Change-point detection for multivariate and non-Euclidean data with local dependency

Hao Chen

Department of Statistics
University of California, Davis
One Shields Avenue
Davis, California 95616
USA
e-mail: hxchen@ucdavis.edu

Abstract: In a sequence of multivariate observations or non-Euclidean data objects, such as networks, local dependence is common and could lead to false change-point discoveries. We propose a new way of permutation – circular block permutation with a random starting point – to address this problem. This permutation scheme is studied on a non-parametric change-point detection framework based on a similarity graph constructed on the observations, leading to a general framework for change-point detection for data with local dependency. Simulation studies show that this new framework retains the same level of power when there is no local dependency, while it controls type I error correctly for sequences with and without local dependency. We also derive an analytic p-value approximation under this new framework. The approximation works well for sequences with length in hundreds and above, making this approach fast-applicable for long data sequences.

AMS 2000 subject classifications: Primary 62G32.
Keywords and phrases: graph-based tests, circular block permutation, non-parametric, scan statistic, high-dimensional data, non-Euclidean data, tail probability, analytic p-value approximation.

1. Introduction

Change-point detection is a widely studied problem in statistics and has its applications in many fields. In the typical formulation, we have observations \( \{y_t : t = 1, \ldots, n\} \) over time (or some other meaningful orderings, such as a one-dimensional spatial domain), and test whether there exists \( \tau \in \{1, \ldots, n - 1\} \) such that the underlying distribution of \( y_t \) changes at \( \tau \). There is a rich literature of this model when \( y_t \)'s are real or integer valued scalars (see Carlstein, Müller and Siegmund (1994); Csörgő and Horváth (1997) for a survey).

As we enter the big data era, change-point analysis for sequences of multivariate observations or non-Euclidean data objects is gaining more and more attentions. For example, in text or sequence analysis, each observation in the sequence could be a vector of word counts over a large dictionary of words.
Network data is ubiquitous nowadays as well. Email, phone and online chat records can be used to construct networks of social interactions among individuals (Kossinets and Watts, 2006; Eagle, Pentland and Lazer, 2009). A large part of these studies is characterizing how the network evolves through time. Here, each observation is a network and one might ask whether there is an abrupt shift in network connectivity at any point in time. In these sequences of complicated data types, it is common that observations are autocorrelated. For example, relationships among people last over an extended time period and the social networks have serial correlations.

A closely related field is time series data analysis. The ARCH model proposed by Engle (1982) and the GARCH model proposed by Bollerslev (1986) and their variants were widely used for studying one-dimensional time series data. There are many generalizations to accommodate multivariate time series data (see for examples Bauwens, Laurent and Rombouts (2006); Silvennoinen and Teräsvirta (2009); Aue et al. (2009) and references therein). These models are useful for low-dimensional data and/or for detecting specific types of changes. For high-dimensional data, tests based on these parametric models cannot be applied or lack power unless some strong assumptions are made on the data to avoid the estimation of the large number of nuisance parameters.

In this work, we restrain the change-point detection problem to locally dependent data, in which we are able to develop a general framework for high-dimensional data and non-Euclidean data objects. We leave the problem of long-range dependence for future studies. The proposed 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. When there is local dependence in the data, the method in Chen and Zhang (2015) could result in more false discoveries than it supposed to have (details see in Section 3). To address this problem, we propose 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. Moreover, simulation studies show that it retains the same level of power when the observations in the sequence are independent.

In the following, we use \( \{y_1, y_2, \ldots, y_n\} \) to denote the data sequence, where \( y_i \) could lie in a high-dimensional or a non-Euclidean space. We focus on the single change-point alternative to illustrate the idea, i.e., there possibly exists at most one change-point. The proposed procedure can be extended to the changed interval alternative and to multiple change-points. Discussions on these extensions see in Section 6.2.

The rest of the paper is organized as follows. Section 2 briefly reviews the method introduced in Chen and Zhang (2015) in utilizing a similarity graph constructed on observations for change-point detection. You can skip reading this section if you are familiar with this method. Section 3 discusses the issues of the method in Chen and Zhang (2015) when the sequence is locally dependent, and proposes a new permutation framework to address the issue. Section 4...
discusses more about the proposed test statistic and derives analytic formulas to calculate the test statistic. Section 5 derives the analytic p-value approximations for the test statistic and check how the approximations work for finite samples. We concludes the paper by discussing the choice of the block size in the new framework and the extension of the proposed framework to multiple change-points in Section 6.

2. A brief review of Chen and Zhang (2015)

In Chen and Zhang (2015), the authors assume that \( y_1, y_2, \ldots, y_n \) are independent. They adapted the edge-count test to the scan statistic framework: Each \( t \) divides the observations into two samples, \( \{y_1, \ldots, y_t\} \) and \( \{y_{t+1}, \ldots, y_n\} \), and the edge-count test is conducted to test whether these two samples are from the same distribution or not. Then, the maximum of the scan statistics over \( t \) is used as the test statistic.

The edge-count test, introduced in Friedman and Rafsky (1979), is a two-sample test that is based on a similarity graph constructed on the pooled observations of two samples. The similarity graph can be a given graph that reflects the similarity between observations (Chen and Zhang, 2013). More generally, it can be constructed based on a similarity measure through a certain criterion, such as a minimum spanning tree (MST) (Friedman and Rafsky, 1979), which is a tree connecting all observations with the total distance across edges minimized, a minimum distance pairing (Rosenbaum, 2005), or a nearest neighbor graph (Henze, 1988).

The edge-count test counts the number of edges in the graph that connect from different samples and reject the null of equal distribution when the count is significantly smaller than its null expectation – when the two samples are from the same distribution, the two samples are well mixed and this count is relatively large, so a small count is an evidence for rejecting the null of equal distribution. Let \( G \) be the similarity graph on all observations in the sequence. The edge-count test statistic at time \( t \) is:

\[
R_G(t) = \sum_{(i,j) \in G} I_{g_i(t) \neq g_j(t)}, \quad g_i(t) = I_{i > t},
\]

where \( I_A \) is the indicator function that takes value 1 if event \( A \) is true and 0 otherwise. Figure 1 illustrates the computation of \( R_G(t) \) on a small artificial data set (the observations are in 2-dimension for illustration purpose).

Under the null hypothesis of no change-point and the assumption that \( y_i \)'s are independent, the joint distribution of \( \{y_i : i = 1, \ldots, n\} \) is the same under permutation. Hence, the null distribution of \( R_G(t) \) is defined to be the permutation distribution, which places \( 1/n! \) probability on each of the \( n! \) permutations of \( \{y_i : i = 1, \ldots, n\} \). When there is no further specification, we denote by \( P_P, E_P, \) \( \text{Var}_P \) probability, expectation, and variance, respectively, under the permutation null distribution.
Fig 1. The computation of $R_G(t)$ for 4 different $t$’s on a small artificial data set of length $n = 20$ with $G$ be the MST on the Euclidean distance. The index of each observation is beside each point. The first 10 points are randomly drawn from $N(0, I_2)$ and the next 10 points are randomly drawn from $N((2, 2)^T, I_2)$. Each $t$ divides the observations into two groups, one group for observations before and at $t$ (shown as circles) and the other group for observations after $t$ (shown as triangles). Edges that connect observations from the two different groups are emboldened in the graph. $G$ does not change as $t$ changes, but the group identities of some observations change, causing $R_G(t)$ to change.

The authors standardized $R_G(t)$ so that it is comparable across $t$. Let

$$Z_G(t) = -\frac{R_G(t) - \mathbb{E}_P(R_G(t))}{\sqrt{\text{Var}_P(R_G(t))}}.$$  \hfill (2.1)

The sign is flipped that a large $Z_G(t)$ indicates a change-point. The analytic expressions for $\mathbb{E}_P(R_G(t))$ and $\text{Var}_P(R_G(t))$ are given in Chen and Zhang (2015). The null hypothesis of no change-point is rejected if the scan statistic

$$\max_{n_0 \leq t \leq n_1} Z_G(t), \quad (n_0, n_1 \text{ prespecified})$$  \hfill (2.2)

is greater than a threshold. When $n$ is small, this threshold could be determined by performing random permutations directly; when $n$ is large, Chen and Zhang (2015) provided accurate analytic formulas to approximate the permutation $p$-value, allowing fast application of the method. The authors also shown through simulations that this graph-based testing framework has better power than likelihood-based methods when the dimension of the data is moderate to high.

3. A circular block permutation framework for locally dependent data

The method in Chen and Zhang (2015) assume that the observations are independent. Under the independence assumption, we can permute the order of the observations to get a pool of sequences that have the same distribution as the original sequence under the null hypothesis of no change-point. However, when there is dependence structure within the sequence, such as autocorrelation, the above argument no longer holds. For example, if we apply the scan statistic (2.2) in Chen and Zhang (2015) to an autocorrelated sequence and use the $p$-value calculated based on permutation, it rejects the null hypothesis more often than it should do (Figure 2, right panel).
Fig 2. Histograms of p-values using the method in Chen and Zhang (2015) in testing homogeneity of 10,000 sequences of no change-point. Left panel: the observations in each sequence are independently generated from multivariate normal distribution \( y_t \overset{iid}{\sim} \mathcal{N}(0, \Sigma), \ d = 10, \Sigma(i,j) = |i-j|^{0.6}, \ n = 200 \). Right panel: each sequence is generated from the multivariate autoregression model \( y_t = \rho y_{t-1} + \varepsilon_t, \ t = 1, \ldots, n, \) with \( y_0 \sim \mathcal{N}(0, \Sigma^{1-\rho^2}), \varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} \mathcal{N}(0, \Sigma), \rho = 0.1, \ d = 10, \ n = 200 \).

When the sequence has dependency over time, the permutation null distribution is no longer a good surrogate to the true null distribution as permutation destroy the local structure. If the dependency structure can be removed from the sequence, the remaining sequence with independent observations can be analyzed through the method in Chen and Zhang (2015). This is, however, not realistic for many applications with high-dimensional/non-Euclidean data sequences. To address the local dependence issue, we propose a new null distribution that serves as a better surrogate to the true null distribution than the permutation null.

3.1. Circular block permutation with a random starting point

The block-resampling bootstrap was proposed by Künsch (1989) and independently by Liu and Singh (1992) as a resampling procedure for weakly dependent stationary observations. The idea is that the dependency structure is preserved within the blocks. This was extended to circular block resampling bootstrap by wrapping the data around in a circle before blocking them (Politis and Romano, 1992). For change-point analysis, the block permutation with fixed blocks starting from the first observation and the circular block bootstrap were studied on dependent data for one-dimensional observations (Kirch, 2006).

In light of these studies, we propose to use circular block permutation with a random starting point to generate a pool of sequences representing realizations from approximate distributions of the original sequence with local dependency under the null of no change. The recipe with block size \( L \) on a sequence of length \( n \) is as follows:
(1) The starting point is chosen uniformly from the \( n \) observations, which is denoted as \( k_0 \). If \( k_0 > 1 \), the first \( k_0 - 1 \) observations are moved to the end of the sequence: \( \{y_{k_0}, \ldots, y_n, y_1, \ldots, y_{k_0-1}\} \).

(2) The new sequence is blocked into \( \left\lfloor n/L \right\rfloor \) blocks of size \( L \) starting from the first observation \( y_{k_0} \). It is possible that the last block has less than \( L \) observations.

(3) The \( \left\lfloor n/L \right\rfloor \) blocks are permuted.

For the above recipe, it is easy to see that the resulting sequence from the circular block permutation with a random starting point is one permutation of the original sequence – each observation appears in the resulting sequence and appears only once. Then, the similarity graph on the resulting sequence is the same as that on the original sequence, which makes theoretical analysis on this framework tractable. In addition, the randomized starting point ensures that the probability of any observation \( y_i \) appears at any location \( j \) in the resulting sequence is \( 1/n \), ensuring unbiasedness.

