Stockholms universitet

Change Point Detection Based on Principal Component Analysis for Multivariate Time Series with Application to Single Molecule Data

Huixin Zhong

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2019:12
Matematisk statistik
September 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2019:12**
http://www.math.su.se

Change Point Detection Based on Principal Component Analysis for Multivariate Time Series with Application to Single Molecule Data

Huixin Zhong*

September 2019

Abstract

Change point detection has been a long-standing problem in statistical analysis. This study aims to develop a nonparametric offline scheme for detecting changes in mean and/or variance in multivariate time series. Based on the principal component analysis (PCA), the multivariate data is projected onto a lower dimensional space such that the multidimensional detection problem is reduced to a one-dimensional one. As a result, we can apply the well formulated univariate methods, namely, the cumulative sum (CUSUM) and cumulative sum of squares (CSS) methods, to test the existence of a change in mean and variance, respectively. The study shows that both methods are reliable to test the existence of a change based on the permutation test. Moreover, the CUSUM-based estimator and the mean square error (MSE) estimator are proposed to determine the most likely change point locations. We compare the performance of the estimators in a simulation study and the results show that the performance of these two estimators depends very much on the properties of data. The MSE estimator outperforms the CUSUM-based estimator if determining a change point location in a time series with a mean change. On the contrary, if determining a change point location with a variance change, the CUSUM-based estimator is preferable. Finally, the CUSUM and CSS methods are combined to detect simultaneous changes in mean and variance. The results on both simulated and real data show that the combined method complements each other well and it can successfully determine the most prominent change point location by comparing the uncertainty in identifying the change point locations.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: huixinsandy.zhong@gmail.com. Supervisor: Chun-Biu Li.

Acknowledgements

This thesis constitutes a master's thesis of 30 ECTS in Mathematical Statistics at the Department of Mathematics at Stockholm University.

I would like to express my deepest gratitude to my supervisor Chun-Biu Li for his tremendous amount of support and valuable guidance during this project. Thank you so much for introducing me to the fascinating topic of change point analysis. It has been a long, difficult but very interesting and inspirational study process. The completion of this thesis would not have been possible without Chun-Biu's greatest support. Furthermore, I would like to thank my family for always understanding and supporting me. Last but not least, I would like to express my great appreciation to my fiance David for his constant support and encouragement throughout this project.

Contents

1 Introduction

1.1 Background

Nowadays, the world is filled with data and changes. One wishes to discover the changes in the statistical properties of data to make better decisions and to avoid possible unnecessary losses. This gives rise to a long-standing problem in statistical analysis, the change-point problem, which is a process of detecting distributional changes in a stochastic process or a time series.

In fact, change point detection has already had a wide range of applications in our daily life. For example, in financial time series modelling, whether an abnormal shifting of the stock price for a company is a statistically significant change in the stock market that will affect investors' decision making. In quality control, one wants to find out whether there is a point where the quality of a product starts to change so that one can address the problem as soon as possible. It has been applied to detect artificial or natural discontinuities and regime shifts in climate [1]. Apart from these, applications can also be found in geology, genetics, medicine, social study and so on [2].

As change point detection is important in a variety of fields, many methods have been developed. Among the several examples mentioned above, methods to detect changes depend on the properties of data. Problems like how the data is obtained, how many variables the dataset consists of and how much distributional information one knows about the data need to be taken into consideration. In this thesis, detecting changes in multidimensional time series is of interest. We give a brief introduction to the change-point problem and clarify the aim of the thesis in the following sections.

1.2 The Change-Point Problem

The change-point problem concerns generally determining whether a single change or multiple changes have been taken place and identifying the locations in which any such changes occur. Suppose a set of independent observations $x_1, x_2, ...$ are drawn from random variables $X_1, X_2, ...$ and an unknown number of changes in distribution at the unknown change point locations $\tau_1, \tau_2, ...$ may occur, the distribution of the random process can be written as [4]:

$$X_i \sim \begin{cases} F_1 & \text{if } i \leq \tau_1 \\ F_2 & \text{if } \tau_1 < i \leq \tau_2 \\ F_3 & \text{if } \tau_2 < i \leq \tau_3 \\ \dots & \end{cases} \tag{1}$$

where F_is are unknown probability distributions in each segment. The goal of change point detection is to recover these segments as accurately as possible.

Depending on the setting of the problem, methods used for detecting the change points are different. We now specify some categorizations related to the change-point problem [4].

Offline or Online Setting: When observations $X_1, X_2, ..., X_{n-1}, X_n$ are received continuously and processed sequentially over time, it is called online or sequential setting. In this case, the length of the random process is not fixed and a decision on whether a change has occurred has to be made as quickly as possible when a new observation has arrived. In contrast, when an entire set of observations is obtained all at once and the retrospective analysis is performed, it is called offline setting. The focuses of the offline setting are testing the existence of a change point at an unknown location and estimating the most likely change point location given the existence of a change point. This study is essentially concerned with offline data setting.

Single or Multiple Change Point Detection: In real life data, multiple change points often take place. Nevertheless, the focus of this thesis is to find one single change point. For those who are also interested in detecting multiple change points and their locations in an offline dataset, we refer to the widely used method, the *binary segmentation procedure* proposed by Vostrika [2].

Parametric or Nonparametric: Change point detection methods differ a lot depending on whether the distributional knowledge of the data is known or not. In real life data, information about the underlying distribution may not be known. In this thesis, we will apply nonparametric methods for change point detection. For an overview of parametric offline methods, we refer the interested readers to [2] and [3].

One-dimensional or Multidimensional Setting: The data we consider in this thesis is multidimensional and we assume that the changes occur simultaneously in different dimensional of the data. The well-known dimensionality reduction method, principal component analysis, will be applied to the data. The resulted data used for further change point analysis is actually one-dimensional. Some basic examples are given in Figure 1, which shows two sequences of two-dimensional normally distributed random variables that undergo one change in mean, variance or both.

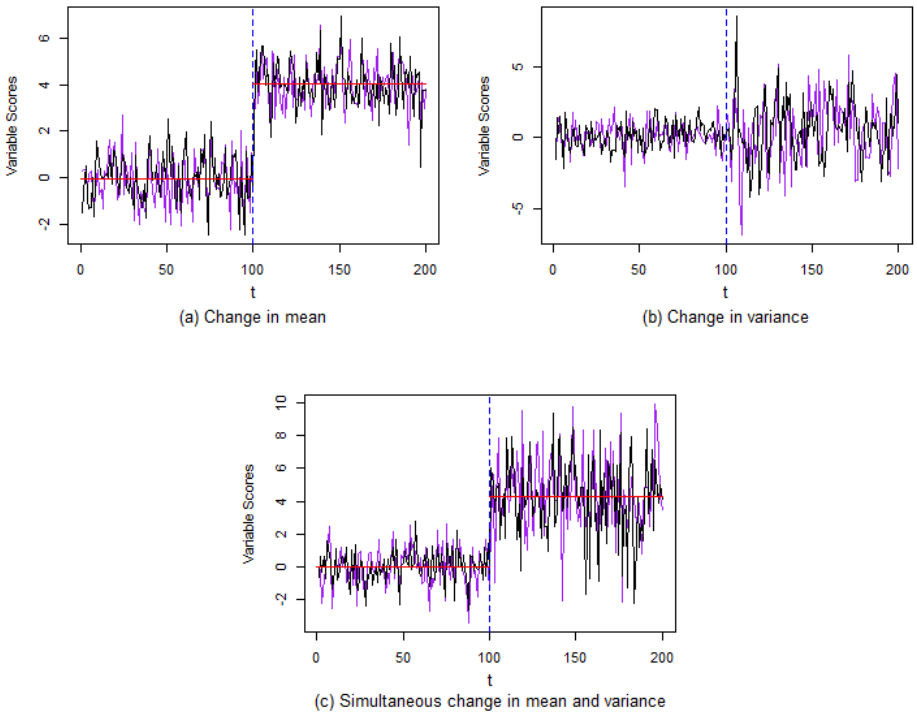


Figure 1: Some basic examples of commonly investigated changes in behaviour to a two-dimensional time series. The traces are plotted in different colors to represent two different variables. Here we are in an offline setting, with one single change point, with a parametric model. The time of the change point is superimposed at position 100, indicated by the blue dashed line. The red segments indicate the mean change.

1.3 Purpose

This study aims to develop a nonparametric scheme for detecting changes in mean and/or variance in multidimensional offline time series based on principal component analysis, which is used for projecting the data into a lower dimensional space. For simplicity, we will only concentrate on finding one single change in two-dimensional time series in this study. Nevertheless, the proposed scheme is general and can be extended to higher dimensional data.

1.4 Outline

The rest of the thesis is organized as follows. Section 2 provides an overview of related work in the field of change point analysis. Section 3 explains the dimensionality reduction method, the principal component analysis and the theoretical analysis of change point detection. Artificial data are generated and a simula-

tion study is conducted in Section 4. Section 5 presents analysis and results from application to real life data (a 2D time series observed in single molecule experiments). Conclusions are made in Section 6. Discussion of the results as well as suggestions for future studies can be found in Section 7.

2 Literature Review

Principal component analysis (PCA) is used as a powerful tool in many aspects. Some researchers have applied PCA for change point detection in multidimensional time series. Kuncheva and Faithfull [5] developed an approach based on PCA feature extraction for change detection in unlabelled, multidimensional streaming data and showed that many datasets benefit significantly from using PCA. Qhatan *et al* [6] proposed a framework that applies PCA's dimensionality reduction property to project the multidimensional streaming data onto the principal components to obtain multiple 1D data streams. Density functions of the projected data are estimated and change-scores are computed to compare the distributions before and after the occurrence of a change. However, these existing approaches are designed for online data streams. The scheme provided in this thesis is also based on PCA's dimensionality reduction property, but it is applicable for offline time series. We will analyze and show that using the first principal component associated with the largest eigenvalue captures the largest variability such that any changes in the original variables are reflected in the first principal component.

As mentioned, the focus of this thesis is to detect one single change point in a two-dimensional offline time series. By using PCA to reduce the data dimension, we can apply to the well formulated univariate methods for detecting changes. The literature on change point detection is rather huge though. In parametric change point detection, the most frequently encountered methods are using the likelihood ratio test given that the underlying distribution of the data is known. An overview of change point analysis for various parametric models can be found in Chen and Gupta [2]. Another strand of literature is based on the so-called cumulative sum (CUSUM) tests for change point detection. The CUSUM method was first introduced to detect one change in one distribution parameter for online data and further developed to fit offline data [4]. Depending on how the CUSUM statistics are constructed, different types of changes can be detected. Taylor developed a cumulative sum-based procedure to detect changes in mean values [7] and Inclan and Tiao discussed the detection of a change in variance [8].

Furthermore, if the change point problem is formulated as testing two hypotheses, e.g., H_0 : no change point exists versus H_1 : there exists a change, two-sample hypothesis tests can be applied [9]. For example, a two-sample t-test to detect a mean change and an F test to detect a variance change can be applied if the

data is assumed to be normally distributed. If no assumption is made on the data, many nonparametric tests can be applied, such as the Wilcoxon rank sum (Mann–Whitney) test for location change, the Mood test for dispersion change and Kolmogorov-Smirnov test for general changes. However, there is an issue in using these tests. That is, it requires the knowledge of the null distribution of the test statistic, which generally does not have an analytic form.

3 Methodology

In this section, the methods used in the study are described in detail. We explore first the dimensionality reduction technique, the principal component analysis. Thereafter the statistical tests and offline change point detection methods based on cumulative sum are presented.

3.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a well-known technique for reducing dimensionality for large datasets, increasing its interpretability while still being able to preserve as much information as possible. It has a lot of other applications such as data visualization, image compression, feature extraction and engineering, and much more. The main idea of PCA is to find a set of orthogonal linearly transformed principal components that are derived from the original data such that the first principal component captures the largest variability of the data, the second principal component captures the second largest variability, and so forth. Hence, the first principal component has always the maximum variance among all principal components [10]. The classic approach measures variability through variance. The principal components can be either based on the eigenvectors of the sample covariance matrix or correlation matrix. With the use of the correlation matrix, variables are standardized to zero mean and unit variance. As a result, finding these principal components reduces to solving an eigenvalue or eigenvector problem.

The goal of using PCA in this thesis is essentially dimensionality reduction. We want to zero out all of the smallest principal components from a multidimensional dataset and obtain a lower-dimensional projection of the data which preserves the largest possible variance. For instance, consider Figure 2, which shows a two-dimensional normally distributed time series generated by two random variables X_1 and X_2, with 100 observations. The blue line represents the first principal component. By eyes, it is clear that the data is most spread out on the line, which indicates a large variance. In other words, if all data points were projected onto this line, the resulting projected data points would have the maximum variability with only one single dimension. The resulting one-dimension time series, which exhibits the most variation and will reflect any

changes in the original multidimensional time series is the desired dataset that will be performed further change point analysis.

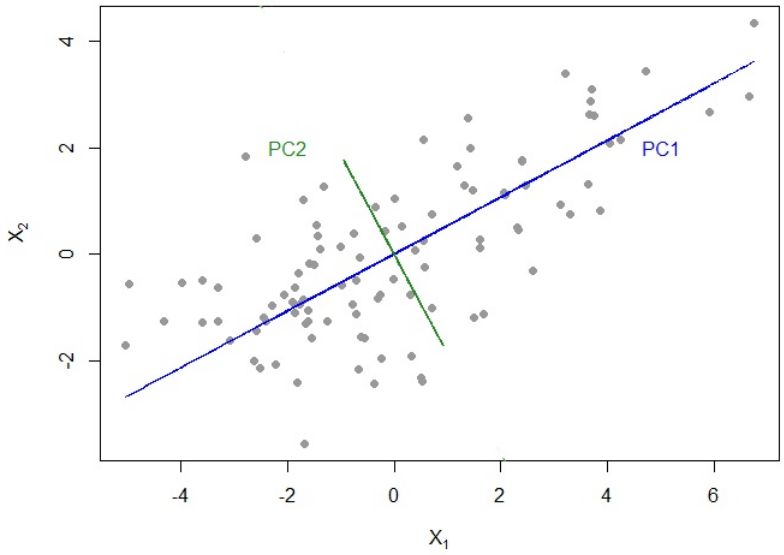


Figure 2: Variable X_1 and X_2 for 100 simulated observations are shown as grey dots in normalized term (mean = 0). The blue line indicates the first principal component, and the green line indicates the second principal component.

