asymptotic properties of these new tests, which further leads to analytic formulas to control the family-wise error rates, making them easy to be applied to large datasets. These new tests outperform the existing tests in various ways as reflected by extensive simulation studies.

In Chapter 4 - **"Change-point Detection in Multiple Sequences of High-dimensional/non-Euclidean Data,"** we study the change-point detection problem in the presence of multiple sequences. We propose a new nonparametric method under the graph-based framework, called the MS-statistic, which can be applied efficiently to high-dimensional/non-Euclidean sequences of observations with a proper control on the type I error. The MS-statistic utilizes the edge-counts information from the similarity graphs for each of the sequences, and is useful in detecting various types of changes in multiple sequences. In particular, the types of changes could be different in different sequences. To approximate the p-values of our test, we derive an analytical formula that is asymptotically distribution-free, making our method fast-applicable to large datasets. Simulation studies show that our new test has significant higher power than existing methods when applied to multiple sequences. The performance and effectiveness of our new method is illustrated by a real data application of the NYC taxi data.

Finally, we conclude the thesis and discuss some future plans and endeavors in Chapter 5. Besides, Chapter A, B, and C are appendices to Chapter 2, 3, and 4, respectively.

Chapter 2

A Fast and Efficient Change-point Detection Framework based on Approximate k-NN Graphs

2.1 Introduction

With advances in technologies, scientists in many fields are collecting massive data for studying complex phenomena over time and/or space. Such data often involve sequences of high-dimensional measurements that cannot be analyzed through traditional approaches. Insights on such data often come from segmentation/change-point analysis, which divides the sequence into homogeneous temporal or spatial segments. They are crucial early steps in understanding the data and in detecting anomalous events. Change-point analysis has been extensively studied for univariate and low-dimensional data (see Basseville et al. (1993); Brodsky and Darkhovsky (1993); Carlstein et al. (1994); Csörgö et al. (1997); Chen and Gupta (2011) for various aspects of classic change-point analysis). However, many modern applications require effective and fast change-point detection for high-dimensional data. For example, Neuropixels recordings Jun et al. (2017), microarrays Zeebaree et al. (2018a), healthcare data Lee et al. (2017), etc.

Let the sequence of observations be $\{\mathbf{y}_t : t = 1, \ldots, n\}$, indexed by some meaningful order, such as time or location. Then the change-point detection problem can be formulated as testing the null hypothesis of homogeneity:

$$H_0 : \mathbf{y}_t \sim F_0,\, t = 1, \ldots, n, \tag{2.1}$$

against the alternative that there exists a change-point τ:

$$H_1 : \exists 1 \leq \tau < n,\, \mathbf{y}_t \sim \begin{cases} F_0,\ t \leq \tau \\ F_1,\ \text{otherwise} \end{cases} \tag{2.2}$$

Here, F_0 and F_1 are two different probability measures. When there are multiple change-points, wild binary segmentation Fryzlewicz (2020); Fryzlewicz et al. (2014) or seeded binary segmentation Kovács et al. (2020) can be incorporated.

Recently, there were quite a number of progresses on parametric change-point detection. For example, Barigozzi et al. (2018) used the piecewise stationary time series factor models for multivariate observations, and assumed the changes are in their second-order structure. Reference Wang et al. (2018) studied the change-point detection and localization problem in dynamic networks by assuming the entries of the adjacency matrices are from inhomogeneous Bernoulli models. Reference Bhattacharjee et al. (2018) considered the change-point detection problem in networks generated by a dynamic stochastic block model mechanism. Reference Londschien et al. (2021) addressed the change-point problem with missing values, and focused on detecting the covariance structure breaks in Gaussian graphical models. These methods work under certain parametric models. However, in many applications, we have little knowledge on F_0 and F_1.

In the context of nonparametric change-point detection for high-dimensional data, kernel-based methods were first explored Harchaoui and Cappé (2007); Harchaoui et al. (2009), and continued to be improved Arlot et al. (2019). However, this kernel approach is difficult to use practically. As we will show in Section 2.4, this method is very sensitive to the choice of a tuning parameter and it is time-consuming to apply the method with a proper type I error control, where type I error is the event that a change-point is falsely detected when the sequence is actually homogeneous. Reference Li et al. (2019) proposed the scan B-statistic for kernel change-point detection, which is computationally efficient and has a fast formula for type I error control; however, it requires a large amount of reference data. In recent years, distance-based methods Matteson and James (2014) and graph-based methods Chen and Zhang (2015); Chu and Chen (2019) were proposed for high-dimensional change-point detection. The distance-based method (ecp) uses all pairwise distances among observations to find change-points, which could also be computationally heavy for large datasets because computing all pairwise distances needs $O(dn^2)$ time for d-dimensional data. In addition, there is no fast analytic formula for type I error control, and thus one needs to draw random permutations to approximate the p-value. The graph-based methods Chen and Zhang (2015); Chu and Chen (2019) utilize the information of a similarity graph constructed on observations to detect change-points. The authors also provided analytic formulas for type I error control, making them faster to run. In addition, the graph-based methods can detect more types of changes compared to

ecp. The ecp method is very sensitive to changes in mean but its performance decays when the changes come in many other forms (see Section 2.4.5).

In this work, we seek further improvement on graph-based methods, especially from an efficiency perspective. Existing graph-based methods for offline change-point detection utilize an undirected graph constructed among observations. Some common choices are the minimum spanning tree (MST), where all observations are connected with the total distance minimized; the minimum distance pairing (MDP), where the observations are partitioned into $n/2$ pairs with the total within-pair distance minimized; the undirected nearest neighbor (NN) graph, where each observation connects to its nearest neighbor; and their denser versions, k-MST, k-MDP, and undirected k-NN graphs. Take the k-MST for example, it is the union of the 1st, ..., kth MSTs, where the 1st MST is the MST, and the jth MST is a spanning tree connecting all observations such that the sum of the edges in the tree is minimized under the constraint that it does not contain any edge in the 1st, ..., $(j-1)$th MSTs. Among these graphs, k-MST is preferred as it in general has a higher power than others Chen and Zhang (2015). Nevertheless, it requires $O(dn^2)$ time to compute the distance matrix among n d-dimensional observations, so it takes at least $O(dn^2)$ time to construct the k-MST from the original data when the pairwise distances were not provided in the beginning, which is usually the case. This could be inefficient when either n or d is large.

Hence, we seek other ways to construct the similarity graph. There are fast existing algorithms to construct the directed approximate k-NN graph Beygelzimer et al. (2019), where each observation finds k other nearby points that might not be the k closest ones. We use the kd-tree algorithm to search for approximate nearest neighbors. A kd-tree is a space-partitioning data structure for organizing points in a high-dimensional space, which is a binary tree constructed through splitting the points by the values on alternating coordiantes as the tree grows. It takes $O(dn \log n)$ time to preprocess a set of n points in $\mathbb{R}^d$ Arya et al. (1998). The nearest neighbor for any given query point can be searched efficiently with the kd-tree. To approximate the nearest neighbors, first traverse the tree to the leaf node that contains the query point, and then search for the nearest neighbors only in nearby areas. It requires only $O(d \log d + \log n)$ time to result in a good approximate nearest neighbor per query Ram and Sinha (2019), so the total computational cost for obtaining a directed approximate k-NN graph with the kd-tree can be achieved at $O\left(dn(\log n + k \log d)\right)$. Simulation studies show that this new approach has power on par with the existing method on k-MST (Section 2.4.5), and the new approach is much faster (Section 2.4.2).

Since the existing offline graph-based change-point detection framework needs the graph to be an undirected graph, we further work out a framework that can deal with the directed approximate k-NN graph, i.e., all the following steps after the graph is constructed: the exact analytic formulas to compute the test statistic, the limiting distribution of the new statistic, and the analytic formula to supervise the false discovery rate efficiently. The time complexity of the method after the directed approximate k-NN graph is obtained is $O(nk^2)$. Thus, the overall time complexity of the

new method is $O\left(dn(\log n + k\log d) + nk^2\right)$. We illustrate the new approach on the analyses of fMRI datasets and Neuropixels datasets (Section 2.5). The former ones have very large dimensions and moderate sample sizes, whereas the latter ones feature very large sample sizes with moderate dimensions.

2.2 Proposed Statistic

Let $G = \{(i,j) : \mathbf{y}_j \text{ is among } \mathbf{y}_i\text{'s } k \text{ approximate nearest neighbors}\}$ be the directed approximate k-NN graph, $R_{G,1}(t)$ be the number of edges on G connecting observations both before t, and $R_{G,2}(t)$ be the number of edges on G connecting observations both after t:

$$R_{G,1}(t) = \sum_{(i,j)\in G} \mathbb{1}_{\{i\le t, j\le t\}},\ R_{G,2}(t) = \sum_{(i,j)\in G} \mathbb{1}_{\{i>t, j>t\}};$$

with $\mathbb{1}_A$ being the indicator function for event A. Here, we use the notations similar to those in Chu and Chen (2019). A key difference is that the graph in Chu and Chen (2019) is undirected, whereas here G is directed. Fig. 2.1 illustrates the computation of $R_{G,1}(t)$ and $R_{G,2}(t)$ on a toy example. We will use these two quantities to construct the test statistics. The rationale is as follows: When all observations are from the same distribution, the distributions of $R_{G,1}(t)$ and $R_{G,2}(t)$ can be figured out under the permutation null distribution that places $1/n!$ probability on each of the $n!$ permutations of $\{\mathbf{y}_i : i = 1,\ldots,n\}$. With no further specification, we use P, E, Var, and Cov to denote probability, expectation, variance, and covariance, respectively, under the permutation null distribution. When there is a change-point at τ, one typical outcome is that observations from the same distribution tend to form edges within themselves, making both $R_{G,1}(\tau)$ and $R_{G,2}(\tau)$ larger than their null expectations. Another common but somewhat counter-intuitive outcome is that observations from one distribution tend to connect within themselves, but observations from the other distribution tend not to connect within themselves, casuing one of $R_{G,1}(\tau)$ and $R_{G,2}(\tau)$ to be larger than its null expectation, and the other smaller than its null expectation. This happens commonly under moderate to high dimensions when the variances of the two distributions differ. The underlying reason is the curse of dimensionality (see Chen and Friedman (2017) for detailed explanations on this phenomenon under the two-sample testing setting).

To cover both possible outcomes under the alternative, we focus on a max-type test statistic in the main context. Three other test statistics (original/weighted/generalized) are discussed in Appendix A.6.

For each candidate t of the true change-point τ, the max-type edge-count statistic is defined as

$$M(t) = \max(Z_w(t), |Z_{\text{diff}}(t)|); \tag{2.3}$$

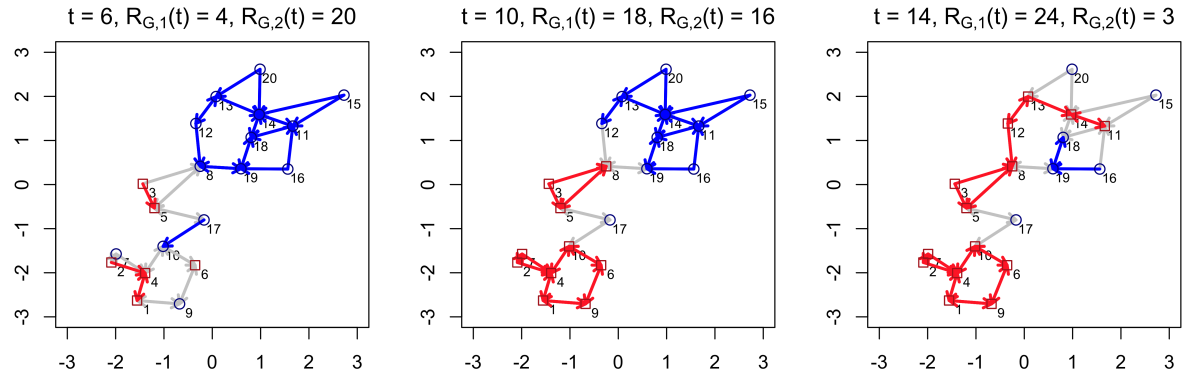


Figure 2.1: The computation of $R_{G,1}(t)$ and $R_{G,2}(t)$ at three different values of t. Here $\mathbf{y}_1, \ldots, \mathbf{y}_{10} \overset{\text{i.i.d.}}{\sim} N((-0.5, -0.5)^T, \mathbb{I}_2)$, and $\mathbf{y}_{11}, \ldots, \mathbf{y}_{20} \overset{\text{i.i.d.}}{\sim} N((0.5, 0.5)^T, \mathbb{I}_2)$, where $\mathbb{I}_2$ is the 2×2 identity matrix. The graph G here is the directed 2-NN on the Euclidean distance. Each t divides the observations into two groups: one group for observations before t (red squares) and the other group for observations after t (blue circles). Red edges connect observations before t and the number of red edges is $R_{G,1}(t)$; blue edges connect observations after t and the number of blue edges is $R_{G,2}(t)$. Notice that as t changes, the group identities change but the graph G does not change.

where

$$\begin{aligned} Z_w(t) &= \frac{R_w(t) - \mathsf{E}(R_w(t))}{\sqrt{\mathsf{Var}(R_w(t))}}, \\ Z_{\text{diff}}(t) &= \frac{R_{\text{diff}}(t) - \mathsf{E}(R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}(R_{\text{diff}}(t))}}; \end{aligned}$$

with

$$\begin{aligned} R_w(t) &= \frac{n-t-1}{n-2} R_{G,1}(t) + \frac{t-1}{n-2} R_{G,2}(t), \\ R_{\text{diff}}(t) &= R_{G,1}(t) - R_{G,2}(t). \end{aligned}$$

The null hypothesis of homogeneity (2.1) is rejected if the test statistic

$$\max_{n_0 \le t \le n_1} M(t); \tag{2.4}$$

with n_0 and n_1 pre-specified, is larger than the critical value for a given significance level, which measures the strength of the evidence that must be presented in the sample to reject the null hypothesis, and has to be determined before conducting the experiment. Statistically, significance level is the probability of rejecting the null hypothesis when it is true, usually set to be 5% or 1%.

Here, the two components, $Z_w(t)$ and $|Z_{\text{diff}}(t)|$, capture the aforementioned two possible outcomes under the

alternative. For better understanding, we illustrate the outcomes through a toy example (Fig. 2.2). When $\{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$ and $\{\mathbf{y}_{t+1}, \ldots, \mathbf{y}_n\}$ are from the same distribution, they are well mixed and $R_{G,1}(t)$ and $R_{G,2}(t)$ would be close to their null expectations (Fig. 2.2 (a)). When they are from different distributions, one common exhibition is that observations from the same distribution are more likely to be connected in G. Fig. 2.2 (b) plots a typical directed 2-NN graph under this alternative and we see that there are more edges connecting within each group. When this happens, $Z_w(t)$ is large. For moderate- to high- dimensional data, another exhibition of the graph is common under the alternative shown in Fig. 2.2 (c). Here, the dimension is $d = 100$, and the blue circles are from a distribution with a larger variance than that of the red squares. We see that $R_{G,1}(t)$ is much larger than its null expectation but $R_{G,2}(t)$ is much smaller than its null expectation (very few blue edges). This happens due to the curse of dimensionality. As the volume of a d-dimensional ball increases exponentially in d, the blue circles from a distribution with a larger variance are sparsely scattered and tend to find their nearest neighbors in red squares. The $Z_{\text{diff}}(t)$ part in our statistic is effective in capturing this pattern. The absolute value is to cover the two possible scenarios in opposite directions showcased in Fig. 2.2 (c) and (d).

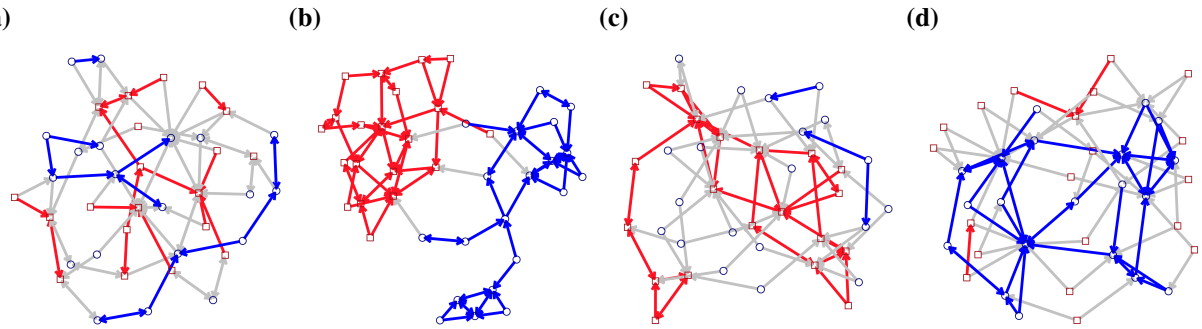


Figure 2.2: Directed 2-NN graphs on 100-dimensional data visualized by the `ggnet2` function. Here, $\mathbf{y}_1, \ldots, \mathbf{y}_{20}$ (red squares) are randomly drawn from a 100-dimensional Gaussian distribution with zero mean and identity covariance matrix, and $\mathbf{y}_{21}, \ldots, \mathbf{y}_{40}$ (blue circles) are randomly drawn from $N_{100}(\boldsymbol{\mu}, a\mathbb{I}_{100})$ with (a) $\boldsymbol{\mu} = 0 \times \mathbf{1}_{100}, a = 1$; (b) $\boldsymbol{\mu} = 0.8 \times \mathbf{1}_{100}, a = 1$; (c) $\boldsymbol{\mu} = 0 \times \mathbf{1}_{100}, a = 1.4$; (d) $\boldsymbol{\mu} = 0 \times \mathbf{1}_{100}, a = 0.8$, where $\mathbf{1}_{100}$ is a length-100 vector whose elements are all one's. The same edge coloring scheme as in Fig. 2.1 is used here.

We next provide the exact analytic formulas for the expectation and variance of $(R_{G,1}(t), R_{G,2}(t))^T$ that are required to compute $M(t)$ so that we do not need to perform the time-consuming permutations to obtain them.

Theorem 1. *The expectation, variance, and covariance of $R_{G,1}(t)$ and $R_{G,2}(t)$ under the permutation null distribution are:*

$$\mathit{E}(R_{G,1}(t)) = nkp_1(t), \quad \mathit{E}(R_{G,2}(t)) = nkq_1(t),$$

$$\mathit{Var}(R_{G,1}(t)) = d_1p_1(t) + d_2p_2(t) + d_3p_3(t) - (nkp_1(t))^2,$$

$$\mathit{Var}(R_{G,2}(t)) = d_1q_1(t) + d_2q_2(t) + d_3q_3(t) - (nkq_1(t))^2,$$

$$Cov\left(R_{G,1}(t), R_{G,2}(t)\right) = d_3 r(t) - \left(nkp_1(t)\right)\left(nkq_1(t)\right),$$

where

$$p_1(t) = \frac{t(t-1)}{n(n-1)}, \quad p_2(t) = \frac{t(t-1)(t-2)}{n(n-1)(n-2)}, \quad p_3(t) = \frac{t(t-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)},$$
$$q_1(t) = \frac{(n-t)(n-t-1)}{n(n-1)}, \quad q_2(t) = \frac{(n-t)(n-t-1)(n-t-2)}{n(n-1)(n-2)},$$
$$q_3(t) = \frac{(n-t)(n-t-1)(n-t-2)(n-t-3)}{n(n-1)(n-2)(n-3)}, \quad r(t) = \frac{t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)},$$

and $d_1 = c^{(1)} + c^{(2)}$, $d_2 = c^{(3)} + c^{(4)} + c^{(5)} + c^{(6)}$, $d_3 = c^{(7)}$, *where* $c^{(1)}, \ldots, c^{(7)}$ *are quantities on the graph G, defined as:*

$$c^{(1)} = nk, \quad c^{(2)} = \sum_{i=1}^{n} \sum_{j \in D_i} \mathbb{1}_{\{(i,j) \in G\}}, \quad c^{(3)} = c^{(4)} = \sum_{i=1}^{n} \sum_{j \in D_i} (k - \mathbb{1}_{\{(i,j) \in G\}}),$$
$$c^{(5)} = nk(k-1), \quad c^{(6)} = \sum_{i=1}^{n} \left(|D_i|^2 - |D_i|\right), \quad c^{(7)} = (nk)^2 - \sum_{m=1}^{6} c^{(m)}.$$

Here, D_i *is the set of indices of observations that point toward observation* $\mathbf{y}_i$*, and* $|D_i|$ *is the cardinality of set* D_i*, or the in-degree of observation* $\mathbf{y}_i$*.*

Remark 1. *The time complexity of computing* $c^{(1)}, \ldots, c^{(7)}$ *in Theorem 1 is* $O(nk)$*. One only has to construct a list of in-degrees for each observation in order to compute these seven quantities, which takes* $O(nk)$ *time.*

Theorem 1 can be proved by combinatorial analysis. The expectations can be obtained easily by the linearity of expectation. For the variances and the covariance, we have to figure out the numbers of the seven possible configurations of pairs of edges as plotted in Fig. 2.3. The quantities $c^{(1)}, \ldots, c^{(7)}$ in Theorem 1 correspond respectively to the numbers of the seven configurations on a directed approximate k-NN graph. For such graphs, the out-degree of every observation node is a constant k, while the in-degree could vary from node to node, which requires one to scan through every edge to obtain the information. A detailed proof of Theorem 1 is in Appendix A.1.

This max-type edge-count statistic $M(t)$ in (2.3) is well-defined under very mild conditions (Theorem 2). The proof is in Appendix A.2.

Theorem 2. *The max-type edge-count statistic* $\{M(t)\}_{t=1,\ldots,n-1}$ *on a directed approximate* k*-NN graph is well-defined when* $n \geq 5$ *and not every observation has the same in-degree (i.e., there exists an* i*,* $1 \leq i \leq n$*, such that* $|D_i| \neq k$*).*

The conditions in Theorem 2 ensure that the variances of $R_w(t)$ and $R_{\text{diff}}(t)$ are not zero. If all the observations have the same in-degree k, then $R_{\text{diff}}(t)$ is a constant. As the distribution of in-degrees could vary, we examine the most

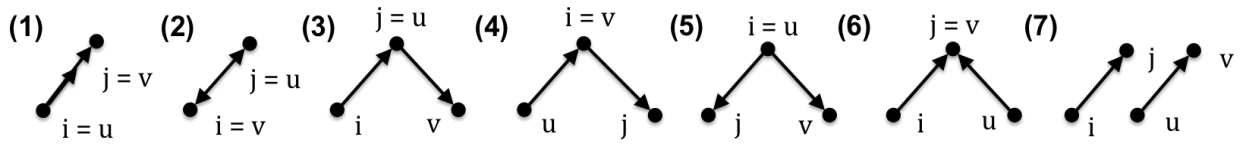


Figure 2.3: Seven possible configurations of two edges $(i, j), (u, v)$ randomly chosen with replacement from a directed graph: (1) two edges degenerate into one ($(i, j) = (u, v)$); (2) two opposite edges (the two end nodes point to each other); (3)-(6) four different configurations with the two edges sharing one node; (7) two edges without any node sharing.

extreme case of which on the directed k-NN graph. Under the worst scenario, $n \geq 5$ ensures that the variance of $R_w(t)$ is positive.

2.3 Analytic type I error control

Given the max-type edge-count statistic, the next question is how large does the critical value need to be to constitute sufficient evidence against the null hypothesis of homogeneity (2.1). This is usually achieved by computing the p-value, which is defined as the probability of observing a more extreme or as extreme test statistic when the null hypothesis is ture. Here, the p-value is defined under the permutation null distribution of the max-type edge-count statisitc. As a relatively large value is the evidence for potential change-points, we are concerned with the tail probability of the test statistics under H_0:

$$\mathsf{P}\Big(\max_{n_0 \leq t \leq n_1} M(t) > b \Big) \tag{2.5}$$

For a small n, the probability (2.5) can be obtained directly by permutation. However, when n is large, doing permutations could be very time-consuming. Hence, we derive analytic formulas to approximate the probability based on the asymptotic properties of the test statistic (2.6). We first work out the limiting distributions of $\{Z_w([nu]^1) : 0 < u < 1\}$ and $\{Z_{\text{diff}}([nu]) : 0 < u < 1\}$ jointly. On a directed graph, we use $e = (e_-, e_+)$ to denote an edge connected from e_- to e_+. Let $A_e = G_{e_-} \cup G_{e_+}$ be the subgraph in G that connect to either node e_- or node e_+, and $B_e = \cup_{e^* \in A_e} A_{e^*}$ be the subgraph in G that connect to any edge in A_e. In the following, we write $a_n = O(b_n)$ when a_n has the same order as b_n, and write $a_n = o(b_n)$ when a_n has order smaller than b_n.

Theorem 3. *For a directed k-NN graph, if $k = O(n^\beta)$, $\beta < 0.25$, $\sum_{e \in G} |A_e||B_e| = o(n^{1.5(\beta+1)})$, $\sum_{e \in G} |A_e|^2 = o(n^{\beta+1.5})$, and $\sum_{i=1}^n |D_i|^2 - k^2 n = O(\sum_{i=1}^n |D_i|^2)$, as $n \to \infty$, $\{Z_w([nu]) : 0 < u < 1\}$ and $\{Z_{\text{diff}}([nu]) : 0 < u < 1\}$ converge to independent Gaussian processes in finite dimensional distributions.*

[1]For a scalar x, we use $[x]$ to denote the largest integer no greater than x.

The covariance functions of the limiting processes $\{Z_w([nu]) : 0 < u < 1\}$ and $\{Z_{\text{diff}}([nu]) : 0 < u < 1\}$ are provided in Appendix A.4.

The complete proof for Theorem 3 is in Appendix A.3. The key idea of the proof is to decouple the dependency resulted from the permutation null distribution and the dependency caused by the graph. More specifically, there are weak dependencies caused by the permutation as one observation appear at one time cannot appear at another time under permutation. To solve this, we take a step back and work on the bootstrap null distribution in which the probability of an observation appearing at one time does not affect by whether it appears at other time(s) or not. Thus, we could focus on dealing with the dependency caused by the graph, and the Stein's method is used to deal with the dependency. The bootstrap null distribution is then connected to the permutation null distribution by conditioning. Based on Theorem 3, the probability (2.5) can be approximated by

$$\begin{aligned} \mathsf{P}\Big(\max_{n_0\le t\le n_1} M(t) > b\Big) &\approx 1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} Z_w(t) < b\Big)\mathsf{P}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff}}(t)| < b\Big) \\ &= 1 - \Big(1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} Z_w(t) > b\Big)\Big) \times \Big(1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff}}(t)| > b\Big)\Big). \end{aligned} \tag{2.6}$$

The two probabilities in (2.6) can be computed similarly as in Chu and Chen (2019):

$$\mathsf{P}\Big(\max_{n_0\le t\le n_1} Z_w(t) > b\Big) \approx b\phi(b)\int_{n_0}^{n_1} S_w(t)C_w(t)\nu\big(\sqrt{2b^2C_w(t)}\big)dt \tag{2.7}$$

$$\mathsf{P}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff}}(t)| > b\Big) \approx 2b\phi(b)\int_{n_0}^{n_1} S_{\text{diff}}(t)C_{\text{diff}}(t)\nu\big(\sqrt{2b^2C_{\text{diff}}(t)}\big)dt \tag{2.8}$$

where the function $\nu(\cdot)$ can be estimated numerically as $\nu(x) \approx \frac{(2/x)(\Phi(x/2)-0.5)}{(x/2)\Phi(x/2)+\phi(x/2)}$ with $\phi(\cdot)$ and $\Phi(\cdot)$ being the probability density function and cumulative distribution function of the standard normal distribution, respectively; $C_w(t)$, $C_{\text{diff}}(t)$ the partial derivative of the covariance function of the process; $S_w(t)$, $S_{\text{diff}}(t)$ the time-dependent skewness correction terms, i.e., for $j = w, \text{diff}$,

$$\begin{aligned} C_j(t) &= \lim_{s\nearrow t}\frac{\partial\rho_j(s,t)}{\partial s}, \quad \rho_j(s,t) = \mathsf{Cov}(Z_j(s), Z_j(t)), \\ S_j(t) &= \frac{\exp\left(\frac{1}{2}(b-\hat\theta_{b,j}(t))^2 + \frac{1}{6}\gamma_j(t)\hat\theta_{b,j}^3(t)\right))}{\sqrt{1+\gamma_j(t)\hat\theta_{b,j}(t)}}; \end{aligned}$$

where

$$\gamma_j(t) = \mathsf{E}\left(Z_j^3(t)\right), \quad \hat\theta_{b,j}(t) = (-1+\sqrt{1+2b\gamma_j(t)})/\gamma_j(t).$$

Moreover, in the above expressions, $C_w(t)$ and $C_{\text{diff}}(t)$ can be derived and simplified to be (details in Appendix A.4)

$$C_w(t) = \frac{n(n-1)(2t^2/n - 2t + 1)}{2t(n-t)(t^2 - nt + n - 1)}, \quad C_{\text{diff}}(t) = \frac{n}{2t(n-t)}.$$

To compute $S_w(t)$ and $S_{\text{diff}}(t)$, we need the third moments of $Z_w(t)$ and $Z_{\text{diff}}(t)$, respectively. Comparing to that under undirected graphs, the computation here is much more complicated. For an undirected graph, there are 8 possible configurations (Fig. 2.4). However, for a directed graph, there are 24 possible configurations (Fig. 2.5). With brute force, it would take $O(|G|^3)$ time to compute the numbers of those configurations for a generic directed graph, which is very computationally expensive. To tackle this problem, we work out efficient formulas that can provide the results in $O(nk^2)$ time for directed k-NN graphs.

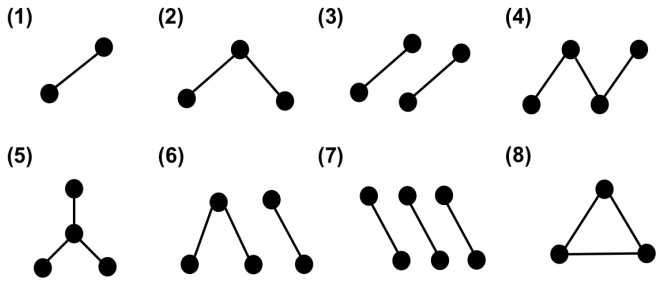


Figure 2.4: Eight possible configurations of three edges randomly chosen with replacement from an undirected graph.

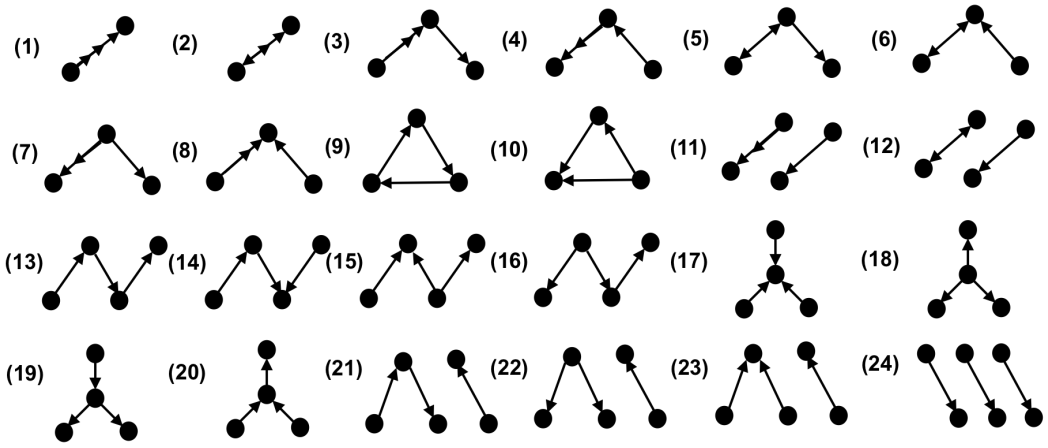


Figure 2.5: Twenty-four possible configurations of three edges randomly chosen with replacement from a directed graph.

Let $G^{(m)}$ be the set of pairs of edges in G having the mth configuration as shown in Fig. 2.3, $m = 1, \ldots, 7$. Let $N^{(l)}$ be the number of occurrence for each of the configurations illustrated in Fig. 2.5, $l = 1, \ldots, 24$, then $\sum_{l=1}^{24} N^{(l)} = |G|^3$. We can obtain $N^{(l)}$'s with effort:

$$N^{(1)} = nk,$$

$$N^{(2)} = 3c^{(2)},$$

$$N^{(3)} = 3c^{(3)},$$

$$N^{(4)} = 3c^{(6)},$$

$$N^{(5)} = 6c^{(2)}(k-1),$$

$$N^{(6)} = 3 \sum_{(i,j),(u,v)\in G^{(2)}} (|D_i| + |D_j| - 2),$$

$$N^{(7)} = 3c^{(5)},$$

$$N^{(8)} = 3c^{(3)},$$

$$N^{(9)} = 2 \sum_{(i,j),(u,v)\in G^{(3)}} \mathbb{1}_{\{(v,i)\in G\}},$$

$$N^{(10)} = 6 \sum_{(i,j),(u,v)\in G^{(3)}} \mathbb{1}_{\{(i,v)\in G\}},$$

$$N^{(11)} = 3c^{(7)},$$

$$N^{(12)} = 3c^{(2)}(nk-2) - (N^{(5)} + N^{(6)}),$$

$$N^{(13)} = 6kc^{(3)} - (N^{(6)} + 3N^{(9)}),$$

$$N^{(14)} = 6 \sum_{(i,j),(u,v)\in G^{(3)}} (|D_v| - 1) - N^{(10)},$$

$$N^{(15)} = 6(k-1)c^{(6)} - N^{(10)},$$

$$N^{(16)} = 6kc^{(5)} - (N^{(5)} + N^{(10)}),$$

$$N^{(17)} = 6 \sum_{i=1}^{n} \binom{|D_i|}{3},$$

$$N^{(18)} = 6n \binom{k}{3},$$

$$N^{(19)} = 3 \sum_{(i,j),(u,v)\in G^{(5)}} |D_i| - N^{(5)},$$

$$N^{(20)} = 3kc^{(6)} - N^{(6)},$$

$$N^{(21)} = 6c^{(3)}(nk-2) - \Big(N^{(5)} + N^{(6)} + N^{(10)} + N^{(14)} + N^{(16)} + 3N^{(9)} + 2N^{(13)} + 2N^{(19)} + 2N^{(20)}\Big),$$

$$N^{(22)} = 3c^{(5)}(nk-2)$$

$$
\begin{aligned}
&\quad - \left(N^{(5)} + N^{(10)} + N^{(15)} + N^{(16)} + N^{(19)} + 3N^{(18)} \right), \\
N^{(23)} &= 3c^{(6)}(nk-2) \\
&\quad - \left(N^{(6)} + N^{(10)} + N^{(14)} + N^{(15)} + N^{(20)} + 3N^{(17)} \right), \\
N^{(24)} &= (nk)^3 - \sum_{l=1}^{23} N^{(l)}.
\end{aligned}
$$

In the above formulas, it takes at most $O(nk^2)$ time to compute $c^{(2)}, c^{(3)}, c^{(5)}$, and $c^{(6)}$, and there are at most nk^2 elements in the sets $G^{(2)}, G^{(3)}$, and $G^{(5)}$, so the numbers of the 24 configurations can be calculated within $O(nk^2)$ time. Indeed, as the rest of the computation is relatively straightforward (see Appendix A.5), the whole analytic p-value approximation procedure for a directed k-NN graph can also be done within $O(nk^2)$.

Now, let's check the performance of (2.6) through simulation studies.

Table 2.1: Critical values for the statistic $\max_{n_0 \le t \le n_1} M(t)$ based on 3-NN's graph at $\alpha = 0.05$.

		Critical Values							
		$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	d	Ana	Per	Ana	Per	Ana	Per	Ana	Per
(C1)	10	3.26	3.26	3.31	3.35	3.39	3.43	3.52	3.60
	100	3.29	3.29	3.36	3.40	3.45	3.52	3.62	3.78
	1,000	3.31	3.31	3.40	3.42	3.51	3.60	3.70	3.94
(C2)	10	3.26	3.26	3.32	3.34	3.39	3.44	3.51	3.60
	100	3.30	3.30	3.35	3.39	3.44	3.51	3.60	3.79
	1,000	3.30	3.30	3.46	3.54	3.59	3.75	3.80	4.27
(C3)	10	3.27	3.28	3.32	3.33	3.40	3.41	3.52	3.58
	100	3.28	3.28	3.35	3.37	3.43	3.48	3.58	3.70
	1,000	3.36	3.41	3.45	3.58	3.58	3.81	3.72	4.07

Table 2.1 shows the performance of the asymptotic p-value approximation of the max-type edge-count statistic (2.6) under different settings. We examine three different distributions (multivariate Gaussian (C1), multivariate t_5 (C2), and multivariate log-normal (C3) distributions) with different data dimensions ($d = 10, 100, 1000$). In Table 2.1, column "Per" is the critical value obtained from doing $10,000$ permutations. This can be deemed as close to the true critical value. Column "Ana" presents the analytical critical values given by plugging (2.6) with (2.7) and (2.8). Here, the length of the sequence is $n = 1000$, and we present the results in four different choices of n_0 with $n_1 = n - n_0$. When $n_0 = 100$ or 75, the analytical approximation works quite well across all distributions and dimensions. As n_0 decreases ($n_0 = 50$ or 25), the analytical critical values become less precise. This is expected as the asymptotic distribution needs both groups to have $O(n)$ observations. When n_0 is small, one group could give a smaller order of observations than the other group. On the other hand, the accuracy of the analytical critical values is less dependent on the distribution of

the data.

2.4 Numerical results

2.4.1 Simulation setting and notations

We compare our proposed test with three state-of-the-art methods: graph-based method with the max-type edge-count statistic on 5-MST (the recommended setting in Chu and Chen (2019), denoted by "5-MST" in the following), the distance-based method (ecp) Matteson and James (2014), and the kernel-based method Arlot et al. (2019). For our method, we use the kd-tree algorithm Beygelzimer et al. (2019) to approximate the directed 5-NN graph (d-a5NN). In the following, we focus on the Euclidean distance. There are implementations for other Minkowski distances in the `ANN` library (`http://www.cs.umd.edu/~mount/ANN/`). We use the directed approximate 5-NN graph as it contains a similar number of edges to the 5-MST to make the comparison with the existing graph-based method fair. The proposed method is denoted by "New" in the following.

2.4.2 Computational efficiency

First, we compare the computational cost of these methods through 10 simulation runs. In each simulation, the observations are generated i.i.d. from a multivariate Gaussian distribution with dimension $d = 500$. The results are presented in Table 2.2. Among all the four methods, our proposed method is the fastest, whereas the kernel method is the slowest to run. For the graph-based methods, in particular, our proposed method on d-a5NN is more than 5 times faster than the method in Chu and Chen (2019) on 5-MST for $n = 2,000$ or above.

Table 2.2: Runtime comparison: Average time cost in seconds (standard deviation) from 100 simulation runs for each choice of n (10 runs for the cells having average runtime greater than 1k seconds). The environment where the experiments are conduected: CPU: Intel(R) Xeon(R) CPU E5-2690 0 @ 2.90GHz / RAM: DDR3 @ 1600MHz / OS: Scientific Linux 6.10 / 2.6.32 Linux.

n	New	5-MST	ecp	kernel
$1,000$	0.8 (0.01)	5.2 (0.5)	13 (1.8)	683 (17)
$2,000$	2.7 (0.01)	21 (3.2)	52 (6.7)	10,224
$5,000$	17 (0.1)	157 (8.3)	482 (87)	>10,000
$10,000$	76 (1.2)	689 (27)	2,073 (387)	-
$20,000$	321 (3.7)	2,193 (189)	6,528 $(1,276)$	-
$30,000$	726 (5.9)	4,757 (106)	>10,000	-

2.4.3 Empirical size

Here, we check the empirical size of these methods under three significance levels ($\alpha = 0.10, 0.05$, and 0.01). Observations are generated i.i.d. from a d-dimensional multivariate Gaussian distribution with no change-point. The results are summerized in Table 2.3, where the p-value of the graph-based methods is obtained through their corresponding analytical formulas, and those for the ecp are based on 999 random permutations. We only report those for $d = 25$ here, as the results are very similar for other choices of d (see Appendix A.7). Both the graph-based methods and ecp could control the type I error well. However, for the kernel method, there is no direct mean to control type I error. We use the `'kcpa'` function in the R package `'ecp'`. In this function, the empirical size is supervised by a tuning parameter C. The larger the C is, the less likely the null is rejected. Unfortunately, it is not straightforward to link the tuning parameter C with the empirical size. As illustrated in Table 2.4, the relation between C and the empirical size depends heavily on the dimension of the observations.

Table 2.3: Fractions of simulation runs (out of 10,000 simulations) that the null hypothesis is rejected when there is no change-point in the sequence ($n = 1,000$). Graph-based methods and ecp at level α.

Method	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
New	0.100	0.051	0.011
5-MST	0.096	0.051	0.012
ecp	0.098	0.050	0.011

Table 2.4: Fractions of simulation runs (out of 10,000 simulations) that the null hypothesis is rejected when there is no change-point in the sequence ($n = 1,000$). Kernel method with tuning parameter C under three different dimensions.

kernel method	$C = 4.8$	$C = 5.2$	$C = 5.6$
$d = 25$	0.943	0.821	0.631
$d = 30$	0.477	0.265	0.134
$d = 35$	0.094	0.036	0.009

2.4.4 Type II error analysis

To get an idea of the performance of the proposed method, we compare the probability of making the type II error, which is the event that the null hypothesis is not rejected when it is false, for the three methods that could control the type I error under some common parametric families. In particular, we consider six scenarios with each coordinate randomly generated from (1) Chi-square distributions with a change in the degree of freedom, (2) Weibull distribution with a change in the scale parameter while the shape parameter is fixed, (3) & (4) Gamma distributions with a change in one of the two parameters, respectively, while the other parameter is fixed, and (5) & (6) Beta distributions with a change in one of the two parameters, respectively, while the other parameter is fixed. In each simulation run, the length

of the sequence is $n = 1,000$, and the change-point is at a quarter of the sequence $\tau = 250$.

- Setting 1 (Chi-square distribution): $F_0 = \chi^2_{\nu_0}$; $F_1 = \chi^2_{\nu_1}$.
- Setting 2 (Weibull distribution): $F_0 = \text{Weibull}(\lambda_0, k_0)$; $F_1 = \text{Weibull}(\lambda_1, k_0)$. Shape parameter fixed at $k_0 = 1$.
- Setting 3 (Gamma distribution-a): $F_0 = \text{Gamma}(\alpha_0, \beta_0)$; $F_1 = \text{Gamma}(\alpha_1, \beta_0)$. Scale fixed at $\beta_0 = 1$.
- Setting 4 (Gamma distribution-b): $F_0 = \text{Gamma}(\alpha_0, \beta_0)$; $F_1 = \text{Gamma}(\alpha_0, \beta_1)$. Shape fixed at $\alpha_0 = 1$.
- Setting 5 (Beta distribution-a): $F_0 = \text{Beta}(\alpha_0, \beta_0)$; $F_1 = \text{Beta}(\alpha_1, \beta_0)$. Second parameter fixed at $\beta_0 = 0.5$.
- Setting 6 (Beta distribution-b): $F_0 = \text{Beta}(\alpha_0, \beta_0)$; $F_1 = \text{Beta}(\alpha_0, \beta_1)$. First parameter fixed at $\alpha_0 = 0.5$.

Table 2.5: Type II error: Numbers of times (out of 100) the null hypothesis is not rejected under $\alpha = 0.05$ for various data dimensions and sizes of change.

S1: Chi-square (d.f. ν change, $\nu_0 = 3$)

d	25	100	500	1000	2000
ν_1	3.27	3.20	3.15	3.12	3.09
New	**.16**	**.32**	**.13**	.11	.13
5-MST	.19	.37	**.13**	**.10**	**.11**
ecp	.95	.95	.95	.95	.96

S2: Weibull (scale λ change, $\lambda_0 = 1, k_0 = 1$)

d	25	100	500	1000	2000
λ_1	1.8	2.4	3.2	4.8	6.2
New	**.19**	**.17**	**.11**	**.18**	**.23**
5-MST	.24	.19	.13	.19	**.23**
ecp	.93	.97	.94	.93	.97

S3: Gamma (shape change, $\alpha_0 = 1, \beta_0 = 1$)

d	25	100	500	1000	2000
α_1	1.09	1.08	1.05	1.04	1.03
New	**.12**	**.15**	.34	.28	.33
5-MST	.19	.18	**.29**	**.22**	**.28**
ecp	.92	.91	.89	.95	.91

S4: Gamma (scale change, $\alpha_0 = 1, \beta_0 = 1$)

d	25	100	500	1000	2000
β_1	1.050	1.040	1.030	1.025	1.020
New	**.32**	**.29**	.25	.11	.12
5-MST	.36	.36	**.24**	**.08**	**.07**
ecp	.95	.93	.92	.90	.89

S5: Beta (shape 1 change, $\alpha_0 = 0.5, \beta_0 = 0.5$)

d	25	100	500	1000	2000
α_1	0.590	0.550	0.530	0.520	0.512
New	.23	**.19**	**.05**	**.03**	**.19**
5-MST	**.21**	.23	.06	.05	.25
ecp	.97	.96	.94	.96	.95

S6: Beta (shape 2 change, $\alpha_0 = 0.5, \beta_0 = 0.5$)

d	25	100	500	1000	2000
β_1	0.590	0.550	0.530	0.520	0.512
New	.25	**.14**	**.04**	**.05**	**.16**
5-MST	**.24**	.18	.06	.09	.24
ecp	.98	.93	.96	.92	.91

The results are shown in Table 2.5. For each dimension, the alternatives are chosen so that the type II error is not too small to be comparable. We see that the type II error of the new test is on the small end for data from different distribution families.

2.4.5 Power comparison

Here, we compare the power of the proposed method to the other two methods. Power is the probability of rejecting the null hypothesis when it is false, i.e., the probability of not making the type II error. We consider six different

scenarios. They are chosen to cover a variety of change types. Scenarios 1-4 emphasize on the Gaussian distribution and cover changes in mean and variance, as well as different parts of the covariance matrix. Scenarios 5 and 6 cover asymmetric distributions and fat-tailed distributions. In the following, a and b are constants. N_d denotes a d-dimensional multivariate Gaussian distribution, $\mathbf{0}_d$ and $\mathbf{1}_d$ denote length-d vectors of all zeros and one's, respectively, $\mathbb{I}_d$ denotes a $d \times d$ identity matrix, and $\mathbf{\Sigma}$ denotes the covariance matrix with $\mathbf{\Sigma}_{ij} = 0.6^{|i-j|}$, where $\mathbf{\Sigma}_{ij}$ is the element of the ith row and the jth column of $\mathbf{\Sigma}$. The L_2 norm of the mean vector in F_1 is given by $||\Delta||_2$ in Table 2.6. In each simulation run, the length of the sequence is $n = 1,000$, and the change-point is at a quarter of the sequence $\tau = 250$.

Table 2.6: Power comparison: Numbers of times (out of 100) the null hypothesis is rejected under $\alpha = 0.05$ for various data dimensions and sizes of change.

S1: MG (Mean and Variance)

d	25	100	500	1000	2000
$\|\|\Delta\|\|_2$	0.10	0.20	0.45	0.63	0.89
b	1.10	1.06	1.03	1.02	1.02
New	**75**	**76**	**65**	**58**	**83**
5-MST	71	66	60	56	**83**
ecp	4	8	10	13	15

S2: MG (5-coordinate)

d	25	100	500	1000	2000
$\|\|\Delta\|\|_2$	0.20	0.40	0.67	0.63	0.89
b	1.8	2.4	3.2	4.8	6.2
New	**94**	**84**	**63**	64	69
5-MST	92	83	60	**65**	**71**
ecp	17	36	33	22	34

S3: MG (Diagonal)

d	25	100	500	1000	2000
b	1.10	1.06	1.03	1.02	1.02
New	**69**	**78**	87	**90**	**99**
5-MST	60	77	**88**	88	**99**
ecp	3	7	4	2	4

S4: MG (Off-diagonal)

d	25	100	500	1000	2000
ρ	0.53	0.50	0.48	0.47	0.46
New	**74**	**88**	**89**	**88**	**89**
5-MST	68	83	87	87	88
ecp	4	9	8	4	3

S5: Chi-square distribution

d	25	100	500	1000	2000
$\|\|\Delta\|\|_2$	0.05	0.10	0.22	0.32	0.45
b	1.10	1.08	1.05	1.04	1.03
New	**71**	**81**	**85**	**85**	**89**
5-MST	66	80	**85**	83	**89**
ecp	4	10	3	7	2

S6: t-distribution

d	25	100	500	1000	2000
$\|\|\Delta\|\|_2$	0.20	0.40	0.67	0.63	0.89
b	1.14	1.08	1.05	1.04	1.03
New	**85**	**75**	**73**	86	85
5-MST	81	73	70	**87**	**87**
ecp	8	15	18	11	11

- Scenario 1 (MG: mean and variance): $F_0 = N_d(\mathbf{0}_d, \mathbf{\Sigma})$; $F_1 = N_d(a \times \mathbf{1}_d, b\mathbf{\Sigma})$.
- Scenario 2 (MG: 5-coordinate): $F_0 = N_d(\mathbf{0}_d, \mathbb{I}_d)$; $F_1 = N_d((a \times \mathbf{1}_5, \mathbf{0}_{d-5})^T, \text{diag}((b \times \mathbf{1}_5, \mathbf{1}_{d-5})^T))$, where $\text{diag}(u)$ is a diagonal matrix with its diagonal vector u.
- Scenario 3 (MG: Diagonal): $F_0 = N_d(\mathbf{0}_d, \mathbb{I}_d)$; $F_1 = N_d(\mathbf{0}_d, b\mathbb{I}_d)$.
- Scenario 4 (MG: Off-diagonal): $F_0 = N_d(\mathbf{0}_d, \mathbf{\Sigma})$; $F_1 = N_d(\mathbf{0}_d, \mathbf{\Sigma'})$, where $\mathbf{\Sigma'}_{ij} = \rho^{|i-j|}$.
- Scenario 5 (Chi-square distribution): $F_0 = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{u}^{\chi^2_{3,c}}$; $F_1 = (b\mathbf{\Sigma})^{\frac{1}{2}}\mathbf{u}^{\chi^2_{3,c}} + a \times \mathbf{1}_d$. Here, $\mathbf{u}^{\chi^2_{3,c}}$ is a length-d vector with each component i.i.d. from the centered χ^2_3 distribution.

- Scenario 6 (t-distribution): $F_0 = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{u}^{t_5}$; $F_1 = (b\mathbf{\Sigma})^{\frac{1}{2}}\mathbf{u}^{t_5} + a \times \mathbf{1}_d$. Here, $\mathbf{u}^{t_5}$ is a length-d vector with each component i.i.d. from the t_5-distribution.

From Table 2.6, we can see that our proposed test has good power under a wide range of alternatives. It is the best or on par with the best in these simulation studies. In sharp contrast, the ecp method suffers from the curse of dimensionality, and could have low power when the change contain sources other than the mean shift.

