

Linköping Studies in Science and Technology  
Licentiate Thesis No. 1970

# Change point detection with respect to variance

**Elias Erdtman**



Linköping Studies in Science and Technology.  
Licentiate Thesis No. 1970

# **Change point detection with respect to variance**

**Elias Erdtman**



Department of Mathematics  
Linköping University, SE-581 83 Linköping, Sweden  
Linköping 2023



This work is licensed under a Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

Linköping Studies in Science and Technology. Licentiate Thesis No. 1970

## **Change point detection with respect to variance**

Copyright © Elias Erdtman, 2023

*Department of Mathematics  
Linköping University  
SE-581 83 Linköping, Sweden*

Email: [elias.erdtman@liu.se](mailto:elias.erdtman@liu.se)

ISSN 0280-7971

ISBN 978-91-8075-239-8 (print version)

ISBN 978-91-8075-240-4 (electronic version)

Printed by LiU-Tryck, Linköping, Sweden 2023

This is a Swedish Licentiate Thesis.

The Licentiate degree comprises 120 ECTS credits of postgraduate studies.

## Abstract

This thesis examines a simple method for detecting a change with respect to the variance in a sequence of independent normally distributed observations with a constant mean. The method filters out observations with extreme values and divides the sequence into equally large subsequences. For each subsequence, the count of extreme values is translated to a binomial random variable which is tested towards the expected number of extremes. The expected number of extremes comes from prior knowledge of the sequence and a specified probability of how common an extreme value should be. Then specifying the significance level of the goodness-of-fit test yields the number of extreme observations needed to detect a change.

The approach is extended to a sequence of independent multivariate normally distributed observations by transforming the sequence to a univariate sequence with the help of the Mahalanobis distance. Thereafter it is possible to apply the same approach as when working with a univariate sequence. Given that a change has occurred, the distribution of the Mahalanobis distance of a multivariate normally distributed random vector with zero mean is shown to approximately follow a gamma distribution. The parameters for the approximated gamma distribution depend only on  $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$  with  $\Sigma_1$  and  $\Sigma_2$  being the covariance matrices before and after the change has occurred. In addition to the proposed approach, other statistics such as the largest eigenvalue, the Kullback-Leibler divergence, and the Bhattacharyya distance are considered.



## Populärvetenskaplig sammanfattning

Det här arbetet diskuterar en enkel metod för detektion av en ökning i variansen (eller en förändring av kovariansmatrisen) i en serie av oberoende normalfördelade observationer. Metoden arbetar genom att dela upp sekvensen i lika stora delar, så kallade fönster. För varje fönster tolkas antalet observationer med extrema värden som en binomialfördelad stokastisk variable som testas mot det förväntade antalet observationer. Det förväntade antalet observationer kommer från tidigare kunskap om sekvensen och en satt signifikansnivå för hur många falsklarmsom får ske.

Metoden generaliseras sedan till sekvenser av oberoende multivariata normalfördelad stokastiska vektorer genom att konstruera Mahalanobisavståndet av observationerna. Avståndsvärdet av alla observationer ger en sekvens av univariat värden där det återigen går att räkna antalet extremer i ett fönster.

Vidare undersöks några andra sätt att testa om kovariansmatrisen har förändrats, till exempel genom att studera förändringen hos egenvärdena eller några informationsmått såsom Kullback-Leibler och Bhattacharyya måtten. Dessa metoder ger lovande resultat och möjligheter till vidare forskning.



## Acknowledgments

First of all, I would like to thank my supervisors Martin Singull and Dietrich von Rosen for their support and encouragement throughout the last few years. I am grateful to Martin for guiding me through the process of conducting research and giving me the opportunity to take part in teaching a new course in statistics. I highly value all the work both Martin and Dietrich have put into pointing out the areas of improvement in the theory and I greatly appreciate your tireless efforts in helping me with my writing.

Furthermore, I have my family and friends to thank for their continuous support and for believing in my capability to step up to this new challenge and go back to studying math. I seriously doubted whether I would be able to do math at the high level that is required after years away from the subject. It was possible!

I would also like to extend my gratitude to my fellow Ph.D. students for the discussions about courses and our studies in general, but mostly for sharing part of our lives during lunches. Finally, a special thank you to Björn Morén who has been my classmate and colleague during all my years at the university. You are an inspiration in pursuing difficult goals and showing that it is possible to achieve them with passion and hard work.

Thank you all!

*Linköping, August 28, 2023*

*Elias Erdtman*



---

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Aims . . . . .	4
1.3	Contributions . . . . .	4
1.4	Outline . . . . .	5
<b>2</b>	<b>Theory and Methods</b>	<b>7</b>
2.1	Model . . . . .	7
2.2	Method . . . . .	10
2.2.1	Extreme observation . . . . .	10
2.2.2	Windows . . . . .	13
2.2.3	Alert algorithm . . . . .	17
<b>3</b>	<b>Results</b>	<b>19</b>
3.1	Univariate case . . . . .	19
3.2	Multivariate case . . . . .	21
3.2.1	Mahalanobis . . . . .	22
3.2.2	Eigenvalues of sample covariance matrix . . . . .	29
3.2.3	Divergence information measure . . . . .	31
3.2.4	Simulation and comparison . . . . .	34
<b>4</b>	<b>Concluding remarks</b>	<b>43</b>
4.1	Summary . . . . .	43
4.2	Conclusion . . . . .	44
4.3	Future research . . . . .	45
	<b>Bibliography</b>	<b>47</b>



# 1

---

## Introduction

This thesis is concerned with detecting a change with respect to the variance in a sequence of observations where the mean is constant. Both univariate and multivariate responses are treated.

### 1.1 Background

Change point detection is about finding a change in a sequence of observations. The observations come from underlying random variables  $\{X_1, \dots, X_N\}$ , where  $X_i$  follows a distribution  $f_i$ . A change point is the point  $n$  where the distribution function changes. In order to decide if a change has occurred, hypothesis testing can be performed. The general hypothesis can be formulated as

$$H_0 : f_1 = f_2 = \dots = f_N,$$

versus the alternative hypothesis

$$H_1 : f_1 = \dots = f_{n_1-1} \neq f_{n_1} = \dots = f_{n_2-1} \neq f_{n_2} = \dots = f_{n_q-1} \neq f_{n_q} = \dots = f_N,$$

where  $n_1, n_2, \dots, n_q$  mark the unknown positions of  $q$  number of changes.

The first to consider the problem of detecting a change in a sequence of observations at an unknown time was Page (1955), who focused on finding a change in the mean by counting observations above and below the expected mean. If the observed value is above the mean, then the count increases by one and if it is below, the count decreases by one. If the value of the count compared to its lowest value goes above a threshold, depending on the length of the series, then a change point is identified.

Among the first to identify change points based on changes in variance was Wichern et al. (1976). They focused on detecting shifts in variance estimates

within a first-order autoregressive model, where the errors follow a normal distribution with a zero mean and a variance  $\sigma_t^2$ . Their algorithm first estimates possible change points by taking the ratio of the sum of the squared observations of two adjacent subsequences, called windows, with a fixed number of observations in the windows. The sequence is then scanned for candidates by systematically sliding these windows through the sequence using a predetermined step size, and subsequently comparing them. The change point candidates are selected as the observations positioned between the windows where the ratio exceeds a threshold based on an F-distribution. Following this selection, the algorithm uses the candidates of change points to estimate the variance of the observations between the points. Subsequently, an iterative process is applied to calculate the maximum likelihood, utilizing the estimated variance as an input. This iterative process aims to enhance the accuracy of change point localization.

Inclán and Tiao (1994) presented a test statistic constructed from a cumulative sum of the squared observation known as CUSUM:  $C_n = \sum_{i=1}^n y_i^2$ , where  $n = 1, \dots, N$ . Here,  $y_i$  represents the  $i^{\text{th}}$  observed value in the sequence of observations, and  $N$  stands for the total number of observations. The observations are assumed to be from a sequence of uncorrelated random variables with a zero mean and variance  $\sigma_n^2$ . With the help of  $C_n$ , they constructed a random walk via  $D_n = \frac{C_n}{C_N} - \frac{n}{N}$ . When testing for equality in variance, it is demonstrated that under the null hypothesis, the test statistic  $\sqrt{\frac{N}{2}}D_n$  follows a known asymptotic distribution. This provides the possibility to establish a threshold by testing a hypothesis with a fixed significance level. If the value of  $D_n$  exceeds the threshold, then a change point has been identified.

The algorithm using CUSUM, described above, was extended by Gombay et al. (1996). This extension involved working with the asymptotic distribution of the CUSUM statistic, particularly under the assumption of a nonzero mean  $\mu$ , which could be unknown. To adjust for an unknown nonzero mean Gombay et al. (1996) use terms  $(y_i - \hat{\mu})^2$  in  $C_n$ . Furthermore, their method demonstrated the possibility of handling independent random variables with a finite fourth-order moment.

Aminikhanghahi and Cook (2017) presented various techniques aimed at identifying change points. Both changes in mean and variance were considered. The methods presented are categorized based on the approach employed to detect these change points. The primary division consists of supervised and unsupervised methods. Supervised methods include approaches using decision trees, support vector machines, and logistic regression. While unsupervised methods are based on likelihood ratio, probabilistic methods, and clustering, to mention a few.

Change point detection methods can also be classified based on the number of observations they are designed to utilize for detecting the change after it occurs. If the complete sequence is required, the method is referred to as an *offline method*. Conversely, if observations are processed as they become accessible, the method is called *online method*.

