Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. 121
DOI: 10.11159/icsta24.121

Challenges in High Dimensional Change Point Analysis and Advanced Approaches

Shrog Albalawi[1], Reza Drikvandi[1]
[1]Department of Mathematical Sciences, Durham University, UK
shrog.f.albalawi@durham.ac.uk; reza.drikvandi@durham.ac.uk

Abstract **-** Change point analysis aims at identifying significant distributional shifts and changes in data sequences. While the problem has been extensively studied for standard low dimensional data, the transition to high dimensional data imposes several challenges. This paper delves into the complexities of change point detection with high dimensional data, discussing the main difficulties associated with high dimensional change point analysis and demonstrating some limitations of the recent methods. The paper also discusses an approach for post detection analysis with high dimensional change points.

Keywords**:** Change point analysis, CUSUM statistic, High dimensional data, Post detection analysis.

1. Introduction

Change point analysis is a class of statistical methods with broad applications in various domains such as medicine, biology, engineering, environmental monitoring, finance, and marketing [1]. The main problem involves examining whether there is a significant distributional change in data before and after a point in time or space. When dealing with a sequence of ordered observations, such as a time series, the primary objective of change point analysis concerns two important questions:

a) Is there a change in the underlying distribution of the data sequence?
b) If a change is indeed detected, where precisely does it occur within the sequence?

These questions encompass the concepts of change point testing and estimation, which constitute fundamental aspects of change point analysis. The classical methods assume an increasing number of observations n towards infinity while maintaining a fixed dimension or number of variables p; however, recent technological advances in data collection and storage capabilities have led to a substantial rise of data with large dimensions. In high dimensional data, the number of variables p is much larger than the sample size n, often expressed as $p \gg n$, with p potentially reaching tens of thousands. This form of data imposes inherent complexities and heterogeneity in the underlying data generation processes. These challenges pose significant obstacles to accurately detect change points in higher dimensions. It is generally challenging to distinguish (small) significant changes from just random variability (i.e., noise) in high dimensional data. Such high dimensional noise can prevent or impede the precise detection of change point locations. Also, post detection analysis is important but remains understudied in the context of high dimensional change points.

2. Change point problem and the AMOC model

Consider a sequence of n random observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$, each p-dimensional and with an unknown probability distribution. Let us denote the unknown distributions by $G_1, G_2, \dots, G_n$, respectively. As discussed in the Introduction, the objective is to investigate whether there is a significant change point, that is, a point in the sequence where the distribution of data changes. The problem of a single change point can generally be formulated as the following hypothesis test:

$$\mathbf{H}_0: G_1 = \cdots = G_n \;\; vs. \;\; \mathbf{H}_1: G_1 = \cdots = G_\tau \neq G_{\tau+1} = \cdots = G_n, \tag{1}$$

where τ is an unknown change point location, with $\tau \in \{1, \dots, n-1\}$. This framework accommodates both parametric and nonparametric scenarios. Depending on the application, one may assume that the distributions belong to a parametric family, say $\{G_i(\boldsymbol{\theta}_i)\}_{i=1}^n$, where $\boldsymbol{\theta}_i$ is a vector of unknown parameters $\boldsymbol{\theta}_i \in \mathbb{R}^p$, or one may work with a nonparametric setting. In the parametric case, one focuses on parameters $\boldsymbol{\theta}_i$. Under the null hypothesis in (1) all the values of $\boldsymbol{\theta}_i$ are the same, whereas the alternative hypothesis says the parameter value $\boldsymbol{\theta}_i$ changes in location τ.