2.4.6 Types of changes the new method can detect

Here, we check the types of changes the proposed method can detect. We investigate five types of changes: mean, variance, covariance, skewness and kurtosis. All experiments are conducted under the setting with $n = 1,000$, $\tau = 250$ and $d = 1,000$. Below describes in details the five scenrios:

- Change in mean: Before the change, all coordinates are from independent standard Gaussian distributions. After the change, the L_2-norm of the mean vector is specified by the x-axis in Fig. 2.6. We study the power for changes in all coordinates ($dc = 1,000$), one coordinate ($dc = 1$), and some subsets of the coordinates ($dc = 200, 50, 10$). We can see from Fig. 2.6 that our test has good power to mean change regardless of the number of coordinates that has a mean shift.

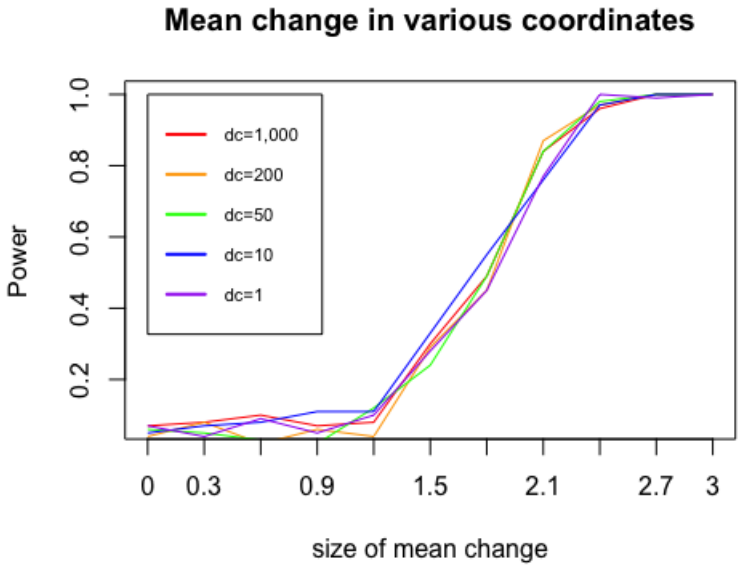


Figure 2.6: Fraction of times (out of 100) that a change-point is detected at a given size of mean change for the changes in various coordinates (dc).

- Change in variance: Before the change, all coordinates are from independent standard Gaussian distributions.

After the change, the determinant of the variance-covariance matrix becomes a^{1000}, where a is specified by the x-axis in Fig. 2.7. We study the power for changes in all coordinates ($dc = 1,000$), one coordinate ($dc = 1$), and some subsets of the coordinates ($dc = 200, 50, 10$). We can see from Fig. 2.7 that our test is generally sensitive to variance change in all cases, and the power is higher when the change comes in fewer coordinates.

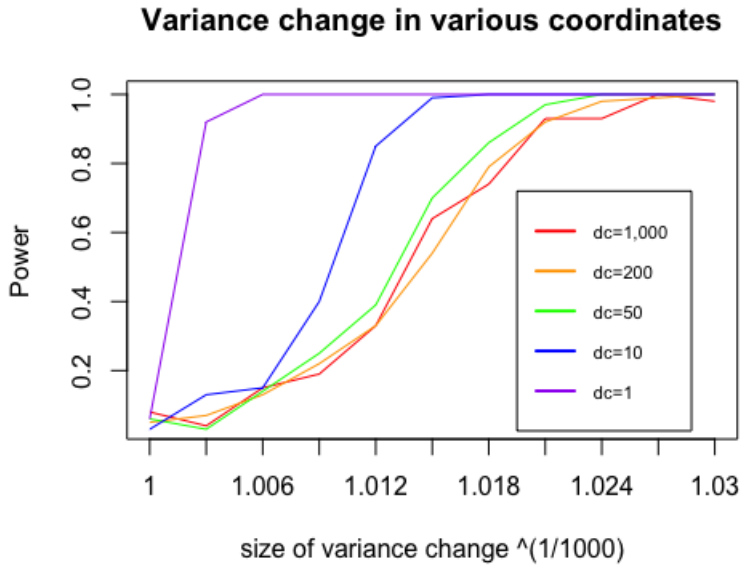


Figure 2.7: Fraction of times (out of 100) that a change-point is detected at a given size of variance change for the changes in various coordinates (dc).

- Change in covariance: Before the change, the observations are from multivariate Gaussian distribution with mean zero and variance-covariance matrix $\mathbf{\Sigma}_{ij} = 0.6^{|i-j|}$, denoted as $\rho_0 = 0.6$. After the change, the variance-covariance matrix becomes $\mathbf{\Sigma}_{ij} = \rho_1^{|i-j|}$ and new correlation coefficent is defined as $\rho_1 = 0.6 - \Delta\rho$. The ten values of the $\Delta\rho$'s used in the experiment are $(0.02, 0.04, \ldots, 0.20)$, corresponding to the x-axis $(1, 2, \ldots, 10)$ in Fig. 2.8. The result shows that our new method is also very sensitive to changes in the covariance structure.

- Change in skewness: Before the change the observations in each coordinate are from independent Gaussian distributions with mean ν and standard deviation $\sqrt{2\nu}$. After the change, observations in each coordinate are from independent Chi-square distributions with degree of freedom ν. The skewness of a χ^2_ν distribution is computed as $\sqrt{8/\nu}$. The ten values of the ν's used in the experiment are chosen so that the skewness of each variable are $(0.2, 0.4, 0.6, \ldots, 2.0)$, corresponding to the x-axis $(1, 2, \ldots, 10)$ in Fig. 2.8. This setting does not change mean and variance while changing the skewness.

- Change in excess kurtosis: Before the change the observations in each coordinate are from independent standard

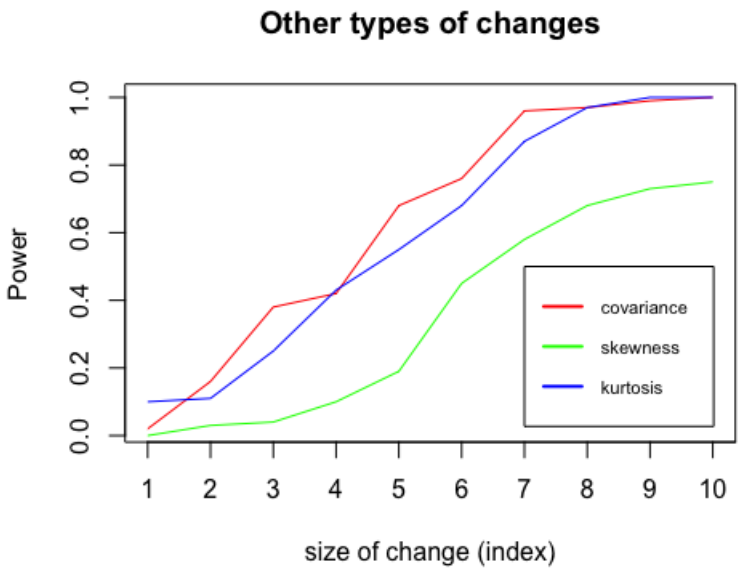


Figure 2.8: Fraction of times (out of 100) that a change-point is detected at a given size of changes in covariance, skewness and excess kurtosis. The sizes of changes increase as the index in the x-axis grows, but the sizes among the three scenarios at each index are not comparable. We plot them on the same figure to save space.

Gaussian distributions. After the change, observations in each coordinate are from independent t distributions with degree of freedom ν. The excess kurtoosis of a t_ν distribution is computed as $\frac{6}{\nu-4}$. The ten values of the ν's used in the experiment are chosen so that the excess kurtosis of each variable are $(0.01, 0.02, 0.03, \ldots, 0.10)$, corresponding to the x-axis $(1, 2, \ldots, 10)$ in Fig. 2.8. The result shows that our method can also detect changes in excess kurtosis.

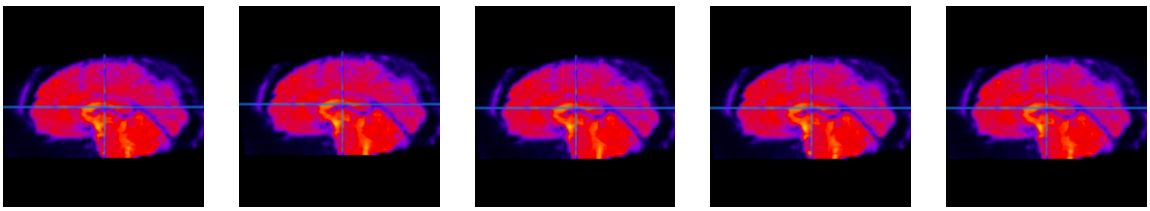
2.5 Real data applications

2.5.1 fMRI data

This fMRI dataset was recorded when the subjects were watching certain pieces of the movie "The Grand Budapest Hotel" by Wes Anderson. It is publicly available at: `https://openneuro.org/datasets/ds003017/versions/1.0.2`. There are in total 25 subjects involved in this experiment, each of them watching 5 pieces of the movie Visconti di Oleggio Castello et al. (2020). Here, we randomly select two such sequences with subject ID SID-000005 and SID-000024 for illustration.

This piece of the movie is about 10 minutes long. The total length of the sequence is $n = 598$ with one time unit as 1 second. Each observation is a 3-dimensional fMRI image with size $96 \times 96 \times 48$. To get an idea of how the fMRI

data look like, Fig. 2.9 shows the profiles of five observations at $t = 150, 250, 350, 450$, and 550, from one certain perspective. There are three different perspectives available, and the other two can be found in Appendix A.8. We rearrange each observation into a length-d vector ($d = 96 \times 96 \times 48 = 442,368$).



$t = 150$ $t = 250$ $t = 350$ $t = 450$ $t = 550$

Figure 2.9: The snapshots of the fMRI images for subject SID-000005 at different timestamps.

In the first sequence (SID-000005), the signal is strong that all three methods find the change-point at around 435 (see Table 2.7). We can see from the heatmap of the pairwise distances of the observations (Fig. 2.10) that the existence of a change-point at around 435 is reasonable. In the second sequence (SID-000024), all three methods find the change-point at around 260, which also appears to be consistent with the heatmap (Fig. 2.11). Among the three methods, the new test is much more efficient to run (Table 2.7, last column). We can also observe that here the computation time for 5-MST is similar to that of ecp as constructing the 5-MST dominates the overall runtime when d is large.

Table 2.7: Results of the estimated change-point locations ($\hat{\tau}$), p-values, and the overall runtimes. For the two graph-based methods, the analytical p-values are reported; for the ecp method, the p-value is based on 999 permutaions.

Subject	Methods	$\hat{\tau}$	p-value	time cost (minutes)
SID-000005	New (d-a5NN)	437	< 0.001	3.8
	5-MST	437	< 0.001	120.9
	ecp	433	0.001	118.4
SID-000024	New (d-a5NN)	260	< 0.001	3.9
	5-MST	260	< 0.001	147.8
	ecp	261	0.001	133.1

2.5.2 Neuropixels data

Neuropixels probes are new technology in neuroscience that can record hundreds of sites in the brain simultaneously Jun et al. (2017); Stringer et al. (2019). Here, we analyze a dataset that records the spiking activities of the neurons in the brain of a mouse while it is awake in darkness during spontaneous behavior. The dataset is publicly available at: `https://figshare.com/articles/dataset/Eight-probe_Neuropixels_recordings_during_spontaneous_behaviors/7739750`. This dataset contains simultaneous recordings from nine brain

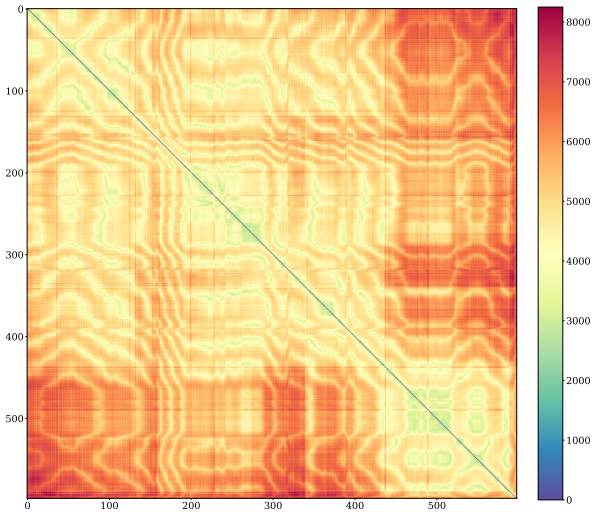


Figure 2.10: Heatmap of pairwise distances of the observations in the sequence (SID-000005).

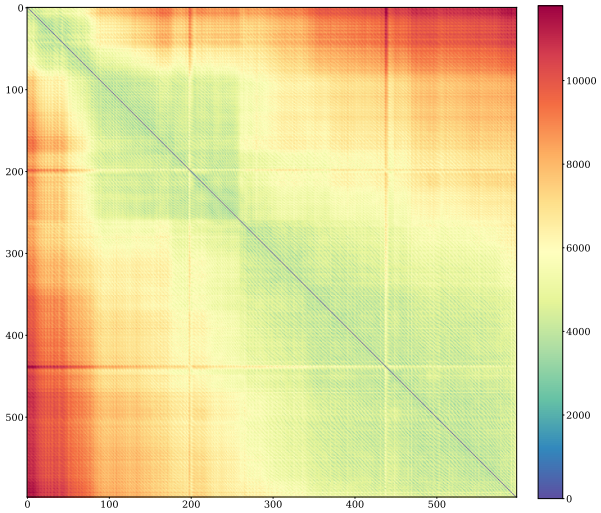


Figure 2.11: Heatmap of pairwise distances of the observations in the sequence (SID-000024).

regions, with each region having hundreds of recording sequences. This dataset was analyzed in Chen et al. (2019), and we follow the same preprocessing procedure there. The lengths of all the sequences are the same $n = 39,053$, and the dimensions are the numbers of recordings which vary from $d = 42$ to $d = 334$. We apply the two graph-based methods to all the nine sueqences. The results are presented in Table 2.8. The ecp method is not applicable to this dataset because the memory space is not enough under the same environment as in Table 2.2.

We see that the new method on the directed approximate 5-NN graph is on average ten times faster than the method in Chu and Chen (2019) on 5-MST. Such improvement can be very imperative especially when analyzing large datasets.

Table 2.8: Results of the estimated change-point locations ($\hat{\tau}$), p-values, and the overall runtimes (in minutes). For the two graph-based methods, the analytical p-values are reported.

Region	Methods	$\hat{\tau}$	p-value	time
Caudate putamen ($d = 176$)	New (d-a5NN)	35,148	< 0.001	7.7
	5-MST	35,056	< 0.001	96.1
Frontal motor ($d = 78$)	New (d-a5NN)	31,081	< 0.001	6.0
	5-MST	32,242	< 0.001	77.8
Hippocampus ($d = 265$)	New (d-a5NN)	4,109	< 0.001	20.7
	5-MST	4,382	< 0.001	159.1
Lateral septum ($d = 122$)	New (d-a5NN)	29,616	< 0.001	11.4
	5-MST	29,636	< 0.001	89.3
Midbrain ($d = 127$)	New (d-a5NN)	20,580	< 0.001	13.9
	5-MST	20,590	< 0.001	105.6
Superior colliculus ($d = 42$)	New (d-a5NN)	23,539	< 0.001	4.0
	5-MST	31,328	< 0.001	65.4
Somatomotor ($d = 91$)	New (d-a5NN)	30,316	< 0.001	7.6
	5-MST	30,312	< 0.001	81.9
Thalamus ($d = 227$)	New (d-a5NN)	28,613	< 0.001	21.7
	5-MST	28,608	< 0.001	146.1
V1 ($d = 334$)	New (d-a5NN)	30,226	< 0.001	17.5
	5-MST	30,338	< 0.001	173.8

2.6 Conclusion

As we enter the era of big data, the importance of the scalibility of statistical methods or data analysis teachniques cannot be overemphasized. Nowadays we are collecting data with exploding sizes (either the dimensionality gets higher or the sequence of observations gets longer). To address this problem, we propose a new nonparametric framework for change-point analysis using the information of approximate k-NN graphs. The time complexity of performing our proposed test is $O\left(dn(\log n + k \log d) + nk^2\right)$, and our method is so far the fastest change-point detection method available with a proper control on the false discovery rate.

In constructing the test statistic, we take into account a pattern caused by the curse of dimensionality. As a result, the new test can detect various types of changes (such as change in mean and/or variance, and change in the covariance structure) in long sequences of moderate- to high- dimensional data. Moreover, our method does not impose any distributional assumption on the data, making it desirable in many real applications where the distribution could be heavy-tailed and/or skewed, the dimension of the data could be much higher than the number of observations, and the change could be global or in a sparse/dense subset of the coordinates.

We apply our method to two large real datasets, the fMRI images and the Neuropixels recordings, with the former having a very large dimension and the latter a very large sample size. Both examples show that our new method improves the computational efficiency upon the existing methods by a significant amount, while its performance remains as reliable as other state-of-the-art methods.

Chapter 3

Graph-based Change-point Detection for Locally Dependent Data

3.1 Introduction

Change-point analysis is regaining attention as we enter the big data era. Massive amount of data are collected in many fields for studying complex phenomena over time and/or space. Such data often involve sequences of high-dimensional or non-Euclidean measurements that cannot be analyzed through traditional approaches. Nowadays it is common that the observation may appear in various types of forms, such as image or network. For example, network data have become increasingly popular as information about e-mail, phone call, and on-line chat records can easily be retrieved. Those information can be used to construct a network of social interactions among individuals (Kossinets and Watts, 2006; Eagle et al., 2009). Image data are also widely collected in many application areas. For instance, in neuroscience, fMRI data are collected for studying brain activities (Kay et al., 2008). Insight on such data often come from segmentation, which divides the sequence into homogeneous temporal or spatial segments. In these data, it is common that there are local dependency along the sequence. For example, social networks and relationships among people lasting over a certain time interval often exhibit serial correlations.

Most change-point analysis assume that observations in the sequences are independet (Zhang et al., 2010; Siegmund et al., 2011). In the field of time series data analysis, most work assume the data to follow a certain parametric model, such as the ARCH and GARCH models are widely used for studying univariate time series data (Bollerslev, 1987, 1988; Akgiray, 1989). For multivariate time series data, there are many generalizations of the original one-dimensional ARCH, GARCH models (Bauwens et al., 2006; Aue et al., 2009). These models are useful for detecting specific types

of changes when the dimension of the data is low. For high-dimensional data, tests based on those parametirc models may have low power or even could not be applied unless strong assumptions are made on the data. Moreover, such parametric models may not work well for detecting more general types of changes, in which case one could refer to non-parametirc methods for change-point detection.

Recently, Chen (2019a) proposed a universal non-parametric framework for dealing with locally dependent data. This framework builds upon an earlier work by Chen and Zhang (2015), in which the authors developed a nonparametric framework for change-point detection for generic data types under the assumption that the observations are independent. This framework is also known as graph-based change-point detection because it utilizes the information on a similarity graph G constructed on observations. When there is local dependence in the data, the method in Chen and Zhang (2015) could result in higher false discovery rates than pre-specified levels. To address this problem, Chen (2019a) proposed to use a new way of permutation - circular block permutation with a random starting point. This new way of permutation retains the local structure and could control type I error correctly when the sequence is locally dependent. The method in Chen (2019a) also retains the same level of power when the observations in the sequence are independent. The author provided a data-driven way to select the blocksize L in circular block permutation. In addition, the author provided analytic formula for controlling type I error, making the approach easy to be applied to large data sets.

In Chen (2019a), the author used the original edge-count two-sample test as the underlying test statistic for change-point detection. It was shown in Chen and Friedman (2017), and Chen et al. (2018) that the edge-count test could be problematic under some common scenarios for high-dimensional data. The issue of the edge-count two-sample test was further studied under the change-point setting in Chu and Chen (2019). In particular, there are two drawbacks of the original edge-count test. First, when the change-point is away from the middle of the sequence, the original edge-count test could have very low power. Second, when the change is not only in mean, the original edge-count test would lead to a biased estimation of change-points. To deal with these problems, Chu and Chen (2019) studied three new scan statistics, the weighted/generalized/max-type edge-count scan statistics, to improve upon the original one. In Chu and Chen (2019), the observations are assumed to be independent, which is insufficient for many applications. However, due the the natural of the circular block permutation (CBP), integrating the circular block permutation framework with the better edge-count test statistics is not straightforward. In Chu and Chen (2019), the generalized edge-count scan statistic was decomposed into two asymptotically independent components: $S(t) = Z^2_{w^0}(t) + Z^2_{\text{diff}}(t)$. In addition, they also showed that $\{Z_{w^0}([nu]) : 0 < u < 1\}$ and $\{Z_{\text{diff}}([nu]) : 0 < u < 1\}$ converge in finite dimensional distributions to two independent Gaussian processes, which made possible the analytic approximation to the p-value of the generalized edge-count test. However, the decomposition in Chu and Chen (2019) no longer holds under CBP, making the follow-up theoretical analysis extremely different.

3.1.1 Our contribution

In this work, we work out a new decomposition for the generalized edge-count scan statistic under CBP. The main result is stated in Theorem 4 below, where $R_{G,1}(t)$ and $R_{G,2}(t)$ are edge-counts obtained from the similarity graph G, and $\mathsf{E}_{\text{CBP}}(\cdot)$ and $\mathsf{Var}_{\text{CBP}}(\cdot)$ are expectation and variance taken under circular block permutation, which are formally defined in Section 3.2.

Theorem 4. *Let $S_{CBP}(t)$ be the generalized edge-count test statistic under CBP, then*

$$S_{CBP}(t) = Z^2_{w,CBP}(t) + Z^2_{diff,CBP}(t)$$

where

$$Z_{w,CBP}(t) = \frac{R_w(t) - \mathsf{E}_{CBP}(R_w(t))}{\sqrt{\mathsf{Var}_{CBP}(R_w(t))}} \quad \textit{with} \quad R_w(t) = q(t)R_{G,1}(t) + p(t)R_{G,2}(t)$$
$$Z_{diff,CBP}(t) = \frac{R_{diff}(t) - \mathsf{E}_{CBP}(R_{diff}(t))}{\sqrt{\mathsf{Var}_{CBP}(R_{diff}(t))}} \quad \textit{with} \quad R_{diff}(t) = R_{G,1}(t) - R_{G,2}(t)$$

with $p(t) = 1 - q(t)$, and $q(t) = c_{G,L}(2t - n) + \frac{1}{2}$ is the weight function whose slope $c_{G,L}$ is given by

$$c_{G,L} = \frac{\frac{1}{2L}\left(\frac{2c_5^{(sub)}|G|}{m^2(m-1)} - \frac{1}{m(m-1)(m-2)}(c_6 + 2c_7 + 2c_9)\right)}{\frac{-4}{m^2}|G|^2 + \frac{1}{m(m-1)}\big(2(c_2 + c_3 + c_5) + 3c_6 + 4(c_4 + c_9)\big) + \frac{1}{m(m-1)(m-2)}\big((7m-8)c_7 + (m-8)c_8\big)}$$

where L is the block size of circular block permutation, $m = n/L$, and $c_1, \cdots, c_9, c_5^{(sub)}$ are defined in Definition 1 in Section 3.3.1.

We further show that $Z_{w,\text{CBP}}(t)$ and $Z_{\text{diff},\text{CBP}}(t)$ are asymptotically independent under mild conditions of the graph. The details are stated in Theorem 7 in Section 3.3.3. Based on the above results, we could derive an analytic formula to approximate the p-value of the generalized edge-count test under CBP, facilitating fast application of the test. We could also define a max-type edge-count test statistic under CBP, based on $Z_{w,\text{CBP}}(t)$ and $Z_{\text{diff},\text{CBP}}(t)$:

$$M_{\text{CBP}}(t) = \max(Z_{w,\text{CBP}}(t), |Z_{\text{diff},\text{CBP}}(t)|).$$

The scan statistics

$$\max_{n_0 \le t \le n_1} S_{\text{CBP}}(t), \quad \max_{n_0 \le t \le n_1} M_{\text{CBP}}(t) \quad (n_0, n_1 \text{ pre-specified})$$

are used to assess the homogeneity of a locally dependent sequence and estimate the location of the change-point if the sequence is deemed to be not homogeneous. They have similar performance while the latter has a more accurate

p-value approximation (details seen in Section 3.5).

The rest of the chapter is organized as follows. In Section 3.2, we define notations and introduce the circular block permutation (CBP) framework. In Section 3.3, we discuss in detail the generalized edge-count scan statistic under CBP. Section 3.4 includes other useful edge-count scan statistics under this new framework. We then study the analytical p-value approximations for those scan statistics in Section 3.5. Section 3.6 uncovers the performances, including type I error control and power comparison, of those new tests with simulation studies. The methods are illustrated in analyzing the NYC taxi data in Section 3.7, and we conclude with discussion in Section 3.8.

3.2 Notations and the CBP framework

Graph-based change-point detection was first proposed in Chen and Zhang (2015). It constructs an undirected similarity graph, such as minimum spanning tree (MST), among obsevations and uses the edge count as test statisitc to detect change-points. The graph-based framework utilizes permutation as the null distribution of test statistic. After the original edge-count scan statistic was studied in Chen and Zhang (2015), Chu and Chen (2019) proposed three new scan statistics: weighted/generalized/max-type edge-count scan statistic to imporve on the original edge-count scan statistic. We breifly review the four edge-count scan statistics in this section.

Let the sequence of observations be $\{\mathbf{y}_i : i = 1, \cdots, n\}$. We use $e = (i, j) \in G$ to denote the edge on a similarity graph G connecting observations $\mathbf{y}_i$ and $\mathbf{y}_j$. (Note that the notation G was used to represent a directed approximate k-NN graph in Chapter 2, but here we use G to denote an undirected similarity graph.) Also, we use $\mathbb{1}_A$ to denote the indicator of event A. Then for every candidate t of the ture change-point τ, we define $R_{G,0}(t)$, $R_{G,1}(t)$, and $R_{G,2}(t)$:

$$R_{G,0}(t) = \sum_{(i,j)\in G} \left(\mathbb{1}_{\{i\le t, j>t\}} + \mathbb{1}_{\{i>t, j\le t\}}\right), \quad R_{G,1}(t) = \sum_{(i,j)\in G} \mathbb{1}_{\{i\le t, j\le t\}}, \quad R_{G,2}(t) = \sum_{(i,j)\in G} \mathbb{1}_{\{i>t, j>t\}}.$$

Here, $R_{G,0}(t)$ is the number of between-group edges, which connects one observation indexed before t and the other after t. On the contrary, $R_{G,1}(t)$ and $R_{G,2}(t)$ are the numbers of within-group edges, which are similar to the edge-count quantities defined in Chapter 2 but here on undirected graphs. The former counts the edges that connect both observations before t and the latter counts the edges that connect both observations after t. In the following, we use $\mathsf{E}_{\mathbb{P}}(\cdot)$ and $\mathsf{Var}_{\mathbb{P}}(\cdot)$ to denote the expectation and variance, respectively, under the permutation null distribution.

- The original edge-count scan statistic (Chen and Zhang, 2015): $\max_{n_0\le t\le n_1} Z_0(t)$

$$Z_0(t) = -\frac{R_{G,0}(t) - \mathsf{E}_{\mathbb{P}}(R_{G,0}(t))}{\sqrt{\mathsf{Var}_{\mathbb{P}}(R_{G,0}(t))}}.$$

- The weighted edge-count scan statistic (Chu and Chen, 2019): $\max_{n_0 \le t \le n_1} Z_{w^0}(t)$

$$Z_{w^0}(t) = \frac{R_{w^0}(t) - \mathsf{E}_{\mathbb{P}}(R_{w^0}(t))}{\sqrt{\mathsf{Var}_{\mathbb{P}}(R_{w^0}(t))}} \quad \text{with} \quad R_{w^0}(t) = \frac{n-t-1}{n-2} R_{G,1}(t) + \frac{t-1}{n-2} R_{G,2}(t).$$

- The generalized edge-count scan statistic (Chu and Chen, 2019): $\max_{n_0 \le t \le n_1} S(t)$

$$S(t) = \Big(R_{G,1}(t) - \mu_1(t), R_{G,2}(t) - \mu_2(t)\Big) \Sigma^{-1}_{\mathbf{R},\mathbb{P}}(t) \begin{pmatrix} R_{G,1}(t) - \mu_1(t) \\ R_{G,2}(t) - \mu_2(t) \end{pmatrix}$$

  where $\mu_1(t)$, $\mu_2(t)$ are the expectations of $R_{G,1}(t)$ and $R_{G,2}(t)$, and $\Sigma_{\mathbf{R},\mathbb{P}}(t)$ is the covariance matrix of $\big(R_{G,1}(t), R_{G,2}(t)\big)$ under permutation. Equivalently, $S(t) = Z^2_{w^0}(t) + Z^2_{\text{diff}}(t)$, where

$$Z_{\text{diff}}(t) = \frac{R_{\text{diff}}(t) - \mathsf{E}_{\mathbb{P}}(R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}_{\mathbb{P}}(R_{\text{diff}}(t))}} \quad \text{with} \quad R_{\text{diff}}(t) = R_{G,1}(t) - R_{G,2}(t).$$

  In addition, $Z_{w^0}(t)$ and $Z_{\text{diff}}(t)$ are asymptotically indpendent.

- The max-type edge-count scan statistic (Chu and Chen, 2019): $\max_{n_0 \le t \le n_1} M(t)$

$$M(t) = \max(Z_{w^0}(t), |Z_{\text{diff}}(t)|)$$

For each of the above four scan statistics, with n_0 and n_1 pre-specified, the null hypothesis is rejected if the scan statistic is greater than a threshold based on a pre-specified significance level.

3.2.1 Circular block permutation

When the sequence of observations are autocorrelated, using the edge-count scan statistics under permutation null could lead to a higher false discovery rate (Chen, 2019a). One reason is that doing permutation could break the locally dependent structure among observations. Therefore, Chen (2019a) proposes to use circular block permutation with a random starting point to generate a pool of sequences that approximate the sample space from which the original sequence under the null hypothesis of no change-point is drawn. In this new framework, observations are assigned into blocks with block size L, and only those blocks are permuted. We use the same recipe in Chen (2019a):

1. Check if the length of the sequence n is a multiple of L. If not, augment the sequence by x pseudo observations so that the length of the augmented sequence is divisible by L. (i.e., $x = L(\lceil n/L \rceil) - n$, with $\lceil n/L \rceil$ being the smallest integer no smaller than n/L.) Those pseudo observations are isolated points on the graph, forming no edge with any other observation.

2. A starting point is randomly selected from the $(n+x)$ observations. We use k_0 to denote the index number of the selected observation. If $k_0 > 1$, then the first $k_0 - 1$ observations are moved to the end of the sequence: $\{\mathbf{y}_{k_0}, \cdots, \mathbf{y}_n, \mathbf{y}_1, \cdots, \mathbf{y}_{k_0-1}\}$.

3. Divide the new sequence into $(n+x)/L$ non-overlapping blocks of size L, starting from the first observation $\mathbf{y}_{k_0}$, and only the $(n+x)/L$ blocks are randomly permuted. Then the outcome is obtained after the permutation.

For simplicity, we use n to denote the length of the augmented seqeunce $(n+x)$ in the following, and use circular block permutation or CBP to denote the above permutation procedures. We use $\mathsf{P}_{\text{CBP}}(\cdot)$, $\mathsf{E}_{\text{CBP}}(\cdot)$, and $\mathsf{Var}_{\text{CBP}}(\cdot)$ to denote the probability, expectation and variance, respectively, under the circular block permutation null distribution.

3.2.2 Edge-count scan statistics under CBP

The following are the edge-count scan statistics (original/weighted/generalized) under the circular block permutation null distribution:

- The original edge-count scan statistic under CBP: $\max_{n_0 \le t \le n_1} Z_{0,\text{CBP}}(t)$

$$Z_{0,\text{CBP}}(t) = -\frac{R_{G,0}(t) - \mathsf{E}_{\text{CBP}}(R_{G,0}(t))}{\sqrt{\mathsf{Var}_{\text{CBP}}(R_{G,0}(t))}}.$$

- The weighted edge-count scan statistic under CBP: $\max_{n_0 \le t \le n_1} Z_{w^0,\text{CBP}}(t)$

$$Z_{w^0,\text{CBP}}(t) = \frac{R_{w^0}(t) - \mathsf{E}_{\text{CBP}}(R_{w^0}(t))}{\sqrt{\mathsf{Var}_{\text{CBP}}(R_{w^0}(t))}}.$$

- The generalized edge-count scan statistic under CBP: $\max_{n_0 \le t \le n_1} S_{\text{CBP}}(t)$

$$S_{\text{CBP}}(t) = \begin{pmatrix} R_{G,1}(t) - \mathsf{E}_{\text{CBP}}(R_{G,1}(t)) \\ R_{G,2}(t) - \mathsf{E}_{\text{CBP}}(R_{G,2}(t)) \end{pmatrix}^T \Sigma^{-1}_{\mathbf{R},\text{CBP}}(t) \begin{pmatrix} R_{G,1}(t) - \mathsf{E}_{\text{CBP}}(R_{G,1}(t)) \\ R_{G,2}(t) - \mathsf{E}_{\text{CBP}}(R_{G,2}(t)) \end{pmatrix} \tag{3.1}$$

  where $\Sigma_{\mathbf{R},\text{CBP}}(t)$ is the covariance matrix of $\big(R_{G,1}(t), R_{G,2}(t)\big)$ under circular block permutation, and

$$Z_{\text{diff,CBP}}(t) = \frac{R_{\text{diff}}(t) - \mathsf{E}_{\text{CBP}}(R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}_{\text{CBP}}(R_{\text{diff}}(t))}}.$$

  However, under CBP with $L > 1$, $S_{\text{CBP}}(t) \neq Z^2_{w^0,\text{CBP}}(t) + Z^2_{\text{diff,CBP}}(t)$. In addition, $Z_{w^0,\text{CBP}}(t)$ and $Z_{\text{diff,CBP}}(t)$ are not asymptotically indpendent. (This can be easily shown using the approach in Appendix C.1 of Chu and Chen (2019) with the covariance matrix of $\big(R_{G,1}(t), R_{G,2}(t)\big)$ being replaced by the results in Theorem 6.)

For each of the above scan statistics, with n_0 and n_1 pre-specified, the null hypothesis is rejected if the scan statistic is greater than a threshold based on a pre-specified significance level.

The analytic expressions for $\mathsf{E}_{\text{CBP}}(R_{G,0}(t))$ and $\mathsf{Var}_{\text{CBP}}(R_{G,0}(t))$ for any pre-specified block size L are derived in Chen (2019a). It has been shown that under CBP, the original edge-count scan statistic has a better control of type I error when the observations are autocorrelated, while maintaining substaintial power when the observations are independent, compared to itself under the permutation null. In the next section, we discuss the other three edge-count scan statistics under circular block permutation.

3.3 Generalized edge-count test under CBP

In this section, we study the generalized edge-count scan statistic under the CBP framework as define in (3.1). In particular, Section 3.3.1 provides analytic expressions for the key components in $S_{\text{CBP}}(t)$ to efficiently compute the test statisitc. Note that $S_{\text{CBP}}(t)$ can no longer be decomposed into $Z^2_{w^0,\text{CBP}}(t) + Z^2_{\text{diff,CBP}}(t)$. We work out a new decomposition of $S_{\text{CBP}}(t)$ in Section 3.3.2, which is essential in deriving the analytic formulas for type I error control of the scan statistic $\max_t S_{\text{CBP}}(t)$.

3.3.1 Analytic expressions for key quantities in $S_{\text{CBP}}(t)$

To compute $S_{\text{CBP}}(t)$ efficiently, we need analytic formulas for $\mathsf{E}_{\text{CBP}}(R_{G,1}(t))$, $\mathsf{E}_{\text{CBP}}(R_{G,2}(t))$, and $\Sigma_{\mathbf{R},\text{CBP}}(t)$, where

$$\Sigma_{\mathbf{R},\text{CBP}}(t) = \begin{pmatrix} \mathsf{Var}_{\text{CBP}}(R_{G,1}(t)) & \mathsf{Cov}_{\text{CBP}}(R_{G,1}(t), R_{G,2}(t)) \\ \mathsf{Cov}_{\text{CBP}}(R_{G,1}(t), R_{G,2}(t)) & \mathsf{Var}_{\text{CBP}}(R_{G,2}(t)) \end{pmatrix}.$$

In other words, we need analytic expressions for the following five quantities:

$$\mathsf{E}_{\text{CBP}}(R_{G,1}(t)), \mathsf{E}_{\text{CBP}}(R_{G,2}(t)), \mathsf{Var}_{\text{CBP}}(R_{G,1}(t)), \mathsf{Var}_{\text{CBP}}(R_{G,2}(t)), \mathsf{Cov}_{\text{CBP}}(R_{G,1}(t), R_{G,2}(t)).$$

First, we discuss how to derive the analytic expressions for $\mathsf{E}_{\text{CBP}}(R_{G,1}(t))$ and $\mathsf{E}_{\text{CBP}}(R_{G,2}(t))$. Let $\pi_{\text{CBP}}(i)$ be the index of observation $\mathbf{y}_i$ after circular block permutation. That is, the original index of observation $\mathbf{y}_i$ is i, and after circular block permutation, new index number $\pi_{\text{CBP}}(i) \in \{1, \cdots, n\}$ is assigned to observation $\mathbf{y}_i$. Also, we define $g_{\pi_{\text{CBP}}(i)}(t) = \mathbb{1}_{\{\pi_{\text{CBP}}(i) > t\}}$, the indicator that the index number of $\mathbf{y}_i$ after circular block permutation is greater than t. Further, we use $e = (i, j)$ to denote the edge connecting observations $\mathbf{y}_i$ and $\mathbf{y}_j$, and let $\delta_{ij} = \min(|i - j|, n - |i - j|)$ be the index difference between $\mathbf{y}_i$ and $\mathbf{y}_j$. In particular, for any given edge $e = (i, j) \in G$, when the block size is L,

we are interested in the following L events: $\{\delta_{ij} = 1\}, \cdots, \{\delta_{ij} = L-1\}$, and $\{\delta_{ij} \geq L\}$.

In the following, we use $[x]$ to denote the largest integer no greater than x, use $(x)_+$ to denote $\max(0, x)$, and use $|\mathcal{A}|$ to denote the number of elements in set $\mathcal{A}$.

Theorem 5. *Let $m = n/L$. For each $t \in \{1, \cdots, n\}$, where t can be written as $t = aL + b$ with $a = [t/L]$ and $b = t - aL$, then*

$$\begin{aligned}
\mathsf{E}_{CBP}(R_{G,1}(t)) &= \sum_{h=1}^{L} p_1(h, a, b)|\mathcal{E}_h|, \\
\mathsf{E}_{CBP}(R_{G,2}(t)) &= \sum_{h=1}^{L} p_2(h, a, b)|\mathcal{E}_h|,
\end{aligned}$$

with

$$\begin{aligned}
p_1(h,a,b) &= (\min(b, L-h) - (b-h)_+)\frac{a(m+a-1)}{n(m-1)} + (b-h)_+\frac{a+1}{n} \\
&\quad +(h-b)_+\frac{a(a-1)}{n(m-1)} + (L-b-h)_+\frac{a}{n} + (b+h-L)_+\frac{a(a+1)}{n(m-1)}, \\
p_2(h,a,b) &= (\min(b, L-h) - (b-h)_+)\frac{(m-a-1)(2m-a-2)}{n(m-1)} + (b-h)_+\frac{m-a-1}{n} \\
&\quad +(h-b)_+\frac{(m-a)(m-a-1)}{n(m-1)} + (L-b-h)_+\frac{m-a}{n} \\
&\quad +(b+h-L)_+\frac{(m-a-1)(m-a-2)}{n(m-1)}, \\
\mathcal{E}_h &= \{(i,j) \in G : \delta_{ij} = h\}, \quad h = 1, \cdots, L-1, \\
\mathcal{E}_L &= \{(i,j) \in G : \delta_{ij} \geq L\}.
\end{aligned}$$

Similar to how $\mathsf{E}_{\text{CBP}}(R_{G,0}(t))$ is derived in Chen (2019a), we partition the edges in G into L categories according to the index difference of the two observations connected by the edge. For each of the L categories, Chen (2019a) studies the probability of having the two nodes in an edge being separated by an index t after CBP. To derive $\mathsf{E}_{\text{CBP}}(R_{G,1}(t))$ and $\mathsf{E}_{\text{CBP}}(R_{G,2}(t))$, we compute the probability that after CBP, both nodes in an edge are placed before t, denoted as $p_1(h, a, b)$, for the former, and the probability that both nodes in an edge are placed after t, denoted as $p_2(h, a, b)$, for the latter. The proof of Theroem 5 is provided in Appendix B.1.

Next we derive the analytic expressions for $\Sigma_{\mathbf{R},\text{CBP}}(t)$. Following Chen (2019a), we only derive $\Sigma_{\mathbf{R},\text{CBP}}(t)$ analytically for those t's that are multiples of L. For other t's that are not divisible by L, we use interpolation (plug-in estimators) to fill in the values. We later compare the variances approximated by the plug-in estimators with those obtained from circular block permutation directly. When t is divisible by L, we further simplify the formulas for the exepctations, $\mathsf{E}_{\text{CBP}}(R_{G,1}(t))$ and $\mathsf{E}_{\text{CBP}}(R_{G,2}(t))$, as stated in Theorem 4 in the following lemma.

Lemma 1. *Let* $c_1^{(sub)} = \frac{1}{L}\sum_{(i,j)\in G}(L-\delta_{ij})_+$, *and* $c_5^{(sub)} = \frac{1}{L}\sum_{(i,j)\in G}\min(\delta_{ij}, L)$, *then for each* $t = aL$, $a \in \{1, \cdots, m-1\}$, *where* $m = n/L$, *we have*

$$\begin{aligned}
\mathsf{E}_{CBP}(R_{G,1}(t)) &= c_1^{(sub)}\frac{a}{m} + c_5^{(sub)}\frac{a(a-1)}{m(m-1)}, \\
\mathsf{E}_{CBP}(R_{G,2}(t)) &= c_1^{(sub)}\frac{m-a}{m} + c_5^{(sub)}\frac{(m-a)(m-a-1)}{m(m-1)}.
\end{aligned}$$

Note that $c_1^{(sub)} + c_5^{(sub)} = |G|$.

Deriving the variances and covariance of $R_{G,1}(t)$ and $R_{G,2}(t)$ under circular block permutation means that for each pair of edges, we have to compute the probability that after CBP, all the nodes involved are placed before or after an index t. In an undirected graph, each pair of edges, denoted by $(i,j),(u,v) \in G$, can be classified into three different configurations as illustrated in Figure 3.1: (a) (i,j) and (u,v) represent the same edge, (b) (i,j) and (u,v) share only one node, denoted as $(i,j),(i,u) \in G$, and (c) (i,j) and (u,v) are separated edges in the graph, denoted as $(i,j),(u,v) \in G$. The notation $(i,j),(u,v) \in G$ means the first edge connecting node $\mathbf{y}_i$ and node $\mathbf{y}_j$, and the second edge connecting node $\mathbf{y}_u$ and node $\mathbf{y}_v$. In addition, under circular block permutation, we have to figure out how the nodes of a pair of edges are blocked. Therefore, we have to consider all the nine possible ways to do the blocking, as illustrated in Figure 3.2. In constrast to Chen (2019a), deriving $\mathsf{Var}_{\text{CBP}}(R_{G,0}(t))$ requires only the probability that the nodes connecting both edges are separated by t under CBP. This probability is positive only when any two nodes of an edge are assigned into different blocks when t is a multiple of L, so there are only three scenarios (scenario 5, 6, 9 in Figure 3.2) to consider in Chen (2019a).

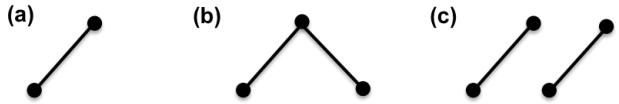


Figure 3.1: Three possible configurations for $(i,j),(u,v) \in G$

Figure 3.2 plots the nine scenarios that the four nodes of a pair of edges $(i,j),(u,v) \in G$ could be blocked into. For simplicity, we plot all the four nodes every time, but the nodes could degenerate into one as long as they are in the same block and not the end points of the same edge. For example, in scenario (1), it could only have two distinct nodes with $i = u$, $j = v$.

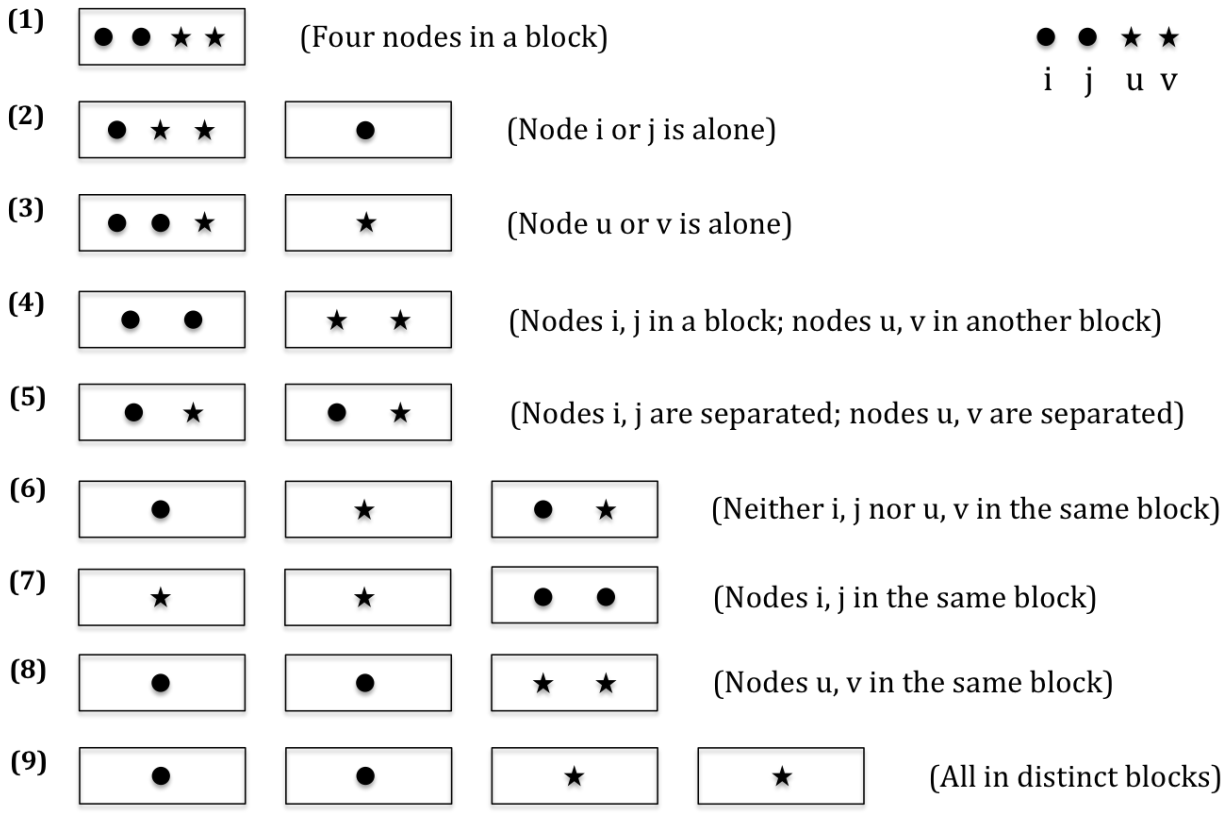


Figure 3.2: Nine scenarios the nodes could be blocked into.

(1) One block: All four nodes are in the same block.

(2) Two blocks: Node u and v are in the same block; node i and j are not.

(3) Two blocks: Node i and j are in the same block; node u and v are not.

(4) Two blocks: Node (i, j) are in one block; node (u, v) are in the other.

(5) Two blocks: Neither (i, j) nor (u, v) is in the same block.

(6) Three blocks: Neither (i, j) nor (u, v) is in the same block.

(7) Three blocks: Node i and j are in the same block; node u and v are not.

(8) Three blocks: Node u and v are in the same block; node i and j are not.

(9) Four blocks: All four nodes are in different blocks.

For circular block permutation with block size L, there are L different ways to block the observations. Also, there are $|G|^2$ pairs of edges in a similarity graph G. Among the L ways to do the blocking, we compute the probability of having each of the nine scenarios for all pairs of edges, and the sum of probabilities of having scenario i is denoted by $c_i, i = 1, \cdots, 9$. The analytic expressions for $c_1, \cdots, c_9$ are given in Definition 1.