To make the theoretical treatment more tractable, we work under the following variant of the framework: We first augment the sequence by \( x \) \((0 \leq x < L)\) pseudo observations by adding them to the end of the sequence so that \( n + x \) is divisible by \( L \). These augmented \( x \) observations have no edge connected to any other observations. Then the recipe described above is applied to this augmented sequence with \( n + x \) observations. This variant is the same as the original version when \( n \) is divisible by \( L \) and works similarly when not. In this variant, all the blocks are of size \( L \), so the theoretical treatments are much more tractable. In the following, we work under this variant and short the framework as ‘circular block permutation’ or ‘CBP’ for simplicity. We also use \( n \) to denote \( n + x \) for simplicity. We use \( P_{\text{CBP}}, E_{\text{CBP}}, \text{Var}_{\text{CBP}} \) to denote probability, expectation, and variance, respectively, under this framework. We consider \( L \) to be fixed for the rest of the paper. A discussion on how to choose \( L \) in a data driven way is in Section 6.1.

### 3.2. Performance under CBP

The standardized edge-count statistic under the circular block permutation can be defined as:

\[
Z_{G, \text{CBP}}(t) = -\frac{R_G(t) - E_{\text{CBP}}(R_G(t))}{\sqrt{\text{Var}_{\text{CBP}}(R_G(t))}}.
\]  

(3.1)

The scan statistic is then defined as:

\[
\max_{n_0 \leq t \leq n_1} Z_{G, \text{CBP}}(t).
\]  

(3.2)

In the following, with no further specification, we set \( n_0 = \lfloor 0.05n \rfloor \) and \( n_1 = n - n_0 \).

---

\( ^1 \)For a real value \( s \), we use \( \lfloor s \rfloor \) to denote the largest integer that is no larger than \( s \).
Figure 3 shows the histograms of $p$-values in testing the homogeneity of sequences when the sequence has no change-point. For each sequence, the $p$-value of the test is obtained through doing 100,000 CBPs, respectively. Now we see that the type I error is correctly controlled for sequences of autocorrelated data.

An immediate follow-up question is whether the improvement in controlling the type I error of the circular block permutation framework come with a sacrifice on its power. To get an idea of this issue, we compare the power of the two frameworks. Table 1 shows the estimated power based on 1,000 simulation runs. In each simulation run, the sequence is generated in the same way as in Figure 2, while here, there is a mean shift in the middle of the sequence with the $L_2$ distance between the means before and after the change to be 2. This specific alternative is chosen so that the tests have moderate power. We see that when data are independent, the power under CBP is similar to that under permutation. When the sequence is autocorrelated, the permutation framework cannot be used, while the CBP framework has power on par with the independent scenario.

|                      | permutation | CBP, $L = 5$ |
|----------------------|-------------|--------------|
| Independent data     | 0.791       | 0.779        |
| Autorrelated data    | –           | 0.775        |

These simulation results show the ability of CBP in controlling the type I error rate and at the same time keeping substantial power for sequences with local dependency. In the above simulation runs, the expectation and variance of $R_G(t)$ under CBP, as well as the $p$-value of the test, are calculated by randomly sampling from the circular block permutation distribution. This is very time consuming if one wants to get a good estimate of them. In the following, we work on analytic expressions (or approximate analytic expressions when the
exact analytic expression is hard to obtain) of these quantities to make this
framework easy to use in practice.

4. Analytic expressions under CBP

Set \( m \equiv n/L \), there are in total \( L \times m \) CBPs and it is very time consuming
to draw all these CBPs when \( m \) is moderate to large. In this section, we derive
exact analytic expression for \( E_{\text{CBP}}(R_G(t)) \) (Section 4.1) and an approximate
analytic expression for \( \text{Var}_{\text{CBP}}(R_G(t)) \) (Section 4.2). In the following, for any
scalar \( s \), we use \((s)_+\) to denote \( \max(0,s) \); and for set \( S \), we use \(|S|\) to denote
the number of elements in the set.

4.1. \( E_{\text{CBP}}(R_G(t)) \)

Let \( \pi_{\text{CBP}}(i) \) be the index of \( y_i \) under the circular block permutation. Then
\[
E_{\text{CBP}}(R_G(t)) = \sum_{(i,j) \in G} P(g_{\pi_{\text{CBP}}(i)}(t) \neq g_{\pi_{\text{CBP}}(j)}(t)) \\
= 2 \sum_{(i,j) \in G} P(\pi_{\text{CBP}}(i) \leq t, \pi_{\text{CBP}}(j) > t).
\]  

The design of the circular block permutation ensures that, besides \( t \), \( n \) and the
block size \( L \), \( P(\pi_{\text{CBP}}(i) \leq t, \pi_{\text{CBP}}(j) > t) \) only depends on \( \delta_{ij} = \min(|i - j|, n - |i - j|) \), which is the smaller index difference between \( y_i \) and \( y_j \) in the
circle formed by linking the end of the sequence to its start. In particular, the
probability depends on which one of the following categories \( \delta_{ij} \) belongs to:
\[
\{
\delta_{ij} = 1\}, \ldots, \{\delta_{ij} = L - 1\}, \{\delta_{ij} \geq L\}.
\]
The probability is the same in each
category and can be calculated exactly for each of them. Hence, we can classify
the edges in \( G \) according to these categories and get Theorem 4.1.

Theorem 4.1. For each \( t \in \{1, \ldots, n\} \), we write \( t \) in the form of \( t = aL + b \)
where \( a = [t/L] \), \( b = t - aL \), then
\[
E_{\text{CBP}}(R_G(t)) = \sum_{k=1}^{L} 2 p(k, a, b)|\mathcal{E}_k|,
\]
where
\[
p(k, a, b) = (\delta_{ij} - b) + \frac{a(m-a)}{m(m-1)} + (b - (L - k)) + \frac{(a+1)(m-a-1)}{m(m-1)} \\
+ (\min(b, L - k) - (b - k)) + \frac{a(m-a-1)+(m-1)}{m(m-1)},
\]
\[
\mathcal{E}_k = \{(i, j) \in G : \delta_{ij} = k\}, \quad k = 1, \ldots, L - 1,
\]
\[
\mathcal{E}_L = \{(i, j) \in G : \delta_{ij} \geq L\}.
\]

Remark 4.2. When \( t \) is divisible by \( L \) \((t = aL)\), we have \( p(k, a, 0) = \delta_{ij} \frac{a(m-a)}{m(m-1)} \).
Then \( E_{\text{CBP}}(R_G(t)) = \frac{2a(m-a)}{m(m-1)} \sum_{k=1}^{L} k|\mathcal{E}_k| \).
Proof of Theorem 4.1. We compute the probability

\[ P(\pi_{\text{CBP}}(i) \leq t, \pi_{\text{CBP}}(j) > t) \]  

(4.2)

under difference scenarios.

When \( \delta_{ij} \geq L \), \( y_i \) and \( y_j \) are always in two different blocks. If \( b = 0 \), then \( \pi_{\text{CBP}}(i) \leq t \) only if the block containing \( y_i \) is placed in the first \( a \) blocks after the circular block permutation, and \( \pi_{\text{CBP}}(j) > t \) only if the block containing \( y_j \) is placed in the last \( m - a \) blocks after the circular block permutation. So the probability (4.2) is

\[
\frac{a}{m} \frac{m-a}{m-1} = \frac{L_a(m-a)}{n(m-1)}.
\]

(If \( b > 0 \), we need to discuss whether the block containing either \( y_i \) or \( y_j \) sits on \( t \), whether \( y_i \) is in the first \( b \) observations in the block, and whether \( y_j \) is in the last \( L - b \) observations in the block. Enumerating all possibilities, the probability (4.2) is

\[
\frac{a}{m} \left( \frac{m-a-1}{m-1} + \frac{1}{m-1} \frac{L-b}{L} \right) + \frac{1}{m} \frac{b}{m} \frac{m-a-1}{m-1} = \frac{L_a(m-a) + b(m-2a-1)}{n(m-1)}.
\]

When \( \delta_{ij} < L \), we also discuss the two scenarios: \( b = 0 \) and \( b > 0 \).

If \( b = 0 \), \( y_i \) and \( y_j \) need to be in different blocks to have \( \pi_{\text{CBP}}(i) \leq t, \pi_{\text{CBP}}(j) > t \). Among the \( L \) possible ways in blocking the sequence, \( \delta_{ij} \) of them have \( y_i \) and \( y_j \) in different blocks, so the probability (4.2) is

\[
\frac{\delta_{ij} a(m-a)}{L m(m-1)} = \frac{\delta_{ij} a(m-a)}{n(m-1)}.
\]

If \( b > 0 \), \( y_i \) and \( y_j \) could be in different blocks or in the same block to satisfy \( \pi_{\text{CBP}}(i) \leq t, \pi_{\text{CBP}}(j) > t \). If they are in the same block, we denote that particular block by \( B_1 \). If they are in different blocks, the two blocks much be adjacent. Among the two blocks, we denote the left block to be \( B_1 \) (to make this argument consistent, the first block of the sequence and the last block of the sequence are considered to be adjacent and the last block of the sequence is considered to be on the left of the first block of the sequence). We then let the adjacent block right of \( B_1 \) to be \( B_2 \). For block \( B_1 \), we further divide it into two sub-regions with \( B_{1,t} \) denoting the first \( b \) location(s) of the block and \( B_{1,r} \), denoting the rest \( L - b \) location(s) of the block. We define \( B_{2,t} \) and \( B_{2,r} \) similarly for block \( B_2 \).

Then, there are four configurations for the placements of \( i \) and \( j \) for each of the two scenarios: (i) \( i \) on the left of \( j \) within \( B_1 \cup B_2 \), and (ii) \( i \) on the right of \( j \) within \( B_1 \cup B_2 \). The four configurations are listed in Tables 2 and 3 for these two scenarios, respectively. Together in the tables are the probability of having each of the configuration out of \( L \) different ways of doing the blocking (Prob. 1 in the tables) and the proportion of the permutations in terms of permuting the blocks so that \( \pi_{\text{CBP}}(i) \leq t, \pi_{\text{CBP}}(j) > t \) given the configuration (Prob. 2 in the tables). For each of the two scenarios, summing over the product of the two probabilities (Prob. 1 and Prob. 2 in the tables) gives (4.2).
Hao Chen/Change-point detection for data with local dependency

For variance, we need to figure out $\text{E}_{\text{CBP}}(R_G^2(t)) = \sum_{(i,j),(u,v) \in G} P(g_{\pi_{\text{CBP}}(i)}(t) \neq g_{\pi_{\text{CBP}}(j)}(t), g_{\pi_{\text{CBP}}(u)}(t) \neq g_{\pi_{\text{CBP}}(v)}(t))$. Then, $\text{Var}_{\text{CBP}}(R_G(t))$ follows as $\text{E}_{\text{CBP}}(R_G^2(t)) - (\text{E}_{\text{CBP}}(R_G(t)))^2$ with the analytic expression for $\text{E}_{\text{CBP}}(R_G(t))$ provided in Section 4.1.

When $i, j, u, v$ are all different, we need to consider $\binom{4}{2} = 6$ index-pairs and whether they could be within a block or not. This is much more complicated than the calculation in $\text{E}_{\text{CBP}}(R_G(t))$ where only one index-pair is considered. In the derivation of $\text{E}_{\text{CBP}}(R_G(t))$, when $t$ is divisible by $L$ ($b = 0$ in the proof), the derivation is much easier. Therefore, we work out the exact analytic expression for $\text{Var}_{\text{CBP}}(R_G(t))$ for $t$ divisible by $L$, which is already very complicated, and do extrapolations for other $t$’s. We then compare the result obtained in this way with doing circular block permutation directly. The following theorem gives exact analytic expression for $\text{Var}_{\text{CBP}}(R_G(t))$ when $t = aL$, $a = 0, \ldots, m$.

