Change-point detection using graph-based methods

Dissertation report of Soham Das under
the guidance of Prof. Anil K Ghosh and Prof. Soumendu Sundar Mukherjee

January 13, 2022

Abstract

Detection of change points in a sequence of multivariate observations is a classical problem in Statistics, and several parametric and nonparametric methods have been proposed for it. Recently, some graph-based methods have also been proposed in the literature. These graph-based approaches can be applied to any data sets of any arbitrary dimensions even when the dimension is larger than the sample size. This project concentrates on analyzing some of these graph-based methods, particularly the issues they face while detecting scale shifts in distribution in a sequence of multivariate observations. We shall try to propose some new graph-based methods to resolve this issue and analyze their theoretical and empirical performance for high dimensional data.

1 Introduction

As we go into the big data age, change-point analysis is gaining traction. Many areas acquire massive volumes of data in order to explore complicated phenomena over time and/or space. Such data frequently consists of a series of high-dimensional or non-Euclidean observations that cannot be examined using conventional methods. Segmentation, which splits the sequence into homogeneous temporal or geographical segments, is widely used to get insights from such data.

Among many other uses, change-point methods can be used to detect change in distribution of a sequence of observations. This problem can be formulated in the following way. Let Y_i be a sequence of observations from $\mathbb{R}^d$ where $i = 1, 2, \cdots, n$ and $d \geq 1$. Let

$$\begin{aligned} Y_i &\sim F_0 \quad \text{for } i = 1, 2, \cdots, \tau \\ \text{and } Y_i &\sim F_1 \quad \text{for } i = \tau + 1, \cdots, n \end{aligned}$$

where F_0 and F_1 are two different distributions and $1 < \tau < n$. We also assume that Y_i's are independent. Here the problem is two-fold. First we need to test for existence of change-point and then we need to estimate the location of the change-point τ.

In the literature, many parametric and nonparametric methods (e.g., likelihood test, kernel methods etc.) have been proposed over the years. Particularly, graph-based methods in nonparametrics have gained much prominence. Two-sample tests based on various types of graphs representing the similarity between observations were first proposed in Friedman and Rafsky (1979) and Rosenbaum (2005). Several nonparametric methods have been constructed for detecting the changes in distribution, and many of them can be used even in high dimension, low sample size

situations. Chen and Zhang (2015) proposed some methods using graph-based two sample tests. In particular, they considered the tests based on minimum distance pairing (Rosenbaum (2005)), nearest neighbors (Schilling (1986); Henze (1988)) and minimum spanning tree Friedman and Rafsky (1979). Some graph-based change-point detection methods were also proposed in Shi et al. (2017) and Sun et al. (2019). Matteson and James (2014) developed a method based on the energy statistic (see, e.g., Székely and Rizzo (2013)). Some kernel-based methods are also available in the literature (see, e.g., Desobry et al. (2005); Harchaoui et al. (2009); Li et al. (2015); Arlot et al. (2019)). But as pointed out by Chen and Zhang (2015), the performance of these methods depends heavily on the choice of the kernel function and the associated tuning parameter called bandwidth. This problem becomes more prominent for high-dimensional data.

In recent times, two articles Chen and Zhang (2015) and Chu and Chen (2019) have proposed graph-based methods for detection and estimation of change-point. In the first article, the authors suggests a scan statistic based on maximum spanning tree(MST), nearest neighbour(NN) and minimum distance pairing(MDP) using inter-point distances. This statistics are nonparametric distribution free only for MDP (see Rosenbaum (2005)). This method uses scan statistic based on counts of between-sample edges. But it fails to estimate change-point correctly when there is scale shift in distribution. In Chu and Chen (2019), a similar scan statistic were introduced to overcome such drawbacks. Both these methods work for any dimension.

Although the statistic proposed in Chu and Chen (2019) works quite well for scale and location shifts in Normal distribution, it does not work well for heavy tailed distributions such as Cauchy distribution and for higher dimension but low sample size (HDLSS) scenario. Section 2 provides some drawbacks in some scenarios through simulations. In Section 3 we discuss how the failure can be attributed to geometric representation of data points in high dimension (see Hall et al. (2005)). Section 4 discusses about a possible way to subdue this problem. Section 5 provides some more simulations to support the new plan. Section 6 contains all proofs of theorems and results used in this report.

2 Existing methods and their drawbacks

First we state the null and alternative hypotheses for our testing problem. And then discuss about existing graph-based method we have mentioned above. Let there are n data points $\{y_i : i = 1, \cdots, n\}$ in $\mathbb{R}^d$, $d \geq 1$. For testing the null hypotheses

$$H_0 : \mathrm{y}_i \sim F_0, i = 1, \ldots, n$$

against the single change-point alternative

$$H_1 : \exists 1 \leq \tau < n, \mathrm{y}_i \sim \begin{cases} F_0, & i \leq \tau \\ F_1, & \text{otherwise} \end{cases}$$

In Chu and Chen (2019), the authors introduced a scan statistic $S(t)$ based on minimum spanning tree(MST) or nearest neighbour graph(NNG). For a sequence of observations $Y_1, Y_2, \cdots, Y_n$, consider a change-point at index t where $1 < t < n$, and split n observations $\{y_i : i = 1, \cdots, n\}$ into two groups $\{y_1, \cdots, y_t\}$ and $\{y_{t+1}, \cdots, y_n\}$. Call these two groups as G_1 and G_2. If t is a true change-point, then observations from G_1 and G_2 would come from two different distributions. According to our assumption, observations from group G_1 and group G_2 come from distributions F_0 and F_1 respectively under the alternative hypothesis.