Definition 1. *We here define $c_1, \cdots, c_9$, representing the sums of probabilities that each of the $|G|^2$ pairs of edges being blocked into scenario i, $i = 1, \cdots, 9$.*

$$
\begin{aligned}
c_1 &= \frac{1}{L}\sum_{h=1}^{L}(L-h)|\mathcal{E}_h| \\
&+ \frac{1}{L}\sum_{(i,j),(i,u)\in G, j\neq u} I(h_0(i,j,u)=3)(L-\max(\delta_{ij},\delta_{iu},\delta_{ju})) \\
&+ \frac{1}{L}\sum_{(i,j),(u,v)\in G, i\neq j\neq u\neq v} I(h_1(i,j,u,v)=6)(L-\delta_{\max}(i,j,u,v)) \\
c_2 &= \frac{1}{L}\sum_{(i,j),(i,u)\in G, j\neq u} (I(h_0(i,j,u)\leq 2, \delta_{iu}<L)(L-\delta_{iu}) \\
&\quad + I(h_0(i,j,u)=3, \max(\delta_{ij},\delta_{iu},\delta_{ju})\neq\delta_{iu})(L-\delta_{iu})) \\
&+ \frac{1}{L}\sum_{(i,j),(u,v)\in G, i\neq j\neq u\neq v} I(h_1(i,j,u,v)=3)(I(\delta_{uv},\delta_{iu},\delta_{iv})+I(\delta_{uv},\delta_{ju},\delta_{jv}))(L-\delta_{\min,2}(i,j,u,v)) \\
&\quad + I(h_1(i,j,u,v)=4)((I(\delta_{ij},\delta_{iu}\geq L)+I(\delta_{ij},\delta_{iv}\geq L))(L-\delta_{\max}(i,j,u,v)+\min(\delta_{iu},\delta_{iv})) \\
&\qquad + (I(\delta_{ij},\delta_{ju}\geq L)+I(\delta_{ij},\delta_{jv}\geq L))(L-\delta_{\max}(i,j,u,v)+\min(\delta_{ju},\delta_{jv})) \\
&\qquad + (I(\delta_{iu},\delta_{iv}\geq L)+I(\delta_{ju},\delta_{jv}\geq L))(L-\delta_{\max}(i,j,u,v)+\delta_{ij})) \\
&\quad + I(h_1(i,j,u,v)=5)(I(\delta_{ij}\geq L)(2L-\delta_{ij}-\delta_{uv}) \\
&\qquad + I(\delta_{iu}\geq L)(I(\delta_{ij}=\delta_{iv}+\delta_{jv})(L-\delta_{uv})+I(\delta_{iv}=\delta_{ij}+\delta_{jv})(L-\delta_{ju})) \\
&\qquad + I(\delta_{iv}\geq L)(I(\delta_{ij}=\delta_{iu}+\delta_{ju})(L-\delta_{uv})+I(\delta_{iu}=\delta_{ij}+\delta_{ju})(L-\delta_{jv})) \\
&\qquad + I(\delta_{ju}\geq L)(I(\delta_{ij}=\delta_{iv}+\delta_{jv})(L-\delta_{uv})+I(\delta_{jv}=\delta_{ij}+\delta_{iv})(L-\delta_{iu})) \\
&\qquad + I(\delta_{jv}\geq L)(I(\delta_{ij}=\delta_{iu}+\delta_{ju})(L-\delta_{uv})+I(\delta_{ju}=\delta_{ij}+\delta_{iu})(L-\delta_{iv}))) \\
&\quad + I(h_1(i,j,u,v)=6)(I(\delta_{\max}(i,j,u,v)=\delta_{ij})(\delta_{ij}-\delta_{uv}) \\
&\qquad + I(\delta_{\max}(i,j,u,v)=\delta_{iu})(I(\delta_{ij}=\delta_{iv}+\delta_{jv})\delta_{iv}+I(\delta_{iv}=\delta_{ij}+\delta_{jv})\delta_{ij}) \\
&\qquad + I(\delta_{\max}(i,j,u,v)=\delta_{iv})(I(\delta_{ij}=\delta_{iu}+\delta_{ju})\delta_{iu}+I(\delta_{iu}=\delta_{ij}+\delta_{ju})\delta_{ij}) \\
&\qquad + I(\delta_{\max}(i,j,u,v)=\delta_{ju})(I(\delta_{ij}=\delta_{iv}+\delta_{jv})\delta_{jv}+I(\delta_{jv}=\delta_{ij}+\delta_{iv})\delta_{ij}) \\
&\qquad + I(\delta_{\max}(i,j,u,v)=\delta_{jv})(I(\delta_{ij}=\delta_{iu}+\delta_{ju})\delta_{ju}+I(\delta_{ju}=\delta_{ij}+\delta_{iu})\delta_{ij})) \\
c_3 &= \frac{1}{L}\sum_{(i,j),(i,u)\in G, j\neq u} I(h_0(i,j,u)\leq 2, \delta_{ij}<L)(L-\delta_{ij}) \\
&\quad + I(h_0(i,j,u)=3, \max(\delta_{ij},\delta_{iu},\delta_{ju})\neq\delta_{ij})(L-\delta_{ij})) \\
&+ \frac{1}{L}\sum_{(i,j),(u,v)\in G, i\neq j\neq u\neq v} I(h_1(i,j,u,v)=3)(I(\delta_{ij},\delta_{iu},\delta_{ju})+I(\delta_{ij},\delta_{iv},\delta_{jv}))(L-\delta_{\min,2}(i,j,u,v)) \\
&\quad + I(h_1(i,j,u,v)=4)((I(\delta_{iu},\delta_{uv}\geq L)+I(\delta_{iv},\delta_{uv}\geq L))(L-\delta_{\max}(i,j,u,v)+\min(\delta_{iu},\delta_{iv}))
\end{aligned}
$$

$$
\begin{aligned}
&+(I(\delta_{ju},\delta_{uv}\geq L)+I(\delta_{jv},\delta_{uv}\geq L))(L-\delta_{\max}(i,j,u,v)+\min(\delta_{ju},\delta_{jv}))\\
&+(I(\delta_{iu},\delta_{ju}\geq L)+I(\delta_{iv},\delta_{jv}\geq L))(L-\delta_{\max}(i,j,u,v)+\delta_{uv}))\\
&+I(h_1(i,j,u,v)=5)(I(\delta_{uv}\geq L)(2L-\delta_{ij}-\delta_{uv})\\
&+I(\delta_{iu}\geq L)(I(\delta_{ij}=\delta_{iv}+\delta_{jv})(L-\delta_{ij})+I(\delta_{iv}=\delta_{ij}+\delta_{jv})(L-\delta_{iv}))\\
&+I(\delta_{iv}\geq L)(I(\delta_{ij}=\delta_{iu}+\delta_{ju})(L-\delta_{ij})+I(\delta_{iu}=\delta_{ij}+\delta_{ju})(L-\delta_{iu}))\\
&+I(\delta_{ju}\geq L)(I(\delta_{ij}=\delta_{iv}+\delta_{jv})(L-\delta_{ij})+I(\delta_{jv}=\delta_{ij}+\delta_{iv})(L-\delta_{jv}))\\
&+I(\delta_{jv}\geq L)(I(\delta_{ij}=\delta_{iu}+\delta_{ju})(L-\delta_{ij})+I(\delta_{ju}=\delta_{ij}+\delta_{iu})(L-\delta_{ju})))\\
&+I(h_1(i,j,u,v)=6)(I(\delta_{\max}(i,j,u,v)=\delta_{uv})(\delta_{uv}-\delta_{ij})\\
&+I(\delta_{\max}(i,j,u,v)=\delta_{iu})(I(\delta_{ij}=\delta_{iv}+\delta_{jv})\delta_{ju}+I(\delta_{iv}=\delta_{ij}+\delta_{jv})\delta_{uv})\\
&+I(\delta_{\max}(i,j,u,v)=\delta_{iv})(I(\delta_{ij}=\delta_{iu}+\delta_{ju})\delta_{jv}+I(\delta_{iu}=\delta_{ij}+\delta_{ju})\delta_{uv})\\
&+I(\delta_{\max}(i,j,u,v)=\delta_{ju})(I(\delta_{ij}=\delta_{iv}+\delta_{jv})\delta_{iu}+I(\delta_{jv}=\delta_{ij}+\delta_{iv})\delta_{uv})\\
&+I(\delta_{\max}(i,j,u,v)=\delta_{jv})(I(\delta_{ij}=\delta_{iu}+\delta_{ju})\delta_{iv}+I(\delta_{ju}=\delta_{ij}+\delta_{iu})\delta_{uv}))
\end{aligned}
$$

$$
\begin{aligned}
c_4 &= \frac{1}{L}\sum_{(i,j),(u,v)\in G,i\neq j\neq u\neq v} I(h_1(i,j,u,v)=2)I(\delta_{ij},\delta_{uv}<L)(L-\delta_{ij}-\delta_{uv}+x(ij,uv))\\
&+I(h_1(i,j,u,v)=3)I(\delta_{ij},\delta_{uv}<L)(2L-\delta_{\max}(i,j,u,v))_+\\
&+I(h_1(i,j,u,v)=4)((I(\delta_{iu},\delta_{iv}\geq L)+I(\delta_{ju},\delta_{jv}\geq L))(L-\delta_{ij})\\
&+(I(\delta_{iu},\delta_{ju}\geq L)+I(\delta_{iv},\delta_{jv}\geq L))(L-\delta_{uv}))\\
&+I(h_1(i,j,u,v)\geq 5)(I(\delta_{\max}(i,j,u,v)=\delta_{iu})I(\delta_{iv}=\delta_{ij}+\delta_{jv})\delta_{jv}\\
&+I(\delta_{\max}(i,j,u,v)=\delta_{iv})I(\delta_{iu}=\delta_{ij}+\delta_{ju})\delta_{ju}\\
&+I(\delta_{\max}(i,j,u,v)=\delta_{ju})I(\delta_{jv}=\delta_{ij}+\delta_{iv})\delta_{iv}\\
&+I(\delta_{\max}(i,j,u,v)=\delta_{jv})I(\delta_{ju}=\delta_{ij}+\delta_{iu})\delta_{iu})\\
c_5 &= \frac{1}{L}\sum_{h=1}^{L} h|\mathcal{E}_h|\\
&+ \frac{1}{L}\sum_{(i,j),(i,u)\in G,j\neq u}(I(h_0(i,j,u)<3,\delta_{ju}<L)(L-\delta_{ju})\\
&+I(h_0(i,j,u)=3)I(\max(\delta_{ij},\delta_{iu},\delta_{ju})\neq\delta_{ju})\min(\delta_{ij},\delta_{iu}))\\
&+ \frac{1}{L}\sum_{(i,j),(u,v)\in G,i\neq j\neq u\neq v} I(h_1(i,j,u,v)=2)(I(\delta_{iu}<L,\delta_{jv}<L)(L-\delta_{iu}-\delta_{jv}+x(iu,jv))\\
&+I(\delta_{iv}<L,\delta_{ju}<L)(L-\delta_{iv}-\delta_{ju}+x(iv,ju)))\\
&+I(h_1(i,j,u,v)=3)(I(\delta_{\min,2}(i,j,u,v)\neq\delta_4(i,j,u,v))(1-I(\delta_{ij},\delta_{uv}<L))(2L-\delta_{\min,3}(i,j,u,v))_+)\\
&+I(h_1(i,j,u,v)=4)(I(\delta_{ij}\geq L)(L+\delta_{uv}-\max(\delta_{iu},\delta_{iv},\delta_{ju},\delta_{jv}))
\end{aligned}
$$

$$
\begin{aligned}
&+I(\delta_{uv} \geq L)(L + \delta_{ij} - \max(\delta_{iu}, \delta_{iv}, \delta_{ju}, \delta_{jv})))\\
&+I(h_1(i,j,u,v) \geq 5)(I(\delta_{\max}(i,j,u,v) = \delta_{ij})\delta_{uv} + I(\delta_{\max}(i,j,u,v) = \delta_{uv})\delta_{ij}\\
&+I(\delta_{\max}(i,j,u,v) = \delta_{iu})I(\delta_{ij} = \delta_{iv} + \delta_{vj})\delta_{jv}\\
&+I(\delta_{\max}(i,j,u,v) = \delta_{iv})I(\delta_{ij} = \delta_{iu} + \delta_{uj})\delta_{ju}\\
&+I(\delta_{\max}(i,j,u,v) = \delta_{ju})I(\delta_{ji} = \delta_{jv} + \delta_{vi})\delta_{iv}\\
&+I(\delta_{\max}(i,j,u,v) = \delta_{jv})I(\delta_{ji} = \delta_{ju} + \delta_{ui})\delta_{iu})\\
c_6 \;=\; &\frac{1}{L}\sum_{(i,j),(i,u)\in G, i\neq u}(I(h_0(i,j,u)=0)L + I(h_0(i,j,u)=1)\min(\delta_{ij},\delta_{iu},\delta_{ju})\\
&+I(h_0(i,j,u)=2)(\max(\delta_{ij},\delta_{iu},\delta_{ju}) - L))\\
+\;&\frac{1}{L}\sum_{(i,j),(u,v)\in G, i\neq j\neq u\neq v} I(h_1(i,j,u,v)=1)I(\delta_{ij}\geq L, \delta_{uv}\geq L)(L - \delta_{\min}(i,j,u,v))\\
&+I(h_1(i,j,u,v)=2)((L-\delta_{iu})_+ + (L-\delta_{iv})_+ + (L-\delta_{ju})_+ + (L-\delta_{jv})_+\\
&+I(\delta_{iu}<L, \delta_{jv}<L)(2\delta_{iu} + 2\delta_{jv} - 2L - 2x(iu,jv))\\
&+I(\delta_{iv}<L, \delta_{ju}<L)(2\delta_{iv} + 2\delta_{ju} - 2L - 2x(iv,ju)))\\
&+I(h_1(i,j,u,v)=3, \delta_{\min,2}(i,j,u,v)\neq\delta_4(i,j,u,v))(I(\delta_{ij}<L,\delta_{uv}<L)(L - \min(\delta_{iu},\delta_{iv},\delta_{ju},\delta_{jv}))\\
&+I(\delta_{ij}<L,\delta_{uv}\geq L)(\delta_{ij} - |\delta_{uv} - 2L|) + I(\delta_{ij}\geq L,\delta_{uv}<L)(\delta_{uv} - |\delta_{ij} - 2L|)\\
&+I(\delta_{ij}\geq L,\delta_{uv}\geq L)(\delta_{\min,3}(i,j,u,v) - L - 2(\delta_{\min,3}(i,j,u,v) - 2L)_+))\\
&+I(h_1(i,j,u,v)=3, \delta_{\min,2}(i,j,u,v)=\delta_4(i,j,u,v))\min(\delta_{ij},\delta_{uv})\\
&+I(h_1(i,j,u,v)=4)((I(\delta_{ij}\geq L) + I(\delta_{uv}\geq L))(\max(\delta_{iu},\delta_{iv},\delta_{ju},\delta_{jv}) - L)\\
&+(I(\delta_{iu},\delta_{iv}\geq L) + I(\delta_{ju},\delta_{jv}\geq L))\delta_{uv} + (I(\delta_{ui},\delta_{uj}\geq L) + I(\delta_{vi},\delta_{vj}\geq L))\delta_{ij})\\
&+I(h_1(i,j,u,v)=5)(1 - I(\delta_{ij}\geq L) - I(\delta_{uv}\geq L))(\delta_{\max}(i,j,u,v) - L)\\
c_7 \;=\; &\frac{1}{L}\sum_{(i,j),(u,v)\in G, i\neq j\neq u\neq v} I(h_1(i,j,u,v)=1, \delta_{ij}<L)(L-\delta_{ij})\\
&+I(h_1(i,j,u,v)=2, \delta_{ij}<L)(L - \delta_{ij} + I(\delta_{uv}<L)(\delta_{uv} - x(ij,uv) + \delta_{ij} - L))\\
&+I(h_1(i,j,u,v)=3, \delta_{ij}<L)(I(\delta_{\min,2}(i,j,u,v)\neq\delta_4(i,j,u,v))((L-\delta_{ij})\\
&+I(\delta_{uv}<L, \delta_{\max}(i,j,u,v)<2L)(\delta_{\min,3}(i,j,u,v) - 2L))\\
&+I(\delta_{\min,2}(i,j,u,v)=\delta_4(i,j,u,v))(\delta_4(i,j,u,v) - \delta_{ij}))\\
&+I(h_1(i,j,u,v)=4)\\
&(I(\delta_{\max}(i,j,u,v)=\delta_{uv})(I(\delta_{iu}\geq L)\delta_{iv} + I(\delta_{iv}\geq L)\delta_{iu} + I(\delta_{ju}\geq L)\delta_{jv} + I(\delta_{jv}\geq L)\delta_{ju})\\
&+(I(\delta_{iu},\delta_{ju}\geq L) + I(\delta_{iv},\delta_{jv}\geq L))(\delta_{\max}(i,j,u,v) - \delta_{ij} - L))
\end{aligned}
$$

$$
\begin{aligned}
& +I(h_1(i,j,u,v)=5,\delta_{uv}\geq L)(\delta_{uv}-L) \\
c_8 \;=\; & \frac{1}{L}\sum_{(i,j),(u,v)\in G, i\neq j\neq u\neq v} I(h_1(i,j,u,v)=1,\delta_{uv}<L)(L-\delta_{uv}) \\
& +I(h_1(i,j,u,v)=2,\delta_{uv}<L)(L-\delta_{uv}+I(\delta_{ij}<L)(\delta_{ij}-x(ij,uv)+\delta_{uv}-L)) \\
& +I(h_1(i,j,u,v)=3,\delta_{uv}<L)(I(\delta_{\min,2}(i,j,u,v)\neq\delta_4(i,j,u,v))((L-\delta_{uv}) \\
& \quad +I(\delta_{ij}<L,\delta_{\max}(i,j,u,v)<2L)(\delta_{\min,3}(i,j,u,v)-2L)) \\
& \quad +I(\delta_{\min,2}(i,j,u,v)=\delta_4(i,j,u,v))(\delta_4(i,j,u,v)-\delta_{uv})) \\
& +I(h_1(i,j,u,v)=4) \\
& \quad (I(\delta_{\max}(i,j,u,v)=\delta_{ij})(I(\delta_{iu}\geq L)\delta_{ju}+I(\delta_{iv}\geq L)\delta_{jv}+I(\delta_{ju}\geq L)\delta_{iu}+I(\delta_{jv}\geq L)\delta_{iv}) \\
& \quad +(I(\delta_{iu},\delta_{iv}\geq L)+I(\delta_{ju},\delta_{jv}\geq L))(\delta_{\max}(i,j,u,v)-\delta_{uv}-L)) \\
& +I(h_1(i,j,u,v)=5,\delta_{ij}\geq L)(\delta_{ij}-L) \\
c_9 \;=\; & \frac{1}{L}\sum_{(i,j),(u,v)\in G, i\neq j\neq u\neq v} I(h_1(i,j,u,v)=0)L+I(h_1(i,j,u,v)=1)\delta_{\min}(i,j,u,v) \\
& +I(h_1(i,j,u,v)=2)((1-I(\delta_{ij},\delta_{uv}<L)-I(\delta_{iu},\delta_{jv}<L)-I(\delta_{iv},\delta_{ju}<L))(\delta_{\min,2}(i,j,u,v)-L) \\
& \quad +I(\delta_{ij},\delta_{uv}<L)x(ij,uv)+I(\delta_{iu},\delta_{jv}<L)x(iu,jv)+I(\delta_{iv},\delta_{ju}<L)x(iv,ju)) \\
& +I(h_1(i,j,u,v)=3)(\delta_{\min,3}(i,j,u,v)-2L)_+
\end{aligned}
$$

with

$$
\begin{aligned}
\mathcal{E}_h &= \{(i,j),(u,v)\in G : i=u, j=v, \delta_{ij}=h\} \\
\mathcal{E}_L &= \{(i,j),(u,v)\in G : i=u, j=v, \delta_{ij}\geq L\} \\
h_0(i,j,u) &= I(\delta_{ij}<L)+I(\delta_{iu}<L)+I(\delta_{ju}<L) \\
h_1(i,j,u,v) &= I(\delta_{ij}<L)+I(\delta_{iu}<L)+I(\delta_{iv}<L)+I(\delta_{ju}<L)+I(\delta_{jv}<L)+I(\delta_{uv}<L) \\
\delta_{\max}(i,j,u,v) &= \max\{\delta_{ij},\delta_{iu},\delta_{iv},\delta_{ju},\delta_{jv},\delta_{uv}\} \\
\delta_r(i,j,u,v) &= \textit{the } r^{th} \textit{ largest value among } \{\delta_{ij},\delta_{iu},\delta_{iv},\delta_{ju},\delta_{jv},\delta_{uv}\},\ r=2,3,4,5 \\
\delta_{\min}(i,j,u,v) &= \min\{\delta_{ij},\delta_{iu},\delta_{iv},\delta_{ju},\delta_{jv},\delta_{uv}\} \\
\delta_{\min,2}(i,j,u,v) &= \textit{the sum of the two smallest values among } \{\delta_{ij},\delta_{iu},\delta_{iv},\delta_{ju},\delta_{jv},\delta_{uv}\} \\
\delta_{\min,3}(i,j,u,v) &= \textit{the sum of the three smallest values among } \{\delta_{ij},\delta_{iu},\delta_{iv},\delta_{ju},\delta_{jv},\delta_{uv}\} \\
s(i,j) &= I(|i-j|<L)\min(i,j)+I(n-|i-j|<L)\max(i,j) \\
\delta_{ij,uv} &= s(u,v)-s(i,j)
\end{aligned}
$$

$$
\begin{aligned}
b_{ij,uv} &= \delta_{ij,uv} \bmod L \\
x(ij,uv) &= (\min(\delta_{ij}, b_{ij,uv}+\delta_{uv}) - b_{ij,uv})_+ + (\min(\delta_{ij}, b_{ij,uv}+\delta_{uv}-L))_+
\end{aligned}
$$

Note that $c_1 + \cdots + c_9 = |G|^2$. The definition for those $c_1, \cdots, c_9$ are completed.

Combining Lemma 1 and Definition 1, the following theorem gives the analytic expressions for $\Sigma_{\mathbf{R},\texttt{CBP}}(t)$. The proof of Theroem 6 is provided in Appendix B.2.

Theorem 6. *Let $d_1 = c_1$, $d_2 = c_2 + c_3 + c_4 + c_5$, $d_3 = c_6 + c_7 + c_8$, and $d_4 = c_9$, where $c_1, \cdots, c_9$ are graph coefficients as defined in Definition 1, for each $t = aL$, $a \in \{1, \cdots, m-1\}$, where $m = n/L$, we have*

$$
\begin{aligned}
\mathsf{Var}_{\mathit{CBP}}(R_{G,1}(t)) &= d_1 p_1(a) + d_2 p_2(a) + d_3 p_3(a) + d_4 p_4(a) - \mathsf{E}_{\mathit{CBP}}(R_{G,1}(t))^2 \\
\mathsf{Var}_{\mathit{CBP}}(R_{G,2}(t)) &= d_1 p_1^*(a) + d_2 p_2^*(a) + d_3 p_3^*(a) + d_4 p_4^*(a) - \mathsf{E}_{\mathit{CBP}}(R_{G,2}(t))^2 \\
\mathsf{Cov}_{\mathit{CBP}}(R_{G,1}(t), R_{G,2}(t)) &= c_4 p_{11}(a) + c_7 p_{12}(a) + c_8 p_{21}(a) + c_9 p_{22}(a) - \mathsf{E}_{\mathit{CBP}}(R_{G,1}(t))\mathsf{E}_{\mathit{CBP}}(R_{G,2}(t))
\end{aligned}
$$

where $\mathsf{E}_{\mathit{CBP}}(R_{G,1}(t))$ and $\mathsf{E}_{\mathit{CBP}}(R_{G,2}(t))$ are from Lemma 1, and

$$
\begin{aligned}
&p_1(a) = \frac{a}{m},\ p_2(a) = \frac{a(a-1)}{m(m-1)},\ p_3(a) = \frac{a(a-1)(a-2)}{m(m-1)(m-2)},\ p_4(a) = \frac{a(a-1)(a-2)(a-3)}{m(m-1)(m-2)(m-3)}, \\
&p_1^*(a) = \frac{(m-a)}{m},\ p_2^*(a) = \frac{(m-a)(m-a-1)}{m(m-1)},\ p_3^*(a) = \frac{(m-a)(m-a-1)(m-a-2)}{m(m-1)(m-2)}, \\
&p_4^*(a) = \frac{(m-a)(m-a-1)(m-a-2)(m-a-3)}{m(m-1)(m-2)(m-3)}, \\
&p_{11}(a) = \frac{a(m-a)}{m(m-1)},\ p_{12}(a) = \frac{a(m-a)(m-a-1)}{m(m-a)(m-2)},\ p_{21}(a) = \frac{a(a-1)(m-a)}{m(m-a)(m-2)}, \\
&p_{22}(a) = \frac{a(a-1)(m-a)(m-a-1)}{m(m-1)(m-2)(m-3)}.
\end{aligned}
$$

3.3.2 Decomposition of $S_{\texttt{CBP}}(t)$

It was shown in Chu and Chen (2019) that the generalized edge-count test statistic, $S(t)$, under permuation null distribution can be decomposed as

$$S(t) = Z_{w^0}^2(t) + Z_{\text{diff}}^2(t)$$

where $Z_{w^0}(t)$ and $Z_{\text{diff}}(t)$ are uncorrelated. However, this decomposition no longer holds under CBP. In other words, when $L > 1$,

$$S_{\texttt{CBP}}(t) \neq Z_{w^0,\texttt{CBP}}^2(t) + Z_{\text{diff},\texttt{CBP}}^2(t).$$

Moreover, $Z_{w^0,\text{CBP}}(t)$ and $Z_{\text{diff},\text{CBP}}(t)$ are not uncorrelated: $\mathsf{Cov}_{\text{CBP}}(Z_{w^0,\text{CBP}}(t), Z_{\text{diff},\text{CBP}}(t)) \neq 0$.

We work out a new decomposition of $S_{\text{CBP}}(t)$ as stated in Theorem 4. Under CBP, $S_{\text{CBP}}(t)$ can be decomposed into the sum of squares of two asymptotcially independent quantities, $Z_{w,\text{CBP}}(t)$ and $Z_{\text{diff},\text{CBP}}(t)$, where the former is similar to $Z_{w^0,\text{CBP}}(t)$, but adopting a different weight function $q(t)$ that depends on the similarity graph. The result is stated in Lemma 2 below with its proof given in Appendix B.3. The new weight function $q(t)$ is in fact the variance-minimizing weight across all weight functions on $R_{G,1}(t)$ and $R_{G,2}(t)$.

Lemma 2. *The weight function $q(t)$ minimizing $\mathsf{Var}_{CBP}(R_{w(t)}(t))$ is linear in t and is given by*

$$q(t) = c_{G,L}(2t - n) + \frac{1}{2}$$

where

$$c_{G,L} = \frac{\frac{1}{2L}\left(\frac{2c_5^{(sub)}|G|}{m^2(m-1)} - \frac{1}{m(m-1)(m-2)}(c_6 + 2c_7 + 2c_9)\right)}{\frac{-4}{m^2}|G|^2 + \frac{1}{m(m-1)}\big(2(c_2 + c_3 + c_5) + 3c_6 + 4(c_4 + c_9)\big) + \frac{1}{m(m-1)(m-2)}\big((7m-8)c_7 + (m-8)c_8\big)}$$

with $m = n/L$, $c_5^{(sub)}$ as defined in Lemma 1, and $c_1, \cdots, c_9$, as defined in Definition 1. In particular, when $L = 1$, $q(t)$ degenerates to $q^0(t) = \frac{n-t-1}{n-2}$.

3.3.3 Asymptotic properties of $S_{\texttt{CBP}}(t)$

In this section, we derive the limiting distributions of $\{Z_{w,\text{CBP}}([nu]) : 0 < u < 1\}$, and $\{Z_{\text{diff},\text{CBP}}([nu]) : 0 < u < 1\}$ under circular block permutation with block size L.

We first introduce some more notations. For an edge $e = (e_-, e_+)$, where $e_- < e_+$ are the indices of nodes connected by the edge e, and let $\delta(e_-, e_+)$ be the index difference between two nodes e_- and e_+. We define the following notations:

$$\begin{aligned}
A_{e,L,0} &= \{e^* : \min(\delta(e_-^*, e_-), \delta(e_-^*, e_+), \delta(e_+^*, e_-), \delta(e_+^*, e_+)) < L\},\\
A_{e,L,1} &= A_{e,L,0} \cup \left(\cup_{\{e':e' \in G_{e_-^*} \cup G_{e_+^*}, \forall e^* \in A_{e,L,0}\}} A_{e',L,0}\right),\\
A_{e,L,2} &= A_{e,L,1} \cup \left(\cup_{\{e':e' \in G_{e_-^*} \cup G_{e_+^*}, \forall e^* \in A_{e,L,1}\}} A_{e',L,1}\right)
\end{aligned}$$

so that $A_{e,L,0}$ is the subgraph in G that connect to any node which is within L index difference from either node of

edge e; and we say that $A_{e,L,1} \supset A_{e,L,0}$ contains all edges that is first-degree related to the edges in $A_{e,L,0}$, and $A_{e,L,2} \supset A_{e,L,1}$ contains all edges that is first-degree related to the edges in $A_{e,L,1}$, or second-degree related to the edges in $A_{e,L,0}$. Further, we define

$$\begin{aligned} A_{i,L,0} &= \{e^* : \min(\delta(e^*_-, i), \delta(e^*_+, i)) < L\}, \\ A_{i,L,1} &= A_{i,L,0} \cup \left(\cup_{\{e':e' \in G_{e^*_-} \cup G_{e^*_+}, \forall e^* \in A_{e,L,0}\}} A_{e',L,0} \right), \\ A_{i,L,2} &= A_{i,L,1} \cup \left(\cup_{\{e':e' \in G_{e^*_-} \cup G_{e^*_+}, \forall e^* \in A_{e,L,1}\}} A_{e',L,1} \right). \end{aligned}$$

so that $A_{i,L,0}$ is the subgraph in G that connect to any node which is within L index difference from node i; and we say that $A_{i,L,1} \supset A_{i,L,0}$ contains all edges that is first-degree related to the edges in $A_{i,L,0}$, and $A_{i,L,2} \supset A_{i,L,1}$ contains all edges that is first-degree related to the edges in $A_{i,L,1}$, or second-degree related to the edges in $A_{i,L,0}$.

Second, we introduce five conditions on the graph that ensure the convergence of the two basic processes: $\{Z_{w,\text{CBP}}([nu]) : 0 < u < 1\}$, and $\{Z_{\text{diff,CBP}}([nu]) : 0 < u < 1\}$. In the following, we write $a_n = O(b_n)$ when a_n has the same order as b_n, and write $a_n = o(b_n)$ when a_n has order smaller than b_n. Theorem 7 states the convergence of the two processes.

Condition 1. $\sum_{(i,j)\in G} I(\delta_{ij} < L) = o(|G|)$

Condition 2. $\sum_{(i,j),(i,u)\in G} I(\delta_{ij} < L, \delta_{iu} < L) = o(|G|)$

Condition 3. $\sum_{(i,j),(i,u)\in G} 1 = O(|G|^{\beta_1})$, $\beta_1 \geq 1$

Condition 4. $\sum_{(i,j),(u,v)\in G} I(\min\{\delta_{iu}, \delta_{iv}, \delta_{ju}, \delta_{jv}\} < L) = O(|G|^{\beta_2})$, $0 < \beta_2 \leq 2$

Condition 5. $\sum_{(i,j),(u,v)\in G} I(\max\{\delta_{ij}, \delta_{uv}, \delta_{iu}, \delta_{iv}, \delta_{ju}, \delta_{jv}\} < L) = o(|G|)$

Theorem 7. *When* $|G| = O(n^a)$, $1 \leq a < 1.5$, $\sum_{e\in G} |A_{e,L,1}||A_{e,L,2}| = o(n|G|^{1/2})$, $\sum_{i=1}^n |A_{i,L,1}||A_{i,L,2}| = o(n^{3/2})$, *and under Conditions 1-5, as* $n \to \infty$, $\{Z_{w,\text{CBP}}([nu]) : 0 < u < 1\}$ *and* $\{Z_{\text{diff,CBP}}([nu]) : 0 < u < 1\}$ *converge in finite dimensional distributions to two independent Gaussian processes. which we denote as* $\{Z^*_{w,\text{CBP}}(u) : 0 < u < 1\}$ *and* $\{Z^*_{\text{diff,CBP}}(u) : 0 < u < 1\}$, *respectively.*

Let $\rho^*_w(u,v) = \mathbf{Cov}_{\text{CBP}}(Z^*_{w,\text{CBP}}(u), Z^*_{w,\text{CBP}}(v))$ and $\rho^*_{\text{diff}}(u,v) = \mathbf{Cov}_{\text{CBP}}(Z^*_{\text{diff,CBP}}(u), Z^*_{\text{diff,CBP}}(v))$ be the covariance functions of the limiting Gaussian processes, $\{Z^*_{w,\text{CBP}}(u) : 0 < u < 1\}$ and $\{Z^*_{\text{diff,CBP}}(u) : 0 < u < 1\}$.

3.4 Other edge-count tests under CBP

Here, we discuss two more edge-count scan statistics that are derivatives (byporducts) of the generalized edge-count scan statistic (3.1) under CBP. They are the modified weighted edge-count scan statistic, and the modified max-type edge-count scan statistic.

3.4.1 The modified weighted edge-count scan statistic

This scan statistic transforms from the weighted edge-count scan statistic proposed in Chu and Chen (2019), which uses a universal weight $q^0(t) = \frac{n-t-1}{n-2}$. Under CBP, the optiaml weight should depend the the block size L and the similarity graph G, as shown in Lemma 2. Hence, the modifed weighted edge-count scan statistic adopts the weight $q(t)$ given by Theorem 4. Let

$$Z_{w,\text{CBP}}(t) = \frac{R_w(t) - \mathsf{E}_{\text{CBP}}(R_w(t))}{\sqrt{\mathsf{Var}_{\text{CBP}}(R_w(t))}} \quad \text{with} \quad R_w(t) = q(t)R_{G,1}(t) + p(t)R_{G,2}(t),$$

then the null hypothesis of homogeneity is rejected if the scan statistic

$$\max_{n_0 \le t \le n_1} Z_{w,\text{CBP}}(t)$$

with n_0 and n_1 pre-specified, is larger than the critical value for a given significance level. The power comparison between $Z_{w,\text{CBP}}(t)$ and $Z_{w^0,\text{CBP}}(t)$ will be discussed in Section 3.6.2.

3.4.2 The modified max-type edge-count scan statistic

As it is $Z_{w,\text{CBP}}(t)$, rather than $Z_{w^0,\text{CBP}}(t)$, together with $Z_{\text{diff},\text{CBP}}(t)$ that are the elementary components of $S_{\text{CBP}}(t)$, it is obvious that $M^0_{\text{CBP}}(t)$ is not the best way to define the max-type edge-count scan statistic under CBP. Therefore, we propose the modified max-type edge-count test statistic as

$$M_{\text{CBP}}(t) = \max(Z_{w,\text{CBP}}(t), |Z_{\text{diff},\text{CBP}}(t)|).$$

The null hypothesis of homogeneity is rejected if the scan statistic

$$\max_{n_0 \le t \le n_1} M_{\text{CBP}}(t)$$

with n_0 and n_1 pre-speicified, is larger than the critical value for a given significance level.

Under CBP, the performance of the modified max-type edge-count test is similar to the generalized edge-count test. Moreover, the former could achieve more accurate analytic p-value approximations. The type I error control for those new edge-count tests are discussed in Section 3.5.

3.5 Analytical p-value approximations

Given the scan statistics, the next question is how large do they need to be to constitute sufficient evidence against the null hypothesis of homogeneity. In other words, we are concerned with the tail probability of the scan statistics under H_0. For the generalized, the modified weighted, and the modified max-type edge-count tests, respectively, they are

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} Z_{w,\text{CBP}}(t) > b\Big) \tag{3.2}$$

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} S_{\text{CBP}}(t) > b\Big) \tag{3.3}$$

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} M_{\text{CBP}}(t) > b\Big) \tag{3.4}$$

Following the methods in Chu and Chen (2019), we derive asymptotic formulas to compute the three underlying probabilities. The formulas use the results from Theorem 7, which says that the rescaled versions of $Z_{w,\text{CBP}}(t)$ and $Z_{\text{diff,CBP}}(t)$ converge in finite dimensional distributions to two independent Gaussion processes. Following similar derivations in Chu and Chen (2019), we approximate (3.2)-(3.4) by equations (3.5)-(3.7):

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} Z_{w,\text{CBP}}(t) > b\Big) \approx b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} h_w^*(x)\nu(b\sqrt{2h_w^*(x)/n})dx \tag{3.5}$$

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} S_{\text{CBP}}(t) > b\Big) \approx \frac{be^{-b/2}}{2\pi}\int_0^{2\pi}\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} u^*(x,\omega)\nu(\sqrt{2bu^*(x,\omega)/n})dxd\omega \tag{3.6}$$

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} M_{\text{CBP}}(t) > b\Big) = 1-\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} Z_{w,\text{CBP}}(t) < b\Big)\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff,CBP}}(t)| < b\Big) \tag{3.7}$$

with

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} Z_{w,\text{CBP}}(t) < b\Big) \approx 1-b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} h_w^*(x)\nu(b\sqrt{2h_w^*(x)/n})dx$$

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff,CBP}}(t)| < b\Big) \approx 1-2b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} h_{\text{diff}}^*(x)\nu(b\sqrt{2h_{\text{diff}}^*(x)/n})dx$$

and the function $\nu(\cdot)$ can be approximated by

$$\nu(x) \approx \frac{(2/x)(\Phi(x/2)-0.5)}{(x/2)\Phi(x/2)+\phi(x/2)}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function and the probability density function of the standard normal distribution, respectively, and $u^*(x,\omega) = h^*_w(x)\sin^2(\omega) + h^*_{\text{diff}}(x)\cos^2(\omega)$, with $h^*_w(x)$ and $h^*_{\text{diff}}(x)$ defined as below

$$\begin{aligned} h^*_w(x) &= \lim_{u\nearrow x} \frac{\partial \rho^*_w(u,x)}{\partial u} \equiv -\lim_{u\searrow x} \frac{\partial \rho^*_w(u,x)}{\partial u}, \\ h^*_{\text{diff}}(x) &= \lim_{u\nearrow x} \frac{\partial \rho^*_{\text{diff}}(u,x)}{\partial u} \equiv -\lim_{u\searrow x} \frac{\partial \rho^*_{\text{diff}}(u,x)}{\partial u}. \end{aligned}$$

In practice, we use the finite-sample equivalent, $h_w(n,x)$ and $h_{\text{diff}}(n,x)$, in place of $h^*_w(x)$ and $h^*_{\text{diff}}(x)$:

$$\begin{aligned} h_w(n,x) &= n \lim_{s\nearrow nx} \frac{\partial \rho_w(s,nx)}{\partial s} \quad \text{with} \quad \rho_w(s,t) := \mathbf{Cov}_{\texttt{CBP}}(Z_{w,\texttt{CBP}}(s), Z_{w,\texttt{CBP}}(t)), \\ h_{\text{diff}}(n,x) &= n \lim_{s\nearrow nx} \frac{\partial \rho_{\text{diff}}(s,nx)}{\partial s} \quad \text{with} \quad \rho_{\text{diff}}(s,t) := \mathbf{Cov}_{\texttt{CBP}}(Z_{\text{diff},\texttt{CBP}}(s), Z_{\text{diff},\texttt{CBP}}(t)), \end{aligned}$$

for any $s,t \in \mathbb{Z}$, $0 < s \le t < n$. Let $C_w(nx) = h_w(n,x)/n$ and $C_{\text{diff}}(nx) = h_{\text{diff}}(n,x)/n$, then for $t = aL$, $C_w(t)$ and $C_{\text{diff}}(t)$ can be derived to be (see Appendix B.5 for the proof)[1]:

$$\begin{aligned} C_w(t) &= \frac{\sum_{i=1}^9 c_i \lambda_i(a)}{L\left(\sum_{i=1}^9 c_i V_i(a) + Res(a)\right)} \\ C_{\text{diff}}(t) &= \frac{4m(m-1)c_1 + 2m(m-2)(2c_2+c_5) + m(m-4)c_6 - 4m(c_4+2c_7+c_9)}{2La(m-a)\Big((m-1)(4c_1+4c_2+2c_5+c_6) - \big(2(2c_2+c_5)+3c_6+4(c_4+2c_7+c_9)\big)\Big)} \end{aligned}$$

where

$$\begin{aligned} V_1(a) &= \Big(aq(t)^2 + (m-a)(1-q(t))^2\Big)/m \\ V_2(a) &= \Big(q(t)^2 a(a-1) + (1-q(t))^2(m-a)(m-a-1)\Big)/\big(m(m-1)\big) \\ V_3(a) &= V_2(a) \\ V_4(a) &= V_2(a) + \Big(2q(t)(1-q(t))a(m-a)\Big)/\big(m(m-1)\big) \\ V_5(a) &= V_2(a) \\ V_6(a) &= \Big(q(t)^2 a(a-1)(a-2) + (1-q(t))^2(m-a)(m-a-1)(m-a-2)\Big)/\big(m(m-1)(m-2)\big) \\ V_7(a) &= V_6(a) + \Big(2q(t)(1-q(t))a(m-a)(m-a-1)\Big)/\big(m(m-1)(m-2)\big) \\ V_8(a) &= V_6(a) + \Big(2q(t)(1-q(t))a(a-1)(m-a)\Big)/\big(m(m-1)(m-2)\big) \\ V_9(a) &= \Big(q(t)^2 a(a-1)(a-2)(a-3) + (1-q(t))^2(m-a)(m-a-1)(m-a-2)(m-a-3) \end{aligned}$$

[1]Note that $C_w(t)$ and $C_{\text{diff}}(t)$ here in Chapter 3 are defined under CBP, hence are different from those defined in Chapter 2 and Chapter 4.

$$+ 2q(t)(1-q(t))a(a-1)(m-a)(m-a-1)\Big)/\big(m(m-1)(m-2)(m-3)\big)$$

$$Res(a) = -\Big(\frac{aq(t)\big(|G|(m-1) - c_5^{(sub)}(m-a)\big)}{m(m-1)} + \frac{(m-a)(1-q(t))\big(|G|(m-1) - c_5^{(sub)}a\big)}{m(m-1)}\Big)^2$$

$$\lambda_1(a) = \Big(2q(t)-1\Big)^2/\big(2m\big)$$

$$\lambda_2(a) = \Big((2q(t)-1)(2aq(t)-2q(t)+1)\Big)/\big(2m(m-1)\big)$$

$$\lambda_3(a) = \Big((2q(t)-1)(2a-2m-2q(t)-2aq(t)+2mq(t)+1)\Big)/\big(2m(m-1)\big)$$

$$\lambda_4(a) = -4\big(2q(t)-1\big)^2/\big(8m(m-1)\big)$$

$$\lambda_5(a) = -\Big(2a-2m-4q(t)-4aq(t)+4mq(t)-2mq(t)^2+4q(t)^2+1)/\big(2m(m-1)\big)$$

$$\lambda_6(a) = \Big(a^2+2amq(t)-2am-8aq(t)+4a+m^2q(t)^2-2m^2q(t)+m^2-6mq(t)^2$$
$$+ 10mq(t)-4m+8q(t)^2-8q(t)+2\Big)/\big(2m(m^2-3m+2)\big)$$

$$\lambda_7(a) = \Big(-2a^2q(t)+a^2-2amq(t)^2+4amq(t)-2am+8aq(t)^2-12aq(t)+4a+m^2q(t)^2-2m^2q(t)$$
$$+ m^2-8mq(t)^2+12mq(t)-4m+8q(t)^2-8q(t)+2\Big)/\big(2m(m^2-3m+2)\big)$$

$$\lambda_8(a) = \Big(2a^2q(t)-a^2+2amq(t)^2-4amq(t)+2am-8aq(t)^2+4aq(t)-m^2q(t)^2+2m^2q(t)$$
$$- m^2+8q(t)^2-8q(t)+2\Big)/\big(2m(m^2-3m+2)\big)$$

$$\lambda_9(a) = -\Big(2a^2+4amq(t)-4am-12aq(t)+6a+2m^2q(t)^2-4m^2q(t)+2m^2-10mq(t)^2+16mq(t)$$
$$- 6m+12q(t)^2-12q(t)+3\Big)/\big(m(m^3-6m^2+11m-6)\big)$$

3.5.1 Numerical results for p-value approximation under CBP

Here, we check how the p-value approximations based on asymptotic results work for finite samples. To do so, we compare the critical values obtained from analytic formulas against the critical values obtained from doing 10,000 circular block permutations directly, under various simulation settings. In each simulation, sequences of length $n = 1,000$ are generated from a given distribution F_0 in $\mathbb{R}^d$. We consider three distributions (multivariate normal, multivariate exponential, and multivariate log-normal) under various dimensions ($d = 10$, $d = 100$, and $d = 1,000$). In Table 3.1, 3.2, and 3.3, we present the results of one selected dimension for each of the three distributions. **(C1)** denotes multivariate normal with $d = 10$, **(C2)** denotes multivariate exponential with $d = 100$, and **(C3)** denotes multivariate log-normal with $d = 1,000$. The complete tables showing all three distributions under these three dimensions with more cases are in Appendix B.4. The analytical approximations depend on constraints on the sequence in which the change-point is searched over (from n_0 to n_1). For simplicity, we let $n_1 = n - n_0$.

Table 3.1: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} Z_{w,\texttt{CBP}}(t)$ based on MST at $\alpha = 0.05$

	Critical Values												Graph	
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$				
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
(C1)	2.99	3.05	3.06	3.03	3.12	3.12	3.08	3.22	3.27	3.15	3.40	3.53	5360	7
	2.99	3.05	3.04	3.03	3.12	3.12	3.08	3.22	3.26	3.14	3.40	3.51	5396	7
(C2)	2.99	3.05	3.07	3.03	3.12	3.15	3.08	3.22	3.30	3.15	3.40	3.59	11750	32
	2.98	3.05	3.05	3.03	3.12	3.14	3.08	3.22	3.27	3.14	3.39	3.57	11572	35
(C3)	2.99	3.04	3.13	3.03	3.11	3.23	3.08	3.21	3.40	3.15	3.39	3.76	41500	113
	2.98	3.04	3.16	3.03	3.11	3.26	3.08	3.20	3.47	3.14	3.37	3.87	65694	164

Table 3.2: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} S_{\texttt{CBP}}(t)$ based on MST at $\alpha = 0.05$. (For the same reason as in Chu and Chen (2019), we do not perform skewness correction on $S_{\texttt{CBP}}(t)$.)

	Critical Values								Graph	
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$			
	A1	**Per**	**A1**	**Per**	**A1**	**Per**	**A1**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
(C1)	13.11	12.86	13.39	13.29	13.71	14.07	14.12	15.32	5360	7
	13.10	13.03	13.39	13.43	13.71	13.95	14.12	15.38	5396	7
(C2)	13.10	13.26	13.39	13.87	13.71	14.84	14.12	17.39	11750	32
	13.10	13.50	13.38	14.22	13.70	15.29	14.11	18.20	11572	35
(C3)	13.10	14.47	13.39	15.89	13.71	17.85	14.12	22.54	41500	113
	13.10	15.06	13.38	16.56	13.70	19.35	14.11	25.12	65694	164

Table 3.3: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} M_{\texttt{CBP}}(t)$ based on MST at $\alpha = 0.05$

	Critical Values												Graph	
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$				
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
(C1)	3.23	3.27	3.26	3.28	3.33	3.33	3.32	3.41	3.44	3.38	3.56	3.64	5360	7
	3.23	3.27	3.25	3.28	3.33	3.34	3.32	3.41	3.42	3.38	3.56	3.63	5396	7
(C2)	3.23	3.29	3.30	3.28	3.37	3.38	3.32	3.47	3.51	3.38	3.65	3.82	11750	32
	3.23	3.30	3.32	3.27	3.38	3.43	3.32	3.48	3.56	3.38	3.67	3.92	11572	35
(C3)	3.32	3.35	3.42	3.28	3.44	3.57	3.32	3.56	3.78	3.38	3.78	4.24	41500	113
	3.32	3.37	3.41	3.27	3.46	3.57	3.32	3.59	3.88	3.38	3.82	4.34	65694	164

The analytical p-value approximation and the permutation p-value both depend on certain characteristics of the structure of the graph G. In the simulations, we use MST constructed on Euclidena distance. As the structure of MST depends on the observations, the critical values vary by simulation runs. We show results for two randomly simulated sequences in each setting. Two characteristics of the graph are reported: the sum of squared node degrees ($\sum_i |G_i|^2$) and the maximum node degree ($d_{\max}$). These quantities give some intuitions on the size and density of the hubs in the graph. In Table 3.1, 3.2, and 3.3, **A1** presents the analytical critical values without skewness correction, **A2** presnets the skewness corrected critical values, and **Per** presents critical values obtained through $10,000$ circular block permutations.

3.5.2 Skewness corrected p-value approximation

From Table 3.1, 3.2, and 3.3, we see that when n_0 is small, approximations on analytical p-value formulas (**A1**) are not that close to the critical values obtained through permutations directly (**Per**). This is because $Z_{w,\text{CBP}}(t)$ and $Z_{\text{diff},\text{CBP}}(t)$ converge to normal distributions slowly when t is close to 1 or n (Figure 3.3). Following Chu and Chen (2019), we adopt skewness correction to improve the p-value approximations (**A2**). The analytic formulas for skewness corrected p-value approximations are:

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} Z_{w,\text{CBP}}(t) > b\Big) \approx b\phi(b)\int_{n_0}^{n_1} S_w(t)C_w(t)\nu\big(\sqrt{2b^2C_w(t)}\big)dt \tag{3.8}$$

$$\mathbf{P}_{\text{CBP}}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff},\text{CBP}}(t)| > b\Big) \approx 2b\phi(b)\int_{n_0}^{n_1} S_{\text{diff}}(t)C_{\text{diff}}(t)\nu\big(\sqrt{2b^2C_{\text{diff}}(t)}\big)dt \tag{3.9}$$

where $C_w(t) = \frac{\partial\rho_w(s,t)}{\partial s}\Big|_{s=t}$, $C_{\text{diff}}(t) = \frac{\partial\rho_{\text{diff}}(s,t)}{\partial s}\Big|_{s=t}$,

$$S_w(t) = \frac{\exp\left(\frac{1}{2}(b-\hat\theta_{b,w}(t))^2 + \frac{1}{6}\gamma_w(t)\hat\theta^3_{b,w}(t)\right))}{\sqrt{1+\gamma_w(t)\hat\theta_{b,w}(t)}} \quad \text{with} \quad \hat\theta_{b,w}(t) = (-1+\sqrt{1+2b\gamma_w(t)})/\gamma_w(t),$$

$$S_{\text{diff}}(t) = \frac{\exp\left(\frac{1}{2}(b-\hat\theta_{b,\text{diff}}(t))^2 + \frac{1}{6}\gamma_{\text{diff}}(t)\hat\theta^3_{b,\text{diff}}(t)\right))}{\sqrt{1+\gamma_{\text{diff}}(t)\hat\theta_{b,\text{diff}}(t)}} \quad \text{with} \quad \hat\theta_{b,\text{diff}}(t) = (-1+\sqrt{1+2b\gamma_{\text{diff}}(t)})/\gamma_{\text{diff}}(t),$$

and $\gamma_w(t) = \mathsf{E}_{\text{CBP}}(Z^3_{w,\text{CBP}}(t))$, $\gamma_{\text{diff}}(t) = \mathsf{E}_{\text{CBP}}(Z^3_{\text{diff},\text{CBP}}(t))$.

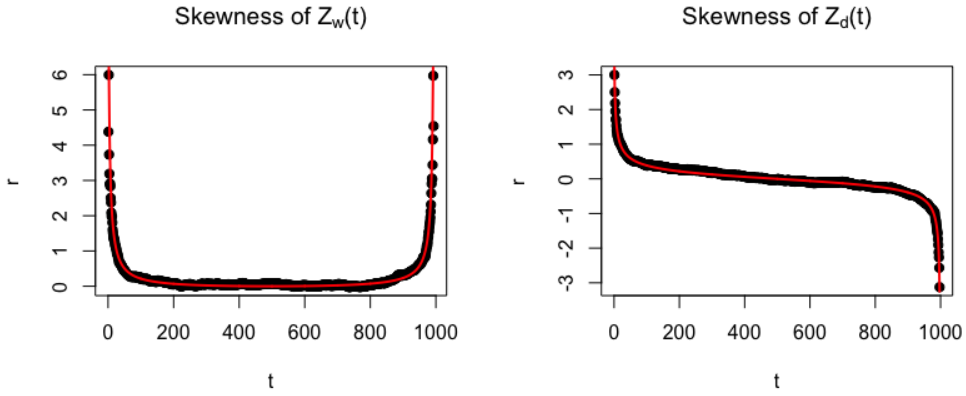


Figure 3.3: Plots of skewness of $Z_{w,\text{CBP}}(t)$ and of $Z_{\text{diff},\text{CBP}}(t)$ against t for a sequence of $1,000$ points randomly generated from $N(0,\mathbb{I}_{100})$. The graph is MST constructed on Euclidean distance. The dots are the estimated $\mathsf{E}_{\text{CBP}}(Z_{w,\text{CBP}}(t))$ and $\mathsf{E}_{\text{CBP}}(Z_{\text{diff},\text{CBP}}(t))$ based on $10,000$ CBP's with $L=5$; the lines represent the analytic values computed by the $\mathsf{E}_{\text{P}}(Z_w(t))$ and $\mathsf{E}_{\text{P}}(Z_{\text{diff}}(t))$ surrogates.