### Table 2

| $B_{1,1}$ | $B_{1,r}$ | $B_{2,l}$ | $B_{2,r}$ | Prob. 1 | Prob. 2 |
|-----------|-----------|-----------|-----------|---------|---------|
| $i$ | $j$ | $\min\{i, j, -i, -j\}$ | $\frac{(i - (L - \delta_i))}{L}$ | $\frac{1}{m}$ | $\frac{(a + 1)(m - a - 1)}{m(m - 1)}$ |
| $i$ | $j$ | $\min\{i, j, -i, -j\}$ | $\frac{L}{L}$ | $\frac{m(m - a - 1)}{m(m - 1)}$ | $\frac{a(m - a)}{m(m - 1)}$ |
| $i$ | $j$ | $\min\{i, j, -i, -j\}$ | $\frac{(j - \delta_i)}{L}$ | $\frac{a(m - a)}{m(m - 1)}$ | $\frac{a(m - a)}{m(m - 1)}$ |

Since $\frac{1}{m} + \frac{a(m-a-1)}{m(m-1)} = \frac{a(m-a)+(m-a-1)}{m(m-1)}$, both summations give, for $\delta_{ij} < L$,

$$(\min(b, L - \delta_{ij}) - (b - \delta_{ij}) + \frac{a(m-a-1)+(m-1)}{m(m-1)}) + (b - (L - \delta_{ij})) + \frac{a(m-a)}{m(m-1)} + (\delta_{ij} - b) + \frac{a(m-a)}{m(m-1)}.$$

Then, the theorem follows as the result for $\delta_{ij} \geq L$ is a special case of the above expression with $\delta_{ij}$ replaced by $L$. \qed

### 4.2. $\text{Var}_{\text{CBP}}(R_G(t))$

For variance, we need to figure out $\text{E}_{\text{CBP}}(R_G^2(t)) = \sum_{(i,j),(u,v) \in G} P(g_{\pi_{\text{CBP}}(i)}(t) \neq g_{\pi_{\text{CBP}}(j)}(t), g_{\pi_{\text{CBP}}(u)}(t) \neq g_{\pi_{\text{CBP}}(v)}(t))$. Then, $\text{Var}_{\text{CBP}}(R_G(t))$ follows as $\text{E}_{\text{CBP}}(R_G^2(t)) - (\text{E}_{\text{CBP}}(R_G(t)))^2$ with the analytic expression for $\text{E}_{\text{CBP}}(R_G(t))$ provided in Section 4.1.
Theorem 4.3. For $t = aL, a = 1, \ldots, m$, we have

\[
\text{Var}_{CBP}(R_{G}(t)) = c_1p_1(a) + c_2p_2(a) + c_3p_3(a) - c_0^2p_1^2(a),
\]

where

\[
p_1(a) = \frac{2a(m-a)}{m(m-1)}, \quad p_2(a) = \frac{1}{2}p_1(a), \quad p_3(a) = \frac{4a(a-1)(m-a)(m-a-1)}{m(m-1)(m-2)(m-3)},
\]

\[
c_0 = \frac{1}{L} \sum_{k=1}^{L} k|\mathcal{E}_k|,
\]

\[
c_1 = \frac{1}{L} \sum_{k=1}^{L} k|\mathcal{E}_k|,
\]

\[
+ \frac{1}{L} \sum_{(i,j), (i,u) \in G; j \neq u} \{(L - \delta_{ju})I_{h_{0}(i,j,u)} < 3, \delta_{ju} < L
\]
\[+ \min(\delta_{ij}, \delta_{iu})I_{h_{0}(i,j,u)} = 3 I_{\max(\delta_{ij}, \delta_{iu}, \delta_{ju})} \neq \delta_{ju}) \}
\]

\[
+ \frac{1}{L} \sum_{(i,j), (i,u), (u,v) \in G; i, j, u, v \text{ all different}} \{(L - \delta_{ju} - \delta_{ju} + x(\delta_{ju} - \delta_{ju}))I_{h_{0}(i,j,u,v)} < L
\]
\[+ (L - \delta_{ju} + x(\delta_{ju} - \delta_{ju}))I_{h_{1}(i,j,u,v)} < L, \delta_{ju} < L)\}
\]

\[
c_2 = \frac{1}{L} \sum_{(i,j), (i,u) \in G; j \neq u} \{L I_{h_{0}(i,j,u)} = 0 + \min(\delta_{ij}, \delta_{iu}, \delta_{ju})I_{h_{0}(i,j,u)} = 1
\]
\[+ (\max(\delta_{ij}, \delta_{iu}, \delta_{ju}) - L)I_{h_{0}(i,j,u)} = 2) \}
\]

\[
+ \frac{1}{L} \sum_{(i,j), (i,u), (u,v) \in G; i, j, u, v \text{ all different}} \{(L - \delta_{iu} - \delta_{ij} + x(\delta_{ij} - \delta_{iu}))I_{h_{0}(i,j,u,v)} < L
\]
\[+ (L - \delta_{iu} + x(\delta_{ij} - \delta_{iu}))I_{h_{1}(i,j,u,v)} < L, L \delta_{uv} < L)\}
\]
For number of edges whose end nodes appear in different blocks over all \( y \)

Hence, there are only \( \omega \)

same node. Then, under a certain blocking

i.e., they could degenerate to be the same edge; or the two edges could share the

\( c \)

edges whose end nodes appear in three distinct blocks with both edges having

\( c \)

edges whose end nodes only appear in two distinct blocks with both edges having

\( \delta \)

\( 1 \)

\( \min \)

Based on the recipe of the circular block permutation, the blocks resulted from a random

2

\( \max \)

, a high level explanation is that it is the average of the

number of pairs of

\( s(i, j) = \left\{ \begin{array}{ll} \min(i, j) & \text{if } |i - j| < L, \\
\max(i, j) & \text{if } n - |i - j| < L. \end{array} \right. \)

\( \delta_{ij, uv} = s(u, v) - s(i, j) \)

\( b_{ij, uv} = \delta_{ij, uv} \mod L \)

\( \delta_{ij, uv} = (\min(\delta_{ij}, b_{ij, uv} + \delta_{uv}) - b_{ij, uv})_+ + (\min(\delta_{ij}, b_{ij, uv} + \delta_{uv} - L) + L)_+ \)

\( O_{\omega} = \delta_{ij, \delta_{iu} \delta_{ju} \delta_{iv} \delta_{jv} < L} + \delta_{ij, \delta_{iu} \delta_{ju} \delta_{iv} \delta_{jv} < L} + \delta_{ij, \delta_{iu} \delta_{ju} \delta_{iv} \delta_{jv} < L} \)

\( c_3 = \frac{1}{L} \sum_{(i, j, (u, v) \in C} \{ LI_{h_1(i, j, u, v)} = 0 + \delta_{min}(i, j, u, v)I_{h_1(i, j, u, v)} = 1 \}

\( h_0(i, j, u) = I_{\delta_{ij} < L} + I_{\delta_{iu} < L} + I_{\delta_{ju} < L} \)

\( h_1(i, j, u, v) = I_{\delta_{ij} < L} + I_{\delta_{iu} < L} + I_{\delta_{ju} < L} + I_{\delta_{iv} < L} + I_{\delta_{ju} < L} \)

\( \delta_{max}(i, j, u, v) = \max(\delta_{ij}, \delta_{iu}, \delta_{iv}, \delta_{ju}, \delta_{jv}, \delta_{uv}) \),

\( \delta_{min}(i, j, u, v) = \min(\delta_{ij}, \delta_{iu}, \delta_{iv}, \delta_{ju}, \delta_{jv}, \delta_{uv}) \),

\( \delta_{min, 2}(i, j, u, v) = \sum \text{of the two smallest values among } \{ \delta_{ij}, \delta_{iu}, \delta_{iv}, \delta_{ju}, \delta_{jv}, \delta_{uv} \} \),

\( \delta_{min, 3}(i, j, u, v) = \sum \text{of the three smallest values among } \{ \delta_{ij}, \delta_{iu}, \delta_{iv}, \delta_{ju}, \delta_{jv}, \delta_{uv} \} \),

\( s(i, j) = \left\{ \begin{array}{ll} \min(i, j) & \text{if } |i - j| < L, \\
\max(i, j) & \text{if } n - |i - j| < L. \end{array} \right. \)

\( \delta_{ij, uv} = s(u, v) - s(i, j) \)

\( b_{ij, uv} = \delta_{ij, uv} \mod L \)

\( x_{ij, uv} = (\min(\delta_{ij}, b_{ij, uv} + \delta_{uv}) - b_{ij, uv})_+ + (\min(\delta_{ij}, b_{ij, uv} + \delta_{uv} - L) + L)_+ \)

\( O_{\omega} = \delta_{ij, \delta_{iu} \delta_{ju} \delta_{iv} \delta_{jv} < L} + \delta_{ij, \delta_{iu} \delta_{ju} \delta_{iv} \delta_{jv} < L} + \delta_{ij, \delta_{iu} \delta_{ju} \delta_{iv} \delta_{jv} < L} \)

\( \sum_{\omega \in \Omega} c_0(\omega) \)

\( c_0, c_1, c_2, c_3 \)

\( \omega \)

\( L \)

\( \omega \)

blocks

Hao Chen/Change-point detection for data with local dependency

Based on the recipe of the circular block permutation, the blocks resulted from a random

starting point at \( y \), are the same as the blocks resulted from a random starting point at \( y_{t+L} \).

Hence, there are only \( L \) different ways of blocking.
their end nodes appearing in different blocks, and \( c_3(\omega) \) is the number of pairs of edges whose end nodes appear in four distinct blocks. It is not hard to see that \( c_1(\omega) + c_2(\omega) + c_3(\omega) = c_0^2(\omega) \). Then,

\[
c_1 + c_2 + c_3 = \frac{1}{L} \sum_{\omega \in \Omega} c_0^2(\omega) \geq \left( \frac{1}{L} \sum_{\omega \in \Omega} c_0(\omega) \right)^2 = c_0^2.
\]

When \( L = 1 \), the equality always holds. Indeed, when \( L = 1 \), the coefficients can be simplified to be \( c_0 = c_1 = |G| \), \( c_2 = \sum_{i=1}^{n} |G_i|^2 - 2|G| \) and \( c_3 = |G|^2 - \sum_{i=1}^{n} |G_i|^2 + |G| \), where \( |G| \) is the number of edges in the graph \( G \), and \( G_i \) is the subgraph in \( G \) that connect to node \( \mathbf{y}_i \). So \( |G_i| \) is the degree of node \( \mathbf{y}_i \). It is clear that \( c_1 + c_2 + c_3 = |G|^2 = c_0^2 \). However, for \( L > 1 \), the equality usually does not hold unless under very special cases that \( c_0(\omega) \) is the same for all \( \omega \in \Omega \).

In Theorem 4.3, we provide the exact analytic expression for \( \text{Var}_{\text{CBP}}(R(t)) \) when \( t \) is divisible by \( L \). Unless \( L = 1 \) that the coefficients, \( c_1, c_2 \) and \( c_3 \), could be greatly simplified, the expressions for these coefficients are in general fairly complicated for \( L > 1 \). It can be imagined that the exact analytic expression for \( \text{Var}_{\text{CBP}}(R(t)) \) when \( t \) is not divisible by \( L \) would be much more complicated. In addition, the computation time for these coefficients when \( L > 1 \) is not negligible as it needs to do some complicated counting. If there are special structures of the graph, the expression might be simplified using the structures. While for a general graph \( G \), the analytic expression could not be further simplified as two edges are involved in computing the variance and we need to consider all the possible combinations of whether the \( \binom{4}{2} = 6 \) pairwise index differences are smaller than \( L \) or not when the four end points of the two edges are distinct. In a typical run for a 1,000-length sequence with local dependence, the computation time for getting all the coefficients with \( L = 5 \) is about 8 second on the 12-inch MacBook (2015), which is acceptable; while the not-derived analytic expressions for the variance under CBP at \( t = aL + b, 0 < b < L \) would be much more complicated (the magnitude of the complication different can be inferred from the exact analytic expression of \( \text{E}_{\text{CBP}}(R(t)) \) in Section 4.1 under \( b = 0 \) and \( b > 0 \)) and require much more time to compute. Combining all the factors, we propose to fill-in \( \text{Var}_{\text{CBP}}(R_G(t)) \) at \( t \) not divisible by \( L \) by extrapolating from the values at \( t = aL, a = 0, 1, \ldots, m \).