When dealing with multiple changes, there are various methods for exploring a sequence. Truong et al. (2020) outline three primary search techniques: Binary Segmentation, Bottom-Up, and the Sliding Window method. Among these, Bi-

nary Segmentation is the most commonly used. It begins to search for the most significant change point in the whole sequence using for example an algorithm based on CUSUM. Once a change is identified at position  $n_1$ , the same procedure is then applied to the subsequences  $x_1, \dots, x_{n_1-1}$  and  $x_{n_1}, \dots, x_N$ . This iterative process continues until no further changes are detected. Binary Segmentation is used for offline methods.

The Bottom-Up approach involves segmenting the entire sequence into successive, equal-sized subsequences known as windows. The adjacent windows are then compared, and those with no significant difference are merged. The method of comparison depends on the algorithm. The procedure is repeated until either all the windows are merged into one or there is a significant difference between the remaining adjacent windows. The change points are identified as the shifts between the remaining windows. While not as widely adopted or extensively studied as Binary Segmentation, the Bottom-Up approach has also shown promise, as demonstrated in Keogh et al. (2001). In general, it is used for offline methods but it can also gather new observations to a new window and compare it with the adjacent window.

The third method described here is the Sliding Window method. The concept revolves around two consecutive windows sliding over the sequence. At each position, they are compared to each other with the objective of determining whether the observations within the windows stem from the same distribution. An important factor to consider for the Sliding Window approach is the size of the windows,  $k$ . If  $k$  is small, substantial differences between windows are necessary in order to notice a change. On the other hand, if  $k$  is excessively large, it loses its ability to locate the change point and requires more observations after the change. The Sliding Window method can be applied to both online and offline methods.

Handling multivariate sequences introduces greater complexity in how changes can occur and more information needs to be checked. Similar to the univariate case, research has primarily concentrated on changes in mean. Examples include studies by Zamba and Hawkins (2006), Wang and Reynolds (2013), and Dette and Gösmann (2018). The methods used are often based on CUSUM or a likelihood ratio.

Chen and Gupta (2012) presented maximum likelihood estimators for change points in normally distributed sequences. Their scope spans from the simplest case involving a univariate sequence with an unknown mean and known variance to a multivariate sequence where both the mean and the covariance matrix are unknown and cover everything in between. All of the maximum likelihood estimators reach the maximum value by varying the change points positions and therefore come with a risk of calculating the estimator for each observation in the sequence.

Additionally, there are studies of high-dimensional cases in which the dimension surpasses the number of observations. In these more intricate scenarios, changes in the mean remain the primary research focus. For instance, studies such as Wang and Samworth (2018) and Enikeeva and Harchaoui (2019) explore sparse changes in the mean vector, where changes only occur in a small subset of dimensions. Research concerning the covariance matrix in high dimensions is an ongoing endeavor, with works by Avanesov and Buzun (2018) and Jirak (2015) offering insights into

this domain.

This work explores a simple algorithm for detecting a relative increase in variance in a sequence of normally distributed observations with a constant mean. The algorithm looks for an increase in variance by counting extreme values in the sequence of observations. If the number of extreme observations is higher than expected, then a change in variance is identified. This approach shares similarities to the technique employed by Page (1955) to find a change in the mean. While Page (1955) counts the number of observations around the mean, in our method, the values outside a threshold are counted instead.

In order to control the algorithm, three probabilities are specified. The first probability determines how "extreme" an observation needs to be for it to be included in the count. The second probability defines the type I error and defines the minimum count of observations within a subsequence necessary to assert the presence of a change. The third and last probability is the power of the test. This helps to establish the smallest size the subsequence can have in order for the test to be significant. To extend the algorithm so that it can handle a multivariate sequence we use the Mahalanobis distance to transform each observation of the multivariate sequence into a sequence of univariate observations.

## 1.2 Aims

The aims of this thesis are:

- to construct a simple algorithm to detect a change in variance in a sequence of observations;
- to define and develop tools that examine how well the algorithm can perform depending on the size of the change;
- to extend the algorithm so that it can handle change point detection in a sequence of a multivariate response, i.e., to detect a change with respect to the covariance matrix.

## 1.3 Contributions

The main contributions of the thesis are as follows:

- A simple change point detection algorithm based on extreme values of a sequence of observations has been derived. The algorithm uses the number of extreme observations to alert whether there has been a change.
- An expression for the distribution of the quadratic form  $\mathbf{Y}'\mathbf{A}^{-1}\mathbf{Y}$ , where  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$  and  $\mathbf{A}$  is a symmetric positive definite matrix. Then  $\mathbf{Y}'\mathbf{A}^{-1}\mathbf{Y}$  is approximately gamma distributed with the shape and rate parameters

$$a = \frac{(\text{tr}(\mathbf{\Psi}))^2}{2\text{tr}(\mathbf{\Psi}^2)}, \quad b = \frac{\text{tr}(\mathbf{\Psi})}{2\text{tr}(\mathbf{\Psi}^2)},$$

where  $\mathbf{\Psi} = \mathbf{A}^{-1/2}\mathbf{\Sigma}\mathbf{A}^{-1/2}$ .

## 1.4 Outline

Chapter 2 introduces the theory for the proposed change point detection algorithm. The theory is summarised into an alert algorithm divided into two stages: an initiation algorithm and a search algorithm.

Chapter 3 presents the results necessary to finalize the alert algorithm for both univariate and multivariate sequences of observations by deriving an expression of the probability for an extreme observation to occur after a change. In the multivariate case, the probability is derived with the help of the Mahalanobis distance. Furthermore, we explore a few different multivariate methods to find a change in variance. All the methods use the same technique for conducting a search by looking at non-overlapping subsequences. The chapter ends with an evaluation of our proposed multivariate method from simulated data.

Finally, Chapter 4 sums up the thesis and includes concluding remarks and ideas for future work.



# 2

---

## Theory and Methods

As noted in the introduction, change point detection is a well-studied field with ongoing research. In this chapter, the theory necessary to understand our proposed algorithm is introduced.

Let us return to the idea of testing hypotheses when identifying a change point. If the random variables come from the same family of distributions and the change occurs in a parameter of the distributions, then the hypothesis could be formulated as a change in that parameter. Since we are looking at one change in variance, denoted  $\sigma_i^2$  for observation  $i$ , the hypothesis is

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_N^2,$$

with the alternative hypothesis being

$$H_1 : \sigma_1^2 = \dots = \sigma_{n-1}^2 \neq \sigma_n^2 = \dots = \sigma_N^2,$$

i.e., we are testing whether the variance changes at observation  $n$ .

### 2.1 Model

Let us start with the following definition.

**Definition 2.1 (SIO).** Let  $\mathbf{X}_{(1,N)} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$  be a sequence of length  $N$  of independent random vectors. A realization of these random vectors is  $\mathbf{x}_{(1,N)} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , i.e., a sequence of independent observations, SIO.

The focus of this thesis is a single change in variance or in the covariance matrix in a sequence of observations consisting of independent normally distributed random vectors with a constant mean. Let  $\mathcal{N}(\mu, \sigma^2)$  denote the normal distribution with mean  $\mu$  and variance  $\sigma^2$ .  $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  denotes the  $p$ -dimensional multivariate normal distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ .

**Definition 2.2 (Change in variance).** Let  $x_{(1,N)}$  be a SIO where observations from  $x_{(1,n-1)}$  follow  $\mathcal{N}(\mu, \sigma_1)$  and those from  $x_{(n,N)}$  follow  $\mathcal{N}(\mu, \sigma_2)$ . Then the measure of change in the variance in  $x_{(1,N)}$  is defined as  $\psi = \sigma_2^2/\sigma_1^2$ . Hence, a change occurs if  $\psi \neq 1$ .

The reason for defining change  $\psi$  as the ratio of the variance between the values before and after the change is that every change is relative to some reference point. Dette and Wied (2016) write about changes in mean and argue that the change needs to be "big enough" for it to be of consequence. For instance, a variance change of 9 would be considered large if the initial variance was 1, but it would be deemed a small change if the initial variance was 91 units. In relative terms, we would observe a change of  $\psi = 10$ , and a change of  $\psi = 1.1$ , respectively. In Example 2.1, we see two SIOs. In the graph on the left, there is no change, while in the one on the right, there is a change at observation 151.

A change in the multivariate case is, in general, more complex. It can occur in a similar manner as in the univariate case; that is, the signal in one variable changes. However, the relation between the different variables may also change. In Example 2.2, the variance in the different variables as well as the relation between them changed.

**Definition 2.3 (Change in the covariance matrix).** Let  $\mathbf{x}_{(1,N)}$  be a SIO where observations from  $\mathbf{x}_{(1,n-1)}$  follow  $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_1)$ , and those of  $\mathbf{x}_{(n,N)}$  follow  $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_2)$ . Then the measure of change in the covariance variance matrix in  $\mathbf{x}_{(1,N)}$  is defined as  $\mathbf{\Psi} = \mathbf{\Sigma}_1^{-1/2} \mathbf{\Sigma}_2 \mathbf{\Sigma}_1^{-1/2}$ .

Note that  $\mathbf{\Psi}$  is a function of  $\mathbf{\Sigma}_1^{1/2}$ , which is defined via the spectral decomposition of  $\mathbf{\Sigma}_1$ ,  $\mathbf{\Sigma}_1 = \mathbf{U} \mathbf{D} \mathbf{U}'$  and  $\mathbf{\Sigma}_1^{1/2} = \mathbf{U} \mathbf{D}^{1/2} \mathbf{U}'$ , where  $\mathbf{U}$  consists of the eigenvectors, and  $\mathbf{D}$  is a diagonal matrix with the eigenvalues of  $\mathbf{\Sigma}_1$ . By defining  $\mathbf{\Psi}$  as  $\mathbf{\Sigma}_1^{-1/2} \mathbf{\Sigma}_2 \mathbf{\Sigma}_1^{-1/2}$ , we set the restrictions to  $\mathbf{\Psi}$  so that it is a symmetric positive definite matrix.