Performing skewness correction requires the computation of $\mathsf{E}_{\text{CBP}}(Z^3_{w,\text{CBP}}(t))$ and $\mathsf{E}_{\text{CBP}}(Z^3_{\text{diff},\text{CBP}}(t))$, which could be very complicated because it invovles the analysis of every set of three edges on the graph. Fortunately, it turns out

that $\mathsf{E}_{\mathbb{P}}(Z_w^3(t))$ and $\mathsf{E}_{\mathbb{P}}(Z_{\text{diff}}^3(t))$ are good approximations of $\mathsf{E}_{\text{CBP}}(Z_{w,\text{CBP}}^3(t))$ and $\mathsf{E}_{\text{CBP}}(Z_{\text{diff,CBP}}^3(t))$, while the former are much easier to be derived. Therefore, we may use $\mathsf{E}_{\mathbb{P}}(Z_w^3(t))$ and $\mathsf{E}_{\mathbb{P}}(Z_{\text{diff}}^3(t))$ to replace $\mathsf{E}_{\text{CBP}}(Z_{w,\text{CBP}}^3(t))$ and $\mathsf{E}_{\text{CBP}}(Z_{\text{diff,CBP}}^3(t))$ in the skewness correction procedure. Figure 3.3 compares the analytic $\mathsf{E}_{\mathbb{P}}(Z_w^3(t))$ and $\mathsf{E}_{\mathbb{P}}(Z_{\text{diff}}^3(t))$ with $\mathsf{E}_{\text{CBP}}(Z_{w,\text{CBP}}^3(t))$ and $\mathsf{E}_{\text{CBP}}(Z_{\text{diff,CBP}}^3(t))$ estimated by $10,000$ CBP's, which shows the feasibility of such approximation. The analytic expressions for $\mathsf{E}_{\mathbb{P}}(Z_w^3(t))$ and $\mathsf{E}_{\mathbb{P}}(Z_{\text{diff}}^3(t))$ are provided in Appendix B.6.

3.6 Performance of new edge-count scan statistics under CBP

This section studies the performace of the new edge-count scan statistics under CBP with various simulation settings.

3.6.1 Type I error control

Figure 3.4, 3.5, and 3.6 show the histograms of p-values in testing the homogeneity of autocorrelated sequences when the sequence has no change-point. For each sequence, the p-value is obtained through doing 1,000 CBP's. Two histograms of p-values for each test statistic are presented: one with permutation and the other CBP with $L = 5$. Here, each sequence is generated from multivariate autoregression model: $\mathbf{y}_t = \rho\mathbf{y}_{t-1} + \epsilon_t,\ t = 1, \cdots, n$, with $\mathbf{y}_0 \sim N(\mathbf{0}, \frac{1}{1-\rho^2}\Sigma)$, $\epsilon_1, \cdots, \epsilon_n \overset{i.i.d.}{\sim} N(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = 0.6^{|i-j|}$, and $\rho = 0.05$, $d = 25$, $n = 100$. The block size $L = 5$ used to account for the autocorrelation is determined by the data-driven method as proposed in Chen (2019a).

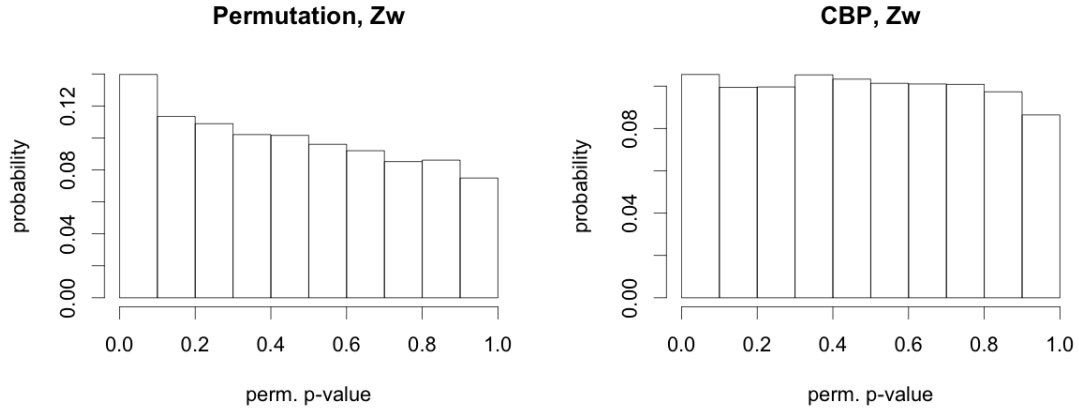


Figure 3.4: Histograms of p-values using $Z_w(t)$ (left) and $Z_{w,\text{CBP}}(t)$ with block size $L = 5$ (right) in testing homogeneity of autocorrelated sequences with no change-point.

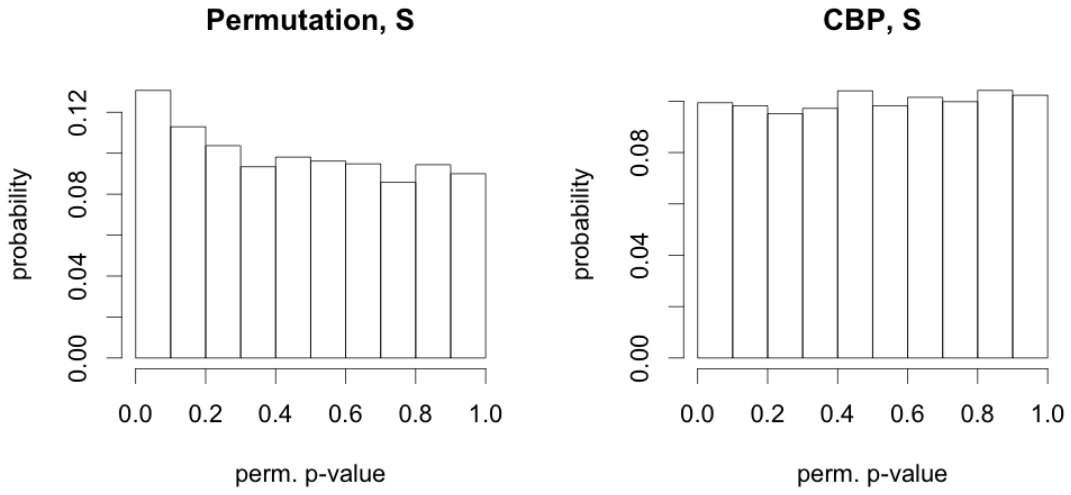


Figure 3.5: Histograms of p-values using $S(t)$ (left) and $S_{\text{CBP}}(t)$ with block size $L = 5$ (right) in testing homogeneity of autocorrelated sequences with no change-point.

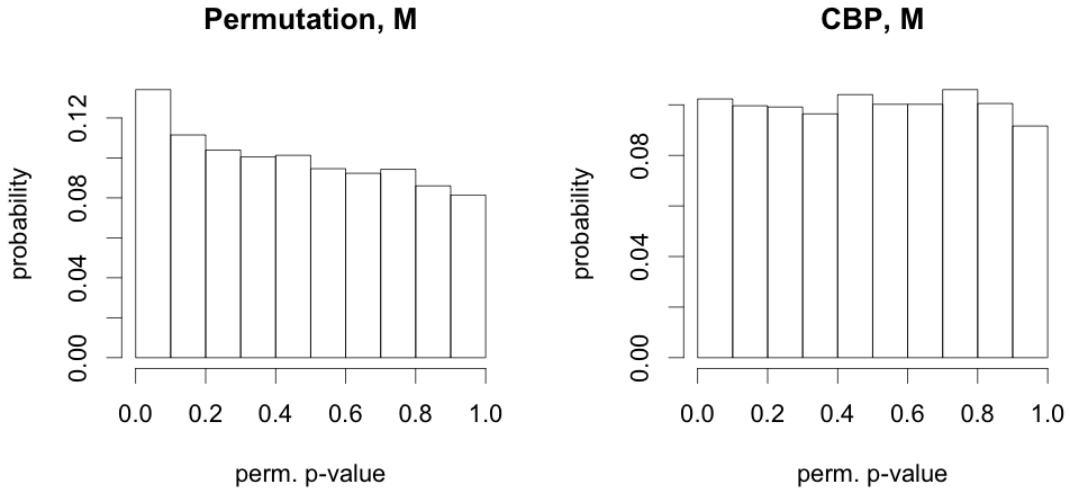


Figure 3.6: Histograms of p-values using $M(t)$ (left) and $M_{\text{CBP}}(t)$ with block size $L = 5$ (right) in testing homogeneity of autocorrelated sequences with no change-point.

It's clear from Figure 3.4, 3.5, and 3.6 that under the null hypothesis of homogeneity, when the observations are autocorrelated, all three tests with CBP would yield approximately uniformly distributed p-values, while the same tests using permutation could lead to higher false discovery rates. Therefore, with circular block permutation, the edge-count scan statisitcs have good control of type I error when the observations are locally dependent.

3.6.2 Power comparison between $Z_{w^0,\texttt{CBP}}(t)$ and $Z_{w,\texttt{CBP}}(t)$

We compare the power of the two edge-count tests: the weighted edge-count test, $Z_{w^0,\texttt{CBP}}(t)$, which adopts weight function $q^0(t)$ for any L under circular block permutation, against the modified weighted edge-count test, $Z_{w,\texttt{CBP}}(t)$, which uses the optimal weight function, $q(t)$ depending on L and the graph G. Note that the larger L is, the more different the two scan statistics are from each other. As a result, to make the comparison meaningful, in each simulation, we randomly generate autocorrelated observations where the block size is chosen to be $L = 20$, and the length of the sequence is $n = 1000$.

Table 3.4 reports the number of times (out of 100) that the null is rejected under $\alpha = 0.05$ for both tests, under various settings of τ, the location of the ture change-point. The numbers in the parentheses indicate the number of times that the change-point is estimated within 20 around the true one, i.e., $\hat{\tau} \in [\tau - 20, \tau + 20]$, where $\hat{\tau}$ is the estimated change-point. The amount of change is chosen so that both tests have moderate powers. Since the weights $q^0(t)$ and $q(t)$ are more close to each other near the middle of the sequence (i.e., t close to $n/2$), and are more separated from one another when t is away from the middle (i.e., t close to 1 or n). We can see that $Z_{w,\texttt{CBP}}(t)$ has higher power than $Z_{w^0,\texttt{CBP}}(t)$ especially when τ is away from 500.

Table 3.4: Power Comparison: Number of times (out of 100) that the null is rejected under $\alpha = 0.05$

τ	150	200	250	300	350	400	450	500
$Z_{w,\texttt{CBP}}(t)$	**68 (30)**	**76 (29)**	**86 (47)**	**94 (48)**	**94 (48)**	98 **(48)**	98 (59)	99 (56)
$Z_{w^0,\texttt{CBP}}(t)$	65 (26)	69 (25)	85 (43)	93 (46)	94 (47)	98 (47)	98 (59)	99 (56)

To further examine the different between the two test, we restrict our τ within 120 to 260. From Table 3.5, we see that in addition to having higher power, the modified weighted edge-count test can estimate the locations of change-points more accurately. Hence we conclude that when τ is near the middle of the sequence, the two tests perform similarly; while $Z_{w,\texttt{CBP}}(t)$ outperforms $Z_{w^0,\texttt{CBP}}(t)$ in the situations where τ is away from the middle.

Table 3.5: Power Comparison: Number of times (out of 100) that the null is rejected under $\alpha = 0.05$

τ	120	140	160	180	200	220	240	260
$Z_{w,\texttt{CBP}}(t)$	**66 (27)**	**65 (24)**	**75 (33)**	68 **(30)**	**77 (36)**	83 **(37)**	**86 (32)**	**90 (40)**
$Z_{w^0,\texttt{CBP}}(t)$	59 (23)	59 (18)	73 (27)	68 (26)	75 (32)	83 (35)	83 (30)	88 (37)

3.6.3 Power comparison among $Z_{0,\texttt{CBP}}(t)$, $Z_{w,\texttt{CBP}}(t)$, $S_{\texttt{CBP}}(t)$, and $M_{\texttt{CBP}}(t)$

Finally, we study the power of all the four edge-count scan statistics (original / modified weighted / generalized / modified max-type) under CBP. Here, a total of six scenarios are investigated, including three types of changes (mean only, mean and variance, and covariance only) at two different locations (center, quarter) of the sequence.

The following simulations use $n = 1000$, $n_0 = 100$, $n_1 = 900$, with observations from multivariate normal distribution. The critical values are determined by 10,000 circular block pemutations. The simulations are studied across various data dimensions, $d = 10, 50, 100, 200, 500$. The baseline parameters before the change happens are $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma_{ij} = \rho_0^{|i-j|}$ with $\rho_0 = 0.6$. Here Σ_{ij} denotes the ith row, jth column of the covariance matrix Σ. The detection is considered a success if the change-point is estimated within 25 from the ture change-point. The significance level is $\alpha = 0.05$.

Table 3.6: *Change in mean vector only.* Observations are generated from multivariate normal distribution with the mean vector changes from $\mathbf{0}$ to $\boldsymbol{\mu}$ after the change-point. The numbers of trials (out of 100) that the null is rejected are reported, and in the parentheses below are the numbers of times the locations of the change-points are sussessfully detected.

Change happens at the center: $\tau = 500$

d	10	50	100	200	500
$\|\|\boldsymbol{\mu}\|\|_2$	0.6	0.8	1.0	1.2	1.5
$Z_{0,\texttt{CBP}}(t)$	**73**	**63**	**91**	**94**	**89**
	(33)	**(38)**	**(73)**	**(78)**	**(76)**
$Z_{w,\texttt{CBP}}(t)$	60	56	88	90	88
	(30)	(21)	(62)	(69)	(72)
$S_{\texttt{CBP}}(t)$	65	51	74	81	75
	(20)	(12)	(40)	(48)	(36)
$M_{\texttt{CBP}}(t)$	59	51	82	78	77
	(24)	(16)	(49)	(64)	(44)

Change happens at a quarter: $\tau = 250$

d	10	50	100	200	500
$\|\|\boldsymbol{\mu}\|\|_2$	0.7	1.0	1.3	1.5	1.8
$Z_{0,\texttt{CBP}}(t)$	**85**	58	87	64	78
	(54)	(23)	(38)	(23)	(16)
$Z_{w,\texttt{CBP}}(t)$	77	**80**	**91**	**92**	**82**
	(46)	**(53)**	**(63)**	**(66)**	**(48)**
$S_{\texttt{CBP}}(t)$	71	62	89	66	74
	(41)	(38)	(59)	(39)	(35)
$M_{\texttt{CBP}}(t)$	75	70	87	83	78
	(39)	(50)	(62)	(59)	(39)

From Table 3.6, we can see that when the change is only in mean vector, and is at the middle of the sequence, the original edge-count scan statistic performs the best. The modified weighted edge-count scan statistic is nearly as good. On the other hand, the generalized / modified max-type edge-count tests have lower power and are less accurate in estimating the change-points under this scenario. However, when the mean change is away from the middle of the sequence, the modified weighted edge-count scan statistic performs the best, and followed by the generalized / modified max-type edge-count tests, while the original edge-count test is relatively not as good as its counterparts.

Table 3.7: *Change in both mean and variance.* Observations are generated from multivariate normal distribution with the mean vector changes from $\mathbf{0}$ to $\boldsymbol{\mu}$, and variance changes from Σ to $\sigma\Sigma$ after the change-point. The numbers of trials (out of 100) that the null is rejected are reported, and in the parentheses below are the numbers of times the locations of the change-points are sussessfully detected.

Change happens at the center: $\tau = 500$

d	10	50	100	200	500
$\|\|\boldsymbol{\mu}\|\|_2$	0.5	0.5	0.5	0.5	0.5
σ	1.05	1.04	1.03	1.02	1.02
$Z_{0,\text{CBP}}(t)$	**78**	81	87	89	93
	(28)	(5)	(2)	(0)	(0)
$Z_{w,\text{CBP}}(t)$	70	48	42	36	34
	(25)	(8)	(5)	(10)	(0)
$S_{\text{CBP}}(t)$	69	88	**93**	**98**	**98**
	(30)	**(52)**	(62)	(51)	**(79)**
$M_{\text{CBP}}(t)$	73	**91**	**93**	86	**98**
	(38)	(48)	**(70)**	**(56)**	(78)

Change happens at a quarter: $\tau = 250$

d	10	50	100	200	500
$\|\|\boldsymbol{\mu}\|\|_2$	0.5	0.5	0.5	0.5	0.5
σ	1.05	1.04	1.03	1.02	1.02
$Z_{0,\text{CBP}}(t)$	36	48	39	41	64
	(0)	(0)	(0)	(0)	(0)
$Z_{w,\text{CBP}}(t)$	52	62	46	26	27
	(21)	(3)	(7)	(11)	(4)
$S_{\text{CBP}}(t)$	**63**	92	91	82	**100**
	(30)	(67)	(63)	(61)	(88)
$M_{\text{CBP}}(t)$	**63**	**95**	**92**	**86**	99
	(24)	**(74)**	**(72)**	**(74)**	**(93)**

Table 3.8: *Change in covariance matrix only.* Observations are generated from multivariate normal distribution with the covariance matrix changes from $\Sigma_{ij} = 0.6^{|i-j|}$ to $\Sigma_{ij} = (0.6 - \Delta\rho)^{|i-j|}$ after the change-point. The numbers of trials (out of 100) that the null is rejected are reported, and in the parentheses below are the numbers of times the locations of the change-points are sussessfully detected.

Change happens at the center: $\tau = 500$

d	10	50	100	200	500
$\Delta\rho$	0.10	0.05	0.06	0.07	0.08
$Z_{0,\text{CBP}}(t)$	65	55	52	61	70
	(5)	(5)	(0)	(3)	(0)
$Z_{w,\text{CBP}}(t)$	40	16	19	12	30
	(8)	(5)	(0)	(2)	(0)
$S_{\text{CBP}}(t)$	93	79	77	88	85
	(74)	(31)	(26)	(61)	(32)
$M_{\text{CBP}}(t)$	**98**	**82**	**91**	**91**	**90**
	(88)	**(45)**	**(45)**	**(70)**	**(49)**

Change happens at a quarter: $\tau = 250$

d	10	50	100	200	500
$\Delta\rho$	0.10	0.05	0.06	0.07	0.08
$Z_{0,\text{CBP}}(t)$	22	23	34	51	32
	(0)	(0)	(0)	(0)	(0)
$Z_{w,\text{CBP}}(t)$	21	28	20	41	21
	(0)	(6)	(0)	(16)	(5)
$S_{\text{CBP}}(t)$	**88**	67	**78**	**92**	90
	(41)	(38)	**(46)**	(61)	**(54)**
$M_{\text{CBP}}(t)$	**88**	**73**	75	91	**94**
	(53)	**(39)**	(37)	**(68)**	(51)

From Table 3.7, we can see that when the change is in both mean and variance, regardless of where the change is, the modified max-type edge-count scan statistic performs the best, followed closely by the modified generalized edge-count scan statistic. On the contrary, the original edge-count test and the modified weighted edge-ocunt test not only have significatly lower power, but could also lead to a biased estimate of change-point. From Table 3.8, we see that even when the change is in covariance matrix, the modified max-type edge-count scan statistic performs very well, and so does the modified generalized edge-count scan statistic. Under this scenario, the other two tests are not recommended.

Through this simulation study, one can clearly observe that the modified weighted edge-count test is designed for the changes in mean while the generalized / modified max-type edge-count tests are used for detecting various types of changes.

3.7 A real data example

We demonstrate our new methods on the yellow taxi trip records, which is publicly available on the NYC Taxi & Limousine Commission (TLC) website (cite website). Tons of detailed information on the taxi trip recoreds are provided on the website, including taxi pickup and drop-off date/times, longitude and latitude coordinates of pickup and drop-off locations, trip distances, fares, rate types, payments types, and driver-reported passenger counts.

Given the abondance of the yellow taxi dataset, there are lots of questions and topics we can pose and explore. Here, we illustrate our methods in detecting changes in travel departing from John F. Kennedy International Airport for a one-year time period from Oct. 1, 2014 to Sep. 30, 2015. For simplicity, the boundary of JFK airport was set to be -73.80 to -73.77 longitude and 40.63 to 40.66 latitude. We restrict our study on trips that began with a pickup location within the territory of JFK airport.

We then extract information on the longitude and latitude drop-off coordinates for those trips departing from JFK airport. The range of their drop-off locations are chosen to be -85.00 to -64.00 longitude and 34.50 to 49.50 latitude. Using longitude/latitude coordinates, we create a 30 by 30 grid on the range of drop-off locations and count the number of daily taxi drop-offs that fall within each cell, where each cell represents a longitude/latitude coordinate range. Then for each day, we have a 30 by 30 matrix such that each element represents the number of taxi drop-offs within each area.

To test whether there are significant changes within the time period, we apply the three new edge-count tests together with the original edge-count test in Chen (2019a) to the taxi data under the CBP framework. Let A_i be the 30×30 matrix on day i, and we denote ν_i to be the vector form of A_i, which is then 900×1. The L_2 norm is used to construct the MST graph representing similarity between days. We first apply the data driven method proposed in Chen (2019a) to determine the size of L to account for the local dependency of the dataset. Figure 3.7 shows the paths of $M_{\text{CBP}}(t)$ from $L = 1$ to $L = 10$. We can see that as L increases, the path of $M_{\text{CBP}}(t)$ moves downward. The figure suggests that $L = 8$ is sufficient for this dataset, because there is a big jump of the paths from $L = 7$ to $L = 8$ and after $L = 8$ the path begins to move slowly. Though such choice of L could be somewhat subjective, it accounts roughly for the locally dependent structure among the data within a week.

For the 365 days period from Oct. 1, 2014 to Sep. 30, 2015, the three new tests (modified weighted/generalized/modified max-type) report a change-point on Day 187, which corresponds to Apr. 5, 2015. Similarly, the original edge-count test reports a change-point on Day 186, Apr. 4, 2015 (Table 3.9). One explanation of this result may lie in weather. Demand for taxi in cold days could be different from that in warm days, so a change-point happened as winter came to an end.

We may compare the results with the methods in Chu and Chen (2019). Instead of circular block permutation with $L = 8$, we use pure permutation, or CBP with $L = 1$ as the null distributions of the scan statistics. When the methods are applied to the whole one-year period or the second segment, all tests agree with the results under CBP framework

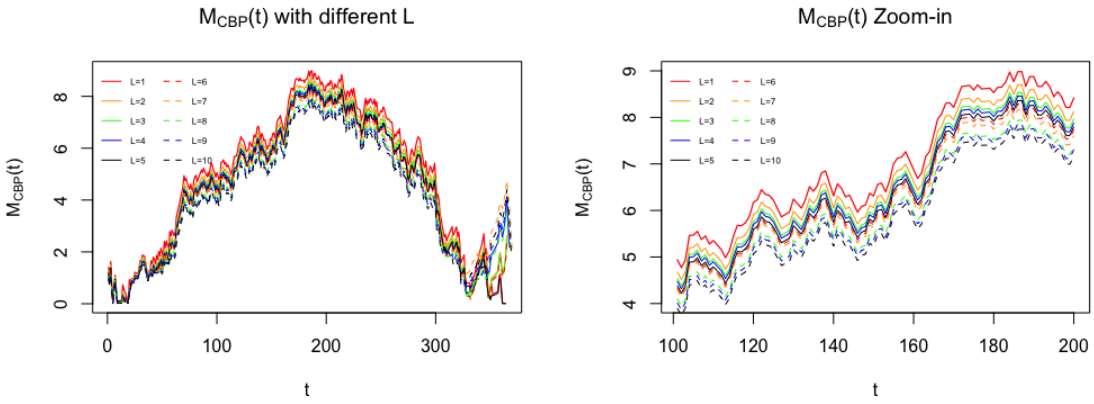


Figure 3.7: Paths for $M_{\text{CBP}}(t)$ with block size $L = 1, \ldots, 10$ and its zoom-in version.

with $L = 8$.

Table 3.9: Change-point results and corresponding p-values (reported in parentheses) for NYC taxi pickups from JFK over the whole one-year period. (Top table: CBP with $L = 8$. Bottom table: Permutation.)

Time period	$Z_{0,\text{CBP}}(t)$	$Z_{w,\text{CBP}}(t)$	$S_{\text{CBP}}(t)$	$M_{\text{CBP}}(t)$
10/1/2014 - 9/30/2015	4/4/2015 **(<0.001)**	4/5/2015 **(<0.001)**	4/5/2015 **(<0.001)**	4/5/2015 **(<0.001)**
Time period	$Z_0(t)$	$Z_{w^0}(t)$	$S(t)$	$M(t)$
10/1/2014 - 9/30/2015	4/2/2015 **(<0.001)**	4/5/2015 **(<0.001)**	4/5/2015 **(<0.001)**	4/5/2015 **(<0.001)**

3.8 Discussion and Conclusion

We propose new graph-based scan statistics for the testing and estimation of change-points that relax independence assumption of the framework proposed by Chu and Chen (2019). To account for locally dependent data, we incorporate the circular block permutation scheme proposed in Chen (2019a) into the new edge-count tests. In particualr, we propose a new decomposition of the generalized edge-count test statistic under CBP. The new component is called the modified weighted edge-count test that adjusts the weighted edge-count test to the circular block permutation setting. Under CBP framework, we find that the optimal weight function should depend on the graph and block size L, while when $L = 1$, this optimal weight coincides with the original weight as proposed in Chu and Chen (2019). The importance of this optimal weight is two-fold. First, it makes the modified weighted edge-count test statistic uncorrelated with the other component in the decomposition, which facilitates the study of the asymptotic distribution of the generalized edge-count scan statistic. Second, the modified weighted edge-count test could achieve higher power especially when the observations are autocorrelated.

The new scan statistics are based on two basic processes, $Z_{w,\texttt{CBP}}(t)$ and $Z_{\text{diff},\texttt{CBP}}(t)$, with the former sensitive to locational alternatives and the latter sensitive to scale alternatives. We show that the two basic processes rescaled by the length of the sequence, $\{Z_{w,\texttt{CBP}}([nu]) : 0 < u < 1\}$ and $\{Z_{\text{diff},\texttt{CBP}}([nu]) : 0 < u < 1\}$, converge to independent Gaussian processes in finite dimensional distributions under some mild conditions of the graph. Even though the covariance functions of the limiting Gaussian processes do depend on the graph when $L \geq 2$, simulation studies show that the limiting processes are robust to the distribution of the observations.

Analytic p-value approximations based on limiting distributions (asymptotic p-value approximation) are derived for all new statistics and the skewness-corrected versions are derived for the modified weighted edge-count statistic and the modified max-type edge-count statistic. The asymptotic p-value approximations provides a ballpark estimate of the p-value. The skewness-corrected versions give more accurate approximations. The modified weighted edge-count scan statistic is designed for changes in mean that are not close to the center of the sequence, while the modified max-type edge-count scan statistic is useful for detecting more generic changes. As a result, in practice, when prior knowledge of the type of changes is unavailable, the modified max-type edge-count test is recommended.

Chapter 4

Change-point Detection in Multiple Sequences of High-dimensional/non-Euclidean Data

4.1 Introduction

We study the problem of detecting simultaneous change-points in multiple sequences of high-dimensional/non-Euclidean observations. With advance in technology, modern data collected in many fields usually come in various forms (such as Neuropixels recordings (Jun et al., 2017), microarrays (Zeebaree et al., 2018b), and so on), rendering traditional change-point detection methods for univariate observations (James et al. (1987), Carlstein et al. (1994)) not very useful in coping with those datasets. Moreover, in many studies and experiments, there are multiple subjects, with each of them represented by a sequence of observations. For example, Chen et al. (2019) studies the Neuropixels data collected from 9 different region of the brain of a mouse; Visconti di Oleggio Castello et al. (2020) records the fMRI sequences of 25 people watching the movie "The Grand Budapest Hotel" by Wes Anderson; Nakai et al. (2021) collects multiple fMRI sequences of the brains from 5 participants while they are listening to different music genres. For the Neuropixels data in the above examples, each observation consists of high-dimensional measurements represented by the probes in certain region of the mouse brain; for the fMRI data, each observation here is a 3D image. In this work, we focus on the detection of simultaneous change-points in sequences of such complex observations.

Let $\{\mathbf{y}_1^{(m)}, \ldots, \mathbf{y}_n^{(m)} : 1 \leq m \leq N\}$ be the observations of N sequences of length n. The task of change-point

detection in multiple sequences can be formulated as the following hypothesis testing problem:

$$H_0 : \mathbf{y}_t^{(m)} \sim F_0^{(m)} \quad \text{for } t = 1, \ldots, n; m = 1, \ldots, N \tag{4.1}$$

against the single change-point alternative:

$$H_a : \text{ For at least one } m, \exists 1 \leq \tau < n, \mathbf{y}_t^{(m)} \sim \begin{cases} F_0^{(m)}, & t \leq \tau \\ F_1^{(m)}, & \text{otherwise} \end{cases}$$

where $F_0^{(m)}$ and $F_1^{(m)}$, $m = 1, \ldots, N$, are different probability measures. When $N = 1$, the problem reduces to change-point detection in a single sequecne, which has been well studied both parametrically and nonparametrically (Heard et al. (2010), Wang et al. (2013), Harchaoui et al. (2009), Matteson and James (2014), Chu and Chen (2019)). In particualr, the graph-based methods in Chu and Chen (2019) have analytic formulas to control the type I error efficiently, and can be easily applied to high-dimensional/non-Euclidean sequences of observations. Nevertheless, when there are multiple sequences ($N > 1$), one usually has to apply these methods to each of the sequences separately. For those methods that are applicable to high-dimensional data, another option is to combine those sequences into one sequence of observations, with each observation a collection of observations from each of the sequences. However, there are some drawbacks and limitations dealing with multiple sequences of observations with the two approaches. First of all, by applying the methods to each of the sequences separately, we have the multiple testing problem. When several tests are performed simultanouesly, it would be difficult to control the real type I error. Second, applying the methods to each the subjects individually fails to take advantage of the sample size (number of sequences, N) because in each subject seuqence, there could be noise that may interfere with the detection of the change-points, and therefore affecting the power of the tests. This could be improved by applying the methods to all the subject sequences aggregately. An alternative approach is to apply the methods to the combined sequence of observations. However, this could also dilute the signals of the change-points, especially when the types or sizes of the changes are different for each sequence of observations, also resulting in low power (see Section 4.4).

In the context of change-point detection in multiple sequences, Zhang et al. (2010) studied the problem of detecting common changes in mean vector in sequences of univariate Gaussian variables. Under the same setting, Siegmund et al. (2011) continued to investigate the scenarios where the mean values of the observations change simultaneously in only a subset of the sequences. Motivated by the problem of detecting of DNA copy number variants, these parametric methods work well in specific applications. However, most data analysis tasks in modern times involve sequences of high-dimensional/non-Euclidean observations, and the types of changes could be very diverse or unpredictable. In such

cases, some nonparametric approaches could be useful. Candidate methods include distance-based ecp (Matteson and James, 2014), and graph-based method gSeg (Chu and Chen, 2019). Nevertheless, when there are multiple sequences, those methods could have low power or even fail to detect some changes in many scenarios (Section 4.4).

In this work, we propose a new edge-count MS-statistic which aims for detecting simultaneous change-points in multiple sequences of high-dimensional/non-Euclidean observations. The new test with the MS-statistic is a nonparametric, graph-based method which utilizes the edge-counts informaiton on the similarity graphs for each sequence of observations. Our MS-statistic improves on the idea of the max-type edge-count test statistic proposed in Chu and Chen (2019), and is able to detect more types of changes and has higher power in the presence of multiple sequences. We also work out an analytic formula to approximate the p-values, so that our method can be fast-applicable to large scale datasets. The performance of our new test is showcased by extensive simulation studies (see Section 4.4), and a real data example on the yellow taxi trip records (see Section 4.5).

The rest of the chapter is arranged as follows. We define the MS-statistic in Section 4.2. The asymptotic properties of the test statisitc, as well as the analytic formula for type I error control are discussed in Section 4.3. In Section 4.4 we present the simulation studies on the power of our test. The real data application is included in Section 4.5. We conclude in Section 4.6.

4.2 The Test Statistics

4.2.1 Test statistic for a single sequence

Let $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be the length-n sequence of observations, and $G = \{(i,j) : \mathbf{y}_i \text{ and } \mathbf{y}_j \text{ are connected}\}$ be the undirected similarity graph constructed among the observations. Similar to the edge-count quantities defined in Chapter 2, we define the within-group edge counts

$$R_{G,1}(t) = \sum_{(i,j)\in G} \mathbb{1}_{\{i\le t, j\le t\}} \quad ; \quad R_{G,2}(t) = \sum_{(i,j)\in G} \mathbb{1}_{\{i>t, j>t\}}.$$

Here, $R_{G,1}(t)$ is the number of edges in G connecting both observations before t, and $R_{G,2}(t)$ is the number of edges in G connecting both observations after t. Figure 4.1 is a toy example that shows the counting of $R_{G,1}(t)$, and $R_{G,2}(t)$. Based on the two quantities, and following the definitions in Chapter 2, we have

$$R_w(t) = \frac{n-t-1}{n-2} R_{G,1}(t) + \frac{t-1}{n-2} R_{G,2}(t) \quad ; \quad R_{\text{diff}}(t) = R_{G,1}(t) - R_{G,2}(t),$$

and $Z_w(t)$ and $Z_{\text{diff}}(t)$ be their standardized versions:

$$Z_w(t) = \frac{R_w(t) - \mathsf{E}(R_w(t))}{\sqrt{\mathsf{Var}(R_w(t))}} \quad ; \quad Z_{\text{diff}}(t) = \frac{R_{\text{diff}}(t) - \mathsf{E}(R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}(R_{\text{diff}}(t))}}.$$

Here E and Var denote respectively the expectation and varinace taken under random permutation that places probability $1/n!$ to each of the possible permutation outcomes of a length-n sequence. Then the max-type edge-count test statistic in Chu and Chen (2019) is defined as

$$\max_{n_0 \le t \le n_1} M(t) \tag{4.2}$$

where $M(t) = \max(Z_w(t), |Z_{\text{diff}}(t)|)$. Then with n_0 and n_1 pre-specified, the null hypothesis of homogeneity (4.1) is rejected if (4.2) is larger than the threshold for a given significance level.

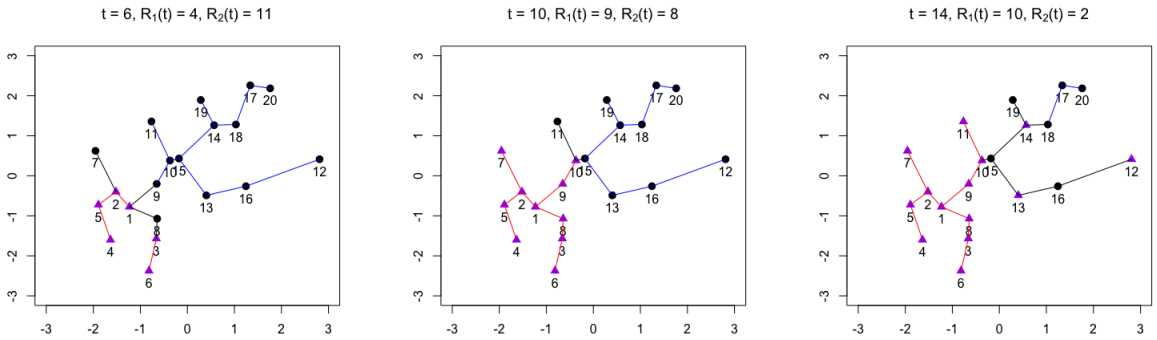


Figure 4.1: The computation of $R_{G,1}(t)$ and $R_{G,2}(t)$ at three different values of t. Here $\mathbf{y}_1, \ldots, \mathbf{y}_{10} \overset{\text{i.i.d.}}{\sim} N((-0.7, -0.7)^T, \mathbb{I}_2)$, and $\mathbf{y}_{11}, \ldots, \mathbf{y}_{20} \overset{\text{i.i.d.}}{\sim} N((0.7, 0.7)^T, \mathbb{I}_2)$, where $\mathbb{I}_2$ is the 2×2 identity matrix. The graph G here is MST on the Euclidean distance. Each t divides the observations into two groups: one group for observations before t (purple triangles) and the other group for observations after t (black circles). Red edges connect observations before t and the number of red edges is $R_{G,1}(t)$; blue edges connect observations after t and the number of blue edges is $R_{G,2}(t)$. Notice that as t changes, the group identities change but the graph G does not change.

The max-type edge-count test statistic (4.2) is desinged to capture the signals from various possible types of changes in the alternative. For example, when there is a change in mean at t in the seuqence, observations before t and observations after t could be separated on the graph. Therefore, there would be more within-group edges, $R_{G,1}(t)$ and $R_{G,2}(t)$, making $R_w(t)$ larger than its null expectation. On the other hand, when there is a change in variance at t, then observations with the smaller variance tends to concentrate within in the inner layer - forming edges mostly within themselves, whereas observations with the larger variance tends to scatter around the outer layer - inevitably form edges with observations in the inner layer. This phenomenon becomes severe especially when the observations are in high dimension, which is also known as The Curse of Dimensionality. If the variance increases after t, then

$R_{G,1}(t)$ would be larger, but $R_{G,2}(t)$ would be smaller compared to their null expectations. Therefore, in this case $R_{\text{diff}}(t)$ would be larger than than its null expectation. If the variance decreases after t, then by the same reasoning $R_{\text{diff}}(t)$ would be smaller than its null expectation. Hence, the absolute value on $Z_{\text{diff}}(t)$ in (4.2) takes into account the two possible directions. (See Chu and Chen (2019) for more discussions.)

4.2.2 New test statistic for multiple sequences

The reason why (4.2) is used to detect change-points for single sequence setting is because when there is a change-point at t, either $Z_w(t)$ or $|Z_{\text{diff}}(t)|$ or both tend to be large, thus making $M(t)$ large (see full discussion in Chu and Chen (2019)). As the size of the edge counts capture the signal of a change-point, when there are multiple sequences, we would like to add up the signals from each of the sequences to detect common change-points. Therefore, let $\{\mathbf{y}_1^{(m)}, \ldots, \mathbf{y}_n^{(m)} : m = 1, \ldots, N\}$ be the N sequences of length-n observations, and $G_m = \{(i,j) : \mathbf{y}_i^{(m)} \text{ and } \mathbf{y}_j^{(m)} \text{ are connected}\}$ be the similarity graph constructed among the m-th sequence of the observations. Then for $m = 1, \ldots, N$, we define

$$R_w^{(m)}(t) = \frac{n-t-1}{n-2} R_{G_m,1}(t) + \frac{t-1}{n-2} R_{G_m,2}(t) \quad ; \quad R_{\text{diff}}^{(m)}(t) = R_{G_m,1}(t) - R_{G_m,2}(t),$$

and $Z_w^{(m)}(t)$ and $Z_{\text{diff}}^{(m)}(t)$ be their standardized versions:

$$Z_w^{(m)}(t) = \frac{R_w^{(m)}(t) - \mathsf{E}(R_w^{(m)}(t))}{\sqrt{\mathsf{Var}(R_w^{(m)}(t))}} \quad ; \quad Z_{\text{diff}}^{(m)}(t) = \frac{R_{\text{diff}}^{(m)}(t) - \mathsf{E}(R_{\text{diff}}^{(m)}(t))}{\sqrt{\mathsf{Var}(R_{\text{diff}}^{(m)}(t))}}.$$

Let $S_w(t) = \sum_{m=1}^N \left(Z_w^{(m)}(t)\right)^2$, and $S_{\text{diff}}(t) = \sum_{m=1}^N \left(Z_{\text{diff}}^{(m)}(t)\right)^2$. We define the new MS-statistic:

$$\max_{n_0 \le t \le n_1} MS(t) \tag{4.3}$$

where $MS(t) = \max(S_w(t), S_{\text{diff}}(t))$. With n_0 and n_1 pre-specified, the null hypothesis of homogeneity (4.1) is rejected if (4.3) is larger than the threshold for a given significance level.

The MS-statistic is motivated to accumulate the signals from the N sequences. When there is a change in mean at t in sequence m, $Z_w^{(m)}(t)$ tends to be larger, and $S_w(t)$ sums up those signals. On the other hand, when there is a change in variance at t in sequence m, $Z_{\text{diff}}^{(m)}(t)$ tends to deviate from zero, and $S_{\text{diff}}(t)$ sums up those signals. We square the statistics of each sequence so that the signals will not cancel out if some of the sequences contain changes in variance with opposite directions.

To illustrate the effectiveness of the MS-statistics, we simulate sample paths for $MS(t)$ with and without a change-point in a sequence of $n = 1,000$ observations. In the setting of no change-point, observations are generated i.i.d. from

100-dimensional Gaussian distribution with mean zero and covariance Σ, whose element in the i-th row and the j-th column is $\Sigma_{ij} = 0.6^{|i-j|}$; in the setting with a change-point at $t = 500$, the first 500 observations are generated from the same distribution as in the no change-point setting, and the last 500 observations have their mean vector shift from zero to $0.15 \times \mathbf{1}_{100}$, where $\mathbf{1}_{100}$ denotes a length-100 vector of all one's. Figure 4.2 depicts the illustration for $MS(t)$.

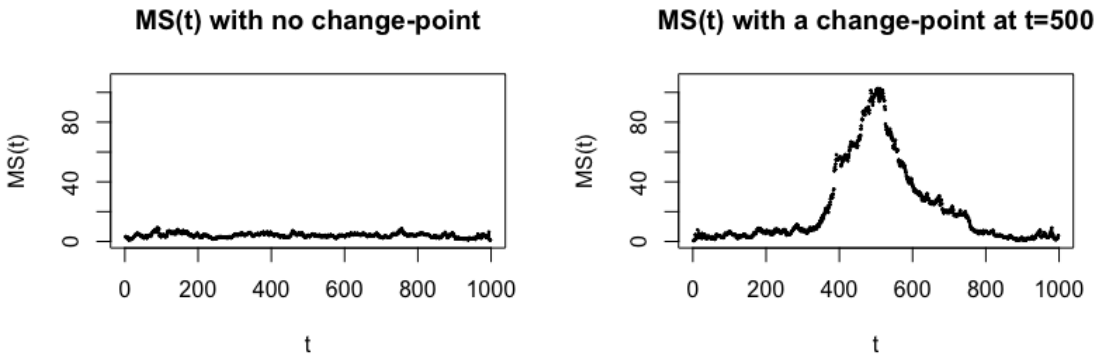


Figure 4.2: Sample paths for $MS(t)$ when there is no change-point (left panel); and when there is a change-point at $\tau = 500$ (right panel).

Another natural formulation of the test statistic would be $S(t) = S_w(t) + S_{\text{diff}}(t)$, called the S-statistic here. The MS-statistic and the S-statistic are similar but have slightly different rejection regions. Figure 4.3 shows the comparison of the performance between the two statistics under locational (change in mean) and scale (change in variance) alternatives, respectively. We see that under both scenarios, the two tests perform similarly with the MS-statistic having slightly higher power than the S-statistic. In this work, we focus on the discussion of the MS-statistic. The other reason for this is that we have derived the analytic formula to approximate type I error for the MS-statistic (see Section 4.3).

4.2.3 Analytic expressions for MS-statistic

To compute the MS-statisitc efficiently, we need analytic expressions for the expectaions and variances for each $R_w^{(m)}(t)$ and $R_{\text{diff}}^{(m)}(t)$ so that we do not have to perform the time-consuming permutations to obtain them. Given the similarity graphs for each of the sequences, G_m, $m = 1, \ldots, N$, we have

$$
\begin{aligned}
\mathsf{E}\left(R_w^{(m)}(t)\right) &= |G_m| \frac{(t-1)(n-t-1)}{(n-1)(n-2)}, \\
\mathsf{E}\left(R_{\text{diff}}^{(m)}(t)\right) &= |G_m| \frac{(2t-n)}{n}, \\
\mathsf{Var}\left(R_w^{(m)}(t)\right) &= \frac{t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)} \left(|G_m| - \frac{\sum_{u=1}^{n} |G_m(u)|^2}{(n-2)} + \frac{2|G_m|^2}{(n-1)(n-2)} \right), \\
\mathsf{Var}\left(R_w^{(m)}(t)\right) &= \frac{t(n-t)}{n(n-1)} \left(\sum_{u=1}^{n} |G_m(u)|^2 - \frac{4|G_m|^2}{n} \right),
\end{aligned}
$$

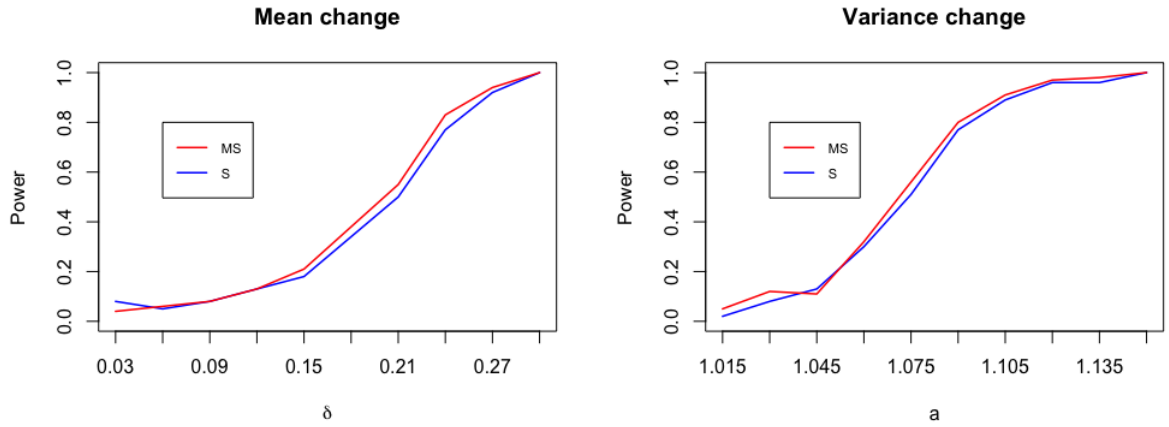


Figure 4.3: Power comparison of MS-statistic and S-statistic. There are $N = 4$ sequences of length $1,000$ with each obseravation generated from 10-dimensional Gaussian distribution with covariance $\Sigma_{ij} = 0.6^{|i-j|}$. The critical values are determined by $10,000$ permutations at $\alpha = 0.05$. The change-point occurs at $\tau = 300$. Left panel: all the mean vectors shift by δ; Right panel: variances become $a\Sigma$ for all sequences.

where $|G_m|$ denotes the total number of edges, and $|G_m(u)|$ denotes the degree of node $\mathbf{y}_u$ in graph G_m. The above results follow directly from Chu and Chen (2019). Note that the expectations only depend on the number of edges in the graphs, when all the graphs have the same numbers of edges, the expectations of $R_w^{(m)}(t)$ for all sequences are the same, and the expectations of $R_{\text{diff}}^{(m)}(t)$ for all sequences are the same. On the other hand, in addition to $|G_m|$, the variances also depend on $\sum_{u=1}^{n} |G_m(u)|^2$, the sum of squared degrees of the graph. Therefore, the variances could be different for each sequence even though the same graph is used (all graphs are k-MST, for example).

4.3 Analytic p-value Approximations

4.3.1 Asymptotic properties of the test statistics

Chu and Chen (2019) shows that under some mild conditions on the graph, the rescaled versions of the two basic processes, $\{Z_w([nv]^1) : 0 < v < 1\}$ and $\{Z_{\text{diff}}([nv]) : 0 < v < 1\}$ converge in finite dimensional distributions to two independent Gaussian processes. Here we list those conditions as in Theorem 4.1, Chu and Chen (2019): (i) $|G| = O(n^\beta), 1 \le \beta < 1.5$, (ii) $\sum_{u=1}^{n} |G(u)|^2 - \frac{4|G|^2}{n} = O(\sum_{u=1}^{n} |G(u)|^2)$, (iii) $\sum_{e \in G} |A_e||B_e| = o(n^{1.5\beta})$, and (iv) $\sum_{e \in G} |A_e|^2 = o(n^{\beta+0.5})$. These conditions ensure that the graph G is dense enough but not too dense. In conditions (i) and (ii), $|G|$ denotes the total number of edges in graph G, and $|G(u)|$ denotes the degree of node $\mathbf{y}_u$ in graph G. In conditions (iii) and (iv), $e = (e_-, e_+)$ denotes an edge with $e_- < e_+$, $A_e = G_{e_-} \cup G_{e_+}$ is the subgraph in G that connect to either node e_- or node e_+, and $B_e = \cup_{e^* \in A_e} A_{e^*}$ is the subgraph in G that connect to any edge in

[1]For a scalar x, we use $[x]$ to denote the largest integer no greater than x.

A_e.

Based on this result, the authors in Chu and Chen (2019) derive asymptotic analytical formulas to approximate the p-value for the max-type test statistic (4.2):

$$\begin{aligned}\mathsf{P}\Big(\max_{n_0\le t\le n_1} M(t) > b\Big) &\approx 1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} Z_w(t) < b\Big)\mathsf{P}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff}}(t)| < b\Big) \\ &= 1 - \left(1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} Z_w(t) > b\Big)\right)\left(1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff}}(t)| > b\Big)\right).\end{aligned} \tag{4.4}$$

where

$$\mathsf{P}\Big(\max_{n_0\le t\le n_1} Z_w(t) > b\Big) \approx b\phi(b)\int_{n_0}^{n_1} C_w(t)\nu\big(\sqrt{2b^2C_w(t)}\big)dt \tag{4.5}$$

$$\mathsf{P}\Big(\max_{n_0\le t\le n_1} |Z_{\text{diff}}(t)| > b\Big) \approx 2b\phi(b)\int_{n_0}^{n_1} C_{\text{diff}}(t)\nu\big(\sqrt{2b^2C_{\text{diff}}(t)}\big)dt \tag{4.6}$$

where the function $\nu(\cdot)$ can be estimated numerically as $\nu(x) \approx \frac{(2/x)(\Phi(x/2)-0.5)}{(x/2)\Phi(x/2)+\phi(x/2)}$ with $\phi(\cdot)$ and $\Phi(\cdot)$ being the probability density function and cumulative distribution function of the standard normal distribution, respectively; and $C_w(t)$, $C_{\text{diff}}(t)$ are the partial derivatives of the covariance function of their corresponding processes, i.e.,

$$\begin{aligned} C_w(t) &= \lim_{s\nearrow t}\frac{\partial\rho_w(s,t)}{\partial s} \quad ; \quad \rho_w(s,t) = \mathsf{Cov}(Z_w(s), Z_w(t)), \\ C_{\text{diff}}(t) &= \lim_{s\nearrow t}\frac{\partial\rho_{\text{diff}}(s,t)}{\partial s} \quad ; \quad \rho_{\text{diff}}(s,t) = \mathsf{Cov}(Z_{\text{diff}}(s), Z_{\text{diff}}(t)).\end{aligned}$$

Moreover, $C_w(t)$ and $C_{\text{diff}}(t)$ can be further derived and simplified to

$$C_w(t) = \frac{n(n-1)(2t^2/n - 2t + 1)}{2t(n-t)(t^2 - nt + n - 1)} \quad ; \quad C_{\text{diff}}(t) = \frac{n}{2t(n-t)}.$$

As we can see, both $C_w(t)$ and $C_{\text{diff}}(t)$ are functions independent of the similarity graph G. In fact, as shown in Theorem 4.3, Chu and Chen (2019), the covariance functions of $Z_w(t)$ and $Z_{\text{diff}}(t)$ are also distribution-free and do not depend on the graph at all.