Figure 4 shows the standard deviation of \( R_G(t) \) under circular block permutation (\( \text{SD}_{\text{CBP}}(R_G(t)) \)) for a 1000-length sequence with local dependence. The blue line in each plot is based on the analytic formula provided in Theorem 4.3 with the values at \( t = aL + b, 0 < b < L \) filled in by extrapolation, and the black line in each plot is based on 900 CBPs (top panels) and 90,000 CBPs (bottom panels). Here, 900 CBPs were chosen as it uses a similar amount of time to that by computing the standard deviation based on Theorem 4.3 and extrapolation for this sequence. We can see clearly from the top panels that 900 CBPs is not enough as the results are fluctuating widely. It is important to have a good estimate of \( \text{SD}_{\text{CBP}}(R_G(t)) \) as it standardize the raw statistic \( R_G(t) \) and a bad estimate could lead to a bad estimate of the change-point location. When we
Fig. 4. Standard deviation of $R_G(t)$ under circular block permutation from the analytic expression given in Theorem 4.3 (the values at $t = aL + b, 0 < b < L$ are extrapolated) and from doing circular block permutation directly. The left panels are plotting the whole line and the right panels are plotting the middle part of the sequence. In each plot, the blue line is based on the analytic formula and the black line is based on directly doing circular block permutation with 900 CBPs for the top panels and 90,000 CBPs for the bottom panels.

increase the number of CBPs to 90,000 (using 100-fold times as for getting the analytic results), the values based on CBPs directly is much better (bottom left panel). However, if we zoom into the middle part of the sequence, we could still see the black line wiggling around, which could cause inaccurate estimate of the change-point location. This toy example shows that we would need even more number of CBPs to get an estimate that is as good as those from the analytic formula with extrapolation.

Therefore, we recommend to use the analytic formula given in Theorem 4.3 to get exact values at $t = aL, a = 0, 1, \ldots, m$ and use extrapolation to get values at $t = aL + b, 0 < b < L$. This approach gives us accurate enough estimate for $\text{Var}_{\text{CBP}}(R_G(t))$ with a reasonable fast enough computing time. In the following, $Z_{G,\text{CBP}}(t)$ is defined with $\text{Var}_{\text{CBP}}(R_G(t))$ computed in this recommended way.

5. Analytic $p$-value approximations

Now, we have a relatively fast analytic way to compute the standardized statistic $Z_{G,\text{CBP}}(t)$. The next question is how large the scan statistic

$$\max_{n_0 \leq t \leq n_1} Z_{G,\text{CBP}}(t)$$
needs to be to constitute sufficient evidence against the null hypothesis of homogeneity, i.e., we are concerned with the tail probability of the scan statistics under $H_0$:

$$P(\max_{n_0 \leq t \leq n_1} Z_{G,\text{CBP}}(t) > b). \quad (5.1)$$

To obtain this tail probability, we can directly perform the circular block permutation, which would be time consuming in obtaining a reasonably accurate estimate. Therefore, we seek to derive an analytic expression for this tail probability. In the rest of this section, we first study the asymptotic distribution of the process $\{Z_{G,\text{CBP}}(t)\}$. We derive approximate analytic expression for the tail probability for the limiting process and then refine the approximation to work for finite $n$.

### 5.1. Asymptotic property of the process

Here, we derive the limiting distribution of $\{Z_{G,\text{CBP}}([mw]L) : \epsilon \leq w \leq 1 - \epsilon\}$ for any $0 < \epsilon < 0.5$ under circular block permutation. We first introduce some notations. We write the nodes connected by an edge $e$ be $e_-$ and $e_+$ with $e_- < e_+$. For node $i$ and edge $e = (e_-, e_+)$, let

$$A_{e,L,0} = \{e^* : \min(\delta(e^*, e_-), \delta(e^*, e_+), \delta(e^*, e_-), \delta(e^*, e_+)) < L\},$$

$$A_{e,L,1} = A_{e,L,0} \cup \left( \bigcup_{e' \in G_{e_-} \cup G_{e_+} \forall e' \in A_{e,L,0}} A_{e',L,0} \right),$$

$$A_{e,L,2} = A_{e,L,1} \cup \left( \bigcup_{e' \in G_{e_-} \cup G_{e_+} \forall e' \in A_{e,L,1}} A_{e',L,1} \right),$$

$$A_{i,L,0} = \{e^* : \min(\delta(e^*, i), \delta(e^*, i)) < L\},$$

$$A_{i,L,1} = A_{i,L,0} \cup \left( \bigcup_{e' \in G_{e_-} \cup G_{e_+} \forall e' \in A_{i,L,0}} A_{e',L,0} \right),$$

$$A_{i,L,2} = A_{i,L,1} \cup \left( \bigcup_{e' \in G_{e_-} \cup G_{e_+} \forall e' \in A_{i,L,1}} A_{e',L,1} \right).$$

Here, $A_{e,L,0}$ is the set of edges whose end nodes could be within the same block with any of the end nodes in $e$ under some circular block permutations. This can be viewed as the neighbors of $e$. Then, $A_{e,L,1}$ is the set of edges whose end nodes could be within the same block with any of the end nodes of the edges in $A_{e,L,0}$, so $A_{e,L,1}$ can be viewed as the set containing neighbors of $A_{e,L,0}$. Similarly, $A_{e,L,2}$ can be viewed as the set containing neighbors of $A_{e,L,1}$. The other three sets, $A_{i,L,0}$, $A_{i,L,1}$ and $A_{i,L,2}$, are defined similarly but initiated from a node $i$. 

For a certain block $i$, let $D_i$ be the number of edges in $G$ that connect a node in block $i$ to another node not in block $i$ under the blocking that block $i$ exists. Let $E_{Ω}$ be the expectation that places probability $\frac{1}{r}$ on each $\omega \in Ω$ with $Ω$ defined in Remark 4.4. 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 $b_n$. The limiting distribution of the stochastic process needs the following conditions.

**Condition 1.** $|G| = O(n^\alpha), 1 \leq \alpha < \frac{3}{2}; \sum_{\epsilon \in G} |A_{ε, L, 1}||A_{ε, L, 2}| = o(n|G|^{\frac{1}{2}}); \sum_{i=1}^{n} |A_{i, L, 1}||A_{i, L, 2}| = o(n^{1.5}).$

**Condition 2.** $E_{Ω}(\sum_{i=1}^{m} D_i^2) - \frac{1}{m}(E_{Ω}(\sum_{i=1}^{m} D_i))^2 = O(E_{Ω}(\sum_{i=1}^{m} D_i^2)).$

**Condition 3.** $E_{Ω}(\sum_{i=1}^{m} D_i^2) = O(|G|).$

**Theorem 5.1.** Under Conditions 1 and 2, or under Conditions 1 and 3, as $n \to \infty$, $\forall \epsilon \in (0, 0.5), \{Z_{CBP}(\nu b | L) : \epsilon < w < 1 - \epsilon \}$ converges in finite dimensional distributions to a Gaussian process, which we denote as $\{Z_{CBP}^*(w) : \epsilon < w < 1 - \epsilon \}$.

The complete proof of the theorem is in Appendix A.2. It utilizes a proof technique that is similar to that in Chen and Zhang (2015).

**Remark 5.2.** Condition 1 under $L = 1$ are relaxed versions of the conditions for the limiting distribution under permutation null distribution in Chen and Zhang (2015)\(^3\). For Conditions 2 and 3, we only need one of them to hold. Condition 2 says that the graph shall not be flat: $D_i$’s shall not be of the same order across $i$’s. However, when $|G| = O(n)$, such a flat graph would also be acceptable as Condition 3 holds.

### 5.2. Analytic formulas

We now examine the asymptotic behavior of the tail probability (5.1). Our approximation require the function $\nu(x)$ defined as $\nu(x) = 2x^{-2} \exp \left\{ -2 \sum_{i=1}^{\infty} \frac{1}{i} \Phi \left( -\frac{1}{2} x \sqrt{i} \right) \right\}, x > 0$ This function is closely related to the Laplace transform of the overshoot over the boundary of a random walk. A simple approximation given in Siegmund and Yakir (2007) is sufficient for numerical purpose: $\nu(x) \approx \frac{(2/\sqrt{\pi}) (\Phi(x/2) - 0.5)}{(x/2) \Phi(x/2) + 0.5 x/2}.$

Based on Theorem 5.1, under Conditions 1 and 2 (or 3), following similar arguments in Chen and Zhang (2015), we have as $n \to \infty$, for $b = O(\sqrt{n})$, $n_0, n_1 = O(n),

\begin{equation}
\begin{align*}
P(\max_{n_0 \leq t \leq n_1} Z_{CBP}(t) > b) & \sim b \nu(b) \sum_{n_0 \leq t \leq n_1} C(t) \nu \left( \sqrt{2b^2 C(t)} \right),
\end{align*}
\end{equation}  

where $C(t) = \frac{1}{T} \frac{\partial \rho(s,t)}{\partial s} \bigg|_{s=t}$, with $\rho(s,t) \equiv \text{Cov}_{\text{CBP}}(Z_{\text{CBP}}(s), Z_{\text{CBP}}(t))$. The following theorem gives an analytic expression for $\text{Cov}_{\text{CBP}}(R_G(t_1), R_G(t_2))$ at $t_1$ and $t_2$ divisible by $L$.

\(^3\)We do not consider $|G| = O(n^\alpha), 0 < \alpha < 1$ here as such a graph only makes use of a tiny portion of the observations and is not effective in practice.
Theorem 5.3. For $t_1 = a_1L < t_2 = a_2L$ where $a_1, a_2 \in \{0, 1, \ldots, m\}$, we have

$$Cov_{CBP}(R_C(t_1), R_C(t_2)) = c_1q_1(a_1, a_2) + c_2q_2(a_1, a_2) + c_3q_3(a_1, a_2) - c_0^2p_1(a_1)p_1(a_2),$$

with $c_0, c_1, c_2, c_3$ provided in Theorem 4.3, and

$$q_1(a_1, a_2) = \frac{2a_1(m-a_2)}{m(m-1)};$$

$$q_2(a_1, a_2) = \frac{a_1(m-a_2)(m-2a_1+2a_2-2)}{m(m-1)(m-2)};$$

$$q_3(a_1, a_2) = \frac{4a_1(m-a_2)((a_1-1)(m-a_1-1)+(a_2-a_1)(m-a_1-2))}{m(m-1)(m-2)(m-3)}.$$

The proof of this theorem is in Appendix A.3.