The following assumption will be used throughout the thesis.

**Assumption A1.**

Let  $x_{(1,N)} = \{x_1, \dots, x_{n-1}, x_n, \dots, x_N\}$  be a SIO from the random variables  $\{X_1, \dots, X_{n-1}, X_n, \dots, X_N\}$ , where  $X_i \sim \mathcal{N}(\mu, \sigma_1^2)$ ,  $i = 1, \dots, n-1$ , and  $X_j \sim \mathcal{N}(\mu, \sigma_2^2)$ ,  $j = n, \dots, N$ .

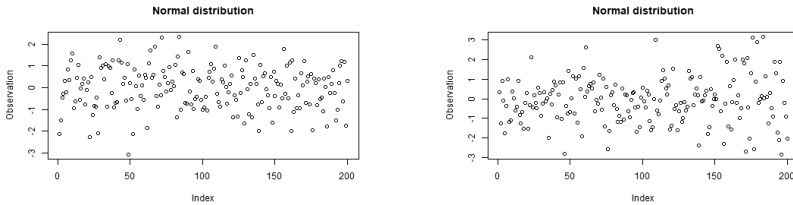
**Assumption A2.**

Let  $\mathbf{x}_{(1,N)} = \{\mathbf{x}_1, \dots, \mathbf{x}_{n-1}, \mathbf{x}_n, \dots, \mathbf{x}_N\}$  be a SIO from the random vectors  $\{\mathbf{X}_1, \dots, \mathbf{X}_{n-1}, \mathbf{X}_n, \dots, \mathbf{X}_N\}$ , where  $\mathbf{X}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_1)$  for  $i = 1, \dots, n-1$  and  $\mathbf{X}_j \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_2)$ ,  $j = n, \dots, N$ .

Our focus lies solely on sequences containing a single change since we are developing an algorithm for real-time monitoring, an online algorithm. In this context, when a new observation arises, it is promptly incorporated into the sequence and assessed as soon as possible. It is assumed that the change will be detected before any other change occurs. This approach assumes a sufficient, but not infinite, number of observations to detect the change, and will therefore not rely on asymptotic results.

**Example 2.1**

Figure 2.1 shows two different SIOs with 200 observations from random variables  $X_i \sim \mathcal{N}(0, \sigma_i^2)$ ,  $i = 1, \dots, 200$ . In the left panel, the variance is constant. In the right panel at observation 150, the variance equals  $\sigma_i^2 = 1$  for  $i = 1, \dots, 150$  and  $\sigma_j^2 = 2$  for  $j = 151, \dots, 200$ .



**Figure 2.1:** In the left panel, we have a sequence of independent normally distributed random variables with mean zero and variance one. To the right, there is a change in the random variables at observation 151, from a variance equal to one to a variance that equals two.

**Example 2.2**

The observations from a bivariate normal distribution form a point cloud. The challenge to identify a change in the covariance matrix is to find out if a set of observations forms a different pattern than expected. Figure 2.2 shows a sequence of 200 observations. The first 150 observations follow  $\mathcal{N}_2(\mathbf{0}, \mathcal{I})$ , whereas the last 50 observations follow  $\mathcal{N}_2(\mathbf{0}, \Sigma)$  with

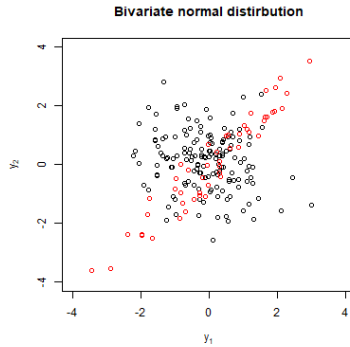
$$\Sigma = \begin{bmatrix} 1.55 & 1.88 \\ 1.88 & 2.53 \end{bmatrix}.$$

Before the change, the observations are presented in black, and after, in red. Here we can clearly see that the pattern changes, the black points are gathered in a circular shape, and the red dots follow a rotated elliptical shape.

To notice a change quickly, the major focus needs to be on the new observations. Both Aminikhanghahi and Cook (2017) and Truong et al. (2020) give an overview of several methods that look at consecutive subsequences of observations in order to try to determine when a change occurs, giving rise to the following definition.

**Definition 2.4 (Window).** Given a SIO of length  $N$ , a window of size  $k$  is a subset of  $k$  consecutive observations  $\mathbf{x}_{(t,t+k)} = \{\mathbf{x}_t, \dots, \mathbf{x}_{t+k}\}$ , where  $t \in [1, N - k]$ .

In the introduction, we gave a short description of three search techniques where Bottom-Up and Sliding Window use windows to scan the complete sequence.



**Figure 2.2:** Observations from a bivariate normal random variable, where the covariant matrix changes. The observations before the change are presented in black and after the change in red.

Our suggested algorithm uses a combination of ideas from the Bottom-Up and the Sliding Window strategies. In the Bottom-Up strategy, we use the concept of having the windows with no common observation and then merging the new window with the default window when no changes are found. From the Sliding Window strategy, we use the forward motion of checking one window at a time. The motion proceeds until there are no more observations in the sequence or until a change is found. To ensure there are no common observations in the windows, and are therefore independent, the step length is the window size.

## 2.2 Method

This section will present our change point algorithm. The objective we are aiming for is to find a change in variance as early as possible after the change has occurred. The change in variance happens in the underlying sequence of normally distributed random variables whose observations give rise to the SIO that we will investigate, in other words, an online change detection algorithm.

In the first part of this section, in Definition 2.6 we will define what we mean by extreme values and denote them as  $\alpha$ -observations. Then we will talk about how they can be used with a window to detect a change by simply counting the  $\alpha$ -observations.

### 2.2.1 Extreme observation

An extreme observation is an observation that occurs with a low probability. This will be called an  $\alpha$ -observation and is defined in the following ways:

**Definition 2.5 ( $\alpha$ -observation).** Let  $X$  be a random variable with zero mean and a known variance. For a given  $\alpha$ , define a constant  $c$  so that

$$\alpha = P(X > c).$$

Then an observation  $x$  is called an  $\alpha$ -observation if

$$x \in (c, \infty).$$

**Definition 2.6 ( $\alpha$ -observation).** Let  $X$  be a random variable with known mean and variance. For a given  $\alpha$ , define constants  $c_1$  and  $c_2$  so that

$$\alpha/2 = P(X < c_1) = P(X > c_2).$$

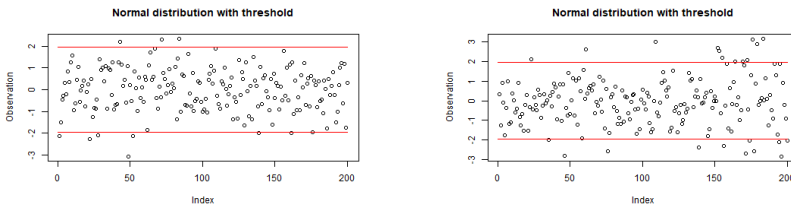
Then an observation  $x$  is called an  $\alpha$ -observation if

$$x \in (-\infty, c_1) \cup (c_2, \infty).$$

The reason that there are two definitions of the  $\alpha$ -observation is that, for some distributions, it does not provide more information with two thresholds. Note that both Definition 2.5 and Definition 2.6 need a priori information to create the thresholds  $c$ ,  $c_1$ , and  $c_2$ . An  $\alpha$ -observation is denoted with its percentage, e.g., when  $\alpha = 0.05$  then an  $\alpha$ -observation is called a 5%-observation. The 5%-observations are expected to be observed in 5% of all observations if there is no change. This means that if there is no change, out of  $N$  observations the expected number of  $\alpha$ -observations will be  $\alpha N$ . If there are significantly more  $\alpha$ -observations in a sequence, then it is likely that there has been an increase in the variance, given that the mean is constant.

### Example 2.3

Let us return to Example 2.1 where we are expecting observations from a random variable  $X$  with the distribution  $\mathcal{N}(0, 1)$ . The thresholds for  $\alpha$ -observations at 5% will be placed at  $\pm 1.96$ . In Figure 2.3 we can see a few 5%-observations in the left SIO, where there is no change. For the right SIO, it is possible to notice an increased frequency of 5%-observations after observation 151.



**Figure 2.3:** The red lines are the thresholds for 5%-observations. All observations above the upper line or below the lower are 5%-observations.

By taking a look at a SIO of size  $N$  we have that if no change has occurred then every single observation has probability  $\alpha$  to be an  $\alpha$ -observation. If the variance has increased then the probability for an  $\alpha$ -observation increases, see Theorem 3.1. Let  $W$  be the random variable representing the number of  $\alpha$ -observations in a SIO of size  $N$ . Then  $W$  will be binomially distributed  $Bin(N, p)$  for some probability  $p$ . If the sequence has no change then  $p = \alpha$ , but if there has been an increase in the variance then  $p > \alpha$ . Thus we can reformulate the hypothesis of a change in variance shown at the beginning of this chapter to a hypothesis as a change in probability for an  $\alpha$ -observation as

$$H_0 : p = \alpha \quad \text{versus} \quad H_1 : p > \alpha. \quad (2.1)$$

To test the hypothesis in (2.1) for a given sequence at a significance level of  $\alpha_W$  it is enough to count the number of  $\alpha$ -observations,  $y$ . Then if  $P(W \geq y | H_0) < \alpha_W$  we can reject  $H_0$  and with the assumption that the mean is constant then draw the conclusion that the variance has increased.

