Change-point detection in high-dimensional data

Dissertation midterm report of Soham Das (Roll - MB2007)

February 21, 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. This project concentrates on analyzing some of these methods for single change-point problem, particularly the issues they face while detecting scale shifts in distribution in a sequence of multivariate observations in high-dimension. We have proposed some modifications to existing methods to resolve this issue and introduced a new statistic. We also discussed their theoretical and empirical performance for high-dimensional data.

1 Introduction and problem description

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 $\{Z_i : i = 1, \ldots, Z_n\}$ be a sequence of observations from $\mathbb{R}^d$, $d \geq 1$. Let

$$\begin{aligned} Z_i &\sim F \quad \text{for } i = 1, 2, \cdots, \tau \\ \text{and } Z_i &\sim G \quad \text{for } i = \tau + 1, \cdots, n \end{aligned}$$

where F and G are two different distributions and $1 < \tau < n$. We also assume that Z_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

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

$$H_0 : \mathrm{Z}_i \sim F, i = 1, \ldots, n$$

against the single change-point alternative

$$H_1 : \exists\ 1 < \tau < n, \mathrm{Z}_i \sim \begin{cases} F, & i \leq \tau \\ G, & \text{otherwise} \end{cases}$$

We shall discuss about performances of two approaches regarding detection of such change-point in high-dimension. One approach is by using distance-based method introduced in Chu and Chen (2019). And the other approach is by using distance-based method that is mentioned in Matteson and James (2014). Although the statistics proposed in these aricles work quite well in low dimension, they fail in some occasions in higher dimension but low sample size (HDLSS) scenario. In this report we shall explore the properties and performances of some statistics particularly in high-dimension and discuss about suitable conditions they need to perform efficiently. Section 2 provides some drawbacks of two existing statistics 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)). In section 4 we propose some statistics for testing and estimation of change-point. Also consistency of the estimators of change-point location are proved here. Section 5 provides some more simulations to support the new statistics. Section 6 summarizes the topics covered in this report and gives some outlines of future plan. Section 7 contains all proofs of theorems and results used in this report.

2 Existing methods and their performance in high-dimension

Matteson and James (2014) used an empirical divergence measure which for our problem can be written as

$$D(t) = \frac{2}{t(n-t)} \sum_{i=1}^{t} \sum_{j=t+1}^{n} \frac{1}{d} \|Z_i - Z_j\| - \frac{1}{t(t-1)} \sum_{i=1}^{t} \sum_{j=1}^{t} \frac{1}{d} \|Z_i - Z_j\| - \frac{1}{(n-t)(n-t-1)} \sum_{i=t+1}^{n} \sum_{j=t+1}^{n} \frac{1}{d} \|Z_i - Z_j\|$$

where $\|\cdot\|$ is a distance metric. Informally speaking, when there is a change-point at index τ, then this

statistic would have higher positive value around τ for usual metrics like L_1 and L_2. Since in our case n is finite and d is large, a scan statistic based on this measure can be useful to estimate and test for change-point. We call the three terms of $D(t)$ as $2T_3$, T_1 and T_2, these are the averages of inter-point distances. In the article, $D(t) = 2T_3(t) - T_1(t) - T_2(t)$ measure was used to detect change-point.

A graph-based method was proposed in Chu and Chen (2019), where the authors introduced a scan statistic, say $V(t)$, based on minimum spanning tree(MST) or nearest neighbour graph(NNG). For a sequence of observations $Z_1, Z_2, \cdots, Z_n$, consider a change-point at index t where $1 < t < n$, and split n observations $\{z_i : i = 1, \cdots, n\}$ into two groups $\{z_1, \cdots, z_t\}$ and $\{z_{t+1}, \cdots, z_n\}$. Call these two groups as G_1 and G_2. Now construct an MST or NNG graph using some distance metric by considering n observations as n vertices of this graph. Define $R_i(t)$ as the number of edges between vertices that belong to same group G_i for each $i = 1, 2$. And define a measure as:

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

where $(\mathbf{E}(R_1(t)), \mathbf{E}(R_2(t)))$ and $\Sigma(t)$ are expectation and covariance matrix of $(R_1(t), R_2(t))$. Note that in this case also, the statistic $G(t)$ is maximum at $t = \tau$.