Now we construct an MST or NNG graph G using Euclidean distance or L_2 distance by considering n observations as n vertices of this graph G. Note that there are three possible types of edges in G: (i) edges between vertices withing group G_1, (ii) edges between vertices withing group G_2 and (iiii) edges between vertices of G_1 & G_2. Now define $R_i(t)$ as the number of edges between vertices which belong to group G_i for each $i = 1, 2$. And define the statistic as:

$$S(t) = \begin{pmatrix} R_1(t) - \mathbf{E}\left(R_1(t)\right) \\ R_2(t) - \mathbf{E}\left(R_2(t)\right) \end{pmatrix}^T \Sigma^{-1}(t) \begin{pmatrix} R_1(t) - \mathbf{E}\left(R_1(t)\right) \\ R_2(t) - \mathbf{E}\left(R_2(t)\right) \end{pmatrix},$$

where $(\mathbf{E}(R_1(t)), \mathbf{E}(R_2(t)))$ and $\Sigma(t)$ are expectation and covariance matrix of $(R_1(t), R_2(t))$ under the permutation null. Under the null hypothesis H_0, the joint distribution of the observations in the sequence is same if we permute the order of the observations. So, under the permutation null distribution that places $1/n!$ probability on each of the $n!$ permutations of $\{y_i : i = 1, \cdots, n\}$. We use this permutation null distribution to calculate the p-value under H_0. The analytic expressions for $\mathbf{E}(R_1(t)), \mathbf{E}(R_2(t))$, and $\Sigma(t)$ were calculated through combinatorial analysis, and they can be obtained straightforwardly following Chen and Friedman (2017).

Note that, under H_0, $S(t)$ should follow same distribution for any $t \in \{1, \cdots, n\}$, and under H_1, $S(t)$ is maximized at the true value of change-point τ. Hence, we estimate change-point by $\hat{\tau} = \arg\max_{1 \le t < n} S(t)$ and p-value of this estimate is calculated under permutation null. To summarize, first we estimate a change-point using the scan statistic, and then find the p-value for this potential change-point. If the p-value is below a specified limit, then we conclude that there is change-point in the observations and consider this potential change-point as a valid change-point estimate. And if the p-value is above the limit, then we say that there is no significant change-point in the observations.

2.1 Simulations

We are interested in the performance of this statistic in a few specific situations. In this part, we have produced a few plots based on simulated data under certain situations where this statistic $S(t)$ works well and where it does not.

First we check the performance of the test statistic for location and scale shift in normally distributed data points. Looking at plots (a),(b),(c),(d) of Figure 1 we conclude that statistic $S(t)$ works quite well for location and scale shift in normally distributed data points for both $n > d$

and $n < d$ cases. After this we do the same exercise on heavy-tailed distribution like Cauchy distribution. From plots in Figure 2, we are certain that statistic $S(t)$ estimates the change-points wrongly. A reason behind such behaviour is that second moments of Cauchy distribution are infinite. So $S(t)$ which is based on L_2 distance fails to detect change-points correctly.