In this work, we consider the case when there are multiple sequences. Suppose there are N sequences, then we have N similarity graphs $(G_1, \ldots, G_N)$, and their corresponding two processes, $Z_w^{(1)}(t), \ldots, Z_w^{(N)}(t)$, and $Z_{\text{diff}}^{(1)}(t), \ldots, Z_{\text{diff}}^{(N)}(t)$. Since both $\rho_w(s,t)$ and $\rho_{\text{diff}}(s,t)$ do not depend on the graph, $Z_w^{(1)}(t), \ldots, Z_w^{(N)}(t)$ share the covariance function, $\rho_w(s,t)$; and $Z_{\text{diff}}^{(1)}(t), \ldots, Z_{\text{diff}}^{(N)}(t)$ share the covariance function, $\rho_{\text{diff}}(s,t)$.

Use the notation $|A|$ to denote the cardinarity of a set A, $G_l(i)$ to denote the set of edges in G_l that incident to node i, so $|G_l(i)|$ is the degree of node i in the graph G_l, and $V_{G_l} := \sum_{i=1}^{n}(|G_l(i)| - \frac{2|G_l|}{n})^2$ that represents the variability

of degrees of the graph G_l. Let G be the union of all graphs $G_1, \cdots, G_N$ ($G = \cup G_l$), $G(i)$ is the set of edges in G incident to node i, $\mathcal{G}(i)$ to be the set of nodes incident to the node i in the graph G, and N_{sq} is the number of squares in the graph G. Define $\phi(e) := \sum_{l=1}^{N} \mathbb{1}_{\{e \in G_l\}}$ that is the number of graphs containing edge e. Let

$$c_i' = \sum_{l=1}^{N} \frac{\left| |G_l(i)| - \frac{2|G_l|}{n} \right|}{\sqrt{V_{G_l}}}, \text{ for } i = 1, \cdots, n$$
$$c_0 = \max_l \{\frac{1}{\sqrt{G_l}}\}.$$

Assume that $\phi(e)$ is uniformly bounded for all edges e, we list sufficient conditions to derive the limiting distribution:

$$\sum_{i=1}^{n} {c_i'}^3 \to 0, \tag{4.7}$$
$$c_0 \sum_{i=1}^{n} {c_i'}^2 |G(i)| \to 0 \tag{4.8}$$
$$c_0^3 \sum_{i=1}^{n} |G(i)|^2 \to 0 \tag{4.9}$$
$$c_0^2 \sum_{i=1}^{n} \sum_{j,k \in \mathcal{G}(i)}^{j \neq k} c_i' c_j' \to 0 \tag{4.10}$$
$$c_0^4 N_{sq} \to 0, \tag{4.11}$$

as n goes to infinity.

Liu et al. (2022) proved the following Theorem.

Theorem 8. *Given N graphs $G_1, \cdots, G_N$ with the same order of cardinality, and $\phi(e) = \sum_{l=1}^{N} \mathbb{1}_{\{e \in G_l\}}$ is bounded for all edges $e \in G$. If for any $1 \leq i < j \leq N$,*

$$\lim_{n \to \infty} \sum_{u=1}^{n} \frac{|G_i(u)||G_j(u)|}{n} - \frac{4|G_i||G_j|}{n^2} = o(1) \quad \textit{and} \quad |G_i \cap G_j| = o(|G_1|),$$

and conditions (4.7)-(4.11) hold, then $\{Z_w^{(1)}([nv]) : 0 < v < 1\}, \ldots, \{Z_w^{(N)}([nv]) : 0 < v < 1\}$ and $\{Z_{diff}^{(1)}([nv]) : 0 < v < 1\}, \ldots, \{Z_{diff}^{(N)}([nv]) : 0 < v < 1\}$ converge in finite dimensional distributions to $2N$ independent Gaussian processes.

Based on Theorem 8, we derive the analytical formulas to approximate the p-values for the MS-statisitc.

4.3.2 Analytical p-value approximation formula for MS-statistic

Here we derive the analytic p-value approximation formula for the MS-statistic, i.e.,

$$\begin{aligned} \mathsf{P}\Big(\max_{n_0\le t\le n_1} MS(t) > b\Big) &\approx 1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} S_w(t) < b\Big)\mathsf{P}\Big(\max_{n_0\le t\le n_1} S_{\text{diff}}(t) < b\Big) \\ &= 1 - \left(1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} S_w(t) > b\Big)\right)\left(1 - \mathsf{P}\Big(\max_{n_0\le t\le n_1} S_{\text{diff}}(t) > b\Big)\right). \end{aligned} \tag{4.12}$$

Again, to compute (4.12), all we need is to compute the two probabilities on the right-hand side. Under the conditions in Theorem 8, we show (Appendix C.1) that the two probabilities can be approximated by

$$\mathsf{P}(\max_{n_0\le t\le n_1} S_w(t) > b) \approx 2b\left(1 - \frac{N-1}{b}\right) f_N^{\chi^2}(b) \int_{n_0}^{n_1} C_w(x)\nu\left(\sqrt{2bC_w(x)}\left(1 - \frac{N-1}{b}\right)\right) dx \tag{4.13}$$

$$\mathsf{P}(\max_{n_0\le t\le n_1} S_{\text{diff}}(t) > b) \approx 2b\left(1 - \frac{N-1}{b}\right) f_N^{\chi^2}(b) \int_{n_0}^{n_1} C_{\text{diff}}(x)\nu\left(\sqrt{2bC_{\text{diff}}(x)}\left(1 - \frac{N-1}{b}\right)\right) dx \tag{4.14}$$

where $f_N^{\chi^2}(b) = \frac{b^{N/2-1}}{2^{N/2}\Gamma(N/2)}\exp(-\frac{b}{2})$ is the density of a chi-square distribution with degree of freedom N.

Remark 2. *In the case of single sequence ($N = 1$), the factor $1 - \frac{N-1}{b} = 1$ vanishes, then equation* (4.13) *and* (4.14) *reduces to* (4.5) *and* (4.6)*, respectively, which can be easily verified by a change of variable.*

Table 4.1 shows the performance of the analytical p-value approximation formula (4.12) under significance levels $\alpha = 0.05$ across various selected data distributions and dimensions based on $10,000$ permutations (multivariate Gaussian distribution with dimension $d = 10$, multivariate t_5-distribution with $d = 100$, and log-normal distribution with $d = 1,000$). The critival values given by the asymptotic analytical formula (4.12) are presented in the last row of Table 4.1, "Analytical." We report the results for the following seven numbers of sequences ($N = 1, 2, 3, 5, 10, 20, 50$). The length of each sequence is $n = 1,000$, and we set $n_0 = 100$, $n_1 = 900$. For the multivairate Gaussian and multivariate t_5-distribution, the observations in each sequence are generated i.i.d. from their corresponding distributions with mean zero and covariance Σ, whose element in the i-th row and the j-th column is $\Sigma_{ij} = 0.6^{|i-j|}$. For the log-normal distribution, each observation consists of d indepednet log-normal randan variables. Complete results for different combinations of data distributions and dimensions are provided in Appendix C.2.

Table 4.1: Critical values for test statistic $\max_{n_0\le t\le n_1} MS(t)$ based on 5-MST at $\alpha = 0.05$.

	$N=1$	$N=2$	$N=3$	$N=5$	$N=10$	$N=20$	$N=50$
Multivariate Gaussian ($d=10$)	11.54	14.87	17.45	21.83	31.42	47.68	89.36
Multivariate t_5 ($d=100$)	11.38	14.81	17.59	22.07	31.64	48.00	90.07
Log-normal ($d=1,000$)	11.98	15.45	18.36	22.99	32.63	49.02	92.19
Analytical	11.31	14.53	17.13	21.59	31.03	47.25	89.54

From Table 4.1, we see that the analytic p-vlaue approximation formula for the MS-statistic gives very accurate estimates of the critical values no matter what the underlying distribution of the data is. The accuracy of the p-value approximation is quite consistent regradless of the number of sequences.

4.4 Power Evaluation

In this section, we study the power of our proposed MS-statistic with three existing methods - the parametric method proposed in Zhang et al. (2010), the the graph-based max-type statistic (Chu and Chen, 2019), and the distance-based method (Matteson and James, 2014) - through simulation studies. In the following, we use "New," "Zhang," "gSeg," and "ecp" to denote the four methods, respectively. Below describes some general settings in the simulations that are adopted throughout Section 4.4. We generate N sequences, with each sequence having length $n = 1,000$, where all observations are independent and have dimension $d = 100$. For all but the ecp method, the change-points are searched over the interval $[n_0 = 100, n_1 = 900]$. Moreover, for our new method (MS-statistic), N 5-MST's are constructed for each of the N sequences on Euclidean distance; for the gSeg method (max-type statistic), one $5N$-MST is consturcted for the combined sequence on Euclidean distance.

Before we dive into the study of the power performance of these methods, we first examine their ability to control type I error at a given significance level for high-dimensional data structures. Figure 4.4 plots the rejection rates of the four methods against various numbers of sequences N. In each scenario, the the rejection rate is computed based on $5,000$ simulations under significance level $\alpha = 0.05$. The threshold (rejection region) for the MS-statistic is derived by plugging in (4.13) and (4.14) into formula (4.12), thresholds for other methods are also determined by their associated techniques.

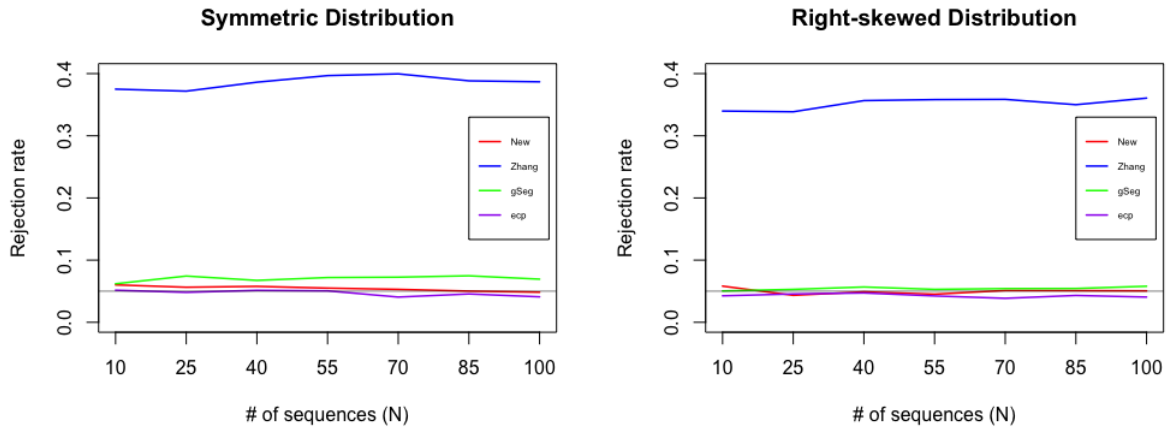


Figure 4.4: Rejection rates of the four methods in N homogeneous sequences. Left panel: observations are from multivariate t_5 distribution. Right panel: observations are from Exp(1) distribution.

From Figure 4.4, we see that the method in Zhang et al. (2010) fails to control the type I error at $\alpha = 0.05$, whereas the other three methods, including our new MS-statistic, have good control on the false discovery rate. The reason is that Zhang et al. (2010) views each dimension as a univariate sequence and assumes those sequences are independent of each other. When the d coordinates in a high-dimensional observation are correlated, the method in Zhang et al. (2010) could have a higher false discovery rate than expected. In the above simulations, the covariance matrix of each observation is Σ, where $\Sigma_{ij} = 0.6^{|i-j|}$. For the right-skewed distribution in the right panel of Figure 4.4, let $\mathbf{u}$ be the length-d vector of i.i.d. random variables from $\text{Exp}(1)$, the obseravtion is generated as $\Sigma^{\frac{1}{2}}\mathbf{u}$. Figure 4.4 illustrates the limitation of the method (Zhang et al., 2010) when applied to more generic data structure. Therefore, we exclude this method in the power comparison below.

4.4.1 Small number of sequences

We then study the power of our proposed MS-statistic (New) with the other two methods that can also control type I error - the graph-based max-type statistic (gSeg), and the distance-based method (ecp). We generate $N = 4$ sequences with length $n = 1,000$ and a change-point occurring at $\tau = 300$. Before the change-point, the observations are i.i.d. from $d = 100$ multivariate t_5 distribution with mean $0 \times \mathbf{1}_d$, and covariance matrix Σ, where $\Sigma_{ij} = 0.6^{|i-j|}$. For the distributions after the change-point, the mean vector for sequence m ($m = 1, 2, 3, 4$) becomes $\Delta_m \times \mathbf{1}_d$, and the covariance matrix for sequence m becomes $a_m\Sigma$. In another setting where only the off-diagonal structure of the covariance matrix changes, the matrix for sequence m becomes $\Sigma^{(m)}$ with $\Sigma^{(m)}_{ij} = \rho_m^{|i-j|}$. We build the 5-MST on Euclidean distance for each of the four sequences in constructing our MS-statistic. For the graph-based max-type statistic (gSeg), we first merge the four sequences into one, and the statistic is computed based on the 20-MST on Euclidean distance among the merged observations. As we use the 5-MST ($k = 5$) for the MS-statistic, we adopt the 20-MST ($k = 20$) for the max-type statistic when there are $N = 4$ sequences to make the comparison fair.

We study the power of the three tests in two settings. Section 4.4.1 discusses the setting where all $N = 4$ sequences have a change-point at $\tau = 300$, while the sizes and types of changes could be differrent. Section 4.4.1 focuses on the setting where the change only exists in one of the four sequences, while the other three sequences remain homogeneous.

Change occurs in all sequences

The simulations are conducted in the following five scenarios, and the results are presented in Table 4.2.

Scenario I: Change in mean vectors only. The distributions before the change, or $F_0^{(m)}$, $m = 1, 2, 3, 4$, are stated in Section 4.4.1. After the change, one coordinate of the mean vector in sequence m shifted by Δ_m, with $(\Delta_1, \Delta_2, \Delta_3, \Delta_4) = (-\delta, -\delta, \delta, \delta)$, two positive and two negative. We study the power of the three tests across different values of δ.

Scenario II: Change in variances in the same direction. The distributions before the change are stated in Section 4.4.1. After the change, the covariance matrix of sequence m changes from Σ to $a_m\Sigma$, with $(a_1, a_2, a_3, a_4) = (2-a, 2-a, a, a)$. We study the power of the three tests across different values of a. After the change-point, two sequences of observations have their variances increase, and the other two decrease.

Scenario III: Change in mean plus change in variances in different directions. The distributions before the change are stated in Section 4.4.1. After the change, all sequences have a mean change of size 0.5 in one coordinate, and the covariance matrix of sequence m changes from Σ to $a_m\Sigma$, with $(a_1, a_2, a_3, a_4) = (2-a, 2-a, a, a)$. We study the power of the three tests across different values of a.

Scenario IV: Change in off-diagonal elements of the covariance matrix. The distributions before the change are stated in Section 4.4.1. After the change, the covariance matrix of sequence m changes from $\Sigma_{ij} = 0.6^{|i-j|}$ to $\Sigma_{ij}^{(m)} = \rho_m^{|i-j|}$, with $(\rho_1, \rho_2, \rho_3, \rho_4) = (\rho, \rho, \rho, \rho)$. Define $\Delta\rho = 0.6 - \rho$. We study the power of the three tests across different values of $\Delta\rho$.

Scenario V: Change in both mean and variance for asymmetric distribution. The changes in mean and covariance before and after the change-point are the same as Scenario III. The underlying distribution here is centered chi-square distribution with degree of freedom 3. That is, $\chi_3^2 - 3$. Let $\mathbf{u}$ be the length-d vector with each component i.i.d. from χ_3^2. Each observation is generated as $\Sigma^{\frac{1}{2}}(\mathbf{u} - 3 \times \mathbf{1}_d)$.

From Table 4.2, we see that wnen the change is only in mean (Scenario I), the ecp performs the best, and the two graph-based methods (gSeg and our method) are not too bad compared to ecp. This is not a surprise since ecp uses all the pairwise distances information while the graph-based methods use only the information of the similarity graphs. In Scenario II to V, we see that ecp has very low power, and our MS-statistic performs significantly better than the other methods. Especially in Scenario II, III and V, when there are changes in variance in the opposite directions, both ecp and gSeg fail to detect such changes. On the other hand, the power of the MS-statistic increases as the size of change increases, meaning that our new method is able to capture the signals of such changes efficiently. Lastly, in Scenario IV, when the change is in the off-diagonal elements of the covariance matrix, both graph-based methods can detect the changes successfully, while our new test has a higher power than gSeg for any given size of change.

Change occurs in one sequence

In Section 4.4.1, we have studied the cases where changes occur in all $N = 4$ sequences. Following the same settings, here we discuss the cases where the change occurs only in one of the four sequences, whereas the other three sequences remain homogeneous. In particular, the five settings here are:

Scenario I: Change in mean in one sequence. After the change-point, one coordinate of the mean vector of one sequence becomes δ.

Table 4.2: Numbers of times (out of 100) that the null hypothesis is rejected under significance level $\alpha = 0.05$. In the parentheses are the numbers of times the estimated change-point $\hat{\tau}$ is within 50 of the true change-point $\tau = 300$, i.e., $\hat{\tau} \in [275, 325]$, which can be interpreted as "accuracy."

Scenario I: Change in mean (only one coordinate: 2 positve 2 negative)

δ	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
New	9 (1)	4 (0)	4 (0)	5 (1)	9 (0)	15 (3)	19 (8)	43 (24)	74 (53)	90 (76)	100 (86)
gSeg	0 (0)	4 (0)	2 (0)	8 (0)	14 (1)	16 (6)	34 (16)	60 (37)	76 (45)	94 (70)	100 (79)
ecp	7 (0)	5 (0)	4 (0)	7 (0)	13 (4)	18 (4)	42 (29)	91 (77)	99 (85)	100 (95)	100 (98)

Scenario II: Change in variance (different directions)

a	1.00	1.02	1.04	1.06	1.08	1.10	1.12	1.14	1.16	1.18	1.20
New	6 (0)	11 (1)	8 (0)	20 (2)	27 (4)	44 (19)	60 (31)	72 (41)	93 (56)	96 (71)	98 (73)
gSeg	9 (0)	4 (0)	3 (0)	6 (0)	8 (0)	8 (0)	6 (1)	6 (0)	10 (0)	1 (0)	7 (0)
ecp	7 (0)	6 (1)	5 (0)	10 (2)	10 (1)	3 (0)	2 (1)	7 (0)	6 (0)	2 (0)	4 (0)

Scenario III: Mean change (fixed) + Variance change (different direction)
(mean 0.5 in one coordinate, variance same as Scenario II)

a	1.00	1.02	1.04	1.06	1.08	1.10	1.12	1.14	1.16	1.18	1.20
New	16 (2)	6 (1)	11 (2)	17 (3)	22 (2)	49 (12)	64 (24)	83 (45)	96 (50)	98 (64)	99 (74)
gSeg	15 (6)	17 (3)	19 (6)	16 (6)	20 (7)	25 (6)	11 (6)	26 (8)	16 (4)	32 (11)	31 (11)
ecp	19 (7)	19 (11)	18 (7)	13 (4)	18 (6)	15 (4)	16 (4)	20 (10)	16 (7)	22 (9)	22 (11)

Scenario IV: Change in off-diagonal elements of the covariance matrix

$\Delta\rho$	0.00	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20
New	6 (0)	7 (1)	5 (0)	17 (4)	27 (10)	40 (12)	63 (28)	91 (51)	97 (64)	100 (81)	100 (87)
gSeg	9 (0)	4 (0)	4 (0)	8 (0)	9 (1)	10 (2)	14 (5)	31 (12)	31 (12)	48 (21)	60 (28)
ecp	7 (0)	6 (1)	5 (0)	3 (0)	5 (1)	4 (0)	4 (1)	5 (0)	9 (1)	4 (0)	9 (0)

Scenario V: Chi-square distribution, changes same as Scenario III

a	1.00	1.02	1.04	1.06	1.08	1.10	1.12	1.14	1.16	1.18	1.20
New	6 (0)	18 (2)	62 (31)	99 (71)	100 (90)	100 (95)	100 (98)	100 (100)	100 (100)	100 (100)	100 (100)
gSeg	4 (1)	3 (0)	4 (0)	6 (2)	7 (0)	7 (1)	11 (1)	9 (2)	16 (2)	13 (1)	19 (10)
ecp	3 (1)	8 (1)	6 (0)	10 (3)	5 (0)	6 (0)	9 (0)	5 (0)	8 (1)	8 (1)	10 (1)

Scenario II: Change in variance in one sequence. After the change-point, the covariance matrix of one sequence changes from Σ to $a\Sigma$.

Scenario III: Change in both mean and variance in one sequence. After the change-point, one coordinate of the mean vector in one sequence shifts by 1.0, and the covariance matrix of the same sequence changes from Σ to $a\Sigma$.

Scenario IV: Change in covariance structure in one sequence. After the change-point, the covariance matrix of one sequence changes from $\Sigma_{ij} = 0.6^{|i-j|}$ to $\Sigma'_{ij} = (0.6 - \Delta\rho)^{|i-j|}$.

Scenario V: Change in both mean and variance for asymmetric distribution in one sequence. The setting is the same as **Scenario V** in Section 4.4.1, but here the change occurs in only one sequence and the other three sequences remain homogeneous.

From Table 4.3, we see that when the change is in only one of the sequences, our MS-statistic is the best among the three methods in all scenarios. The new test is sensitive to various types of changes and has the highest power. Even in Scenario I, when the change is only in mean, our new test has higher power than ecp. In other scenarios where gSeg or ecp could fail, our MS-statistic demonstrates its effectiveness in detecting different types of changes.

4.4.2 Large number of sequences

Here we study the power performance of the three tests (our method, gSeg, and ecp) under simulations with $N = 100$ sequences. Other parameters are the same: $n = 1,000$, $d = 100$, $n_0 = 100$, $n_1 = 900$, $k = 5$. Before the change-point, the observations are i.i.d. from multivariate t_5 distribution with mean zero and covariance Σ as defined before. In Section 4.4.1, all the sequences have the same type of changes. Here, we allow the sequences to have different types of chagnes. For each sequence in the simulations, the change could equally likely be in one of the three: change in mean, change in variance, and change in covariance. If the change is in mean, then after the change-point, one coordinate of the mean vector becomes $\pm\delta$, with plus or minus being assigned randomly with probability a half; if the change is in variance, the covariance matrix becomes $(1 \pm \Delta a)\Sigma$ after the change-point, with plus or minus being assigned randomly with probability a half; if the change is in covariance, the coefficient ρ changes from 0.6 to $0.6 - \Delta\rho$. Table 4.4 below lists the values of δ, Δa, and $\Delta\rho$, as the size of change increases. We run the experiments under three scenarios: (i) change occurs in all 100 seuqences, (ii) change occurs in 20 sequences, and (iii) change occurs in only 1 sequence. The results are shown in Figure 4.5.

From Figure 4.5, we see that when the changes occur in all $N = 100$ sequences, all the three methods are able to detect the changes. In addition, our MS-statistic has the higher power and accuracy at any given size of change. When the changes occur in 20 of the sequences, or only in one of the sequences, both gSeg and ecp fail to detect such changes. On the other hand, our new method still has high power and decent accuracy in those two scenarios.

In Section 4.4, we learn from extensive simulation studies that our proposed MS-statistic is capable of detecting

Table 4.3: Numbers of times (out of 100) that the null hypothesis is rejected under significance level $\alpha = 0.05$. In the parentheses are the numbers of times the estimated change-point $\hat{\tau}$ is within 50 of the true change-point $\tau = 300$, i.e., $\hat{\tau} \in [275, 325]$, which can be interpreted as "accuracy."

Scenario I: Change in mean in one sequence (only one coordinate)

δ	0.00	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	1.80	2.00
New	3 (0)	14 (0)	8 (0)	10 (2)	26 (8)	52 (30)	86 (57)	97 (84)	100 (92)	100 (98)	100 (100)
gSeg	8 (0)	3 (1)	5 (0)	5 (0)	9 (0)	21 (5)	34 (18)	74 (41)	81 (52)	93 (71)	99 (87)
ecp	5 (0)	9 (0)	4 (1)	10 (1)	5 (0)	17 (6)	44 (33)	86 (71)	100 (94)	100 (94)	100 (97)

Scenario II: Change in variance in one sequence

a	1.00	1.04	1.08	1.12	1.16	1.20	1.24	1.28	1.32	1.36	1.40
New	3 (0)	15 (0)	9 (0)	13 (3)	26 (10)	47 (21)	52 (21)	72 (35)	87 (44)	93 (48)	98 (67)
gSeg	8 (0)	4 (2)	6 (0)	6 (2)	8 (2)	10 (3)	15 (3)	24 (5)	20 (8)	36 (7)	38 (11)
ecp	5 (0)	7 (1)	6 (0)	11 (4)	14 (3)	11 (3)	15 (1)	17 (1)	16 (6)	18 (9)	27 (5)

Scenario III: Mean change + Variance change (one sequence)
(mean 1.0 in one coordinate, variance same as Scenario II)

a	1.00	1.04	1.08	1.12	1.16	1.20	1.24	1.28	1.32	1.36	1.40
New	48 (28)	47 (25)	51 (29)	68 (38)	66 (36)	75 (43)	81 (54)	93 (57)	97 (61)	94 (55)	99 (70)
gSeg	19 (4)	27 (10)	28 (7)	32 (12)	26 (11)	25 (11)	35 (11)	43 (14)	47 (13)	52 (10)	61 (17)
ecp	19 (5)	18 (8)	33 (15)	25 (9)	30 (16)	32 (13)	40 (26)	46 (23)	49 (24)	57 (37)	67 (39)

Scenario IV: Change in covariance structure in one sequence

$\Delta\rho$	0.00	0.03	0.06	0.09	0.12	0.15	0.18	0.21	0.24	0.27	0.30
New	3 (0)	10 (1)	7 (3)	8 (0)	16 (4)	28 (10)	46 (16)	69 (39)	89 (62)	99 (82)	100 (95)
gSeg	8 (0)	5 (1)	7 (0)	4 (0)	4 (0)	3 (0)	13 (0)	14 (2)	9 (4)	8 (0)	18 (0)
ecp	5 (0)	10 (0)	4 (0)	10 (0)	4 (1)	3 (0)	4 (0)	7 (1)	7 (1)	5 (1)	3 (0)

Scenario V: Chi-square distribution, change occurs in one sequence

a	1.00	1.02	1.04	1.06	1.08	1.10	1.12	1.14	1.16	1.18	1.20
New	3 (0)	7 (1)	12 (3)	42 (17)	67 (30)	87 (58)	98 (66)	100 (71)	100 (86)	100 (83)	100 (90)
gSeg	7 (0)	8 (1)	9 (0)	18 (5)	24 (3)	31 (6)	46 (13)	60 (30)	73 (30)	81 (44)	82 (42)
ecp	4 (0)	6 (0)	8 (0)	6 (0)	8 (0)	4 (1)	7 (1)	8 (0)	7 (0)	8 (0)	6 (1)

Table 4.4: Values of parameters used in each simulation run for Scenario (i) and (ii). Larger values, in particular ($\delta \to 4\delta, \Delta a \to 3\Delta a, \Delta\rho \to 2\Delta\rho$) are used for Scenario (iii).

size index	0	1	2	3	4	5	6	7	8	9	10
δ	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Δa	0.00	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20
$\Delta\rho$	0.00	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20

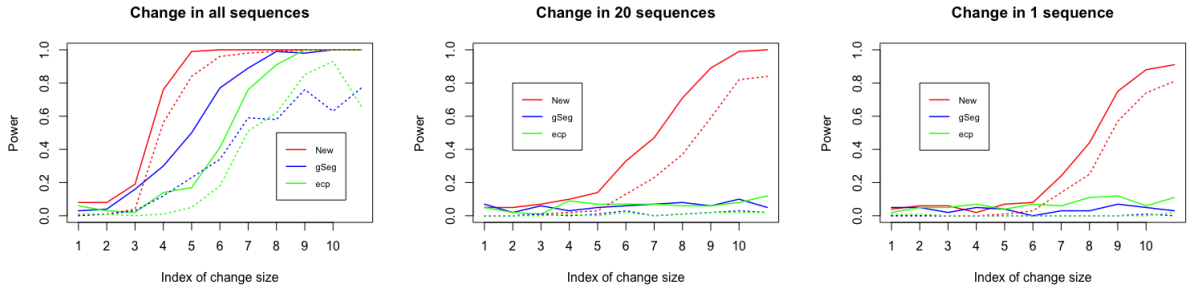


Figure 4.5: Power of the three methods for changes occur in different number of sequences when the total number of sequences is $N = 100$. The power is computed based on 100 simulations at significance level $\alpha = 0.05$. The dotted lines represent the number of times that the change-point location is correctly detected. That is, $\hat{\tau} \in [275, 325]$ where the ture change-point is at $\tau = 300$.

various types of changes, from mean to variance to covariance structure. In addition, its performance is robust to different underlying distributions, symmetric or asymmetric. Our MS-statistic has contant high power no matter the changes are in all sequences, a subset of the sequences, or only a few of the sequences.

4.5 Real Data Application

To illustrate the effectiveness of the MS-statistic, we apply our method to the dataset of the yellow taxi trip records, which is publicly available on the NYC Taxi & Limousine Commission (TLC) website (`http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`). The trip records contain abundant information, such as taxi pickup and drop-off dates and times, longitude and latitude coordinates of pickup and drop-off locations, trip distances, fares, rate types, payments types, and driver-reported passenger counts.

The dataset is so rich that many studies can root on. Here, we demonstrate the new approach in detecting changes in travel to the John F. Kennedy International Airport over the year. The boundary of JFK airport was set to be 40.63 to 40.66 latitude and 73.80 to 73.77 longitude for easy reference. We use the year-end data from 2011 to 2015. The reason that we include the data up to the year of 2015 is because the longitutde and latitude coordinates are no longer available in 2016 and after.

For those trips having a destination at JFK airport, i.e., a drop-off coordinates within the boundary of the airport,

we extract information on where those trips begin, their longitude and latitude pickup coordinates. Using the longitude/latitude coordinates, we create a 30 by 30 grid of New York City and count the number of taxi pickups that fall within each cell, where each cell represents a rectangular longitude, latitude coordinate range. Then for each day, we have a 30 by 30 matrix representing the number of taxi pickups in each location. Therefore, we have five sequences, from year 2011 to 2015, respectively. Each sequence consists of 365 such matrices from January 1st and December 31st. In the year of 2012, which has 366 days, the matrix for Feburary 29th is excluded from the sequence.

We apply the MS-statistic on the five sequences of observations, where each observation is a 30 by 30 matrix. We use the L_2 norm distance among the observations to construct the MST graphs for each sequences. The MS-statistic is then computed based on those derived graphs.

As there might be more than one change-points in the sequences, if the first change-point is found, we apply the method again to the two subsequences separated by the first change-point. With significance level $\alpha = 0.01$, the method is then applied iteratively until no more change-points can be found. After the initial set of candidate change-points are determined. A change-point refinement procedure is then performed to prune the quality of the change-points.

Suppose the initial set contains K change-points denoted by $1 < t_1 < t_2 < \cdots < t_K < 365$. Let $t_0 \equiv 1$ and $t_{K+1} \equiv 366$, we perform the following step to first refine the estiamtes of change-points. For each change-point t_k, $k = 1, \ldots, K$, the change-point detection approach is applied to the interval $[t_{k-1}, t_{k+1})$ to refine the estimate, and the refined change-point is denoted by $t_k^{(1)}$.

If the set of refined change-points are different from the initial set, we further check if there is any change-point in each sub-interval $[t_k^{(1)}, t_{k+1}^{(1)})$, $k = 0, 1, \ldots, K$, with the type I error being controlled at $0.01/K$. Let $K^{(2)}$ be the number of candidate change-points after the re-searching and we denote these change-points by $1 < t_1^{(2)} < t_2^{(2)} < \cdots < t_{K^{(2)}}^{(2)} < 365$. Now we treat this new set of change-points as the initial set and repeat the refinement and re-searching procedures until the candidate set converges.

After the candidate set is finalized, we do one last step to prune the change-points, that is, $t_k^{(2)}$, $k = 1, \ldots, K^{(2)}$ is kept only if the observations in the time interval $[t_{k-1}^{(2)}, t_{k+1}^{(2)})$ is significatly non-homonegeous. In particular, here we use the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001) to control the false discover rate at $\alpha = 0.01$. Throughout the process, n_0 is chosen to be either 5 or 5% of the sequence length, whichever is larger, and n_1 is chosen symmetrically. That is, let n_s be the length of any sequence the graph-based method is applied to, then $n_0 = \max(5, \lfloor 0.05 n_s \rfloor)$, and $n_1 = n_s - n_0$.

The result of the set of change-points found by the MS-statisitc is presented in Table 4.5. We see that in general both methods detect similar change-point locations.

To perform sanity check on those change-pooints found. We plot the heatmaps of the five sequences (Figure 4.6) on L_2 norm distance. We further zoom in those heatmaps by partitioning the whole sequences into three non-overlapping

Table 4.5: Final sets of change-points (and their corresponding dates) found by the MS-statistic after refinement and pruning, under significance level $\alpha = 0.01$.

	Set of detected change-points								
MS-statisitc	44 Feb.13	127 May7	146 May26	183 Jul.2	188 Jul.7	249 Sep.6	301 Oct.28	352 Dec.18	357 Dec.23

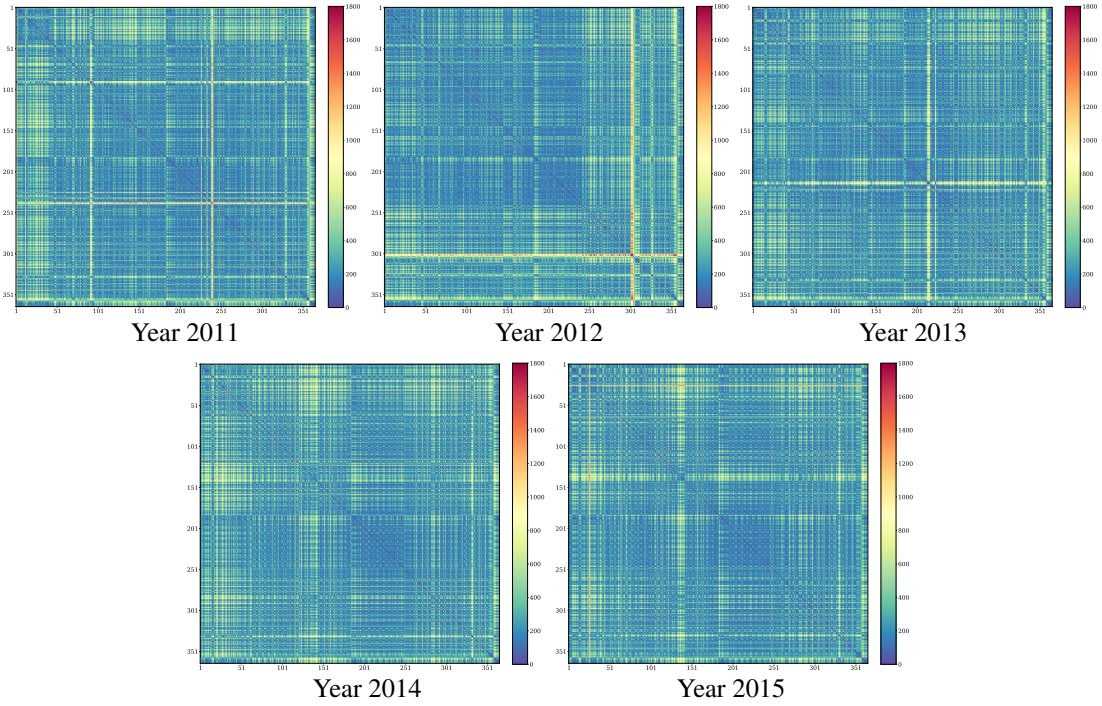


Figure 4.6: Heatmaps of the five seuqences on L_2 norm distances, from day 1 to day 365.

subseqeunces to get a closer look at their relative distances. The three subsequences are from day 1 to day 150 (Figure 4.7), day 151 to day 300 (Figure 4.8), and day 301 to 365 (Figure 4.9), respectively.

From the first subsequences (Figure 4.7, day 1 to day 150), we see that the first change-point at Feb. 13 (day 44) reflects mainly the change in the year of 2011, and the second change-point at May 7 (day 127) possibly reflects weak changes in the year of 2013, 2014 and 2015. From the second subsequences (Figure 4.8, day 151 to day 300), the change in the interval July 2 to July 7 (day 183 to 188) is common in all the five years, and the next change-point at Sep. 6 (day 249) is mostly contributed by the year of 2012 and 2013. Finally in the third subsequences (Figure 4.9, day 301 to day 365), the last change-point of Dec. 23 (day 357) is obvious in all sequences, after which are the Christmas holidays. The other change-point in this segment Dec. 18 (day 352), though weaker, is also suggested by all the five sequences.

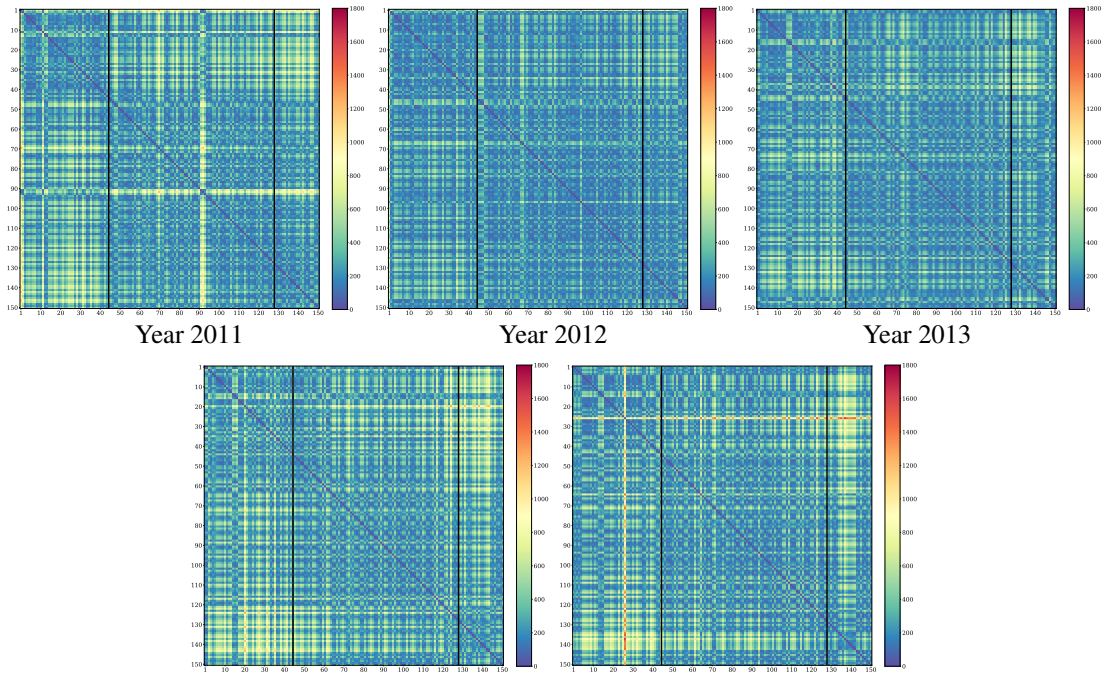


Figure 4.7: Heatmaps of the five seuqences on L_2 norm distances, from day 1 to day 150. Black lines indicate the locations of change-points detected by the MS-statistic.

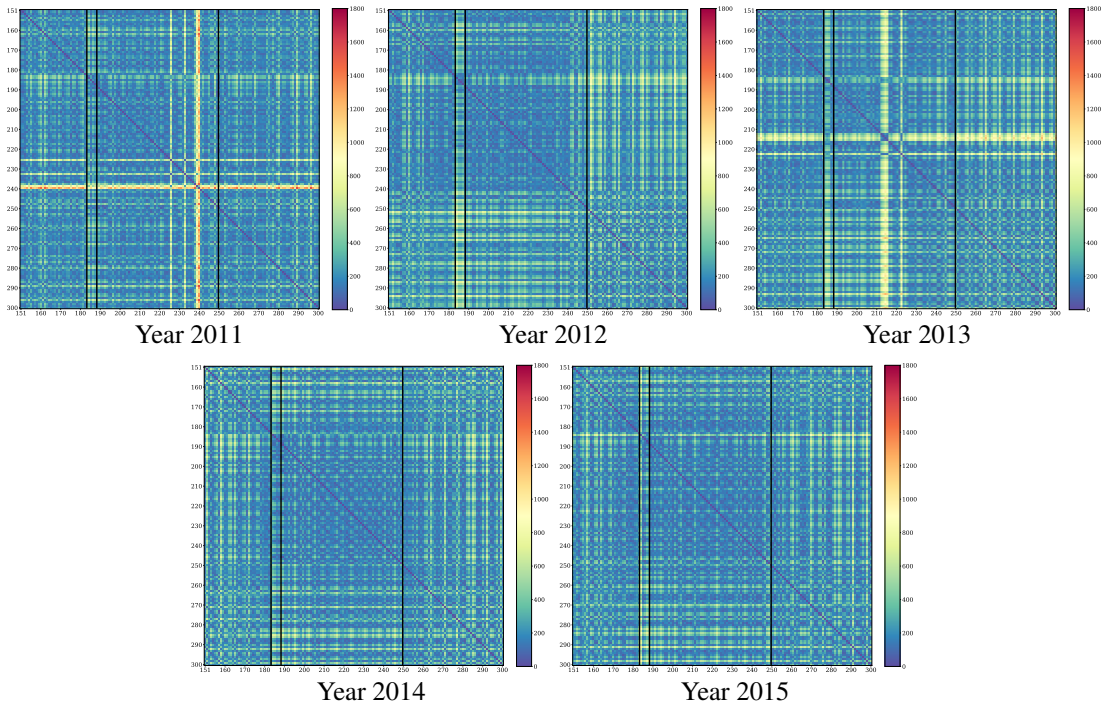


Figure 4.8: Heatmaps of the five seuqences on L_2 norm distances, from day 151 to day 300. Black lines indicate the locations of change-points detected by the MS-statistic.

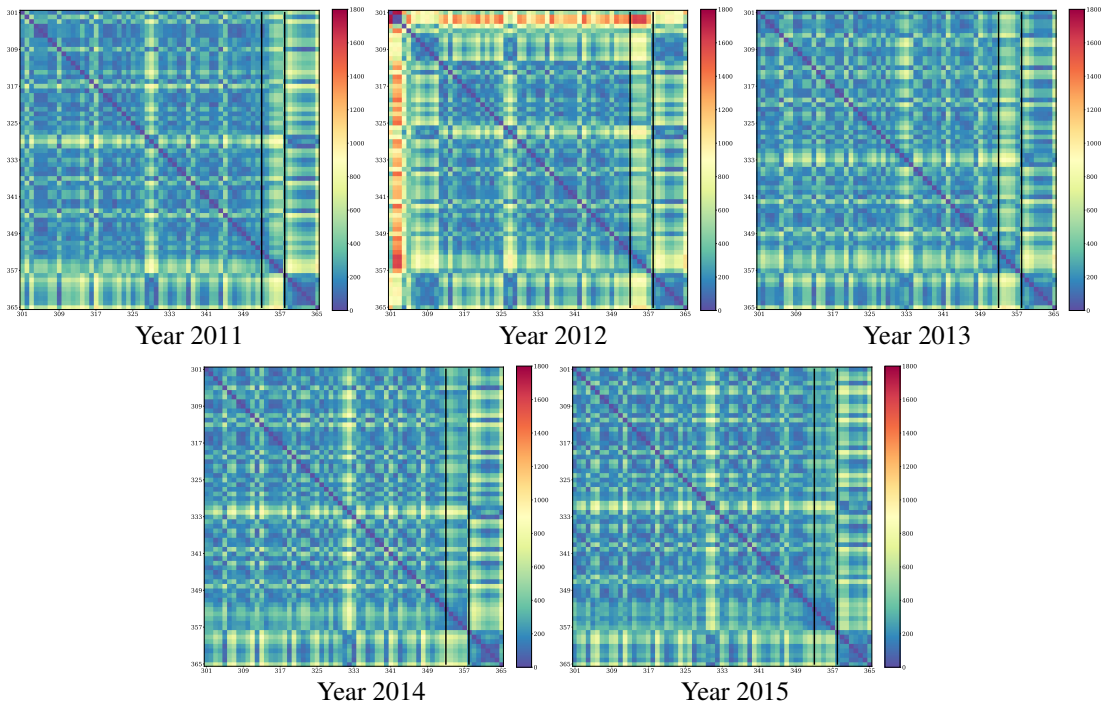


Figure 4.9: Heatmaps of the five seuqences on L_2 norm distances, from day 301 to day 365. Black lines indicate the locations of change-points detected by the MS-statistic.

4.6 Conclusion

We propose a new nonparametric, graph-based change-point detection method using the MS-statistic to detect simultaneous change-points in multiple sequences of observations. Our method can be applied to datasets with any number of sequences in arbitrary dimensions, with a fast type I error control. The method can also be applied to non-Euclidean data as long as a proper dissimilarity measure can be defined among the observations. In constructing our MS-statistic, we utilize the edge-counts information from the similarity graphs for each sequence of the data observations so that it is sensitive to changes that happen simultaneously in all the sequences, as well as in a small subset of the sequences. To make our method instant-applicable to large scale datasets, we derive the asymptotic analytic fromula to approximate the p-values of our test. This asymptotic formula is distribution-free and performs well in finite sample as shown by the simulations in Section 4.3.

Existing change-point detection methods for multiple sequences are limited many ways. For example, the method in Zhang et al. (2010) is effective in detecting locational alternatives in sequences of univariate Gaussian observations. The nonparametric ecp method (Matteson and James, 2014) can be applied to high-dimensional observations but is only effective in detecting changes in the mean vectors. In addition, when the mean changes are only presented in a subset of the sequences, the ecp method could also have low power. The graph-based (gSeg) max-type statistic (Chu and Chen, 2019) is more flexible than ecp but could fail when the sequences exhibit different types of changes, or the changes

are only in few of the sequences. Our proposed MS-statistic improves upon those two methods by collecting signals of changes from individual sequences separately, and hence is more powerful in the presence of multiple sequences. Moreover, the MS-statistic can detect changes that happen in all of the sequences, as well as in only a small subset of the sequences. We apply our new method to the yellow taxi trip records in Section 4.5. This example demonstrates the effectiveness of our method in detecting change-points in multiple sequences of observations.

Chapter 5

Conclusions

5.1 Summary of Contributions

Graph-based change-point detection is a nonparametric framework that utilizes the edge-count information on a similarity graph constructed on the observations. It can be easily applied to data of arbitrary dimension or even non-Euclidean data as long as a proper dissimilarity measure can be defined among the observations. We propose three versions of the graph-based methods that tackle various difficulties one may encounter in many real applications, making the methods more flexible and compatible with a wide range of modern change-point analysis tasks.

The first version improves the time efficiency of the algorithms, which makes the method more desirable especially when dealing with long data sequences or high-dimensional data. In the example of the fRMI sequence of length-598 and dimensionality more than 400 thousands, our new method is more than 30 folds faster than other compelling nonparametric methods. We incorporates the approximate k-NN information into the graph-based framework and the total time complexity of our new method can be achieved in $O\left(dn(\log n + k \log d) + nk^2\right)$ time. We work out the analytic expressions of the edge-count test statistics, and derive the analytic p-value approximation formulas under directed k-NN graphs. Our new method has proper control on the false discovery rate, and has power higher than or at least on par with other competitive nonparametric methods. In addition, out new test is sensitive to various types of alternatives, such as changes in mean, variance, covariance, skewness and kurtosis.

The second version handles data with local dependency. We incorporates the circular block permutation (CBP) framework into the three (weighted/generalized/max-type) edge-count test statistics that improve on the original one. The main finding is that under CBP, a new weight function that depends on the graph and the block size used in CBP should be adopted. We derive the expression of this new optimal weight, and redefine the three edge-count test statistics accordingly. To make the methods instant-applicable, we derive the analytic expressions to compute those test statistics

under CBP, and their corresponding formulas for type I error control. Simulation studies show that the new tests have good control on the false discovery rate for autocorrelated data, and the power of the new weighted edge-count test is higher than that of the old one.

The third version provides a new tool for change-point analysis in multiple sequences of high-dimensional/non-Euclidean observations. We construct similarity graphs for each sequence of observations, and design a new MS-statistic that collects and accumulates signals from all the graphs. The new test can detect simultaneous changes even if the types of changes in each of the sequences are different. In addition, the test is powerful in detecting changes that occur in all or a subset of the sequences. We derive the analytic formulas for type I error control based on the asymptotic properties of the test statistic. The good performance of the analytic formulas is robust to the number of sequences, data dimensionality, and the underlying distributions of the data observations.

All three methods embrace the challenges for modern data analysis because they are all sensitive in detecting various types of changes, and have analytic formulas to control the type I error. The three methods can be adopted individually or as a combination. For example, in the task of change-point detection in multiple sequences, we may construct the directed approximate k-NN graphs for each sequence, and compute the MS-statistic under circular block permutation. Any combination of two is also an option depending on the needs of the applications. Despite the concentration on single change-point detection, when there are multiple change-points, seeded binary segmentation (Kovács et al., 2020) or wild binary segmentation (Fryzlewicz et al., 2014; Fryzlewicz, 2020) can be integrated.

5.2 Future Directions

There are lots of studies on the offline change-point detection problems, where the length of the observations is fixed and the goal is to detect whether there is a change-point within the sequence. However, in many real applications, it is equally important to consider the online change-point detection problems, where the historical data may not contain any change-point. However, as new data points are constantly collected, the goal is to detect as soon as possible whenever a change-point occurs. For example, to monitor whether a machine or device is working properly, we may use sensors to collect signals from the subjects, hoping that an alert would be made if the signals start behaving abnormally. In contrast to controlling the type I error in the offline setting, the online change-point detection supervises the average run length under the null hypothesis. Chen (2019b) has proposed a framework for graph-based online change-point detection. When new observations come in, the similarity graph changes accordingly. The method models the dynamic of the similarity graphs under the independence assumption. When data are autocorrelated, the average run length under the null hypothesis of no change-point could be much shorter than the target length we want to control. My current endeavors include conducting research on online change-point detection methods for locally dependent data.

The challenge lies in modeling the graph dynamic with some unknown dependence structure, which we may not have enough data observations to estimate in the online setting.

Also, for the graph-based change-point detection framework, the choice of dissimilarity measures and similarity graphs have not been much explored. For the former, in addition to the commonly seeing Euclidean distance or the L_1 norm distance, any other proper measures can be adopted as advised by domain experts. Depending on what data we are working with, and/or what types of changes we are concerned with, a problem-specific similarity measure could improve the performance of change-point analysis. For the latter, take the directed approximate k-NN as an example. The choice of k remains an open question. We want the graph to be dense enough to provide substantial information but not so dense that too many noises may be included. In addition to the unweighted k-NN that treats all the k edges from a certain node equally. We may also add weights to those edges to enrich the information of similarity graphs.

Besides the above-mentioned research directions that I will explore in the near future, my long-term goal is to conduct research that provides powerful tools for analyzing modern complex data and investigate their potential in various fields, including financial data analysis, risk modeling, and applications in bioscience.

