Proceedings of the 5th International Conference on Statistics: Theory and Applications (ICSTA'23)
Brunel University, London, United Kingdom – August 03-05, 2023
Paper No. 142
DOI: 10.11159/icsta23.142

High Dimensional Change Points: Challenges and Some Proposals

Lupeng Zhang[1], Reza Drikvandi[1]
[1]Department of Mathematical Sciences, Durham University, UK
lupeng.zhang@durham.ac.uk; reza.drikvandi@durham.ac.uk

Abstract – Change point analysis is being widely applied in various fields such as economics, finance, engineering, genetics and medical research. The main objective is to detect significant changes in the distribution of a data sequence. The change point problem for low dimensional data is well studied in the literature, however change point detection is challenging in high dimensional situations. The classical methods fail to work in high dimensional data where the number of variables is much larger than the number of observations. This paper discusses some main challenges with high dimensional change points and shows the limitations of the recent methods for high dimensional change points in dealing with such challenges. The paper presents some proposals to address those challenges.

Keywords**:** Change point, CUSUM statistic, Dissimilarity distance, High dimensional data

1. Introduction

High dimensional data are becoming popular due to the recent technological developments facilitating data collection and processing management. Change point analysis is important but challenging in high dimensional settings. Considering a collection of time-ordered observations, change point analysis initially aims to address two problems: whether and where a change in the underlying distribution of the observations may occur. Let $\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n}$ be a sequence of n independent p-dimensional random observations, where $p \gg n$, with distribution functions $F_1, F_2, \dots, F_n$. The fundamental problem of a single change point can generally be formulated as the following hypothesis testing problem

$$H_0: F_1 = F_2 = \cdots = F_n \quad vs. \quad H_1: F_1 = \cdots = F_\tau \neq F_{\tau+1} = \cdots = F_n, \tag{1}$$

where τ is the unknown change point location with $\tau \in \{1, \dots, \mathrm{n}-1\}$.

There is a growing literature on high dimensional change point analysis. On the one hand, recent nonparametric methods tend to provide robust change point detection in high dimensional settings. A nonparametric divergence measure based on the α^{th}absolute moment of Euclidean distance was used by [1] to develop a divisive and agglomerative algorithm for testing (1). Denoting $\boldsymbol{Y_\tau} = \{\boldsymbol{X_1}, \dots, \boldsymbol{X_\tau}\}$, $\boldsymbol{Z_\tau} = \{\boldsymbol{X_{\tau+1}}, \dots, \boldsymbol{X_n}\}$ and using a constraint value $\alpha \in (0, 2)$, the change point location τ can be estimated as

$$\hat{\tau} = arg \max_{1 \le \tau \le \mathrm{n}-1} \hat{Q}(\boldsymbol{Y_\tau}, \boldsymbol{Z_\tau}; \alpha), \tag{2}$$

where $\hat{\mathrm{Q}}(\boldsymbol{Y_\tau}, \boldsymbol{Z_\tau}; \alpha) = \frac{\tau(\mathrm{n}-\tau)}{\mathrm{n}}\left(\frac{2}{\tau(n-\tau)}\Sigma_{i=1}^{\tau}\Sigma_{j=1}^{n-\tau}\|\boldsymbol{Y_i} - \boldsymbol{Z_j}\|_2^{\alpha} - \binom{\tau}{2}^{-1}\Sigma_{1 \le i < k \le \tau}\|\boldsymbol{Y_i} - \boldsymbol{Y_k}\|_2^{\alpha} - \binom{n-\tau}{2}^{-1}\Sigma_{1 \le j < k \le \tau}\|\boldsymbol{Z_j} - \boldsymbol{Z_k}\|_2^{\alpha}\right)$ is the empirical divergence measure based on the Euclidean distance. [2] suggested to use interpoint distances to detect change points in high dimensional data. [3] developed a procedure based on the spatial and temporal dependence of data. On the other hand, recent parametric methods for high dimensional change points, which often require the normality assumption and sparsity, focus on dimensionality reduction of high dimensional data. This then enables one to apply the standard methods of low dimensional data to the transformed data. [4] used random projection to transform an $n \times p$ data matrix into an $\mathrm{n} \times 1$ univariate data sequence. [5] suggested a geometric mapping approach to project the data onto a two-dimensional space using distance and angle measures.