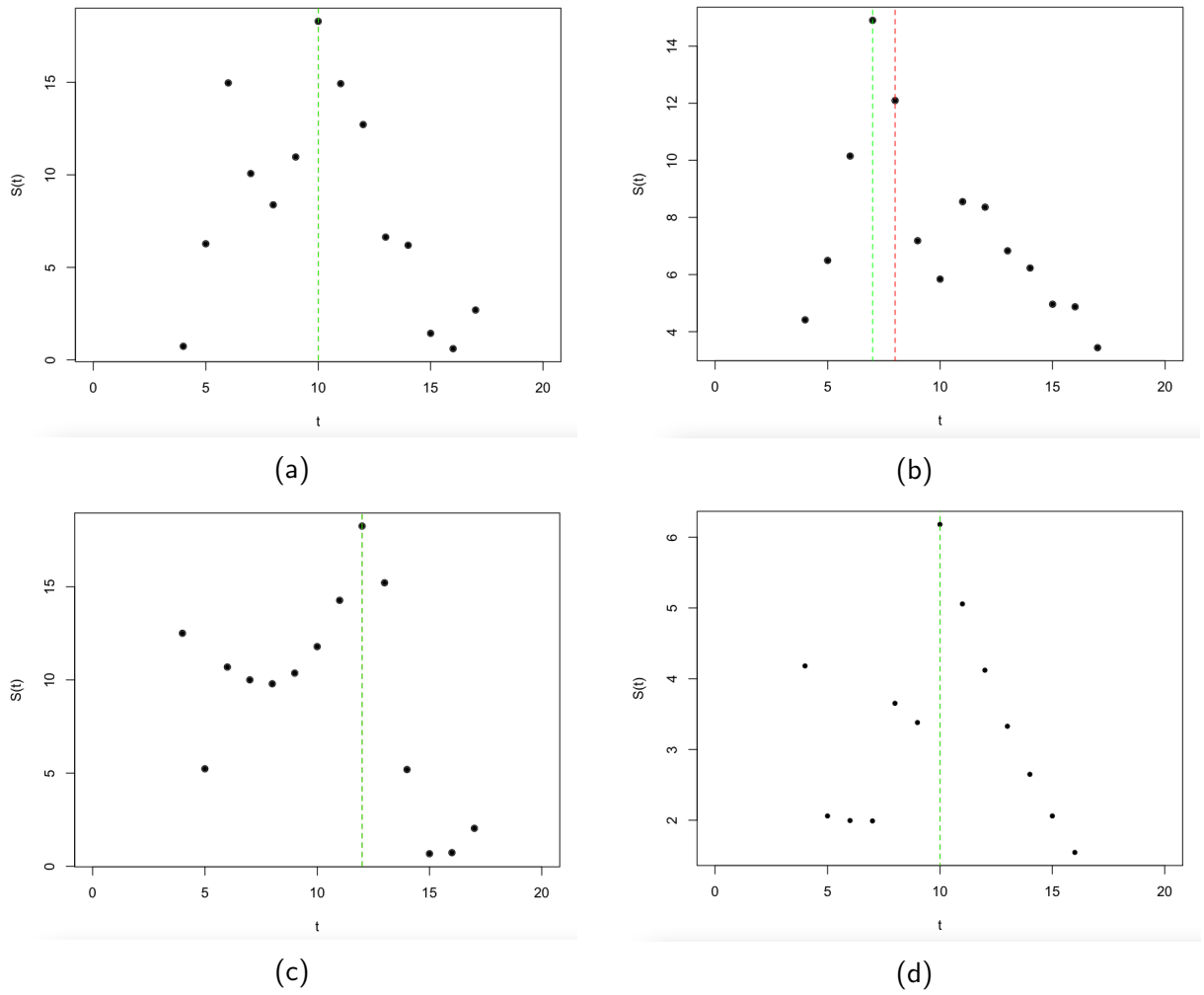


Figure 1: The vertical red and green lines show original and estimated change-points. (a) location shift when $n > d$: 10 observations generated from $N_5(\mathbf{0}_5, \boldsymbol{I}_5)$ and 10 observations generated from $N_5(\mathbf{2}_5, \boldsymbol{I}_5)$ for dimension $d = 5$, (b) scale shift when $n > d$: 8 observations generated from $N_5(\mathbf{0}_5, \boldsymbol{I}_5)$ and 12 observations generated from $N_5(\mathbf{0}_5, 4\boldsymbol{I}_5)$ for dimension $d = 5$, (c) location shift when $n < d$: 12 observations generated from $N_{50}(\mathbf{0}_{50}, \boldsymbol{I}_{50})$ and 8 observations generated from $N_{50}(\mathbf{2}_{50}, \boldsymbol{I}_{50})$ for dimension $d = 50$, (d) scale shift when $n < d$: 8 observations generated from $N_{50}(\mathbf{0}_{50}, \boldsymbol{I}_{50})$ and 12 observations generated from $N_{50}(\mathbf{0}_{50}, 4\boldsymbol{I}_{50})$ for dimension $d = 50$, where $\mathbf{2}_p$ is a vector of length p with all entries equal to 2 and $\boldsymbol{I}_p$ is identity matrix of size p.

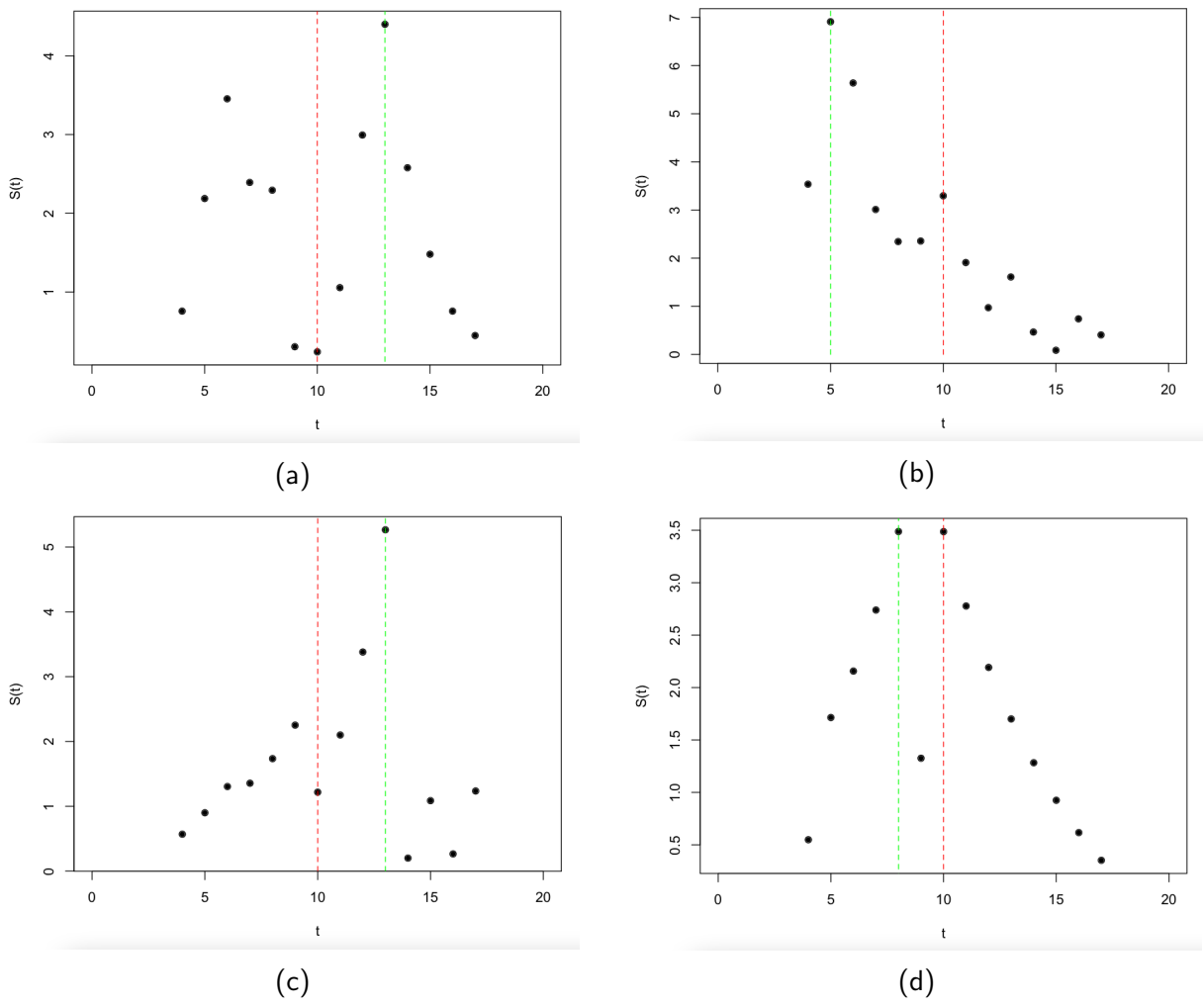
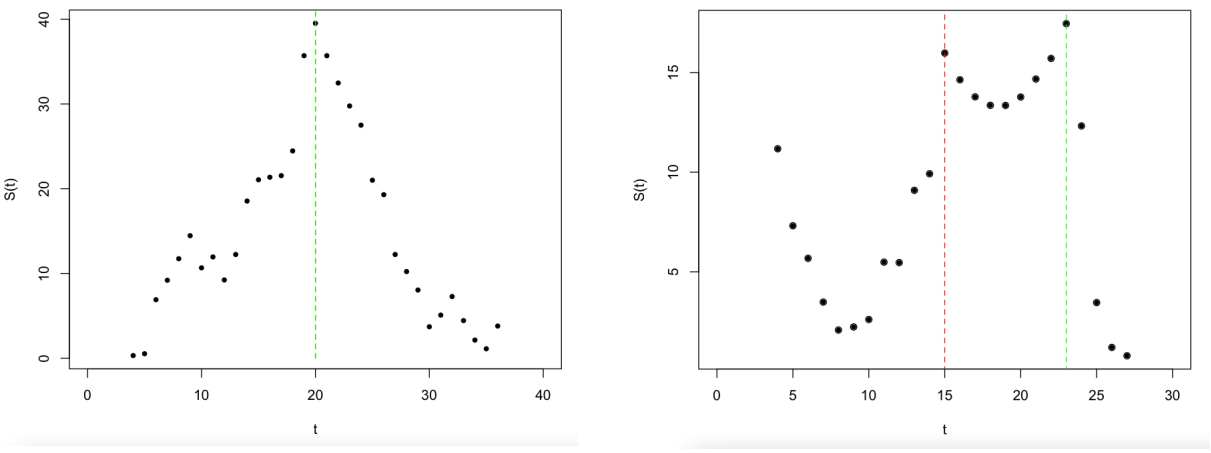