3.1.1 The Computational Procedure

As mentioned above, the derivation of principal components can be based either on a covariance matrix or a correlation matrix. There have been lots of discussions on which method is preferable and the interested readers are referred to Chapter 2 in [12]. In this thesis, we use the covariance matrix method without further discussion. The entire dimensionality reduction can be divided into the following three steps [10, 11, 12, 13]:

- Centralize a given multidimentional time series by subtracting the mean value of each variable from the data and then derive the covariance matrix.
- Compute the eigenvalues and eigenvectors of the covariance matrix to obtain the direction of the first principal component.
- Project each point of the original dataset onto the first principal component line.

Mathematically, consider an n-dimensional random variable $\mathbf{X} = \{X_1, \cdots, X_n\}$. Without loss of generality, the column vectors $\mathbf{X}$ are assumed to be centralized

with zero mean. Then the covariance matrix $\mathbf{S}$ can be obtained by

$$\mathbf{S} = \frac{1}{k-1} \sum_{i=1}^{k} {X_{il}}^T X_{il} \qquad \text{for } i = 1, \cdots k. \tag{2}$$

where l denotes the l-th column of $\mathbf{X}$ for $l \in \{1, \cdots n\}$ and k denotes the number of data points.

Let a vector of principal components $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_p\}$ represent p ($1 \leq p \leq n$) uncorrelated linear components of $\mathbf{X}$, that is, each principal component vector is defined as

$$Y_j = \boldsymbol{X}\boldsymbol{a}_j \qquad\qquad \text{for } j = 1, \cdots p. \tag{3}$$

$$Y_j = \begin{bmatrix} y_{1j} \\ \vdots \\ y_{kj} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \\ x_{k1} & & x_{kn} \end{bmatrix} \begin{bmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{bmatrix}$$

where $\boldsymbol{a_j}$ is the normalized eigenvectors of covariance matrix $\mathbf{S}$, such that the variance of Y_j is maximized. Since the covariance matrix $\mathbf{S}$ is real and symmetric, the eigenvectors a_j are orthogonal to each other and they can be obtained by solving a set of equations:

$$(\mathbf{S} - \lambda_j I_n)\boldsymbol{a}_j = \mathbf{0} \tag{4}$$

where λ_j are the eigenvalues of $\mathbf{S}$, $\lambda_1 = Var(Y_1) > \lambda_2 = Var(Y_2) > \cdots > \lambda_p = Var(Y_p)$ and I_n is a $n \times n$ identity matrix. A detailed derivation of eigenvalues and eigenvactors can be found in Appendix. The eigenvectors are rearranged with descending eigenvalues. A projection matrix is formed by the sorted p eigenvectors associated with the p largest eigenvalues λ_j. The largest eigenvector corresponding to the very largest eigenvalue defines the direction of the first principal component. The projection of each data point onto the direction line creates the first principal component, e.g., $Y_1 = \mathbf{X}\boldsymbol{a}_1$. The second principal component can be produced using the second largest eigenvector, and so forth.

Thus, the entire dimensionality reduction is actually done by an orthogonal linear transformation of the original time series. Equation (3) can be rewritten as

$$\mathbf{Y}_{k \times p} = \mathbf{X}_{k \times n}\ \mathbf{A}_{n \times p} \tag{5}$$

where $\mathbf{A}$ is the orthogonal matrix having the sorted p eigenvectors as its columns.

3.1.2 Example of Dimensionality Reduction

This section gives examples to illustrate how the dimensionality reduction referred to above is done in practice. We consider again the 100 simulating observations drawn from a bivariate normal distribution in the previous section. In Figure 3a, the blue line represents the direction of the first principal component of the data, and in Figure 3b, the 100 observations have been projected onto the line. Mathematically, this process is done by first using equation (2) and (4) to find the largest eigenvector

$$A_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = \begin{pmatrix} 0.882 \\ 0.472 \end{pmatrix}$$

Apply thereafter equation (3) or (5) to obtain the first principal component

$$Y_1 = 0.882 \times (X_1 - \bar{X_1}) + 0.472 \times (X_2 - \bar{X_2}) \tag{6}$$

where $\bar{X_1}$ indicates the mean of all X_1 values in this variable, and $\bar{X_2}$ indicates the mean of all X_2 values. Since the X_1 and X_2 are vectors of length 100, and so is the first principal component Y_1 whose values can be seen in Figure 3b.

Since the example dataset is two-dimensional, we can construct up to two distinct principal components. The second principal component Y_2 is computed using the second largest eigenvector which is orthogonal to the first eigenvector A_1. Applying again equation (3) or (5), the second principal component is given as follows

$$Y_2 = 0.882 \times (X_2 - \bar{X_2}) - 0.472 \times (X_1 - \bar{X_1}) \tag{7}$$

By construction, the first principal component captures the most variability of the original data, which can also be shown by the much larger variability of Y_1 (on the x axis) compared to Y_2 (on the y-axis) in Figure 3b. Another illustration can be shown by Figure 4 which displays the first principal component Y_1 versus the variables both X_1 and X_2. A strong linear relationship between the first principal component and the two variables is shown on the scatterplots. In other words, the first principal component captures most of the fluctuations of the data. On the other hand, very little relationship between the second principal component and the two variables can be observed in Figure 5, which displays scatterplots of Y_2 versus X_1 and X_2. All of these suggest that one only needs to keep the first or the first few principal components in order to accurately represent the original data.

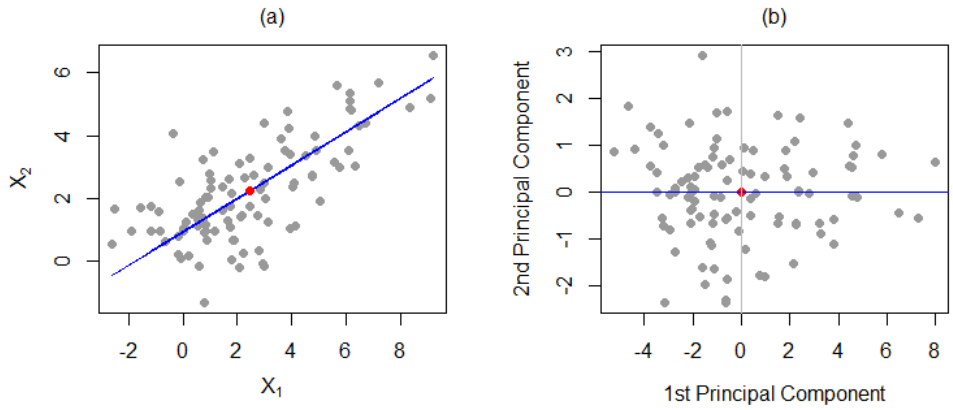


Figure 3: Variable X_1 and X_2 for 100 simulated observations are shown as grey dots. The mean of (X_1, X_2) is indicated with a red dot, denoted $(\bar{X}_1, \bar{X}_2)$. **(a)** The direction of the first principal component is shown in blue color in the original data setting (before data centering). It is the dimension of the data that preserves the maximum data variability. **(b)** A scatterplot of the 1st PC versus the 2nd PC resulting by rotating Fig.a. We see that the first principal component direction coincides with the x-axis.

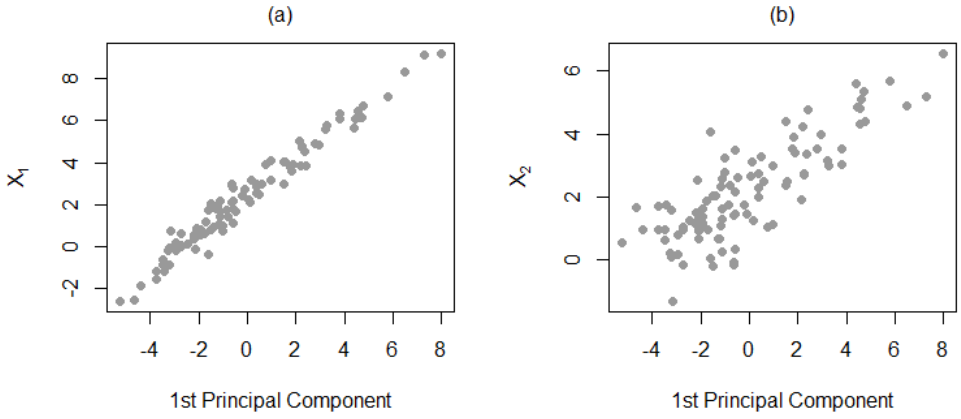


Figure 4: **(a)** Scatter plot of 1st PC versus X_1. **(b)** Scatter plot of 1st PCX_2

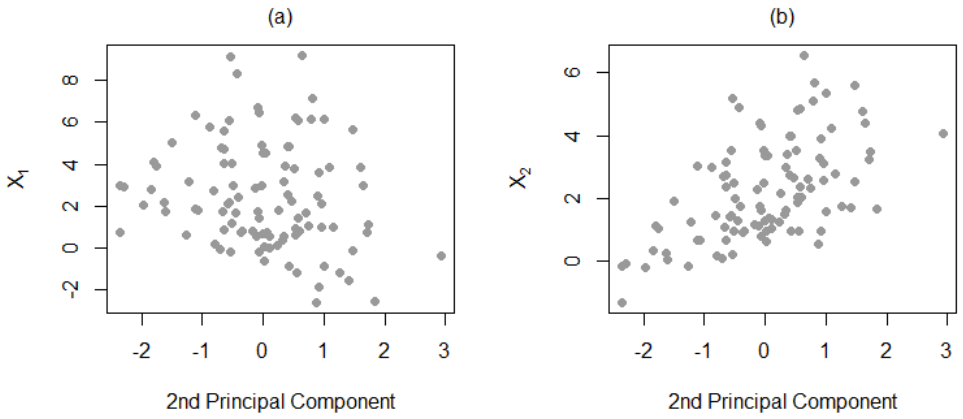


Figure 5: **(a)** Scatter plot of 2nd PC versus X_1. **(b)** Scatter plot of 2nd PCX_2

3.2 Comparing Distributions using Kullback-Leibler Divergence

As we saw in the previous section, the first principal component captures the largest variability of the original multidimensional data, which is the reason why it is selected for further change point analysis. However, how can one be sure that any changes in the original data are reflected in the first principal component? In this section, we analyze how different types of changes can be observed in the first principal component. We start to illustrate this problem graphically by using a concrete example and then perform a theoretical analysis by using the widely used Kullback–Leibler divergence.

3.2.1 Graphical Analysis

Consider a time series with 200 observations generated from a bivariate normal distribution, undergo one change in mean and variance simultaneously. The distributions before and after the change are: $q(x) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$ and $p(x) \sim N\left(\begin{pmatrix}3\\3\end{pmatrix}, \begin{pmatrix}6 & 3\\3 & 6\end{pmatrix}\right)$, respectively. Figure 6a shows the trace of the original 2D time series and Figure 6b shows the trace of the first principal component. From these, we can see by eyes that there is a sudden jump close to time point 100 in both figures and at the same time the variance of the traces starts to change. In other words, changes in the original 2D time series are reflected on the first principal component.

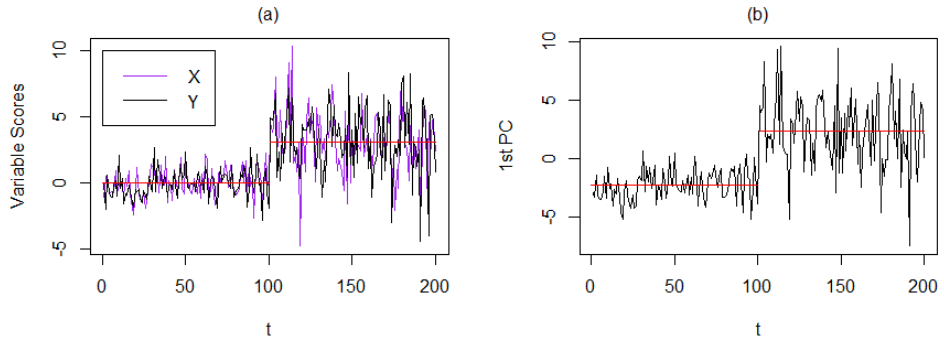


Figure 6: An example of traces with a simultaneous change in mean and variance. The time series consists of 200 observations with a change at 100. (**a**) The existence of an sudden jump in the original bivariate time series. (**b**) The existence of an sudden jump in the first principal component. The two red segments on both figures indicate a change in mean value.

3.2.2 Theoretical Analysis

In change point analysis, changes are detected by measuring the difference between the distributions before and after the change. The Kullback–Leibler divergence is a measure of how one distribution is different from another distribution and defined as [14]:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{8}$$

where $D_{KL}(P \parallel Q) \geq 0$, $p(x)$ and $q(x)$ denote the probability density function (PDF) of P and Q. A Kullback–Leibler divergence equals to zero if and only if two distributions in question are identical. In change point setting, $q(x)$ represents the PDF before the change and $p(x)$ represents the PDF after the change.

In fact, Qahtan *et al* [6] has already proved this problem by using Kullback–Leibler divergence. We adapt their analysis and show how change in mean and change in variance are reflected in the first principal component here.

For simplicity, Qahtan *et al* assume that the data without change has mean vector $\boldsymbol{\mu_1} = [0, 0]^T$ and variables have the same variance such that the covariance matrix $\Sigma_1 = \begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix}$ for $-1 < \rho < 1, \quad \rho \neq 0$. The following lemma whose proof can be found in Appendix will be used.

Lemma 1: *For two univariate normal distributions, $p(x)$ and $q(x)$, where $q(x) \sim N(\mu_1, \sigma_1^2)$ and $p(x) \sim N(\mu_2, \sigma_2^2)$, the KL-divergence from $q(x)$ to $p(x)$*

is given as follows:

$$D_{KL}(p \parallel q) = \frac{1}{2}\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) - \frac{1}{2} + \frac{\sigma_2^2 + (\mu_1 - \mu_2)^2}{2\sigma_1^2} \tag{9}$$

If PCA is not applied, data are the same as projection on the original coordinates $\boldsymbol{u_1} = [1, 0]^T$ and $\boldsymbol{u_2} = [0, 1]^T$. If PCA is applied, data are projected on an orthogonal matrix having the sorted eigenvectors as its columns (described in Section 3.1.1). In this case, we obtain the first eigenvector $\boldsymbol{v_1} = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$ associated with the largest eigenvalue $\lambda_1 = \sigma^2 + \sigma^2\rho$, and the second eigenvector $\boldsymbol{v_2} = \left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$ associated with the second eigenvalue $\lambda_2 = \sigma^2 - \sigma^2\rho$.