2. Some challenges in high dimensional change point detection

In the high dimensional settings, it is challenging to estimate the covariance matrix due to the curse of dimensionality. A reliable estimation of covariance matrix requires exponentially growing sample sizes, but the low sample sizes in high

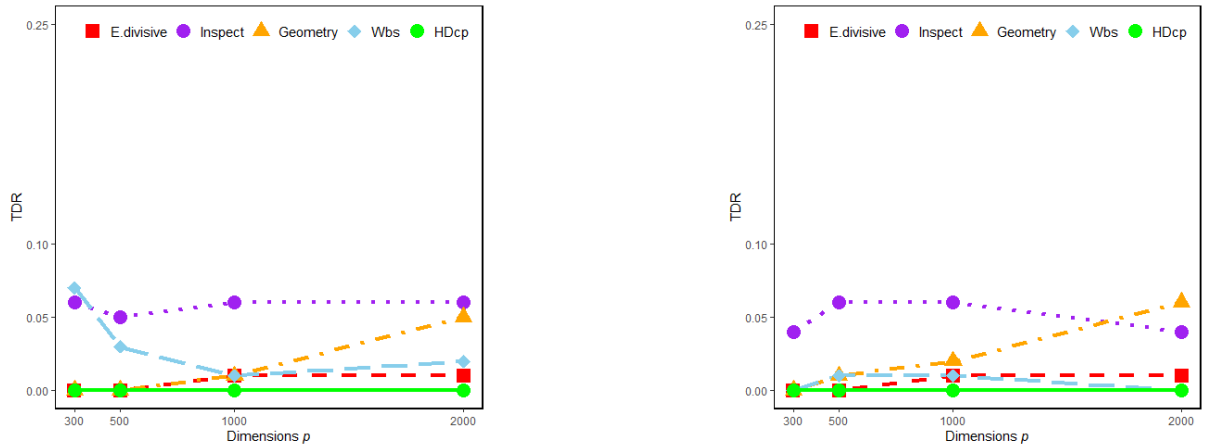
dimensional settings makes this challenging. Consider the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$. Simple algebra shows the nonexistence of $\widehat{\boldsymbol{\Sigma}}^{-1}$ as

$$rank\left(\widehat{\boldsymbol{\Sigma}}\right) \leq \min\left(rank\left(\boldsymbol{X}^{\boldsymbol{T}}\right), rank(\boldsymbol{X})\right) \leq \min(n, p) = n \ll p. \quad (3)$$

The nonexistence of the inverse covariance matrix makes the routine change point detection, for example based on likelihood tests, challenging in high dimensions.

Parametric methods often rely on the normality assumption as well as the sparsity of change points. The sparsity here refers to a change in only a small number of variables but with a large magnitude. The power of parametric methods is substantially affected when those conditions are violated in practice, especially when the change is small.

Most of the methods in the literature can only detect a change in the mean or variance of observations. To detect other distributional changes is even more challenging in high dimensional situations. One example is when there is a change in the shape of distribution while the mean and variance remain the same. To illustrate this challenging problem, we here assess the performance of five recent methods for high dimensional change points, namely E.divisive [1], HDcp [3], Inspect [4], Geometry [5] and WBS [6] (wbs with observation means), in two simulation scenarios when there is a true change in the shape of distribution but having the same mean and variance. Fig. 1 shows the detection performance of these methods. We can see that all these methods perform poorly in detecting such change in the shape of distribution.



(a) Change in the shape of distribution from $\mathrm{N}(0, 2)$ to $t(4)$. (b) Change in the shape of distribution from $\mathrm{N}(1, 1)$ to $Pois(1)$.

Fig. 1: The true detection rate (TDR) over 500 replications for five recent methods in detecting a true change in the shape of distribution while the mean and variance remain the same.

3. Some proposals for high dimensional change point detection

The cumulative sum (CUSUM) statistic is frequently used for change point detection [7]. It is defined as

$$\boldsymbol{C}(k) = \sqrt{\frac{k(n-k)}{n}}\left(\frac{1}{n-k}\sum_{i=k+1}^{n}\boldsymbol{X_i} - \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{X_i}\right), \quad (4)$$

where $k \in \{1, \ldots, n-1\}$. The CUSUM statistic quantifies the difference between the sample means before and after the k^{th} observation for each possible k. An appropriate test statistic based on the CUSUM is as follows

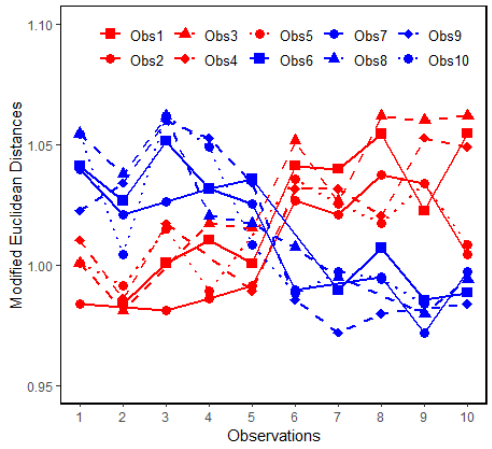
$$\mathrm{T} = \max_{1 \leq k \leq n-1} \|\boldsymbol{C}(k)\|_q, \quad (5)$$

in which $\|\cdot\|_q$ is the L^q-norm (often *q=1* or *q=2*). Clearly, large values of T will suggest the rejection of no change point.