So for both these methods, we can use the scan statistic $\max_{1<t<n} D(t)$ and $\max_{1<t<n} G(t)$ to detect the change-point. Note that, under H_0 when all observations come from same distribution F, $D(t)$ and $G(t)$ should have same distribution for any $t \in \{1, \cdots, n\}$, but under H_1, they are maximized at the true value of change-point τ. We do a permutation test for the existence of change-point. 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 using the permutation distribution that places $1/n!$ probability on each of the $n!$ permutations of $\{z_i : i = 1, \cdots, n\}$, we simulate large number of permutations of z_i's and calculate the statistic to get its permutation null distribution. We use this null distribution to calculate the p-value for the observed data under H_0. If the test is significant, i.e., there is significant evidence for the existence of change-point at some location in the observed data, then we estimate the location of change-point by $\hat{\tau} = \arg\max_{1<t<n} G(t)$ and $\hat{\tau} = \arg\max_{1<t<n} D(t)$, and p-value of this estimate is calculated under permutation null. Need to mention that when t is close to 1 or n, then the statistic behaves erratically. So in order to have a better estimate, we ignore p^* portion from both ends while scanning through 1 to n. Now the statistic would look like $\underset{p^*<t<n(1-p^*)}{\operatorname{argmax}} D(t)$ for some small fraction p^*.

To summarize, first we test the existence of the change-point using permutation test on the scan statistic, and then if the test is significant, we estimate the change-point using that scan statistic for observed data. If the test is not significant then we conclude that there is not much evidence for the existence of change-point in the data.

Our area of interest is detection of change-point particularly in high-dimension. Although these methods work quite well for usual low-dimensional case (when $n > d$ i.e., sample size is greater than the dimension), in high-dimension but low sample-size (when $n < d$) setup, their performances are not sound and in some cases they are really bad. Next we have presented a simulation based performance

check for these methods in HDLSS setup.

We shall mention two interesting scenarios to identify the drawbacks of the aforementioned statistics: one is when there is change in scale parameter and the other one is when there is change in distribution but means and variances are same. Let $Z_i \sim F$ for $1 \leq i \leq \tau$ and $Z_i \sim G$ for $\tau + 1 \leq i \leq n$.
Scenario 1: Take $F \equiv N_d(\mathbf{0}, \boldsymbol{I})$ and $G \equiv N_d(\mathbf{0}, 2\boldsymbol{I})$ where $n = 30$, $d = 100$.
Scenario 2: Take $F \equiv N_d(\mathbf{0}, \boldsymbol{I})$ and each coordinates of G is $\frac{1}{\sqrt{3}} t_3$ (student's t-distribution with 3 degrees of freedom) independently where $n = 30$, $d = 100$.

Distributions	Distance-based method	Graph-based method
$N(\mathbf{0}, \boldsymbol{I})$ vs $N(\mathbf{0.5}, \boldsymbol{I})$	1.00	0.92
$N(\mathbf{0}, \boldsymbol{I})$ vs $N(\mathbf{0}, 2\boldsymbol{I})$	0.49	0.53
$N(\mathbf{0}, \boldsymbol{I})$ vs $\frac{1}{\sqrt{3}} t_3(\mathbf{0}, \boldsymbol{I})$	0.06	0.10

Table 1: Although the distance-based and graph-based methods working properly in both testing and estimation for location shift, these tests have very low power in high-dimension while countering scale shifts. Also when there are two distributions with same mean and variance then these tests have low power. In such cases it was also observed that estimations of change-point location by using these statistics are not good either.

Powers of two tests indicate that the above statistics do not work well for some situations in high-dimension. Note that in the scenario 2, two distributions F and G are taken in such a way that average of coordinate-wise variances are same for F and G.

3 Properties of data points in high dimension

In this section, we explain why the measures $D(t)$ and $G(t)$ failed for low sample size in high-dimension. This happened because of certain properties of points in high-dimension. 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*

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

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 proposition from Hall et al. (2005).

Proposition 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 $\left(2\sigma^2 d\right)^{1/2}$ 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 Proposition 1 along with some more assumptions that

$$\frac{1}{d}\sum_{k=1}^{d} \mathrm{var}\left(X^{(k)}\right) \to \sigma_1^2, \qquad \frac{1}{d}\sum_{k=1}^{d} \mathrm{var}\left(Y^{(k)}\right) \to \sigma_2^2 \qquad \text{and} \qquad \frac{1}{d}\sum_{k=1}^{d}\left\{E\left(X^{(k)}\right) - E\left(Y^{(k)}\right)\right\}^2 \to \mu^2$$

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 $i \neq j$, In short,

$$\frac{1}{d}\|X_i - X_j\|_2^2 \to 2\sigma_1^2, \qquad \frac{1}{d}\|Y_i - Y_j\|_2^2 \to 2\sigma_2^2, \qquad \frac{1}{d}\|X_i - Y_j\|_2^2 \to \mu^2 + \sigma_1^2 + \sigma_2^2.$$