Appendix A

Appendix to Chapter 2

A.1 Proof of Theorem 1

Let $\pi(i)$ be the index of observation $\mathbf{y}_i$ after permutation, where $i = 1, \cdots, n$. Then

$$\mathsf{E}(R_{G,1}(t)) = \mathsf{E}\left(\sum_{(i,j)\in G} \mathbb{1}_{\{i\le t, j\le t\}}\right) = \sum_{(i,j)\in G} \mathsf{P}(\pi(i)\le t, \pi(j)\le t) = nk\tfrac{t(t-1)}{n(n-1)},$$
$$\mathsf{E}(R_{G,2}(t)) = \mathsf{E}\left(\sum_{(i,j)\in G} \mathbb{1}_{\{i> t, j> t\}}\right) = \sum_{(i,j)\in G} \mathsf{P}(\pi(i)> t, \pi(j)> t) = nk\tfrac{(n-t)(n-t-1)}{n(n-1)}.$$

and for variances, it suffices to derive $\mathsf{E}(R^2_{G,1}(t))$ and $\mathsf{E}(R^2_{G,2}(t))$ since

$$\mathsf{Var}(R_{G,1}(t)) = \mathsf{E}(R^2_{G,1}(t)) - \mathsf{E}(R_{G,1}(t))^2,$$
$$\mathsf{Var}(R_{G,2}(t)) = \mathsf{E}(R^2_{G,2}(t)) - \mathsf{E}(R_{G,2}(t))^2.$$

For a pair of edges $(i,j),(u,v)\in G$, the probability of having $\{\pi(i),\pi(j),\pi(u),\pi(v)\le t\}$ is equivalent to having all four nodes being placed before t after permutation. This probability only depends on the number of distinct nodes in the pair of edges. On a directed approximate k-NN graph, there are $c^{(1)}+c^{(2)}$ pairs of edges that share two nodes, and the probability of $\{\pi(i),\pi(j),\pi(u),\pi(v)\le t\}$ is: $p_1(t) = \frac{t(t-1)}{n(n-1)}$. There are $c^{(3)}+c^{(4)}+c^{(5)}+c^{(6)}$ pairs of edges that share only one node, and for all the three distinct nodes being placed before t after permutation, the probability is: $p_2(t) = \frac{t(t-1)(t-2)}{n(n-1)(n-2)}$. Finally, there are $c^{(7)}$ pairs of edges that share no node, and the probability of having all the four

nodes before t is: $p_3(t) = \frac{t(t-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)}$. Hence,

$$\mathsf{E}(R^2_{G,1}(t)) = \mathsf{E}\left(\sum_{(i,j),(u,v)\in G} \mathbb{1}_{\{i,j,u,v\leq t\}}\right) = \sum_{(i,j),(u,v)\in G} \mathsf{P}(\pi(i),\pi(j),\pi(u),\pi(v)\leq t)$$
$$= \left(c^{(1)}+c^{(2)}\right)\frac{t(t-1)}{n(n-1)} + \left(c^{(3)}+c^{(4)}+c^{(5)}+c^{(6)}\right)\frac{t(t-1)(t-2)}{n(n-1)(n-2)} + c^{(7)}\frac{t(t-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)}.$$

For the event $\{\pi(i),\pi(j),\pi(u),\pi(v) > t\}$, we require all the nodes to be placed after t, instead. Therefore, the corresponding probabilities are: $q_1(t) = \frac{(n-t)(n-t-1)}{n(n-1)}$, $q_2(t) = \frac{(n-t)(n-t-1)(n-t-2)}{n(n-1)(n-2)}$, and $q_3(t) = \frac{(n-t)(n-t-1)(n-t-2)(n-t-3)}{n(n-1)(n-2)(n-3)}$, respectively. Hence,

$$\mathsf{E}(R^2_{G,2}(t)) = \mathsf{E}\left(\sum_{(i,j),(u,v)\in G} \mathbb{1}_{\{i,j,u,v>t\}}\right) = \sum_{(i,j),(u,v)\in G} \mathsf{P}(\pi(i),\pi(j),\pi(u),\pi(v) > t)$$
$$= \left(c^{(1)}+c^{(2)}\right)\frac{(n-t)(n-t-1)}{n(n-1)} + \left(c^{(3)}+c^{(4)}+c^{(5)}+c^{(6)}\right)\frac{(n-t)(n-t-1)(n-t-2)}{n(n-1)(n-2)}$$
$$+ c^{(7)}\frac{(n-t)(n-t-1)(n-t-2)(n-t-3)}{n(n-1)(n-2)(n-3)}.$$

For the covariance between $R_{G,1}(t)$ and $R_{G,2}(t)$, we have

$$\mathsf{Cov}(R_{G,1}(t), R_{G,2}(t)) = \mathsf{E}(R_{G,1}(t)R_{G,2}(t)) - \mathsf{E}(R_{G,1}(t))\mathsf{E}(R_{G,2}(t)).$$

When $(i,j),(u,v)\in G$ share at least one node, it is not possible to have the edge (i,j) connecting both observatiosn before t and the edge (u,v) connecting both observations after t. Thus,

$$\mathsf{E}(R_{G,1}(t)R_{G,2}(t)) = \mathsf{E}\left(\sum_{(i,j),(u,v)\in G} \mathbb{1}_{\{i,j\leq t,u,v>t\}}\right)$$
$$= \sum_{(i,j),(u,v)\in G} \mathsf{P}(\pi(i),\pi(j)\leq t,\pi(u),\pi(v) > t) = c^{(7)}\frac{t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)},$$

and $\mathsf{Cov}(R_{G,1}(t), R_{G,2}(t))$ follows accordingly.

A.2 Proof of Theorem 2

Given that $R_w(t) = \frac{n-t-1}{n-2}R_{G,1}(t) + \frac{t-1}{n-2}R_{G,2}(t)$ and $R_{\text{diff}}(t) = R_{G,1}(t) - R_{G,2}(t)$, we have

$$\mathsf{Var}(R_w(t)) \quad = \quad \left(\frac{n-t-1}{n-2}\right)^2 \mathsf{Var}(R_{G,1}(t)) + \left(\frac{t-1}{n-2}\right)^2 \mathsf{Var}(R_{G,2}(t))$$

$$+2\left(\tfrac{n-t-1}{n-2}\right)\left(\tfrac{t-1}{n-2}\right)\mathsf{Cov}(R_{G,1}(t),R_{G,2}(t)),$$

$$\mathsf{Var}(R_{\text{diff}}(t)) \quad = \quad \mathsf{Var}(R_{G,1}(t))+\mathsf{Var}(R_{G,2}(t))-2\mathsf{Cov}(R_{G,1}(t),R_{G,2}(t)).$$

Plugging in the results from Theorem 1, $\mathsf{Var}(R_w(t))$ can be derived and further simplified to be

$$\begin{aligned}\mathsf{Var}(R_w(t)) \quad &= \quad \frac{t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)^2(n-3)} \\ &\quad \times\left(\frac{2n^2k^2}{n-1}+(n-4)\left(c^{(1)}+c^{(2)}\right)-\left(c^{(3)}+c^{(4)}+c^{(5)}+c^{(6)}\right)\right).\end{aligned} \tag{A.1}$$

From Theorem 1, we always have $2c^{(2)}+c^{(3)}+c^{(4)}+c^{(5)}+c^{(6)} = 3nk^2-2nk+\sum_{i=1}^n|D_i|^2$ for a directed approximate k-NN, so we have

$$\begin{aligned}&\frac{2n^2k^2}{n-1}+(n-4)(c^{(1)}+c^{(2)})-(c^{(3)}+c^{(4)}+c^{(5)}+c^{(6)}) \\ &\quad=\frac{2n^2k^2}{n-1}+(n-4)c^{(1)}+(n-2)c^{(2)}-\left(2c^{(2)}+c^{(3)}+c^{(4)}+c^{(5)}+c^{(6)}\right) \\ &\quad=\frac{2n^2k^2}{n-1}+(n-4)nk+(n-2)c^{(2)}-\left(3nk^2-2nk+\sum_{i=1}^n|D_i|^2\right) \\ &\quad\geq\frac{2n^2k^2}{n-1}+(n-4)nk+(n-2)2k^2-\left(3nk^2-2nk+(n-1)^2k+k^2\right) \\ &\quad=\left(\frac{n^2-4n+5}{n-1}\right)k^2-k.\end{aligned} \tag{A.2}$$

For a directed approximate k-NN, (A.2) is minimized when k nodes has in-degree $n-1$, one node has in-degree k, and all other nodes have in-degree zero. In this case, $\sum_{i=1}^n|D_i|^2=(n-1)^2k+k^2$, and $c^{(2)}$ is at least $2k^2$. Therefore, when $n\geq 5$, we have

$$\mathsf{Var}(R_w(t))=\frac{t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)^2(n-3)}\left(\left(\frac{n^2-4n+5}{n-1}\right)k^2-k\right)>0.$$

Plugging in the results from Theorem 1, $\mathsf{Var}(R_{\text{diff}}(t))$ can be derived and further simplified to be

$$\mathsf{Var}(R_{\text{diff}}(t))=\frac{t(n-t)}{n(n-1)}\left(\sum_{i=1}^n|D_i|^2-nk^2\right). \tag{A.3}$$

For a directed approximate k-NN, we must have $\sum_{i=1}^n|D_i|=nk$. Under this constraint, $\sum_{i=1}^n|D_i|^2$ is minimized when all the observations have the same in-degree, i.e., $|D_i|=k, \forall i=1,\cdots,n$. Thus, $\mathsf{Var}(R_{\text{diff}}(t))>0$ as long as there exists one i such that $|D_i|\neq k$.

A.3 Proof of Theorem 3

Here, we show that $\{Z_w([nu]) : 0 < u < 1\}$ and $\{Z_{\text{diff}}([nu]) : 0 < u < 1\}$ converge to independent Gaussian processes in finite dimensional distributions under the conditions in Theorem 3. We first show that

$$(Z_w([nu_1]), \ldots, Z_w([nu_L]), Z_{\text{diff}}([nu_1]), \ldots, Z_{\text{diff}}([nu_L]))$$

converges to a multivariate Gaussian distribution as $n \to \infty$ for and $0 < u_1 < u_2 < \cdots < u_L < 1$ and for any fixed L. We then show that $\mathsf{Cov}(Z_w(u), Z_{\text{diff}}(v)) = 0$ for any $0 < u, v < 1$ as $n \to \infty$. For notation simplicity, let $t_l = [nu_l]$, $l = 1, \ldots, L$.

To prove $(Z_w(t_1), \ldots, Z_w(t_L), Z_{\text{diff}}(t_1), \ldots, Z_{\text{diff}}(t_L))$ converges to a multivariate Gaussian distribution, we revisit the permutation distribution. In permutation distribution, we permute the order of the observations. Let $\pi(i)$ be the observed time of $\mathbf{y}_i$ after permutation. Then $(\pi(1), \pi(2), \ldots, \pi(n))$ is a permutation of $1, \ldots, n$. To obtain the permutation distribution, we can do it in two steps: (1) For each i, $\tilde{\pi}(i)$ is sampled uniformly from $1, \ldots, n$; (2) only those $(\tilde{\pi}(1), \tilde{\pi}(2), \ldots, \tilde{\pi}(n))$ such that each value in $\{1, \ldots, n\}$ is sampled exactly once are retained. We can see that each permutation has the same occurrence probability after these two steps. We call the distribution resulting from only performing the first step the bootstrap distribution, and use $\mathsf{P}_{\text{B}}, \mathsf{E}_{\text{B}}, \mathsf{Var}_{\text{B}}, \mathsf{Cov}_{\text{B}}$ to denote the probability, expectation, variance, and covariance under bootstrap distribution, respectively. In this section, the corresponding quantities with the subscript P are used to denote the equivalences under the permutation distribution.

Let $d^{(1)} = \left(c^{(1)} + c^{(2)}\right)$, $d^{(2)} = \left(c^{(3)} + c^{(4)} + c^{(5)} + c^{(6)}\right)$, $d^{(3)} = c^{(7)}$. Given that the $\tilde{\pi}(i)$'s are independent under the bootstrap null distribution, we have

$$\begin{aligned}
&\mathsf{E}_{\text{B}}(R_{G,1}(t)) = \frac{t^2}{n^2}|G|, \\
&\mathsf{E}_{\text{B}}(R_{G,2}(t)) = \frac{(n-t)^2}{n^2}|G|, \\
&\mathsf{Var}_{\text{B}}(R_{G,1}(t)) = \frac{t^2}{n^2}d^{(1)} + \frac{t^3}{n^3}d^{(2)} + \frac{t^4}{n^4}d^{(3)} - \left(\frac{t^2}{n^2}|G|\right)^2, \\
&\mathsf{Var}_{\text{B}}(R_{G,2}(t)) = \frac{(n-t)^2}{n^2}d^{(1)} + \frac{(n-t)^3}{n^3}d^{(2)} + \frac{(n-t)^4}{n^4}d^{(3)} - \left(\frac{(n-t)^2}{n^2}|G|\right)^2, \\
&\mathsf{Cov}_{\text{B}}(R_{G,1}(t), R_{G,2}(t)) = \frac{t^2(n-t)^2}{n^4}d^{(3)} - \left(\frac{t(n-t)}{n^2}|G|\right)^2.
\end{aligned}$$

For $R_w(t) = \frac{n-t-1}{n-2}R_{G,1}(t) + \frac{t-1}{n-2}R_{G,2}(t)$, we have

$$\mathsf{E}_{\text{B}}(R_w(t)) = \frac{t^2(n-t-1) + (n-t)^2(t-1)}{n^2(n-2)}|G| := \mu_w^{\text{B}}(t)|G|,$$

$$\begin{aligned}\mathsf{Var}_{\mathrm{B}}(R_w(t)) = \frac{d^{(1)}}{n^4(n-2)^2}\Big(t^2(n+t)(n-t)(n-t-1)^2 - \big((n-t)^4 - n^2(n-t)^2\big)(t-1)^2 \\ - 2t^2(n-t)^2(t-1)(n-t-1)\Big) + \frac{d^{(2)}}{n^4(n-2)^2}\Big(t(n-t)(2t^2-2nt+n)^2\Big) \\ + \frac{|G|^2}{n^4(n-2)^2}\left((n-t)^2(t-1) - t^2(n-t-1)\right)^2 := \left(\sigma_w^{\mathrm{B}}(t)\right)^2.\end{aligned}$$

For $R_{\text{diff}}(t) = R_{G,1}(t) - R_{G,2}(t)$, we have

$$\begin{aligned}\mathsf{E}_{\mathrm{B}}(R_{\text{diff}}(t)) &= \frac{2t-n}{n}|G| := \mu_{\text{diff}}^{\mathrm{B}}(t)|G|, \\ \mathsf{Var}_{\mathrm{B}}(R_{\text{diff}}(t)) &= \frac{t(n-t)}{n^2}\left(\sum_{i=1}^{n}|D_i|^2 + 3nk^2\right) := \left(\sigma_w^{\mathrm{B}}(t)\right)^2.\end{aligned}$$

Let

$$\begin{aligned}Z_w^{\mathrm{B}}(t) &= \frac{R_w(t) - \mathsf{E}_{\mathrm{B}}(R_w(t))}{\sqrt{\mathsf{Var}_{\mathrm{B}}(R_w(t))}}, \\ Z_{\text{diff}}^{\mathrm{B}}(t) &= \frac{R_{\text{diff}}(t) - \mathsf{E}_{\mathrm{B}}(R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}_{\mathrm{B}}(R_{\text{diff}}(t))}}, \\ X^{\mathrm{B}}(t) &= \frac{n^{\mathrm{B}}(t) - t}{\sqrt{t(1-t/n)}}, \text{ where } n^{\mathrm{B}}(t) = \sum_{i=1}^{n}\mathbb{1}_{\{\tilde{\pi}(i)\le t\}}.\end{aligned}$$

To prove $(Z_w(t_1), \ldots, Z_w(t_L), Z_{\text{diff}}(t_1), \ldots, Z_{\text{diff}}(t_L))$ converges to a multivariate Gaussian distribution under the conditions on the graph in Theorem 1, as $n \to \infty$, we only need to prove the following two lemmas:

Lemma 3. *When $k = O(n^\beta)$, $\beta < 0.25$, $\sum_{e\in G}|A_e||B_e| = o(n^{1.5(\beta+1)})$, $\sum_{e\in G}|A_e|^2 = o(n^{\beta+1.5})$, and $\sum_{i=1}^n |D_i|^2 - nk^2 = O(\sum_{i=1}^n |D_i|^2)$, for $0 < u_1, u_2, \ldots, u_L < 1$, as $n \to \infty$, under the bootstrap distribution,*

$$\left(Z_w(t_1), \ldots, Z_w(t_L), Z_{\textit{diff}}(t_1), \ldots, Z_{\textit{diff}}(t_L), X^B(t_1), \ldots, X^B(t_L)\right) \tag{A.4}$$

is multivariate normal and the covariance matrix of

$$\left(X^B(t_1), X^B(t_2), \ldots, X^B(t_L)\right)$$

is positive definite.

Lemma 4. *We have,*

1. $\dfrac{\mathsf{Var}_B(R_w(t))}{\mathsf{Var}_P(R_w(t))} \to 1$

2. $\dfrac{\mathsf{Var}_B(R_{diff}(t))}{\mathsf{Var}_P(R_{diff}(t))} \to c_1$
3. $\dfrac{\mathsf{E}_B(R_w(t)) - \mathsf{E}_P(R_w(t))}{\sqrt{\mathsf{Var}_P(R_w(t))}} \to 0$
4. $\dfrac{\mathsf{E}_B(R_{diff}(t)) - \mathsf{E}_P(R_{diff}(t))}{\sqrt{\mathsf{Var}_P(R_{diff}(t))}} \to 0$

where c_1 is a positive constant.

From Lemma 3

$$\big(Z_w(t_1), \ldots, Z_w(t_L), Z_{\text{diff}}(t_1), \ldots, Z_{\text{diff}}(t_L) | X^{\text{B}}(t_1), X^{\text{B}}(t_2), \ldots, X^{\text{B}}(t_L)\big)$$

is multivariate normal under the bootstrap distribution. Since

$$\big(Z_w(t_1), \ldots, Z_w(t_L), Z_{\text{diff}}(t_1), \ldots, Z_{\text{diff}}(t_L) | X^{\text{B}}(t_1) = 0, \ldots, X^{\text{B}}(t_L) = 0\big)$$

under the bootstrap distribution, and

$$Z_w(t) = \frac{\mathsf{Var}_{\text{B}}(R_w(t))}{\mathsf{Var}_{\text{P}}(R_w(t))} \left(Z_w^{\text{B}}(t) + \frac{\mathsf{E}_{\text{B}}(R_w(t)) - \mathsf{E}_{\text{P}}(R_w(t))}{\sqrt{\mathsf{Var}_{\text{B}}(R_w(t))}} \right),$$
$$Z_{\text{diff}}(t) = \frac{\mathsf{Var}_{\text{B}}(R_{\text{diff}}(t))}{\mathsf{Var}_{\text{P}}(R_{\text{diff}}(t))} \left(Z_{\text{diff}}^{\text{B}}(t) + \frac{\mathsf{E}_{\text{B}}(R_{\text{diff}}(t)) - \mathsf{E}_{\text{P}}(R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}_{\text{B}}(R_{\text{diff}}(t))}} \right).$$

Then together with Lemma 4, we conclude that

$$(Z_w([nu_1]), \ldots, Z_w([nu_L]), Z_{\text{diff}}([nu_1]), \ldots, Z_{\text{diff}}([nu_L]))$$

is multivariate Gaussian under the permutation null distribution.

To prove (A.4), we only need to show that $\sum_{l=1}^{L}(a_l Z_w^{\text{B}}(t_l) + b_l Z_{\text{diff}}^{\text{B}}(t_l) + c_l X^{\text{B}}(t_l))$ is normal for any fixed a_l, b_l, and c_l for the non-degenerating case that $\mathsf{Var}_{\text{B}}(W) := \mathsf{Var}_{\text{B}}(a_l Z_w^{\text{B}}(t_l) + b_l Z_{\text{diff}}^{\text{B}}(t_l) + c_l X^{\text{B}}(t_l)) > 0$. We prove the Gaussianity of $\sum_{l=1}^{L}(a_l Z_w^{\text{B}}(t_l) + b_l Z_{\text{diff}}^{\text{B}}(t_l) + c_l X^{\text{B}}(t_l))$ by Stein's method.

Consider the sums of the form $W = \sum_{i \in \mathcal{J}} \xi_i$ where $\mathcal{J}$ is an index set and ξ are random variables with $\mathsf{E}(\xi) = 0$ and $\mathsf{E}(W^2) = 1$. The following assumption restricts the dependence among $\{\xi_i : i \in \mathcal{J}\}$.

Assumption 1. *(Chen and Shao (2005), page 17) For each $j \in \mathcal{J}$, there exists $S_i \subset T_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{S_i^c}$ and ξ_{S_i} is independent of $\xi_{T_i^c}$.*

We will use the following specific form of Stein's method.

Theorem 9. *(Chen and Shao (2005), Theorem 3.4) Under Assumption 1, we have*

$$\sup_{h\in Lip(1)} |\mathsf{E}(h(W)) - \mathsf{E}(h(Z))| \leq \delta$$

where $Lip(1) = \{h : \mathbb{R} \to \mathbb{R}, ||h'|| \leq 1\}$*, Z has* $N(0,1)$ *distribution and*

$$\delta = 2\sum_{i\in\mathcal{J}} (\mathsf{E}|\xi_i\eta_i\theta_i| + |\mathsf{E}(\xi_i\eta_i)|\mathsf{E}|\theta_i|) + \sum_{i\in\mathcal{J}} \mathsf{E}|\xi_i\eta_i^2|$$

with $\eta_i = \sum_{j\in S_i} \eta_j$ *and* $\theta_i = \sum_{j\in T_i} \theta_j$*, where* S_i *and* T_i *are defined in Assumption 1.*

We adopt the same notations with the index set $\mathcal{J} = \{e : e \in G\} \cup \{1, \ldots, n\}$. Let

$$\begin{aligned}\xi_{e,l} = &\frac{1}{\sigma_w^{\mathrm{B}}(t_l)} \left(\tfrac{n-t_l-1}{n-2} \mathbb{1}_{\{\tilde{\pi}(e_-)\leq t_l, \tilde{\pi}(e_-)\leq t_l\}} + \tfrac{t_l-1}{n-2} \mathbb{1}_{\{\tilde{\pi}(e_-)> t_l, \tilde{\pi}(e_-)> t_l\}} - \mu_w^{\mathrm{B}}(t_l) \right) \\ &+ \frac{1}{\sigma_{\text{diff}}^{\mathrm{B}}(t_l)} \left(\mathbb{1}_{\{\tilde{\pi}(e_-)\leq t_l, \tilde{\pi}(e_-)\leq t_l\}} - \mathbb{1}_{\{\tilde{\pi}(e_-)> t_l, \tilde{\pi}(e_-)> t_l\}} - \mu_{\text{diff}}^{\mathrm{B}}(t_l) \right)\end{aligned}$$

and

$$\xi_{i,l} = \frac{\mathbb{1}_{\{\tilde{\pi}(i)\leq t_l\}} - \frac{t_l}{n}}{\sqrt{t_l(1-t_l/n)}}.$$

Let $\xi_e = \sum_l a_l \xi_{e,l}$, and $\xi_i = \sum_l b_l \xi_{i,l}$. Then $W = \sum_{j\in\mathcal{J}} \xi_j = \sum_l (a_l Z_w^{\mathrm{B}}(t_l) + b_l Z_{\text{diff}}^{\mathrm{B}}(t_l) + c_l X^{\mathrm{B}}(t_l))$. We have $\mathsf{E}_{\mathrm{B}}(W) = 0$ and $\mathsf{E}_{\mathrm{B}}(W^2) = 1$. Let $a = \max(\max_l a_l, \max_l b_l, \max_l c_l)$, $\sigma_{\mathrm{B}} = \min(\min_l a_l \sigma_w^{\mathrm{B}}(t_l), \min_l b_l \sigma_{\text{diff}}^{\mathrm{B}}(t_l))$, and $\sigma_0 = \min_l(\sqrt{t_l(1-t_l/n)})$. Then

$$|\xi_e| \leq \frac{2aL}{\sigma_{\mathrm{B}}}, \forall e \in G; \quad |\xi_i| \leq \frac{aL}{\sigma_0}, \forall i \in \{1, \ldots, n\}.$$

For $e \in G$, let

$$\begin{aligned} S_e &= \{A_e, e_-, e_+\}, \\ T_e &= B_e \cup \{\text{Nodes in } A_e\}, \end{aligned}$$

where A_e, B_e are defined between line 160 and line 161 in the main context. Then S_e and T_e satisfy Assumption 1.

For $i = 1, \ldots, n$, let

$$\begin{aligned} S_i &= \{e \in G_i\} \cup \{i\}, \\ T_i &= \{e \in G_{i,2}\} \cup \{\text{Nodes in } G_i\}, \end{aligned}$$

where $G_{i,2}$ is the subgraph of G including all edges connect to any node in G_i. Then S_i and T_i satisfy Assumption 1.

For a directed graph we have the following relations,

$$\begin{aligned}
|S_e| &= |A_e| + 2, \\
|T_e| &\leq |B_e| + 2|A_e|, \\
|S_i| &= |G_i| + 1, \\
|T_i| &\leq |G_{i,2}| + 2|G_i|.
\end{aligned}$$

By Theorem 9, we have $\sup_{h \in Lip(1)} |\mathsf{E}(h(W)) - \mathsf{E}(h(Z))| \leq \delta$ for $Z \sim N(0,1)$, where

$$\begin{aligned}
\delta &= \frac{1}{\sqrt{\mathsf{Var}_{\mathrm{B}}(W)}} \left(2 \sum_{j \in \mathcal{J}} (\mathsf{E}_{\mathrm{B}} |\xi_j \eta_j \theta_j| + |\mathsf{E}_{\mathrm{B}}(\xi_j \eta_j)| \mathsf{E}_{\mathrm{B}} |\theta_j|) + \sum_{j \in \mathcal{J}} \mathsf{E}_{\mathrm{B}} |\xi_j \eta_j^2| \right) \\
&\leq \frac{a^3 L^3}{\sqrt{\mathsf{Var}_{\mathrm{B}}(W)}} \left(10 \sum_{e \in G} \frac{1}{\sigma_{\mathrm{B}}} \left(\frac{|A_e|}{\sigma_{\mathrm{B}}} + \frac{2}{\sigma_0} \right) \left(\frac{|B_e|}{\sigma_{\mathrm{B}}} + 2 \frac{|A_e|}{\sigma_0} \right) \right) \\
&\quad + \frac{a^3 L^3}{\sqrt{\mathsf{Var}_{\mathrm{B}}(W)}} \left(5 \sum_{i=1}^{n} \frac{1}{\sigma_0} \left(\frac{|G_i|}{\sigma_{\mathrm{B}}} + \frac{1}{\sigma_0} \right) \left(\frac{|G_{i,2}|}{\sigma_{\mathrm{B}}} + 2 \frac{|G_i|}{\sigma_0} \right) \right)
\end{aligned}$$

Since σ_{B} is at least of order $|G|^{0.5}$, and $\sigma_0 = O(n^{0.5})$, as long as the following results hold, we have $\delta \to 0$ as $n \to \infty$. The results we need are

$$\sum_{e \in G} |A_e||B_e| = o(|G|^{1.5}), \tag{A.5}$$

$$\sum_{e \in G} |A_e|^2 = o(|G| n^{0.5}), \tag{A.6}$$

$$\sum_{e \in G} |B_e| = o(|G| n^{0.5}), \tag{A.7}$$

$$\sum_{e \in G} |A_e| = o(|G|^{0.5} n), \tag{A.8}$$

$$\sum_{i=1}^{n} |G_i||G_{i,2}| = o(|G| n^{0.5}), \tag{A.9}$$

$$\sum_{e \in G} |G_i|^2 = o(|G|^{0.5} n), \tag{A.10}$$

$$\sum_{e \in G} |G_{i,2}| = o(|G|^{0.5} n), \tag{A.11}$$

$$\sum_{e \in G} |G_i| = o(n^{1.5}). \tag{A.12}$$

To show the above 9 results, recall the conditions in Theorem 1 are

1. $k = O(n^{\beta}), \beta < 0.25$,
2. $\sum_{e \in G} |A_e||B_e| = o(n^{1.5(\beta+1)})$,
3. $\sum_{e \in G} |A_e|^2 = o(n^{\beta+1.5})$.

Note that when the third condition holds, β in the first condition must be less than 0.25 because

$$\begin{aligned} 2\sum_{e \in G} |A_e|^2 &= \sum_{i=1}^{n} \sum_{j \in \mathcal{V}_{G_i}} \left(|G_i| + |G_j| - 1 - \mathbb{1}_{\{(j,i) \in G\}}\right)^2 \\ &= \sum_{i=1}^{n} \sum_{j \in \mathcal{V}_{G_i}} \left(|G_i|^2 + |G_j|^2 + 2|G_i||G_j|\right) + o(\sum_{e \in G} |A_e|^2) \\ &= 2\sum_{i=1}^{n} |G_i|^3 + 2\sum_{i=1}^{n} \sum_{j \in \mathcal{V}_{G_i}} |G_i||G_j| + o(\sum_{e \in G} |A_e|^2); \end{aligned}$$

here we use $\mathcal{V}_{G_i}$ to denote the vertex set of G_i. On the right hand side, the order of the second term is at most that of the first term, so $\sum_{e \in G} |A_e|^2$ and $\sum_{i=1}^{n} |G_i|^3$ must be of the same order. In addition, by Cauchy-Schwartz, we have

$$\sqrt{\sum_{i=1}^{n} |G_i|^3 \sum_{i=1}^{n} 1^2} \geq \sum_{i=1}^{n} |G_i|^{3/2} \geq n \left(\frac{cn^{\beta+1}}{n}\right)^{3/2} = O(n^{1.5\beta+1});$$

where $c = 2|G|/n^{\beta+1}$ is a constant, and the second inequality holds because $\sum_{i=1}^{n} |G_i|^{3/2}$ is minimized when all the nodes have the same degree. Therefore, $\sum_{i=1}^{n} |G_i|^3$ is at least $O(n^{3\beta+1})$. For condition 3 to hold, we require $\beta + 1.5 < 1.5\beta + 1$, so $\beta < 0.25$.

- Substituting $|G|$ by $O(n^{\beta+1})$, then (A.5), (A.6), and (A.12) follow immediately.
- Since $|B_e| \leq \sum_{e^* \in A_e} |A_{e^*}|$, we have $\sum_{e \in G} |B_e| \leq \sum_{e \in G} \sum_{e^* \in A_e} |A_{e^*}| = \sum_{e \in G} |A_e|^2$. So (A.6) ensures (A.7).
- By Cauchy-Schwartz, we have $\sum_{e \in G} |A_e| \leq \sqrt{\sum_{e \in G} |A_e|^2 \sum_{e \in G} 1^2} = o(n^{\beta+1.25})$. Since $(\beta + 1.25) - (0.5(\beta + 1) + 1) = 0.5\beta - 0.25 < 0$ when $\beta < 0.25$, (A.8) holds.
- Since $|G_{i,2}| \leq \sum_{j \in \mathcal{V}_{G_i}} |G_j|$, we have $\sum_{i=1}^{n} |G_{i,2}| \leq \sum_{i=1}^{n} \sum_{j \in \mathcal{V}_{G_i}} |G_j| \leq \sum_{(i,j) \in G} (|G_i| + |G_j|) \leq 2\sum_{e \in G} |A_e|$. So (A.8) ensures (A.11).
- $\sum_{e \in G} |A_e| = \sum_{(i,j) \in G} (|G_i| + |G_j| - 1 - \mathbb{1}_{\{(j,i) \in G\}}) = \sum_{i=1}^{n} |G_i|^2 - |G| - c^{(2)}$. Since $|G| = o(n^{0.5\beta+1.5})$

when $|G| = O(n^{\beta+1})$ with $\beta < 0.25$, (A.8) and (A.10) are equivalent.

- For (A.9), since $\sum_{i=1}^n |G_i||G_{i,2}| \leq \sum_{i=1}^n |G_i| \sum_{j\in\mathcal{V}_{G_i}} |G_j| = \sum_{i=1}^n \sum_{j\in\mathcal{V}_{G_i}} |G_i||G_j| \leq 4\sum_{(i,j)\in G} |G_i||G_j| \leq 4\sum_{e\in G} |A_e|^2$, thus (A.6) ensures (A.9).

Hence, (A.5) - (A.12) can be derived from conditions in Theorem 1.

The proof for the second part of Lemma 3 (showing that the covariance matrix of $(X^{\mathrm{B}}(t_1), X^{\mathrm{B}}(t_2), \ldots, X^{\mathrm{B}}(t_L))$ is positive definite) can be done in the same way as that in the Appendix of Chen and Zhang (2015) and is omitted here.

Now we prove Lemma 4: Note that since $\sum_{i=1}^n |D_i|^2 \leq 2\sum_{e\in G} |A_e|^2 = o(n^{\beta+1.5})$, the leading term in both $\mathsf{Var}_{\mathrm{P}}(R_w(t))$ and $\mathsf{Var}_{\mathrm{B}}(R_w(t))$ is only the term with $d^{(1)}$. We have

$$\lim_{n\to\infty} \frac{\mathsf{Var}_{\mathrm{P}}(R_w(t))}{\mathsf{Var}_{\mathrm{B}}(R_w(t))} = \lim_{n\to\infty} \frac{(\frac{t}{n})^2(1-\frac{t}{n})^2 d^{(1)}}{(\frac{t}{n})^2(1-\frac{t}{n})^2\Big(\big(1-(\frac{t}{n})^2\big) - (1-\frac{t}{n})^2 + 1 - 2(\frac{t}{n})(1-\frac{t}{n})\Big) d^{(1)}} = 1.$$

Since, $\mathsf{E}_{\mathrm{B}}(R_w(t)) - \mathsf{E}_{\mathrm{P}}(R_w(t)) = \frac{t(n-t)}{n^2(n-1)}|G|$, when $|G| = O(n^{\beta+1})$, $\beta < 0.25$, we have

$$\lim_{n\to\infty} \frac{\mathsf{E}_{\mathrm{B}}(R_w(t)) - \mathsf{E}_{\mathrm{P}}(R_w(t))}{\sqrt{\mathsf{Var}_{\mathrm{P}}(R_w(t))}} = \lim_{n\to\infty} \frac{|G|^{1/2}}{n} = 0.$$

Similarly, for $R_{\text{diff}}(t)$, when $\sum_{i=1}^n |D_i|^2 - nk^2 = O(\sum_{i=1}^n |D_i|^2)$,

$$\lim_{n\to\infty} \frac{\mathsf{Var}_{\mathrm{P}}(R_{\text{diff}}(t))}{\mathsf{Var}_{\mathrm{B}}(R_{\text{diff}}(t))} = \lim_{n\to\infty} \frac{\frac{t(n-t)}{n(n-1)}\left(\sum_{i=1}^n |D_i|^2 - nk^2\right)}{\frac{t(n-t)}{n^2}\left(\sum_{i=1}^n |D_i|^2 + 3nk^2\right)} = \lim_{n\to\infty} \frac{\sum_{i=1}^n |D_i|^2 - nk^2}{\sum_{i=1}^n |D_i|^2 + 3nk^2} = c_1.$$

The above limit converges to a positive constant, because $\sum_{i=1}^n |D_i|^2 - nk^2 \geq 0$, and $\sum_{i=1}^n |D_i|^2 = nk^2$ if and only if $|D_i| = k, \forall i = 1, \ldots, n$.

Since $\mathsf{E}_{\mathrm{B}}(R_{\text{diff}}(t)) - \mathsf{E}_{\mathrm{P}}(R_{\text{diff}}(t)) = 0$, we have

$$\lim_{n\to\infty} \frac{\mathsf{E}_{\mathrm{B}}(R_{\text{diff}}(t)) - \mathsf{E}_{\mathrm{P}}(R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}_{\mathrm{P}}(R_{\text{diff}}(t))}} = 0.$$

Finally, we have to show that as $n \to \infty$, $\mathsf{Cov}_{\mathrm{P}}(Z_w(s), Z_{\text{diff}}(t)) = 0$, $\forall s, t$. Without loss of generality, let $s < t$ and $\lim_{n\to\infty} s/n = u$, $\lim_{n\to\infty} t/n = v$. Since,

$$\mathsf{Cov}_{\mathrm{P}}(Z_w(s), Z_{\text{diff}}(t)) = \frac{\mathsf{E}_{\mathrm{P}}(R_w(s)R_{\text{diff}}(t)) - \mathsf{E}_{\mathrm{P}}(R_w(s))\mathsf{E}_{\mathrm{P}}(R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}_{\mathrm{P}}(R_w(s))\mathsf{Var}_{\mathrm{P}}(R_{\text{diff}}(t))}},$$

and

$$\begin{aligned}
\mathsf{E}_{\mathsf{P}}(R_w(s)R_{\text{diff}}(t)) &= \tfrac{n-s-1}{n-2}\Big(\mathsf{E}_{\mathsf{P}}(R_{G,1}(s)R_{G,1}(t)) - \mathsf{E}_{\mathsf{P}}(R_{G,1}(s)R_{G,2}(t))\Big) \\
&\quad + \tfrac{s-1}{n-2}\Big(\mathsf{E}_{\mathsf{P}}(R_{G,2}(s)R_{G,1}(t)) - \mathsf{E}_{\mathsf{P}}(R_{G,2}(s)R_{G,2}(t))\Big) \\
&= \frac{(n-2t)(s-1)(s-n+1)}{n(n-1)(n-2)}|G|^2, \\
\mathsf{E}_{\mathsf{P}}(R_w(s)) &= \left(\frac{n-s-1}{n-2}\frac{s(s-1)}{n(n-1)} + \frac{s-1}{n-2}\frac{(n-s)(n-s-1)}{n(n-1)}\right)|G|, \\
\mathsf{E}_{\mathsf{P}}(R_{\text{diff}}(s)) &= \left(\frac{s(s-1)}{n(n-1)} + \frac{(n-s)(n-s-1)}{n(n-1)}\right)|G|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\lim_{n\to\infty}\left(\mathsf{E}_{\mathsf{P}}(R_w(s)R_{\text{diff}}(t)) - \mathsf{E}_{\mathsf{P}}(R_w(s))\mathsf{E}_{\mathsf{P}}(R_{\text{diff}}(t))\right) \\
&\quad = (1-2v)u(u-1)|G|^2 - u(1-u)|G|(v^2-(1-v)^2)|G| = 0.
\end{aligned}$$

A.4 Derivations of $C_w(t)$ and $C_{\text{diff}}(t)$

As $C_w(t)$ and $C_{\text{diff}}(t)$ are the partial derivatives of the covariances of $Z_w(t)$ and $Z_{\text{diff}}(t)$, respectively. We first derive $\mathsf{Cov}(Z_w(s), Z_w(t))$ and $\mathsf{Cov}(Z_{\text{diff}}(s), Z_{\text{diff}}(t))$ analytically. We may start with deriving $C_{\text{diff}}(t)$ as it is relatively simple, and the derivation of $C_w(t)$ follows exactly the same procedure. Notice that

$$\begin{aligned}
\mathsf{Cov}(R_{\text{diff}}(s), R_{\text{diff}}(t)) &= \mathsf{Cov}(R_{G,1}(s), R_{G,1}(t)) - \mathsf{Cov}(R_{G,1}(s), R_{G,2}(t)) \\
&\quad - \mathsf{Cov}(R_{G,2}(s), R_{G,1}(t)) + \mathsf{Cov}(R_{G,2}(s), R_{G,2}(t)).
\end{aligned}$$

We can derive the above four covariances through combinatorial analysis similar to what we have done in proving Theorem 1:

$$\begin{aligned}
\mathsf{Cov}(R_{G,1}(s), R_{G,1}(t)) &= d^{(1)}\tfrac{s(s-1)}{n(n-1)} + d^{(2)}\tfrac{s(s-1)(t-2)}{n(n-1)(n-2)} + d^{(3)}\tfrac{s(s-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)} \\
&\quad - \left(n^2k^2p_1(s)p_1(t)\right),
\end{aligned}$$

$$\mathsf{Cov}(R_{G,1}(s), R_{G,2}(t)) = d^{(3)}\tfrac{s(s-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)} - \left(n^2k^2p_1(s)q_1(t)\right),$$

$$\mathsf{Cov}(R_{G,2}(s), R_{G,1}(t)) = d^{(1)}\tfrac{(t-s)(t-s-1)}{n(n-1)} + d^{(2)}\tfrac{s(t-s)(n-s-1)+(t-s)(t-s-1)(n-s-2)}{n(n-1)(n-2)}$$

$$+ d^{(3)} \tfrac{s(s-1)(n-s)(n-s-1)+2s(t-s)(n-s-1)(n-s-2)+(t-s)(t-s-1)(n-s-2)(n-s-3)}{n(n-1)(n-2)(n-3)} - \left(n^2k^2q_1(s)p_1(t)\right),$$

$$\mathsf{Cov}(R_{G,2}(s), R_{G,2}(t)) = d^{(1)} \tfrac{(n-t)(n-t-1)}{n(n-1)} + d^{(2)} \tfrac{(n-t)(n-t-1)(n-s-2)}{n(n-1)(n-2)} + d^{(3)} \tfrac{(n-t)(n-t-1)(n-s-2)(n-s-3)}{n(n-1)(n-2)(n-3)} - \left(n^2k^2q_1(s)q_1(t)\right).$$

where $d^{(1)} = \left(c^{(1)} + c^{(2)}\right)$, $d^{(2)} = \left(c^{(3)} + c^{(4)} + c^{(5)} + c^{(6)}\right)$, $d^{(3)} = c^{(7)}$. Let $|G_i|$ be the total degree of node i, then on a directed k-NN, $|G_i| = |D_i| + k$, for $i = 1, \ldots, n$. We can derive the following relations: $d^{(2)} = \sum_{i=1}^n |G_i|^2 - 2d^{(1)}$, and $d^{(3)} = n^2k^2 - \sum_{i=1}^n |G_i|^2 + d^{(1)}$. The above four covariances and be rearranged and simplified to

$$\mathsf{Cov}(R_{G,1}(s), R_{G,1}(t)) = \tfrac{s(s-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)} \left(d^{(1)} + \tfrac{t-2}{n-t-1} \left(\sum_{i=1}^n |G_i|^2 + \tfrac{t-3}{n-t} n^2k^2\right)\right) - \left(n^2k^2p_1(s)p_1(t)\right),$$

$$\mathsf{Cov}(R_{G,1}(s), R_{G,2}(t)) = \tfrac{s(s-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)} \left(n^2k^2 - \sum_{i=1}^n |G_i|^2 + d^{(1)}\right) - \left(n^2k^2p_1(s)q_1(t)\right),$$

$$\mathsf{Cov}(R_{G,2}(s), R_{G,1}(t)) = (a_1(s,t) - 2a_2(s,t) + a_3(s,t))\, d^{(1)} + (a_2(s,t) - a_3(s,t)) \sum_{i=1}^n |G_i|^2 + a_3(s,t)n^2k^2 - \left(n^2k^2q_1(s)p_1(t)\right),$$

$$\mathsf{Cov}(R_{G,2}(s), R_{G,2}(t)) = \tfrac{s(s-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)} \left(d^{(1)} + \tfrac{n-s-2}{s-1} \left(\sum_{i=1}^n |G_i|^2 + \tfrac{n-s-3}{s} n^2k^2\right)\right) - \left(n^2k^2q_1(s)q_1(t)\right),$$

where $a_1(s,t) = \frac{(t-s)(t-s-1)}{n(n-1)}$, $a_2(s,t) = \frac{s(t-s)(n-s-1)+(t-s)(t-s-1)(n-s-2)}{n(n-1)(n-2)}$, and $a_3(s,t) = \frac{s(s-1)(n-s)(n-s-1)+2s(t-s)(n-s-1)(n-s-2)+(t-s)(t-s-1)(n-s-2)(n-s-3)}{n(n-1)(n-2)(n-3)}$.

By plugging in and simplifying the expressions, we have

$$\mathsf{Cov}(R_{\text{diff}}(s), R_{\text{diff}}(t)) = \frac{s(n-t)}{n(n-1)} \left(\sum_{i=1}^n |G_i|^2 - 4nk^2\right).$$

Therefore, with the expression of $\mathsf{Var}(R_{\text{diff}}(t))$ given by (A.3), and note that on a directed k-NN, $\sum_{i=1}^{n}|G_i|^2 = \sum_{i=1}^{n}|D_i|^2 + 3nk^2$, we have

$$\begin{aligned}\rho_{\text{diff}}(s,t) &= \mathsf{Cov}(Z_{\text{diff}}(s), Z_{\text{diff}}(t)) = \frac{\mathsf{Cov}(R_{\text{diff}}(s), R_{\text{diff}}(t))}{\sqrt{\mathsf{Var}(R_{\text{diff}}(s))\mathsf{Var}(R_{\text{diff}}(t))}} \\ &= \frac{\frac{s(n-t)}{n(n-1)}\left(\sum_{i=1}^{n}|G_i|^2 - 4nk^2\right)}{\sqrt{\frac{s(n-s)}{n(n-1)}\left(\sum_{i=1}^{n}|G_i|^2 - 4nk^2\right)\frac{t(n-t)}{n(n-1)}\left(\sum_{i=1}^{n}|G_i|^2 - 4nk^2\right)}} \\ &= \frac{s(n-t)}{\sqrt{st(n-s)(n-t)}}.\end{aligned}$$

Since $C_{\text{diff}}(t) = \lim_{s\nearrow t}\frac{\partial\rho_{\text{diff}}(s,t)}{\partial s}$, by taking the partial derivative and plugging in $s = t$, we have

$$C_{\text{diff}}(t) = \frac{n}{2t(n-t)}.$$

Remark 3. *Let $\rho^*_{\mathit{diff}}(u,v) = \mathsf{Cov}(Z^*_{\mathit{diff}}(u), Z^*_{\mathit{diff}}(v))$ be the covariance function of the limiting process, where $\{Z^*_{\mathit{diff}}(u) : 0 < u < 1\}$ denotes $\{Z_{\mathit{diff}}([nu]) : 0 < u < 1\}$. The expression of $\rho^*_{\mathit{diff}}(u,v)$ does not depend on G at all. For $u \le v$, we have*

$$\rho^*_{\mathit{diff}}(u,v) = \lim_{n\to\infty}\mathsf{Cov}(Z_{\mathit{diff}}(s), Z_{\mathit{diff}}(t)) = \frac{u(1-v)}{\sqrt{u(1-u)v(1-v)}}.$$

Similarly, for $u > v$, we have

$$\rho^*_{\mathit{diff}}(u,v) = \frac{v(1-u)}{\sqrt{u(1-u)v(1-v)}}.$$

For the derivation of $C_w(t)$, notice that

$$\begin{aligned}\mathsf{Cov}(R_w(s), R_w(t)) &= \mathsf{Cov}\left(\tfrac{n-s-1}{n-2}R_{G,1}(s) + \tfrac{s-1}{n-2}R_{G,2}(s), \tfrac{n-t-1}{n-2}R_{G,1}(t) + \tfrac{t-1}{n-2}R_{G,2}(t)\right) \\ &= (\tfrac{n-s-1}{n-2})(\tfrac{n-t-1}{n-2})\mathsf{Cov}(R_{G,1}(s), R_{G,1}(t)) + (\tfrac{n-s-1}{n-2})(\tfrac{t-1}{n-2})\mathsf{Cov}(R_{G,1}(s), R_{G,2}(t)) \\ &\quad + (\tfrac{s-1}{n-2})(\tfrac{n-t-1}{n-2})\mathsf{Cov}(R_{G,2}(s), R_{G,1}(t)) + (\tfrac{s-1}{n-2})(\tfrac{t-1}{n-2})\mathsf{Cov}(R_{G,2}(s), R_{G,2}(t)).\end{aligned}$$

We can then derive $\mathsf{Cov}(Z_w(s), Z_w(t))$ in a similar way, since

$$\mathsf{Cov}(Z_w(s), Z_w(t)) = \frac{\mathsf{Cov}(R_w(s), R_w(t))}{\sqrt{\mathsf{Var}(R_w(s))\mathsf{Var}(R_w(t))}}$$

with the expression of $\mathsf{Var}(R_w(t))$ given by (A.1), and after some simplification, we have

$$\rho_w(s,t) = \mathsf{Cov}(Z_w(s), Z_w(t)) = \frac{\mathsf{Cov}(R_w(s), R_w(t))}{\sqrt{\mathsf{Var}(R_w(s))\mathsf{Var}(R_w(t))}}$$

$$
\begin{aligned}
&= \frac{\frac{d^{(1)}(n-1)(n-2)-\sum_{i=1}^{n}|G_i|^2(n-1)+2n^2k^2}{n(n-1)^2(n-2)^2}s(s-1)(n-t)(n-t-1)}{\frac{d^{(1)}(n-1)(n-2)-\sum_{i=1}^{n}|G_i|^2(n-1)+2n^2k^2}{n(n-1)^2(n-2)^2}\sqrt{st(s-1)(t-1)(n-s)(n-t)(n-s-1)(n-t-1)}} \\
&= \frac{s(s-1)(n-t)(n-t-1)}{\sqrt{st(s-1)(t-1)(n-s)(n-t)(n-s-1)(n-t-1)}}
\end{aligned}
$$

Since $C_w(t) = \lim_{s \nearrow t} \frac{\partial \rho_w(s,t)}{\partial s}$, by taking the partial derivative and plugging in $s = t$, we have

$$
C_w(t) = \frac{n(n-1)(2t^2/n - 2t + 1)}{2t(n-t)(t^2 - nt + n - 1)}.
$$

Remark 4. *Let $\rho_w^*(u,v) = \mathsf{Cov}(Z_w^*(u), Z_w^*(v))$ be the covariance function of the limiting process, where $\{Z_w^*(u) : 0 < u < 1\}$ denotes $\{Z_w([nu]) : 0 < u < 1\}$. The expression of $\rho_w^*(u,v)$ does not depend on G at all. For $u \le v$, we have*

$$
\rho_w^*(u,v) = \lim_{n\to\infty} \mathsf{Cov}(Z_w(s), Z_w(t)) = \frac{u(1-v)}{v(1-u)}.
$$

Similarly, for $u > v$, we have

$$
\rho_w^*(u,v) = \frac{v(1-u)}{u(1-v)}.
$$

A.5 Derivations of $\mathsf{E}(Z_w^3(t))$ and $\mathsf{E}(Z_{\mathbf{diff}}^3(t))$

For the 24 configurations listed in Section 2.3 in the main context, $N^{(1)}, \ldots, N^{(24)}$, we can further classify then into 8 categories by shape regardless of the directions of the edges. Let

$$
\begin{aligned}
C_1 &= \sum_{l=1}^{2} N^{(l)}, \quad C_2 = \sum_{l=3}^{8} N^{(l)}, \quad C_3 = \sum_{l=11}^{12} N^{(l)}, \quad C_4 = \sum_{l=13}^{16} N^{(l)}, \\
C_5 &= \sum_{l=17}^{20} N^{(l)}, \quad C_6 = \sum_{l=21}^{23} N^{(l)}, \quad C_7 = N^{(24)}, \quad C_8 = \sum_{l=9}^{10} N^{(i)}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\mathsf{E}\left(R_{G,1}^3(t)\right) &= C_1 p_1(t) + (C_2 + C_8) p_2(t) + (C_3 + C_4 + C_5)\, p_3(t) \\
&\quad + C_6 \left(p_3(t)\tfrac{t-4}{n-4}\right) + C_7 \left(p_3(t)\tfrac{(t-4)(t-5)}{(n-4)(n-5)}\right) \\
\mathsf{E}\left(R_{G,2}^3(t)\right) &= C_1 q_1(t) + (C_2 + C_8) q_2(t) + (C_3 + C_4 + C_5)\, q_3(t) \\
&\quad + C_6 \left(q_3(t)\tfrac{n-t-4}{n-4}\right) + C_7 \left(q_3(t)\tfrac{(n-t-4)(n-t-5)}{(n-4)(n-5)}\right) \\
\mathsf{E}\left(R_{G,1}^2(t)\right) &= \left(c^{(1)} + c^{(2)}\right) p_1(t) + \left(c^{(3)} + c^{(4)} + c^{(5)} + c^{(6)}\right) p_2(t) + c^{(7)} p_3(t)
\end{aligned}
$$

$$\mathsf{E}\left(R_{G,2}^2(t)\right) = \left(c^{(1)}+c^{(2)}\right)q_1(t)+\left(c^{(3)}+c^{(4)}+c^{(5)}+c^{(6)}\right)q_2(t)+c^{(7)}q_3(t)$$

and the expectations of all possible cross products of $R_{G,1}(t)$ and $R_{G,2}(t)$:

$$\begin{aligned}
\mathsf{E}\left(R_{G,1}^2(t)R_{G,2}(t)\right) &= \tfrac{C_3}{3}r(t)+\tfrac{C_6}{3}\left(r(t)\tfrac{t-2}{n-4}\right)+C_7\left(r(t)\tfrac{(t-2)(t-3)}{(n-4)(n-5)}\right)\\
\mathsf{E}\left(R_{G,1}(t)R_{G,2}^2(t)\right) &= \tfrac{C_3}{3}r(t)+\tfrac{C_6}{3}\left(r(t)\tfrac{n-t-2}{n-4}\right)+C_7\left(r(t)\tfrac{(n-t-2)(n-t-3)}{(n-4)(n-5)}\right)\\
\mathsf{E}\left(R_{G,1}(t)R_{G,2}(t)\right) &= c^{(7)}r(t)
\end{aligned}$$

Then $\mathsf{E}\left(Z_w^3(t)\right)$ follows as

$$\begin{aligned}
\mathsf{E}\left(Z_w^3(t)\right) &= \mathsf{E}\left(\left(\frac{R_w(t)-\mathsf{E}\left(R_w(t)\right)}{\mathsf{Var}\left(R_w(t)\right)^{1/2}}\right)^3\right)\\
&= \frac{\mathsf{E}\left(R_w^3(t)\right)-3\mathsf{E}\left(R_w^2(t)\right)\mathsf{E}\left(R_w(t)\right)+3\mathsf{E}\left(R_w(t)\right)\mathsf{E}\left(R_w(t)\right)^2-\mathsf{E}\left(R_w(t)\right)^3}{\mathsf{Var}\left(R_w(t)\right)^{3/2}}\\
&= \frac{\mathsf{E}\left(R_w^3(t)\right)-3\mathsf{E}\left(R_w^2(t)\right)\mathsf{E}\left(R_w(t)\right)+2\mathsf{E}\left(R_w(t)\right)^3}{\mathsf{Var}\left(R_w(t)\right)^{3/2}},
\end{aligned}$$

with

$$\begin{aligned}
\mathsf{E}\left(R_w^3(t)\right) &= (\tfrac{n-t-1}{n-2})^3\mathsf{E}\left(R_{G,1}^3(t)\right)+3(\tfrac{n-t-1}{n-2})^2(\tfrac{t-1}{n-2})\mathsf{E}\left(R_{G,1}^2(t)R_{G,2}(t)\right)\\
&\quad +3(\tfrac{n-t-1}{n-2})(\tfrac{t-1}{n-2})^2\mathsf{E}\left(R_{G,1}(t)R_{G,2}^2(t)\right)+(\tfrac{t-1}{n-2})^3\mathsf{E}\left(R_{G,2}^3(t)\right).
\end{aligned}$$

Similarly, $\mathsf{E}\left(Z_{\text{diff}}^3(t)\right)$ follows as

$$\begin{aligned}
\mathsf{E}\left(Z_{\text{diff}}^3(t)\right) &= \mathsf{E}\left(\left(\frac{R_{\text{diff}}(t)-\mathsf{E}\left(R_{\text{diff}}(t)\right)}{\mathsf{Var}\left(R_{\text{diff}}(t)\right)^{1/2}}\right)^3\right)\\
&= \frac{\mathsf{E}\left(R_{\text{diff}}^3(t)\right)-3\mathsf{E}\left(R_{\text{diff}}^2(t)\right)\mathsf{E}\left(R_{\text{diff}}(t)\right)+3\mathsf{E}\left(R_{\text{diff}}(t)\right)\mathsf{E}\left(R_{\text{diff}}(t)\right)^2-\mathsf{E}\left(R_{\text{diff}}(t)\right)^3}{\mathsf{Var}\left(R_{\text{diff}}(t)\right)^{3/2}}\\
&= \frac{\mathsf{E}\left(R_{\text{diff}}^3(t)\right)-3\mathsf{E}\left(R_{\text{diff}}^2(t)\right)\mathsf{E}\left(R_{\text{diff}}(t)\right)+2\mathsf{E}\left(R_{\text{diff}}(t)\right)^3}{\mathsf{Var}\left(R_{\text{diff}}(t)\right)^{3/2}},
\end{aligned}$$

with

$$\mathsf{E}\left(R_{\text{diff}}^3(t)\right)=\mathsf{E}\left(R_{G,1}^3(t)\right)-3\mathsf{E}\left(R_{G,1}^2(t)R_{G,2}(t)\right)+3\mathsf{E}\left(R_{G,1}(t)R_{G,2}^2(t)\right)-\mathsf{E}\left(R_{G,2}^3(t)\right).$$

A.6 Other edge-count statistics on a directed approximate k-NN graph

A.6.1 Generalized edge-count test statistic

The generalized edge-count test statistic is defined as

$$S(t) = \begin{pmatrix} R_{G,1}(t) - \mathsf{E}\left(R_{G,1}(t)\right) \\ R_{G,2}(t) - \mathsf{E}\left(R_{G,2}(t)\right) \end{pmatrix}^T \Sigma^{-1}(t) \begin{pmatrix} R_{G,1}(t) - \mathsf{E}\left(R_{G,1}(t)\right) \\ R_{G,2}(t) - \mathsf{E}\left(R_{G,2}(t)\right) \end{pmatrix}, \tag{A.13}$$

where $\Sigma(t)$ is the covariance matrix of $(R_{G,1}(t), R_{G,2}(t))$ under permutation. The null hypothesis of homogeneity (2.1) is rejected if the test statistic

$$\max_{n_0 \le t \le n_1} S(t) \tag{A.14}$$

with n_0 and n_1 pre-specified, is larger than the critical value for a given significance level.