Case 1: Change of Variance
With changes in the variance of the two variables, the covariance matrix after the change can be $\Sigma_2 = \begin{pmatrix} \sigma^2 + \tau_1 & \sqrt{\sigma^2 + \tau_1}\sqrt{\sigma^2 + \tau_2}\rho \\ \sqrt{\sigma^2 + \tau_1}\sqrt{\sigma^2 + \tau_2}\rho & \sigma^2 + \tau_2 \end{pmatrix}$ where τ_i denote changes in variance with $\tau_i \geq 0,\ i = 1, 2$.

In the original coordinate system, before the change in variance of the variables, the projection on $\boldsymbol{u_1}$ has the distribution $q_1^u \sim N(0,\ \boldsymbol{u_1}^T\Sigma_1\boldsymbol{u_1}) = N(0,\ \sigma^2)$. After the change, $p_1^u \sim N(0,\ \boldsymbol{u_1}^T\Sigma_2\boldsymbol{u_1}) = N(0,\ \sigma^2 + \tau_1)$. Then, applying *Lemma 1*, we have

$$D_{KL}(p_1^u \parallel q_1^u) = \frac{1}{2}\log\left(\frac{\sigma^2}{\sigma^2 + \tau_1}\right) + \frac{\tau_1}{2\sigma^2}$$

If PCA is applied, the first principal component is obtained after projecting the data on the first eigenvector. Before the change, the first principal component has the distribution $q_1^v \sim N(0,\ \boldsymbol{v_1}^T\Sigma_1\boldsymbol{v_1}) = N(0,\ \lambda_1)$. After the change, $p_1^v \sim N(0,\ \boldsymbol{v_1}^T\Sigma_2\boldsymbol{v_1}) = N\left(0,\ \sigma^2 + \sqrt{\sigma^2 + \tau_1}\sqrt{\sigma^2 + \tau_2}\rho + \frac{1}{2}(\tau_1 + \tau_2)\right)$. The KL-divergence is calculated as

$$D_{KL}(p_1^v \parallel q_1^v) = \frac{1}{2}\left(\log\frac{2\lambda_1^2}{2(\sigma^2 + r) + \tau_1 + \tau_2}\right) + \frac{2(\sigma^2 + r) - 2\lambda_1^2 + \tau_1 + \tau_2}{2\lambda_1^2}$$

where $r = \sqrt{\sigma^2 + \tau_1}\sqrt{\sigma^2 + \tau_2}\rho$. We see that $D_{KL}(p_1^v \parallel q_1^v) > 0$ as long as $\tau_1, \tau_2 \geq 0$, that is, we can observe changes in the variances of the original variables in the first principal component.

Case 2: Change of Mean
Suppose that the mean vector of the data changes from $\boldsymbol{\mu_1} = [0, 0]^T$ to $\boldsymbol{\mu_2} = [\tau_1, \tau_2]^T$ where τ_i denote mean shifts with $\tau_i \geq 0,\ i = 1, 2$.

In the original coordinate system, the projections on $\boldsymbol{u_1}$ before and after the mean shift have distributions $q_1^u \sim N(0,\ \sigma^2)$ and $p_1^u \sim N(\tau_1,\ \sigma^2)$, respectively.

The KL-divergence is given as

$$D_{KL}(p_1^u \parallel q_1^u) = \frac{\tau_1^2}{2\sigma^2}$$

If PCA is applied, the first principal component before and after the change has distributions $q_1^v \sim N(0,\ \lambda_1)$ and $p_1^v \sim N(\boldsymbol{\mu_2}^T \boldsymbol{v_1},\ \lambda_1) = N((\tau_1 + \tau_2)/\sqrt{2},\ \lambda_1)$, respectively. The KL-divergence is obtained as

$$D_{KL}(p_1^v \parallel q_1^v) = \frac{(\tau_1 + \tau_2)^2}{4\lambda_1}.$$

Clearly, we see that any positive τ_1 or τ_2 will give us a non-negative value of $D_{KL}(p_1^v \parallel q_1^v)$, which means that mean changes in the original data can be observed in the first principal component.

By using the Kullback–Leibler divergence, we have theoretically shown that changes in the original multidimensional time series are reflected in the first principal component. Accordingly, the problem of detecting changes in multidimensional time series is reduced to detecting changes in one-dimensional time series.

3.3 Change Point Analysis

The previous section has shown that the problem of detecting changes in multidimensional time series is reduced to detecting changes in one-dimensional time series after applying PCA to a multidimensional time series. In this section, we present statistical tests and change point detection methods for one single change point in a one-dimensional time series.

In offline setting, we concern mainly two problems: one is whether there is a change point, and the other is where it is most likely located if there exists such a point. As mentioned in Section 2, there are several methods to deal with offline change point detection. In this thesis, we adapt a cumulative sum-based change point algorithm developed first by Taylor [7] and later by Li [15] to detect a change in mean. To detect a change in variance, we make a slight modification to Taylor's algorithm by squaring the value of the data.

3.3.1 Cumulative Sum-Based Detection

Detection of Changes in Mean

To give an intuition of the cumulative sum tests, we use the first principal component dataset presented in Figure 6 (Section 3.2) to illustrate. The first principal component, denoted by $h(t)(1 < t < T)$, contains 200 observations, see the trace in Figure 7a. From this, the cumulative sums of the principal component values

are calculated as follows

$$CUSUM(t) = \sum_{k=1}^{t} h(k) \quad \text{for } k,\ t = 1, \cdots, T \tag{10}$$

where T is the length of a time series.

After the projection, we know that the origin coincides with the sample mean. The coordinates of the first principal component measure actually how far, on average, the data points are from the sample mean along the direction of the first principal component. This explains that the values of $h(t)$ sum to zero and why the cumulative sum over all values always ends at zero. The CUSUM curve for $h(x)$ is shown in Figure 7b. Since the principal component values range from negative value to positive value over time, it is expected that values added to the cumulative sum will be negative and the sum will steadily decrease. If a change occurs, the CUSUM curve will be broken into two segments with clearly different slopes. This means that there is a value that maximizes the contrast between these two segments and this value indicates a significant change in the mean value. We define the test statistic as the difference between the extreme values of $CUSUM(t)$, denoted by

$$T_D = CUSUM(t)_{max} - CUSUM(t)_{min} \tag{11}$$

To determine the significance of T_D, a threshold under the null hypothesis of no change needs to be derived. Here, we apply a permutation test which requires no knowledge of the underlying distribution. A detailed description of the test procedure is presented in Section 3.3.2.

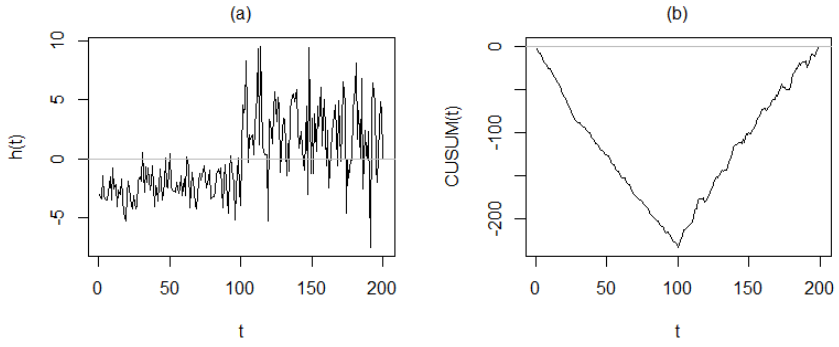


Figure 7: (**a**) The trace of the first principal component $h(t)$. (**b**) The CUSUM curve of $h(t)$.

Detection of Changes in Variance

To detect changes in variance, we modify slightly Taylor's CUSUM algorithm.

Instead of using the first principal component $h(t)$ directly, we use the squared value of $h(t)$, denoted by $h(t)^2$ (see the trace Figure 8a). From this, the Cumulative Sums of Squares of $h(t)$ are calculated:

$$CSS(t) = \sum_{k=1}^{t}((h(k))^2 - \overline{(h(k))^2}) \quad \text{for } k,\ t = 1, \cdots, T \tag{12}$$

$$\text{where} \qquad \overline{(h(k))^2} = \frac{\sum_{k=1}^{T}(h(k))^2}{T}$$

The squared values of $h(t)$ are first centered to zero mean such that the CSS are the CUSUM of the differences between the squared values of $h(t)$ and their average. Figure 8b displays the $CSS(t)$ curve. We see that the fluctuation of the $CSS(t)$ curve is as immense as the $CUSUM(t)$ curve (Figure 7b) and the slope shows a dramatic change indicating that the variance changes from a smaller value to a larger value. The difference between the extreme values of $CSS(t)$ is again defined as the test statistic. A statistical test can thereafter be applied to determine the significance of the difference.

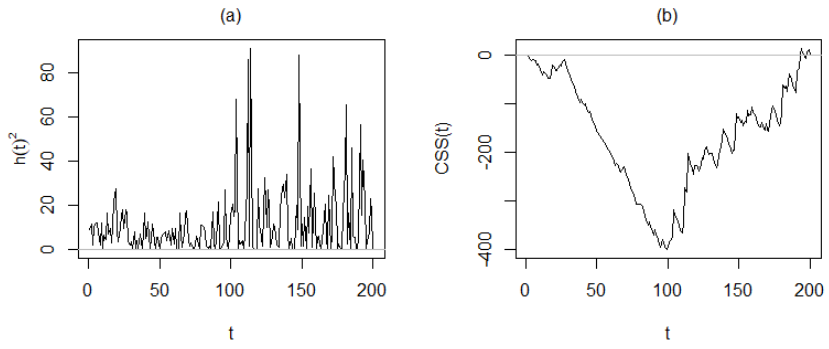


Figure 8: (**a**) The trace of the squared $h(t)$. **(b)** The CSS curve of $h(t)$.

3.3.2 Significance Testing the Existence of Change Points

The change point problem can be interpreted as a statistical hypothesis test, where the null hypothesis is that no change point exists versus the alternative hypothesis that there exists at least a change point. A permutation test is conducted to test whether a change in mean is statistically significant under the null hypothesis. We illustrate the procedure as follows.

1. Randomly permute the original data values i.e., $\mathbf{h} = \{h(t), t = 1, \cdots, T\}$ to generate a new permutation of $\{1, \cdots, T\}$, denoted by $h(t)_p$. From this, we can estimate how much the T_D value would vary if no change points existed. An example of the trace of $h(t)_p$ is shown Figure 9a.

2. Compute the $CUSUM(t)$ based on $h(t)_p$ and calculate thereafter the permuted test statistic, T_p. By comparing this value with the original difference T_D, we can check if this value is consistent with the situation when no change point exists. The CUSUM curve of the original time series $h(t)$ and CUSUM curves of 3 randomly permuted traces, $h(t)_p$ are displayed in Figure 9b. The CUSUM curves oscillate around 0 when no change occurs. The difference between the extreme CUSUM values of the permuted trace is much smaller compared to the CUSUM curve of the original time series.

3. Repeat (1)-(2) N (usually ~ 1000) times in order to obtain the distribution of T_p, i.e., $P(T_p) = \{(T_p)^{(1)}, (T_p)^{(2)}, \cdots, (T_p)^{(N)}\}$. We can then reject the null hypothesis and declare the existence of a change if T_D is greater than a predetermined threshold, e.g., at a significance level of 0.05.

We note that the distribution of T_p is based on N times randomly selected permutations rather than all possible permutations of the original time series. The total number of permutations ($T!$) grows out of control and this is not computation feasible. In general, setting $N = 1000$ times is sufficient for most situations [7].

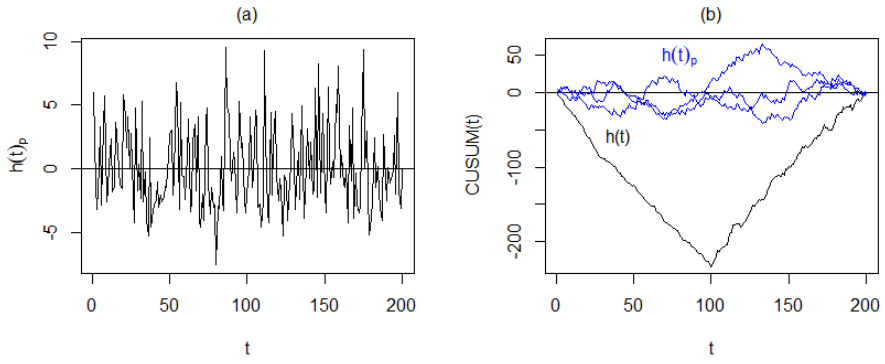


Figure 9: (**a**) A permutation of $h(t)$ obtained by randomly reordering its time order. (**b**) CUSUM curve of $h(t)$ and 3 randomly permuted traces $h(t)_p$, shown in blue color.

To test whether a change in variance is statistically significant, the same permutation test described above is implemented. Instead of generating the time order of the data values of $h(t)$, we randomly permute the time order of the squared $h(t)$, denoted as $(h(t)^2)_p$, see Figure 10a. This process is repeated N times. For each permutation, the $CSS(t)$ and the corresponding test statistic are calculated. In Figure 10b, we see that the CSS curve of the permuted traces (in blue color) fluctuate around 0 while the CSS curve of the original time series has a drastic change in direction. This sudden change indicates a change in variance.

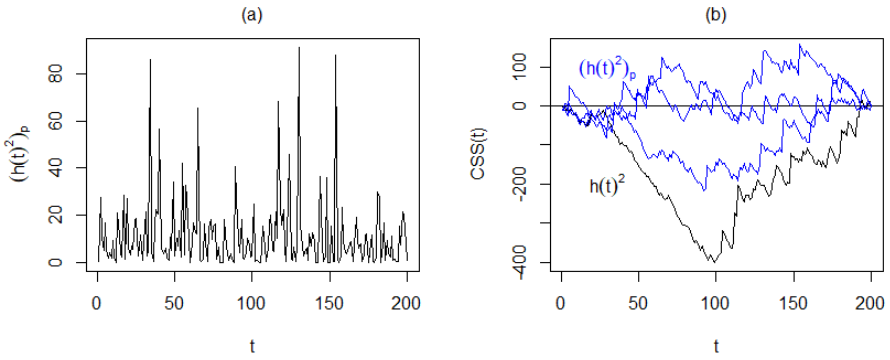


Figure 10: **(a)** A permutation of $h(t)^2$ obtained by randomly reordering its time order. **(b)** The CSS curve of $h(t)$ and 3 randomly permuted $h(t)^2$ traces, $(h(t)^2)_p$, shown in blue color.