All convergences are convergence in probability. To illustrate this in words, $\|X_i - X_j\|_2$ is asymptotically of order $(2\sigma_1^2 d)^{1/2}$ when d is large and $i \neq j$. Similarly, $\|Y_i - Y_j\|_2$ and $\|X_i - Y_j\|_2$ are asymptotically of order $(2\sigma_2^2 d)^{1/2}$ and $\left((\mu^2 + \sigma_1^2 + \sigma_2^2)d\right)^{1/2}$ respectively when d is large. Distances between any pair of points are approximately same in above two scenarios when $\mu = 0$ and $\sigma = \tau$ for large dimension d. As a result, graph-based method cannot distinguish between two classes, and the measure $G(t)$ do not work. Similarly when $\mu = 0$, distance-based measure $D(t) \to 2\sqrt{\mu^2 + \sigma_1^2 + \sigma_2^2} - \sqrt{2}\sigma_1 - \sqrt{2}\sigma_2 = 0$ at $t = \tau$ as $d \to \infty$, and fails to detect change-point. This property of HDLSS helps us to explain Table 1. Notice that in these examples, $G(t)$ and $D(t)$ depend on the average of coordinate-wise variation.

Next we explore general asymptotic property of distance-based statistic. Let us denote

$$T_1 = \frac{1}{t(t-1)}\sum_{i=1}^{t}\sum_{j=1}^{t}\frac{1}{d^{1/p}}\|Z_i - Z_j\|_p \qquad T_2 = \frac{1}{(n-t)(n-t-1)}\sum_{i=t+1}^{n}\sum_{j=t+1}^{n}\frac{1}{d^{1/p}}\|Z_i - Z_j\|_p$$

$$T_3 = \frac{1}{t(n-t)}\sum_{i=1}^{t}\sum_{j=t+1}^{n}\frac{1}{d^{1/p}}\|Z_i - Z_j\|_p$$

where $\|\cdot\|_p$ is the usual L_p norm, $\|x\|_p^p = \sum_{k=1}^{d}|x^{(k)}|^p$ for d-dimensional vector $x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$

and $p = 1, 2$. Also denote

$$\delta_{FG} = \|X_i - Y_j\|_p \qquad \delta_{FF} = \|X_i - X_j\|_p \qquad \delta_{GG} = \|Y_i - Y_j\|_p.$$

where $X_i, X_j \sim F$ and $Y_i, Y_j \sim G$ independently. From Theorem 1 we know

$$T_1 - \frac{1}{d^{1/p}}\delta_{FF} \xrightarrow{P} 0 \qquad T_2 - \frac{1}{d^{1/p}}\delta_{GG} \xrightarrow{P} 0 \qquad T_3 - \frac{1}{d^{1/p}}\delta_{FG} \xrightarrow{P} 0$$

Clearly, $\delta_{FG}, \delta_{FF}, \delta_{GG}$ would depend on the L_p-norm.

4 Main results

In this section we propose a new distance-based statistics based on T_1, T_2, T_3 for both L_1 and L_2 norms in addition to the previous statistic $D(t)$. Consider following statistics

Statistic 1: $\max_{1<t<n} S_1(t)$ where $S_1(t) = 2T_3(t) - T_1(t) - T_2(t)$

Statistic 2: $\max_{1<t<n} S_2(t)$ where $S_2(t) = |T_3(t) - T_1(t)| + |T_3(t) - T_2(t)|$

Note that whatever statistic we consider as favorable, procedure of testing and estimation for change-point remains same: first test for change-point using permutation test using scan statistic $\widehat{\tau} = \max_t S_i(t)$, $i = 1, 2$ as mentioned in Section 2 and then estimate location of change-point.

Functionality of these two statistics depends on 'signals' that are different for different metric, but the forms of these signals are same for each statistic. $(2\delta_{FG} - \delta_{FF} - \delta_{GG})$ and $|\ \delta_{FG} - \delta_{FF}\ | + |\ \delta_{FG} - \delta_{GG}\ |$ are the signals for S_1 and S_2 respectively. Roughly speaking, if these signals are positive we can successfully use corresponding statistic. In that case, finding suitable conditions for making these signals positive should be our goal.

In order to use Theorem 1 we need the moment assumptions of the distributions.

(A1) Distributions F and G in $\mathbb{R}^d$ are such that they have uniformly bounded fourth moment for all the coordinates and ρ-mixing property holds for the coordinates of Z when $Z \sim F$ or G.

(A2) Distributions F and G in $\mathbb{R}^d$ are such that they have uniformly bounded second moment for all the coordinates and ρ-mixing property holds for the coordinates of Z when $Z \sim F$ or G.