Then, for $t = aL$, $a \in \{1, \ldots, m - 1\}$, we have

$$C(t) = \frac{m(m-1)(h_1(a,m)c_1 + h_2(a,m)c_2 + h_3(a,m)c_3)}{2La(m-a)(h_2(a,m)(2c_1+c_2)+h_3(a,m)c_3+h_4(a,m)c_4)}$$

where

$$h_1(a,m) = 2m(m-2)(m-3),$$

$$h_2(a,m) = (m-3)((m-2a)^2-2m),$$

$$h_3(a,m) = -4(m-2a)^2+4m,$$

$$h_4(a,m) = m(m-1)(m-2)(m-3)$$

$$h_5(a,m) = 4m(m-1)(a-1)(m-a-1),$$

$$h_6(a,m) = -4a(m-a)(m-2)(m-3).$$

Based on the above results, the $p$-value approximation is reasonably well for low to moderate dimension, but not that well when the dimension is high (details see in Section 5.3, Table 4). The reason is that when the dimension is high, the convergence of $Z_{CBP}(t)$ to the Gaussian distribution is low when $t$ closes to the two ends and, for finite samples, $Z_{CBP}(t)$ could be quite skewed. Figure 5 plots the skewness of $Z_{CBP}(t)$ estimated from 100,000 random circular block permutations for two data sequences generated based on model M2 under scenarios 2 and 3 provided in Section 5.3. We can see that the skewness is quite severe when $t$ is close to the two ends of the sequence and when the dimension is high. Chen and Zhang (2015) discussed the same issue under the permutation framework, here, we adopt a similar treatment for the circular block permutation by adding an extra term $S(t)$ to correct for the skewness. This term varies across $t$ to address for the different extend of the skewness across $t$.

$$P(\max_{n_0 \leq t \leq n_1} Z_{CBP}(t) > b) \approx b\phi(b) \sum_{n_0 \leq t \leq n_1} S(t)C(t)\nu\left(\sqrt{2b^2C(t)}\right), \quad (5.3)$$

where $S(t) = \exp\left(\frac{1}{2}(b-\hat{\theta}_6(t))^2 + \frac{1}{2}\gamma(t)\hat{\theta}_6^3(t)\right)/\sqrt{1+\gamma(t)\hat{\theta}_6(t)}$, with $\gamma(t) = E_{CBP}(Z^3_{CBP}(t))$ and $\hat{\theta}_6(t) = (-1 + \sqrt{1 + 2b\gamma(t)})/\gamma(t)$. 
To get an exact analytic expression for \( \mathbb{E}_{\text{CBP}}(Z_{\text{CBP}}^3(t)) \), one needs to figure out all possible configurations of 3 edges, and whether any of the six end nodes are within a block or not. Thus, even for only calculation \( t = aL, a = 1, 2, \ldots, m \), the analytic expression is very complicated and needs quite some time to run in R. It turns out that \( \gamma(t) \) can be reasonably well approximated by \( \mathbb{E}_{\text{P}}(Z_{\text{P}}^3(t)) \), which can be instantly computed in R with its exact analytic expression provided in Chen and Zhang (2015). Figure 5 plots \( \mathbb{E}_{\text{P}}(Z_{\text{P}}^3(t)) \) (red line) on top of estimated \( \mathbb{E}_{\text{CBP}}(Z_{\text{CBP}}^3(t)) \) (black dots), and we can see that \( \mathbb{E}_{\text{P}}(Z_{\text{P}}^3(t)) \) provides a good estimate to \( \mathbb{E}_{\text{CBP}}(Z_{\text{CBP}}^3(t)) \). Hence, when we apply (5.3) to approximate the \( p \)-value, we use \( \mathbb{E}_{\text{P}}(Z_{\text{P}}^3(t)) \) to estimate \( \gamma(t) \) in computing \( S(t) \).

Fig 5. Skewness \( \mathbb{E}_{\text{CBP}}(Z_{\text{CBP}}^3(t)) \) estimated from 100,000 random circular block permutations (black dots), and \( \mathbb{E}_{\text{P}}(Z_{\text{P}}^3(t)) \) computed from its exact analytic expression (red line).

\[ d = 100 \quad d = 1,000 \]

5.3. Check analytic \( p \)-value approximations

Here, we check how accurate the analytic formulas provided in 5.2 in approximating the \( p \)-values. We compare the analytic \( p \)-value approximations obtained through asymptotic results (5.2) (denoted by “A1”) and after skewness correction (5.3) (denoted by “A2”) with the \( p \)-value estimated from 100,000 random circular block permutations (denoted by “CBP”). We generate data from autoregressive and/or moving average models and consider the following three scenarios:

Scenario 1: \( d = 10 \), noises generated from the Gaussian distribution.
Scenario 2: \( d = 100 \), noises generated from \( t_5 \) distribution.
Scenario 3: \( d = 1,000 \), noises generated from the Laplace distribution.

We use function \texttt{arima.sim} to first generate \( d \) independent sequences of time series data of length \( n = 1,000: \{z_1, z_2, \ldots, z_n\} \in \mathbb{R}^d \). Then, let \( y_t = \Sigma^{1/2} z_t \), where the \((i,j)\) element of \( \Sigma \) is \( 0.6^{|i-j|} \). The methods are applied to the data sequence \( \{y_1, y_2, \ldots, y_n\} \). We consider five autoregressive and/or moving average models in generating \( z_t \)’s.
• M1: AR(1) with parameter 0.1.
• M2: AR(2) with parameters 0.1 and 0.05.
• M3: MA(1) with parameter 0.1.
• M4: MA(2) with parameters 0.1 and 0.05.
• M5: ARMA(1,1) with parameters 0.1 and 0.1.

We checked for several block sizes: \( L = 2, 5, 10, 20 \). The critical values based on “A1”, “A2”, and “CBP” are provided in Table 4. Here, results under “CBP” can be viewed as good estimates of the true critical values even though they are time-consuming to obtain. We compare results under “A1” and “A2” with those under “CBP”. We see that, under scenario 1, when the dimension is low, both analytic \( p \)-value approximations work reasonably well with those under “A2” very close the corresponding ones under “CBP”. Under scenarios 2 and 3, the dimension is higher, we see that the analytic \( p \)-value approximation only based on the asymptotic results is doing poorly, while the analytic \( p \)-value approximation after skewness correction are still close to those obtained through 100,000 circular block permutations. The same pattern goes for different models and different choices of \( L \)’s. Based on these simulation studies, we see that the analytic \( p \)-value approximation after skewness correction (5.3) provides reasonably accurate estimates to the \( p \)-value under the circular block permutation framework.
6. Discussion

In this section, we discuss the choice of $L$ through a data driven way, and the extension of the proposed framework to accommodate multiple change-points.

6.1. Choice of $L$

One practical question in applying this circular block permutation framework is the choice of $L$. Here, we provide a data driven way to choose $L$. Figure 6 plots the $Z_{G,CBP}(t)$ under different choices of $L$'s.

In the top panel, the observations are independent: $y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$, and the $(i,j)$th element of $\Sigma$ is $0.6|i-j|$. The dimension of the data is 100 and the length of the sequence is 1,000. We see that the scan statistic $Z_{G,CBP}(t)$ under $L = 1$ (permutation), $L = 2$, $L = 5$ are almost the same. On the other hand, in the middle and bottom panels, the data sequence is from a multivariate autoregression model: $y_t = \rho y_{t-1} + \varepsilon_t$, $t = 1, \ldots, n$, with $y_0 \sim \mathcal{N}(0, \frac{1}{1-\rho^2} \Sigma), \varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$. Here, $\rho = 0.1$ for the middle panel, and $\rho = 0.2$ for the bottom panel. First of all, when the observations are not independent, the scan statistic under $L = 2$ no longer overlaps with that under permutation ($L = 1$). The stronger the autocorrelation is, the larger the discrepancy between the two curves under $L = 1$ and $L = 2$. Secondly, as $L$ increases, the curves becomes more similar. For example, in the middle panel, the two curves under $L = 5$ and $L = 6$ almost overlap with each other; in the bottom panel, the curve under $L = 6$ is also close to that under $L = 5$, showing that $L$ around this range is close to enough to take care of the local dependence in the sequence. Hence, we could choose $L$ to be the value that the scan statistic no longer changes or changes in a negligible amount.

To be more specific, we plot the maximum scan statistic $\max_{0 \leq t \leq n_1} Z_{G,CBP}(t)$ over $L$ for the three data sequences and they are shown in Figure 7 (left panels). We see that the maximum scan statistic scatter around under the independent case (top panel), and the maximum scan statistic is decreasing for $L$ from 1 to 10 under the autocorrelated cases (middle panel and bottom panel). In the right panel, the ratio of the maximum scan statistic under $L + 1$ over that under $L$ is plotted. In each ratio plot, a horizontal line at 0.99 is added, which appears to be a reasonable threshold to use in choosing $L$. We could set $L$ to be the smallest value such that the ratio goes above 0.99. Under this criterion, we could set $L = 2$, $L = 6$, and $L = 8$ for the three data sequences, respectively.

6.2. Extension to multiple change-points

We focused on the single change-point alternative so far. The proposed framework could be extended to the changed interval alternative up to some modifications: For any pair of times $t_1 < t_2$, we could construct the test statistic $Z_{G,CBP}(t_1, t_2)$ to test $\{y_{t_1}, \ldots, y_{t_2-1}\}$ against $\{y_{t_2}, \ldots, y_n, y_1, \ldots, y_{t_1-1}\}$ in a
Fig 6. Plots of the scan statistic $Z_{G, CBP}(t)$ under different choices of $L$’s. Data generated from the multivariate autoregression model $y_t = \rho y_{t-1} + \epsilon_t$, $t = 1, \ldots, n$, with $y_0 \sim N(0, \frac{1}{1-\rho^2} \Sigma), \epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0, \Sigma), d = 100, n = 1,000$. The $(i,j)$th element of $\Sigma$ is $0.6^{|i-j|}$. Top panel: $\rho = 0$; middle panel: $\rho = 0.1$; bottom panel: $\rho = 0.2$. 

- Top panel: $\rho = 0$ 
- Middle panel: $\rho = 0.1$ 
- Bottom panel: $\rho = 0.2$
Fig 7. Plots of the maximum scan statistic \( \max_{n_0 \leq t \leq n_1} Z_{G,\text{CBP}}(t) \) under different choices of \( L \)'s based on the same data sets in Figure 6. Top panel: \( \rho = 0 \); middle panel: \( \rho = 0.1 \); bottom panel: \( \rho = 0.2 \). In the right panel, the ratio of \( \max_{n_0 \leq t \leq n_1} Z_{G,\text{CBP}}(t) \) under \( L + 1 \) over that under \( L \) is plotted. The horizontal line is at 0.99.
similar way as $Z_{G,\text{CBP}}(t)$, which tests \{y_1, \ldots, y_t\} against \{y_{t+1}, \ldots, y_n\}. The scan statistic for the changed-interval alternative can then be defined as

$$\max_{1 \leq t_1 < t_2 \leq n, \, l_0 \leq t_2 - t_1 \leq l_1} Z_{G,\text{CBP}}(t_1, t_2), \quad (l_0, l_1 \text{ prespecified}).$$

Other treatments could follow accordingly.

For more complicated scenarios with possibly multiple change-points, we could apply the scan statistic for the single change-point alternative recursively in a binary segmentation procedure (Vostrikova, 1981), or apply the scan statistic for the changed-interval alternative recursively in a circular binary segmentation procedure (Olshen et al., 2004).

Acknowledgments

Hao Chen is supported in part by NSF award DMS-1513653.

References

Aue, A., Hörmann, S., Horváth, L., Reimherr, M. et al. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics* **37** 4046–4087.

Bauwens, L., Laurent, S. and Rombouts, J. V. (2006). Multivariate GARCH models: a survey. *Journal of applied econometrics* **21** 79–109.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* **31** 307–327.

Carlstein, E. G., Müller, H. G. and Siegmund, D. (1994). Change-point problems. Inst of Mathematical Statistic.

Chen, L. H. Y. and Shao, Q. M. (2005). Stein’s method for normal approximation. *An introduction to Stein’s method* **4** 1–59.

Chen, H. and Zhang, N. R. (2013). Graph-based tests for two-sample comparisons of categorical data. *Statistica Sinica* **1479–1503.

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

Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis* **18**. John Wiley & Sons Inc.

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** 15274–15278.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society* 987–1007.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* **7** 697–717.
Girón, J., Ginebra, J. and Riba, A. (2005). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician* **59** 19–30.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics* **772–783.

Kirch, C. (2006). Resampling methods for the change analysis of dependent data. *PhD thesis, University of Cologne*.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science* **311** 88–90.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* 1217–1241.

Liu, R. Y. and Singh, K. (1992). Moving blocks bootstrap and jackknife capture weak dependence. Exploring the Limits of Bootstrap **225–248**.

Olshen, A. B., Venkatraman, E., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.

Politis, D. N. and Romano, J. P. (1992). A circular block-resampling procedure for stationary data. Exploring the limits of bootstrap 263–270.

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 515–530.

Siegmund, D. and Yakir, B. (2007). *The statistics of gene mapping*. Springer Science & Business Media.

Silvennoinen, A. and Teräsvirta, T. (2009). Multivariate GARCH models. Handbook of Financial Time Series TG Andersen, RA Davis, JP. Kreiss and T. Mikosch, eds.

Tsirigos, A. and Rigoutsos, I. (2005). A new computational method for the detection of horizontal gene transfer events. *Nucleic acids research* **33** 922–933.