It can be shown that $S(t) = Z_w^2(t) + Z_{\text{diff}}^2(t)$, and based on Theorem 3, we can approximate the tail probability

$$\mathsf{P}\Big(\max_{n_0 \le t \le n_1} S(t) > b\Big) \approx \frac{be^{-b/2}}{2\pi} \int_0^{2\pi} \int_{n_0}^{n_1} u(t,\theta)\nu\big(\sqrt{2bu(t,\theta)}\big)dtd\theta$$

where

$$u(t,\theta) = C_w(t)\sin^2(\theta) + C_{\text{diff}}(t)\cos^2(\theta).$$

Under the same setting as in Section 2.3, we check the performance of the analytic p-value approximation for the generalized edge-count test statistic. The results are presented in Table A.1.

A.6.2 Weighted edge-count test statistics

The weighted edge-count test statistic is $Z_w(t)$, and the null hypothesis of homogeneity (2.1) is rejected if the test statistic

$$\max_{n_0 \le t \le n_1} Z_w(t) \tag{A.15}$$

with n_0 and n_1 pre-specified, is larger than the critical value for a given significance level.

Table A.1: Critical values for the test statistics $\max_{n_0 \le t \le n_1} S(t)$ on the 3-NN's graph at $\alpha = 0.05$.

	$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$
Ana	13.10	13.38	13.70	14.11

Distributions and dimensions		Critical Values			
		$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$
		Per	Per	Per	Per
Multivariate Gaussian	$d = 10$	12.94	13.32	13.91	15.02
	$d = 100$	13.37	14.07	14.91	16.81
	$d = 1,000$	13.39	14.28	15.55	18.62
Multivariate t with $df = 5$	$d = 10$	13.07	13.37	13.92	15.16
	$d = 100$	13.32	14.06	15.13	17.43
	$d = 1,000$	14.59	16.10	18.13	22.93
Multivariate log-normal	$d = 10$	12.96	13.12	13.66	14.94
	$d = 100$	13.26	13.81	14.67	16.39
	$d = 1,000$	15.00	16.56	18.68	24.30

Based on Theorem 3, the tail probability (with skewness correction) can be approximated by

$$\mathsf{P}\Big(\max_{n_0 \le t \le n_1} Z_w(t) > b\Big) \approx b\phi(b) \int_{n_0}^{n_1} S_w(t) C_w(t) \nu\big(\sqrt{2b^2 C_w(t)}\big) dt.$$

Under the same setting as in Section 2.3, we check the performance of the analytic p-value approximation for the weighted edge-count test statistic. The results are presented in Table A.2.

Table A.2: Critical values for the test statistics $\max_{n_0 \le t \le n_1} Z_w(t)$ on the 3-NN's graph at $\alpha = 0.05$.

Distributions and dimensions		Critical Values							
		$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
		Ana	Per	Ana	Per	Ana	Per	Ana	Per
Multivariate Gaussian	$d = 10$	3.06	3.05	3.12	3.12	3.21	3.25	3.37	3.46
	$d = 100$	3.04	3.03	3.10	3.10	3.19	3.20	3.34	3.43
	$d = 1,000$	3.03	3.02	3.10	3.10	3.18	3.20	3.32	3.44
Multivariate t with $df = 5$	$d = 10$	3.06	3.07	3.13	3.15	3.21	3.25	3.37	3.46
	$d = 100$	3.03	3.03	3.10	3.12	3.18	3.23	3.32	3.47
	$d = 1,000$	3.02	3.05	3.08	3.14	3.16	3.26	3.31	3.49
Multivariate log-normal	$d = 10$	3.07	3.06	3.14	3.13	3.23	3.25	3.38	3.44
	$d = 100$	3.05	3.05	3.12	3.12	3.20	3.22	3.34	3.43
	$d = 1,000$	3.02	3.05	3.07	3.19	3.14	3.37	3.29	3.59

A.6.3 Original edge-count test statistics

Let $R_{G,0}(t)$ be the number of between-group edges on a directed similarity graph G. That is

$$R_{G,0}(t) = \sum_{(i,j)\in G} \left(\mathbb{1}_{\{i\le t, j>t\}} + \mathbb{1}_{\{i>t, j\le t\}} \right)$$

The original edge-count test statistic, which is the prototype of the graph-based change-point detection proposed in Chen and Zhang (2015), is defined as

$$Z_0(t) = -\frac{R_{G,0}(t) - \mathsf{E}(R_{G,0}(t))}{\sqrt{\mathsf{Var}(R_{G,0}(t))}} \tag{A.16}$$

and the null hypothesis of homogeneity (2.1) is rejected if the test statistic

$$\max_{n_0\le t\le n_1} Z_0(t) \tag{A.17}$$

with n_0 and n_1 pre-specified, is larger than the critical value for a given significance level.

Similarly, we can approximate (with skewness correction) the tail probability by

$$\mathsf{P}\Big(\max_{n_0\le t\le n_1} Z_0(t) > b \Big) \;\approx\; b\phi(b) \int_{n_0}^{n_1} S_0(t) C_0(t) \nu\big(\sqrt{2b^2 C_0(t)}\big) dt$$

where

$$\begin{aligned} S_0(t) &= \frac{\exp\left(\frac{1}{2}(b-\hat\theta_{b,0}(t))^2 + \frac{1}{6}\gamma_0(t)\hat\theta_{b,0}^3(t)\right))}{\sqrt{1+\gamma_0(t)\hat\theta_{b,0}(t)}}, \\ C_0(t) &= \lim_{s\nearrow t} \frac{\partial \rho_0(s,t)}{\partial s}, \quad \rho_0(s,t) = \mathsf{Cov}(Z_0(s), Z_0(t)); \end{aligned}$$

with $\gamma_0(t) = \mathsf{E}\left(Z_0^3(t)\right)$ and $\hat\theta_{b,0}(t) = (-1+\sqrt{1+2b\gamma_0(t)})/\gamma_0(t)$. Under the same setting as in Section 2.3, we check the performance of the analytic p-value approximation for the original edge-count test statistic. The results are presented in Table A.3.

Table A.3: Critical values for the test statistics $\max_{n_0 \leq t \leq n_1} Z_0(t)$ on the 3-NN's graph at $\alpha = 0.05$.

Distributions and dimensions		Critical Values							
		$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
		Ana	Per	Ana	Per	Ana	Per	Ana	Per
Multivariate Gaussian	$d = 10$	2.90	2.91	2.92	2.94	2.95	2.97	2.97	2.99
	$d = 100$	2.68	2.63	2.68	2.63	2.68	2.63	2.69	2.63
	$d = 1,000$	2.70	2.67	2.72	2.67	2.73	2.67	2.73	2.67
Multivariate t with $df = 5$	$d = 10$	2.90	2.93	2.92	2.96	2.95	2.98	2.96	3.00
	$d = 100$	2.65	2.61	2.65	2.61	2.66	2.61	2.66	2.61
	$d = 1,000$	2.51	2.47	2.51	2.47	2.51	2.47	2.51	2.47
Multivariate log-normal	$d = 10$	2.90	2.95	2.93	2.98	2.95	3.02	2.97	3.05
	$d = 100$	2.76	2.72	2.77	2.72	2.78	2.72	2.78	2.72
	$d = 1,000$	2.37	2.35	2.37	2.35	2.37	2.35	2.37	2.35

A.7 Additional results on empirical size

Table A.4: Empirical size: Fractions of simulation runs (out of 10,000 simulations) that the null hypothesis is rejected when there is no change-point in the sequence ($n = 1,000$).

($d = 30$)	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
New	0.103	0.052	0.011
5-MST	0.098	0.053	0.011
ecp	0.099	0.051	0.012

($d = 35$)	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
New	0.097	0.050	0.011
5-MST	0.095	0.055	0.011
ecp	0.103	0.049	0.011

A.8 The fMRI Data Profiles

Here we show the three perspectives of the images for the two subjects we use in Section 2.5.1 of Chpater 2 (ID SID-000005 and SID-000024) at $t = 150, 250, 350, 450$ and 550. The complete sequences of the images are available at `https://openneuro.org/datasets/ds003017/versions/1.0.2`.

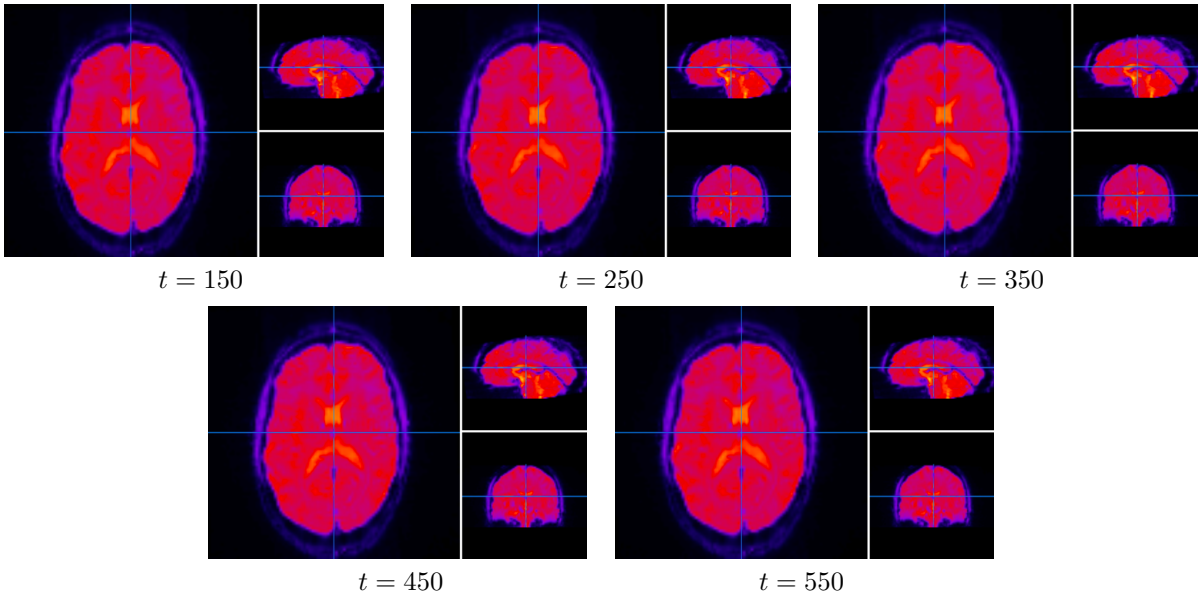


Figure A.1: The three perspectives of the fMRI images for subject SID-000005 at $t = 150, 250, 350, 450, 550$.

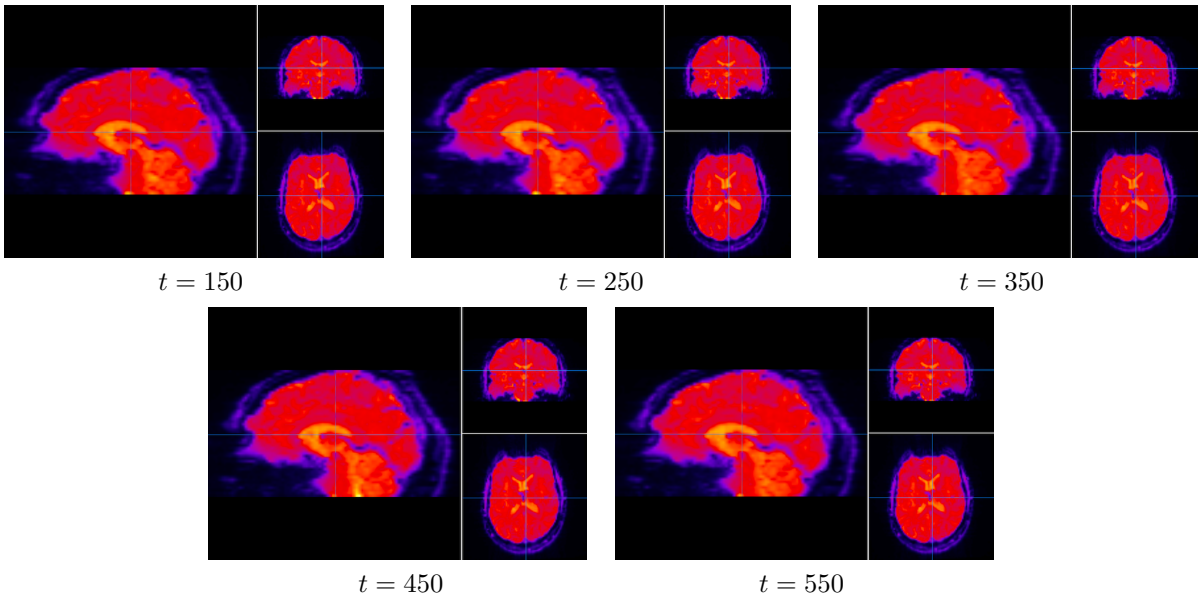


Figure A.2: The three perspectives of the fMRI images for subject SID-000024 at $t = 150, 250, 350, 450, 550$.

Appendix B

Appendix to Chapter 3

B.1 Proof of Theorem 5

Since,

$$R_{G,1}(t) = \sum_{(i,j)\in G} \mathbb{1}_{\{g_{\pi_{\texttt{CBP}}(i)}(t)=g_{\pi_{\texttt{CBP}}(j)}(t)=0\}}$$

we have its expectation,

$$\mathsf{E}_{\texttt{CBP}}(R_{G,1}(t)) = \sum_{(i,j)\in G} \mathsf{P}(\pi_{\texttt{CBP}}(i) \leq t, \pi_{\texttt{CBP}}(j) \leq t)$$

Therefore, for each edge in G formed by $\mathbf{y}_i$ and $\mathbf{y}_j$, we have to compute the probability that they are both indexed before t after circular block permutation, or equivalently, the probability of having the event $(\pi_{\texttt{CBP}}(i) \leq t, \pi_{\texttt{CBP}}(j) \leq t)$. To compute this probability, we only need to consider the particular case where $\delta_{ij} < L$ and $b > 0$. The reason is that from the proof of $\mathsf{E}_{\texttt{CBP}}(R_{G,0}(t))$ in Chen (2019a), we learn that if $\delta_{ij} \geq L$, we can plug in $\delta_{ij} = L$, and if $b = 0$, we can plug in $b = 0$ to the formula we obtain in this particular case.

For an edge with $\delta_{ij} < L$ and $b > 0$, there are six configurations that make the event $(\pi_{\texttt{CBP}}(i) \leq t, \pi_{\texttt{CBP}}(j) \leq t)$ possible: (1) $\mathbf{y}_i$ and $\mathbf{y}_j$ are in the same block with both of them in the left side of the block, (2) $\mathbf{y}_i$ and $\mathbf{y}_j$ are in the same block with $\mathbf{y}_i$ in the left side and $\mathbf{y}_j$ in the right side of the block, (3) $\mathbf{y}_i$ and $\mathbf{y}_j$ are in the same block with both of them in the right side of the block, (4) $\mathbf{y}_i$ and $\mathbf{y}_j$ are in two consecutive blocks with both $\mathbf{y}_i$ and $\mathbf{y}_j$ in the left side of their blocks, (5) $\mathbf{y}_i$ and $\mathbf{y}_j$ are in two consecutive blocks with $\mathbf{y}_i$ in the right side of its block and $\mathbf{y}_j$ in the left side of

its block, and (6) $\mathbf{y}_i$ and $\mathbf{y}_j$ are in two consecutive blocks with both $\mathbf{y}_i$ and $\mathbf{y}_j$ in the right side of their blocks. (By symmetry, we can only consider the situation where $\mathbf{y}_i$ is to the left of $\mathbf{y}_j$.)

Table B.1: Different Configurations for $\delta_{ij} < L$, $b > 0$ and i to the left of j. For each configuration, Prob.1 is the probability of having this configuration among L different ways to do the blocking, and Prob.2 is the probability of having $(\pi_{\texttt{CBP}}(i) \leq t, \pi_{\texttt{CBP}}(j) \leq t)$ after permutation given the configuration. (In this table, δ_{ij} is shortened as δ)

$B_{1,l}$	$B_{1,r}$	$B_{2,l}$	$B_{2,r}$	Prob.1	Prob.2
i,j				$\frac{(b-\delta)_+}{L}$	$\frac{a+1}{m}$
i	j			$\frac{\min(b,L-\delta)-(b-\delta)_+}{L}$	$\frac{a}{m}$
	i,j			$\frac{(L-b-\delta)_+}{L}$	$\frac{a}{m}$
i		j		$\frac{(b+\delta-L)_+}{L}$	$\frac{a(a+1)}{m(m-1)}$
	i	j		$\frac{\min(b,L-\delta)-(b-\delta)_+}{L}$	$\frac{a^2}{m(m-1)}$
	i		j	$\frac{(\delta-b)_+}{L}$	$\frac{a(a-1)}{m(m-1)}$

Summing over the products of Prob.1 and Prob.2 in Table B.1, we obtain that

$$\begin{aligned}\mathsf{P}(\pi_{\texttt{CBP}}(i) \leq t, \pi_{\texttt{CBP}}(j) \leq t) = (\min(b, L-\delta) - (b-\delta)_+)(\frac{a(m+a-1)}{n(m-1)}) + (b-\delta)_+(\frac{a+1}{n}) \\ +(\delta - b)_+(\frac{a(a-1)}{n(m-1)}) + (L-b-\delta)_+(\frac{a}{n}) + (b+\delta-L)_+(\frac{a(a+1)}{n(m-1)})\end{aligned}$$

The proof for $\mathsf{E}_{\texttt{CBP}}(R_{G,1}(t))$ is completed here.

$\mathsf{E}_{\texttt{CBP}}(R_{G,2}(t)) = \sum_{(i,j)\in G} \mathsf{P}(\pi_{\texttt{CBP}}(i) > t, \pi_{\texttt{CBP}}(j) > t)$, by the same token, for an edge formed by $\mathbf{y}_i$ and $\mathbf{y}_j$, we have to compute the probability that both $\mathbf{y}_i$ and $\mathbf{y}_j$ are indexed after t under circular block permutation. The same six scenarios apply, and Prob.1 remains the same, but Prob.2 will differ.

Table B.2: Different Configurations for $\delta_{ij} < L$, $b > 0$ and i to the left of j. For each configuration, Prob.1 is the probability of having this configuration among L different ways to do the blocking, and Prob.2 is the probability of having $(\pi_{\texttt{CBP}}(i) > t, \pi_{\texttt{CBP}}(j) > t)$ after permutation given the configuration. (In this table, δ_{ij} is shortened as δ)

$B_{1,l}$	$B_{1,r}$	$B_{2,l}$	$B_{2,r}$	Prob.1	Prob.2
i,j				$\frac{(b-\delta)_+}{L}$	$\frac{m-a-1}{m}$
i	j			$\frac{\min(b,L-\delta)-(b-\delta)_+}{L}$	$\frac{m-a-1}{m}$
	i,j			$\frac{(L-b-\delta)_+}{L}$	$\frac{m-a}{m}$
i		j		$\frac{(b+\delta-L)_+}{L}$	$\frac{(m-a-1)(m-a-2)}{m(m-1)}$
	i	j		$\frac{\min(b,L-\delta)-(b-\delta)_+}{L}$	$\frac{(m-a-1)^2}{m(m-1)}$
	i		j	$\frac{(\delta-b)_+}{L}$	$\frac{(m-a)(m-a-1)}{m(m-1)}$

Summing over the products of Prob.1 and Prob.2 in Table B.2, we obtain that

$$\begin{aligned}\mathsf{P}(\pi_{\text{CBP}}(i) > t, \pi_{\text{CBP}}(j) > t) = (\min(b, L-\delta) - (b-\delta)_+)(\frac{(m-a-1)(2m-a-2)}{n(m-1)}) + (b-\delta)_+(\frac{m-a-1}{n}) \\ +(\delta - b)_+(\frac{(m-a)(m-a-1)}{n(m-1)}) + (L-b-\delta)_+(\frac{m-a}{n}) + (b+\delta-L)_+(\frac{(m-a-1)(m-a-2)}{n(m-1)})\end{aligned}$$

The proof for $\mathsf{E}_{\text{CBP}}(R_{G,2}(t))$ is completed.

B.2 Proof of Theorem 6

For variances, since

$$\begin{aligned}\mathsf{Var}_{\text{CBP}}(R_{G,1}(t)) &= \mathsf{E}_{\text{CBP}}\left(R_{G,1}^2(t)\right) - \left(\mathsf{E}_{\text{CBP}}(R_{G,1}(t))\right)^2 \\ \mathsf{Var}_{\text{CBP}}(R_{G,2}(t)) &= \mathsf{E}_{\text{CBP}}\left(R_{G,2}^2(t)\right) - \left(\mathsf{E}_{\text{CBP}}(R_{G,2}(t))\right)^2\end{aligned}$$

we only have to work out

$$\begin{aligned}\mathsf{E}_{\text{CBP}}\left(R_{G,1}^2(t)\right) &= \sum_{(i,j),(u,v)\in G} P(g_{\pi_{\text{CBP}}(i)}(t) = g_{\pi_{\text{CBP}}(j)}(t) = g_{\pi_{\text{CBP}}(u)}(t) = g_{\pi_{\text{CBP}}(v)}(t) = 0) \\ &= \sum_{(i,j),(u,v)\in G} P(\pi_{\text{CBP}}(i), \pi_{\text{CBP}}(j), \pi_{\text{CBP}}(u), \pi_{\text{CBP}}(v) \le t) \\ \mathsf{E}_{\text{CBP}}\left(R_{G,2}^2(t)\right) &= \sum_{(i,j),(u,v)\in G} P(g_{\pi_{\text{CBP}}(i)}(t) = g_{\pi_{\text{CBP}}(j)}(t) = g_{\pi_{\text{CBP}}(u)}(t) = g_{\pi_{\text{CBP}}(v)}(t) = 1) \\ &= \sum_{(i,j),(u,v)\in G} P(\pi_{\text{CBP}}(i), \pi_{\text{CBP}}(j), \pi_{\text{CBP}}(u), \pi_{\text{CBP}}(v) > t)\end{aligned}$$

To prove $\mathsf{Var}_{\text{CBP}}(R_{G,1}(t))$, it suffices to show that

$$\mathsf{E}_{\text{CBP}}\left(R_{G,1}^2(t)\right) = d_1 p_1(a) + d_2 p_2(a) + d_3 p_3(a) + d_4 p_4(a)$$

Since

$$\mathsf{E}_{\text{CBP}}\left(R_{G,1}^2(t)\right) = \sum_{(i,j),(u,v)\in G} \mathsf{P}(\pi_{\text{CBP}}(i), \pi_{\text{CBP}}(j), \pi_{\text{CBP}}(u), \pi_{\text{CBP}}(v) \le aL)$$

Now, d_i, $i = 1, 2, 3, 4$, represent the sum of probability that the four nodes of each of the $|G|^2$ pairs of edges being blocked into i different blocks. No matter how many distinct blocks the four nodes are blocked into, the probability of having all four nodes indexed before t, for $t = aL$, is equivalent to having all blocks containing these four nodes ending up within the first a blocks after permutation.

Given a pair of edges $(i, j), (u, v) \in G$, it must belong to one the the following fours events, after doing the blocking, but before permutation:

1. If the four nodes are blocked into one single block (scenario 1), then the probability of having this block within the first a blocks after permutation is: $p_1(a) = \frac{a}{m}$
2. If the four nodes are blocked into two different blocks (scenario 2,3,4,5), then the probability of having these two blocks within the first a blocks after permutation is: $p_2(a) = \frac{a(a-1)}{m(m-1)}$
3. If the four nodes are blocked into three different blocks (scenario 6,7,8), then the probability of having these three blocks within the first a blocks after permutation is: $p_3(a) = \frac{a(a-1)(a-2)}{m(m-1)(m-2)}$
4. If all of the four nodes are three different blocks (scenario 9), then the probability of having these three blocks within the first a blocks after permutation is: $p_4(a) = \frac{a(a-1)(a-2)(a-3)}{m(m-1)(m-2)(m-3)}$

Denote the above four events E_1, E_2, E_3, E_4, respectively, and use A to denote the event $\{\pi_{\text{CBP}}(i), \pi_{\text{CBP}}(j), \pi_{\text{CBP}}(u), \pi_{\text{CBP}}(v) \leq aL\}$ just for here. Then

$$\mathsf{E}_{\text{CBP}}\left(R_{G,1}^2(t)\right) = \sum_{(i,j),(u,v)\in G} \Big(\sum_{r=1}^{4} P(A|E_r)P(E_r)\Big)$$

The proof for $\mathsf{Var}_{\text{CBP}}(R_{G,1}(t))$ is completed here.

The proof for $\mathsf{Var}_{\text{CBP}}(R_{G,2}(t))$ can be done in the same way but all the blocks involved should be ended up within the last $(m - a)$ blocks after circular block permutation. Therefore, here we skip the proof.

To prove $\mathsf{Cov}_{\text{CBP}}(R_{G,1}(t), R_{G,2}(t))$, it is sufficient to show that

$$\mathsf{E}_{\text{CBP}}(R_{G,1}(t)R_{G,2}(t)) = c_4 p_{11}(a) + c_7 p_{12}(a) + c_8 p_{21}(a) + c_9 p_{22}(a)$$

Since

$$\mathsf{E}_{\text{CBP}}(R_{G,1}(t)R_{G,2}(t)) = \sum_{(i,j),(u,v)\in G} \mathsf{P}(\pi_{\text{CBP}}(i), \pi_{\text{CBP}}(j) \leq aL; \pi_{\text{CBP}}(u), \pi_{\text{CBP}}(v) > aL)$$

Here we only consider those t's with $t = aL$. For the event $\{\pi_{\text{CBP}}(i), \pi_{\text{CBP}}(j) \leq aL; \pi_{\text{CBP}}(u), \pi_{\text{CBP}}(v) > aL\}$ to be possible, any node from the first edge (i, j) cannot be in the same block with any node from the second edge (u, v), for a given pairs of edges $(i, j), (u, v) \in G$. When the blocking is determined, scenarios 1,2,3,5,6 have zero probability of having $\{\pi_{\text{CBP}}(i), \pi_{\text{CBP}}(j) \leq aL; \pi_{\text{CBP}}(u), \pi_{\text{CBP}}(v) > aL\}$ after permutation. As a result, we only need to take scenarios 4,7,8,9 into consideration.

The probability of having $\{\pi_{\text{CBP}}(i), \pi_{\text{CBP}}(j) \leq aL; \pi_{\text{CBP}}(u), \pi_{\text{CBP}}(v) > aL\}$ of each scenarios 4,7,8,9:

1. Scenario 4: Nodes (i, j) are in one block; nodes (u, v) are in another:

   In this scenario, the block containing (i, j) must be in the first a blocks and the block containing (u, v) must be in the last $m - a$ blocks after permutation. Therefore, the probability is: $p_{11}(a) = \frac{a(m-a)}{m(m-1)}$

2. Scenario 7: Nodes (i, j) are in one block; node u and node v are in two other blocks:

   In this scenario, the block containing (i, j) must be in the first a blocks and the other two blocks containing node u and node v must both be in the last $m - a$ blocks after permutation. Therefore, the probability is: $p_{12}(a) = \frac{a(m-a)(m-a-1)}{m(m-a)(m-2)}$

3. Scenario 8: Nodes (u, v) are in one block; node i and node j are in two other blocks:

   In this scenario, the block containing (u, v) must be in the last $m - a$ blocks and the other two blocks containing node i and node j must both be in the first a blocks after permutation. Therefore, the probability is: $p_{21}(a) = \frac{a(a-1)(m-a)}{m(m-a)(m-2)}$

4. Scenario 9: All four nodes $(i, j), (u, v)$ are in different blocks:

   In this scenario, the two blocks containing node i or node j must be in the first a blocks and the other two block containing node u and node v must be in the last $m - a$ blocks after permutation. Therefore, the probability is $p_{22}(a) = \frac{a(a-1)(m-a)(m-a-1)}{m(m-1)(m-2)(m-3)}$

The proof for $\mathsf{Cov}_{\text{CBP}}(R_{G,1}(t), R_{G,2}(t))$ is completed here.

B.3 Proof of Lemma 2

In Chu and Chen (2019), $R_{w^0}(t)$ is defined as $R_{w^0}(t) = q^0(t)R_{G,1}(t) + (1 - q^0(t))R_{G,2}(t)$ where $q^0(t) = \frac{n-t-1}{n-2}$, which is chosen to minimize $\mathsf{Var}_{\mathrm{P}}(R_{w(t)}(t))$ under permutation. However, under circular block permutation, the weight function that minimizes $\mathsf{Var}_{\mathrm{CBP}}(R_{w(t)}(t))$ is no longer this $q^0(t)$ as proposed in Chu and Chen (2019). Here we derive analytically the optimal weight function that minimize $\mathsf{Var}_{\mathrm{CBP}}(R_{w(t)}(t))$ for each $t = aL$, $a \in \{1, \cdots, m-1\}$, and for the other t's, we use interpolation to fill in the blanks. While the optimal weight function under permutation, $q^0(t) = \frac{n-t-1}{n-2}$, is independent of the graph, we will show that the optimal weight function under circular block permutation indeed depends on the graph when $L \geq 2$.

Since, for $t = aL$, $a \in \{1, \cdots, m-1\}$, and any weight function $w(t)$, we have

$$R_{w(t)}(t) = w(t)R_{G,1}(t) + (1 - w(t))R_{G,2}(t)$$

Therefore, the variance of $R_{w(t)}(t)$ under circular block permutation can be expressed as

$$\begin{aligned}
\mathsf{Var}_{\mathrm{CBP}}(R_{w(t)}(t)) &= \mathsf{Var}_{\mathrm{CBP}}(w(t)R_{G,1}(t) + (1 - w(t))R_{G,2}(t)) \\
&= w(t)^2\mathsf{Var}_{\mathrm{CBP}}(R_{G,1}(t)) + (1 - w(t))^2\mathsf{Var}_{\mathrm{CBP}}(R_{G,2}(t)) + 2w(t)(1 - w(t))\mathsf{Cov}_{\mathrm{CBP}}(R_{G,1}(t), R_{G,2}(t))
\end{aligned}$$

Searching over all possible $w(t)$'s, we want to find the optimal weight that minimizes $\mathsf{Var}_{\mathrm{CBP}}(R_{w(t)}(t))$. The optimal weight function $q(t)$ must satisfy the first order condition. Therefore, we have

$$q(t) = \frac{\mathsf{Var}_{\mathrm{CBP}}(R_{G,2}(t)) - \mathsf{Cov}_{\mathrm{CBP}}(R_{G,1}(t), R_{G,2}(t))}{\mathsf{Var}_{\mathrm{CBP}}(R_{G,1}(t)) + \mathsf{Var}_{\mathrm{CBP}}(R_{G,2}(t)) - 2\mathsf{Cov}_{\mathrm{CBP}}(R_{G,1}(t), R_{G,2}(t))}$$

For $t = aL$, we can plug in the expressions for $\mathsf{Var}_{\mathrm{CBP}}(R_{G,1}(t))$, $\mathsf{Var}_{\mathrm{CBP}}(R_{G,2}(t))$, and $\mathsf{Cov}_{\mathrm{CBP}}(R_{G,1}(t), R_{G,2}(t))$ as given by Theorem 7 and 5. Then the numerator is a polynomial in a of order 3, and the denominator is a polynomial in a of order 2. In fact, the optimal weight function is a straight line, linear in a, and can be written as

$$q(t) = \frac{x_1 a + x_2 a^2 + x_3 a^3}{y_1 a + y_2 a^2} = Ca + B$$

Where C is the slope and B is the intercept of the weight function. For the above fraction to be linear in a, we must have the following relations:

$$x_1 - y_1 B = 0$$

$$\frac{x_2 - y_2 B}{y_1} \quad = \quad \frac{x_3}{y_2} = C$$

With careful calculation, one can derive that

$$\begin{aligned}
x_1 &= \frac{3}{m}\left(c_1^{(sub)}\right)^2 + \frac{7m-6}{m(m-1)} c_1^{(sub)} c_5^{(sub)} + \frac{4m-3}{m(m-1)}\left(c_5^{(sub)}\right)^2 \\
&\quad - \left(\frac{1}{m}c_1 + \frac{2m-1}{m(m-1)}(c_2+c_3+c_5) + \frac{3m-1}{m(m-1)}(c_4+c_8) + \frac{3m^2-6m+2}{m(m-1)(m-2)}c_6 + \frac{4m^2-7m+2}{m(m-1)(m-2)}(c_7+c_9)\right) \\
y_1 &= \frac{4}{m}|G|^2 - \left(\frac{2}{m-1}(c_2+c_3+c_5) + \frac{4}{m-1}(c_4+c_9) + \frac{3}{m-1}c_6 + \frac{5m-8}{(m-1)(m-2)}c_7 + \frac{3m-8}{(m-1)(m-2)}c_8\right) \\
x_2 &= \frac{-2}{m}\left(c_1^{(sub)}\right)^2 - \frac{7m-4}{m^2(m-1)} c_1^{(sub)} c_5^{(sub)} - \frac{5m-2}{m^2(m-1)}\left(c_5^{(sub)}\right)^2 \\
&\quad + \frac{1}{m(m-1)}(c_2+c_3+c_5+2c_4+2c_8) + \frac{3}{m(m-2)}c_6 + \frac{5m-4}{m(m-1)(m-2)}(c_7+c_9) \\
y_2 &= \frac{-4}{m^2}|G|^2 + \frac{1}{m(m-1)}\big(2(c_2+c_3+c_5) + 3c_6 + 4(c_4+c_9)\big) + \frac{1}{m(m-1)(m-2)}\big((7m-8)c_7 + (m-8)c_8\big) \\
x_3 &= \frac{2c_5^{(sub)}|G|}{m^2(m-1)} - \frac{1}{m(m-1)(m-2)}(c_6 + 2c_7 + 2c_9)
\end{aligned}$$

where $m = n/L$, $|G| = c_1^{(sub)} + c_5^{(sub)}$, with $c_1^{(sub)}$ and $c_5^{(sub)}$ being defined in lemma 2, and $c_1, \cdots, c_9$, as defined in Definition 1, are coefficients depending on the similarity graph, satisfying $c_1 + \cdots + c_9 = |G|^2$.

Note that regardless of the values of those c_i's, $q(t) = q(aL) = \frac{1}{2}$ at $a = \frac{m}{2}$ always holds. Moreover, for $t = aL$, $a \in \{1, \cdots, m-1\}$, $q(t) = q(aL)$ is linear in a, hence $q(t)$ is automatically defined for every t, $t \in \{1, \cdots, n-1\}$. Hence, lemma (2) can be obtained.

B.4 More results on p-value approximation under CBP

1. **Multivariate Gaussian Distributions**:

Table B.3: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} Z_{w,\mathrm{CBP}}(t)$ based on MST at $\alpha = 0.05$

	Critical Values												Graph	
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$				
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i\rvert^2$	$d_{\max}$
$d = 10$	2.99	3.05	3.06	3.03	3.12	3.12	3.08	3.22	3.27	3.15	3.40	3.53	5360	7
	2.99	3.05	3.04	3.03	3.12	3.12	3.08	3.22	3.26	3.14	3.40	3.51	5396	7
$d = 100$	2.98	3.05	3.07	3.03	3.12	3.15	3.08	3.22	3.28	3.14	3.40	3.59	13960	44
	2.98	3.05	3.03	3.03	3.12	3.12	3.08	3.22	3.26	3.14	3.39	3.54	10732	43
$d = 1,000$	2.98	3.05	3.07	3.03	3.12	3.16	3.08	3.22	3.35	3.14	3.39	3.67	20352	82
	2.98	3.05	3.02	3.03	3.12	3.13	3.08	3.22	3.27	3.14	3.39	3.55	17236	65

Table B.4: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} S_{\mathtt{CBP}}(t)$ based on MST at $\alpha = 0.05$.

	Critical Values								Graph	
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$			
	A1	**Per**	**A1**	**Per**	**A1**	**Per**	**A1**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 10$	13.11	12.86	13.39	13.29	13.71	14.07	14.12	15.32	5360	7
	13.10	13.03	13.39	13.43	13.71	13.95	14.12	15.38	5396	7
$d = 100$	13.10	13.70	13.38	14.30	13.70	15.32	14.11	17.89	13960	44
	13.09	13.16	13.38	13.82	13.70	14.93	14.11	17.49	10732	43
$d = 1,000$	13.10	13.94	13.38	14.96	13.71	16.93	14.11	21.61	20352	82
	13.10	13.55	13.38	14.61	13.70	16.12	14.11	19.76	17236	65

Table B.5: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} M_{\mathtt{CBP}}(t)$ based on MST at $\alpha = 0.05$

	Critical Values												Graph	
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$				
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 10$	3.23	3.27	3.26	3.28	3.33	3.33	3.32	3.41	3.44	3.38	3.56	3.64	5360	7
	3.23	3.27	3.25	3.28	3.33	3.34	3.32	3.41	3.42	3.38	3.56	3.63	5396	7
$d = 100$	3.23	3.30	3.33	3.27	3.38	3.43	3.32	3.49	3.58	3.38	3.67	3.88	13960	44
	3.23	3.30	3.27	3.27	3.38	3.36	3.32	3.48	3.53	3.38	3.67	3.87	10732	43
$d = 1,000$	3.23	3.35	3.36	3.28	3.43	3.50	3.32	3.56	3.74	3.38	3.78	4.22	20352	82
	3.23	3.32	3.31	3.28	3.41	3.43	3.32	3.52	3.63	3.38	3.72	4.04	17236	65

2. **Exponential Distributions**:

Table B.6: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} Z_{w,\mathtt{CBP}}(t)$ based on MST at $\alpha = 0.05$

	Critical Values												Graph	
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$				
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 10$	2.98	3.05	3.03	3.02	3.12	3.12	3.08	3.22	3.21	3.14	3.39	3.46	5212	8
	2.99	3.05	3.03	3.03	3.12	3.12	3.08	3.22	3.22	3.14	3.40	3.43	5016	6
$d = 100$	2.99	3.05	3.07	3.03	3.12	3.15	3.08	3.22	3.30	3.15	3.40	3.59	11750	32
	2.98	3.05	3.05	3.03	3.12	3.14	3.08	3.22	3.27	3.14	3.39	3.57	11572	35
$d = 1,000$	2.99	3.05	3.12	3.03	3.12	3.25	3.08	3.21	3.42	3.15	3.39	3.83	38302	120
	2.98	3.04	3.08	3.03	3.11	3.20	3.08	3.21	3.38	3.14	3.38	3.86	43376	112

Table B.7: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} S_{\mathtt{CBP}}(t)$ based on MST at $\alpha = 0.05$.

	Critical Values								Graph	
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$			
	A1	**Per**	**A1**	**Per**	**A1**	**Per**	**A1**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 10$	13.10	12.67	13.38	13.09	13.70	13.64	14.11	14.88	5212	8
	13.10	12.97	13.38	13.25	13.70	13.79	14.11	15.10	5016	6
$d = 100$	13.10	13.26	13.39	13.87	13.71	14.84	14.12	17.39	11750	32
	13.10	13.50	13.38	14.22	13.70	15.29	14.11	18.20	11572	35
$d = 1,000$	13.10	14.36	13.39	15.61	13.71	17.99	14.12	23.13	38302	120
	13.10	14.64	13.38	16.18	13.70	18.43	14.11	23.41	43376	112

Table B.8: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} M_{\mathtt{CBP}}(t)$ based on MST at $\alpha = 0.05$

	Critical Values												Graph	
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$				
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 10$	3.23	3.27	3.27	3.27	3.33	3.33	3.32	3.41	3.42	3.38	3.55	3.58	5212	8
	3.23	3.26	3.25	3.28	3.32	3.32	3.32	3.40	3.41	3.38	3.54	3.64	5016	6
$d = 100$	3.23	3.29	3.30	3.28	3.37	3.38	3.32	3.47	3.51	3.38	3.65	3.82	11750	32
	3.23	3.30	3.32	3.27	3.38	3.43	3.32	3.48	3.56	3.38	3.67	3.92	11572	35
$d = 1,000$	3.23	3.35	3.39	3.28	3.44	3.55	3.32	3.57	3.80	3.38	3.79	4.28	38302	120
	3.23	3.36	3.41	3.27	3.45	3.59	3.32	3.58	3.81	3.38	3.80	4.32	43376	112

3. **Log-normal Distributions**:

Table B.9: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} Z_{w,\mathtt{CBP}}(t)$ based on MST at $\alpha = 0.05$

	Critical Values												Graph	
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$				
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 10$	2.99	3.05	3.02	3.03	3.12	3.12	3.08	3.22	3.23	3.15	3.40	3.51	5028	6
	2.99	3.05	3.03	3.03	3.12	3.12	3.08	3.22	3.26	3.14	3.40	3.48	5116	7
$d = 100$	2.99	3.05	3.06	3.03	3.12	3.14	3.08	3.22	3.25	3.15	3.40	3.55	9436	27
	2.98	3.05	3.03	3.03	3.12	3.10	3.08	3.22	3.24	3.14	3.39	3.51	10674	42
$d = 1,000$	2.99	3.04	3.13	3.03	3.11	3.23	3.08	3.21	3.40	3.15	3.39	3.76	41500	113
	2.98	3.04	3.16	3.03	3.11	3.26	3.08	3.20	3.47	3.14	3.37	3.87	65694	164

Table B.10: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} S_{\text{CBP}}(t)$ based on MST at $\alpha = 0.05$.

	Critical Values								Graph	
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$			
	A1	**Per**	**A1**	**Per**	**A1**	**Per**	**A1**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 10$	13.10	12.92	13.39	13.32	13.71	13.97	14.12	15.30	5028	6
	13.10	12.88	13.39	13.23	13.71	13.84	14.11	15.07	5116	7
$d = 100$	13.10	13.43	13.39	13.96	13.71	14.87	14.12	16.98	9436	27
	13.10	13.54	13.38	14.25	13.70	15.58	14.11	18.41	10674	42
$d = 1,000$	13.10	14.47	13.39	15.89	13.71	17.85	14.12	22.54	41500	113
	13.10	15.06	13.38	16.56	13.70	19.35	14.11	25.12	65694	164

Table B.11: Critical values for the scan statistics $\max_{n_0 \le t \le n_1} M_{\text{CBP}}(t)$ based on MST at $\alpha = 0.05$

	Critical Values												Graph	
	$n_0 = 100$			$n_0 = 75$			$n_0 = 50$			$n_0 = 25$				
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 10$	3.23	3.26	3.26	3.28	3.32	3.34	3.32	3.40	3.43	3.38	3.55	3.65	5028	6
	3.23	3.27	3.26	3.28	3.32	3.34	3.32	3.40	3.44	3.38	3.55	3.63	5116	7
$d = 100$	3.23	3.29	3.30	3.28	3.36	3.38	3.32	3.46	3.52	3.38	3.64	3.80	9436	27
	3.23	3.32	3.31	3.28	3.40	3.42	3.32	3.52	3.58	3.38	3.72	3.92	10674	42
$d = 1,000$	3.32	3.35	3.42	3.28	3.44	3.57	3.32	3.56	3.78	3.38	3.78	4.24	41500	113
	3.32	3.37	3.41	3.27	3.46	3.57	3.32	3.59	3.88	3.38	3.82	4.34	65694	164

B.5 Analytic expressions for $C_w(t)$ and $C_{\text{diff}}(t)$ under CBP

In this section, we will derive the analytic expressions for $C_w(t)$ and $C_{\text{diff}}(t)$, in order to compute the asymptotic p-value approximations. Here we only work in detail with $C_w(t)$, since $C_{\text{diff}}(t)$ can be derived in a similar way with the weight functions $q(t), p(t)$ being replaced by $1, -1$, respectively. Throughout Section B.5, notations $R_{G,1}(\cdot)$ and $R_{G,2}(\cdot)$ are abbreviated to $R_1(\cdot)$ and $R_2(\cdot)$ for simplicity.