3.3.3 Threshold Setting and p-value

In hypothesis testing, there are two ways to determine whether there is enough evidence from the sample to reject the null hypothesis. One approach is to specify a threshold T_h corresponding to a given significance level such that the null hypothesis is rejected if the given test statistic is greater than the threshold. The other approach is based on p-value and one can reject the null hypothesis if the p-value is smaller than the given significance level.

If the first approach is used, then the threshold is selected to bound the probability of type I error $\alpha \in (0, 1)$, also known as the rate of false positives, that is, the error of rejecting the null hypothesis when it is actually true. In change point detection, it is the probability of detecting a change when no change has occurred. In general, one tries to minimize the chance of type I error's occurrence as no one wants to accept an invalid hypothesis. Traditionally, one usually sets α equal to 0.05 or 0.01, which is also called significance level of the test [16]. In change point detection, one should choose the significance level carefully. If the chosen type I error value is too small, a large number of change points will not be detected. An extremely small value of type I error is suggested to avoid. It is worth noting that the type II error, the probability of missing to detect a change point even it actually does exist is not controlled by our cumulative sum-based algorithms. For example, for detecting a change in mean, the upper one-sided test is used because the test statistic T_D is always positive. The threshold value can be found such that the probability of finding $T_p \geq T_h$ (shaded area in Figure 11a) equals to the significance level or type I error, e.g., $\alpha = 5\%$. In the original time series, the null hypothesis assuming no change point is rejected if $T_D \geq T_h$.

In many statistical software packages, the second approach is more commonly used. Instead of comparing the test statistic T_D with the pre-specified threshold, the associated p-value is usually computed to be compared with a pre-determined significance level, α. The p-value is defined as the probability of obtaining an event that equals to or more extreme than the one in the data, assuming that the null hypothesis is true [16].

In this thesis, we use the p-value approach to determine the result. Consequently, a p-value will be computed from the permutation test we described above. As we do not consider all possible permutations, we will not obtain an exact p-value, but rather an approximate p-value which is calculated as follows:

$$\text{p-value} = \frac{\text{The number of } \{N : T_p^{(N)} \geq T_D\}}{N} \tag{13}$$

A histogram of the permuted test statistic T_p is displayed in Figure 11b. The blue dashed line indicates where the original sample falls ($T_D = 232.86$). Using definition (13) we obtain a p-value equal to 0. An approximate p-value of 0 indicates that the change in the mean value is significant.

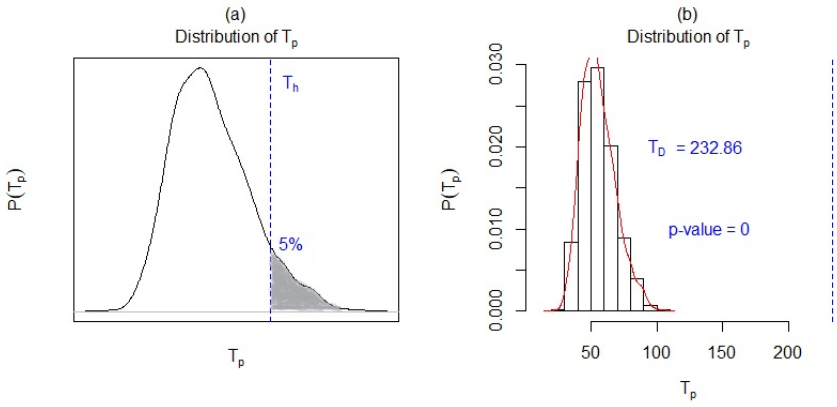


Figure 11: **(a)** The distribution of T_p obtained by generating random permutations from original dataset $h(t)$. **(b)** Histogram of the permuted test statistic T_p for 1000 permuted samples.

3.3.4 Estimating the Location of a Change Point and Its uncertainty

Once a change has been detected, the location of the most likely change point t^* $(1 < t^* < T)$ can be estimated. Here we present two methods for locating the most probable change point. One is the cumulative sum-based estimator and the other is the mean square error (MSE) estimator. We will conduct a small experiment to compare these two methods in a later section.

Cumulative Sum-based Estimator
For mean change detection, the most likely change point location is at which the maximum absolute value of cumulative sum $CUSUM(t^*)$ is attained [7]:

$$CUSUM(t^*) = \max_{t=1,\cdots,T} |CUSUM(t)| \tag{14}$$

For variance change detection, the most likely change point location is at which the maximum absolute value of the cumulative sum of squares ($CSS(t^*)$) is attained:

$$CSS(t^*) = \max_{t=1,\cdots,T} |CSS(t)| \tag{15}$$

$CUSUM(t^*)$ and $CSS(t^*)$ are the most extreme points on the CUSUM and CSS curves, respectively. The point t^* estimates the last point before the change. The point $t^* + 1$ estimates the first point after the change. For the example time series $h(t)$, the most extreme point on the CUSUM curve and CSS curve attains $CUSUM(100)$ and $CSS(99)$, respectively, see Figure 12.

MSE Estimator
For a given time t^* $(1 < t^* < T)$, the mean square error (MSE) estimator is defined as [7]:

$$MSE(t^*) = \sum_{k=1}^{t^*} \left(h(k) - \overline{h(k)_1}\right)^2 + \sum_{k=t^*+1}^{T} \left(h(k) - \overline{h(k)_2}\right)^2 \tag{16}$$

$$\text{where} \quad \overline{h(k)_1} = \frac{\sum_{k=1}^{t^*} h(k)}{t^*} \quad \text{and} \quad \overline{h(k)_2} = \frac{\sum_{k=t^*+1}^{T} h(k)}{T - t^*}$$

The idea behind the MSE estimator is as follows: By separating the time series $h(t)$ at $t = t^*$ $(t^* = 3, \cdots, T-2)$ into two segments and then estimating the mean of each segment, one can estimate how well the two estimated mean values fit the data. The best estimator of the change point location t_{ch} is given as the time where $MSE(t^*)$ reaches its minimum. Figure 13a shows the resulting $MSE(t^*)$ for the example time series $h(t)$. The time point that minimizes $MSE(t^*)$ is 100 for detecting a change in mean. For estimating a change point location for a detected change in variance, one can simply substitute $h(t)$ by $h(t)^2$. A plot of the resulting $MSE(t^*)$ for the squared $h(t)$ is shown in Figure 13b.

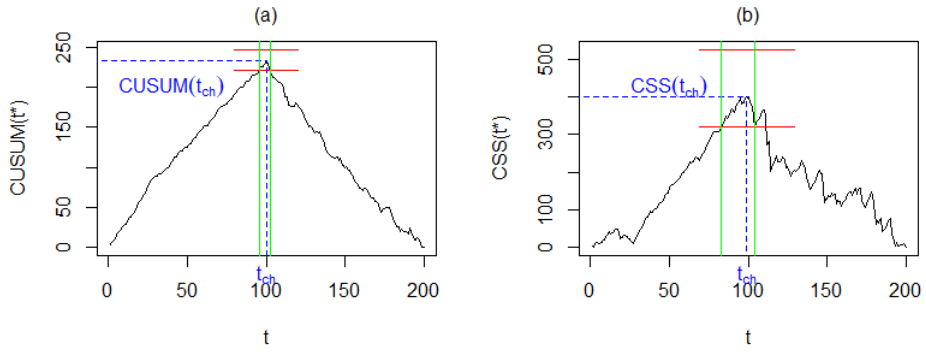


Figure 12: Determining the most likely change point location using the cumulative sum-based estimator. **(a)** Plot of the absolute value of the CUSUM for $h(t)$. **(b)** Plot of the absolute value of the CSS for $h(t)$. The vertical dashed blue line in both plots indicates the estimated change point location where $CUSUM(t^*)$ or $CSS(t^*)$ obtains its maximum absolute value. The two red lines represent a 68% percentiles bootstrap confidence interval for $CUSUM(t_{ch})$ or $CSS(t_{ch})$ (error bars). The uncertainty in corresponding change point location is indicated by the green lines which are determined as the time interval whose absolute CUSUM and CSS values are enclosed by the confidence interval of $CUSUM(t_{ch})$ and $CSS(t_{ch})$, respectively.

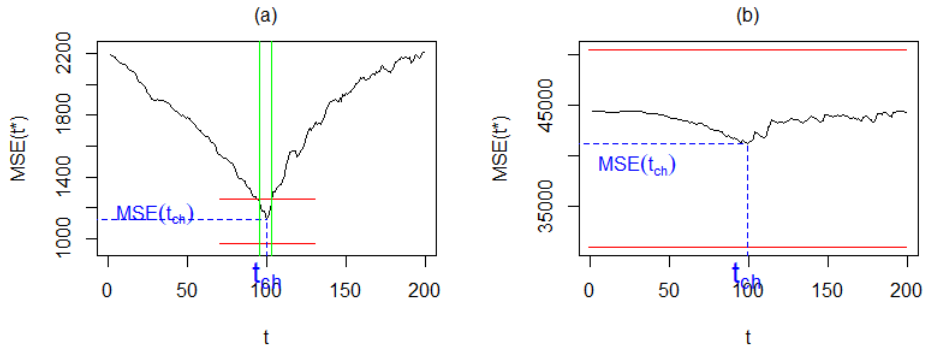


Figure 13: Determining the most likely change point location using the mean square error estimator. **(a)** Plot of the resulting $MSE(t^*)$ for $h(t)$. **(b)** Plot of the resulting $MSE(t^*)$ for $h(t)^2$. The vertical dashed blue line in both plots indicates the estimated change point location where $MSE(t^*)$ obtains its minimum value. The two red lines represent a 68% percentiles bootstrap confidence interval for $MSE(t_{ch})$ (error bars). The uncertainty in corresponding change point location is indicated by the green lines which are determined as the time interval whose MSE values are enclosed by the confidence interval of $MSE(t_{ch})$.

Uncertainty Estimation

When evaluating $CUSUM(t^*)$, $CSS(t^*)$ or $MSE(t^*)$ for determining a change point location, one can hardly avoid a sampling error. Here we present a method based on bootstrapping, a process of resampling with replacement, to estimate the uncertainty in the estimated change point location (t_{ch}) due to this error, which is equivalent to construct a confidence interval for the corresponding estimate using the bootstrap [15]. By determining the uncertainty in the change point location, we can see how prominent a change point it is. Suppose that the change point location is determined to be at $t^* = t_{ch}$ and the data is divided into two segments: $\{h(1), \cdots, h(t_{ch})\}$ and $\{h(t_{ch}+1), \cdots, h(T)\}$. For illustration, we let the value of the $CPL(t_{ch})$ represent any of the values of $\{CUSUM(t_{ch}), CSS(t_{ch}), MSE(t_{ch})\}$. The bootstrap procedure for estimating the uncertainty in $CPL(t_{ch})$ is implemented as follows:

1. Resample the first segment $\{h(1), \cdots, h(t_{ch})\}$ with replacement.
2. Resample the second segment $\{h(t_{ch}+1), \cdots, h(T)\}$ with replacement.
3. Apply formula (14), (15) or (16) to evaluate the bootstrap based $CPL(t_{ch})$. Observe that one needs to first combine the first segment from step 1 with the second segment from step 2 before applying formula (14) or (15) when evaluating the bootstrap based $CUSUM(t_{ch})$ or $CSS(t_{ch})$.
4. Repeat $(1)-(3)$ independently N (i.e., 1000) times to obtain a bootstrap distribution of $CPL(t_{ch})$, i.e. $\{CPL(t_{ch})^{(1)}_{boot}, \cdots, CPL(t_{ch})^{(N)}_{boot}\}$.
5. A two-sided $(1-\alpha)\times 100\%$ bootstrap confidence interval can then be constructed for $CPL(t_{ch})$, e.g., using the standard error (SE) or the percentiles of the bootstrap distribution if the bootstrap distribution is approximately smooth and symmetric. Estimation of the standard error to some point estimate is simply the standard deviation of the bootstrap distribution [17]. For instance, a 68% confidence interval for the mean square error estimator is generated by $CPL(t_{ch}) \pm 1.0 \times \text{SE}$. A percentile bootstrap confidence interval is created by selecting the endpoints from the middle of the bootstrap distribution corresponding to e.g., 68% confidence level.

The constructed bootstrap confidence intervals can be displayed graphically as error bars (shown as red lines in Figure 12 and Figure 13). If the value of $CPL(t^*)$ falls inside the confidence interval of $CPL(t_{ch})$, any time instant t^* closed to t_{ch} can be a potential change point. Therefore, the uncertainty in the change point location can be determined as the time interval whose CPL values are enclosed by the confidence interval of $CPL(t_{ch})$, shown as the double green lines in Figure 12 and Figure 13. It can be easily seen in both figures that a narrower confidence interval for the value of change point location estimator gives rise to a smaller uncertainty in the change point location. In Figure 13b we notice that the error bar for the value of $MSE(t_{ch})$ is surprisingly big which

means that there is large uncertainty in the change point location. This looks unreasonable. At this moment, we do not know the answer. An investigation into the problem will be made in later sections.

4 Simulation Study

We consider three scenarios in which three different types of changes take place in a two-dimensional time series:

1. Change in mean only
2. Change in variance only
3. Simultaneous change in mean and variance

Let $N(\boldsymbol{\mu}, \Sigma)$ denote the bivariate normal distribution with mean vector $\boldsymbol{\mu} = [\mu, \mu]^T$ and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$ where σ_1, σ_2 are standard deviations and ρ, $-1 < \rho < 1$ is the correlation coefficient between variables. Each simulation applies the CUSUM or CSS method to a normally distributed time series with one single change point after performing the principal component analysis. Throughout the study, observations before and after the occurrence of a change are generated from $\boldsymbol{N_1}\left(\boldsymbol{\mu_1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \Sigma_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$, and $\boldsymbol{N_2}\left(\boldsymbol{\mu},\ \Sigma_1\right)$ or $\boldsymbol{N_2}\left(\boldsymbol{\mu_1},\ \Sigma_\sigma\right)$, respectively. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma_\sigma}$ varies in each simulation depending on the type of changes we want to study. Since we do not detect changes in correlation, the value of ρ is set to 0.5 for all simulations.