4.1 Assumptions for L_2 norm

Under the general assumption **(A1)**, the signals are needed to be positive. From Section 3 we know that when $X_i, X_j \sim F$ and $Y_i, Y_j \sim G$

$$\frac{1}{d}\|X_i - X_j\|_2^2 \xrightarrow{P} 2\sigma_1^2, \qquad \frac{1}{d}\|Y_i - Y_j\|_2^2 \xrightarrow{P} 2\sigma_2^2, \qquad \frac{1}{d}\|X_i - Y_j\|_2^2 \xrightarrow{P} \mu^2 + \sigma_1^2 + \sigma_2^2.$$

where μ is the average of differences in means and σ_1, σ_2 are standard deviations of distributions F and G. Assumption **(A1)** ensures that they exists finitely. So signal for S_1,

$$2\delta_{FG} - \delta_{FF} - \delta_{GG} = 2\sqrt{\mu^2 + \sigma_1^2 + \sigma_2^2} - \sqrt{2}\sigma_1 - \sqrt{2}\sigma_2 \geq 0$$

where equality holds if and only if $\sigma_1 = \sigma_2$ and $\mu = 0$. That means S_1 with L_2 norm is effective unless the mean and average variances over all coordinates for distributions F and G are same.
Signal for S_2,

$$|\delta_{FG} - \delta_{FF}| + |\delta_{FG} - \delta_{GG}| = \left|\sqrt{\mu^2 + \sigma_1^2 + \sigma_2^2} - \sqrt{2}\sigma_1\right| + \left|\sqrt{\mu^2 + \sigma_1^2 + \sigma_2^2} - \sqrt{2}\sigma_2\right| \geq 0$$

where equality holds if and only if $\mu = 0$ and $\sigma_1 = \sigma_2$.
(A3) F and G are two d-variate distributions such that coordinate-wise means or average variances over all coordinates for distributions F and G are not same.
Condition for L_2: Under assumption **(A1)** for L_2 distance based statistics, **(A3)** is a necessary and sufficient condition to detect change-point using S_1 and S_2.

In fact, S_2 works better than S_1 since

$$|2\delta_{FG} - \delta_{FF} - \delta_{FF}| \leq |\delta_{FG} - \delta_{FF}| + |\delta_{FG} - \delta_{FF}| \tag{1}$$

But the hurdle mentioned in Section 2 still persists when we use L_2 distance - it performs badly for scale shifts and it needs stronger assumption like **(A3)** to work properly. Using L_1 distance instead of L_2 can have some improvements. In case of L_1 distance, the signals are positive even for weaker conditions than **(A3)**.

4.2 Assumptions for L_1 norm

We need to find asymptotic properties for L_1 similar to what we did in Section 3 for L_2. Lets start with the following proposition.

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

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

We use Theorem 1 and Proposition 3 to have asymptotic property for L_1-distance in the following result.

Proposition 3. *Under assumption **(A2)** if $X = (X^{(1)}, X^{(2)}, \cdots) \sim F$ and $Y = (Y^{(1)}, Y^{(2)}, \cdots) \sim G$ with $F^{(k)}, G^{(k)}$ as marginal distributions of $X^{(k)}, Y^{(k)}$ respectively for $1 \leq k \leq n$, then*

$$\frac{1}{d}\sum_{k=1}^{d}\left|X^{(k)} - Y^{(k)}\right| - \frac{1}{d}\delta_{FG} \to 0$$

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

$$\delta_{FG} = \sum_{k=1}^{d} \mathbf{E}\left|X^{(k)} - Y^{(k)}\right| = \sum_{k=1}^{d} \int_{-\infty}^{+\infty} \left[F^{(k)}(x) + G^{(k)}(x) - 2F^{(k)}(x)G^{(k)}(x)\right] dx.$$

Notice that δ_{FG} involves marginal distribution of every coordinates of X and Y. This proposition immediately implies a few relations for δ_{FG}.

$$2\delta_{FG} - \delta_{FF} - \delta_{GG} = \sum_{k=1}^{d} \int_{-\infty}^{+\infty} 2\left[F^{(k)}(x) - G^{(k)}(x)\right]^2 dx \geq 0.$$

where equality holds if and only if $F^{(k)} = G^{(k)}$ for every $1 \leq k \leq d$. So the signal for S_1 using L_1 distance is positive when the coordinate-wise marginal distributions are not same. Clearly, as mentioned before, when this condition holds, then the signal for S_1 is positive, and the statistic can be used to detect change-point.

Also we have

$$\delta_{FG} - \delta_{FF} = \sum_{k=1}^{d} \int_{-\infty}^{+\infty} \left[G^{(k)}(x) - F^{(k)}(x)\right]\left[1 - 2F^{(k)}(x)\right] dx$$

$$\delta_{FG} - \delta_{GG} = \sum_{k=1}^{d} \int_{-\infty}^{+\infty} \left[F^{(k)}(x) - G^{(k)}(x)\right]\left[1 - 2G^{(k)}(x)\right] dx$$

Inequality (1) confirms that signal for S_2 is positive and sometimes better than that of S_1. This implies S_2 works when marginal distributions are not same.

