ISSN 1749-3889 (print), 1749-3897 (online)
International Journal of Nonlinear Science
Vol.27(2019) No.2,pp.67-80

Change Point Detection Based on Directed K-Nearest Neighbour Graph

Yuze Zhou *
Graduate School of Art and Science, Columbia University, New York, 10027, USA
(Received 9 October 2018, accepted 25 January 2019)

Abstract: In this paper, we propose a new graph-based method for change point detection. We first construct an alternative directed K-nearest neighbour graph and propose a new χ^2 like statistics with hypothesis testing methods for change point detection, and then give a numerical approximation method to calculate the p-value. As an application, we conduct change point detection for the character-event matrix of Chinese novel called *A Dream in Red Mansions*.

Keywords: Change Point; Detection; K-Nearest; Neighbour Graph

1 Introduction

Change-point models are usually used in various research fields that detect the lack of homogeneity in a sequence of observations [1–3]. In a typical situation, the observations $\{y_i, i = 1, 2, \cdots, n\}$ are assumed to have the same distribution F_0 if no change points have ever occurred, while in the other case, the observations will shift to another distribution F_1 for i after τ. Statistics for detecting change points based on undirected K-nearest neighbor graph [4–6] and a χ^2 like statistics [7–9] have both been proposed, however, it loses power when encountered by peculiar cases like variance shift. In summary, the change-point problems can be represented in the following statistical formulation [10–12]: A sequence of observations $\{y_i, i = 1, 2, \cdots, n\}$ indexes are marked by some meaningful orderings $\{1, 2, \cdots n\}$, such as time and rankings. We concern with testing the null hypothesis

$$H_0 : \quad y_i \sim F_0,$$

against the alternative hypothesis

$$H_\alpha : \quad \begin{array}{ll} y_i \sim F_0, & i \leq \tau, \\ y_i \sim F_1, & i > \tau, \end{array}$$

where F_0 and F_1 are two different distributions.

Change-point detection problems are mostly studied under the assumption that y_i comes from the same distribution sampled independently and identically, which can be easily violated in many cases. However, the *i.i.d* assumption is very important as it makes theoretical study possible.

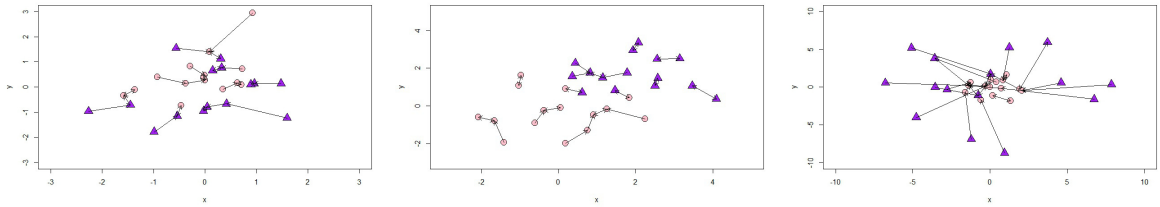
In general, the types of distribution of y_i are not strictly required. However, some distance structures are required so that the observation y_i can be represented in the graph, where the edges in the graph connect observations that are close to each other.

KNNG (k-nearest neighbor graph) is an algorithm widely used in statistical estimation and pattern recognition, the idea is simple [13–15]. For a set $\hat{P}$ consisting of n observations from a metric space (usually a Euclidean space), a KNNG graph is a directed graph with $\hat{P}$ being its vertex and with a directed edge connecting $\hat{p}$ to $\hat{q}$ whenever $\hat{q}$ is among the k vertex which is close to $\hat{p}$. In most literatures of machine learning, KNNG is often represented in a undirected version. An edge connecting $\hat{p}$ with $\hat{q}$ exists, as long as $\hat{p}$ is among the nearest neighbors of $\hat{q}$ or $\hat{q}$ is among the nearest neighbors $\hat{p}$. However, in the special case of change-point problems, a directed KNNG is much more powerful than a undirected KNNG as it maintains more information and can handle more cases.

E-mail address: yuzezhou123@163.com

IJNS.2019.04.15/1034

Here we specially choose three different cases to demonstrate how a directed KNNG can represent the dissimilarities among different observations as show in Fig. 1. They are the general case with no change points, the mean shift case and the variance change case, respectively. For each case, we have 30 observations in all, where the pink dots represent observations sampled before $\tau = 15$, and the purple triangles represent observations sampled after $\tau = 15$. Moreover, some statements for these three cases are as follows.



(a) Case 1-No Change Point (b) Case 2-Mean Shift (c) Case 3-Variance Change

Figure 1: Dissimilarities shown by a directed KNNG among different observations.

Case 1 *All the dots and triangles are sampled from the distribution $N(0, I_2)$, which means no change point has ever occurred. All observations and edges are scattered equally and evenly, see Fig. 1(a).*

Case 2 *The pink dots are samples from the distribution $N(0, I_2)$ and the purple triangles are samples from the distribution $N((2, 2), I_2)$, where the two distributions have the same variance but different means. Compared to the Case 1, the number of edges connecting pink dots to purple triangles, as well those connecting purple triangles to pink dots, both decrease in this case significantly, see Fig. 1(b).*

Case 3 *The pink dots are sample from the distribution $N(0, I_{30})$ and the purple triangles are sampled from the distribution $N(0, 4I_{30})$, where the two distributions have the same mean but different variances. The dimensions of the observations are reduced by MDS (Multi-dimensional Scaling) method so that they can be plotted on a two-dimensional graph. Compared to the Case 1, the number of edges connecting pink dots to purple triangles decrease significantly. However, the number of edges connecting purple triangles to pink dots increase, see Fig. 1(c).*

In the Case 2, a directed KNNG does not have any advantages over a undirected KNNG, as the number of edges connecting the pink dots and the purple triangles also decrease in the mean shift case. However, in the Case 3, a directed KNNG outperforms the undirected graph. The reason is that, even though the number of edges connecting pink dots to purple triangles decrease and the number of those connecting purple triangles to pink dots increase, the sum of both of them, which is the number of edges connecting the pink dots and the purple triangles in a undirected KNNG, do not change much.

This paper is organized as follows. In Section 2, a new test statistics is given. In Section 3, a hypothesis testing through asymptotic permutation null distribution is given. In Section 4, a numerical analysis is given. In Section 5, a application to real life data is given. At last, a conclusion is given.

2 The new test statistics

2.1 The basic statistics and properties

First of all, let us derive the basic statistics which are required for constructing the new statistic [3]. For each different t, all the observations can be divided into two different groups, observations sampled before t and those sampled after t. Let G be a KNNG, including both the vertexes and edges. We use i or j to represent a vertex in the graph G, and also use (i, j) to represent an edge in G that connecting i to j, if existed. For any event x, let I_x be the indicator function that takes value 1 if x happens and takes value 0 if x does not occur. Then we can derive the two basic statistics $G_{12}(t)$ and $G_{21}(t)$:

$$G_{12}(t) = \sum_{(i,j)\in G} I_{g_i(t)}(1 - I_{g_j(t)}), \tag{1}$$

and

$$G_{21}(t) = \sum_{(i,j)\in G} (1 - I_{g_i(t)}) I_{g_j(t)}, \tag{2}$$

where $I_{g_i(t)} = I_{i \le t}$.