The analysis of data, both simulated and real life, are performed using the statistical software R (R Development Core Team) [18]. We used $N = 1000$ iterations when performing the permutation test, which was conducted at a $\alpha = 0.05$ significance level.

4.1 Scenarios Study

A variety of data according to the three scenarios described above are simulated to study how well the change point detection methods work. The PCA is applied to all of the data before performing the change point analysis. We should emphasize here again that the focus of our methods is to test the existence of a change point and then find the most prominent change point location if the existence of a change point is shown to be statistically significant. For this reason, we will study how different sizes of parameter change affects the detection of a change point and what factors affect the accuracy of a change point.

4.1.1 Change in Mean

We simulate two sets of time series without replications to study how well the CUSUM method works.

Set 1: Four 2D time series are generated with distributions $\boldsymbol{N_1}(\boldsymbol{\mu_1},\ \Sigma_1)$, $\boldsymbol{N_2}(\boldsymbol{\mu},\ \Sigma_1)$ where $\boldsymbol{\mu} = [1,1]^T, [2,2]^T, [3,3]^T$ or $[4,4]^T$, respectively. The sample size $T = 300$ is fixed for each simulation.

Set 2: Three 2D time series are generated as $\boldsymbol{N_1}(\boldsymbol{\mu_1},\ \Sigma_1)$, and $\boldsymbol{N_2}(\boldsymbol{\mu},\ \Sigma_1)$ where $\boldsymbol{\mu} = [2,2]^T$. The sample sizes vary at $T = \{200, 400, 600\}$.

The approximated p-value obtained by the permutation test, the change point location using two different methods ($CUSUM(t^*)$ and $MSE(t^*)$ estimator) and the 95% percentiles bootstrap confidence interval for the time of change which indicates how well the time of change has been pinpointed, are recorded in Table 1 and Table 2. To easily compare the effect of the sample size, we normalize the time intervals in the range $[0, 1]$ in Table 2 by dividing confidence limits by the sample size T.

Table 1 shows clearly that the CUSUM method can successfully detect the existence of a change point even with a small change in mean, though with a wider confidence interval. As the size of mean change increases, the confidence interval for the change point location is getting narrower using both methods (see the last two 95% Time Interval columns). Also, shown by Figure 14 and Figure 15, the $CUSUM(t^*)$ and $MSE(t^*)$ curves are getting steeper with the increase of mean change. This means that the change point location can be more accurately pinpointed with a larger mean change. We note that the size of the confidence interval for the time of change depends on the sample size of the data, as shown by the normalized time interval in Table 2. In Figure 16 and Figure 17, we see that the size of confidence intervals becomes smaller as the increase of the sample size.

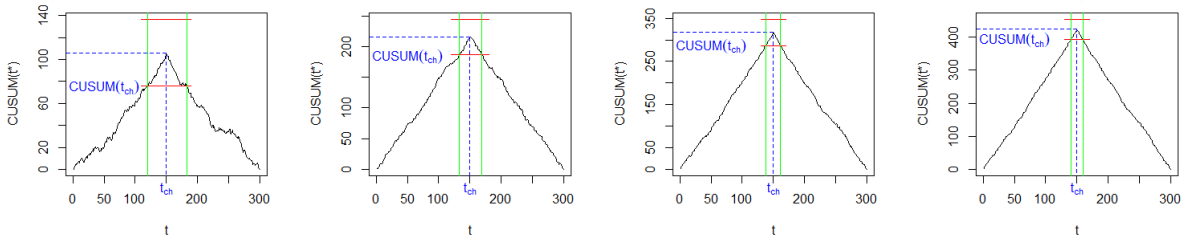
Moreover, we observe that the change point locations determined by the $MSE(t^*)$ estimator results in less uncertainty compared to the $CUSUM(t^*)$ estimator (see Figure 14, 15, 16, 17). This may simply be because of the sampling error as the bootstrapping method will not produce identical results each time it is performed. It could also be because it does not have an easy true correspondent between extreme values of the fluctuation with the change points if the $CUSUM(t^*)$ estimator is used, as several extreme values may occur due to the sampling error. In such a case, using $CUSUM(t^*)$ estimator will result in a relatively larger uncertainty in the change point location compared to the $MSE(t^*)$ estimator which estimates how well two estimated mean values fit the data each time the data is split into two segments.

μ	Significance (p-value)	Location $CUSUM(t^*)$	Location $MSE(t^*)$	95% Time Interval $CUSUM(t_{ch})$	95% Time Interval $MSE(t_{ch})$
1	0.00	150	150	(120, 183)	(120, 171)
2	0.00	150	150	(133, 169)	(144 , 158)
3	0.00	150	150	(137, 161)	(147, 153)
4	0.00	150	150	(140, 159)	(148, 152)

Table 1: Results of change point analysis on four 2D time series which undergo a mean change at 150. Each time series consists of 300 observations, with distributions $\boldsymbol{N_1}(\boldsymbol{\mu_1}, \Sigma_1)$, $\boldsymbol{N_2}(\boldsymbol{\mu}, \Sigma_1)$ where $\boldsymbol{\mu} = [1,1]^T, [2,2]^T, [3,3]^T$ or $[4,4]^T$, respectively. The 95% time interval is enclosed by the 95% percentiles bootstrap confidence interval for the value of $CUSUM(t_{ch})$ or $MSE(t_{ch})$.

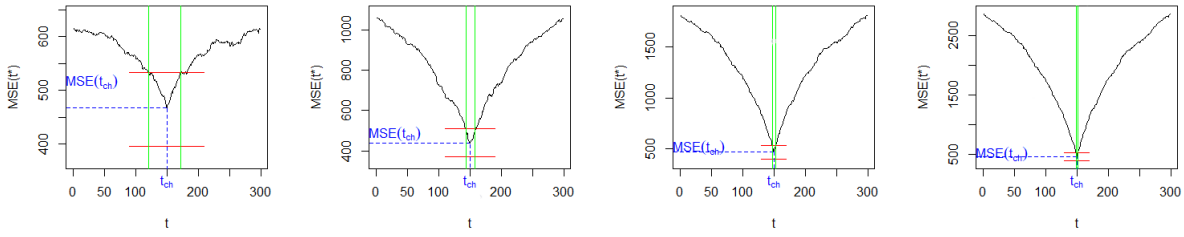
T	Significance (p-value)	Location $CUSUM(t^*)$	Location $MSE(t^*)$	95% Time Interval $CUSUM(t_{ch})$ (normalized)	95% Time Interval $MSE(t_{ch})$ (normalized)
200	0.00	102	102	(87, 114) (0.44, 0.57)	(92, 108) (0.46, 0.54)
400	0.00	200	200	(174, 223) (0.44, 0.56)	(187 , 209) (0.47, 0.52)
600	0.00	300	300	(275, 325) (0.46, 0.54)	(285, 313) (0.48, 0.52)

Table 2: Results of change point analysis on three 2D time series which undergo a mean change at 100, 200 and 300, respectively. All these three time series have distributions before and after a change $\boldsymbol{N_1}(\boldsymbol{\mu_1}, \Sigma_1)$, and $\boldsymbol{N_2}(\boldsymbol{\mu}, \Sigma_1)$ where $\boldsymbol{\mu} = [2,2]^T$, respectively. The 95% time interval is enclosed by the 95% confidence interval for the value of $CUSUM(t_{ch})$ or $MSE(t_{ch})$. The second range in the time interval cells are normalized confidence intervals with respect to the sample sizes.



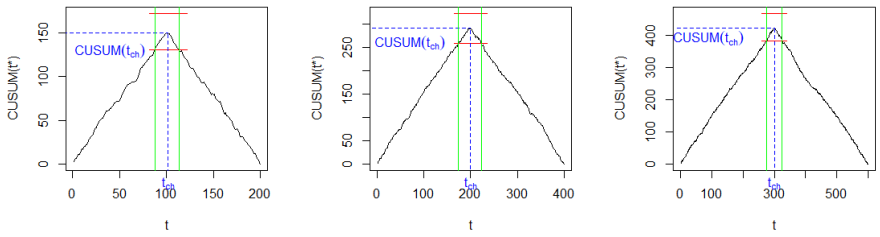
(a) $\mu = 1$ (b) $\mu = 2$ (c) $\mu = 3$ (d) $\mu = 4$

Figure 14: Determining the change point location using the $CUSUM(t^*)$ estimator and its uncertainty using 95% bootstrap confidence interval for four time series with various mean change. (a)-(d) are resulting $CUSUM(t^*)$. The red lines represent the error bars and the green lines represent the time interval enclosed by the error bars.



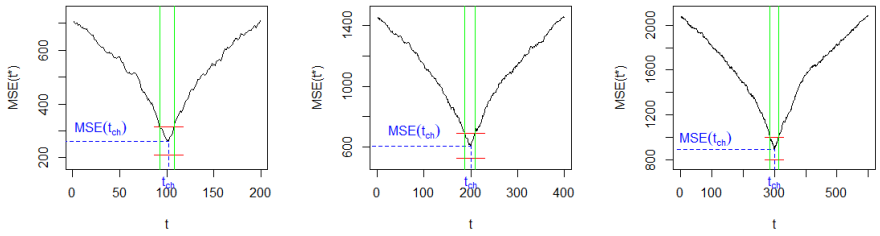
(a) $\mu = 1$ (b) $\mu = 2$ (c) $\mu = 3$ (d) $\mu = 4$

Figure 15: Determine the change point location using the $MSE(t^*)$ estimator and its uncertainty using 95% bootstrap confidence interval for four time series with various mean change. (a)-(d) are resulting $MSE(t^*)$.



(a) $T = 200$ (b) $T = 400$ (c) $T = 600$

Figure 16: Determining the change point location using the $CUSUM(t^*)$ estimator for three time series with sample size $T = \{200, 400, 600\}$ and its uncertainty using 95% bootstrap confidence interval. (a)-(d) are resulting $CUSUM(t^*)$.



(a) $T = 200$ (b) $T = 400$ (c) $T = 600$

Figure 17: Determining the change point location using the $MSE(t^*)$ estimator for three time series with sample size $T = \{200, 400, 600\}$ and its uncertainty using 95% bootstrap confidence interval. (a)-(c) are resulting $MSE(t^*)$ curves. All time series have the same distributions.

4.1.2 Change in Variance

The same type of simulation study is also performed to evaluate how well the CSS method performs.

Set 1: Four time series are generated with distributions $\boldsymbol{N_1}(\boldsymbol{\mu_1},\ \Sigma_1)$ and $\boldsymbol{N_2}(\boldsymbol{\mu_1},\ \Sigma_\sigma)$ where $\sigma_1^2 = \sigma_2^2 = \{3, 6, 9, 12\}$, respectively. The sample size $T = 300$ is fixed for each simulation.

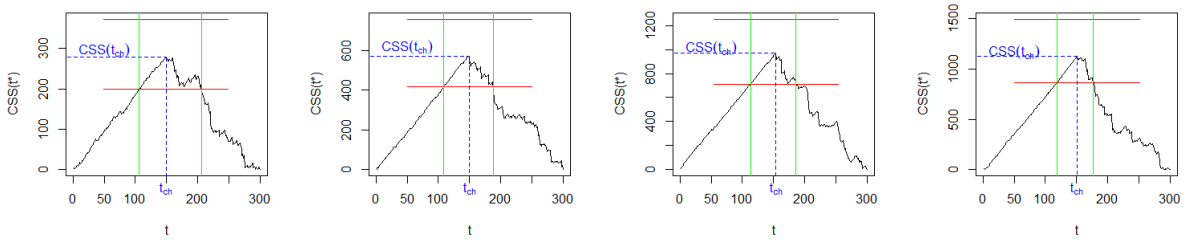
Set 2: Three time series are generated as $\boldsymbol{N_1}(\boldsymbol{\mu_1},\ \Sigma_1)$ and $\boldsymbol{N_2}(\boldsymbol{\mu_1},\ \Sigma_\sigma)$ where $\sigma_1^2 = \sigma_2^2 = 6$. The sample sizes vary at $T = \{300, 600, 900\}$.

Table 3 shows that the CSS method can also successfully detect the existence of a change in variance even with a small change in variance. As the size of variance change increases, the confidence interval for the change point location is getting narrower using the $CSS(t^*)$ estimator (see the column of the 95% Time Interval for $CSS(t_{ch})$ and Figure 18). We observe also that the size of the confidence interval for the value of $CSS(t_{ch})$ becomes smaller as the sample size increases (see Figure 20). Both error bar and its corresponding time interval are getting narrower with an increase in the sample size.

On the other hand, the $MSE(t^*)$ estimator appears to be an unreliable method to locate a change point in variance. As shown by Figure 19, the confidence intervals for the value of $MSE(t_{ch})$ are so extremely wide that the error bars cannot even touch the $MSE(t^*)$ curves. This means that using the $MSE(t^*)$ estimator to determine the change point location in variance will result in large uncertainty in the determined location. This is quite unexpected. In Figure 19, the values of the error bar increase dramatically as the size of the variance increases. We suspect that it may be because of the simulation setting itself. Other data simulations may give more reasonable results. However, due to time limitations, we will not carry out further investigation into this problem at this moment.

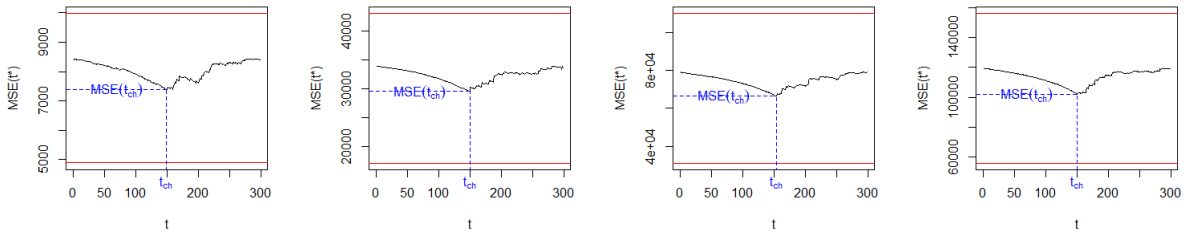
σ^2	Significance (p-value)	Location $CSS(t^*)$	Location $MSE(t^*)$	95% Time Interval $CSS(t_{ch})$
3	0.00	149	149	(105, 206)
6	0.00	150	150	(108, 188)
9	0.00	154	154	(113, 186)
12	0.00	151	151	(118, 177)

Table 3: Results of change point analysis on four 2D time series which undergo a variance change at 150. Each time series consists of 300 observations, with distributions $\boldsymbol{N_1}(\boldsymbol{\mu_1},\ \Sigma_1)$, $\boldsymbol{N_2}(\boldsymbol{\mu_1},\ \Sigma_\sigma)$ where $\sigma_1^2 = \sigma_2^2 = \{3, 6, 9, 12\}$, respectively. The 95% time interval is enclosed by the 95% confidence interval for the value of $CSS(t_{ch})$.