(A4) F and G are two d-variate distributions such that for every coordinate medians of marginal distributions for F and G are same but scale parameters are different.

Condition for L_1**:** Under assumption **(A2)** for L_1 distance based statistics, **(A4)** is a necessary and sufficient condition to detect change-point using S_1 and S_2.

To summarize, we have learnt that S_2 works better than S_1 and using L_1 distance requires less condition on underlying distributions than that of L_2. So using S_2 with L_1 gives best testing and estimation results so far. Relative performances of these statistics have been confirmed through simulations in Section 5.

4.3 Consistency

Theorem 2. *If* τ *is the original change point and the estimated change-point is* $\widehat{\tau} = \operatorname{argmax} S_i(t)$ *then under respective assumptions for all four statistics* $\widehat{\tau} \xrightarrow{P} \tau$ *as* $d \to \infty$ *for* $i = 1, 2$.

While proving consistency we can understand that the signals for the statistics are at the core of estimation of $\widehat{\tau}$. That is why a stronger signal always results in a better estimation.

Also in Section 5 we have empirically shown the consistency of the tests for L_1 distance.

5 Some more simulations

Distributions	Max $S_1(t)$	Max $S_2(t)$
$\mathbf{N}(\mathbf{0}, \boldsymbol{I})$ vs $\mathbf{N}(\mathbf{0.5}, \boldsymbol{I})$	1.000	0.975
$\mathbf{N}(\mathbf{0}, \boldsymbol{I})$ vs $\mathbf{N}(\mathbf{0}, 2\boldsymbol{I})$	0.445	1.000
$\mathbf{N}(\mathbf{0}, \boldsymbol{I})$ vs $\frac{1}{\sqrt{3}} t_3(\mathbf{0}, \boldsymbol{I})$	0.330	0.880

Table 2: This table presents relative performance of S_1 and S_2 using L_1 distance. We have taken dimension $d = 100$, sample size $n = 30$. Considered 200 simulations in each case. According to the table both S_1 and S_2 performs better for L_1 than L_2.

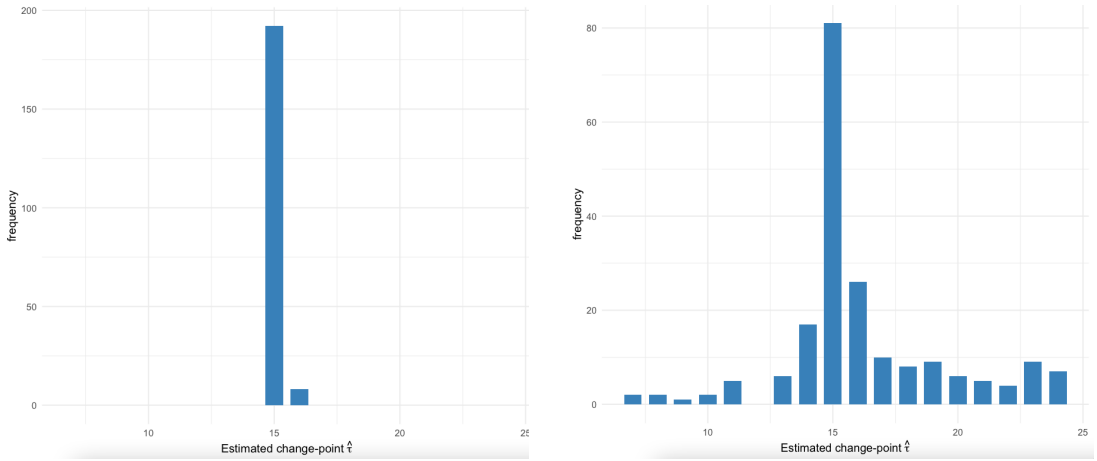

Figure 1: Barplot of estimated change-point location $\hat{\tau}$ for S_2 using L_1 distance. Here dimension $d = 100$, sample size $n = 30$. Two distributions are $F \equiv \mathsf{N}(\mathbf{0}, \boldsymbol{I})$ and $G \equiv \mathsf{N}(\mathbf{0}, 2\boldsymbol{I})$ (for the plot on the left) and $F \equiv \mathsf{N}(\mathbf{0}, \boldsymbol{I})$ and $G \equiv \frac{1}{\sqrt{3}} t_3(\mathbf{0}, \boldsymbol{I})$ (for the plot on the right). In both cases, S_1 estimates τ very well.