A common aspect of change point detection revolves around identifying significant changes in the mean of observations. This problem has been extensively studied in low dimensional settings. In the recent literature, several methods for detecting changes in the mean of high dimensional observations have been proposed including, but not limited to, [2], [3], [4], [5] and [6]. Also, a few methods have been proposed for detecting changes in the variance or covariance of data (e.g., [7] and [8]). Let $\boldsymbol{X} = (X_1, X_2, \dots, X_p) \sim G(\boldsymbol{x})$ be a random vector of observations. The data consists of n ordered observations, denoted by $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$, where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})$ is the i^{th}observation. The change point model in this case is as follows:

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{\delta\mu}_\tau 1\{i \geq \tau\} + \boldsymbol{\epsilon_i}, \qquad i = 1, \dots, n, \tag{2}$$

where the vector $\boldsymbol{\mu}$ represents the mean of observations before the change point, and $\tau \in \{1, \dots, n-1\}$ denotes the unknown location of the mean shift. The term $\boldsymbol{\delta\mu}_\tau = (\delta_1, \dots, \delta_p)$ corresponds to the mean shift vector after the change point location τ, if such a change exists. The vector $\boldsymbol{\epsilon}_i$ denotes the error term with $\mathbb{E}(\boldsymbol{\epsilon}_i) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$. Model (2) is commonly referred to as At Most One Change point (AMOC) model. To test if there is a significant change point in the mean of observations, we can then carry out the following hypothesis test:

$$\boldsymbol{H_0}: \boldsymbol{\delta\mu}_\tau = \boldsymbol{0} \text{ and } \tau = n \;\; vs. \;\; \boldsymbol{H_1}: \exists \tau \in \{1, \dots, n-1\} \text{ such that } \boldsymbol{\delta\mu}_\tau \neq \boldsymbol{0}. \tag{3}$$

For testing (3), the CUSUM statistic is frequently used which is defined as (e.g., [7])

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

in which $k \in \{1, \dots, n-1\}$ is a candidate search location. By evaluating at each candidate search location k, one can construct the following test statistic (see [7]) to test the null hypothesis $\boldsymbol{H_0}$ in (3):

$$T_n = \max_{1 \leq k \leq n-1} \boldsymbol{C}(k)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{C}(k). \tag{5}$$

The test statistic T_n checks all potential locations where a change point can occur by measuring the magnitude of mean differences before and after each candidate search location. Large values of T_n indicate that there are substantial deviations in the means, suggesting that the null hypothesis $\boldsymbol{H_0}$ should be rejected. The following section outlines some major challenges in detecting change points when data has a very large dimension.

3. Challenges and methods for high dimensional change point detection

In the test statistic (5), one needs the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$, however the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}$ is singular when the number of variables is larger than the number of observations. This is because $\widehat{\boldsymbol{\Sigma}}$ does not have full column rank as $rank(\boldsymbol{X^T X}) \leq min(\mathrm{rank}(\boldsymbol{X^T}), \mathrm{rank}(\boldsymbol{X})) \leq min(n, p) = n < p$. So, it is not possible to use the test statistic (5). [3] suggested a non-parametric method for high dimensional change points using the so-called energy distance based on the Euclidean norm. We briefly explain this method known as ecp. Let $\boldsymbol{Z}_\tau$ and $\boldsymbol{Y}_\tau$ denote independent samples corresponding to two distinct partitions of the dataset: $\boldsymbol{Z}_\tau = \{X_1, \dots, X_\tau\}$ and $\boldsymbol{Y}_\tau = \{X_{\tau+1}, X_{\tau+2}, \dots, X_n\}$. Additionally, a parameter $\alpha \in (0,2)$ controls the divergence measure characteristics. The objective is to estimate the change points location, denoted as τ, leading to the estimate $\hat{\tau}$, by maximizing the function $\hat{\mathcal{E}}(\boldsymbol{Z}_\tau, \boldsymbol{Y}_\tau; \alpha)$, where $\hat{\mathcal{E}}(\boldsymbol{Z}_\tau, \boldsymbol{Y}_\tau; \alpha) = \frac{2}{\tau(n-\tau)}\sum_{i=1}^{\tau}\sum_{j=1}^{\tau+1} \|\boldsymbol{Z}_i - \boldsymbol{Y}_j\|_2^\alpha - \binom{\tau}{2}^{-1}\sum_{1 \leq i < k \leq \tau} \|\boldsymbol{Z}_i - \boldsymbol{Z}_k\|_2^\alpha - \binom{n-\tau}{2}^{-1}\sum_{1 \leq j < k \leq \tau} \|\boldsymbol{Y}_j - \boldsymbol{Y}_k\|_2^\alpha$. [3] showed that this method performs well for change point detection. However, a recent study by [6] revealed some limitations with the energy distance statistic based on the Euclidean norm. They showed that the energy distance struggles to capture distinctions beyond the first two moments (mean and variance) in high dimensional settings. Moreover, it cannot detect change points when the mean and variance are the