---

#### Example 2.4

---

Returning to the data in Example 2.1 and 2.3 to check if a change is noticed by counting the  $\alpha$ -observations in each of them. There we have two SIOs both with 200 observations and  $\alpha$  is set to 0.05 which gives the thresholds of  $\pm 1.96$ . Let the random variable  $W_i \sim Bin(N, p_i)$ ,  $i = 1, 2$ , represent the number of  $\alpha$ -observations in the two SIOs ( $i = 1$  is the left image and  $i = 2$  is the right image). The counted 5%-observation for the first SIO is  $w_1 = 9$  and for the second SIO, we have  $w_2 = 22$  5%-observations. To check if a change has occurred the test in (2.1) will be applied with  $\alpha = 0.05$  and a significance level of  $\alpha_W = 0.05$ , i.e.,

$$H_0 : p_i = \alpha \quad \text{versus} \quad H_1 : p_i > \alpha, \quad i = 1, 2,$$

where  $H_0$  for  $i = 1, 2$  can be rejected if  $P(W_i \geq w_i | H_0) < \alpha_W$ .

$$P(W_1 \geq 9 | H_0) = 1 - \sum_{i=0}^8 \binom{200}{i} \alpha^i (1 - \alpha)^{(200-i)} \approx 0.6730 > \alpha_W,$$

$$P(W_2 \geq 22 | H_0) = 1 - \sum_{i=0}^{21} \binom{200}{i} \alpha^i (1 - \alpha)^{(200-i)} \approx 0.0005 < \alpha_W.$$

Hence, for the first sequence, we cannot reject the hypothesis that  $p_1 = \alpha$ . However, for the second sequence, the hypothesis  $p_2 = \alpha$  can be rejected. Therefore, it can be stated that there has been an increase in the variance in the second SIO, but not in the first.

---

In Example 2.4 we can see that the change is noticed in the second sequence when just a fraction of the observations are coming from random variables with a larger variance. To increase the certainty of the test we would like to have a smaller part of the observations from before the change has happened.

## 2.2.2 Windows

From Example 2.4 we get an indication that we could have identified the change closer to the change point if we would have done the test earlier and focused on the newer observations. This motivates a division of the SIO into windows where each window is a SIO whose number of  $\alpha$ -observations can be represented as  $W \sim Bin(k, p)$ , where  $k$  is the size of the window and  $p$  the probability for being an  $\alpha$ -observation.

Specifying the significance level,  $\alpha_W$ , of the test (2.1), makes it possible to set a threshold for the number of  $\alpha$ -observation that is expected in a window of size  $k$  when there is no change. If the number of observed  $\alpha$ -observations exceeds the threshold then the null hypothesis (2.1) can be rejected. The window threshold is defined as

**Definition 2.7 (Window threshold).** Let  $W \sim Bin(k, p)$  and  $\alpha_W$  be the significance level of the hypothesis

$$H_0 : p = \alpha \quad \text{versus} \quad H_1 : p > \alpha.$$

Then the window threshold,  $m_k$ , is the lowest number of  $\alpha$ -observations in a window where  $H_0$  can be rejected and is given by

$$m_k = \min_m P(W \geq m | H_0) < \alpha_W.$$

Note that with the window threshold, we control the type I error to be less than  $\alpha_W$ .

---

### Example 2.5

---

Returning to Example 2.4. We would like to use a smaller window for the possibility to notice the change earlier. By examining the second sequence in Figure 2.3 we can see that there are more than four 5%-observations within 15 observations from the change at 151. Table 2.1 gives us the window threshold for  $\alpha = 0.05$ . Here we can see that we could reject the null hypothesis in (2.1) with a significance level of  $\alpha_W = 0.01$ .

**Table 2.1:** Calculated  $m_k$ -value for different window sizes,  $k$ , 5%-observations with a significance level of  $\alpha_W \in \{0.05, 0.025, 0.01\}$ .

$k$	10	15	20	25	50	100	200	250
$m_k, \alpha_W = 0.05$	3	3	4	4	6	10	16	19
$m_k, \alpha_W = 0.025$	3	4	4	5	7	11	17	21
$m_k, \alpha_W = 0.01$	4	4	5	5	8	12	19	22

---

As the location of the change is uncertain, we will need to conduct an unknown amount of tests. Depending on the window size then for the same amount of observation, there will be more tests done with a smaller window, and the probability to make a type I error increases.

A general way to handle multiple testing is to work with the so-called Family-wise error rate, FWER. It is defined as the probability that at least one test of a group of related tests rejects a true  $H_0$ , a type I error. Miller (1966) denotes this probability family rate and he emphasizes that it doesn't make a distinction on the number of errors that occurs in the group of tests. By specifying FWER to  $\alpha_{FWER}$  there will be a limitation on the significance level of one window  $\alpha_W$ . The limit on  $\alpha_W$  to keep the FWER to  $\alpha_{FWER}$  could be specified in several ways. As an example if we have  $l$  independent test called  $A_1, \dots, A_l$  and a probability of a false positive as  $\alpha_{FWER}$ , the FWER, we have  $1 - \alpha_{FWER} = P(\bigcap_{i=1}^l A_i) = \prod_{i=1}^l P(A_i) = (1 - \alpha_W)^l$ , hence  $\alpha_W = 1 - (1 - \alpha_{FWER})^l$ . This is called the Šidák correction. We are aiming for a simpler solution that is slightly more stringent, the Bonferroni correction where there is no need to assume independence. The significance level for the tests of the Bonferroni correction is easily derived with the help of the inequality  $P(\bigcup_{i=1}^l A_i) \leq \sum_{i=1}^l P(A_i)$ . With the FWER,  $\alpha_{FWER} = P((\bigcap_{i=1}^l A_i)') = P(\bigcup_{i=1}^l A_i) \leq \sum_{i=1}^l P(A_i) = l\alpha_W$ , hence the inequality will be fulfilled if  $\alpha_W = \alpha_{FWER}/l$ .

For online monitoring, we have the risk of infinitely many tests before a change occurs. This means that it is impossible to set a limit on the number of type I errors with a fixed  $\alpha_W$ . To handle the problem with the infinite number of tests we limit the construction of the algorithm so the type I error is adjusted for a specified amount of observations, the expected maximum number of observations, EMO, before a change. We will denote it as  $N_{max}$ . With the EMO specified we have the number of tests by dividing the  $N_{max}$  with the window size. The significance level adjusted with Bonferroni corrections for  $k$ , denoted  $\alpha_W(k)$ , equal

$$\alpha_W(k) = \alpha_{FWER} \frac{k}{N_{max}}. \quad (2.2)$$

The significance levels depending on the window sizes can be taken into account in the window threshold by updating Definition 2.7 as follows.

**Definition 2.8 (Family-wise window threshold).** Let  $W \sim Bin(k, p)$  and  $\alpha_W(k)$  be the significance level of the hypothesis

$$H_0 : p = \alpha \quad \text{versus} \quad H_1 : p > \alpha,$$

for window size  $k$ . Then the window threshold,  $m_k$ , is the lowest number of observations in a window where  $H_0$  can be rejected. The threshold is given by

$$m_k = \min_m P(W \geq m | H_0) < \alpha_W(k).$$

The next part to address is how to choose the window size. Our focus here will be on windows that are disjoint and therefore independently distributed. There are several factors that play a role in the choice of window size. The first factor is that since we are aiming for an online method we would like to have as few observations after the change as possible. The second factor is that we want to control the type II error, which decreases when more observation is included in the window. By combining the definition of window threshold and the probability

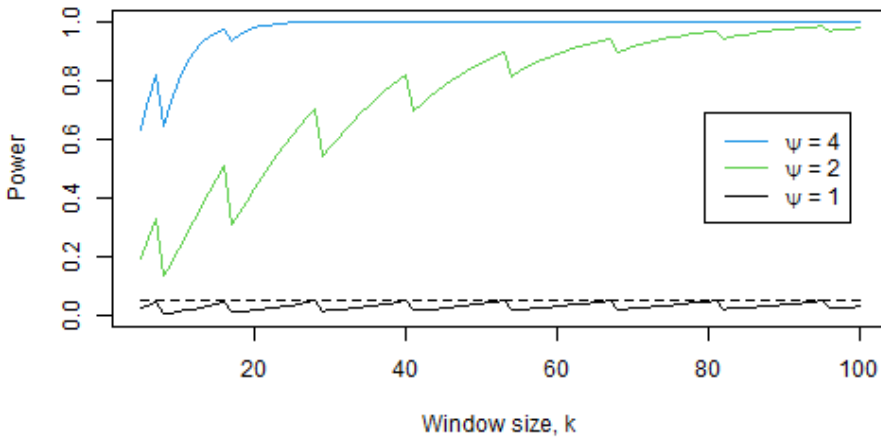
for an  $\alpha$ -observations for a given change, see Theorem 3.1, we can calculate the probability to reject  $H_0$  in (2.1) for a given size of the change,  $\psi$ , i.e., the power of the test for a window.

**Definition 2.9 (Window probability).** Let the window threshold,  $m_k$ , be defined as in Definition 2.8, and let  $W \sim \text{Bin}(k, p_\psi)$  represent the number of  $\alpha$ -observations in a window of size  $k$ , where  $p_\psi$  is the probability for an  $\alpha$ -observation with change  $\psi = \sigma_2^2/\sigma_1^2$ . Then the window probability,  $p_w(k, \psi)$  is defined as

$$p_w(k, \psi) = P(W \geq m_k).$$

### Example 2.6

From Definition 2.9 we have the possibility to find the probability of identifying a change that depends on the variance ratio,  $\psi$  and the window size. This is illustrated for  $\psi \in \{1, 2, 4\}$  with an increasing window size in Figure 2.4. In other words, the power of testing according to (2.1) if a window consists of observation from a distribution with a higher variance. In the graph, we can see an increasing trend depending on the size of the window but also sharp drops in the probability at specific window sizes. These drops happen when the threshold,  $m_k$ , increases to keep the probability for false positives less than 5%.