In other sense, $G_{12}(t)$ counts the number of the edges that connects observations sampled before t and at t to those sampled after t. Contrarily, $G_{21}(t)$ counts the number of the edges that connects observations that sample after t to those sampled before t and at t.

Under the null hypothesis H_0 and the *i.i.d* assumption, the joint distribution of $\{y_i, i = 1, 2, \cdots, n\}$ is the same under the permutation distribution. We define the null distribution of $G_{12}(t)$ and $G_{21}(t)$ to be the permutation distribution, which places $1/n!$ probability on each of the $n!$ permutations of $\{y_i, i = 1, 2, \cdots, n\}$. Let $\pi(i)$ be the time of observing $y_i^{'}$ after permutation, then for the permuted sequence, $g_j(t)$ becomes $I_{\pi(i)\le t}$. Note that the graph G is determined by the values of $y_i^{'}$, instead of their order of appearance, so it remains constant under permutation. We denote P, E, Var, and cov by the probability, expectation, variance, and covariance respectively, under the permutation null distribution. The following lemmas holds naturally, and the proofs omit.

Lemma 1 *Under the permutation null distribution, the expectations of* $G_{12}(t)$ *and* $G_{21}(t)$ *are* $E(G_{12}(t)) = K\frac{t(n-t)}{n-1}$ *and* $E(G_{21}(t)) = K\frac{t(n-t)}{n-1}$ *,respectively.*

Lemma 2 *Under the permutation null distribution, the variances and covariances of* $G_{12}(t)$ *and* $G_{21}(t)$ *are*

$$\begin{aligned} Var(G_{12}(t)) &= Var(\sum_{i=1}^{t}\sum_{j=t+1}^{n} A_{ij}^{+}) \\ &= t(n-t)Var(A_{ij}^{+}) \\ &+ 2tC_{n-t}^{2} cov(A_{ij}^{+}, A_{il}^{+}) \\ &+ 2(n-t)C_{t}^{2} cov(A_{ij}^{+}, A_{lj}^{+}) \\ &+ 4C_{t}^{2} * C_{n-t}^{2} cov(A_{ij}^{+}, A_{kl}^{+}), \end{aligned}$$

$$\begin{aligned} Var(G_{21}(t)) &= Var(\sum_{i=t+1}^{n}\sum_{j=1}^{t} A_{ij}^{+}) \\ &= t(n-t)Var(A_{ij}^{+}) \\ &+ 2(n-t)C_{t}^{2} cov(A_{ij}^{+}, A_{il}^{+}) \\ &+ 2tC_{n-t}^{2} cov(A_{ij}^{+}, A_{lj}^{+}) \\ &+ 4C_{t}^{2}C_{n-t}^{2} cov(A_{ij}^{+}, A_{kl}^{+}), \end{aligned}$$

and

$$\begin{aligned} cov(G_{12}(t), G_{21}(t)) &= cov(\sum_{i=1}^{t}\sum_{j=t+1}^{n} A_{ij}^{+}, \sum_{k=t+1}^{n}\sum_{l=1}^{t} A_{kl}^{+}) \\ &= \sum_{i=1}^{t}\sum_{j=t+1}^{n}\sum_{k=t+1}^{n}\sum_{l=1}^{t} cov(A_{ij}^{+}, A_{kl}^{+}) \\ &= t(n-t)cov(A_{ij}^{+}, A_{ji}^{+}) \\ &+ 2tC_{n-t}^{2} cov(A_{ij}^{+}, A_{ki}^{+}) \\ &+ 2(n-t)C_{t}^{2} cov(A_{ki}^{+}, A_{ij}^{+}) \\ &+ 4C_{t}^{2}C_{n-t}^{2} cov(A_{ij}^{+}, A_{kl}^{+}), \end{aligned}$$

where $A_{ij}^{+} = I$ (y_j *is among the k nearest neighbours of* y_i).

Lemma 3 *Under the permutation null distribution, the variances and covariances of different pairs of A_{ij}^+ are*

$$\begin{aligned}
Var(A_{ij}^+) &= \frac{K(n-1-K)}{(n-1)^2},\\
cov(A_{ij}^+, A_{ji}^+) &= \frac{\sum_{i,j=1}^n a_{ij}^+, a_{ji}^+}{n*(n-1)} - (\frac{K}{n-1})^2,\\
cov(A_{ij}^+, A_{il}^+) &= -\frac{K(n-1-K)}{(n-1)^2(n-2)},\\
cov(A_{ij}^+, A_{ki}^+) &= \frac{nK^2 - \sum_{i,j=1}^n a_{ij}^+, a_{ji}^+}{n(n-1)(n-2)} - (\frac{K}{n-1})^2,\\
cov(A_{ij}^+, A_{lj}^+) &= \frac{\sum_{i,j,l=1}^n a_{ij}^+ * a_{lj}^+ - Kn}{n(n-1)(n-2)} - (\frac{K}{n-1})^2,
\end{aligned}$$

and

$$cov(A_{ij}^+, A_{lk}^+) = \frac{K(nK-K+1)}{(n-1)(n-2)(n-3)} - \frac{2nK^2 + \sum_{i,j,l=1}^n a_{ij}^+ a_{lj}^+ - \sum_{i,j=1}^n a_{ij}^+ a_{ji}^+}{n(n-1)(n-2)(n-3)} - (\frac{K}{n-1})^2,$$

respectively. In the equations above, $\sum_{i,j=1}^n a_{ij}^+ a_{ji}^+$ is the number of pairs of edges which connect the same nodes but in different directions and $\sum_{i,j,l=1}^n a_{ij}^+ a_{lj}^+$ is the number of pairs of edges which share the same ending nodes, including those also sharing the same starts.

2.2 A χ^2 like test statistic

2.2.1 Transformation of the basic statistic

In order to derivie a χ^2 like statistic, firstly we have to transform the basic statistics $G_{12}(t)$ and $G_{21}(t)$ so that they become uncorrelated, hence it is possible to obtain the statistics from the sum of squares of them. The transformation is mainly based on the Cholesky decomposition of their covariance matrix. To start the transformation, let us make several important symbols clearly.

$$r_0 = \frac{\sum_{i,j=1}^n a_{ij}^+ a_{ji}^+}{n},\ r_1 = \frac{\sum_{i,j,l=1}^n a_{ij}^+ a_{lj}^+}{n},\ p = \frac{t}{n}.$$

Then we can transform the two original statistics into two new statistics $\hat{G_{12}}(t)$ and $\hat{G_{21}}(t)$ are as follows:

$$\begin{pmatrix} \hat{G_{12}}(t) \\ \hat{G_{21}}(t) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{(r_1-K^2)p(1-p)}} & 0 \\ 0 & \frac{1}{\sqrt{(r_0+K)p^2(1-p)^2}} \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1-p & p \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{n}}(G_{12}(t) - E(G_{12}(t))) \\ \frac{1}{\sqrt{n}}(G_{21}(t) - E(G_{21}(t))) \end{pmatrix}, \quad (3)$$

where $G_{12}(t)$ and $G_{21}(t)$ refer to Eq.(1) and Eq.(2), respectively.