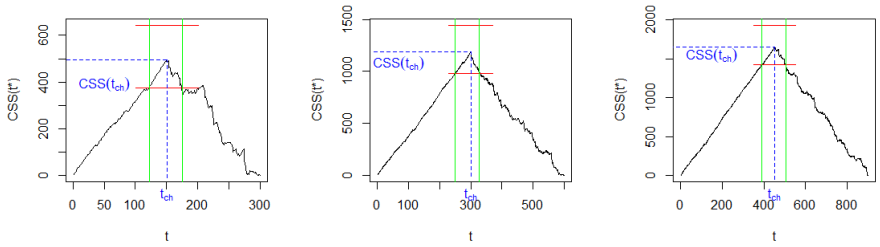
(a) $\sigma^2 = 3$ (b) $\sigma^2 = 6$ (c) $\sigma^2 = 9$ (d) $\sigma^2 = 12$

Figure 18: Determining the change point location using the $CSS(t^*)$ estimator and its uncertainty using 95% bootstrap confidence interval. (a)-(d) are resulting $CSS(t^*)$.



(a) $\sigma^2 = 3$ (b) $\sigma^2 = 6$ (c) $\sigma^2 = 9$ (d) $\sigma^2 = 12$

Figure 19: Determining the change point location using the $MSE(t^*)$ estimator and its uncertainty using 95% bootstrap confidence interval. (a)-(d) are resulting $MSE(t^*)$.



(a) $T = 300$ (b) $T = 600$ (c) $T = 900$

Figure 20: Determining the change point location using the $CSS(t^*)$ estimator and its uncertainty using 95% bootstrap confidence interval. (a)-(c) are resulting $CSS(t^*)$. Sample sizes vary at $T = \{300, 600, 900\}$

4.1.3 Simultaneous Change in Mean and Variance

The cumulative sum (CUSUM) and cumulative sum of squares (CSS) methods are designed to test changes only in mean and variance respectively. Nevertheless one can combine these two methods to find a simultaneous change in mean and variance. This means that we apply both of the methods to a time series individually. If only one of the methods gives a positive result (e.g. there exists a statistically significant change point in the time series.), the time of the detected change point is the most likely change point location. On the other hand, if both methods give a positive result, only one of the change point location is identified as the most prominent change point location. The confidence interval of the change point location which tells the uncertainty the change point location is to be used as a criterion here. The change point with less uncertainty in the estimated change point location is selected as the most prominent change point in the time series. Now we provide an example to show how this combined method works.

Let a 2D time series consisting of 300 observations has distributions before and after a change: $N_1\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$ and $N_2\left(\begin{pmatrix}3\\3\end{pmatrix}, \begin{pmatrix}6 & 3\\3 & 6\end{pmatrix}\right)$, respectively. The traces of the bivariate time series and the 1st PC are shown in Figure 21. As expected, both CUSUM and CSS methods show that there is a significant change in mean and variance, respectively. We use the cumulative sum-based estimator to determine the change point location and find that the change point location determined by $CUSUM(t^*)$ estimator is 150 while the change point location determined by $CSS(t^*)$ estimator is 151. It seems that either of the change points can be the simultaneous change point location as they are so close to each other. However, if we look at Figure 22, the confidence interval for the value of $CSS(151)$ is wider. This indicates that the time of the change in variance cannot be accurately pinpointed. The change point location determined by the $CUSUM(t^*)$ estimator is more prominent and therefore 150 is chosen as the most prominent change point in mean and variance for this simulation.

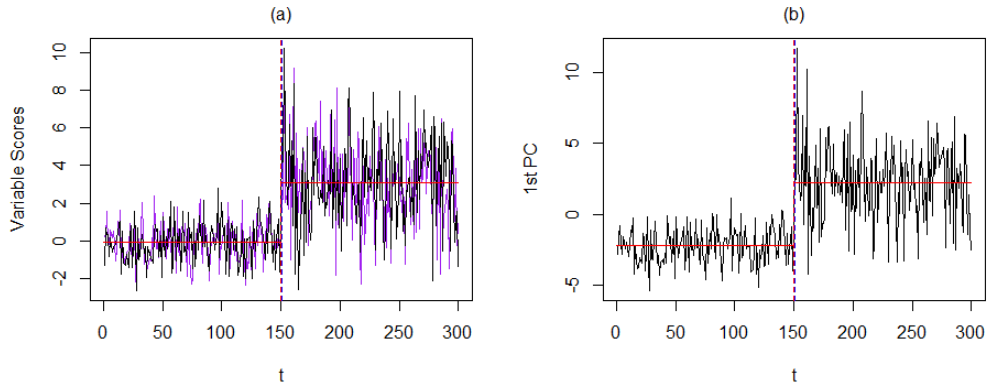
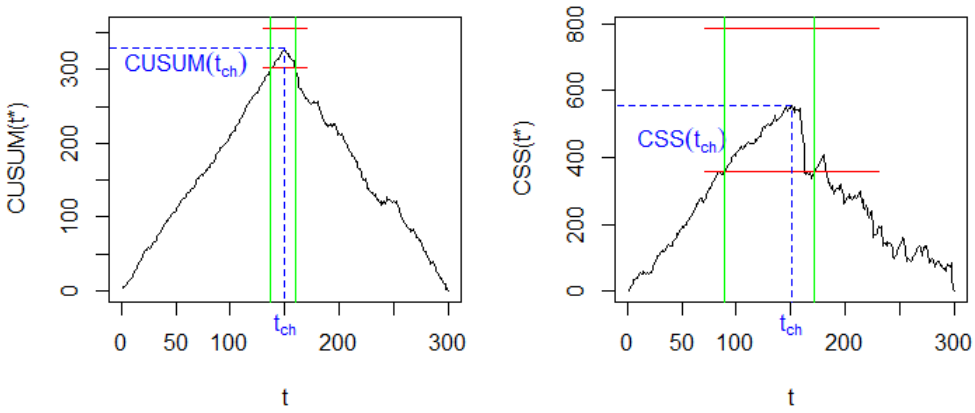


Figure 21: **(a)** The trace of the 2D time series. **(b)** The trace of the 1st PC. The vertical dashed red line indicates the change point location at 150 determined by $CUSUM(t^*)$ estimator. The vertical dashed blue line indicates the change point location at 151 determined by $CSS(t^*)$ estimator.



(a) $CUSUM(t^*)$ curve (b) $CSS(t^*)$ curve

Figure 22: **(a)**The resulting $CUSUM(t^*)$ curve for the value of $CUSUM(150)$ and its uncertainty. **(b)**The resulting $CSS(t^*)$ curve for the value of $CSS(151)$ and its uncertainty. The red lines represent the 95% error bars determined by bootstrapping and the green lines represent the corresponding time intervals.

4.2 Performance Evaluation of the Two Change Point Location Estimators

In the previous section, we have seen that the $MSE(t^*)$ estimator slightly outperforms the $CUSUM(t^*)$ estimator for locating a change point in a time series with a mean change in terms of the uncertainty in the determined location. On the other hand, the $CSS(t^*)$ estimator is much better than $MSE(t^*)$ estimator for determining a change point in a time series with a variance change. Here we use other criteria to evaluate the performance of these two estimators.

This time, we generate a set of independent time series with 1000 replications for each simulation setting. Each time series consists of 300 observations with a true change point location at 150 as before. The number of all detected change points, denoted by $\boldsymbol{n_1}$ is noted. We use the number of correctly identified change point locations, denoted by $\boldsymbol{n_2}$, and the number of incorrectly identified change point locations, denoted by $\boldsymbol{n_3}$ as the criteria [19]. Moreover, the number of almost correctly identified change point locations which differ three observations from the true change point location, denoted by $\boldsymbol{n_4}$ are also calculated. That is, for a true change point locating at 150, the almost correctly identified change point locations are defined to be in this range: $\{147, 148, 149, 151, 152, 153\}$. Note that six points are designed to be an acceptable range in this experiment. In general, this range may not be suitable for other experiments. By including the number of almost correctly detected change point locations, we can get an overview of how close an incorrectly identified change point location lies to the true change point location.

The simulation setting is the same as previous sections: various sizes of the parameter changes are assigned to see how the size of change affects the ability of locating the change point. For change in mean, the distribution after the change is given as $\boldsymbol{N_2}(\boldsymbol{\mu},\ \Sigma_1)$ with $\boldsymbol{\mu} = [1,1]^T, [2,2]^T, [3,3]^T$ or $[4,4]^T$. For change in variance, the distribution after the change is set to $\boldsymbol{N_2}(\boldsymbol{\mu_1},\ \Sigma_\sigma)$ with $\sigma_1^2 = \sigma_2^2 = \{3, 6, 9, 12\}$, respectively. The simulation results are concluded in Table 4 and Table 5.

Change in Mean							
μ	#	$CUSUM(t^*)$	$MSE(t^*)$	μ	#	$CUSUM(t^*)$	$MSE(t^*)$
1	n_1	1000	1000	3	n_1	1000	1000
	n_2	365	355		n_2	915	912
	n_3	635	645		n_3	85	88
	n_4	433	430		n_4	84	87
2	n_1	1000	1000	4	n_1	1000	1000
	n_2	733	730		n_2	978	978
	n_3	267	270		n_3	22	22
	n_4	251	253		n_4	22	22

Table 4: Detected change point locations from 1000 simulations using $CUSUM(t^*)$ and $MSE(t^*)$ estimator after having tested the existence of one single change point using cumulative sum (CUSUM) method on 1st PCs. Each sample has $T = 300$ observations. The distributions before and after the change are $\boldsymbol{N_1}(\boldsymbol{\mu_1}, \Sigma_1)$ and $\boldsymbol{N_2}(\boldsymbol{\mu}, \Sigma_1)$ with $\boldsymbol{\mu} = [1,1]^T, [2,2]^T, [3,3]^T$ or $[4,4]^T$, respectively.

As it is shown in Table 4, the CUSUM test is truly reliable for testing the existence of changes in mean value because the number of all detected change points ($\boldsymbol{n_1}$) is 1000 dispite of different sizes of mean shifts. Given the existence of a change point in mean, the $CUSUM(t^*)$ and $MSE(t^*)$ estimators for locating the most likely change point have similar performance across different size of mean change using both criteria. Both methods give approximately the same number of correctly identified change point locations ($\boldsymbol{n_2}$) and the number of incorrectly detected change point locations ($\boldsymbol{n_3}$). For small changes (e.g., $\mu = 1$), both estimators are not able to locate a large amount of true change point. Only around 36% of the detected change points have the true location at 150. As the size of mean shift increases, the number of correctly identified true change point locations is increasing. For a mean change larger than 2, more than 90% of the detected change points have the true location and almost all of the incorrectly identified change point locations lie pretty close to the true change point location ($\boldsymbol{n_3} \approx \boldsymbol{n_2}$ within a difference of three observations).

Change in Variance							
σ^2	#	$CSS(t^*)$	$MSE(t^*)$	σ^2	#	$CSS(t^*)$	$MSE(t^*)$
3	n_1	1000	1000	9	n_1	1000	1000
	n_2	166	149		n_2	285	268
	n_3	834	851		n_3	715	732
	n_4	338	301		n_4	372	347
6	n_1	1000	1000	12	n_1	1000	1000
	n_2	255	233		n_2	316	297
	n_3	745	767		n_3	684	703
	n_4	370	345		n_4	362	351

Table 5: Detected change points from 1000 simulations using $CSS(t^*)$ or $MSE(t^*)$ estimator after having tested the existence of one single change point using cumulative sum squares (CSS) method on 1st PC. Each sample has $T = 300$ observations. The distributions before and after the change are $\boldsymbol{N_1}(\boldsymbol{\mu_1},\ \Sigma_1)$ and $\boldsymbol{N_2}(\boldsymbol{\mu_1},\ \Sigma_\sigma)$ with $\sigma^2 = \{3, 6, 9, 12\}$, respectively.

For testing the existence of a change in variance, Table 5 shows that the CSS method is also reliable. At the significance level of 0.05, the null hypothesis was rejected for each of the simulations. In terms of locating the exact change point, the $CSS(t^*)$ estimator appears to slightly outperform the $MSE(t^*)$ estimator since more true change point locations can be identified by the $CSS(t^*)$ estimator across different sizes of variance change. However, neither $CSS(t^*)$ nor $MSE(t^*)$ estimator seems to be a very favourable method to determine the true change point location in general. Even the size of variance change is 12, both methods fail to detect a large amount of the true change point locations, and around 50% of the number of incorrectly identified change point locations do not lie into the 6 points acceptable range.

4.3 Computational Cost Analysis

The computational cost of the whole change point analysis depends mainly on the following three subroutines:

1. The PCA routine for reducing multidimensional data to a one-dimensional data, which is called every time before performing the change point analysis, has a quadratic complexity concerning the data dimensionality. Since we only want to extract the first PC associated with the largest eigenvalue, the data dimensionality reduction has relatively not too much effect on the running time of the whole change point analysis scheme.

2. Both CUSUM and CSS methods for testing the existence of a change point have en inefficient running time due to performing the computer-intensive permutation test. Although we have avoided the factorial increase in the number of permutations by using approximate permutation, a large number of repeated tests are still needed to achieve satisfactory accuracy. 1000

random permutations were used in our analyses and gave us quite reliable results. Fewer permutations can be specified but this could lead to less accuracy of the methods. Therefore, the computational cost of our scheme is effected more by this step.

3. Two methods to locate the most probable change point were introduced. Depending on which method to be used, the running time of the scheme can differ a lot. The $MSE(t^*)$ estimator is not consistent with the change point existence testing methods and performed independently only when a change is reported. A mean square error is calculated each time the data is split into two segments. The time complexity of the method is linearly depending on the length of the time series. On the other hand, the $CUSUM(t^*)$ and $CSS(t^*)$ estimators are consistent with the corresponding change-point existence testing methods. The maximum absolute value of the CUSUM or CSS, where the change point is located, is attained almost by free. Therefore their running time on locating the change point is remarkably more effective than the $MSE(t^*)$ estimator.