**Figure 2.4:** The figure shows the probability of detecting a change when the observations in the window have a variance that has increased with a multiple of 2 or 4 of the initial variance.

In Figure 2.4 we can see that we have higher power to find a change with a larger window. The drawback is that with each observation needed for the window, we will be one observation further away from the change. By using a smaller window we will have the possibility to do more tests with the same amount of observations which could increase the chance of detecting a change early compared to a larger window but also increase the risk of type I error. Therefore, there is a compromise between maintaining a consistent FWER by implementing a stricter significance level for smaller windows, versus utilizing a larger window with more observations by default.

Let us denote  $w(k, \psi)$  as the number of windows examined after a change before the change was identified. With the window probability  $p(k, \psi)$  the probability for a specific value of  $w(k, \psi)$  is expressed as  $(1 - p(k, \psi))^{w(k, \psi) - 1} p(k, \psi)$  which is the probability mass function for the geometric distribution with the probability  $p(k, \psi)$ . Hence,  $w(k, \psi)$  is an observation from a geometric distributed random variable. This will help us to pick a good window size.

To choose the window size that uses the smallest number of observations after a change we need to specify what power we would like to have. Let  $\beta$  be the power of finding the change with the smallest number of observations. Then the following theorem will give the size of the window that is used.

**Theorem 2.1**

Let  $\beta$  be the power for a change to be identified and  $p(k, \psi)$  the windows probability in Definition 2.9. Then the window size that gives the smallest number of observations after a change with the specified power is calculated as

$$k_{min} = \arg \min_k k \left\lceil \frac{\ln(1 - \beta)}{\ln(1 - p_w(k, \psi))} \right\rceil.$$

**Proof:** Let  $W(k, \psi)$  be the random variable for the number of windows examined after a change for it to be identified. Then

$$\begin{aligned} \beta &\geq P(W(k, \psi) \leq w(k, \psi)) = \sum_{i=1}^{w(k, \psi)} (1 - p(k, \psi))^{i-1} p(k, \psi) \\ &= 1 - (1 - p(k, \psi))^{w(k, \psi)}. \end{aligned}$$

Hence,

$$(1 - p(k, \psi))^{w(k, \psi)} \geq 1 - \beta$$

and

$$w(k, \psi) \geq \frac{\ln(1 - \beta)}{\ln(1 - p(k, \psi))}. \quad (2.3)$$

Since  $w(k, \psi)$  is an integer we could round up the value and keep the inequality as

$$w(k, \psi) \geq \left\lceil \frac{\ln(1 - \beta)}{\ln(1 - p(k, \psi))} \right\rceil.$$

Then  $w(k, \psi)$  is the smallest number of windows for each  $k$  and  $\psi$  needed to have a certainty of  $\beta$  of identifying the change. The number of observations needed

for each window size is given by  $k w(k, \psi)$ , hence the window size that gives the smallest number of observations after a change equals

$$k_{min} = \arg \min_k k c_w(k, \psi) = \arg \min_k k \left[ \frac{\ln(1 - \beta)}{\ln(1 - p_w(k, \psi))} \right].$$

□

### 2.2.3 Alert algorithm

Our proposed algorithm is a straightforward method for detecting changes in the variance of a sequence. It involves breaking the sequence into non-overlapping windows of fixed size, denoted by  $k$ . Within each window, we filter out the  $\alpha$ -observations, which are observations that are unlikely to occur. Next, we compare the number of  $\alpha$ -observation in the windows in sequential order to the window threshold,  $m_k$ . If the number of  $\alpha$ -observations in a window exceeds the threshold,  $m_k$ , then we can reject the null hypothesis, denoted by  $H_0$  in hypothesis (2.1). This indicates that a change in variance has occurred inside or before the window.

To have the algorithm adjustable for different situations we have created a general algorithm. The algorithm consists of two stages. The first stage is the initiation from our assumption about the SIO and the limitation with type I and type II errors with the specified EMO gives us the window threshold and size. The second stage searches through the SIO and evaluates it for a change with the help of an alert function. An alert function  $\delta$  is defined as a function that maps a sequence to either 1 for an alert of a change or 0 when nothing outside what is expected happens. The alert function for our algorithm is created with the help of a sequence  $z_{(1,N)} = \{z_1, \dots, z_N\}$ , where

$$z_i = \begin{cases} 1, & \text{if } x_i \text{ is an } \alpha\text{-observation,} \\ 0, & \text{otherwise.} \end{cases} \quad (2.4)$$

We will denote our alert function  $\delta_\alpha$  and it will be defined as

$$\delta_\alpha(z_{(1,k)}) = \begin{cases} 1, & \text{if } \sum_1^k z_i \geq m_k, \\ 0, & \text{otherwise.} \end{cases} \quad (2.5)$$

The first stage is the initiation algorithm and it is seen in Algorithm 2.1. With the simple adjustment of Algorithm 2.1 it can also be used in particular multivariate cases. The required modification involves applying Corollary 3.1 with  $\Psi$  to compute the probability of an  $\alpha$ -observation, denoted as  $p_\Psi$ . Instead of calculating the thresholds, we compute  $\mathbf{S}_1 = \frac{1}{n_0} \sum_{i=1}^{n_0} x_i^2$ , serving as an estimation for  $\Sigma_1$ , where  $n_0$  is the number of observation of a SIO without any changes.

---

**Algorithm 2.1 (Initiation)**


---

**Required:**  $\alpha$ ;  $\beta$ ;  $\alpha_{FWER}$ ;  $\mathcal{N}_{max}$ ;  $\psi$

**Data:**  $x_{(1,n_0)}$  a SIO known to have no change

**Output:**

- thresholds  $c_1$  and  $c_2$ ;
- window threshold  $m_k$ ;
- window size  $k$ .

**Steps:**

- 1) calculate  $m_k$  from Definition 2.8 with  $\alpha$ ,  $\alpha_{FWER}$  and  $\mathcal{N}_{max}$
  - 2) calculate  $p_\psi$  from Theorem 3.1 with  $\alpha$  and  $\psi$
  - 3) calculate  $p(k, \psi)$  from Definition 2.9 with  $m_k$  and  $p_\psi$
  - 4) calculate  $k$  from Theorem 2.1 with  $p(k, \psi)$  and  $\beta$
  - 5) estimate threshold for  $\alpha$ -observation  $c_1$  and  $c_2$  with  $x_{(1,n_0)}$
- 

The next stage is the search algorithm. It uses the output from the first stage for its setup of searching through the SIO. The search algorithm is specified in Algorithm 2.2.

---

**Algorithm 2.2 (Search)**


---

**Required:** thresholds  $c_1$  and  $c_2$ ; window threshold  $m_k$ ; window size  $k$

**Data:**  $y_{(1,N)}$  an unknown SIO

**Steps:**

- 1) **set**  $i = 1$
  - 2) **construct**  $z_{(1,k)}$  from  $y_{(i,i+k-1)}$
  - 3) **if**  $\delta(z_{(1,k)}) = 1$  **then**
    - **go to end**
  - else if**  $\delta(w_i) = 0$  **then**
    - **set**  $i = i + k$
    - **go to step 2**
  - 4) **end**
- 

With the algorithms, we have now a framework that can be used for both univariate and multivariate SIOs. We will explore using the framework of creating an alert function with a search algorithm with non-overlapping windows with Roy's largest eigenvalue criteria, Kullback-Leibler divergence, and Bhattacharyya distance.

# 3

---

## Results

This chapter will present the results needed to set up the alert function in both the univariate and multivariate cases. In the multivariate case, we also consider the largest eigenvalue of the sample covariance matrix to detect a change and some divergence information measures, such as the Kullback-Leibler divergence and the Bhattacharyya distance.

### 3.1 Univariate case

As we have seen in Example 2.3 the probability for an  $\alpha$ -observations increases when the variance ratio  $\psi$  increases. To be able to know how large changes can be identified the size of the change needs to be investigated. In the following theorem, we will derive an expression for the probability for an observation to be an  $\alpha$ -observation that only depends on  $\alpha$  and the relative increase in variance,  $\psi$ .

#### Theorem 3.1

Let  $X \sim \mathcal{N}(\mu, \sigma_1^2)$ ,  $Y \sim \mathcal{N}(\mu, \sigma_2^2)$ , and for a given probability  $\alpha$ , define the constants  $c_1$  and  $c_2$  for an  $\alpha$ -observation according to Definition 2.6. Furthermore, also assume that the ratio  $\psi = \sigma_2^2/\sigma_1^2$  characterizes the variance change. Then the probability for an  $\alpha$ -observation to occur depends on  $\psi$  and is given by

$$p_\psi = P((Y < c_1) \cup (Y > c_2)) = 2\Phi\left(\frac{1}{\sqrt{\psi}}\Phi^{-1}(\alpha/2)\right), \quad (3.1)$$

where  $\Phi$  is the standard normal distribution function.