The transformation of the two original statistics $G_{12}(t)$ and $G_{21}(t)$ seem to be a little awkward and unnatural at the first sight, and the two new statistics seems to contain little information related to the graph. However, $\hat{G_{12}}(t)$ and $\hat{G_{21}}(t)$ display very good asymptotic qualities as we will show later.

2.2.2 Derivation of the χ^2 like test statistic

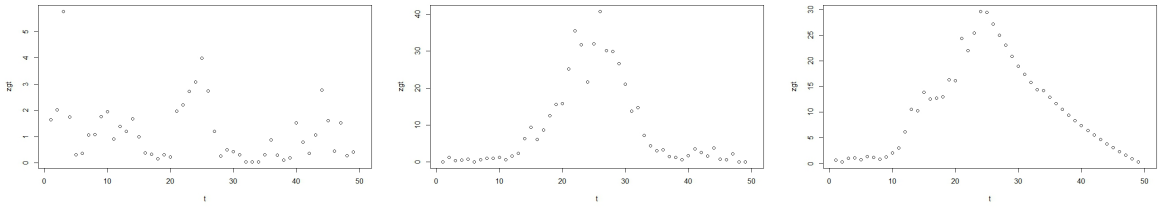
Now firstly let us define a new statistic for constructing the test statistic:

Definition 1 *Let*

$$Z_G(t) := \hat{G_{12}}^2(t) + \hat{G_{21}}^2(t),$$

where $\hat{G_{12}}^2(t)$ and $\hat{G_{21}}^2(t)$ refer to Eq.(3)*.*

Theoretically, if a change point ever occurs, at least one of the two statistics $G_{12}(t)$ and $G_{21}(t)$ will be deviated from its permutation expectation, and $\hat{G_{12}}^2(t)$ and $\hat{G_{21}}^2(t)$ can be deviated from their expectation, thus the value of $Z_G(t)$ will increase. This result is further confirmed through simulations, see Fig. 2.



(a) Case 1-No Change Point (b) Case 2-Mean Shift (c) Case 3-Variance Change

Figure 2: Observations for $Z_G(t)$.

In the Fig. 2(a), all observations are sampled from the distribution $N(0, I_2)$, which means no change point has ever occurred. $Z_G(t)$ remains comparably low less than 5 all the time. In the Fig. 2(b), the first 25 observations are sampled from the distribution $N(0, I_2)$ and the next 25 observations are sampled from the distribution $N((2, 2), I_2)$, where the two distributions have the same variance but different means. Compared to Case 1, $Z_G(t)$ increases significantly more than 40 at around the 25th observation. In the Fig. 2(c), the first 25 observations are sampled from the distribution $N(0, I_{30})$ and the next 25 are sampled from the distribution $N(0, 4I_{30})$, where the two distributions have the same mean but different variances. Compared to Case 1, $Z_G(t)$ increases significantly more than 30 at around the 25th observation.

From the three examples above, we can find out that in most cases, if a change point occurs at t, the value of $Z_G(t)$ will greatly increase at around t, which motivates us to derive the scan statistic to test H_0 versus H_α: $\max\limits_{n_0 \le t \le n_1} Z_G(t)$, where n_0 and n_1 are some pre-specified constraints for the range of t. The null hypothesis is rejected if the maximal value of $Z_G(t)$ is greater than some thresholds.

3 Hypothesis testing through asymptotic of permutation null distribution

Since the value for the statistics $Z_G(t)$ becomes much larger around the change point, we come up with the idea of rejecting the null hypothesis H_0 when $Z_G(t)$ exceeds certain thresholds. We concern with the tail probability of $\max\limits_{n_0 \le t \le n_1} Z_G(t)$ under the null hypothesis, that is,

$$P(\max_{n_0 \le t \le n_1} Z_G(t) > b).$$

The null distributions of $\max\limits_{n_0 \le t \le n_1} Z_G(t)$ are defined as the permutation null distribution. For small n, we can sample directly from the permutation distribution to approximate the test statistic. However, permutation is computationally prohibitive when n becomes larger. Therefore, we have to derive analytic expressions for the tail probabilities to make the method applicable. Moreover, the distribution of $Z_G(t)$ is quite complicated, thus it is impossible to derive the exact expression. So in the rest of this section, we will give the analytic approximations for tail probability.

3.1 Asymptotic permutation null distribution

Here we will derive the limiting distribution of $G_{12}([nu])$ and $G_{21}([nu])$ as well as $\hat{G_{12}}([nu])$ and $\hat{G_{21}}([nu])$, where $0 < u < 1$. Before the discussion of the limiting distribution, we have to introduce some notations for any edge e in the graph firstly, let

$$\begin{aligned} K_e &= \left\{ e^{'} : \ e^{'} \ shares \ a \ node \ with \ e \right\}, \\ N_e &= \{ nodes \ in \ K_e \}, \\ L_e &= \left\{ e^{''} : \ e^{''} \ has \ a \ node \ in \ N_e \right\}, \\ |K_e| &= the \ number \ edges \ in \ K_e, \end{aligned}$$

and

$$|L_e| = the \ number \ edges \ in \ L_e,$$

respectively. In order to define the asymptotic null distribution, an assumption is imposed on the obtained KNNG.

Condition 1 $\sum_{e\in G}|K_e||L_e| \sim o(n^\alpha), 1<\alpha<1.5.$

In a specific graph, $\sum_{e\in G}|K_e||L_e|$ indicates how large the 'hub' of edges we can get, and α indicates the order of the 'hub'. In most cases, 1.5 is a comparably big upper-bound that can be easily satisfied. If the Condition 1 is satisfied, we have the following results.

Theorem 1 *Assume the Condition 1 holds, both*

$$\frac{1}{\sqrt{n}}(G_{12}([nu]) - E(G_{12}([nu])))$$

and

$$\frac{1}{\sqrt{n}}(G_{21}([nu]) - E(G_{12}([nu])))$$

converge in distribution to Gaussian processes.

Proof. Under the bootstrap null distribution, we denote $E_B(G_{12}(t))$ and $E_B(G_{21}(t))$ by the expectations of $G_{12}(t)$ and $G_{21}(t)$, and denote $Var_B(G_{12}(t))$ and $Var_B(G_{21}(t))$ by the variances of $G_{12}(t)$ and $G_{21}(t)$, respectively.

$$E_B(G_{12}(t)) = E_B(G_{21}(t)) = K\frac{t(n-t)}{n},$$

$$\begin{aligned} Var_B(G_{12}(t)) &= nK\frac{t(n-t)(n^2-nt+t^2)}{n^4} \\ &\quad - (2nK^2 - \sum_{i,j=1}^{n} a_{ij}^+ a_{ji}^+)\frac{t^2(n-t)^2}{n^4} \\ &\quad + nK(K-1)\frac{t^3(n-t)}{n^4} \\ &\quad + (\sum_{i,j,l=1}^{n} a_{ij}^+ a_{lj}^+ - nK)\frac{t(n-t)^3}{n^4}, \end{aligned}$$

and

$$\begin{aligned} Var_B(G_{21}(t)) &= nK\frac{t(n-t)(n^2-nt+t^2)}{n^4} \\ &\quad - (2nK^2 - \sum_{i,j=1}^{n} a_{ij}^+ a_{ji}^+)\frac{t^2(n-t)^2}{n^4} \\ &\quad + nK(K-1)\frac{t*(n-t)^3}{n^4} \\ &\quad + (\sum_{i,j,l=1}^{n} a_{ij}^+ a_{lj}^+ - nK)\frac{t^3(n-t)}{n^4}. \end{aligned}$$