Figure 2: The vertical red and green lines show original and estimated change-points. Here we have used Cauchy distribution for (a),(b),(c),(d) with exactly same parameters and dimensions as in Figure 1

In our third set of simulations, we do the same exercise for normally distributed data points, but this time the coordinates of each Y_i do not have same variance, i.e., the diagonal entries of covariance matrix of Y_i are not same. Let $Y_i \sim N_{40}(\mathbf{0}_{40}, \Sigma)$ for $1 \leq i \leq \tau$ and $y_i \sim N_{40}(\mathbf{0}_{40}, \Gamma)$ for $\tau + 1 \leq i \leq n$ where $\Sigma = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$, $\Gamma = \begin{bmatrix} B & 0 \\ 0 & A \end{bmatrix}$ and $A = I_{20}$, $B = 9I_{20}$. Note that in this case, since covariance matrices Σ and Γ are different, there is a change-point at τ. But in Figure 3, as we can see, the existing method using $S(t)$ correctly estimates change-point when $n > d$ but fails to do so when $n < d$. The reason behind this anomaly is explained in Section 3.



(a) (b)

Figure 3: (a) $n > d$ case: 50 observations generated from $N_{40}(\mathbf{0}_{40}, \Sigma)$ and 50 observations generated from $N_{40}(\mathbf{0}_{40}, \Gamma)$ for dimension $d = 40$, (b) $n < d$ case: 10 observations generated from $N_{40}(\mathbf{0}_{40}, \Sigma)$ and 10 observations generated from $N_{40}(\mathbf{0}_{40}, \Gamma)$ for dimension $d = 40$.

3 Geometry of data points on high dimension

In this section, we try to explain why the statistic $S(t)$ failed for low sample size in high-dimension in Figure 3. We start with a theorem.

Theorem 1. *If a sequence of real-valued random variables $\{X_i : i \geq 1\}$ has uniformly bounded second moment and ρ-mixing property holds for $\{X_i : i \geq 1\}$, i.e. there is a function $\rho : \mathbb{N} \to \mathbb{R}^+ \cup \{0\}$ such that*

$$\sup_{1 \leq q < q' \leq \infty, |q-q'| > r} |\mathrm{Cor}\{X_q, X_{q'}\}| \leq \rho(r) \overset{r \to \infty}{\longrightarrow} 0,$$

then

$$\frac{1}{d}\sum_{i=1}^{d} X_i - \frac{1}{d}\sum_{i=1}^{d} \mathbb{E}X_i \overset{\mathbb{P}}{\longrightarrow} 0.$$

We can use this theorem to prove the following result from Hall et al. (2005).

Result 1. *For $\{X_i : i \geq 1\}$ vectors in $\mathbb{R}^d$, assume the following.*

(a) The fourth moments of the entries of the data vectors are uniformly bounded.

(b) For a constant σ^2,

$$\frac{1}{d}\sum_{k=1}^{d} \mathrm{var}\left(X^{(k)}\right) \to \sigma^2$$

(c) ρ-mixing property holds for $\{X_i : i \geq 1\}$.

Then it follows by a law of large numbers that the distance between X_i and X_j, for any $i \neq j$, is approximately equal to $(2\sigma^2 d)^{1/2}$ as $d \to \infty$, in the sense that

$$\frac{1}{d^{1/2}} \left\{ \sum_{k=1}^{d} \left(X_i^{(k)} - X_j^{(k)} \right)^2 \right\}^{1/2} \to \left(2\sigma^2 \right)^{1/2},$$

This result tells some interesting asymptotic property of high-dimension. For d-dimensional point $X = (X^{(1)}, \cdots, X^{(d)})$ and $Y = (Y^{(1)}, \cdots, Y^{(d)})$ if we assume conditions in Result 1 along with some more assumptions that

$$\frac{1}{d} \sum_{k=1}^{d} \text{var} \left(Y^{(k)} \right) \to \tau^2, \qquad \frac{1}{d} \sum_{k=1}^{d} \text{var} \left(X^{(k)} \right) \to \sigma^2$$

and

$$\frac{1}{d} \sum_{k=1}^{d} \left\{ E \left(X^{(k)} \right) - E \left(Y^{(k)} \right) \right\}^2 \to \mu^2$$

for some constants τ and μ, and if $\{X_i : i \geq 1\}$ are i.i.d. copies of X and $\{Y_j : j \geq 1\}$ are i.i.d. copies of Y, then for any $i \neq j$,

$$\frac{1}{d^{1/2}} \left\{ \sum_{k=1}^{d} \left(X_i^{(k)} - X_j^{(k)} \right)^2 \right\}^{1/2} \to \left(2\sigma^2 \right)^{1/2}, \qquad \frac{1}{d^{1/2}} \left\{ \sum_{k=1}^{d} \left(Y_i^{(k)} - Y_j^{(k)} \right)^2 \right\}^{1/2} \to \left(2\tau^2 \right)^{1/2},$$

$$\frac{1}{d^{1/2}} \left\{ \sum_{k=1}^{d} \left(X_i^{(k)} - Y_j^{(k)} \right)^2 \right\}^{1/2} \to \left(\sigma^2 + \tau^2 + \mu^2 \right)^{1/2}.$$

In short,

$$\frac{1}{d} \|X_i - X_j\|_2^2 \to 2\sigma^2, \qquad \frac{1}{d} \|Y_i - Y_j\|_2^2 \to 2\tau^2, \qquad \frac{1}{d} \|X_i - Y_j\|_2^2 \to \sigma^2 + \tau^2 + \mu^2.$$