**Proof:** Notice that the constants  $c_1$  and  $c_2$  have the following values

$$\alpha/2 = P(X < c_1) = P\left(\frac{X - \mu}{\sigma_1} < \frac{c_1 - \mu}{\sigma_1}\right),$$

and

$$c_1 = \sigma_1 \Phi^{-1}(\alpha/2) + \mu.$$

Furthermore,

$$\alpha/2 = P(X > c_2) = 1 - P\left(\frac{X - \mu}{\sigma_1} < \frac{c_2 - \mu}{\sigma_1}\right)$$

with

$$c_2 = \sigma_1 \Phi^{-1}(1 - \alpha/2) + \mu.$$

Since  $(Y < c_1) \cap (Y > c_2) = \{\emptyset\}$  we have that  $P((Y < c_1) \cup (Y > c_2)) = P(Y < c_1) + P(Y > c_2)$ . Then using the expression for  $c_1$  in  $P(Y < c_1)$  equals

$$\begin{aligned} P(Y < c_1) &= \Phi\left(\frac{c_1 - \mu}{\sigma_2}\right) = \Phi\left(\frac{\sigma_1}{\sigma_2} \Phi^{-1}(\alpha/2) + \frac{\mu - \mu}{\sigma_2}\right) \\ &= \Phi\left(\frac{\sigma_1}{\sigma_2} \Phi^{-1}(\alpha/2)\right), \end{aligned}$$

and the expression for  $c_2$  in  $P(Y > c_2)$  equals

$$\begin{aligned} P(Y > c_2) &= 1 - P(Y < c_2) = 1 - \Phi\left(\frac{c_2 - \mu}{\sigma_2}\right) \\ &= 1 - \Phi\left(\frac{\sigma_1}{\sigma_2} \Phi^{-1}(1 - \alpha/2) + \frac{\mu - \mu}{\sigma_2}\right) \\ &= 1 - \Phi\left(-\frac{\sigma_1}{\sigma_2} \Phi^{-1}(\alpha/2)\right) \\ &= \Phi\left(\frac{\sigma_1}{\sigma_2} \Phi^{-1}(\alpha/2)\right). \end{aligned}$$

With  $\psi = \sigma_2^2/\sigma_1^2$  we have that

$$P((Y < c_1) \cup (Y > c_2)) = P(Y < c_1) + P(Y > c_2) = 2\Phi\left(\frac{1}{\sqrt{\psi}} \Phi^{-1}(\alpha/2)\right),$$

which concludes the proof.  $\square$

---

### Example 3.1

---

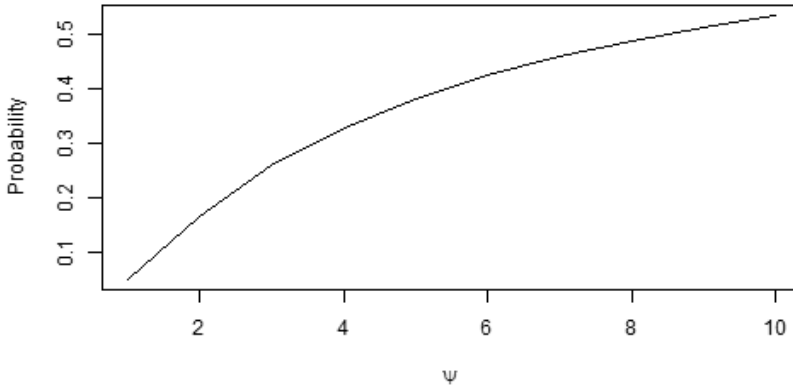
Figure 3.1 shows the probability for an observation to be a 5%-observation when the variance ratio increases. Note that with  $\psi = 3$  about a quarter of the observations are 5%-observations. When  $\psi$  increases 9 times then more than 50% are 5%-observations.

---

With Theorem 3.1 we can use Algorithm 2.1. Then with our specification of the  $\alpha$ -observation we have that (2.4) can be calculate as

$$z_i = \begin{cases} 1, & \text{if } (x_i < c_1) \cup (x_i > c_2), \\ 0, & \text{otherwise.} \end{cases}$$

This makes it easy to calculate  $\delta_\alpha(z_{(1,k)})$  and therefore it is suggested to use Algorithm 2.2 to search for a change.



**Figure 3.1:** The graph shows the probability for a 5%-observation to occur when the variance ratio  $\psi$ , defined in Theorem 3.1, increases from 1 to 10.

---

### Example 3.2

---

We would like to decide the size of a window for 5%-observations that will make the algorithm have a power of 90% using as few observations as possible. Our setting is that the FWER is  $\alpha_{FWER} = 0.05$  with an EMO,  $N_{max} = 200$ , and with a power of at least  $\beta = 0.9$  probability. We have assumed the change to triple the variance,  $\psi = \sigma_2^2/\sigma_1^2 = 3$ .

From Theorem 3.1 we deduce that the probability for a 5%-observation to occur is  $p_{\psi=3} = 0.258$ . Using  $\alpha_W(k) = k\alpha_{FWER}/N_{max}$  we can employ Definition 2.8 to determine the thresholds  $m_k$  for different window sizes,  $k$ , see Table 2.1. From here we use Definition 2.9 to calculate  $p_w(k, \psi)$ . From Theorem 2.1 the least amount of observations after a change with a specified power is derived. Table 3.1 shows that the window size associated with the lowest number of observations is  $k = 34$ , accompanied by a window threshold of 6 observations. Then the power is 0.904 which is the smallest window size with a power greater than 0.9. We can also see how many windows a smaller window size needs to have to fulfill the conditions of type I error and power.

---

## 3.2 Multivariate case

Working with multivariate SIO makes the problem of detecting a change in the covariance matrix more complex since the change can happen in any of the dimensions or in a combination of them. One idea is to apply a method for the univariate case on each dimension. Then with knowledge of the univariate method, it will be

**Table 3.1:** The table shows how many observations are needed after a change to be noticed with a probability  $\beta = 0.9$  when 5%-observations is used with a FWER  $\alpha_{FWER} = 0.05$  and EMO before a change on  $N_{max} = 200$ .

Window size $k$	5	13	24	34	36	50
Window threshold, $m_k$	3	4	5	6	6	7
Number of windows	20	4	2	1	1	1
Number of observations	100	52	48	34	36	50
Probability, $p_w(k, 3)$	0.112	0.442	0.780	0.904	0.931	0.986

possible to understand what is going on and how to implement the method. In this idea, there are two major drawbacks. First, the time complexity will increase in at least a linear scale with the number of dimensions. Second, it can miss the more complex changes when variables are affecting each other. The next way of creating a method for finding change points is to use a suitable map from the multivariate observation to a one-dimensional value and then use a univariate method. The last method is to work with the multivariate sequence directly.

Our method for the multivariate case will pursue the projection down to a univariate sequence and then use the method described in Chapter 2. We are also exploring the use of the largest eigenvalue, Kullback-Leibler divergence, and Bhattacharyya distance which are working with the sequence directly.

### 3.2.1 Mahalanobis

In this section, we take a closer look at using the Mahalanobis distance to map the multivariate observations to one dimension. To be able to use the same framework presented in Chapter 2 and in Section 3.1 we need another way to calculate the probability for an  $\alpha$ -observation. In the univariate case, this is done with Theorem 3.1. We need a corresponding theorem for the projected sequence to create an alert algorithm for the multivariate case.

We will use  $\mathbf{x}$  as observation from the random vector  $\mathbf{X}$  that has the covariance matrix,  $\Sigma_1$ , and  $\mathbf{y}$  as observations from the random vector  $\mathbf{Y}$  with the covariance matrix,  $\Sigma_2$ . Let us look at the squared Mahalanobis distance. It gives us a measure of the distance to  $\boldsymbol{\mu}$  for an observation depending on the expected "layout" of the statistic field. The Mahalanobis distance is defined as follows:

**Definition 3.1 (Mahalanobis distance).** Given a covariance matrix  $\Sigma$  of size  $p \times p$  and a mean  $\boldsymbol{\mu}$ , and random vector  $\mathbf{X}$  then the Mahalanobis distance of  $\mathbf{X}$  is given by

$$d_M(\mathbf{X}; \boldsymbol{\mu}, \Sigma) = \sqrt{(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})}.$$

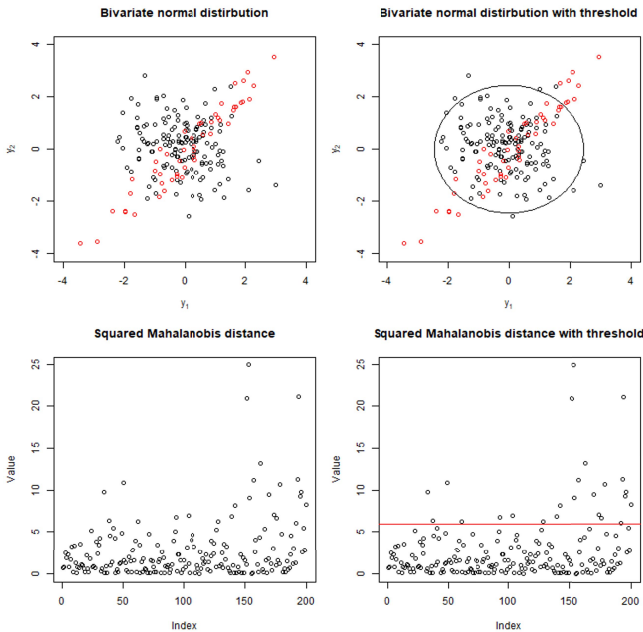
Furthermore, denote  $d_M(\mathbf{X}; \mathbf{0}, \Sigma) = d_M(\mathbf{X}; \Sigma)$ . We also know that if  $\mathbf{X}$  is distributed as  $\mathcal{N}_p(\mathbf{0}, \Sigma)$  then  $\mathbf{Z} = \Sigma^{-1/2} \mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathcal{I})$  and  $d_M^2(\mathbf{X}; \Sigma) = \mathbf{X}' \Sigma^{-1} \mathbf{X} = \mathbf{Z}' \mathbf{Z}$  is distributed as  $\chi^2(p)$ . Therefore we will use  $\chi^2(p)$  as the distribution of the random variable to define a multivariate  $\alpha$ -observation as in Definition 2.5.