Vostrikova, L. J. (1981). Detecting disorder in multidimensional random processes. In *Soviet Mathematics Doklady* **24** 55–59.

**Appendix A: Proofs for theorems**

**A.1. Proof of Theorem 4.3**

We have

\[
\mathbb{E}_{\text{CBP}}(R_G^2(t)) = \sum_{(i,j), (u,v) \in G} P(g_{\pi_{\text{CBP}}(i)}(t) \neq g_{\pi_{\text{CBP}}(j)}(t), g_{\pi_{\text{CBP}}(u)}(t) \neq \pi_{\pi_{\text{CBP}}(v)}(t))
\]

\[
= \sum_{(i,j) \in G} P(g_{\pi_{\text{CBP}}(i)}(t) \neq g_{\pi_{\text{CBP}}(j)}(t))
\]

\[+ \sum_{(i,j),(i,u) \in G; j \neq u} P(g_{\pi_{\text{CBP}}(i)}(t) \neq g_{\pi_{\text{CBP}}(j)}(t), g_{\pi_{\text{CBP}}(i)}(t) \neq \pi_{\pi_{\text{CBP}}(u)}(t))
\]
\[ + \sum_{(i, j, u, v) \in G} P(g_{\text{CBP}}(i)(t) \neq g_{\text{CBP}}(j)(t), g_{\text{CBP}}(u)(t) \neq \pi_{\text{CBP}}(v)(t)). \]

The first part of the summation is \( E_{\text{CBP}}(R_G(t)) \). In the following, we figure out the second and third part of the summation.

**A.1.1.** \( P(g_{\text{CBP}}(i)(t) \neq g_{\text{CBP}}(j)(t), g_{\text{CBP}}(i)(t) \neq \pi_{\text{CBP}}(u)(t)), j \neq u \)

When \( \delta_{ij}, \delta_{iu}, \delta_{ju} \geq L, i, j, u \) are all in different blocks, and the probability is \( p_2(a) \). When at least one of \( \delta_{ij}, \delta_{iu}, \delta_{ju} \geq L \) is less than \( L \), we need to consider scenarios that some of the indices could be in the same block. In the following, we consider when the event \( \{g_{\text{CBP}}(i)(t) \neq g_{\text{CBP}}(j)(t), g_{\text{CBP}}(i)(t) \neq \pi_{\text{CBP}}(u)(t)\} \) could happen.

When \( \delta_{ij} < L, \delta_{iu}, \delta_{ju} \geq L \), since \( \pi_{\text{CBP}}(i) \) and \( \pi_{\text{CBP}}(j) \) need to be in different blocks, \( i \) and \( j \) need to be in different blocks. And the probability is

\[ \frac{\delta_{iu}}{L} p_2(a). \]

Similarly, when \( \delta_{iu} < L, \delta_{ij}, \delta_{ju} \geq L \), the probability is

\[ \frac{\delta_{iu}}{L} p_2(a). \]

When \( \delta_{ju} < L, \delta_{ij}, \delta_{iu} \geq L \), since \( \pi_{\text{CBP}}(j) \) and \( \pi_{\text{CBP}}(u) \) can be in the same block, so the probability is

\[ \frac{l - \delta_{ju}}{L} p_1(a) + \frac{\delta_{iu}}{L} p_2(a). \]

When \( \delta_{ij}, \delta_{iu} < L, \delta_{ju} \geq L \), then \( \delta_{ju} = \delta_{ij} + \delta_{iu} \). \( j \) and \( u \) will always be in different blocks. We need \( i \) to be in a block different from \( j \)'s and \( u \)'s block so that the probability is positive. So the probability is

\[ \frac{\delta_{ju} - L}{L} p_2(a). \]

When \( \delta_{ij}, \delta_{ju} < L, \delta_{iu} \geq L \), then \( \delta_{iu} = \delta_{ij} + \delta_{ju} \). \( i \) and \( j \) need to be in different blocks, so the probability is

\[ \frac{\delta_{ij} - (\delta_{ju} - L)}{L} p_1(a) + \frac{\delta_{ju} - L}{L} p_2(a) = \frac{l - \delta_{ju}}{L} p_1(a) + \frac{\delta_{iu} - L}{L} p_2(a). \]

Similarly for \( \delta_{iu}, \delta_{ju} < L \).

When \( \delta_{ij}, \delta_{iu}, \delta_{ju} < L \), \( i \) needs to be in different block from \( j \) and \( u \). So the probability is

\[ \frac{\min(\delta_{ij}, \delta_{iu}, \delta_{ju})}{L} I(\max(\delta_{ij}, \delta_{iu}, \delta_{ju}) \neq \delta_{ju}) p_1(a). \]
A.1.2. \( P(g_{\pi_{\text{CBP}}(i)}(t) \neq g_{\pi_{\text{CBP}}(j)}(t), g_{\pi_{\text{CBP}}(u)}(t) \neq \pi_{\text{CBP}}(v)(t)), i, j, u, v \) all different

When \( i, j, u, v \) are all different, there are \( \binom{4}{2} = 6 \) index-pairs among them. If all pairwise index distances are greater than or equal to \( L \), this probability is \( p_3(a) \). In the following, we consider scenarios that at least one of the six index distances is less than \( L \).

1) One distance < \( L \).
   If \( \delta_{ij} < L \) and all other pairwise distances are \( \geq L \), then this probability is
   \[
   \frac{\delta_{ij}}{L} p_3(a).
   \]
   Similarly, if only \( \delta_{uv} < L \), the probability is
   \[
   \frac{\delta_{uv}}{L} p_3(a).
   \]
   If only \( \delta_{iu} < L \), the probability is
   \[
   \frac{L-\delta_{iu}}{L} p_2(a) + \frac{\delta_{iu}}{L} p_3(a).
   \]
   Similar for only \( \delta_{iv}, \delta_{jv}, \delta_{ju} < L \).

2) Two distances < \( L \).
   If only \( \delta_{ij}, \delta_{uv} < L \), then the probability is
   \[
   \frac{(\min(\delta_{ij}, b_{ij,uv}+\delta_{uv})-\delta_{ij,uv})_+ + (\min(\delta_{ij}, b_{ij,uv}+\delta_{uv})-L)_+)}{L} p_3(a).
   \]
   If only \( \delta_{ij}, \delta_{iu} < L \), we must have \( \delta_{ui} + \delta_{ij} = \delta_{uj} \), the probability is
   \[
   \frac{L-\delta_{iu}}{L} p_2(a) + \frac{\delta_{ij}+\delta_{iu}-L}{L} p_3(a).
   \]
   Similar for other 7 similar cases.
   If only \( \delta_{iu}, \delta_{iv} < L \), then we have \( \delta_{uv} = \delta_{ui} + \delta_{iv} \), the probability is
   \[
   \frac{2L-\delta_{uv}}{L} p_2(a) + \frac{\delta_{iu}-L}{L} p_3(a).
   \]
   Similar for other 3 similar cases.
   If only \( \delta_{iu}, \delta_{jv} < L \), the probability is
   \[
   \frac{l-\delta_{iu}-\delta_{jv}+x(iu,jv)}{L} p_1(a) + \frac{\delta_{iu}+\delta_{jv}-2x(iu,jv)}{L} p_2(a) + \frac{x(iu,jv)}{L} p_3(a).
   \]
   Similar for the case that only \( \delta_{iv}, \delta_{ju} < L \).

3) Three distances < \( L \).
   If only \( \delta_{ij}, \delta_{uv}, \delta_{iu} < L \), then the order of the four indices must be \( (j, i, u, v) \) or the reverse, and \( \delta_{jv} = \delta_{ji} + \delta_{iu} + \delta_{uv} \). The probability is
   \[
   \frac{L-\delta_{iu}}{L} p_2(a) + \frac{(\delta_{iv}-2L)_+}{L} p_3(a).
   \]
   Similar for other 3 similar cases: replace \( \delta_{iu} \) by one of \( (\delta_{iv}, \delta_{ju}, \delta_{jv}) \).
If only $\delta_{ij}, \delta_{iu}, \delta_{ju} < L$, then the order of the four indices must be $(u, i, j, v)$ or the reverse, and $\delta_{uv} = \delta_{ui} + \delta_{ij} + \delta_{ju}$. The probability is

$$\frac{2L - \delta_{uv} + (\delta_{uv} - 2L)\cdot p_1(a) + \delta_{uv} - 2L - 2(\delta_{uv} - 2L)\cdot p_2(a) + (\delta_{uv} - 2L)\cdot p_3(a)}{L}.$$ 

Similar for other 3 similar cases: only $\delta_{ij}, \delta_{iv}, \delta_{ju} < L$; only $\delta_{uv}, \delta_{ui}, \delta_{uj} < L$.

If only $\delta_{iu}, \delta_{iv}, \delta_{ju} < L$, then the order of the four indices must be $(j, u, i, v)$ or the reverse, and $\delta_{jv} = \delta_{ju} + \delta_{ui} + \delta_{iv}$. The probability is

$$\frac{2L - \delta_{jv} + (\delta_{jv} - 2L)\cdot p_1(a) + \delta_{jv} - 2L - 2(\delta_{jv} - 2L)\cdot p_2(a) + (\delta_{jv} - 2L)\cdot p_3(a)}{L}.$$ 

Similar for other 3 similar cases: choose 3 out of $(\delta_{iu}, \delta_{iv}, \delta_{ju}, \delta_{jv})$.

If only $\delta_{ij}, \delta_{iu}, \delta_{ju} < L$, the probability is

$$\frac{\delta_{ij}}{L} p_2(a).$$

Similar for other 3 similar cases: only $\delta_{ij}, \delta_{iv}, \delta_{ju} < L$; only $\delta_{uv}, \delta_{ui}, \delta_{iv} < L$; only $\delta_{uv}, \delta_{ju}, \delta_{jv} < L$.

It’s not possible for 4 cases with only three distances smaller than $l$ that share one index. For example, only $\delta_{ij}, \delta_{iu}, \delta_{iv} < L$.

4) Four distances $< L$.

For scenarios that 4 distances $< L$ and 2 distances $\geq L$, it is impossible for 3 cases: only $\delta_{ij}, \delta_{uv} \geq L$; only $\delta_{iu}, \delta_{iv} \geq L$; only $\delta_{iv}, \delta_{ju} \geq L$.

If only $\delta_{ij}, \delta_{iu} \geq L$, it can be $(i, j, u, v)$ or $(i, j, v, u)$ or their reverses. For both orders, the probability is

$$\frac{L - \delta_{iu}}{L} p_1(a) + \frac{\delta_{iu} - L}{L} p_2(a).$$

Similar for 7 other similar cases.

If only $\delta_{iu}, \delta_{iv} \geq L$, it can be $(i, j, u, v) (i, j, v, u)$ or their reverses. The probability is

$$\frac{\delta_{iu}}{L} p_2(a).$$

Similar for 3 other similar cases.

5) Five distances $< L$.

If only $\delta_{ij} \geq L$, the probability is

$$\frac{\delta_{iu}}{L} p_1(a).$$

Similar for if only $\delta_{uv} \geq L$.

If only $\delta_{iu} \geq L$, it can be $(i, j, v, u)$ or $(i, j, u, v)$ or their reverses. The probability is

$$\frac{\delta_{iu} - L}{L} p_2(a) + \frac{\delta_{iu}}{L} I(\delta_{ij} = \delta_{iv} + \delta_{ij}) p_1(a).$$

Similar for other 3 similar cases.

6) Six distances $< L$.

If all 6 distances are $< L$. 
If $\delta_{ij}$ is the maximum distance, the probability is
$$\frac{\delta_{uv}}{L}p_1(a).$$

Similar for the case that $\delta_{uv}$ is the maximum distance.

If $\delta_{iu}$ is the maximum distance, the probability is
$$\frac{\delta_{jv}}{L}I(\delta_{ij} = \delta_{iv} + \delta_{jv})p_1(a).$$

Similar for 3 other similar cases.

Summing all possible scenarios, we get the result stated in Lemma 4.3.