To illustrate this in words, $\|X_i - X_j\|_2$ is asymptotically of order $(2\sigma^2 d)^{1/2}$ when d is large and $i \neq j$. Similarly, $\|Y_i - Y_j\|_2$ and $\|X_i - Y_j\|_2$ are asymptotically of order $(2\tau^2 d)^{1/2}$ and $((\sigma^2 + \tau^2 + \mu^2)d)^{1/2}$ respectively when d is large. That is why classification of two class of points just based on L_2-distance between points is not effective in high-dimension, particularly when $\sigma = \tau$ and $\mu = 0$. Inter-point distance between points are approximately same in such special situation when $\mu = 0$ and $\sigma = \tau$ for large dimension d. This property of HDLSS helps us to explain plot (b) in Figure 3. Notice that in this setup, $S(t)$ depends on the average of coordinate-wise variation.

In the example (b) of Figure 3, points from groups G_1 and G_2 follow $N_{40}(\mathbf{0}_{40}, \Sigma)$ and $N_{40}(\mathbf{0}_{40}, \Gamma)$ respectively. So, $\sum_{k=1}^{d} \text{var} \left(X^{(k)} \right) = tr(\Sigma)$ and $\sum_{k=1}^{d} \text{var} \left(X^{(k)} \right) = tr(\Gamma)$ are equal

and

$$\frac{1}{d}\sum_{k=1}^{d}\left\{E\left(X^{(k)}\right)-E\left(Y^{(k)}\right)\right\}^2=0.$$

Hence, L_2-distances of within-group points $\|X_i-X_j\|_2$, $\|Y_i-Y_j\|_2$ and between-group points $\|X_i-Y_j\|_2$ are asymptotically equal as $d\to\infty$. And as a result, $S(t)$ fails to detect the change in distribution. On the other hand, if there is difference in location parameter with same covariance matrix, then $\sum_{k=1}^{d}\text{var}\left(X^{(k)}\right)=tr(\Sigma)$ and $\sum_{k=1}^{d}\text{var}\left(X^{(k)}\right)=tr(\Gamma)$ are equal but

$$\frac{1}{d}\sum_{k=1}^{d}\left\{E\left(X^{(k)}\right)-E\left(Y^{(k)}\right)\right\}^2>0.$$

Hence, L_2-distances of within-group points $\|X_i-X_j\|_2$ and $\|Y_i-Y_j\|_2$ are asymptotically equal, but since $\mu\neq 0$, $S(t)$ can detect the change in distribution. In the next section, we have used L_1-distance to circumvent this issue.

4 Introducing L_1 distance

Section 3 identifies shortcomings of using L_2-distance while constructing statistic $S(t)$ with observations in high-dimension using MDP or NND graphs. We can use L_1-distance instead to avoid this issue. Expected L_1-distance between two pints depend on their coordinate-wise marginal distributions. We consider the same statistic $S(t)$ as before, but this time the graph is constructed using L_1-distance; we call this new statistic $V(t)$. In the next section, we inspect the performance of this statistic $V(t)$. Before moving to simulations, we discuss some results relating the statistic $V(t)$.

Theorem 2. *If X and Y has distribution functions F and G respectively on $\mathbb{R}$, then*

$$E|X-Y|=\int_{-\infty}^{+\infty}F(x)(1-G(x))dx+\int_{-\infty}^{+\infty}G(x)(1-F(x))dx.$$

Next we use Theorem 1 and 2 to explore similar property for L_1-distance as in Result 1. Under usual conditions in Theorem 1 we can say that as $d\to\infty$,

$$\frac{1}{d}\sum_{k=1}^{d}\left|X_i^{(k)}-X_j^{(k)}\right|-\frac{1}{d}\sum_{k=1}^{d}\mathbf{E}\left|X_i^{(k)}-X_j^{(k)}\right|\to 0.$$

This motivates the following result

Result 2. *If $X=(X^{(1)},X^{(2)},\cdots)$ and $Y=(Y^{(1)},Y^{(2)},\cdots)$ with $F^{(k)},G^{(k)}$ as marginal*

distributions of $X^{(k)}, Y^{(k)}$ *respectively for* $1 \le k \le n$*, then for* $i \ne j$*,*

$$\frac{1}{d}\sum_{k=1}^{d}\left|X_i^{(k)} - X_j^{(k)}\right| - \Psi_d(F,F) \to 0 \qquad \frac{1}{d}\sum_{k=1}^{d}\left|Y_i^{(k)} - Y_j^{(k)}\right| - \Psi_d(G,G) \to 0$$

as $d \to \infty$*, where*

$$\Psi_d(F,G) = \frac{1}{d}\sum_{k=1}^{d}\mathbf{E}\left|X_i^{(k)} - Y_j^{(k)}\right| = \frac{1}{d}\sum_{k=1}^{d}\int_{-\infty}^{+\infty}\left[F^{(k)}(x) + G^{(k)}(x) - 2F^{(k)}(x)G^{(k)}(x)\right]dx.$$

Notice that $\Psi_d(F,G)$ compares distribution of every coordinates of X and Y. On the other hand, $S(t)$ used asymptotic average of coordinate-wise variation i.e., σ^2 and τ^2; it would detect change-point only when there is difference in average variation (while $\mu = 0$). That is why, in Figure 3, statistic $S(t)$ fails miserably when $\mu = 0$, but statistic $V(t)$ (using L_1-distance) works well (see Section 5).