**Example 3.3**

The collection of images in Figure 3.2 shows a bivariate normal distribution with 200 observations with a change at observation 151. The change was from a covariance matrix equal to the identity matrix to the covariance matrix

$$\Sigma = \begin{bmatrix} 1.55 & 1.88 \\ 1.88 & 2.53 \end{bmatrix},$$

that has eigenvalues 4 and 0.1. The circle thresholds are the square root of the  $\chi^2$ -distribution with two degrees of freedom at 95%. The threshold on the Mahalanobis distance is directly from the  $\chi^2$ -distribution.



**Figure 3.2:** Upper row is a bivariate normal distribution with a change at observation 151 of 200 and a threshold at 95%. The lower row is the squared Mahalanobis distance of the same observation with the same threshold.

With the Mahalanobis distance as a mapping from  $p$  dimensions down to one dimension, we would like to know what kind and size of changes are possible to notice in the new univariate SIO. What we are looking for is a similar theorem to Theorem 3.1 that could give the probability for an  $\alpha$ -observation for  $d_M^2(\mathbf{Y}; \Sigma_1)$ , when  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \Sigma_2)$  compared to  $d_M^2(\mathbf{X}; \Sigma_1)$ . For simplicity, we assume that  $\mu = \mathbf{0}$ .

The following theorem from Graybill (1976) gives the distribution of  $d_M^2(\mathbf{Y}; \Sigma_1)$  as a sum of weighted  $\chi^2$ -distributions, where the weights are the eigenvalues of  $\Psi = \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}$ .

**Theorem 3.2**

Let  $\mathbf{Y} \sim N_p(\mathbf{0}, \Sigma_2)$ , then the random variable  $d_M^2(\mathbf{Y}; \Sigma_1)$  has the same distribution as  $\sum_{i=1}^p d_i w_i$  where  $d_i$  are the eigenvalues of  $\Psi = \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}$  and  $w_i$  are independent  $\chi^2$  random variables with one degree of freedom.

Since  $\chi^2$ -distribution is a special case of the gamma distribution we have that  $d_i w_i$  is distributed as  $G(1/2, 1/(2d_i^2))$ , where  $G(a, b)$  denotes the gamma distribution with the shape parameter  $a$  and the rate parameter  $b$ . Hence,  $d_M^2(\mathbf{Y}; \Sigma_1)$  is distributed as a sum of independent gamma distributed random variables.

Moschopoulos (1985) gives an expression of the exact probability density function of a sum of independent gamma-distributed random variables. To write down the density function the following expressions are needed

$$C = \prod_{i=1}^p \left( \frac{b_i}{b_1} \right), \quad \gamma_k = \sum_{i=1}^p a_i \frac{(1 - b_i/b_1)^k}{k}, \quad \rho = \sum_{i=1}^p a_i, \quad (3.2)$$

and  $\delta_k$  that is recursively defined as

$$\delta_k = \frac{1}{k+1} \sum_{i=1}^{k+1} i \gamma_i \delta_{k+1-i}, \quad k = 0, 1, \dots, \quad \delta_0 = 1, \quad (3.3)$$

and the gamma function  $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ .

**Theorem 3.3**

Let  $X_1, \dots, X_p$  be a set of mutually independent gamma variables with parameters  $a_i > 0$  and  $\beta_i > 0$ . Then the density of  $Y = \sum_{i=1}^p X_i$  can be expressed as

$$g(y) = C \sum_{k=0}^{\infty} \frac{\delta_k y^{\rho+k-1} b_1^{\rho+k} e^{-yb_1}}{\Gamma(\rho+k)}, \quad y > 0, \quad (3.4)$$

where  $\rho$ ,  $C$  is given in (3.2) and  $\delta_k$  in (3.3).

**Proof:** See Moschopoulos (1985). □

The recursive  $\delta_k$  and the infinite sum in the density presented in Theorem 3.3 are demanding to use. This opens up for using an approximation of the sum of gamma variables. Feng et al. (2016) present a way to make an approximation of a sum of correlated gamma variables to one new gamma variable. Let  $\xi_i, i = 1, \dots, p$ , be correlated gamma distributed random variables with the shape parameter  $a$ , parameter  $\beta_i = a/\Omega_i$ , where  $\Omega_i = E(\xi_i)$  and let  $\mathbf{V}$  be the covariance matrix of the random vector  $(\xi_1, \dots, \xi_p)'$ . Then the distribution of  $\xi = \sum_{i=1}^p \xi_i$  can be approximated as

$$G \left( \left( \sum_{i=1}^p \Omega_i \right)^2 \middle/ \sum_{j=1}^p \sum_{l=1}^p \mathbf{V}_{j,l}, \quad \sum_{i=1}^p \Omega_i \middle/ \sum_{j=1}^p \sum_{l=1}^p \mathbf{V}_{j,l} \right), \quad (3.5)$$

where  $\mathbf{V}_{j,l}$  is the  $(j, l)$  entry in the matrix  $\mathbf{V}$ . For the approximation, it is used that  $\xi_i$  can be represented as a weighted sum of independent gamma distributed

random variables,  $w_k, k = 1, \dots, p$ , with unit variance. The weights depend on the covariance matrix  $\mathbf{V}$ , and the parameters of  $w_k$  are derived by matching the first and second moments of all  $\xi_i$  with the respective weighted sums of  $w_k$ .

The main result of this section is the corresponding multivariate theorem for the univariate Theorem 3.1, the probability for being an  $\alpha$ -observation but in the multivariate setting. The theorem presents the approximate distribution of  $d_M(\mathbf{Y}, \boldsymbol{\Sigma}_1)$ , where  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_2)$ . A corollary states the probability of being an  $\alpha$ -observation. To prove the results we need two functions, the Hadamard product and a function that maps a vector to a diagonal matrix.

The Hadamard product is defined as the element-wise matrix multiplication of two equal-sized matrices,  $\mathbf{A}$  and  $\mathbf{B}$ , such that

$$\mathbf{A} \circ \mathbf{B} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1p}b_{1p} \\ a_{21}b_{21} & a_{22}b_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1}b_{n1} & \dots & \dots & a_{np}b_{np} \end{pmatrix}.$$

The diagonalization function is defined as

$$\text{diag}(\mathbf{x}) = \begin{pmatrix} x_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & x_p \end{pmatrix}.$$

These functions help us to formulate the following result from Jitkrittum (2013).

**Lemma 3.1**

Let  $\mathbf{A}$  and  $\mathbf{B}$  be matrices of size  $n \times p$ , and  $\mathbf{x}$  and  $\mathbf{y}$  vector of length  $n$  and  $p$ , respectively. Then

$$\mathbf{x}'(\mathbf{A} \circ \mathbf{B})\mathbf{y} = \text{tr}((\text{diag}(\mathbf{x})\mathbf{A})'(\mathbf{B}\text{diag}(\mathbf{y}))).$$

**Proof:** Let us start noticing

$$(\mathbf{A} \circ \mathbf{B})\mathbf{y} = \left( \sum_{i=1}^p a_{1i}b_{1i}y_i \quad \sum_{i=1}^p a_{2i}b_{2i}y_i \quad \dots \quad \sum_{i=1}^p a_{ni}b_{ni}y_i \right)'$$

Then we have that

$$\begin{aligned} \mathbf{x}'(\mathbf{A} \circ \mathbf{B})\mathbf{y} &= \sum_{i=1}^p a_{1i}b_{1i}y_i x_1 + \sum_{i=1}^p a_{2i}b_{2i}y_i x_2 + \dots + \sum_{i=1}^p a_{ni}b_{ni}y_i x_n \\ &= \sum_{j=1}^n \sum_{i=1}^p a_{ji}b_{ji}y_i x_j. \end{aligned}$$

Looking at the factors of the right-hand side

$$(\text{diag}(\mathbf{x})\mathbf{A})' = \begin{pmatrix} x_1 a_{11} & x_2 a_{21} & \dots & x_n a_{n1} \\ x_1 a_{12} & x_2 a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ x_1 a_{1p} & \dots & \dots & x_n a_{np} \end{pmatrix}$$

and

$$\mathbf{B} \operatorname{diag}(\mathbf{y}) = \begin{pmatrix} y_1 b_{11} & y_2 b_{12} & \cdots & y_p b_{1p} \\ y_1 b_{21} & y_2 b_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ y_1 b_{n1} & \cdots & \cdots & y_p b_{np} \end{pmatrix}.$$

Now we are ready to write down the full expression of the right-hand side and compare it with the left-hand side as follows:

$$\begin{aligned} \operatorname{tr}((\operatorname{diag}(\mathbf{x})\mathbf{A})'(\mathbf{B} \operatorname{diag}(\mathbf{y}))) &= \sum_{j=1}^n x_j a_{j1} y_1 b_{j1} + \sum_{j=1}^n x_j a_{j2} y_2 b_{j2} + \cdots \\ &\quad + \sum_{j=1}^n x_j a_{jp} y_p b_{jp} \\ &= \sum_{i=1}^p \sum_{j=1}^n x_j a_{ji} y_i b_{ji} = \sum_{j=1}^n \sum_{i=1}^p a_{ji} b_{ji} y_i x_j \\ &= \mathbf{x}'(\mathbf{A} \circ \mathbf{B})\mathbf{y}. \end{aligned}$$

□

The next result gives an expression for the covariance matrix of  $(Y_1^2, Y_2^2, \dots, Y_p^2)$ .