Distributions	Statistic	$d = 50$	$d = 100$	$d = 200$	$d = 500$	$d = 1000$
$\mathbf{N}(\mathbf{0}, \boldsymbol{I})$ vs $\mathbf{N}(\mathbf{0}, 2\boldsymbol{I})$	$S_1(t)$	0.019	0.410	0.695	0.995	1.000
	$S_2(t)$	1.000	1.000	1.000	1.000	1.000
$\mathbf{N}(\mathbf{0}, \boldsymbol{I})$ vs $\frac{1}{\sqrt{3}} t_3(\mathbf{0}, \boldsymbol{I})$	$S_1(t)$	0.235	0.300	0.635	0.950	1.000
	$S_2(t)$	0.640	0.885	0.980	1.000	1.000

Table 3: This table provides empirical evidence of consistency for the tests using S_1 and S_2 with L_1 distance. Here sample size $n = 30$ and considered 200 simulations in each case for the calculation of p-values.

6 Summary of report and an outline of future work

In this report, we have addressed the problem of using existing L_2-distance based statistic in single change-point detection problem in high-dimension when sample size n is less than the dimension d. We have introduced another statistic which performs better than the previous one, and proved their consistency in estimating the location of change-point both for L_1 and L_2 distances. We have showed that using L_1-distance is better since it requires weaker assumptions on the underlying distributions. This makes L_1-distance based statistics uniformly better-performing than the L_2 distance based ones.

At the same time we also need to mention that moment finiteness assumptions for both L_1 and L_2 distances are required for these methods. A possible way-out would be using metrics that are uniformly bounded. In that case, we do not need moment assumptions like **(A1)** or **(A2)**. In high-dimensional $n < d$ setup, we have considered the case where $d \to \infty$ keeping n fixed. We can also try the case where both d and n goes to infinity such that d/n remains bounded.

7 Proofs

7.1 Proof of Theorem 1

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

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

From ρ-mixing property we have $\rho(r) \to 0$ as $r \to \infty$, hence for any $\varepsilon > 0$, $\exists R_\varepsilon \in \mathbb{N}$ such that $\rho(r) < \frac{\varepsilon}{2M}$ for every $r \geq R_\varepsilon$. Now, for $d > \dfrac{6MR_\varepsilon}{\varepsilon}$,

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

Now for a fixed $1 \leq q \leq d$, we can have at most $2R_\varepsilon$ many q' such that $0 < |q - q'| \leq R_\varepsilon$. So the second term in right-hand side can have at most $2dR_\varepsilon$ many summands. Similarly, for a fixed $1 \leq q \leq d$, we can have at most d^2 many q' such that $|q - q'| \geq R_\varepsilon$. Now using the fact that $\left|\mathrm{Cor}\left(X_q, X_{q'}\right)\right| \leq 1$ for the second term and that every summand in the third term is less than $\frac{\varepsilon}{2M}$, we get

$$\mathrm{Var}\left(\frac{1}{d}\sum_{i=1}^{d} X_i\right) \leq \frac{M}{d} + \frac{2MR_\varepsilon}{d} + \frac{\varepsilon}{2}$$

We can take large d to show that the right-hand side is less than ε. Since $\varepsilon > 0$ is arbitrary, this implies $\mathrm{Var}\left(\frac{1}{d}\sum_{i=1}^{d} X_i\right) \longrightarrow 0$ as $d \to \infty$. For a sequence of real random variables $\{Y_n : n \geq 1\}$ with 0 mean, if $\lim_{n\to\infty} \mathrm{Var}\,(Y_n) = 0$, then for any $\delta > 0$, by Chebyshev's Inequality,

$$\mathbb{P}\left(|Y_n| > \delta\right) \leq \frac{\mathrm{Var}\,(Y_n)}{\delta^2} \longrightarrow 0 \qquad \Rightarrow |Y_n| \xrightarrow{\mathrm{P}} 0$$

as $n \to \infty$. We can apply this on $\frac{1}{d}\sum_{i=1}^{d} X_i$ to show

$$\left|\frac{1}{d}\sum_{i=1}^{d} X_i - \frac{1}{d}\sum_{i=1}^{d} \mathbb{E}X_i\right| \xrightarrow{\mathrm{P}} 0$$

This completes our proof. □

7.2 Proof of Proposition 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

$$\frac{1}{d}\sum_{k=1}^{d} Var\left(U_i^{(k)}\right) \to \sigma_1^2, \qquad \frac{1}{d}\sum_{k=1}^{d} Var\left(V_j^{(k)}\right) \to \sigma_2^2, \qquad \frac{1}{d}\sum_{k=1}^{d}\left\{\mathbf{E}\left(U_i^{(k)}\right) - \mathbf{E}\left(V_j^{(k)}\right)\right\}^2 \to \mu^2.$$

$$\begin{aligned}\frac{1}{d}\sum_{k=1}^{d}\mathbf{E}\left(U_i^{(k)} - V_j^{(k)}\right)^2 &= \frac{1}{d}\sum_{k=1}^{d} Var\left(U_i^{(k)}\right) + \frac{1}{d}\sum_{k=1}^{d} Var\left(V_j^{(k)}\right) + \frac{1}{d}\sum_{k=1}^{d}\left\{\mathbf{E}\left(U_i^{(k)}\right) - \mathbf{E}\left(V_j^{(k)}\right)\right\}^2 \\ &\to \sigma_1^2 + \sigma_2^2 + \mu^2.\end{aligned}$$