5 Application to Single Molecule Data

The data we used in the simulation study were generated from the normal distribution. But in real life, the distribution of the data may not be known. Since the change point analysis scheme we proposed here is designed to distributional-free data. It is worthy to apply the scheme on a non-normally distributed real life time series and see how it performs.

5.1 Single Molecule Data

The real life data describe rotary traces obtained by observing freely rotary motions of a rotary motor protein, single F_1-ATPase (F_1) [15]. The rotations of F_1 robustly construct the dwells statistics which is displayed graphically in Figure 23b. One can see that there are several abrupt jumps over time, and therefore change point analysis can be applied to detect the dwell-time. The given time series consists of three types of dwells statistics which are resulted by the rotations: x-coordinates, y-coordinates, and angle. Detailed information about the experiment of observing the rotary motions of F_1-ATPase and how the data were collected can be found in Li *et al*[15]. Here we extract the first 1000 observations from the x, y-coordinates variables and study only its general statistical properties from the perspective of change point analysis. The interpretation of the results related to the experiment itself will not be made.

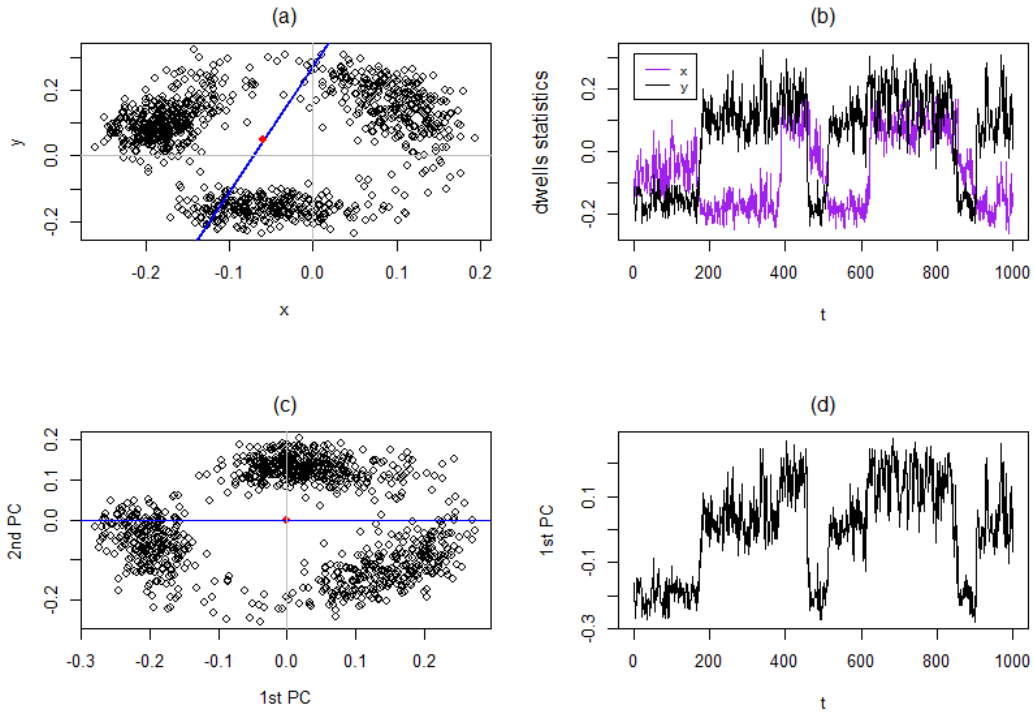


Figure 23: A subset of the two-dimensional time series observed in single molecule experiments: the first 1000 observations. (a) A scatterplot for the original data with a blue line indicating the direction of the first principal component. (b) The trace of the original 2D time series. (c) A scatterplot of 1st PC versus 2nd PC. (d) The trace of the 1st PC.

5.2 Data Analysis

Figure 23a above shows the scatterplot of the extracted two-dimensional time series and the traces are plotted in Figure 23b. We can see by eyes that there exist multiple changes in both mean and variance. The principal component analysis is applied to the data to obtain the principal components. Figure 23c shows that the 1st PC captures the majority of the variation in the data. We observe that possible changes in the original 2D time series are reflected in the trace of 1st PC, see Figure 23d.

Next, we apply the permutation-based CUSUM and CSS methods individually to detect if there exists a change in mean and/or variance of the variables. It turns out that both tests give statistically significant results. Then, the change point location estimators are calculated to determine the change points. We apply both methods and find that the change point in mean determined by $MSE(t^*)$ and $CUSUM(t^*)$ are at 177 and 178, respectively. The change point in variance determined by $MSE(t^*)$ and $CSS(t^*)$ are at 173 and 174, respectively. In this case, both methods locate approximately the same change point. If we look at their uncertainty in Figure 24, we find that the cumulative sum-based estimator gives a smaller uncertainty in the estimated change point location (compare plot (a), (c) with plot (b), (d), respectively). For this reason, we use the estimated locations determined by the cumulative sum based estimator (mean change at 178 and variance change at 173). Lastly, we compare the uncertainty in the value of $CUSUM(178)$ (Figure 24b) with the uncertainty in the value of $CSS(173)$ (Figure 24b). We see that the confidence interval for the value of $CUSUM(178)$ is slightly narrower. Therefore, the time point 178 is chosen as the most prominent change point location for a simultaneous change in mean and variance, see Figure 25 with the determined change point location on the original 2D traces and the 1st PC trace.

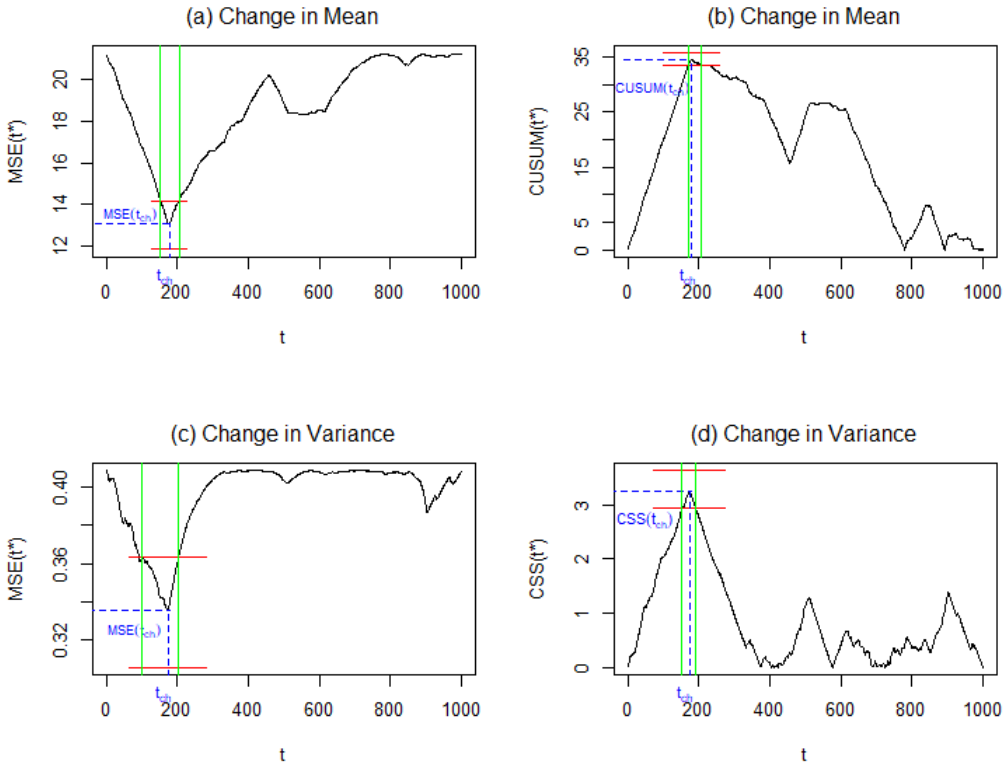


Figure 24: Determining change point location and its uncertainty by computing the 95% confidence interval. **(a)** The resulting $MSE(t^*)$. **(b)** The resulting $CUSUM(t^*)$. **(c)** The resulting $MSE(t^*)$. **(d)** The resulting $CSS(t^*)$

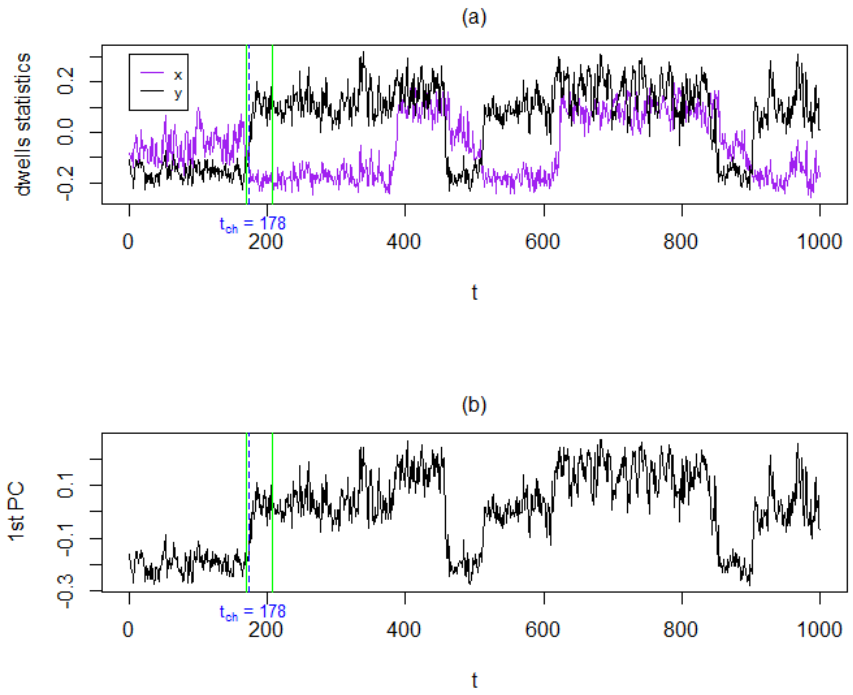


Figure 25: (a) The trace of the original 2D time series. (b) The trace of the 1st PC. The blue dashed line indicates the most prominent change point location. The solid green lines represent the 95% confidence interval for the change point location.

6 Conclusion

In this thesis, we have developed a nonparametric offline scheme for detecting changes in mean and/or variance in multivariate time series (with focus on two-dimensional time series). The scheme is based on principal component analysis, which is used for projecting the multivariate data onto a lower dimensional space. By using PCA to reduce the data dimension, we can successfully apply to the well formulated univariate methods, namely, the CUSUM and CSS methods, to test the existence of a change in mean and variance, respectively. The simulation study shows that the CUSUM and CSS methods are reliable to test the existence of a change point by using a permutation test.

To estimating the most likely change point location, two methods were proposed. One is the cumulative-sum based estimator which is consistent with the change point existence testing methods. The other is the mean square error (MSE) estimator which is on the contrary inconsistent with the existence test-

ing methods and therefore it will cause additional computational complexity in implementation. The study shows that the performance of the estimators depends very much on the properties of data. In terms of determining the true location of the change point, both methods perform similarly well if it is a mean change. However, if it is a variance change, both methods are shown to fail to identify a large number of change points. In terms of the resulted uncertainty in the determined change point location, the $MSE(t^*)$ estimator outperforms the $CUSUM(t^*)$ estimator if it is a mean change, and the $CSS(t^*)$ estimator outperforms the $MSE(t^*)$ estimator if it is a variance change. That is, in short, the MSE estimator is preferable if determining a change point location in a time series with a mean change while the CUSUM-based estimator is preferable if determining a change point location in a time series with a variance change.

We combine the CUSUM and CSS methods to detect simultaneous changes in mean and variance. The results on both simulated and real life data show that the combined method complements each other well and it can successfully determine the most prominent change point location by comparing the uncertainty in identifying the change point locations, though it costs additional computational complexity for running through a time series twice.

7 Discussion

First, it is worth to note that this thesis focuses on detecting one single change in mean or variance on two-dimensional offline data. As introduced at the beginning of the thesis, the proposed change point analysis scheme is able to generalize to higher dimensional data by keeping only the first principal component. The problem of how many principal components should be retained has been discussed a lot when it comes to reducing the data dimensionality. In this thesis, if we always choose the first PC associated with the largest eigenvalues, the first PC will always capture the most variation of the original data. Researchers are encouraged to perform a simulation study and evaluate its performance.

Furthermore, for simplification, we detected only one single change point in this thesis. This is a restriction that does not usually present in real life data and multiple change points do often exist over a longer time in many situations. We have provided a simple scheme such that one can easily apply the well-known binary segmentation procedure to detect multiple changes recursively. Interesting readers can find the procedure which has been described explicitly in both Chen and Gupta [2] and Eckley I.A.*et al* [3].

The proposed change point analysis scheme is further restricted by PCA's limitation of handling only linear data. Some modern nonlinear versions of linear principal component methods, such as *kernel principal component* has been well

studied [11]. Detecting changes in nonlinear data based on a nonlinear PCA method could be an interesting topic in future study. Here, we should notice that the change point detection methods described in this thesis are not restricted to PCA. PCA is served as a dimensionality reduction tool and the change point detection methods can be applied to one-dimensional nonlinear data independently.

Despite restrictions in this thesis, we consider that the change point analysis scheme we developed is simple to use and interpret. It successfully reduces the computational cost in detecting changes in multidimensional time series by only regarding a one-dimensional principal component. Both CUSUM and CSS methods are powerful at detecting smaller changes. The provided bootstrap based error bar and associated time interval better characterize the changes. If one is concerned with both change in mean and change in variance, both methods can be combined to complement each other. However, to gain more insight into the combined method, one may want to evaluate its performance in comparison with other existing change point detection methods.

A further investigation on the $MSE(t^*)$ estimator for locating a change point in a time series with a possible variance change could be made. When we applied the $MSE(t^*)$ estimator to locate a variance change point in the single molecule data, the resulted error bar in the determined change point location was reasonable (see Figure 24c). However, when the $MSE(t^*)$ estimator was applied to the simulated data, the results were quite unreasonable (see Figure 19). So, it could be possibly because of the setting of simulation. It would be very interesting to see if the results could be improved if another simulation setting is used.

We consider that the main disadvantage of the provided scheme comes to the expense of computational complexity due to the permutation test. As the number of data points increases, the computational cost becomes more expensive. Less computer intensive nonparametric statistical tests such as the Wilcoxon Rank Sum test, Mood test could be worth to try. But one should keep in mind that two sample hypothesis testing procedure has its drawback as only being suitable to apply to independent data. Also, the null distribution of the test statistic needs to be derived.