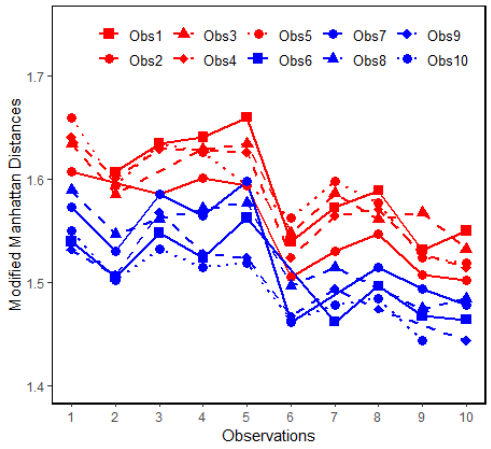
We propose to construct a CUSUM-type statistic based on the dissimilarity distances between observations. The CUSUM statistic based on the distances between observations would measure the average distance differences before

and after all possible change point locations. Unlike the observation means in (4), the dissimilarity distances could help detect more general types of change points. Because the Euclidean distance suffers converge issues in high dimensions [8], [8], we use a modified version of the Euclidean distance defined as $\mathrm{d}(\boldsymbol{X_i}, \boldsymbol{X_j}) = p^{-\frac{1}{2}} \|\boldsymbol{X_i} - \boldsymbol{X_j}\|_2$ to construct the proposed CUSUM statistic. As shown in [8], the scale $p^{-\frac{1}{2}}$ helps the modified Euclidean distance converge asymptotically as the dimension of data diverges. In Fig. 2, we present a visualisation of the modified Euclidean distance for detecting a change in the mean of high dimensional observations, which shows that this approach can distinguish between observations having a different distribution. We use the test statistic (5) with either of $q=1$ and $q=2$, and then apply a permutation procedure [9] to carry out the significance test as all the observations are exchangeable under the null hypothesis of no change point.

The proposal can be applied using any appropriate dissimilarity measure for high dimensional data. Another useful choice is the modified Manhattan distance with L^1-norm defined as $\mathrm{d}(\boldsymbol{X_i}, \boldsymbol{X_j}) = p^{-1} \|\boldsymbol{X_i} - \boldsymbol{X_j}\|_1$. Fig. 2 also shows the result of the modified Manhattan distance but for detecting a change in the shape of distribution while the mean and variance remain the same. The result supports the effectiveness of the modified Manhattan distance for this challenging problem. Note that our proposal does not require the normality assumption or any other distribution for data or sparsity.



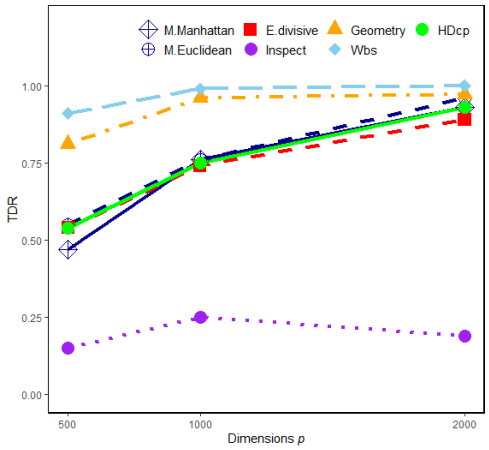
(a) Change in the mean from $\mathrm{N}(0, 0.5)$ to $\mathrm{N}(0.3, 0.5)$.



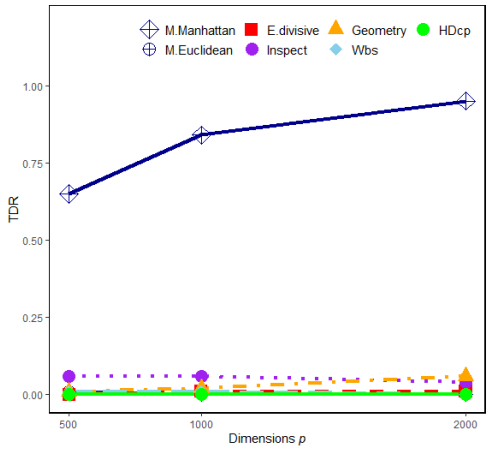
(b) Change in the shape of distribution from $\mathrm{N}(0, 2)$ to $t(4)$.