Let

$$\begin{aligned} W_1 &:= \frac{G_{12}(t) - E(G_{12}(t))}{\sqrt{Var(G_{12}(t))}}, \qquad W_2 := \frac{G_{21}(t) - E(G_{21}(t))}{\sqrt{Var(G_{21}(t))}}, \\ W_{1B} &:= \frac{G_{12}(t) - E_B(G_{12}(t))}{\sqrt{Var_B(G_{12}(t))}}, \qquad W_{2B} := \frac{G_{21}(t) - E_B(G_{21}(t))}{\sqrt{Var_B(G_{21}(t))}}. \end{aligned} \tag{4}$$

By some algebraic computation, we can get the following relation between the Permutation Null Distribution and the Bootstrap Null Distribution from Eq.(4):

$$\begin{aligned} W_1 &= \frac{\sqrt{Var_B(G_{12}(t))}}{\sqrt{Var(G_{12}(t))}}W_{1B} - \frac{K\frac{t(n-t)}{n(n-1)}}{\sqrt{Var(G_{12}(t))}}, \\ W_2 &= \frac{\sqrt{Var_B(G_{21}(t))}}{\sqrt{Var(G_{21}(t))}}W_{2B} - \frac{K\frac{t(n-t)}{n(n-1)}}{\sqrt{Var(G_{21}(t))}}. \end{aligned} \tag{5}$$

In the next, we will show that (W_{1B}, W_{2B}) converge to a two-dimensional Gaussian distribution as n grows larger under some restrictive conditions. First we let two functions about an edge e of form

$$I_e := I(e \text{ connects a node sampled before } t \text{ to a node sampled after } t),$$
$$\tilde{I}_e := I(e \text{ connects a node sampled after } t \text{ to a node sampled before } t).$$

In order to prove that (W_{1B}, W_{2B}) converges to a two-dimensional Gaussian distribution, we only need to show that $W_B := aW_{1B} + bW_{2B}$ converges to a Gaussian distribution for any pair of real numbers (a, b) such that the variance of W_B is larger than 0.

Given the notations of I_e and $\tilde{I}_e$, we can write W_B in another representation of form

$$W_B = \sum_{e\in G}\left(a\frac{I_e - \frac{t(n-t)}{n^2}}{\sqrt{Var_B(G_{12}(t))}} + b\frac{\tilde{I}_e - \frac{t(n-t)}{n^2}}{\sqrt{Var_B(G_{21}(t))}}\right).$$

Also let

$$\zeta_e := a\frac{I_e - \frac{t(n-t)}{n^2}}{\sqrt{Var_B(G_{12}(t))}} + b\frac{\tilde{I}_e - \frac{t*(n-t)}{n^2}}{\sqrt{Var_B(G_{21}(t))}}.$$

For each $i \in \mathcal{J}$, there exits $\mathcal{S}_i \subset \mathcal{T}_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{S_i^c}$ and ξ_{S_i} is independent of $\xi_{T_i^c}$. Thus, from [2], we have

$$\sup_{h\in Lip(i)} |E[h(W)] - E[h(Z)]| < \delta, \tag{6}$$

where $Lip(1) = \{h : \mathbb{R} \to \mathbb{R}\}$, $Z \sim N(0,1)$, and $\delta = 2\sum_{i\in\mathcal{J}}(E(|\xi_i\eta_i\theta_i|) + |E(\xi_i\eta_i)| \cdot E(|\theta_i|) + \sum_{i\in\mathcal{J}} E(|\xi_i\eta_i^2)|$ with $\eta_i = \sum_{j\in\mathcal{S}_i}\xi_j$ and $\theta_i = \sum_{j\in\mathcal{T}_i}\xi_j$. Let $\eta_e = \sum_{e'\in K_e}\zeta_{e'}$ and $\theta_e = \sum_{e''\in L_e}\zeta_{e''}$. By Eq.(6), we have

$$\sup_{h\in Lip(1)} E|h(\frac{W_B}{\sqrt{Var(W_B)}}) - h(Z)| < \delta, \qquad Z \sim N(0,1),$$

where

$$\delta = \frac{1}{\sqrt{(Var(W_b))^3}} * (2\sum_{e\in G}(E_B|\zeta_e\eta_e\theta_e| + |E_B(\zeta_e\eta_e)|(E_B|\theta_e|)) + \sum_{e\in G} E_B|\zeta_e\eta_e|^2).$$

Also let $m := max(|a|, |b|)$ and $\sigma := \sqrt{min(Var_B(G_{12}(t)), Var_B(G_{21}(t)))}$, then for any edge $e \in G$, it folllows that

$$\begin{aligned}
|\zeta_e| &\le |a\frac{I_e - \frac{t(n-t)}{n^2}}{\sqrt{Var_B(G_{12}(t))}}| + |b\frac{\tilde{I}_e - \frac{t(n-t)}{n^2}}{\sqrt{Var_B(G_{21}(t))}}| \\
&\le \frac{|a|}{\sqrt{Var_B(G_{12}(t))}} + \frac{|b|}{\sqrt{Var_B(G_{21}(t))}} \\
&\le \frac{2m}{\sigma}.
\end{aligned}$$

Consequently, $|\eta_e| \le |K_e|\frac{2m}{\sigma}$, $|\theta_e| \le |L_e|\frac{2m}{\sigma}$ and

$$\begin{aligned}
\delta &\le \frac{1}{\sqrt{(Var(W_b))^3}}\left[2\sum_{e\in G}(E_B|\zeta_e||\eta_e||\theta_e| + (E_B|\zeta_e||\eta_e|)(E_B|\theta_e|)) + \sum_{e\in G} E_B(|\zeta_e||\eta_e|^2)\right] \\
&\le \frac{1}{\sqrt{(Var(W_b))^3}}\left[2\sum_{e\in G}\left((\frac{2m}{\sigma})^3|K_e||L_e| + (\frac{2m}{\sigma})^2|K_e| * (\frac{2m}{\sigma})|L_e|\right)\right] \\
&\le \frac{1}{\sqrt{(Var(W_b))^3}}\frac{8m^3}{\sigma^3}\left(6\sum_{e\in G}|K_e||L_e| + \sum_{e\in G}|K_e|^2\right) \\
&\le \frac{56m^3}{\sqrt{(Var(W_b))^3}}\frac{\sum_{e\in G}|K_e||L_e|}{\sigma^3}.
\end{aligned}$$