To summarize, if $X = (X^{(1)}, X^{(2)}, \cdots, X^{(d)})$ and $Y = (Y^{(1)}, Y^{(2)}, \cdots, Y^{(d)})$ follow distributions $F = (F^{(1)}, F^{(2)}, \cdots, F^{(d)})$ and $G = (G^{(1)}, G^{(2)}, \cdots, G^{(d)})$ respectively, and X_i's, Y_i's are i.i.d. copies of X and Y respectively, then $\|X_i - X_j\|_1$, $\|Y_i - Y_j\|$ and $\|X_i - Y_j\|$ are asymptotically $\Psi_d(F,F), \Psi_d(G,G)$ and $\Psi_d(F,G)$ as $d \to \infty$. Note that $F^{(k)}$ is marginal distribution of k-th coordinate of X. Now our method can detect a change-point when these three values of Ψ are not same. Lets call these as A, B, C.

$$A = \Psi_d(F,F), \quad B = \Psi_d(G,G), \quad C = \Psi_d(F,G)$$

From the expression of A, B, C we can see that $C \ge (A+B)/2$ and equality holds if and only if the marginals of F and G are same, i.e., $F^{(k)} = G^{(k)}$ for all $1 \le k \le d$.

$$2\Psi_d(F,G) - \Psi_d(F,F) - \Psi_d(G,G) = \frac{2}{d}\sum_{k=1}^{d}\int_{-\infty}^{+\infty}\left[F^{(k)}(x) - G^{(k)}(x)\right]^2 dx \ge 0.$$

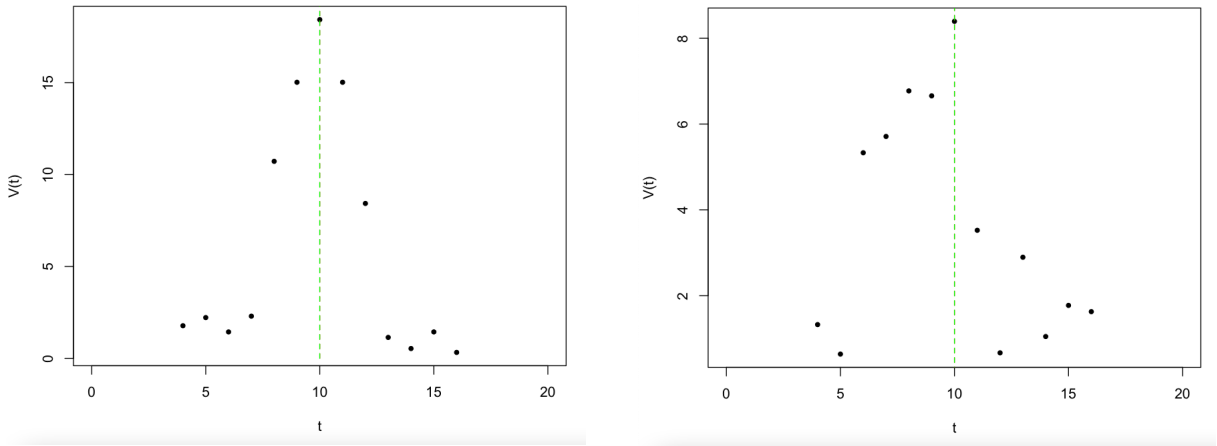
This implies that given the marginals of two distributions are different for at least one coordinate, we can detect the change-point. More importantly, under such condition, $C > (A+B)/2$ or $C > \min\{A, B\}$ (at least one of A and B is strictly less than C). This inequality is particularly important for detecting the change-point successfully.

Now we apply Kruskal's algorithm for creation of the Minimum Spanning Tree. To tell in short, Kruskal's algorithm keeps joining the shortest edge (edge with minimum weight) until a tree is formed including all the vertices by avoiding any cycle at the intermediate steps. This algorithm helps to understand how our statistic $V(t)$ depends on the MST. Under the condition of different marginal distributions, without losing generality, lets assume $A = \min\{A, B\}$. Then $A < C$ and $A < B$. No relation is specified between B and C. In such a scenario, while creating

the MST, the Kruskal's algorithm would join all the vertices that come from distribution F among themselves to create a tree (because all such points are at distance A from each other and A is the smallest possible edge-distance in the MST). And then it would connect other vertices with this tree to create the whole MST. This means all the vertices from group G_1 are connected to each other making the value of $R_(t)$ to be $t-1$, which is deviation from the expected value of $R_1(t)$. This deviation increases the value of the statistic $V(t)$. At the true change-point τ this deviation is maximum, and thus $V(t)$ has a maxima at $t = \tau$.

5 Some more simulations

This section is dedicated to simulations regrading statistic $V(t)$. In Figure 5 we have simulation for statistic $V(t)$ similar to what we saw in Figure 1. It shows that $V(t)$ is effective for location and scale shift for both $n > d$ and $n < d$ cases for normally distributed data points. In Figure 4, we test its performance under the scenario that we discussed in Section 3. Good performance of $V(t)$ is confirmed in all these occasions from these plots.



(a) (b)

Figure 4: (a) both location and scale shift: 10 observations generated from $N_{40}(\mathbf{0}_{40}, \Sigma)$ and 10 observations generated from $N_{40}(\mathbf{0}_{40}, \Gamma)$ for dimension $d = 40$, (b) pure scale shift: 10 observations generated from $N_{40}(\mathbf{0}_{40}, \Sigma)$ and 10 observations generated from $N_{40}(\mathbf{0}_{40}, \Gamma)$ for dimension $d = 40$.