Fig. 2: The modified Euclidean distance and the modified Manhattan distance for 10 observations under different change point scenarios with p=2000, where the change occurs at location 5. For a better visualisation, we removed the cases when $d_{ii} = 0$.

We here conduct some simulations with 500 replications for two challenging high dimensional change point scenarios. As in line with the challenges discussed in Section 2, we evaluate the performance of our proposals based on the two distances, L^2-norm and L^1-norm, for those two scenarios and compare them with those five recent methods E.divisive [1], HDcp [3], Inspect [4], Geometry [5] and WBS [6] (wbs with observation means). We consider the high dimensional settings $n = 50$ and $p \in \{500, 1000, 2000\}$, where the true change point is set at location τ=30 of $n = 50$. For a fair comparison, we generate the data from normal distribution $N(0, 0.5)$ to satisfy the normality assumption for the parametric methods and set a mean shift of $\Delta\mu = 0.1$ right after observation $\boldsymbol{X_{30}}$. Since it is unrealistic that all p variables change in a high dimensional problem, we here keep $p/4$ variables unchanged. The simulation results are presented in Fig. 3. The results indicate that the modified Manhattan distance performs much better than all the other methods for detecting the change in the shape of distribution while mean and variance remain the same, and furthermore the modified Euclidean distance performs reasonably well compared to the other methods when detecting the mean shift. As the dimension increases, the performance of the two methods improves. Again, all those recent methods fail to detect such change in the shape of distribution when the mean and variance remain the same. The Inspect method does not even show a good power for detecting the change in mean of observations as this method requires a large sample size and a high level of sparsity.



(a) The proposed methods and other methods with the change in the mean by $\Delta\mu = 0.1$.



(b) The proposed methods and other methods with the change in the shape of distribution from $\mathrm{N}(0, 2)$ to $t(4)$.

Fig. 3: The true detection rate (TDR) over 500 replications of the proposed methods with different distances compared to the other recent methods for detecting changes in the mean as well as in the shape of distribution while mean and variance remain the same.

4. Concluding remarks

This paper discussed some major challenges with change point analysis in high dimensional data. These include the curse of dimensionality, the dependency on data nature such as sparsity and normality of observations, and the difficulty of detecting a change in the shape of distribution while the mean and variance remain the same. A CUSUM-type statistic based on dissimilarity distances between high dimensional observations was proposed to address those challenges. Numerical results showed the advantages of the proposals in comparison with some of the recent methods for high dimensional change points. We demonstrated the problem for a single high dimensional change point here because of the space limitation, however our ongoing research concerns the problem for multiple change points in high dimensional data. This involves the use of the recurring binary segmentation and the wild binary segmentation.

References

[1] D. S. Matteson and N. A. James, "A nonparametric approach for multiple change point analysis of multivariate data". *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 334–345, Jan. 2014.

[2] J. Li, "Asymptotic distribution-free change-point detection based on interpoint distances for high-dimensional data". *Journal of Nonparametric Statistics*, vol. 32, no. 1, pp. 157–184, Jan. 2020.

[3] J. Li, M. Xu, P.-S. Zhong, and L. Li, "Change point detection in the mean of high-dimensional time series data under dependence". *arXiv:1903.07006 [stat]*, Mar. 2019.

[4] T. Wang and R. J. Samworth, "High dimensional change point estimation via sparse projection". *Journal of the Royal Statistical Society: Series B*, vol. 80, no. 1, pp. 57–83, Aug. 2017.

[5] T. Grundy, R. Killick, and G. Mihaylov, "High-dimensional changepoint detection via a geometrically inspired mapping". *Statistics and Computing*, vol. 30, no. 4, pp. 1155–1166, Mar. 2020.

[6] P. Fryzlewicz, "Wild binary segmentation for multiple change-point detection". *The Annals of Statistics*, vol. 42, no. 6, Dec. 2014.

[7] B. Liu, X. Zhang, and Y. Liu, "High dimensional change point inference: Recent developments and extensions". *Journal of Multivariate Analysis*, vol. 188, p. 104833, Mar. 2022.

[8] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data". *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 3, pp. 427–444, Jun. 2005.

[9] R. Drikvandi, A. Khodadadi, and G. Verbeke, "Testing variance components in balanced linear growth curve models". *Journal of Applied Statistics*, vol. 39, no. 3, pp. 563-572, Mar. 2012.