Since $Var(W_B) = a^2 + b^2 + 2abcorr_B(G_{12}(t), G_{21}(t))$, and it is of constant order, the order of δ only depends on the orders of $\sum_{e\in G} |K_e||L_e|$ and σ. For simplicity, let $p_n := \frac{t}{n}$ and $q_n := \frac{n-t}{n}$. Since $\sum_{i,j,l=1}^{n} a_{ij}^{+}a_{lj}^{+} \geq nK^2$), then

$$\begin{aligned}
Var_B(G_{12}(t)) =& nKp_nq_n(1-p_nq_n) - (2nK^2 - \sum_{i,j=1}^{n} a_{ij}^{+}a_{ji}^{+})p_n^2q_n^2 \\
&+ nK(K-1)p_n^3q_n + (\sum_{i,j,l=1}^{n} a_{ij}^{+}a_{lj}^{+} - nK)p_nq_n^3 \\
\geq& nKp_nq_n(1-p_nq_n) - 2nK^2p_n^2q_n^2 + nK^3p_n^3q_n - nKp_n^3q_n + (nK^2-nK)p_nq_n^3 \\
\geq& n(Kp_nq_n)(1 + Kp_n + Kq_n - p_nq_n - Kp_nq_n - p_n^2 - q_n^2).
\end{aligned}$$

Because for any pair of (p_n, q_n), $(1+Kp_n+Kq_n-p_nq_n-Kp_nq_n-p_n^2-q_n^2)$ is always larger than 0, so $Var_B(G_{21}(t))$ is at least of order $O(n)$, and it is also the same with $Var_B(G_{21}(t))$, then $\sigma = \sqrt{min(Var_B(G_{12}(t)), Var_B(G_{21}(t)))} \geq O(n^{0.5})$. Consequently, as long as the order of $\sum_{e\in G} |K_e||L_e|$ is smaller than $o(n^{1.5})$, δ will surely converge to 0 and (W_{1B}, W_{2B}) will converge to a two-dimensional Gaussian distribution.

Since $Var_B(G_{12}(t))$ and $Var(G_{12}(t))$ are of the same order, both $\frac{\sqrt{Var_B(G_{12}(t))}}{\sqrt{Var(G_{12}(t))}}$ and $\frac{\sqrt{Var_B(G_{21}(t))}}{\sqrt{Var(G_{21}(t))}}$ converge to a constant. Moreover, since $\frac{K\frac{t(n-t)}{n(n-1)}}{\sqrt{Var(G_{12}(t))}}$ and $\frac{K\frac{t(n-t)}{n(n-1)}}{\sqrt{Var(G_{21}(t))}}$ converge to 0, by Eq.(5), (W_1, W_2) converge to a two-dimensional Gaussian distribution. The proof is complete. ■

From Theorem 1, we have further the following lemma.

Lemma 4 *Assume the Condition 1 holds, $\hat{G_{12}}([nu])$ and $\hat{G_{21}}([nu])$ converge to two independent Gaussian processes $\hat{G_{12}}^*(u)$ and $\hat{G_{21}}^*(u)$ respectively, which have the covariance structure: for $0 < u < v < 1$,*

$$Var\begin{pmatrix} \hat{G_{12}}^*(u) \\ \hat{G_{21}}^*(u) \\ \hat{G_{12}}^*(v) \\ \hat{G_{21}}^*(v) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \frac{u(1-v)}{\sqrt{uv(1-u)(1-v)}} & 0 \\ 0 & 1 & 0 & \frac{u(1-v)}{(1-u)v} \\ \frac{u(1-v)}{\sqrt{uv(1-u)(1-v)}} & 0 & 1 & 0 \\ 0 & \frac{u(1-v)}{(1-u)v} & 0 & 1 \end{pmatrix}.$$

From the results above, we find out that $Z_G([nu])$ will also converge to a stochastic process $Z_G^*(u)$. It is also fairly easy to find out that for $0 < u < v < 1$, both $Z_G^*(u)$ and $Z_G^*(v)$ conform to χ^2 distributions.

3.2 Asymptotic approximations to the p-value

3.2.1 Approximations of the tail probability

In this section, we concern with approximating the value of $P(\max_{n_0\leq t\leq n_1} Z_G(t) > b)$ through the limiting process $Z_G^*(u)$. As $\hat{G_{12}}^*(u)$ and $\hat{G_{21}}^*(u)$ are two independent Gaussian processes with unit variances, we can derive the approximation as [5]:

$$\begin{aligned}
P(\max_{n_0\leq t\leq n_1} Z_G(t) > b) &\approx P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} Z_G^*(t) > b) = P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{12}}^*(u)^2 + \hat{G_{21}}^*(u)^2 > b) \\
&\approx P(\{\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{12}}^*(u)^2 > \frac{b}{2}\} \bigcap \{\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{21}}^*(u)^2 > \frac{b}{2}\}) \\
&= P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{12}}^*(u)^2 > \frac{b}{2}) P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{21}}^*(u)^2 > \frac{b}{2}),
\end{aligned}$$

where $P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{12}}^*(u)^2 > \frac{b}{2})$ and $P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{21}}^*(u)^2 > \frac{b}{2})$ are approximated as:

$$P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{12}}^*(u)^2 > \frac{b}{2}) = P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} |\hat{G_{12}}^*(u)| > \sqrt{\frac{b}{2}}) \approx 2P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{12}}^*(u) > \sqrt{\frac{b}{2}}),$$

$$P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)^2>\frac{b}{2})=P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}|\hat{G_{21}}^*(u)|>\sqrt{\frac{b}{2}})\approx 2P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)>\sqrt{\frac{b}{2}}).$$

From the derivation above, the key to approximating the tail probability $P(\max_{n_0\leq t\leq n_1}Z_G(t)>b)$ is the estimation of $P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)>\sqrt{\frac{b}{2}})$ and $P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)>\sqrt{\frac{b}{2}})$.

3.2.2 Asymptotic approximations to the p-values

At first, we concern with approximating the value of $P(\max_{n_0\leq t\leq n_1}Z_G(t)>b)$, i.e., the p-value, through the limiting process $Z_G^*(u)$ for $0<u<1$. Since $\hat{G_{12}}^*(u)$ and $\hat{G_{21}}^*(u)$ are two independent Gaussian processes with unit variances, we can derive the approximation as [5]:

$$\begin{aligned}P(\max_{n_0\leq t\leq n_1}Z_G(t)>b)&\approx P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}Z_G^*(t)>b)\\&=P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)^2+\hat{G_{21}}^*(u)^2>b)\\&\approx P(\{\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)^2>\frac{b}{2}\}\bigcap\{\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)^2>\frac{b}{2}\})\\&=P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)^2>\frac{b}{2})P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)^2>\frac{b}{2}),\end{aligned}$$

where $P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)^2>\frac{b}{2})$ and $P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)^2>\frac{b}{2})$ are approximated as

$$P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)^2>\frac{b}{2})=P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}|\hat{G_{12}}^*(u)|>\sqrt{\frac{b}{2}})\approx 2P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)>\sqrt{\frac{b}{2}})$$

and

$$P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)^2>\frac{b}{2})=P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}|\hat{G_{21}}^*(u)|>\sqrt{\frac{b}{2}})\approx 2P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)>\sqrt{\frac{b}{2}}),$$

respectively.