References

[1] Beaulieu, C., Chen, J., and Sarmiento, J.L. (2012). *Change-point analysis as a tool to detect abrupt climate variations*. Philosophical Transactions of The Royal Society, Volume 370, avaiable at `https://royalsocietypublishing.org/doi/full/10.1098/rsta.2011.0383`

[2] Chen, J. and Gupta, A.K. (2012). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. New York:Springer, 2nd Edition, Birkhäuser Boston.

[3] Eckley, I.A., Fearnhead, P., and Killick, R. (2011). *Analysis of change point models*. In: Bayesian time-series models. Cambridge University Press, Cambridge, pp. 203-224.

[4] Brodsky, B. and Darkhovsky, B. (1993). *Nonparametric Methods in Change–Point Problems*. Kluwer Academic Publishers, 1st Edition, The Netherlands.

[5] Kuncheva, L.I and Faithfull, W.J (2014). *PCA feature extraction for change detection in multidimensional unlabelled streaming data*. Published in: IEEE Transactions on Neural Networks and Learning Systems, 25:69-80.

[6] Qahtan, A.A., Alharbi, B., Wang, S. and Zhang, X. (2015) *A PCA-Based change detection framework for multidimensional data streams*, In: SIGKDD, pp. 935-944.

[7] Taylor, W.A. (2000). *Change-Point Analysis: A Powerful New Tool For Detecting Changes*. Baxter Healthcare Corporation, Round Lake, IL 60073, available at `http://www.variation.com/cpa/tech/changepoint.html`

[8] Inclan, C. and Tiao, G. C., (1994). *Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance*. Journal of the American Statistical Association, Vol.89, No.427, pp. 913-923.

[9] Ross, G.J.(2015). *Parametric and Nonparametric Sequential Change Detection in R: The cpm package*. Journal of Statistical Software.

[10] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

[11] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer, 2nd Edition.

[12] Jolliffe, I.T.(2002). *Principal Component Analysis, Springer Series in Statistics*. Springer-Verlag New York, Inc, 2nd Edition.

[13] Vidal, R., Ma Y., and Sastry S.S. (2016). *Generalized Principal Component Analysis*. Springer-Verlag New York.

[14] Cover, T.M, and Thomas, J.A. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc.

[15] Li, C.-B. *et al.* (2015) *ATP hydrolysis assists phosphate release and promotes reaction ordering in F1-ATPase*. Nat. Commun. 6:10223, doi: 10.1038/ncomms10223.

[16] Held, L. and Sabanés B.D. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Springer-Verlag Berlin Heidelberg.

[17] Efron, B. (1981), *Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods*. Biometrika 68, 589 - 599.

[18] R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[19] Andersson, J., (2014). *Locating Multiple Change-Points Using a Combination of Methods*, TRITA-MAT-E 2014:30, ISRN-KTH/MAT/E–14/30–SE, avaiable at `www.kth.se/sci`.

Appendix

A.1 Derivation of eigenvalues and eigenvectors

In this section, we explain the procedure of finding eigenvalues and eigenvectors using a trivial two-dimensional covariance matrix example. Let the covariance matrix $\boldsymbol{S}$ be in the following form

$$\boldsymbol{S} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \tag{17}$$

where σ_1, σ_2 are the standard deviations and ρ is the correlation coefficient between two variables.

To obtain the eigenvalues ($\boldsymbol{\lambda}$) and the associated eigenvectors ($\boldsymbol{v}$), the following equations need to be soloved:

$$\boldsymbol{S}\boldsymbol{v} - \boldsymbol{\lambda}\boldsymbol{v} = \boldsymbol{0} \Leftrightarrow (\boldsymbol{S} - \boldsymbol{\lambda}\boldsymbol{I})\boldsymbol{v} = \boldsymbol{0} \tag{18}$$

where $\boldsymbol{v}$ and $\boldsymbol{\lambda}$ are both unknown. Therefore $\boldsymbol{S} - \boldsymbol{\lambda}\boldsymbol{I}$ must be a singular matrix for nontrivial eigenvectors. So, it is possible to find all $\boldsymbol{\lambda}$s' first because it is known from linear algebra that determinant of a singular matrix is 0, which is $\det(\boldsymbol{S} - \boldsymbol{\lambda}\boldsymbol{I}) = \boldsymbol{0}$, where

$$S - \lambda I = \begin{pmatrix} \sigma_1^2 - \lambda & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 - \lambda \end{pmatrix} \tag{19}$$

Depending on the value of the correlation coefficient between two variables, the calculations of eigenvalues and eigenvectors are different. Therefore we divide the calculation into two cases.

Case 1: $\boldsymbol{\rho \neq 0}$
When $\rho \neq 0$, we have the following,

$$\begin{aligned} \det(S - \lambda I) &= (\sigma_1^2 - \lambda)(\sigma_2^2 - \lambda) - \sigma_1^2\sigma_2^2\rho^2 \\ &= \sigma_1^2\sigma_2^2 - \lambda\sigma_1^2 - \lambda\sigma_2^2 + \lambda^2 - \sigma_1^2\sigma_2^2\rho^2 \\ &= \lambda^2 - \lambda(\sigma_1^2 + \sigma_2^2) + \sigma_1^2\sigma_2^2 - \sigma_1^2\sigma_2^2\rho^2 \\ &= \lambda^2 - \lambda(\sigma_1^2 + \sigma_2^2) + \sigma_1^2\sigma_2^2(1 - \rho^2) \end{aligned} \tag{20}$$

The eigenvalues can then be found by solving the following equation

$$\lambda^2 - \lambda(\sigma_1^2 + \sigma_2^2) + \sigma_1^2\sigma_2^2(1 - \rho^2) = 0 \tag{21}$$

By some calculations, we obtain

$$\begin{aligned} \lambda_{1,2} &= \frac{1}{2}\left((\sigma_1^2 + \sigma_2^2) \pm \sqrt{(\sigma_1^2 + \sigma_2^2)^2 - 4 \cdot \sigma_1^2\sigma_2^2(1 - \rho^2)}\right) \\ &= \frac{1}{2}\left((\sigma_1^2 + \sigma_2^2) \pm \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + (2\sigma_1\sigma_2\rho)^2}\right) \end{aligned} \tag{22}$$

Since the component $\sqrt{(\sigma_1^2 - \sigma_2^2)^2 + (2\sigma_1\sigma_2\rho)^2} > 0$, we let

$$\lambda_1 = \frac{1}{2}\left((\sigma_1^2 + \sigma_2^2) + \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + (2\sigma_1\sigma_2\rho)^2}\right) \tag{23}$$

which is always larger than

$$\lambda_2 = \frac{1}{2}\left((\sigma_1^2 + \sigma_2^2) - \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + (2\sigma_1\sigma_2\rho)^2}\right). \tag{24}$$

Thus, we say that λ_1 is the largest eigenvalue of $\boldsymbol{S}$.

Now, we look for the eigenvector, v_1, corresponding to the largest eigenvalue of $\boldsymbol{S}$, λ_1. The eigenvector v_1 is obtained by soloving the following equation

$$(\boldsymbol{S} - \lambda_1) \cdot \boldsymbol{v_1} = 0 \tag{25}$$

$$\begin{bmatrix} \sigma_1^2 - \lambda_1 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 - \lambda_1 \end{bmatrix} \cdot \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

From the top row of the equations we get

$$(\sigma_1^2 - \lambda_1)v_{11} + \sigma_1\sigma_2\rho v_{21} = 0$$

$$\Leftrightarrow v_{11} = -\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_1} v_{21} \tag{26}$$

If we let $v_{21} = t$, then $v_{11} = \frac{-\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_1} t$. So all eigenvectors corresponding to λ_1 are multipliers of

$$\boldsymbol{v_1} = \begin{bmatrix} -\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_1} \\ 1 \end{bmatrix} \text{ which normalises to } \boldsymbol{v_1} = \begin{bmatrix} \frac{-\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_1}}{\sqrt{\left(-\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_1}\right)^2 + 1}} \\ \frac{1}{\sqrt{\left(-\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_1}\right)^2 + 1}} \end{bmatrix} \tag{27}$$

The other eigenvector $\boldsymbol{v_2}$ corresponding to the second eigenvalue λ_2 is calculated analogously.

$$\boldsymbol{v_2} = \begin{bmatrix} -\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_2} \\ 1 \end{bmatrix} \text{ which normalises to } \boldsymbol{v_2} = \begin{bmatrix} \frac{-\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_2}}{\sqrt{\left(-\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_2}\right)^2 + 1}} \\ \frac{1}{\sqrt{\left(-\frac{\sigma_1\sigma_2\rho}{\sigma_1^2 - \lambda_2}\right)^2 + 1}} \end{bmatrix} \tag{28}$$

Case 2: $\rho \to 0$

When ρ approaches 0, the covariance term $\sigma_1\sigma_2\rho$ in $\boldsymbol{S}$ approaches also 0. In this case, the covariance matrix $\boldsymbol{S}$ becomes a diagonal matrix, which is $\boldsymbol{S} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$ as $\rho \to 0$. This means that the variances must be equal to the eigenvalues $\boldsymbol{\lambda}$. This is illustrated by Figure 26 which shows an example of a two-dimensional dataset whose covariance matrix is given by $\boldsymbol{S} = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$, where the eigenvectors are shown in blue and green.

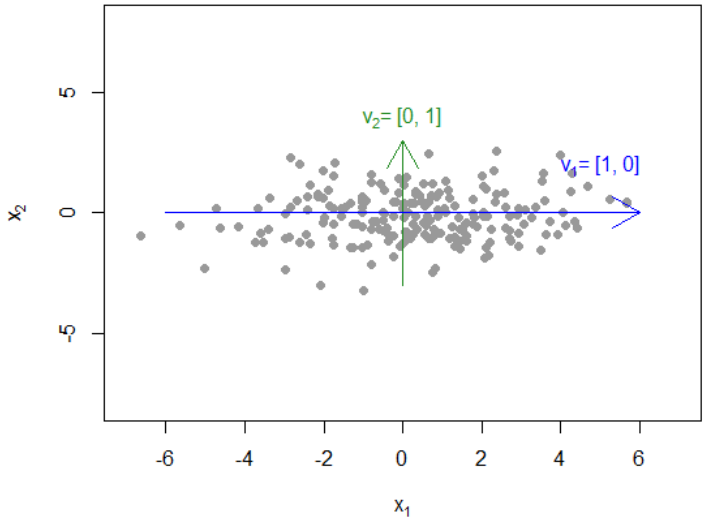


Figure 26: Eigenvectors of a diagonal covariance matrix.

Mathematically, suppose that $\sigma_1 \geq \sigma_2$, then λ_1 is the largest eigenvalue. The corresponding eigenvector is obtained by

$$\begin{bmatrix} \sigma_1^2 - \sigma_1^2 & 0 \\ 0 & \sigma_2^2 - \sigma_1^2 \end{bmatrix} \cdot \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

From the second row of equations, we get

$$v_{21}(\sigma_2^2 - \sigma_1^2) = 0.$$

This means that v_{21} is necessary to be 0 while v_{11} is a trivial value before normalization. Thus,

$$\boldsymbol{v_1} = \begin{bmatrix} v_{11} \\ 0 \end{bmatrix} \text{ which normalises to } \boldsymbol{v_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The other eigenvector $\boldsymbol{v_2}$ corresponding to λ_2 is obtained analogously,

$$\boldsymbol{v_2} = \begin{bmatrix} 0 \\ v_{21} \end{bmatrix} \text{ which normalises to } \boldsymbol{v_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Observe that the variables are uncorrelated in this case, which can be interpreted as that no linear relationship exists between variables.

A.2 Proof of Lemma 1

***Lemma 1**: For two univariate normal distributions, $p(x)$ and $q(x)$, where $q(x) \sim N(\mu_1, \sigma_1^2)$ and $p(x) \sim N(\mu_2, \sigma_2^2)$, the KL-distance from q(x) to p(x) is given as follows*:

$$D_{KL}(p \parallel q) = \frac{1}{2}\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) + \frac{\sigma_2^2 + (\mu_1 - \mu_2)^2}{2\sigma_1^2} - \frac{1}{2} \tag{29}$$

Proof:

$$\begin{aligned}
D_{KL}(p \parallel q) &= \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right) dx \\
&= \int_{-\infty}^{\infty} p(x)\log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}\right) dx \\
&= \int_{-\infty}^{\infty} p(x)\log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_2^2}}}{\frac{1}{\sqrt{2\pi\sigma_1^2}}}\right) + \int_{-\infty}^{\infty} p(x)\log\left(\frac{\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)}{\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}\right) \\
&= \int_{-\infty}^{\infty} p(x)\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{1/2} dx + \int_{-\infty}^{\infty} p(x)\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2} + \frac{(x-\mu_1)^2}{2\sigma_1^2}\right] dx \\
&= \frac{1}{2}\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) + \frac{1}{2\sigma_2^2}\left[-\int_{\infty}^{-\infty} p(x)(x-\mu_2)^2 dx\right] + \frac{1}{2\sigma_1^2}\left[\int_{\infty}^{-\infty} p(x)(x-\mu_1)^2 dx\right] \\
&= \frac{1}{2}\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) - \frac{\sigma_2^2}{2\sigma_2^2} + \frac{1}{2\sigma_1^2}\underbrace{\left[\int_{\infty}^{-\infty} (x-\mu_2+\mu_2-\mu_1)^2 p(x) dx\right]}_{(1)} \\
&= \frac{1}{2}\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) - \frac{1}{2} + \frac{\sigma_2^2 + (\mu_2-\mu_1)^2}{2\sigma_1^2}
\end{aligned}$$

where

$$\begin{aligned}
(1) &= \int_{-\infty}^{\infty} (x-\mu_2)^2 p(x) dx + 2(x-\mu_1)\int_{-\infty}^{\infty} \underbrace{(x-\mu_2)}_{=0} p(x) dx + (\mu_2-\mu_1)^2 \int_{-\infty}^{\infty} p(x) dx \\
&= \sigma_2^2 + (\mu_2-\mu_1)^2
\end{aligned}$$

Note that, by definition,

$$\int_{-\infty}^{\infty} p(x)dx = 1, \quad \text{and}$$

$$E_2[(x-\mu_2)^2] = \int_{-\infty}^{\infty} (x-\mu_2)^2 p(x)dx = \sigma_2^2$$

.