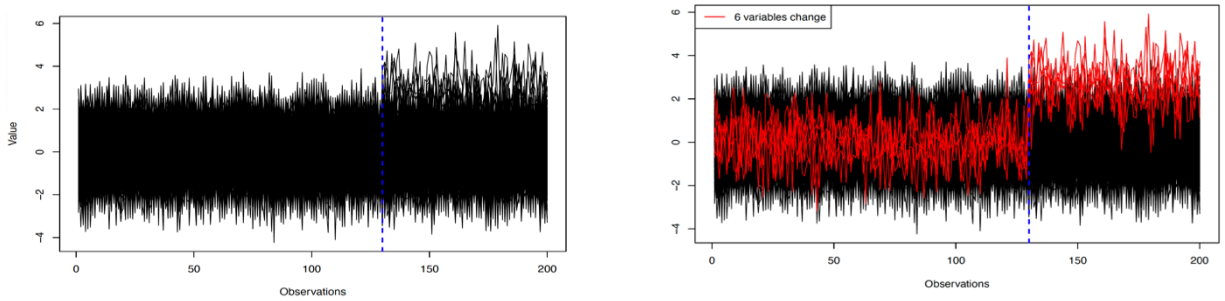
same while the distribution is different. Recently, [2] introduced a parametric method based on random projection of data into a one-dimensional subspace. They assumed $\boldsymbol{X}_t \sim N_p(\boldsymbol{\mu}_t, \sigma^2 \boldsymbol{I}_p)$, where $\{\boldsymbol{X}_t\}_{t=1}^n$ represents a high dimensional sample, $\boldsymbol{\mu}_t$ is the mean vector of observations $\{\boldsymbol{X}_t\}$ at time t, and σ^2 denotes the common variance. They used a projection vector $\boldsymbol{a} \in \mathbb{R}^p$, with $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} = 1$, to transform the p-dimensional observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ into the one-dimensional observations:

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}_t \sim N(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\mu}_t, \sigma^2), \qquad t = 1, \ldots, n.$$

[2] applied the low-dimensional CUSUM statistic (4) to estimate change point locations. However, the random projection method requires sparsity assumption as well as the normality assumption, which are restrictive in high dimensional settings.

4. Advanced approaches for high dimensions: post detection analysis

As a motivating example, we consider daily stock prices from multiple companies, such as S&P 500 stock price data [8], where detecting a significant price shift on a particular day exemplifies change point detection. Once a change point in stock prices is detected, the focus shifts onto validating the change reliability and determining the key variables driving this change. Identifying influential variables for post detection analysis is understudied in high dimensional change points. The identification of change points in high dimensional data is especially challenging when changes are due to a very sparse subset of variables. For example, as illustrated in Fig. 1(a), discerning which variables cause the observed changes can be daunting. Fig. 1(b) demonstrates the scenario where only six variables drive a significant change. So, it is important to address the intricacies of post change point detection in scenarios where changes are sparse and concentrated within a few variables. The goal is to pinpoint the specific variables that underlie significant shifts in the data landscape.



(a) Without knowing the truly important variables.

(b) With knowing the truly important variables.

Fig. 1 Change in mean of observations due to 6 variables out of 500, starting at change point location 130.

Representing the observed data as an $n \times p$ matrix, we write each column of the data matrix as $\boldsymbol{V}_j = (X_{1j}, \ldots, X_{nj})$, representing the j^{th} variable in the data for $j = 1, \ldots, p$. One can then write $\boldsymbol{X} = [\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p]$. We denote by $\boldsymbol{X}_{-j}$ the data matrix $\boldsymbol{X}$ when the j^{th} column $\boldsymbol{V}_j$ is excluded. Recall the AMOC model (2) which primarily focuses on mean change. If there is a significant change point, say $\hat{\tau}$, we investigate which variable(s) would cause this change. We remove each variable $\boldsymbol{V}_j$ at a time and then apply the change point method to $\boldsymbol{X}_{-j}$. We denote the estimated change point with this by $\hat{\tau}_{-j}$. We then define $\Delta_j = |\hat{\tau} - \hat{\tau}_{-j}|$. If Δ_j is very small then variable $\boldsymbol{V}_j$ is not important, but if Δ_j is large, say $\Delta_j > t$ for some threshold value t, for example $t = 2$, we can conclude that the variable $\boldsymbol{V}_j$ is an important variable for the change.