From the derivation above, the key to approximating the tail probability $P(\max_{n_0\leq t\leq n_1}Z_G(t)>b)$ is the estimation of $P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)>\sqrt{\frac{b}{2}})$ and $P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)>\sqrt{\frac{b}{2}})$.

In the following, we mainly concern with approximating $P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)>\sqrt{\frac{b}{2}})$ and $P(\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)>\sqrt{\frac{b}{2}})$ by taking advantage of its covariance structure. Our approximation involves the function $\upsilon(x)$, which is defined as

$$\upsilon(x)=2x^{-2}exp\{-2\sum_{m=1}^{\infty}m^{-1}\Phi(-\frac{1}{2}xm^{\frac{1}{2}})\},$$

where $\Phi(x)$ is the probability distribution function of standard normal distribution. This function is closely related to the Laplace transform of the overshoot over the boundary of a random walk. A simple approximation is sufficient for numerical purposes [6]:

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

where $\phi(x)$ is the probability density function of the standard normal distribution. Then the tail probability of $\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{12}}^*(u)$ and $\max_{\frac{n_0}{n}\leq t\leq\frac{n_1}{n}}\hat{G_{21}}^*(u)$ could be obtained from the following theorem.

Theorem 2 *Assume that* $\frac{n_0}{n} \to x_0$, $\frac{n_1}{n} \to x_1$ *and* $\sqrt{\frac{b}{2n}} \to b_0$ *hold as* $n \to \infty$. *Then we have*

$$P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{12}}^*(u) > \sqrt{\frac{b}{2}}) \sim \sqrt{\frac{b}{2}}\phi(\sqrt{\frac{b}{2}}) \int_{x_0}^{x_1} g_{12}(x) v(b_0\sqrt{2g_{12}(x)})dx \tag{7}$$

and

$$P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{21}}^*(u) > \sqrt{\frac{b}{2}}) \sim \sqrt{\frac{b}{2}}\phi(\sqrt{\frac{b}{2}}) \int_{x_0}^{x_1} g_{21}(x) v(b_0\sqrt{2g_{21}(x)})dx \tag{8}$$

as $n \to \infty$, *where* $g_{12}(x) = \frac{1}{2x(1-x)}$ *and* $g_{21}(x) = \frac{1}{x(1-x)}$.

Proof. We approximate $P(\max_{n_0 \le t \le n_1} \hat{G_{12}}(t) > b)$ by the Gaussian process that it converges to $P(\max_{n_0 \le t \le n_1} \hat{G_{12}}^*(\frac{t}{n}) > b)$, where

$$\begin{aligned}
&P(\max_{n_0 \le t \le n_1} \hat{G_{12}}^*(\frac{t}{n}) > b) \\
=&\sum_{n_0 \le t \le n_1} \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+b)^2}{2}} P(\max_{n_0 \le s} \hat{G_{12}}^*(\frac{s}{n}) < b | \hat{G_{12}}^*(\frac{t}{n}) = x + b)dx \\
=&\frac{\phi(b)}{b} \sum_{n_0 \le t \le n_1} \int_0^\infty e^{-x} e^{-\frac{x^2}{2b^2}} P(b(\hat{G_{12}}^*(\frac{s}{n}) - \hat{G_{12}}^*(\frac{t}{n})) < -x | \hat{G_{12}}^*(\frac{t}{n}) = b + \frac{x}{b})dx.
\end{aligned}$$

If $x \sim o(b^2)$, $\frac{x^2}{b^2}$ is negligible to x and $\frac{x}{b}$ is negligible to b, then

$$P(\max_{n_0 \le t \le n_1} \hat{G_{12}}^*(\frac{t}{n}) > b) = \frac{\phi(b)}{b} \sum_{n_0 \le t \le n_1} \int_0^\infty e^{-x} P(b(\hat{G_{12}}^*(\frac{s}{n}) - \hat{G_{12}}^*(\frac{t}{n})) < -x | \hat{G_{12}}^*(\frac{t}{n}) = b)dx.$$

From [7], we have

$$b(\hat{G_{12}}^*(u) - \hat{G_{12}}^*(v)) | \hat{G_{12}}^*(v) = b \sim N((\rho_G^*(u,v) - 1)b^2, (1 - \rho_G^*(u,v)^2)b^2),$$

and by the second order Taylor expansion, it follows that

$$\begin{aligned}
\rho_G^*(u,v) &= 1 + f'_{v,-}(0)(u-v) + \frac{1}{2} f''_{v,-}(0)(u-v)^2, \\
\rho_G^*(u,v)^2 &= 1 + 2f'_{v,-}(0)(u-v) + [f'_{v,-}(0)^2 + f''_{v,-}(0)](u-v)^2.
\end{aligned} \tag{9}$$

After some algebraic calculation, it can be easily shown that $f'_{v,-}(0) = \frac{1}{2v(1-v)}$. By Eq.(9) and neglecting the second order terms, we can approximate the distribution as

$$b(\hat{G_{12}}^*(u) - \hat{G_{12}}^*(v)) | \hat{G_{12}}^*(v) = b \sim N(-f'_{v,-}(0)|v-u|b^2, 2f'_{v,-}(0)|v-u|b^2).$$

Let $W_m^{(t)}$ be a random walk with $W_1^{(t)} \sim N(\frac{1}{n} f'_{v,-}(0)b^2, 2\frac{1}{n} f'_{v,-}(0)b^2)$, then we have

$$P(\max_{n_0 \le s < t} b(\hat{G_{12}}^*(u) - \hat{G_{12}}^*(v)) < -x | \hat{G_{12}}^*(v) = b) \sim P(\max_{n_0 \le s < t} -W_{t-s}^{(t)} < -x) \sim P(\min_{t \ge 1} W_m^{(t)} > x).$$

Let

$$g_{12}(x) := \lim_{u \nearrow x} \frac{\partial \rho_{g_{12}}(u,x)}{\partial u}, \;\; g_{21}(x) := \lim_{u \nearrow x} \frac{\partial \rho_{g_{21}}(u,x)}{\partial u}, \tag{10}$$

where

$$\rho_{g_{12}}(u,x) = cov(\hat{G_{12}}^*(u), \hat{G_{12}}^*(x)), \;\; \rho_{g_{21}}(u,x) = cov(\hat{G_{21}}^*(u), \hat{G_{21}}^*(x)).$$

By Eq.(10), we have

$$g_{12}(x) = \lim_{u \nearrow x} \frac{\partial \rho_{g_{12}}(u,x)}{\partial u} = f'_{x,-}(0) = \frac{1}{2x(1-x)}.$$

Since $\int_0^\infty e^{\frac{-2\mu x}{\sigma}} P(\min_{m\geq 1} W_m > x)dx = \mu v(\frac{2\mu}{\sigma})$ if a random walk W_m satisfies $W_1 \sim N(\mu, \sigma)$, we have

$$\begin{aligned} &P(\max_{n_0\leq t\leq n_1} \hat{G_{12}}^*(\frac{t}{n}) > b) \\ \sim &\lim_{n\to\infty} \frac{\phi(b)}{b} \sum_{n_0\leq t\leq n_1} \frac{b^2}{n} g_{12}(\frac{t}{n}) v(\frac{b}{\sqrt{n}}\sqrt{2g_{12}(\frac{t}{n})}) \\ \sim &b\phi(b)\int_{x_0}^{x_1} g_{12}(x) v(\frac{b}{\sqrt{n}}\sqrt{2g_{12}(x)})dx, \end{aligned}$$

where $x_0 = \lim_{n\to\infty} \frac{n_0}{n}$ and $x_1 = \lim_{n\to\infty} \frac{n_1}{n}$. Thus, it follows that Eq.(7) holds.