Applying Slutsky's theorem we get

$$\frac{1}{d}\|U_i - V_j\|_2^2 = \frac{1}{d}\sum_{k=1}^{d}\left(U_i^{(k)} - V_j^{(k)}\right)^2 \to \sigma_1^2 + \sigma_2^2 + \mu^2.$$

□

7.3 Proof of Proposition 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 $c \geq 0$.

$$(X-Y)^+ \geq c \implies X \geq Y + c$$
$$\implies \int_{-\infty}^{\infty} I(X \geq u > Y)dx \geq \int_{-\infty}^{\infty} I(Y + c \geq u > Y)dx \geq \int_0^c I(Y + c \geq u > Y)dx = c$$

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

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

□

7.4 Proof of Theorem 2

Proof. We have sample of size n with original change-point at τ. Suppose for statistic $S(t)$ we have to show that $\widehat{\tau} = \text{argmax}_t\, S(t) \to \tau$ as $d \to \infty$. For every $t \in \{1, \ldots, n\}$ denote the limiting values of $S(t)$ by $E(t)$ such that $S(t) - E(t) \xrightarrow{P} 0$ as $d \to \infty$. Now we use the lemma on convergence of maxima: if $S(t)$ converges to $E(t)$ uniformly for all t then $\text{argmax}_t\, S(t) \to \text{argmax}_t\, E(t)$. Interestingly, since our t takes finitely many values, proving uniform convergence reduces to proving simple pointwise convergence. Also that since we are taking E to be the pointwise limit of statistic S, it is just enough to show that τ is the maxima of $E(t)$.

Note that the proof is consistent with both the metric L_1 and L_2. So we are just presenting a general proof that is applicable to both. Also we have not included the factor $\frac{1}{d^{1/p}}$ in the expression of $S(t)$ to keep the proofs simple (otherwise we should write

$d^{1/p}S(t)$ in each step instead of just $S(t)$, where $p = 1$ or 2 whichever applicable).

7.4.1 Consistency for S_1

Fix $t \in \{1, \ldots, n\}$. If test for existence of change-point is significant, we assume alternative hypothesis is true, i.e., τ is the original change-point and point before and after τ have different distributions, namely F and G. Now we divide the observations at t to compute statistic $S_1(t)$. Now we break this into two cases: one is $t \leq \tau$ and another one is $t \geq \tau$.
Case 1: When $t \leq \tau$, then as $d \to \infty$,

$$\begin{aligned} S_1(t) &= 2T_3(t) - T_1(t) - T_2(t) \\ &\xrightarrow{P} \frac{2}{t(n-t)}\left[(t(\tau-t)\delta_{FF} + t(n-\tau)\delta_{FG}\right] - \delta_{FF} \\ &\quad - \frac{1}{(n-t)(n-t-1)}\left[(n-\tau)(n-\tau-1)\delta_{GG} + (\tau-t)(\tau-t-1)\delta_{FF} + 2(n-\tau)(\tau-t)\delta_{FG}\right] \\ &= \frac{(n-\tau)(n-\tau-1)}{(n-t)(n-t-1)}\left[2\delta_{FG} - \delta_{FF} - \delta_{GG}\right] = E_1(t) \end{aligned}$$

Note that the last expression $E_1(t)$ is non-negative for all $t \leq \tau$ and it is strictly increasing. So $E_1(t) < E_1(\tau)$ for all $t < \tau$.
Case 2: When $t \geq \tau$, then as $d \to \infty$,

$$\begin{aligned} S_1(t) = 2T_3(t) - T_1(t) - T_2(t) &\xrightarrow{P} \frac{2}{t(n-t)}\left[(n-t)(t-\tau)\delta_{GG} + \tau(n-t)\delta_{FG}\right] - \delta_{GG} \\ &\quad - \frac{1}{t(t-1)}\left[(t-\tau)(t-\tau-1)\delta_{GG} + \tau(\tau-1)\delta_{FF} + 2\tau(t-\tau)\delta_{FG}\right] \\ &= \frac{\tau(\tau-1)}{t(t-1)}\left[2\delta_{FG} - \delta_{FF} - \delta_{GG}\right] = E_1(t) \end{aligned}$$

Here $E_1(t)$ is non-negative for all $t \geq \tau$ and it is strictly decreasing. So $E_1(t) < E_1(\tau)$ for all $t > \tau$.

This completes the proof of consistency of estimator for S_1.