### A.2. Proof of Theorem 5.1

To prove $\{Z_G, \text{CBP}([mw]L) : \epsilon \leq w \leq 1 - \epsilon\}$ converges to a Gaussian process in finite dimensional distributions, we only need to show that

$$(Z_G, \text{CBP}([mw_1]L), Z_G, \text{CBP}([mw_2]L), \ldots, Z_G, \text{CBP}([mw_K]L))$$

converges to multivariate Gaussian as $n \to \infty$ for any $\epsilon \leq w_1 < w_2 < \cdots < w_K \leq 1 - \epsilon$ for any fixed $K$. For simplicity, let $t_k = [mw_k]L, k = 1, \ldots, K$.

To prove the above results, we take one step back. For circular block permutation with a random starting point, in the last step of the recipe, the action is to permute the $m = n/L$ blocks. Let $\pi(i)$ be the observed time of block $i$ after this block permutation, then $(\pi(1), \ldots, \pi(m))$ is a permutation of $1, \ldots, m$. We can do the last step in the following two-step approach: (1) For each $i$, $\tilde{\pi}(i)$ is sampled uniformly from $1$ to $m$; (2) only those that each value in $\{1, \ldots, m\}$ is sampled once are retained. It is easy to see that each block permutation has the same occurrence probability after these two steps.

We call the distribution resulting from only performing the first step the circular block bootstrap with a random starting point, short as CBB, and use $P_{\text{CBB}}, E_{\text{CBB}}, \text{Var}_{\text{CBB}}$ to denote the probability, expectation, and variance, respectively.

Let
$$Z_G, \text{CBB}(t) = \frac{-R_G(t) - E_{\text{CBB}}(R_G(t))}{\sqrt{\text{Var}_{\text{CBB}}(R_G(t))}},$$
$$X_{\text{CBB}}(t) = \frac{n_{\text{CBB}}(t) - (t/L)}{\sqrt{(1-t/n)t/L}}, \quad \text{where } n_{\text{CBB}}(t) = \sum_{i=1}^{m} I_{\tilde{\pi}(i) \leq t/L}.$$

Then following the similar arguments for obtaining Theorems 4.1 and 4.3, we have that, for $t = aL, a \in \{1, \ldots, m\}$,

$$E_{\text{CBB}}(R_G(t)) = c_0 p_1, \text{CBB}(a),$$
$$\text{Var}_{\text{CBB}}(R_G(t)) = c_1 p_1, \text{CBB}(a) + c_2 p_2, \text{CBB}(a) + c_3 p_3, \text{CBB}(a) - c_0^2 p_1, \text{CBB}(a) := (\sigma_{\text{CBB}}(t))^2,$$

where $p_1, \text{CBB}(a) = \frac{2a(m-a)}{m^2}$, $p_2, \text{CBB}(a) = \frac{a(m-a)}{m^2}$, $p_3, \text{CBB}(a) = \frac{4a^2(m-a)^2}{m^4}$.

We next prove the following two lemmas.
Lemma A.1. Under Condition 1, as $n \to \infty$,

$$(Z_{G,CBB}(t_1), \ldots, Z_{G,CBB}(t_K), X_{CBB}(t_1), \ldots, X_{CBB}(t_K)) \quad (A.1)$$

converges to a multivariate Gaussian distribution under CBB and the covariance matrix of $(X_{CBB}(t_1), X_{CBB}(t_2), \ldots, X_{CBB}(t_K))$ is positive definite.

Lemma A.2. Under Condition 2 or Condition 3, we have for $k = 1, \ldots, K$,

1. $\frac{\text{Var}_{CBB}(RG(t_k))}{\text{Var}_{CBB}(RG(t_k))} \to r([m w_k])$, with $r(a)$ a constant only depends on $a$.
2. $\frac{E_{CBB}(RG(t_k)) - E_{CBB}(RG(t_k))}{\sqrt{\text{Var}_{CBB}(RG(t_k))}} \to 0$.

From Lemma A.1, $(Z_{G,CBB}(t_1), Z_{G,CBB}(t_2), \ldots, Z_{G,CBB}(t_K))$ conditioning on $(X_{CBB}(t_1), X_{CBB}(t_2), \ldots, X_{CBB}(t_K))$ converges to multivariate normal under CBB. Since $(Z_{G,CBB}(t_1), Z_{G,CBB}(t_2), \ldots, Z_{G,CBB}(t_K)|X_{CBB}(t_1) = 0, X_{CBB}(t_2) = 0, \ldots, X_{CBB}(t_K) = 0)$ under CBB has the same distribution as $(Z_{G,CBB}(t_1), Z_{G,CBB}(t_2), \ldots, Z_{G,CBB}(t_K))$ under CBP, and notice that

$$Z_{G,CBP}(t) = \frac{\text{Var}_{CBB}(RG(t_k))}{\text{Var}_{CBB}(RG(t_k))} \left( Z_{G,CBB}(t) - \frac{E_{CBB}(RG(t_k)) - E_{CBB}(RG(t_k))}{\sqrt{\text{Var}_{CBB}(RG(t_k))}} \right).$$

Given Lemma A.2, we conclude that $(Z_{G,CBP}(t_1), Z_{G,CBP}(t_2), \ldots, Z_{G,CBP}(t_K))$ converges to a multivariate Gaussian distribution under CBP. Next, we prove the two lemmas.

A.2.1. Proof for Lemma A.1

To show that (A.1) converges to a multivariate Gaussian distribution, we only need to show that $\sum_{k=1}^{K}(c_{1k}Z_{G,CBB}(t_k) + c_{2k}X_{CBB}(t_k))$ converges to a normal distribution for any fixed $\{c_{1k}\}$ and $\{c_{2k}\}$. If $\text{Var}_{G,CBB}(\sum_{k=1}^{K}(c_{1k}Z_{G,CBB}(t_k) + c_{2k}X_{CBB}(t_k)) = 0$, $\sum_{k=1}^{K}(c_{1k}Z_{G,CBB}(t_k) + c_{2k}X_{CBB}(t_k)$ is degenerating. For non-degenerating case, let

$$\sigma_0^2 = \text{Var}_{G,CBB}(\sum_{k=1}^{K}(c_{1k}Z_{G,CBB}(t_k) + c_{2k}X_{CBB}(t_k)).$$

We prove the Gaussianity of $\sum_{k=1}^{K}(c_{1k}Z_{G,CBB}(t_k) + c_{2k}X_{CBB}(t_k))$ by the Stein’s method. Consider sums of the form $W = \sum_{i \in J} \xi_i$, where $J$ is an index set and $\xi$ are random variables with $E\xi_i = 0$, and $E(W^2) = 1$. The following assumption restricts the dependence between $\{\xi_i : i \in J\}$.

Assumption A.3. (Chen and Shao, 2005, p. 17) For each $i \in J$ there exists $K_i \subset L_i \subset J$ such that $\xi_i$ is independent of $\xi_{K_i}$ and $\xi_{K_i}$ is independent of $\xi_{L_i}$.

We will use the following existing theorem in proving Theorem 5.1.
Theorem A.4. (Chen and Shao, 2005, Theorem 3.4) Under Assumption A.3, we have

\[
\sup_{h \in \text{Lip}(1)} |E(h(W)) - E(h(Z))| \leq \delta
\]

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

\[
\delta = 2 \sum_{i \in \mathcal{J}} |E[\xi_i \eta_i \theta_i] + |E[\xi_i \eta_i]E[\theta_i]| + \sum_{i \in \mathcal{J}} E[|\xi_i \eta_i^2|]
\]

with \(\eta_i = \sum_{j \in K_i} \xi_j \) and \(\theta_i = \sum_{j \in L_i} \xi_j\), where \(K_i\) and \(L_i\) are defined in Assumption A.3.

We adopt the same notations with the index set \(\mathcal{J} = \{ e \in G \} \cup \{ 1, \ldots, n \}\). Denote the end nodes of an edge \(e\) to be \(e_-\) and \(e_+\). Let

\[
\xi_{e,k} = \frac{I_{g_{(e_-)}(t_k) \neq g_{(e_+)}(t_k) - p_{1,CBB}(t_k)}}{CBB(t_k)}.
\]

Since \(I_{g_{(e_-)}(t_k) \neq g_{(e_+)}(t_k) - p_{1,CBB}(t_k)} \in \{0, 1\}\) and \(p_{1,CBB}(t_k) \in [0, 1]\), we have that

\[
|\xi_{e,k}| \leq \frac{1}{\sigma_{CBB}(t_k)}.
\]

Let

\[
\xi_{i,k} = \frac{I_{g_{(e)}(t_k) \leq t_k/I - w_k}}{\sqrt{mw_k(1-w_k)}}
\]

Similarly, we have

\[
|\xi_{i,k}| \leq \frac{1}{\sqrt{mw_k(1-w_k)}}
\]

Let \(\xi_e = \sum_{k=1}^K c_k \xi_{e,k}/\sigma_0\), \(\xi_i = \sum_{k=1}^K c_{2k} \xi_{i,k}/\sigma_0\), and \(W = \sum_{j \in \mathcal{J}} \xi_j = \sum_{k=1}^K (c_{1k} Z_{G,CBB}(t_k) + c_{2k} X_{CBB}(t_k))/\sigma_0\). Then \(E_{CBB}(W) = 0, E_{CBB}(W^2) = 1\).

Let \(c_M = \max(\sum_{k=1}^K |c_{1k}|, \sum_{k=1}^K |c_{2k}|)\), \(\sigma_1 = \sigma_0 \times \min_k \sigma_{CBB}(t_k)\), \(\sigma_2 = \sigma_0 \times \min_k \sqrt{mw_k(1-w_k)}\), then

\[
|\xi_e| \leq \frac{c_M}{\sigma_1}, \forall e \in G; \quad |\xi_i| \leq \frac{c_M}{\sigma_2}, \forall i \in \{ 1, \ldots, n \}.
\]

Based on Remark 4.4, \(c_1 + c_2 + c_3 \geq c_0^2\), we have, with \(a_k = [mw_k]\),

\[
\text{Var}_{CBB}(R_G(t_k)) \geq c_1 (p_{1,CBB}(a_k) - p_{1,CBB}^2(a_k)) + c_2 (p_{2,CBB}(a_k) - p_{2,CBB}^2(a_k))
\]

\[
= c_1 a_k (m-a_k/m(m-1)) + (c_1 + c_2) a_k^2 (m-a_k/m(m-1)) \left(1 - \frac{4a_k(m-a_k)}{m(m-1)}\right)
\]

\[
\geq c_1 a_k (m-a_k/m(m-1)) - (c_1 + c_2) a_k^2 (m-a_k/m(m-1))\tau.
\]

It can be shown that \(c_2 = o(mc_1)\) (detailed arguments seen in the proof for Lemma A.2). So \(\text{Var}_{CBB}(R_G(t_k))\) is at least of order \(c_1\), which is at least of order \(|G|\). It is clear that \(\sigma_2 = O(n)\).

For any \(A\) a set of edge(s), let \(V(A)\) to be the set that contains all nodes being connected by at least one edge in \(A\). Notice that \(|V(A)| \leq 2|A|\) as the worst scenario occurs when all edge(s) in \(A\) are disconnected.
With \( A_{e,L,1} \) and \( A_{e,L,2} \) defined in Section 5.1, for \( e \in G \), let
\[
S_e = A_{e,L,1} \cup V(A_{e,L,1}),
\]
\[
T_e = A_{e,L,2} \cup V(A_{e,L,2}),
\]
Then \( S_e \) and \( T_e \) satisfy Assumption A.3.

With \( A_{i,L,1} \) and \( A_{i,L,2} \) defined in Section 5.1, for \( i = 1, \ldots, n \), let
\[
S_i = A_{i,L,1} \cup V(A_{i,L,1}),
\]
\[
T_i = A_{i,L,2} \cup V(A_{i,L,2}),
\]
Then \( S_i \) and \( T_i \) satisfy Assumption A.3.