With the same argument, we obtain Eq.(8). The proof is complete. ■

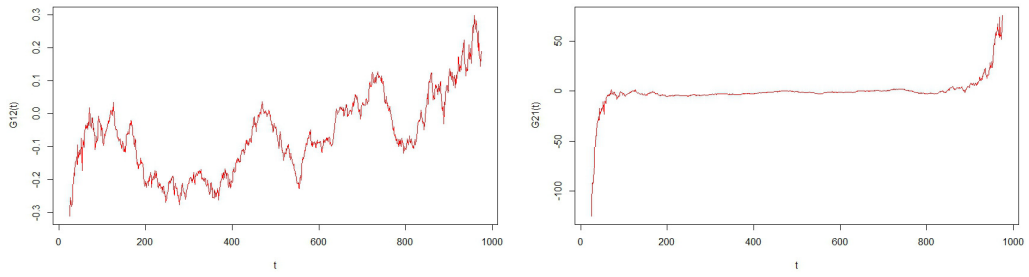
If the Condition 1 is satisfied, the whole tail probability $P(\max_{n_0\leq t\leq n_1} Z_G(t) > b)$ could be approximated by

$$\begin{aligned} P(\max_{n_0\leq t\leq n_1} Z_G(t) > b) &\approx 4P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{12}}^*(u) > \sqrt{\frac{b}{2}}) P(\max_{\frac{n_0}{n}\leq t\leq \frac{n_1}{n}} \hat{G_{21}}^*(u) > \sqrt{\frac{b}{2}}) \\ &\approx 2b\phi^2(\sqrt{\frac{b}{2}}) \int_{x_0}^{x_1} g_{12}(x) v(b_0\sqrt{2g_{12}(x)})dx \int_{x_0}^{x_1} g_{21}(x) v(b_0\sqrt{2g_{21}(x)})dx. \end{aligned}$$

From above, $g_{12}(x)$ and $g_{21}(x)$ have some special meanings, they both indicate the limitation of the partial derivations of the covariances of the two independent Gaussian processes $\hat{G_{12}}^*(u)$ and $\hat{G_{21}}^*(u)$.

3.3 Skewness correction

Generally speaking, the methodology proposed above requires the convergence of both $\hat{G_{12}}(t)$ and $\hat{G_{21}}(t)$ to Gaussian distribution to be fast. However, as shown by simulation results, the convergence might be quite slow when $\frac{t}{n}$ comes closer to 0 or 1 with a bigger skewness in most cases, which requires further correction of the original approach. This phenomenon is shown in Fig. 3, where the samples are sampled from the 10-dimensional standard Gaussian distribution.



(a) $E\hat{G_{12}}(t)^3$ (b) $E\hat{G_{21}}(t)^3$

Figure 3: Simulation results for convergence.

From Fig. 3, we know that both $\hat{G_{12}}(t)$ and $\hat{G_{21}}(t)$ become highly left-skewed when $\frac{t}{n}$ comes close to 0 and highly right-skewed when $\frac{t}{n}$ comes closer to 1, so that the p-value estimation is inaccurate at the end. It needs skewness correction in the tail probability approximation for them.

We first consider the approximation of the marginal probability $P(\hat{G_{12}}(t) \in b + dx/b)$. Since $\hat{G_{12}}(t)$ is standardized, namely, $E(\hat{G_{12}}(t)) = 0$, and $Var(\hat{G_{12}}(t)) = 1$, we make use of the cumulative generating function $\psi(\theta) = logE_P(e^{\theta G_{12}(t)})$, and its value and first derivatives at 0 could be easily obtained by the standardization. By change of measure, $d\mathcal{Q}_\theta = e^{\theta G_{12}(t) - \psi(\theta)} dP$, we can approximate $P(\hat{G_{12}}(t) \in b + dx/b)$ by

$$\frac{1}{\sqrt{2\pi(1+\gamma\theta_b)}} exp(-\theta_b b - x\theta_b/b + \theta_b^2(1+\gamma\theta_b/3)/2),$$

where θ_b is specially chosen to satisfy $\dot{\psi}(\theta_b) = b$. By Taylor expansion of $\psi(\theta)$ at b, we can approximate θ_b by

$$\theta_b \approx (-1 + \sqrt{1 + 2\gamma b})/\gamma,$$

where $\gamma(t) = E(G_{12}(t)^3)$.

After skewness correction being finished, we can approximate $P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{12}}^*(u) > \sqrt{\frac{b}{2}})$ by adding the skewness term into the integral, from which we get the following approximation of the tail probability.

Lemma 5 *Approximation of* $P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{12}}^*(u) > \sqrt{\frac{b}{2}})$ *with skewness correction is*

$$P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{12}}^*(u) > \sqrt{\frac{b}{2}}) \approx \sqrt{\frac{b}{2}}\phi(\sqrt{\frac{b}{2}}) \int_{\frac{n_0}{n}}^{\frac{n_1}{n}} S_{G_{12}}(nx) g_{12}(x) \upsilon(b_0\sqrt{2g_{12}(x)}) dx,$$

where $S_{G_{12}}(t) = \frac{exp((1/2)(b-\theta_b(t))^2 + (1/6)\gamma(t)\theta_b(t)^3)}{\sqrt{1+\gamma(t)\theta_b(t)}}$ *with* $\gamma(t) = E(G_{12}(t)^3)$ *and* $\theta_b(t) = (-1 + \sqrt{1 + 2\gamma(t)b})/\gamma(t)$.

Here as we have already derived the skewness correction method for approximating $P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{12}}^*(u) > \sqrt{\frac{b}{2}})$, skewness correction for $P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{21}}^*(u) > \sqrt{\frac{b}{2}})$ can also be carried out in the same way by simply replacing the third-moment term $\gamma(t)$ and the other terms likewise.

Lemma 6 *Approximation of* $P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{21}}^*(u) > \sqrt{\frac{b}{2}})$ *with skewness correction is*

$$P(\max_{\frac{n_0}{n} \le t \le \frac{n_1}{n}} \hat{G_{12}}^*(u) > \sqrt{\frac{b}{2}}) \approx \sqrt{\frac{b}{2}}\phi(\sqrt{\frac{b}{2}}) \int_{\frac{n_0}{n}}^{\frac{n_1}{n}} S_{G_{21}}(nx) g_{21}(x) \upsilon(b_0\sqrt{2g_{21}(x)}) dx,$$

where $S_{G_{21}}(t) = \frac{exp((1/2)(b-\theta_b(t))^2 + (1/6)\gamma(t)\theta_b(t)^3)}{\sqrt{1+\gamma(t)\theta_b(t)}}$ *with* $\gamma(t) = E(G_{21}(t)^3)$ *and* $\theta_b(t) = (-1 + \sqrt{1 + 2\gamma(t)b})/\gamma(t)$.