7.4.2 Consistency for S_2

Proof of this one is similar to the last one. Fix $t \in \{1, \ldots, n\}$. If τ is the original change-point and point before and after τ have different distributions F and G. We divide the observations at t to compute statistic $S_2(t)$.

Case 1: When $t \leq \tau$, then as $d \to \infty$,

$$
\begin{aligned}
|T_3(t) - T_1(t)| &\xrightarrow{P} \left| \frac{1}{t(n-t)} \left[(t(\tau - t)\delta_{FF} + t(n-\tau)\delta_{FG} \right] - \delta_{FF} \right| = \frac{(n-\tau)}{(n-t)} |d_{FG} - d_{FF}| \\
|T_3(t) - T_2(t)| &\xrightarrow{P} \left| \frac{1}{t(n-t)} \left[(t(\tau - t)\delta_{FF} + t(n-\tau)\delta_{FG} \right] \right. \\
&\qquad \left. - \frac{1}{t(t-1)} \left[(t-\tau)(t-\tau-1)\delta_{GG} + \tau(\tau-1)\delta_{FF} + 2\tau(t-\tau)\delta_{FG} \right] \right| \\
&\leq \frac{(n-\tau)(\tau - t)}{(n-t)(n-t-1)} |d_{FF} - d_{FG}| + \frac{(n-\tau)(n-\tau-1)}{(n-t)(n-t-1)} |d_{FG} - d_{GG}| \\
&\leq \frac{(n-\tau)}{(n-t)} |d_{FF} - d_{FG}| + \frac{(n-\tau)(n-\tau-1)}{(n-t)(n-t-1)} |d_{FG} - d_{GG}| \\
&\qquad \left[\text{Assuming } t \leq \tau - 1 \text{ or } \frac{n-\tau}{n-t-1} \leq 1 \right]
\end{aligned}
$$

So when $t < \tau$ adding two expressions above we get

$$
\begin{aligned}
S_2(t) = |T_3(t) - T_1(t)| + |T_1(t) - T_2(t)| \xrightarrow{P} E_2(t) &\leq |d_{FF} - d_{FG}| + \frac{(n-\tau)(n-\tau-1)}{(n-t)(n-t-1)} |d_{FG} - d_{GG}| \\
&< |d_{FF} - d_{FG}| + |d_{FG} - d_{GG}|
\end{aligned}
$$

and $E_2(\tau) = |d_{FF} - d_{FG}| + |d_{FG} - d_{GG}|$. This implies $E_2(t) < E_2(\tau)$ when $t < \tau$.
Case 2: When $t \geq \tau$, then as $d \to \infty$,

$$
\begin{aligned}
|T_3(t) - T_1(t)| &\xrightarrow{P} \left| \frac{1}{t(n-t)} \left[(n-t)(t-\tau)\delta_{GG} + \tau(n-t)\delta_{FG} \right] - \delta_{GG} \right| = \frac{\tau}{t} |\delta_{FG} - \delta_{GG}| \\
|T_3(t) - T_2(t)| &\xrightarrow{P} \left| \frac{1}{t(n-t)} \left[(n-t)(t-\tau)\delta_{GG} + \tau(n-t)\delta_{FG} \right] \right. \\
&\qquad \left. - \frac{1}{t(t-1)} \left[(t-\tau)(t-\tau-1)\delta_{GG} + \tau(\tau-1)\delta_{FF} + 2\tau(t-\tau)\delta_{FG} \right] \right| \\
&\leq \frac{\tau(t-\tau)}{t(t-1)} |d_{GG} - d_{FG}| + \frac{\tau(\tau-1)}{t(t-1)} |d_{FG} - d_{FF}| \\
&\leq \frac{(t-\tau)}{t} |d_{GG} - d_{FG}| + \frac{\tau(\tau-1)}{t(t-1)} |d_{FG} - d_{FF}| \\
&\qquad \left[\text{Assming } t \geq \tau + 1 \text{ or } \frac{\tau}{t-1} \leq 1 \right]
\end{aligned}
$$

So when $t > \tau$,

$$
\begin{aligned}
S_2(t) = |T_3(t) - T_1(t)| + |T_1(t) - T_2(t)| \xrightarrow{P} E_2(t) &\leq |d_{FF} - d_{FG}| + \frac{\tau(\tau-1)}{t(t-1)} |d_{FG} - d_{FF}| \\
&< |d_{FF} - d_{FG}| + |d_{FG} - d_{GG}|
\end{aligned}
$$

and $E_2(\tau) = |d_{FF} - d_{FG}| + |d_{FG} - d_{GG}|$. This implies $E_2(t) < E_2(\tau)$ when $t > \tau$.

This completes the proof of consistency of estimator for S_2. □

References

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.

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.

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.

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*.