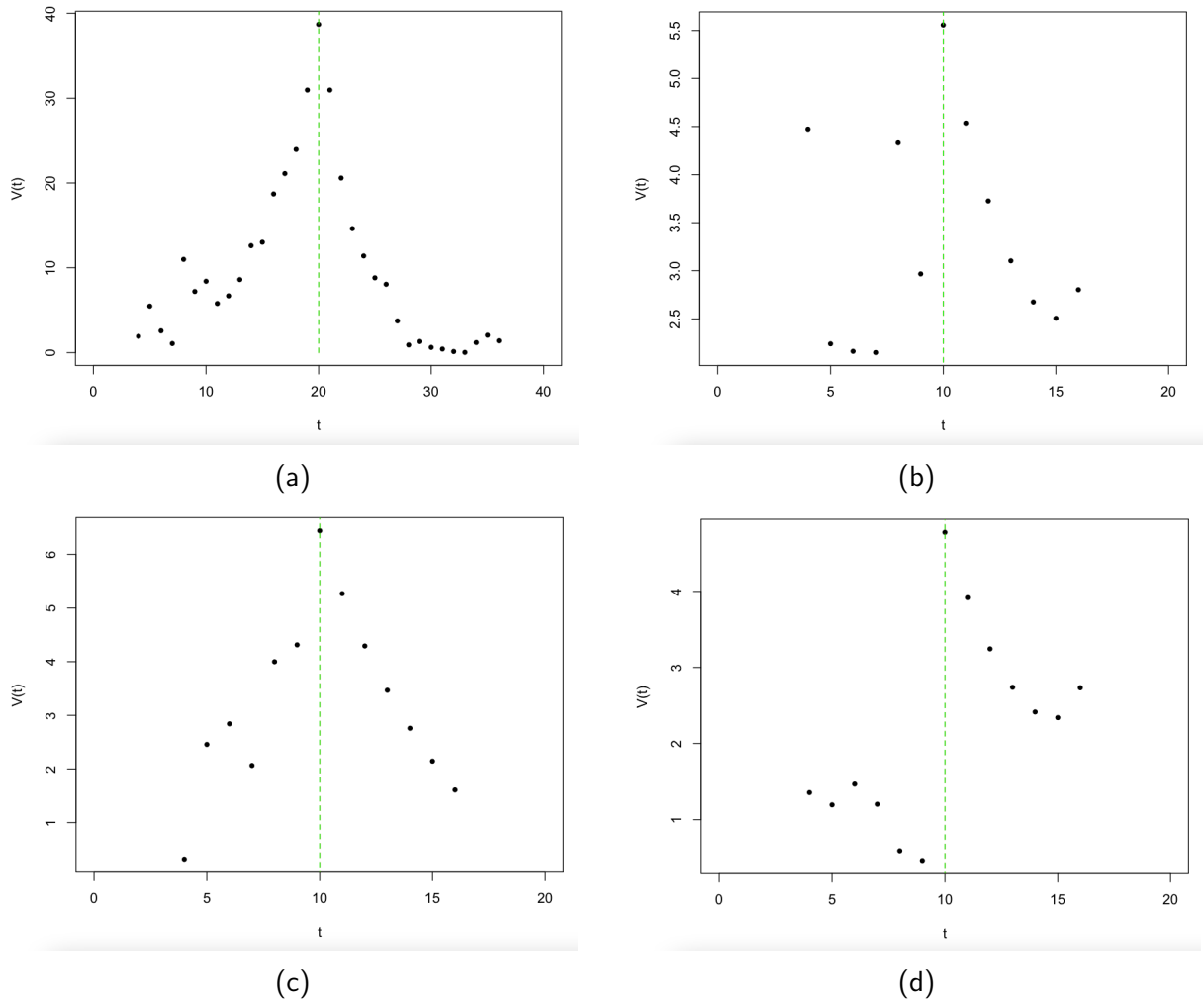


Figure 5: The vertical red and green lines show original and estimated change-points. (a) location shift when $n > d$: 50 observations generated from $N_{10}(\mathbf{0}_{10}, \boldsymbol{I}_{10})$ and 50 observations generated from $N_{10}(\mathbf{2}_{10}, \boldsymbol{I}_{10})$ for dimension $d = 10$, (b) scale shift when $n > d$: 50 observations generated from $N_{10}(\mathbf{0}_{10}, \boldsymbol{I}_{10})$ and 50 observations generated from $N_{10}(\mathbf{0}_{10}, 4\boldsymbol{I}_{10})$ for dimension $d = 10$, (c) location shift when $n < d$: 10 observations generated from $N_{50}(\mathbf{0}_{50}, \boldsymbol{I}_{50})$ and 10 observations generated from $N_{50}(\mathbf{2}_{50}, \boldsymbol{I}_{50})$ for dimension $d = 50$, (d) scale shift when $n < d$: 10 observations generated from $N_{50}(\mathbf{0}_{50}, \boldsymbol{I}_{50})$ and 10 observations generated from $N_{50}(\mathbf{0}_{50}, 4\boldsymbol{I}_{50})$ for dimension $d = 50$, where $\mathbf{2}_p$ is a vector of length p with all entries equal to 2 and $\boldsymbol{I}_p$ is identity matrix of size p.

6 Proofs

6.1 Proof of Theorem 1

Proof. Since, $\{X_i : i \geq 1\}$ has uniformly bounded second moments, hence, $\exists 0 < C < \infty$ such that $\text{Var}(X_i) < C, \forall i \geq 1$. Hence, for a fixed d,

$$\text{Var}\left(\frac{1}{d}\sum_{i=1}^{d} X_i\right) = \frac{1}{d^2}\sum_{i=1}^{d}\text{Var}(X_i) + \frac{1}{d^2}\sum_{q\neq q'}\left(\text{Cor}(X_q, X_{q'}) \times \sqrt{\text{Var}(X_q)} \times \sqrt{\text{Var}(X_{q'})}\right)$$
$$\leq \frac{C}{d} + \frac{C}{d^2}\sum_{q\neq q'}\text{Cor}(X_q, X_{q'})$$

Since, $\rho(r) \to 0$ as $r \to \infty$, for any $\epsilon > 0, \exists R_e \in \mathbb{N}$ such that $\rho(r) < \frac{\epsilon}{2C}$ for every $r \geq R_c$. Now, for $d > \frac{6CR_\varepsilon}{\varepsilon}$,

$$\text{Var}\left(\frac{1}{d}\sum_{i=1}^{d} X_i\right) \leq \frac{C}{d} + \frac{C}{d^2}\sum_{0<|q-q'|\leq R_e}\text{Cor}(X_q, X_{q'}) + \frac{C}{d^2}\sum_{|q-q'|>R_e}\text{Cor}(X_q, X_{q'})$$