We here conduct some simulations to evaluate the performance and effectiveness of this approach. In the simulations, we examine both low and high dimensional settings under different scenarios. We consider sample size $n = 100$ and four different dimensions $p \in \{50, 100, 200, 500\}$. The true change point here is set at location $\tau = 65$. We generate the data from a normal distribution $N(0, 0.5)$ and set a mean shift of $\boldsymbol{\delta}\mu_\tau = \{1, 1.5\}$ right after the observation X_{65}. Here, we focus on a simple case where changes are attributed to one variable—specifically the second variable. We apply the ecp method [3] on the simulated data. The simulation results over 200 replications, which are presented in Tables 1 and 2, show the proportion of correctly detected important variables over the 200 replications for each scenario. The frequency of truly detected important variables is reasonably well especially when the mean shift increases. Note that if the mean shift is small,

the signal gets dominated by the high dimensional noise, making the identification of a significant change point and important variables causing it difficult. We note that the results are less satisfactory when the dimension is very large compared to the sample size. This is because of the high level of sparsity. When dealing with a larger number of variables, the computational time of this approach becomes demanding. Also, another limitation is that if several variables are equally important, removing one variable while retaining the others may still result in the same change point. To overcome such challenges for data with very high dimensions and with very sparse changes, we can use clustering techniques to group variables based on their similarity characteristics. Further research is required to investigate this approach in such high dimensional cases.

Table 1: The proportion of correctly identified important variables over 200 simulation replications. Note that the true change point is set at location $\tau = 65$ accompanied by a mean shift of $\boldsymbol{\delta}\mu_\tau = 1$.

n	p	Average $\hat{\tau}_j$	No. of truly important variables	Proportion of correctly identified important variables
100	50	66	1	53%
	100	64	1	24%
	200	66	1	8%
	500	59	1	5%

Table 2: The proportion of correctly identified important variables over 200 simulation replications. Note that the true change point is set at location $\tau = 65$ accompanied by a mean shift of $\boldsymbol{\delta}\mu_\tau = 1.5$.

n	p	Average $\hat{\tau}_j$	No. of truly important variables	Proportion of correctly identified important variables
100	50	65	1	94%
	100	65	1	93%
	200	65	1	77%
	500	64	1	27%

5. Conclusion

The main challenges with change point analysis in high dimensional data were discussed, particularly by focusing on difficulties in accurately estimating patterns and detecting shifts when dealing with a very large number of variables. To identify important variables causing change points, a post detection analysis was investigated by removing variables each at a time. It was discussed that employing grouping techniques emerges as an efficient approach in high dimensions.

References

[1] L. Zhang and R. Drikvandi. "High dimensional change points: challenges and some proposals". *Proceedings of the 5th International Conference on Statistics: Theory and Applications (ICSTA'23)*, Aug. 2023.

[2] 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. 2018.

[3] 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.

[4] H. Cho and P. Fryzlewicz. "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation". *Journal of the Royal Statistical Society Series B,* vol. 77, no. 2, pp. 475-507, 2015.

[5] M. Jirak. "Uniform change point tests in high dimension". *Annals of Statistics*, vol. 43, no. 6, pp. 2451–248, 2015.

[6] S. Chakraborty and X. Zhang. "High-dimensional change-point detection using generalized homogeneity metrics". arXiv preprint *arXiv:2105.08976*. 2021.

[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] R. Drikvandi and R. Modarres. "A distribution-free method for change point detection in non-sparse high dimensional data". *Journal of Computational and Graphical Statistics*, https://doi.org/10.1080/10618600.2024.2365733, Jun. 2024.