We need to derive $\frac{\partial}{\partial s}\mathsf{Cov}_{\text{CBP}}(Z_{w,\text{CBP}}(s), Z_{w,\text{CBP}}(t))$, which we denote as $\mathsf{Cov}'_{\text{CBP}}(Z_{w,\text{CBP}}(s), Z_{w,\text{CBP}}(t))$, $0 < s \le t < n$, for the remaining context, in order to compute the asymptotic p-value approximation for extended weighted edge-count test. As usual, we will compute $C_w(t) = \frac{\partial}{\partial s}\mathsf{Cov}_{\text{CBP}}(Z_{w,\text{CBP}}(s), Z_{w,\text{CBP}}(t))\big|_{s=t}$ only for those t's with $t = aL$, where a is an integer; for other t's, we compute the estimation of this quantity by pluggin in $t = aL$ with non-integer valued a directly. Since

$$\mathsf{Cov}_{\text{CBP}}(Z_{w,\text{CBP}}(s), Z_{w,\text{CBP}}(t)) = \frac{\mathsf{Cov}_{\text{CBP}}(R_w(s), R_w(t))}{\sqrt{\mathsf{Var}_{\text{CBP}}(R_w(s))\mathsf{Var}_{\text{CBP}}(R_w(t))}}$$

$$\mathsf{Cov}'_{\text{CBP}}(Z_{w,\text{CBP}}(s), Z_{w,\text{CBP}}(t)) = \frac{\mathsf{Cov}'_{\text{CBP}}(R_w(s), R_w(t))}{\sqrt{\mathsf{Var}_{\text{CBP}}(R_w(s))\mathsf{Var}_{\text{CBP}}(R_w(t))}} - \frac{1}{2}\frac{\mathsf{Cov}_{\text{CBP}}(R_w(s), R_w(t))}{\sqrt{\mathsf{Var}_{\text{CBP}}(R_w(s))\mathsf{Var}_{\text{CBP}}(R_w(t))}}\frac{\mathsf{Var}'_{\text{CBP}}(R_w(s))}{\mathsf{Var}_{\text{CBP}}(R_w(s))}$$

We've already derived $\mathsf{Var}_{\text{CBP}}(R_w(a))$, for $t = aL$, where a is an integer. For other t's that are not multiples of L, here we compute $\mathsf{Var}_{\text{CBP}}(R_w(t))$ by plugging in $a = t/L$ directly, instead of doing interpolation. To compute $\mathsf{Cov}'_{\text{CBP}}(Z_{w,\text{CBP}}(s), Z_{w,\text{CBP}}(t))$, we have to further derive $\mathsf{Var}'_{\text{CBP}}(R_w(t))$, $\mathsf{Cov}_{\text{CBP}}(R_w(s), R_w(t))$, and $\mathsf{Cov}'_{\text{CBP}}(R_w(s), R_w(t))$.

B.5.1 $\mathsf{Var}'_{\text{CBP}}(R_w(t)) = \frac{d}{dt}\mathsf{Var}_{\text{CBP}}(R_w(t))$

We here consider $\mathsf{Var}_{\text{CBP}}(R_w(aL))$ and $\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_w(aL))$, where t is a multiple of L, i.e., $t = aL$.
Note that $\frac{d}{dt}\mathsf{Var}_{\text{CBP}}(R_w(t)) = \frac{1}{L}\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_w(aL))$

$$\begin{aligned}
\mathsf{Var}_{\text{CBP}}(R_w(aL)) &= (q(aL))^2\mathsf{Var}_{\text{CBP}}(R_1(aL)) + (1-q(aL))^2\mathsf{Var}_{\text{CBP}}(R_2(aL)) \\
&\quad +2q(aL)(1-q(aL))\mathsf{Cov}_{\text{CBP}}(R_1(aL), R_2(aL))
\end{aligned}$$

$$\begin{aligned}
\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_w(aL)) &= 2q(aL)\frac{dq(aL)}{da}\mathsf{Var}_{\text{CBP}}(R_1(aL)) + (q(aL))^2\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_1(aL)) \\
&\quad +2(1-q(aL))(-\frac{dq(aL)}{da})\mathsf{Var}_{\text{CBP}}(R_2(aL)) + (1-q(aL))^2\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_2(aL)) \\
&\quad +2\frac{dq(aL)}{da}(1-q(aL))\mathsf{Cov}_{\text{CBP}}(R_1(aL), R_2(aL)) + 2q(aL)(-\frac{dq(aL)}{da})\mathsf{Cov}_{\text{CBP}}(R_1(aL), R_2(aL)) \\
&\quad +2q(aL)(1-q(aL))\frac{d}{da}\mathsf{Cov}_{\text{CBP}}(R_1(aL), R_2(aL)) \\
&= 2cq(aL)\mathsf{Var}_{\text{CBP}}(R_1(aL)) + (q(aL))^2\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_1(aL)) \\
&\quad +2c(q(aL)-1)\mathsf{Var}_{\text{CBP}}(R_2(aL)) + (1-q(aL))^2\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_2(aL)) \\
&\quad +2c(1-q(aL))\mathsf{Cov}_{\text{CBP}}(R_1(aL), R_2(aL)) + 2(-c)q(aL)\mathsf{Cov}_{\text{CBP}}(R_1(aL), R_2(aL)) \\
&\quad +2q(aL)(1-q(aL))\frac{d}{da}\mathsf{Cov}_{\text{CBP}}(R_1(aL), R_2(aL))
\end{aligned}$$

Denote $\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_1(aL))$, $\frac{d}{da}\mathsf{Var}_{\text{CBP}}(R_2(aL))$, and $\frac{d}{da}\mathsf{Cov}_{\text{CBP}}(R_1(aL), R_2(aL))$ as $\mathsf{Var}'_{\text{CBP}}(R_1(aL))$, $\mathsf{Var}'_{\text{CBP}}(R_2(aL))$, and $\mathsf{Cov}'_{\text{CBP}}(R_1(aL), R_2(aL))$, respectively. The three quantities can be derived by taking derivative w.r.t. a directly.

$$\begin{aligned}
\mathsf{Var}'_{\text{CBP}}(R_1(aL)) &= d_1\frac{1}{m} + d_2\frac{2a-1}{m(m-1)} + d_3\frac{3a^2-6a+2}{m(m-1)(m-2)} + d_4\frac{4a^3-18a^2+22a-6}{m(m-1)(m-2)(m-3)} \\
&\quad -2\mathsf{E}_{\text{CBP}}[R_1(aL)]\mathsf{E}'_{\text{CBP}}[R_1(aL)] \\
\mathsf{Var}'_{\text{CBP}}(R_2(aL)) &= d_1\frac{(-1)}{m} + d_2\frac{(-2m+2a+1)}{m(m-1)} + d_3\frac{-3(m-a)^2+6(m-a)-2}{m(m-1)(m-2)}
\end{aligned}$$

$$+d_4\frac{-4(m-a)^3+18(m-a)^2-22(m-a)+6}{m(m-1)(m-2)(m-3)}-2\mathsf{E}_{\mathtt{CBP}}[R_2(aL)]\mathsf{E}'_{\mathtt{CBP}}[R_2(aL)]$$

$$\begin{aligned}\mathsf{Cov}'_{\mathtt{CBP}}(R_1(aL),R_2(aL)) &= c_4\frac{m-2a}{m(m-1)}+c_7\frac{m^2-4ma+3a^2-m+2a}{m(m-1)(m-2)}+c_8\frac{2ma-3a^2-m+2a}{m(m-1)(m-2)}\\ &\quad +c_9\frac{2am^2-6ma^2+4a^3+2ma-m^2+m-2a}{m(m-1)(m-2)(m-3)}-\mathsf{E}'_{\mathtt{CBP}}[R_1(aL)]\mathsf{E}_{\mathtt{CBP}}[R_2(aL)]\\ &\quad -\mathsf{E}_{\mathtt{CBP}}[R_1(aL)]\mathsf{E}'_{\mathtt{CBP}}[R_2(aL)]\end{aligned}$$

In the above expressions, $\mathsf{E}'_{\mathtt{CBP}}[R_1(aL)]$ and $\mathsf{E}'_{\mathtt{CBP}}[R_2(aL)]$ can also be derived simply by taking the derivatives of $\mathsf{E}_{\mathtt{CBP}}[R_1(aL)]$ and $\mathsf{E}_{\mathtt{CBP}}[R_2(aL)]$ w.r.t. a, and thus we have

$$\begin{aligned}\mathsf{E}'_{\mathtt{CBP}}[R_1(a)] &= c_1^{(sub)}\frac{1}{m}+c_5^{(sub)}\frac{2a-1}{m(m-1)}\\ \mathsf{E}'_{\mathtt{CBP}}[R_2(a)] &= c_1^{(sub)}\frac{(-1)}{m}+c_5^{(sub)}\frac{(-2m+2a+1)}{m(m-1)}\end{aligned}$$

Finally, $\mathsf{Var}'(R_w(t))$ can be obtained by $\frac{d}{dt}\mathsf{Var}_{\mathtt{CBP}}(R_w(t))=\frac{1}{L}\frac{d}{da}\mathsf{Var}_{\mathtt{CBP}}(R_w(aL))$.

B.5.2 $\mathsf{Cov}_{\mathtt{CBP}}(R_w(s),R_w(t))$ and its partial derivative

In this subsection, we derive the covariance of $R_w(s)$ and $R_w(t)$ under circular block permutation, denoted as $\mathsf{Cov}_{\mathtt{CBP}}(R_w(s),R_w(t))$, and its partial derivative w.r.t. the first argument s, denoted as $\frac{\partial}{\partial s}\mathsf{Cov}_{\mathtt{CBP}}(R_w(s),R_w(t))$, for future usage. As what has just been discussed, one need to evaluate this quantitiy at $s=t$ to compute asymptotic p-value approximation for modified weighted edge-count test.

Since we only compute $\mathsf{Cov}_{\mathtt{CBP}}(R_w(s),R_w(t))$ analytically at those s's and t's that are multiples of L. Let $s=a_1L$, and $t=a_2L$, then $\frac{\partial}{\partial s}\mathsf{Cov}_{\mathtt{CBP}}(R_w(s),R_w(t))=\frac{1}{L}\frac{\partial}{\partial a_1}\mathsf{Cov}_{\mathtt{CBP}}(R_w(a_1L),R_w(a_2L))$. We first derive $\mathsf{Cov}_{\mathtt{CBP}}(R_w(a_1L),R_w(a_2L))$, and then consider its partial derivative $\frac{\partial}{\partial a_1}\mathsf{Cov}_{\mathtt{CBP}}(R_w(a_1L),R_w(a_2L))$.

Let $p(aL)=1-q(aL)$, then for $0<a_1\le a_2<m$,

$$\begin{aligned}&\mathsf{Cov}_{\mathtt{CBP}}(R_w(a_1L),R_w(a_2L))\\ =\ &\mathsf{Cov}_{\mathtt{CBP}}(q(a_1L)R_1(a_1L)+p(a_1L)R_2(a_1L),q(a_2L)R_1(a_2L)+p(a_2L)R_2(a_2L))\\ =\ &q(a_1L)q(a_2L)\mathsf{Cov}_{\mathtt{CBP}}(R_1(a_1L),R_1(a_2L))+q(a_1L)p(a_2L)\mathsf{Cov}_{\mathtt{CBP}}(R_1(a_1L),R_2(a_2L))\\ &+p(a_1L)q(a_2L)\mathsf{Cov}_{\mathtt{CBP}}(R_1(a_2L),R_2(a_1L))+p(a_1L)p(a_2L)\mathsf{Cov}_{\mathtt{CBP}}(R_2(a_1L),R_2(a_2L))\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial}{\partial a_1}\mathsf{Cov}_{\text{CBP}}(R_w(a_1L), R_w(a_2L)) \\
= \;& cq(a_2L)\mathsf{Cov}_{\text{CBP}}(R_1(a_1L), R_1(a_2L)) + q(a_1L)q(a_2L)\frac{\partial}{\partial a_1}\mathsf{Cov}_{\text{CBP}}(R_1(a_1L), R_1(a_2L)) \\
&+cp(a_2L)\mathsf{Cov}_{\text{CBP}}(R_1(a_1L), R_2(a_2L)) + q(a_1L)p(a_2L)\frac{\partial}{\partial a_1}\mathsf{Cov}_{\text{CBP}}(R_1(a_1L), R_2(a_2L)) \\
&-cq(a_2L)\mathsf{Cov}_{\text{CBP}}(R_1(a_2L), R_2(a_1L)) + p(a_1L)q(a_2L)\frac{\partial}{\partial a_1}\mathsf{Cov}_{\text{CBP}}(R_1(a_2L), R_2(a_1L)) \\
&-cp(a_2L)\frac{\partial}{\partial a_1}\mathsf{Cov}_{\text{CBP}}(R_2(a_1L), R_2(a_2L))
\end{aligned}
$$

From the expressions above, $\mathsf{Cov}_{\text{CBP}}(R_1(a_1L), R_1(a_2L))$, $\mathsf{Cov}_{\text{CBP}}(R_1(a_1L), R_2(a_2L))$, $\mathsf{Cov}_{\text{CBP}}(R_1(a_2L), R_2(a_1L))$, $\mathsf{Cov}_{\text{CBP}}(R_2(a_1L), R_2(a_2L))$, are in need to compute $\mathsf{Cov}_{\text{CBP}}(R_w(a_1L), R_w(a_2L))$; and the four quantities, along with their partial derivatives w.r.t. a_1, are in need to compute $\frac{\partial}{\partial a_1}\mathsf{Cov}_{\text{CBP}}(R_w(a_1L), R_w(a_2L))$

Lemma 5. *With $c_1, \cdots, c_9$ stated before, then for $0 < a_1 \leq a_2 < m$,*

$$
\begin{aligned}
\mathit{Cov}_{\mathit{CBP}}(R_1(a_1L), R_1(a_2L)) &= c_1\frac{a_1}{m} + (c_2+c_5)\frac{a_1(a_1-1)}{m(m-1)} + (c_3+c_4)\frac{a_1(a_2-1)}{m(m-1)} + (c_6+c_8)\frac{a_1(a_1-1)(a_2-2)}{m(m-1)(m-2)} \\
&\quad + c_7\frac{a_1(a_2-1)(a_2-2)}{m(m-1)(m-2)} + c_9\frac{a_1(a_1-1)(a_2-2)(a_2-3)}{m(m-1)(m-2)(m-3)} - \mathit{E}_{\mathit{CBP}}[R_1(a_1L)]\mathit{E}_{\mathit{CBP}}[R_1(a_2L)] \\
\mathit{Cov}_{\mathit{CBP}}(R_1(a_1L), R_2(a_2L)) &= c_4\frac{a_1(m-a_2)}{m(m-1)} + c_7\frac{a_1(m-a_2)(m-a_2-1)}{m(m-1)(m-2)} + c_8\frac{(m-a_2)a_1(a_1-1)}{m(m-1)(m-2)} \\
&\quad + c_9\frac{a_1(a_1-1)(m-a_2)(m-a_2-1)}{m(m-1)(m-2)(m-3)} - \mathit{E}_{\mathit{CBP}}[R_1(a_1L)]\mathit{E}_{\mathit{CBP}}[R_2(a_2L)] \\
\mathit{Cov}_{\mathit{CBP}}(R_1(a_2L), R_2(a_1L)) &= c_1\frac{(a_2-a_1)}{m} + c_2\frac{(a_2-a_1)(a_2-1)}{m(m-1)} + c_3\frac{(a_2-a_1)(m-a_1-1)}{m(m-1)} \\
&\quad + c_4\frac{a_1(m-a_1)+(a_2-a_1)(m-a_1-1)}{m(m-1)} + c_5\frac{(a_2-a_1)(a_2-a_1-1)}{m(m-1)} \\
&\quad + c_6\frac{a_1(a_2-a_1)(m-a_1-1)+(a_2-a_1)(a_2-a_1-1)(m-a_1-2)}{m(m-1)(m-2)} \\
&\quad + c_7\frac{a_1(m-a_1)(m-a_1-1)+(a_2-a_1)(m-a_1-1)(m-a_1-2)}{m(m-1)(m-2)} \\
&\quad + c_8\frac{(m-a_2)a_2(a_2-1)+(a_2-a_1)(a_2-1)(a_2-2)}{m(m-1)(m-2)} \\
&\quad + c_9\frac{a_1(a_1-1)(m-a_1)(m-a_1-1)}{m(m-1)(m-2)(m-3)} + c_9\frac{2a_1(a_2-a_1)(m-a_1-1)(m-a_1-2)}{m(m-1)(m-2)(m-3)} \\
&\quad + c_9\frac{(a_2-a_1)(a_2-a_1-1)(m-a_1-2)(m-a_1-3)}{m(m-1)(m-2)(m-3)} - \mathit{E}_{\mathit{CBP}}[R_2(a_1L)]\mathit{E}_{\mathit{CBP}}[R_1(a_2L)] \\
\mathit{Cov}_{\mathit{CBP}}(R_2(a_1L), R_2(a_2L)) &= c_1\frac{(m-a_2)}{m} + (c_2+c_4)\frac{(m-a_2)(m-a_1-1)}{m(m-1)} + (c_3+c_5)\frac{(m-a_2)(m-a_2-1)}{m(m-1)} \\
&\quad + (c_6+c_7)\frac{(m-a_2)(m-a_2-1)(m-a_1-2)}{m(m-1)(m-2)} + c_8\frac{(m-a_2)(m-a_1-1)(m-a_1-2)}{m(m-1)(m-2)}
\end{aligned}
$$

$$+c_9\frac{(m-a_2)(m-a_2-1)(m-a_1-2)(m-a_1-3)}{m(m-1)(m-2)(m-3)} - \mathsf{E}_{CBP}[R_2(a_1L)]\mathsf{E}_{CBP}[R_2(a_2L)]$$

For the following content we denote $\frac{\partial}{\partial a_1}\mathsf{Cov}_{CBP}(R_1(a_1L), R_1(a_2L))$, $\frac{\partial}{\partial a_1}\mathsf{Cov}_{CBP}(R_1(a_1L), R_2(a_2L))$, $\frac{\partial}{\partial a_1}\mathsf{Cov}_{CBP}(R_1(a_2L), R_2(a_1L))$, $\frac{\partial}{\partial a_1}\mathsf{Cov}_{CBP}(R_2(a_1L), R_2(a_2L))$ *as* $\mathsf{Cov}'_{CBP}(R_1(a_1L), R_1(a_2L))$, $\mathsf{Cov}'_{CBP}(R_1(a_1L), R_2(a_2L))$, $\mathsf{Cov}'_{CBP}(R_1(a_2L), R_2(a_1L))$, $\mathsf{Cov}'_{CBP}(R_2(a_1L), R_2(a_2L))$, *respectively.*

$$\begin{aligned}
\mathsf{Cov}'_{CBP}(R_1(a_1L), R_1(a_2L)) &= c_1\frac{1}{m} + (c_2+c_5)\frac{2a_1-1}{m(m-1)} + (c_3+c_4)\frac{a_2-1}{m(m-1)} + (c_6+c_8)\frac{(2a_1-1)(a_2-2)}{m(m-1)(m-2)} \\
&\quad + c_7\frac{(a_2-1)(a_2-2)}{m(m-1)(m-2)} + c_9\frac{(2a_1-1)(a_2-2)(a_2-3)}{m(m-1)(m-2)(m-3)} - \mathsf{E}'_{CBP}[R_1(a_1L)]\mathsf{E}_{CBP}[R_1(a_2L)] \\
\mathsf{Cov}'_{CBP}(R_1(a_1L), R_2(a_2L)) &= c_4\frac{m-a2}{m(m-1)} + c_7\frac{(m-a_2)(m-a_2-1)}{m(m-1)(m-2)} + c_8\frac{(m-a_2)(2a_1-1)}{m(m-1)(m-2)} \\
&\quad + c_9\frac{(2a_1-1)(m-a_2)(m-a_2-1)}{m(m-1)(m-2)(m-3)} - \mathsf{E}'_{CBP}[R_1(a_1L)]\mathsf{E}_{CBP}[R_2(a_2L)] \\
\mathsf{Cov}'_{CBP}(R_1(a_2L), R_2(a_1L)) &= c_1\frac{(-1)}{m} + c_2\frac{(-(a_2-1))}{m(m-1)} + c_3\frac{(-(m-a_1-1)-(a_2-a_1))}{m(m-1)} + c_4\frac{1-a_2}{m(m-1)} \\
&\quad + c_5\frac{2a_1-2a_2+1}{m(m-1)} + c_6\frac{(-a_2m+2a_1a_2+m-a_2^2+4a_2-4a_1-2)}{m(m-1)(m-2)} \\
&\quad + c_7\frac{(2m-4a_1+3a_2-2ma_2+2a_1a_2-2)}{m(m-1)(m-2)} + c_8\frac{(-(a_2-1)(a_2-2)}{m(m-1)(m-2)} \\
&\quad + \frac{c_9}{m(m-1)(m-2)(m-3)}\{(a_1-1)(m-a_1)(m-a_1-1) + a_1(m-a_1)(m-a_1-1) \\
&\quad - a_1(a_1-1)(m-a_1-1) - a_1(a_1-1)(m-a_1) + 2(a_2-a_1)(m-a_1-1)(m-a_1-2) \\
&\quad - 2a_1(m-a_1-1)(m-a_1-2) - 2a_1(a_2-a_1)(m-a_1-2) - 2a_1(a_2-a_1)(m-a_1-1) \\
&\quad - (a_2-a_1-1)(m-a_1-2)(m-a_1-3) - (a_2-a_1)(m-a_1-2)(m-a_1-3) \\
&\quad - (a_2-a_1)(a_2-a_1-1)(m-a_1-3) - (a_2-a_1)(a_2-a_1-1)(m-a_1-2)\} \\
&\quad - \mathsf{E}'_{CBP}[R_2(a_1L)]\mathsf{E}_{CBP}[R_1(a_2L)] \\
\mathsf{Cov}'_{CBP}(R_2(a_1L), R_2(a_2L)) &= (c_2+c_4)\frac{(-(m-a_2))}{m(m-1)} + (c_6+c_7)\frac{(-(m-a_2)(m-a_2-1))}{m(m-1)(m-2)} \\
&\quad + c_8\frac{(-2(m-a_1)+3)(m-a_2)}{m(m-1)(m-2)} + c_9\frac{(-2(m-a_1)+5)(m-a_2)(m-a_2-1)}{m(m-1)(m-2)(m-3)} \\
&\quad - \mathsf{E}'_{CBP}[R_2(a_1L)]\mathsf{E}_{CBP}[R_2(a_2L)]
\end{aligned}$$

Finally, $\mathsf{Cov}'_{\text{CBP}}(Z_{w,\text{CBP}}(s), Z_{w,\text{CBP}}(t))$ can be obtained by plugging in everything, and it follows that

$$C_w(t) = \mathsf{Cov}'_{\text{CBP}}(Z_{w,\text{CBP}}(s), Z_{w,\text{CBP}}(t))\big|_{s=t}$$

Similarly, replacing $q(aL)$ with 1 and $p(aL)$ with -1, we get the expression for $C_{\text{diff}}(t)$:

$$C_{\text{diff}}(t) = \mathsf{Cov}'_{\text{CBP}}(Z_{\text{diff,CBP}}(s), Z_{\text{diff,CBP}}(t))\big|_{s=t}$$

B.6 Approximations for $\mathsf{E}_{\text{CBP}}[Z^3_{w,\text{CBP}}(t)]$ and $\mathsf{E}_{\text{CBP}}[Z^3_{\text{diff,CBP}}(t)]$

We have to compute $\mathsf{E}_{\text{CBP}}[Z^3_{w,\text{CBP}}(t)]$, and $\mathsf{E}_{\text{CBP}}[Z^3_{\text{diff,CBP}}(t)]$ to perform skewness correction. However, analytical expressions for $\mathsf{E}_{\text{CBP}}[Z^3_{w,\text{CBP}}(t)]$, $\mathsf{E}_{\text{CBP}}[Z^3_{\text{diff,CBP}}(t)]$ could be hard to derive, hence we use $\mathsf{E}_{\mathbb{P}}[Z^3_w(t)]$ and $\mathsf{E}_{\mathbb{P}}[Z^3_{\text{diff}}(t)]$, in other words, circular block permutation with $L = 1$, as surrogates for the quantities $\mathsf{E}_{\text{CBP}}[Z^3_{w,\text{CBP}}(t)]$ and $\mathsf{E}_{\text{CBP}}[Z^3_{\text{diff,CBP}}(t)]$ of our interest.

Here we only show in detail the derivation of $\mathsf{E}_{\mathbb{P}}[Z^3_w(t)]$, since $\mathsf{E}_{\mathbb{P}}[Z^3_{\text{diff}}(t)]$ can be derived similarly with the weight functions being replaced by 1 and -1.

$$\begin{aligned}
\mathsf{E}_{\mathbb{P}}[Z^3_w(t)] &= \mathsf{E}_{\mathbb{P}}[\frac{(R_w(t) - \mathsf{E}_{\mathbb{P}}[R_w(t)])^3}{(\mathsf{Var}_{\mathbb{P}}(R_w(t))^{3/2}}] \\
&= \frac{\mathsf{E}_{\mathbb{P}}[R^3_w(t)] - 3\mathsf{E}_{\mathbb{P}}[R^2_w(t)]\mathsf{E}_{\mathbb{P}}[R_w(t)] + 3\mathsf{E}_{\mathbb{P}}[R_w(t)]\mathsf{E}_{\mathbb{P}}[R_w(t)]^2 - \mathsf{E}_{\mathbb{P}}[R_w(t)]^3}{(\mathsf{Var}_{\mathbb{P}}(R_w(t))^{3/2}} \\
&= \frac{\mathsf{E}_{\mathbb{P}}[R^3_w(t)] - 3\mathsf{E}_{\mathbb{P}}[R^2_w(t)]\mathsf{E}_{\mathbb{P}}[R_w(t)] + 2\mathsf{E}_{\mathbb{P}}[R_w(t)]^3}{\mathsf{E}_{\mathbb{P}}[R^2_w(t)] - \mathsf{E}_{\mathbb{P}}[R_w(t)]^2}
\end{aligned}$$

To compute $\mathsf{E}_{\mathbb{P}}[Z^3_w(t)]$, we need $\mathsf{E}_{\mathbb{P}}[R^3_w(t)]$, $\mathsf{E}_{\mathbb{P}}[R^2_w(t)]$, and $\mathsf{E}_{\mathbb{P}}[R_w(t)]$.

When $L = 1$, circular block permutation is equivalent to random permutation, so the weight function $q(t)$ degenerates to $q^0(t) = \frac{n-t-1}{n-2}$, the weight function proposed in Chu and Chen (2019). Therefore,

$$\begin{aligned}
R_w(t) &= q(t)R_{G,1}(t) + p(t)R_{G,2}(t) \\
R^2_w(t) &= (q(t))^2R^2_{G,1}(t) + 2(q(t))(p(t))R_{G,1}(t)R_{G,2}(t) + (p(t))^2R^2_{G,2}(t) \\
R^3_w(t) &= (q(t))^3R^3_{G,1}(t) + 3(q(t))^2(p(t))R^2_{G,1}(t)R_{G,2}(t) + 3(q(t))(p(t))^2R_{G,1}(t)R^2_{G,2}(t) + (p(t))^3R^3_{G,2}(t)
\end{aligned}$$

To obtain $\mathsf{E}_{\mathbb{P}}[Z^3_w(t)]$, it suffices to compute the expectations of $R_{G,1}(t)$, $R_{G,2}(t)$, $R^2_{G,1}(t)$, $R^2_{G,2}(t)$, $R_{G,1}(t)R_{G,2}(t)$, $R^3_{G,1}(t)$, $R^3_{G,2}(t)$, $R^2_{G,1}(t)R_{G,2}(t)$, and $R_{G,1}(t)R^2_{G,2}(t)$. The expectations of the first five variables can be obtained directly from the previous formulas with $L = 1$. Here we further discuss how to compute $\mathsf{E}_{\mathbb{P}}[R^3_{G,1}(t)]$, $\mathsf{E}_{\mathbb{P}}[R^3_{G,2}(t)]$, $\mathsf{E}_{\mathbb{P}}[R^2_{G,1}(t)R_{G,2}(t)]$, and $\mathsf{E}_{\mathbb{P}}[R_{G,1}(t)R^2_{G,2}(t)]$.

When considering the cube of the number of edge-counts, we need to consider each set of edges that consisting three edges. Those three edges in a set may nor may not share nodes, so there are 8 configurations that those edges can have. The 8 configurations are:

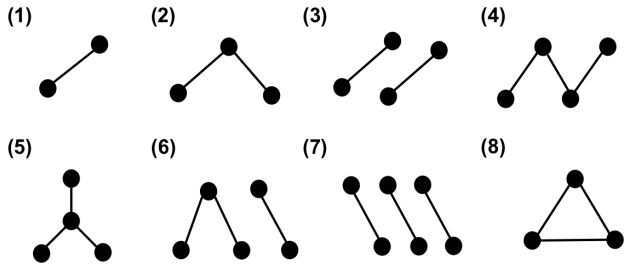


Figure B.1: The 8 possible configurations for a set of three edges

(1) The three edges consist only two distinct nodes

(2) The three edges consist only three distinct nodes

(3) Two of the edges consist two distinct nodes; the other edge is separated

(4) Chain-shaped: The first two edges share one node; the second and third edges share another node

(5) Star-shaped: The three edges share one node

(6) Two edges share one node, and share no node with the third edge

(7) No pair of the three edges share any node (six distinct nodes)

(8) Triangle: The three edges form a triangle

Denote the number of sets of three edges having configuration 1 to 8 as $C_1, \cdots, C_8$, respectively. For any undirected graph G, $C_1, \cdots, C_8$ can be computed analytically. (Observe that $C_1 + \cdots + C_8 = |G|^3$)

$$
\begin{aligned}
C_1 &= |G| \\
C_2 &= 3\sum_i |G_i|(|G_i - 1) \\
C_3 &= 3|G|(|G| - 1) - 3\sum_i |G_i|(|G_i - 1) \\
C_4 &= 6\sum_{(i,j)\in G} (|G_i| - 1)(|G_j| - 1) - 6\sum_{(i,j)\in G} |\{k : (i,k),(j,k) \in G\}| \\
C_5 &= \sum_i |G_i|(|G_i - 1)(|G_i - 2) \\
C_6 &= 3\sum_i |G_i|(|G_i - 1)(|G| - |G_i|) + 6\sum_{(i,j)\in G} |\{k : (i,k),(j,k) \in G\}| - 12\sum_{(i,j)\in G} (|G_i| - 1)(|G_j| - 1)
\end{aligned}
$$

$$
\begin{aligned}
C_7 &= |G|(|G|-1)(|G|-2) + 6\sum_{(i,j)\in G}(|G_i|-1)(|G_j|-1) - 2\sum_{(i,j)\in G}|\{k:(i,k),(j,k)\in G\}| \\
&\quad -\sum_i |G_i|(|G_i-1)(3|G|-2|G_i|-2) \\
C_8 &= 2\sum_{(i,j)\in G}|\{k:(i,k),(j,k)\in G\}|
\end{aligned}
$$

The expectations can be computed by applying proper probabilities to $C_1, \cdots, C_8$.

$$
\begin{aligned}
\mathsf{E}_{\mathbb{P}}[R^3_{G,1}(t)] &= C_1\frac{t(t-1)}{n(n-1)} + (C_2+C_8)\frac{t(t-1)(t-2)}{n(n-1)(n-2)} + (C_3+C_4+C_5)\frac{t(t-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)} \\
&\quad + C_6\frac{t(t-1)(t-2)(t-3)(t-4)}{n(n-1)(n-2)(n-3)(n-4)} + C_7\frac{t(t-1)(t-2)(t-3)(t-4)(t-5)}{n(n-1)(n-2)(n-3)(n-4)(n-5)} \\
\mathsf{E}_{\mathbb{P}}[R^3_{G,2}(t)] &= C_1\frac{(n-t)(n-t-1)}{n(n-1)} + (C_2+C_8)\frac{(n-t)(n-t-1)(n-t-2)}{n(n-1)(n-2)} \\
&\quad + (C_3+C_4+C_5)\frac{(n-t)(n-t-1)(n-t-2)(n-t-3)}{n(n-1)(n-2)(n-3)} \\
&\quad + C_6\frac{(n-t)(n-t-1)(n-t-2)(n-t-3)(n-t-4)}{n(n-1)(n-2)(n-3)(n-4)} \\
&\quad + C_7\frac{(n-t)(n-t-1)(n-t-2)(n-t-3)(n-t-4)(n-t-5)}{n(n-1)(n-2)(n-3)(n-4)(n-5)} \\
\mathsf{E}_{\mathbb{P}}[R^2_{G,1}(t)R_{G,2}(t)] &= C_3\frac{t(t-1)(n-t)(n-t-1)}{3n(n-1)(n-2)(n-3)} + C_6\frac{t(t-1)(t-2)(n-t)(n-t-1)}{3n(n-1)(n-2)(n-3)(n-4)} \\
&\quad + C_7\frac{t(t-1)(t-2)(t-3)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)(n-4)(n-5)} \\
\mathsf{E}_{\mathbb{P}}[R_{G,1}(t)R^2_{G,2}(t)] &= C_3\frac{t(t-1)(n-t)(n-t-1)}{3n(n-1)(n-2)(n-3)} + C_6\frac{t(t-1)(n-t)(n-t-1)(n-t-2)}{3n(n-1)(n-2)(n-3)(n-4)} \\
&\quad + C_7\frac{t(t-1)(n-t)(n-t-1)(n-t-2)(n-t-3)}{n(n-1)(n-2)(n-3)(n-4)(n-5)}
\end{aligned}
$$

Finally, $\mathsf{E}_{\mathbb{P}}[Z^3_w(t)]$ can be computed by plugging in everything.

Appendix C

Appendix to Chapter 4

C.1 Proof of Equation (4.13) and Equation (4.14)

We want to prove:

$$\begin{aligned}
\mathsf{P}(\max_{n_0\le t\le n_1} S_w(t) > b) &\approx 2b\left(1-\frac{N-1}{b}\right) f_N^{\chi^2}(b)\int_{n_0}^{n_1} h_w(x)\nu\left(\sqrt{2bh_w(x)}\left(1-\frac{N-1}{b}\right)\right)dx \\
\mathsf{P}(\max_{n_0\le t\le n_1} S_d(t) > b) &\approx 2b\left(1-\frac{N-1}{b}\right) f_N^{\chi^2}(b)\int_{n_0}^{n_1} h_d(x)\nu\left(\sqrt{2bh_d(x)}\left(1-\frac{N-1}{b}\right)\right)dx
\end{aligned}$$

Here we use $h_w(x)$ to denote $C_w(x)$ in Equation (4.13), and $h_d(x)$ to denote $C_{\text{diff}}(x)$ in Equation (4.14). Let

$$S_w^*(t/n) = \sum_{j=1}^{N}\left(Z_w^{*(j)}(t/n)\right)^2 \quad ; \quad S_d^*(t/n) = \sum_{j=1}^{N}\left(Z_d^{*(j)}(t/n)\right)^2.$$

We only show the proof in detail for the result of $S_w(t)$, as the result of $S_d(t)$ can be done in exactly the same way. First, since $S_w(t)$ is always positive, we may look at its square root, instead:

$$\begin{aligned}
&\mathsf{P}(\max_{n_0\le t_1\le n_1} S_w^*(t_1/n) > b^2) = \mathsf{P}(\max_{n_0\le t_1\le n_1}\sqrt{S_w^*(t_1/n)} > b) \\
&= \sum_{n_0\le t_1\le n_1}\int_0^\infty \mathsf{P}(\sqrt{S_w^*(t_1/n)}\in b+dx)\mathsf{P}(\max_{t_1<t_2\le n_1}\sqrt{S_w^*(t_2/n)} < b|\sqrt{S_w^*(t_1/n)} = b+x) \\
&\approx f_N^\chi(b)\sum_{n_0\le t_1\le n_1}\int_0^\infty (1+\frac{x}{b})^{N-1}e^{-bx}\mathsf{P}(\max_{t_1<t_2\le n_1}\sqrt{S_w^*(t_2/n)} < b|\sqrt{S_w^*(t_1/n)} = b+x)dx \\
&= \frac{1}{b}f_N^\chi(b)\sum_{n_0\le t_1\le n_1}\int_0^\infty (1+\frac{x}{b^2})^{N-1}e^{-x}\mathsf{P}(\max_{t_1<t_2\le n_1}\sqrt{S_w^*(t_2/n)} < b|\sqrt{S_w^*(t_1/n)} = b+\frac{x}{b})dx
\end{aligned}$$

$$\approx \frac{1}{b} f_N^{\chi}(b) \sum_{n_0 \le t_1 \le n_1} \int_0^{\infty} e^{-x\left(1-(N-1)/b^2\right)} \mathsf{P}\Big(\max_{t_1 < t_2 \le n_1} \sqrt{S_w^*(t_2/n)} < b | \sqrt{S_w^*(t_1/n)} = b + \frac{x}{b} \Big) dx$$

where $f_N^{\chi}(\cdot)$ is the density function of a χ distribution with degree of freedom N.

$$\begin{aligned} \mathsf{P}(\sqrt{S_w^*(t_1/n)} = b + x) &= f_N^{\chi}(b+x) = \frac{(b+x)^{N-1} e^{-\frac{(b+x)^2}{2}}}{\Gamma(N/2) 2^{N/2-1}} \\ &= \frac{b^{N-1} e^{-\frac{b^2}{2}}}{\Gamma(N/2) 2^{N/2-1}} (1 + \frac{x}{b})^{N-1} e^{-bx - \frac{x^2}{2}} \approx f_N^{\chi}(b) (1 + \frac{x}{b})^{N-1} e^{-bx}. \end{aligned}$$

Now we focus on the probability inside the integral:

$$\begin{aligned} &\mathsf{P}\Big(\max_{t_1 < t_2 \le n_1} \sqrt{S_w^*(t_2/n)} < b | \sqrt{S_w^*(t_1/n)} = b + \frac{x}{b} \Big) dx \\ &\quad \approx \mathsf{P}\Big(\max_{t_1 < t_2 \le n_1} b \left(\sqrt{S_w^*(t_2/n)} - \sqrt{S_w^*(t_1/n)} \right) < -x | \sqrt{S_w^*(t_1/n)} = b \Big) dx \end{aligned}$$

Conditioning on $\sqrt{S_w^*(t_1/n)} = b$ is equivalent to conditioning on

$$(Z_w^{*(1)}(t_1/n), Z_w^{*(2)}(t_1/n), \ldots, Z_w^{*(N)}(t_1/n)) = (b, 0, \ldots, 0).$$

Let $r = \frac{t_2}{n} - \frac{t_1}{n}$ and $\Delta_{jr} = Z_w^{*(j)}(t_2/n) - Z_w^{*(j)}(t_1/n)$. By Taylor expansion, we have

$$\begin{aligned} \sqrt{S_w^*(t_2/n)} &= \sqrt{\sum_{j=1}^{N} \left(Z_w^{*(j)}(t_1/n) + \Delta_{jr} \right)^2} = \sqrt{S_w^*(t_1/n) + 2b\Delta_{1r} + \sum_{j=1}^{N} \Delta_{jr}^2} \\ &= \sqrt{S_w^*(t_1/n)} + \frac{1}{2\sqrt{S_w^*(t_1/n)}} \left(2b\Delta_{1r} + \sum_{j=1}^{N} \Delta_{jr}^2 \right) - \frac{1}{2} \frac{1}{4b^3} \left(4b^2 \Delta_{1r}^2 + 4b\Delta_{1r} \sum_{j=1}^{N} \Delta_{jr}^2 \right) + O(r^2) \\ &= b + \Delta_{1r} + \frac{1}{2b} \left(\sum_{j=2}^{N} \Delta_{jr}^2 \right) + O(r^2) \end{aligned}$$

Therefore, we may use the following approximation:

$$b \left(\sqrt{S_w^*(t_2/n)} - \sqrt{S_w^*(t_1/n)} \right) = b \left(\Delta_{1r} + \frac{1}{2b} \left(\sum_{j=2}^{N} \Delta_{jr}^2 \right) \right).$$

Since each of the $Z_w^{*(j)}(\cdot)$ is a Gaussian process with covariance function $\rho_w^*(u, v)$, we have

$$\Delta_{1r} \sim N(b\rho - b, 1 - \rho^2), \quad \text{and} \quad \Delta_{jr} \sim N(0, 1 - \rho^2), \quad \text{for} \quad j \neq 1.$$

Let $h_w^*(u) = \lim_{v \searrow u} \dfrac{\partial \rho^*(u,v)}{\partial v}$, then by Taylor exapansion, we have the following approximation:

$$\rho_w^*(u,v) = 1 - h_w^*(u)r + O(r^2), \quad \text{and} \quad (\rho_w^*(u,v))^2 = 1 - 2h_w^*(u)r + O(r^2).$$

Therefore, $b\Delta_{1r}$ is approximately normally distributed with mean $b^2(1 - h_w^*(u)r) - b^2 = -b^2 h_w^*(u)r$ and variance $2b^2 h_w^*(u)r$; and for $j \neq 1$, $\frac{b}{2b}\Delta_{jr}^2$ has mean $h_w^*(u)r$ and variance $O(r^2)$. The variability from the $(N-1)$ Δ_{jr}^2 can be ignored. Finally, let $W_m^{(t_1)}$ be a random walk with $W_1^{(t_1)} \sim N(\mu^{(t_1)}, (\sigma^{(t_1)})^2)$, where

$$\mu^{(t_1)} = \frac{1}{n}\left(b^2 - (N-1)\right) h_w^*(u), \quad \text{and} \quad (\sigma^{(t_1)})^2 = \frac{1}{n} 2b^2 h_w^*(u).$$

$$\begin{aligned}
&\mathsf{P}(\max_{t_1 < t_2 \leq n_1} b\left(\sqrt{S_w^*(t_2/n)} - \sqrt{S_w^*(t_1/n)}\right) < -x | \sqrt{S_w^*(t_1/n)} = b)dx \\
&\quad \approx \mathsf{P}(\max_r \left(b\Delta_{1r} + \frac{b}{2b}\left(\sum_{j=2}^N \Delta_{jr}^2\right)\right) < -x)dx \approx \mathsf{P}(\min_{m \geq 1} W_m^{(t_1)} > x)dx
\end{aligned}$$

Using the fact

$$\int_0^\infty \exp(-2\mu x/\sigma^2)\mathsf{P}(\min_{m \geq 1} W_m > x)dx = \mu\nu(2\mu/\sigma),$$

for a random walk $W_1 \sim N(\mu, \sigma^2)$ with $\mu > 0$ (Siegmund, 1992), and here we have

$$\frac{2\mu}{\sigma^2} = 1 - \frac{N-1}{b^2}, \quad \text{and} \quad \frac{2\mu}{\sigma} = \sqrt{2b^2 h_w^*(u)/n}(1 - \frac{N-1}{b^2}).$$

Therefore,

$$\begin{aligned}
&\mathsf{P}(\max_{n_0 \leq t \leq n_1} \sqrt{S_w^*(t/n)} > b) \\
&\quad \approx \frac{1}{b}\left(b^2 - (N-1)\right) f_N^\chi(b) \int_{x_0}^{x_1} \frac{1}{n} h_w^*(x) \nu\left(\sqrt{2b^2 h_w^*(x)/n}\left(1 - \frac{N-1}{b^2}\right)\right) dx \\
&\quad = \frac{1}{b}\left(b^2 - (N-1)\right) f_N^\chi(b) \int_{n_0}^{n_1} h_w(x) \nu\left(\sqrt{2b^2 h_w(x)}\left(1 - \frac{N-1}{b^2}\right)\right) dx \\
&\quad = \frac{1}{b}\left(b^2 - (N-1)\right) 2b f_N^{\chi^2}(b^2) \int_{n_0}^{n_1} h_w(x) \nu\left(\sqrt{2b^2 h_w(x)}\left(1 - \frac{N-1}{b^2}\right)\right) dx \\
&\quad = 2b^2\left(1 - \frac{(N-1)}{b^2}\right) f_N^{\chi^2}(b^2) \int_{n_0}^{n_1} h_w(x) \nu\left(\sqrt{2b^2 h_w(x)}\left(1 - \frac{N-1}{b^2}\right)\right) dx
\end{aligned}$$

Replace b^2 with b we have:

$$\mathsf{P}(\max_{n_0 \leq t \leq n_1} S_w(t) > b) \quad \approx \quad 2b\left(1 - \frac{N-1}{b}\right) f_N^{\chi^2}(b) \int_{n_0}^{n_1} h_w(x) \nu\left(\sqrt{2b h_w(x)}\left(1 - \frac{N-1}{b}\right)\right) dx.$$

C.2 More results on p-value approximation for MS-statistic

Table C.1: Critical values for test statistic $\max_{n_0 \le t \le n_1} MS(t)$ based on 5-MST at $\alpha = 0.05$.

Multivariate Gaussian	$N = 1$	$N = 2$	$N = 3$	$N = 5$	$N = 10$	$N = 20$	$N = 50$
$d = 10$	11.54	14.87	17.45	21.83	31.42	47.68	89.36
$d = 100$	11.55	14.70	17.16	21.88	31.52	47.59	89.68
$d = 1,000$	11.51	14.81	17.35	22.08	31.69	47.58	89.75
Analytical	11.31	14.53	17.13	21.59	31.03	47.25	89.54

Table C.2: Critical values for test statistic $\max_{n_0 \le t \le n_1} MS(t)$ based on 5-MST at $\alpha = 0.05$.

Multivariate t_5	$N = 1$	$N = 2$	$N = 3$	$N = 5$	$N = 10$	$N = 20$	$N = 50$
$d = 10$	11.52	14.63	17.18	21.97	31.28	47.49	89.46
$d = 100$	11.38	14.81	17.59	22.07	31.64	48.00	90.07
$d = 1,000$	12.30	16.04	18.96	23.66	33.69	49.95	92.92
Analytical	11.31	14.53	17.13	21.59	31.03	47.25	89.54

Table C.3: Critical values for test statistic $\max_{n_0 \le t \le n_1} MS(t)$ based on 5-MST at $\alpha = 0.05$.

Log-normal	$N = 1$	$N = 2$	$N = 3$	$N = 5$	$N = 10$	$N = 20$	$N = 50$
$d = 10$	11.62	14.69	17.44	21.80	31.50	47.52	89.64
$d = 100$	11.34	14.62	17.39	22.02	31.70	47.77	90.01
$d = 1,000$	11.98	15.45	18.36	22.99	32.63	49.02	92.19
Analytical	11.31	14.53	17.13	21.59	31.03	47.25	89.54

Bibliography

Akgiray, V. (1989). Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts. *Journal of business*, pages 55–80.

Arlot, S., Celisse, A., and Harchaoui, Z. (2019). A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20(162):1–56.

Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R., and Wu, A. Y. (1998). An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923.

Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087.

Barigozzi, M., Cho, H., and Fryzlewicz, P. (2018). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics*, 206(1):187–225.

Basseville, M., Nikiforov, I. V., et al. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice hall Englewood Cliffs.

Bauwens, L., Laurent, S., and Rombouts, J. V. (2006). Multivariate garch models: a survey. *Journal of applied econometrics*, 21(1):79–109.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2019). *Fast Nearest Neighbor Search Algorithms and Applications*. R package FNN: kd-tree fast k-nearest neighbor search algorithms.

Bhattacharjee, M., Banerjee, M., and Michailidis, G. (2018). Change point estimation in a dynamic stochastic block model. *arXiv preprint arXiv:1812.03090*.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, pages 542–547.

Bollerslev, T. (1988). On the correlation structure for the generalized autoregressive conditional heteroskedastic process. *Journal of Time Series Analysis*, 9(2):121–131.

Brodsky, B. and Darkhovsky, B. (1993). Applications of nonparametric change-point detection methods. In *Nonparametric Methods in Change-Point Problems*, pages 169–182. Springer.

Carlstein, E. G., Müller, H.-G., and Siegmund, D. (1994). Change-point problems. IMS.

Chen, H. (2019a). Change-point detection for multivariate and non-euclidean data with local dependency. *arXiv preprint arXiv:1903.01598*.

Chen, H. (2019b). Sequential change-point detection based on nearest neighbors. *The Annals of Statistics*, 47(3):1381–1407.

Chen, H., Chen, S., and Deng, X. (2019). A universal nonparametric event detection framework for neuropixels data. *bioRxiv*.

Chen, H., Chen, X., and Su, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 113(523):1146–1155.

Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association*, 112(517):397–409.

Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, 43(1):139–176.

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

Chen, L. H. and Shao, Q.-M. (2005). Stein's method for normal approximation. *An introduction to Stein's method*, 4:1–59.

Chu, L. and Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. *The Annals of Statistics*, 47(1):382–414.

Csörgö, M., Csörgö, M., and Horváth, L. (1997). Limit theorems in change-point analysis.

Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278.

Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*, pages 1–44.

Fryzlewicz, P. et al. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

Girón, J., Ginebra, J., and Riba, A. (2005). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician*, 59(1):19–30.

Harchaoui, Z. and Cappé, O. (2007). Retrospective mutiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772. IEEE.

Harchaoui, Z., Moulines, E., and Bach, F. R. (2009). Kernel change-point analysis. In *Advances in neural information processing systems*, pages 609–616.

Heard, N. A., Weston, D. J., Platanioti, K., Hand, D. J., et al. (2010). Bayesian anomaly detection methods for social networks. *The Annals of Applied Statistics*, 4(2):645–662.

James, B., James, K. L., and Siegmund, D. (1987). Tests for a change-point. *Biometrika*, 74(1):71–83.

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, C., et al. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *science*, 311(5757):88–90.

Kovács, S., Li, H., Bühlmann, P., and Munk, A. (2020). Seeded binary segmentation: A general methodology for fast and optimal change point detection. *arXiv preprint arXiv:2002.06633*.

Lee, C., Luo, Z., Ngiam, K. Y., Zhang, M., Zheng, K., Chen, G., Ooi, B. C., and Yip, W. L. J. (2017). Big healthcare data analytics: Challenges and applications. In *Handbook of large-scale distributed computing in smart healthcare*, pages 11–41. Springer.

Li, S., Xie, Y., Dai, H., and Song, L. (2019). Scan b-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544.

Liu, Y.-W. and Chen, H. (2022). A fast and efficient change-point detection framework based on approximate k-nearest neighbor graphs. *IEEE Transactions on Signal Processing*, 70:1976–1986.

Liu, Y.-W., Zhu, Y., and Chen, H. (2022). Change-point detection in multiple sequences of high-dimensional/non-euclidean data. *Working paper*.

Londschien, M., Kovács, S., and Bühlmann, P. (2021). Change-point detection for graphical models in the presence of missing values. *Journal of Computational and Graphical Statistics*, pages 1–12.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

Nakai, T., Koide-Majima, N., and Nishimoto, S. (2021). Correspondence of categorical and feature-based representations of music in the human brain. *Brain and behavior*, 11(1):e01936.

Ram, P. and Sinha, K. (2019). Revisiting kd-tree for nearest neighbor search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 1378–1388.

Siegmund, D. (1992). Tail approximations for maxima of random fields. In *Probability Theory: Proceedings of the 1989 Singapore Probability Conference*, pages 147–158. Singapore.

Siegmund, D., Yakir, B., and Zhang, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *The Annals of Applied Statistics*, pages 645–668.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., and Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437).

Visconti di Oleggio Castello, M., Chauhan, V., Jiahui, G., and Gobbini, M. I. (2020). The grand budapest hotel: an fmri dataset in response to a socially-rich, naturalistic movie. *bioRxiv*.

Wang, D., Yu, Y., and Rinaldo, A. (2018). Optimal change point detection and localization in sparse dynamic networks. *arXiv preprint arXiv:1809.09602*.

Wang, H., Tang, M., Park, Y., and Priebe, C. E. (2013). Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing*, 62(3):703–717.

Zeebaree, D. Q., Haron, H., and Abdulazeez, A. M. (2018a). Gene selection and classification of microarray data using convolutional neural network. In *2018 International Conference on Advanced Science and Engineering (ICOASE)*, pages 145–150. IEEE.

Zeebaree, D. Q., Haron, H., and Abdulazeez, A. M. (2018b). Gene selection and classification of microarray data using convolutional neural network. In *2018 International Conference on Advanced Science and Engineering (ICOASE)*, pages 145–150. IEEE.

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645.