Now, note that, for a fixed $1 \leq q \leq d$, we can have at most $2R_t$ many q' such that $0 < |q - q'| \leq R_c$. So, the second term in RHS can have at most $2dR_t$ many summands. Similarly, for a fixed $1 \leq q \leq d$, we can have at most d^2 many q' such that $|q - q'| \geq R_c$. Now, using the fact that $|\text{Cor}(X_q, X_{q'})| \leq 1$ and all the summands in the third term are $< \frac{\epsilon}{2C}$, we get,

$$\text{Var}\left(\frac{1}{d}\sum_{i=1}^{d} X_i\right) \leq \frac{C}{d} + \frac{2CR_e}{d} + \frac{\epsilon}{2}$$

which is lesser than ϵ by our choice of d. Since $\epsilon > 0$ was chosen arbitrarily, this implies, $\text{Var}\left(\frac{1}{d}\sum_{i=1}^{d} X_i\right) \overset{d\to\infty}{\longrightarrow} 0$. For a sequence of real random variables $\{Y_n : n \geq 1\}$ with 0 mean, if $\lim_{n\to\infty}\text{Var}(Y_n) = 0$, then, for any $\delta > 0$, by Chebyshev's Inequality,

$$\mathbb{P}(|Y_n| > \delta) \leq \frac{\text{Var}(Y_n)}{\delta^2} \overset{n\to\infty}{\longrightarrow} 0 \Rightarrow Y_n \xrightarrow{\text{P}} 0$$
$$\Rightarrow \left|\frac{1}{d}\sum_{i=1}^{d} X_i - \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}X_i\right| \xrightarrow{\text{P}} 0$$

This completes the proof. □

6.2 Proof of Result 1

Proof. If ρ-mixing property holds for U and V, then invoking Theorem 1 we get as $d \to \infty$

$$\frac{1}{d}\sum_{k=1}^{d}\left\{\left(U_i^{(k)} - V_j^{(k)}\right)^2 - \mathbf{E}\left(U_i^{(k)} - V_j^{(k)}\right)^2\right\} \to 0.$$

Also assuming for U and V

$$\frac{1}{d}\sum_{k=1}^{d} Var\left(U_i^{(k)}\right) \to \sigma^2, \qquad \frac{1}{d}\sum_{k=1}^{d} Var\left(V_j^{(k)}\right) \to \tau^2, \qquad \frac{1}{d}\sum_{k=1}^{d}\left\{\mathbf{E}\left(U_i^{(k)}\right) - \mathbf{E}\left(V_j^{(k)}\right)\right\}^2 \to \mu^2.$$

$$\begin{aligned}\frac{1}{d}\sum_{k=1}^{d}\mathbf{E}\left(U_i^{(k)} - V_j^{(k)}\right)^2 &= \frac{1}{d}\sum_{k=1}^{d} Var\left(U_i^{(k)}\right) + \frac{1}{d}\sum_{k=1}^{d} Var\left(V_j^{(k)}\right) + \frac{1}{d}\sum_{k=1}^{d}\left\{\mathbf{E}\left(U_i^{(k)}\right) - \mathbf{E}\left(V_j^{(k)}\right)\right\}^2 \\ &\to \sigma^2 + \tau^2 + \mu^2.\end{aligned}$$

Applying Slutsky's theorem we get

$$\frac{1}{d}\|U_i - V_j\|_2^2 = \frac{1}{d}\sum_{k=1}^{d}\left(U_i^{(k)} - V_j^{(k)}\right)^2 \to \sigma^2 + \tau^2 + \mu^2.$$

Now for $(U, V) = (X, X), (Y, Y), (X, Y)$, where $i \neq j$ if $(U, V) = (X, X)$ or $(U, V) = (Y, Y)$

$$\frac{1}{d}\|X_i - X_j\|_2^2 \to 2\sigma^2, \qquad \frac{1}{d}\|Y_i - Y_j\|_2^2 \to 2\tau^2, \qquad \frac{1}{d}\|X_i - Y_j\|_2^2 \to \sigma^2 + \tau^2 + \mu^2.$$

□

6.3 Proof of Theorem 2

Proof. Let $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$. We shall prove that $(X - Y)^+ = \int_{-\infty}^{\infty} I(X \leq u < Y)dx$ by proving that the events $\{(X - Y)^+ > c\}$ and $\{\int_{-\infty}^{\infty} I(X \leq u < Y)dx > c\}$ are same for any $c \in [0, \infty)$. Fix $c \geq 0$,

$$\begin{aligned}(X - Y)^+ \geq c &\implies X \geq Y + c \\ &\implies \int_{-\infty}^{\infty} I(X \geq u > Y)dx \geq \int_{-\infty}^{\infty} I(Y + c \geq u > Y)dx \geq \int_{0}^{c} I(Y + c \geq u > Y)dx = c\end{aligned}$$

and

$$(X-Y)^+ < c \implies Y > X - c$$
$$\implies \int_{-\infty}^{\infty} I(X \geq u > Y)dx < \int_{-\infty}^{\infty} I(X \geq u > X - c)dx = c.$$

This proves that the events $\{(X-Y)^+ > c\}$ and $\{\int_{-\infty}^{\infty} I(X \leq u < Y)dx > c\}$ are same for any $c \in [0, \infty)$. Similarly, for $(X-Y)^-$ we can show that

$$(X-Y)^- = (Y-X)^+ = \int_{-\infty}^{\infty} I(Y \leq u < X)dx.$$

Combining these we get

$$|X-Y| = (X-Y)^+ + (X-Y)^- = \int_{-\infty}^{\infty} [I(X \leq u < Y) + I(Y \leq u < X)]dx.$$

If we assume that $X \sim F$ and $Y \sim G$ independently. So taking expectation on both sides and using Fubini's theorem we have

$$E|X-Y| = \int_{-\infty}^{+\infty} F(x)(1-G(x))dx + \int_{-\infty}^{+\infty} G(x)(1-F(x))dx.$$

□

References

Arlot, S., Celisse, A., and Harchaoui, Z. (2019). A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162).

Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association*, 112(517):397–409.

Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, 43(1):139–176.

Chu, L. and Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. *The Annals of Statistics*, 47(1):382–414.

Desobry, F., Davy, M., and Doncarli, C. (2005). An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444.

Harchaoui, Z., Moulines, E., and Bach, F. R. (2009). Kernel change-point analysis. In *Advances in neural information processing systems*, pages 609–616.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783.

Li, S., Xie, Y., Dai, H., and Song, L. (2015). Scan b-statistic for kernel change-point detection. *arXiv preprint arXiv:1507.01279*.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

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(4):515–530.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806.

Shi, X., Wu, Y., and Rao, C. R. (2017). Consistent and powerful graph-based change-point test for high-dimensional data. *Proceedings of the National Academy of Sciences*, 114(15):3873–3878.

Sun, Y.-W., Papagiannouli, K., and Spokoiny, V. (2019). Online graph-based change-point detection for high dimensional data. *arXiv preprint arXiv:1906.03001*.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272.