4 Numerical analysis

Numerical analysis is conducted to test the validity of the new test statistic. For a given p-value $\widetilde{p}$, the critical value c is chosen to satisfy the condition

$$P(\max_{n_0 \le t \le n_1} Z_G(t) > c) = \widetilde{p}.$$

The true critical value could be replaced by the permutation critical value when the simulations times are sufficiently large. If the new test statistic performs well, then the estimated asymptotic critical value would not be much different from the permutation critical value.

The table 1 below shows both the true and the estimated critical value with the given p-value of 0.05. 1000 samples are generated from the ten-dimensional Gaussian distribution $N(0, I_{10})$.

Table 1: The true and the estimated critical value with the given p-value of 0.05.

Permutation Critical Value	Asymptotic Critical Value	Asymptotic Critical Value with Skewness Correction	t_1	t_2
14.36	15.05	14.65	26	975
14.81	15.05	15.03	26	975
14.44	15.05	14.80	26	975
13.99	14.45	14.15	51	950
13.47	14.45	13.77	51	950
13.67	14.45	13.91	51	950

From the table 1, we know that the asymptotic critical value can give a rough estimate of the true critical value and that skewness correction can improve the validity of the original estimation to some extent, even though not significantly.

5 Application to real life data

5.1 The character-event matrix for *a dream in red mansions*

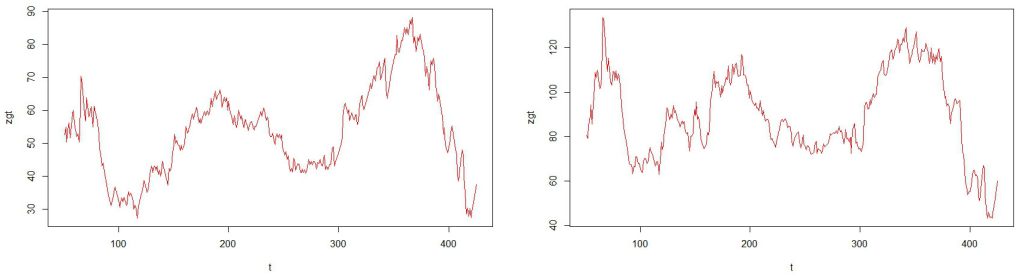
As is known to all, a dream in red mansions is one of the most famous novels in the history of Chinese literature, which is characterized by tortuous plots and vivid character images. Here we check up the character-event matrix and perform change point detection on it.

The character-event matrix is a matrix with 475 rows and 374 columns. Each row represents an event happened and is arranged according to time, each column represents a character in the novel. Every element in the matrix only takes value 0 or 1. For example, the (i, j)th element in the matrix takes value 1 if the jth character occurs in the ith event and takes value 0 if the jth character did not occur in the ith event.

Then the matrix could be treated as a sequence of 475 observations sampled in a 374-dimensional space with each observation representing an event. If the ith dimension of an observation takes value 1, then the ith character will occur in the event corresponding to that observation.

5.2 Result of application

We conduct change point detection for the matrix through two different graphs constructed upon it, the 1-NNG, the 3-NNG, and the corresponding test statistic $Z_G(t)$ is also computed and its maximal value is compared with the theoretically approximated p-value.



(a) **1-NNG** (b) **3-NNG**

Figure 4: Change point detection of $Z_G(t)$.

From Fig. 4, in both cases, the constraints for the starting point n_0 is chosen as 50, while n_1 is chose as 425. The critical value b obtained from p-value approximation such that

$$P(\max_{\frac{50}{475}\leq u\leq\frac{425}{475}} Z_G^*(t) > b) = 0.01$$

is 37.1. However, in the 1-NNG case, the

$$\max_{50\leq t\leq 425} Z_G(t)$$

is 88.3 and in the 3-NNG case, it becomes even greater and reaches 133.7. In both cases, we can reach the conclusion that a significant change point occurs in the character-event data. Moreover, $Z_G(t)$ reaches to three peaks across the sequence of events in both cases. In the 1-NNG case, $Z_G(t)$ reaches to a peak at the 66th event, the 192th event and the 352th event; while in the 3-NNG case, it reaches the peak at the 66th event, the 193th event and the 367th event. The three peaks happen at around the 16th chapter, the 50th chapter and the 82th chapter, and reasonable answers can be found out to account for these changes.

6 Conclusion

The new method for change point detection based on directed nearest neighbour graph is a great improvement for the old method based on the undirected nearest neighbour graph. The new method adopted the χ^2 like statistics for hypothesis

testing that is frequently used in multi-sequence change point detections. Combining the advantage of the graph-based method with that of the χ^2 like statistics, we solve the case of variance shift of change point detection, where both of the previous methods lost their powers, especially in high dimensions. The new method has great applications in detecting the disorder in sequential data as demonstrated by the analysis of the text of the novel 'Dream of Red Mansions'.

References

[1] E.G. Carlstein, H.G. M$\ddot{u}$ller and D. Siegmund. Change-point problems. *IMS Lecture Notes.* 1994.

[2] H. Chen and J.H. Friedman. A new graph-based two-sample test for multivariate and object data. *Publications of the American Statistical Association*, 112(2017): 397–409.

[3] N.R. Zhang, D. Siegmund, H.L. Ji and J.Z. Li. Detecting simultaneous change points in multiple sequences. *Biometrika*, 97(2010): 631–645.

[4] L.I. Vostrikova. Detection of the disorder in multi-dimensional random-processes. *Doklady Akademii Nauk SSSR*, 259(1981): 270–274.

[5] D. Siegmund. Tail approximations for maixma of random fields. *Probability Theory: Proceedings of the 1989 Singapore Probability Conference*, 1992: 147–158.

[6] H. Chen and N.R. Zhang. Graph-based test for two-sample comparisons of categorical data. *Statistica Sinica*, 23(2013): 1479–1503.

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

[8] H. Chen. Sequential change point detection based on nearest neighbours. *arXiv preprint arXiv:1604.03611*, 2016.

[9] I.P. Tu and D. Siegmund. The maximum of a function of a Markov Chain and application to linkage analysis. *Advances in Applied Probability*, 31(1999): 510–531.

[10] B. James, K.L. James and D. Siegmund. Tests for a change point. *Biometrika*, 74(1987): 71–83.

[11] M. Radonović, A. Nanopoulos and M. Ivanović. Hubs in Space: Popular nearest neighbours in high-dimensional data. *Journal of Machine Learning Research*, 11(2010): 2487–2531.

[12] D. Siegmund. Approximate tail probabilities for the maxima of some random fields. *The Annals of Probability*, 1988: 487–501.

[13] N.R. Zhang, D. Siegmund, H.L. Ji and J.Z. Li. Detecting simultaneous change-points in multiple sequences. *Biometrika*, 97(2010): 631–645.

[14] I.A. Ibragimov and Y.A. Rozanov. Gaussian Random Processes. Springer Verlag. 1978.

[15] A. Tartakovsky, I. Nikiforov and M. Basseville. Sequential analysis: Hypothesis testing and changepoint detection. Chapman and Hall/CRC. 2014.