By Theorem A.4, we have \( \sum_{h \in L \cap (1)} |E(h(W) - E(h(Z))| \leq \delta \) for \( Z \sim N(0,1) \), where
\[
\delta = 2 \sum_{i \in J} (E[\xi_i \eta_i] + |E(\xi_i \eta_i)|E[\theta_i]) + \sum_{i \in J} E[\xi_i \eta_i^2]
\]
\[
= 2 \sum_{e \in G} (E[\xi_e \eta_e] + |E(\xi_e \eta_e)|E[\theta_e]) + \sum_{e \in G} E[\xi_e \eta_e^2]
\]
\[
+ 2 \sum_{i=1}^n (E[\xi_i \eta_i \theta_i] + |E(\xi_i \eta_i)|E[\theta_i]) + \sum_{i=1}^n E[\xi_i \eta_i^2]
\]
\[
\leq \frac{5CM}{\sigma_1} (\frac{\|A_{e,L,1}\|_{CM}}{\sigma_1} + |V(A_{e,L,1})|\frac{CM}{\sigma_2}) \left( \frac{\|A_{e,L,2}\|_{CM}}{\sigma_1} + |V(A_{e,L,2})|\frac{CM}{\sigma_2} \right)
\]
\[
+ \frac{5CM}{\sigma_2} \left( \frac{\|A_{i,L,1}\|_{CM}}{\sigma_1} + |V(A_{i,L,1})|\frac{CM}{\sigma_1} \right) \left( \frac{\|A_{i,L,2}\|_{CM}}{\sigma_1} + |V(A_{i,L,2})|\frac{CM}{\sigma_1} \right)
\]
\[
\leq \frac{5CM}{\sigma_1} \left( \frac{CM}{\sigma_1} + \frac{2CM}{\sigma_2} \right)^2 \|A_{e,L,1}\|_{CM} \|A_{e,L,2}\|_{CM} + \frac{5CM}{\sigma_2} \left( \frac{CM}{\sigma_1} + \frac{2CM}{\sigma_2} \right)^2 \|A_{i,L,1}\|_{CM} \|A_{i,L,2}\|_{CM}
\]

When \( \sum_{e \in G} \|A_{e,L,1}\|_{CM} \|A_{e,L,2}\|_{CM} = o(n|G|^{0.5}) \) and \( \sum_{i=1}^n \|A_{i,L,1}\|_{CM} \|A_{i,L,2}\|_{CM} = o(n^{1.5}) \), we have \( \delta \to 0 \) as \( n \to \infty \).

Let \( \Sigma_X \) be the covariance matrix of \( (X_{CBB}(t_1), X_{CBB}(t_2), \ldots, X_{CBB}(t_K)) \).

Follow similar arguments in Chen and Zhang (2015), \( |\Sigma_X| = \prod_{k=1}^K (1 - t_k/t_{k+1}) \prod_{k=1}^K (1 - t_k/n) \).

So \( \Sigma_X \) is positive definite.

A.2. Proof for Lemma A.2

For \( t = aL, a = |mw_k|, k = 1, \ldots, K \), we have
\[
\text{Var}_{CBB}(R_G(t)) = c_1 p_1(a) + c_2 p_2(a) + c_3 p_3(a) - c_0 p_1(a)^2
\]
\[
= c_1 \frac{4a^2(m-a)^2}{m^2(m-1)^2} + (2c_1 + c_2) \frac{-4a^2(m-a)^2 + m(m-1)a(m-a)}{m^2(m-1)^2}
\]
\[
+ (c_1 + c_2 + c_3 - c_0) \frac{4a^2(m-a)^2}{m^2(m-1)^2},
\]
\[
\text{Var}_{CBP}(R_G(t)) = c_1 p_1(a) + c_2 p_2(a) + c_3 p_3(a) - c_0 p_1(a)^2
\]
\[
= c_1 \frac{a(m-a)(m-a-1)}{m(m-1)(m-2)(m-3)}
\]
From Remark 4.4, we know that $c_1 + c_2 + c_3 - c_0^2 \geq 0$. Using the same notations in Remark 4.4, under a certain blocking $\omega$, let $|G_{b_i}(\omega)|$ be the number of edges in $G$ that connect a node in block $i$ to another node not in block $i$. Then $c_0(\omega) = \frac{1}{2} \sum_{i=1}^{m} |G_{b_i}(\omega)|$, and $c_2(\omega) + 2c_1(\omega) = \sum_{i=1}^{m} |G_{b_i}(\omega)|^2$. By Cauchy-Schwarz inequality, we have that

$$c_2(\omega) + 2c_1(\omega) \geq \left(\sum_{i=1}^{m} |G_{b_i}(\omega)|^2\right) \frac{c_0^2(\omega)}{m} = \frac{4c_0^2(\omega)}{m}.$$ 

Then,

$$2c_1 + c_2 = \frac{1}{t} \sum_{\omega \in \Omega} 2c_1(\omega) + c_2(\omega) \geq \frac{1}{t} \sum_{\omega \in \Omega} \frac{4c_0^2(\omega)}{m} \geq \frac{4c_0^2}{m}.$$

So $2c_1 + c_2 - \frac{4c_0^2}{m} \geq 0$.

Also, for any node $j$ in block $i$, it is clear that $|G_{b_j}(\omega)| \leq |A_{j,L,1}|$, so $2c_1(\omega) + c_2(\omega) \leq \frac{1}{t^2} \sum_{j=1}^{n} |A_{j,n,1}|^2$. Then $2c_1 + c_2 \leq \frac{1}{t^2} \sum_{j=1}^{n} |A_{j,n,1}|^2 \leq \frac{1}{t^2} \sum_{j=1}^{n} |A_{j,n,1}| |A_{j,n,2}| = o(n^{1.5})$.

Notice that $a(m-a) = O(m^2)$ for $a = \lfloor mw_k \rfloor, k = 1, \ldots, K$, and $0 < a(m-a) \leq m^2$,

$$-4a^2(m-a)^2 + m(m-1)a(m-a) \in \left[-\frac{m^3}{4}, \frac{m^2(m-1)^2}{16}\right].$$

If $a$ is in a range such that $-4a^2(m-a)^2 + m(m-1)a(m-a) = O(m^3)$, notice that $c_0 = O(|G|)$, $2c_1 + c_2 = o(n^{1.5}), \frac{4c_0^2}{m} = o(mc_0)$, and $c_1 \geq c_0$, so term $(2c_1 + c_2 - \frac{4c_0^2}{m} + m(m-1)a(m-a))$ is dominated by term $c_1 \frac{4c_0^2(m-a)^2}{m(m-1)^2}$, and term $(2c_1 + c_2 - \frac{4c_0^2}{m} + m(m-1)a(m-a)) - \frac{4a^2(m-a)^2 + m(m-1)a(m-a)}{m(m-1)(m-2)(m-3)}$ is dominated by term $c_1 \frac{a(m-a)(m-a-1)(m-a-2)}{m(m-1)(m-2)(m-3)}$

$$\lim_{m \to \infty} \frac{\text{Var}_{\text{CBP}}(R_0(t))}{\text{Var}_{\text{CBP}}(R_G(t))} = 1.$$

If $a$ is in a range such that $-4a^2(m-a)^2 + m(m-1)a(m-a)$ is of order higher than $m^3$, then $-4a^2(m-a)^2 + m(m-1)a(m-a)$ must be positive. Under Condition 2, we have $2c_1 + c_2 - \frac{4c_0^2}{m} = O(2c_1 + c_2)$, then $\lim_{m \to \infty} \frac{\text{Var}_{\text{CBP}}(R_0(t))}{\text{Var}_{\text{CBP}}(R_G(t))}$ is a positive constant. Under Condition 3, we have $2c_1 + c_2 = O(c_1)$, then $\lim_{m \to \infty} \frac{\text{Var}_{\text{CBP}}(R_0(t))}{\text{Var}_{\text{CBP}}(R_G(t))}$ is also a positive constant.

For $t = aL, a = \lfloor mw_k \rfloor, k = 1, \ldots, K$, notice that $E_{\text{CBP}}(R_0(t)) - E_{\text{CBP}}(R_G(t)) = -c_0 \frac{2a(m-a)}{m(m-1)}$. From the above arguments, we know that $\text{Var}_{\text{CBP}}(R_G(t))$ is at least of order $|G|$, and $c_0 = O(|G|)$, then $\frac{\text{E_{CBP}}(R_0(t)) - E_{\text{CBP}}(R_G(t))}{\sqrt{\text{Var}_{\text{CBP}}(R_G(t))}}$ is at most of order $|G|^{1/2}m^{-1} = m^{0.5a-1}$, which converges to 0 as $m \to \infty$ since $a < \frac{n}{7}$. 

$$Hao Chen/Change-point detection for data with local dependency 32$$
A.3. Proof for Theorem 5.3

We need to figure out \( E_{\text{CBP}}(R_G(t_1)R_G(t_2)) \) for \( t_1 = a_1L < t_2 = a_2L \) with \( a_1, a_2 \in \{1, \ldots, m\} \). We have

\[
E_{\text{CBP}}(R_G(t_1)R_G(t_2)) = \sum_{(i,j),(u,v) \in G} P(g_{\pi_{\text{CBP}}(i)}(t_1) \neq g_{\pi_{\text{CBP}}(j)}(t_1), g_{\pi_{\text{CBP}}(u)}(t_2) \neq \pi_{\pi_{\text{CBP}}(v)}(t_2))
\]

\[
= \sum_{(i,j) \in G} P(g_{\pi_{\text{CBP}}(i)}(t_1) \neq g_{\pi_{\text{CBP}}(j)}(t_1), g_{\pi_{\text{CBP}}(i)}(t_2) \neq g_{\pi_{\text{CBP}}(j)}(t_2))
+ \sum_{(i,j),(i,u) \in G; j \neq u} P(g_{\pi_{\text{CBP}}(i)}(t_1) \neq g_{\pi_{\text{CBP}}(j)}(t_1), g_{\pi_{\text{CBP}}(i)}(t_2) \neq \pi_{\pi_{\text{CBP}}(u)}(t_2))
+ \sum_{(i,j),(u,v) \in G; i, j, u, v \text{ all different}} P(g_{\pi_{\text{CBP}}(i)}(t_1) \neq g_{\pi_{\text{CBP}}(j)}(t_1), g_{\pi_{\text{CBP}}(u)}(t_2) \neq \pi_{\pi_{\text{CBP}}(v)}(t_2)).
\]

Notice that, when \( i, j, u, v \) are all in different blocks, we have

\[
P(g_{\pi_{\text{CBP}}(i)}(t_1) \neq g_{\pi_{\text{CBP}}(j)}(t_1), g_{\pi_{\text{CBP}}(i)}(t_2) \neq g_{\pi_{\text{CBP}}(j)}(t_2)) = \frac{2a_1(m-a_2)}{m(m-1)} = q_1(a_1, a_2),
\]

\[
P(g_{\pi_{\text{CBP}}(i)}(t_1) \neq g_{\pi_{\text{CBP}}(j)}(t_1), g_{\pi_{\text{CBP}}(i)}(t_2) \neq \pi_{\pi_{\text{CBP}}(u)}(t_2)) = \frac{a_1(m-a_2)(m-2a_1+2a_2-2)}{m(m-1)(m-2)} = q_2(a_1, a_2),
\]

\[
P(g_{\pi_{\text{CBP}}(i)}(t_1) \neq g_{\pi_{\text{CBP}}(j)}(t_1), g_{\pi_{\text{CBP}}(u)}(t_2) \neq \pi_{\pi_{\text{CBP}}(v)}(t_2)) = \frac{4a_1(m-a_2)(a_1-1)(m-a_1-1)+a_2(a_1-1)(m-a_1-2)}{m(m-1)(m-2)(m-3)} = q_3(a_1, a_2).
\]

When some or all \( i, j, u, v \) are in the same block, we could follow exactly the same procedure in the proof for Theorem 4.3 while replacing \( p_1(a), p_2(a), p_3(a) \) by \( q_1(a_1, a_2), q_2(a_1, a_2), q_3(a_1, a_2) \), respectively. Hence, the result follows.