**Lemma 3.2**

Let  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$  then the covariance matrix  $\mathbf{V}$  for  $(Y_1^2, Y_2^2, \dots, Y_p^2)'$  equals

$$\mathbf{V} = 2\Sigma \circ \Sigma.$$

**Proof:** Let us denote the elements in  $\Sigma$  by  $\sigma_{jl}$ . The covariance between  $Y_j$  and  $Y_l$  is calculated as  $\operatorname{Cov}(Y_j^2, Y_l^2) = E(Y_j^2 Y_l^2) - E(Y_j^2)E(Y_l^2)$ . We have since  $\sigma_{ll} Y_l^2 \sim \chi^2(1)$

$$E(Y_l^2) = \sigma_{ll},$$

so  $E(Y_j^2)E(Y_l^2) = \sigma_{jj}\sigma_{ll}$ . For  $E(Y_j^2 Y_l^2)$  we will use the moment generating function for a multivariate normal distribution with  $\boldsymbol{\mu} = \mathbf{0}$ , which equals  $M(\mathbf{t}) = e^{\frac{\mathbf{t}'\Sigma\mathbf{t}}{2}}$ , and  $E(Y_j^2 Y_l^2) = \frac{\partial^4}{\partial t_j^2 \partial t_l^2} M(\mathbf{t}) \Big|_{t_j=t_l=0}$ . This gives

$$E(Y_j^2 Y_l^2) = \frac{\partial^4}{\partial t_j^2 \partial t_l^2} M(\mathbf{t}) \Big|_{t_j=t_l=0} = \cdots = \sigma_{jj}\sigma_{ll} + 2\sigma_{jl}^2,$$

and  $E(Y_j^2)E(Y_l^2) = \sigma_{jj}\sigma_{ll}$ , hence  $\operatorname{Cov}(Y_j^2, Y_l^2) = 2\sigma_{jl}^2$ . Thus,

$$\mathbf{V} = 2\Sigma \circ \Sigma.$$

□

Now we are ready to formulate and prove the probability of detecting an  $\alpha$ -observation when there has been a change between two covariance matrices.

**Theorem 3.4**

Let  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_2)$  then  $d_M^2(\mathbf{Y}; \mathbf{\Sigma}_1) = \mathbf{Y}'\mathbf{\Sigma}_1^{-1}\mathbf{Y}$  is approximately distributed as  $G(a, b)$ , where

$$a = \frac{(\text{tr}(\mathbf{\Psi}))^2}{2\text{tr}(\mathbf{\Psi}^2)} \quad \text{and} \quad b = \frac{\text{tr}(\mathbf{\Psi})}{2\text{tr}(\mathbf{\Psi}^2)},$$

with  $\mathbf{\Psi} = \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}$ .

**Proof:** Let  $\mathbf{Z} = \mathbf{\Sigma}_1^{-1/2}\mathbf{Y}$  then  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2})$ , where  $\mathbf{\Sigma}_1^{1/2}$  is a symmetric square root. Let  $\sigma_i^2$  be the variance for  $Z_i$  and therefore also diagonal elements of  $\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}$ , which gives us that  $\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^p Z_i^2$ , where  $Z_i^2 \sim G(a_0 = 1/2, b_i = 1/(2\sigma_i^2))$ . From Feng et al. (2016) we have that a sum of correlated Gamma random variables can be approximated by another Gamma random variable as in (3.5). From Lemma 3.2 we have that the covariance matrix of  $(Z_1^2, \dots, Z_p^2)'$  is  $\mathbf{C} = 2\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2} \circ \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}$ . Then by writing the sum of all elements of the covariance matrix as  $\mathbf{1}'\mathbf{C}\mathbf{1}$  and combining it with Lemma 3.1 yields

$$\begin{aligned} \mathbf{1}'\mathbf{C}\mathbf{1} &= \mathbf{1}'\left(2\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2} \circ \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}\right)\mathbf{1} \\ &= 2\text{tr}\left(\left(\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}\right)'\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}\right) \\ &= 2\text{tr}\left(\left(\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}\right)^2\right). \end{aligned}$$

In (3.5) adding the expected values  $\Omega_i = E(Z_i^2) = \sigma_i^2$  and  $\mathbf{1}'\mathbf{C}\mathbf{1}$  implies that

$$\begin{aligned} a &= \frac{(\sum_{i=1}^p \Omega_i)^2}{\sum_{j=1}^p \sum_{l=1}^p \mathbf{V}(j, l)} = \frac{(\text{tr}(\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}))^2}{2\text{tr}((\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2})^2)} = \frac{(\text{tr}(\mathbf{\Psi}))^2}{2\text{tr}(\mathbf{\Psi}^2)}, \\ b &= \frac{\sum_{i=1}^p \Omega_i}{\sum_{j=1}^p \sum_{l=1}^p \mathbf{V}(j, l)} = \frac{\text{tr}(\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2})}{2\text{tr}((\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2})^2)} = \frac{\text{tr}(\mathbf{\Psi})}{2\text{tr}(\mathbf{\Psi}^2)}. \end{aligned}$$

Hence,  $\mathbf{Z}'\mathbf{Z} = \mathbf{Y}'\mathbf{\Sigma}_1^{-1}\mathbf{Y} \sim G(a, b)$ . □

**Corollary 3.1**

Let  $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_1)$  and  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_2)$  with  $\alpha = P(\mathbf{X}'\mathbf{\Sigma}_1^{-1}\mathbf{X} > c)$ . Then an approximation of the probability for an  $\alpha$ -observation is given by

$$P(\mathbf{Y}'\mathbf{\Sigma}_1^{-1}\mathbf{Y} > c) \approx 1 - \frac{\gamma(a, bc)}{\Gamma(a)},$$

where  $a$  and  $b$  are as in Theorem 3.4,  $c = \chi_{1-\alpha}^2(p)$ , and  $\gamma(a, x)$  is the incomplete gamma function defined by

$$\gamma(a, x) = \int_0^x t^{a-1}e^{-t} dt.$$

**Proof:** Note that  $\Sigma_1^{-1/2}\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathcal{I}_p)$ . Hence  $\mathbf{X}'\Sigma_1^{-1}\mathbf{X} \sim \chi^2(p)$  and

$$\alpha = P(\mathbf{X}'\Sigma_1^{-1}\mathbf{X} > c),$$

with

$$c = \chi_{1-\alpha}^2(p).$$

From Theorem 3.4 we have that  $d_M^2(\mathbf{Y}; \Sigma_1) = \mathbf{Y}'\Sigma_1^{-1}\mathbf{Y}$  is approximately  $G(a, b)$  so the probability of an  $\alpha$ -observation is

$$P(\mathbf{Y}'\Sigma_1^{-1}\mathbf{Y} > c) \approx 1 - \frac{\gamma(a, bc)}{\Gamma(a)}.$$

□

In Example 3.1, we could see that the probability for an  $\alpha$ -observation to occur in the univariate case increases when the variance ratio increases. In the following example, this is extended to the bivariate case.

---

**Example 3.4**

---

With the help of Corollary 3.1 we can calculate the approximated probability for an  $\alpha$ -observation. In Figure 3.3 we can find the probability for a 5%-observation from a bivariate normal distribution with  $\boldsymbol{\mu} = \mathbf{0}$ . The covariance matrix before the change equals the identity matrix, and after the change the covariance matrix has the structure

$$\Sigma_2 = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where  $\rho$  takes values in the range from -0.99 to 0.99 and  $\sigma^2 = 1, \dots, 9$ . Since  $\Sigma_1 = \mathcal{I}$  we have that  $\Psi = \Sigma_2$ . From  $\Psi$  it follows that the parameter for the approximated gamma distribution of  $\mathbf{Y}'\Sigma_1^{-1}\mathbf{Y}$  will be

$$a = \frac{(\text{tr}(\Psi))^2}{\text{tr}(\Psi^2)} = \frac{1}{1 + \rho^2}, \quad b = \frac{\text{tr}(\Psi)}{\text{tr}(\Psi^2)} = \frac{1}{2\sigma^2(1 + \rho^2)}.$$

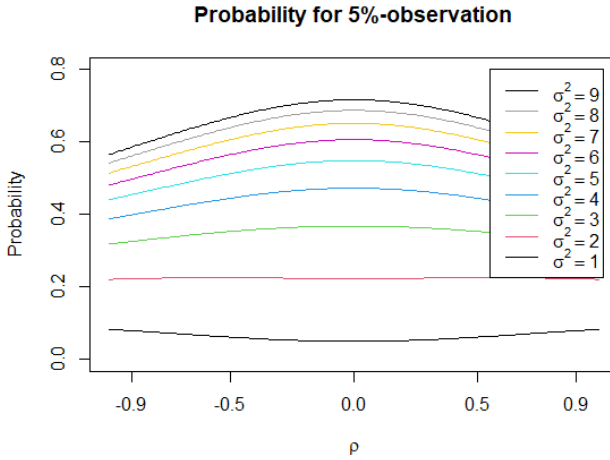
What we see in Figure 3.3 is  $P(\mathbf{Y}'\Sigma_1^{-1}\mathbf{Y} > \chi_{0.95}^2(2))$ . This can be compared with the univariate case in Example 3.1 which showed the probability for an  $\alpha$ -observation. Another thing to notice is that the probability mostly depends on the multiple,  $\sigma^2$ , of the matrix rather than the structure of the covariance matrix.

---

Example 3.4 shows the probability for one  $\alpha$ -observation for the bivariate normal random vector. The probability can then be used to calculate the window probability and the optimal window size according to Algorithm 2.1.

To adjust Algorithm 2.1 for use in the multivariate case we will need to update step 2 to use Corollary 3.1 with  $\Psi$  and estimate  $\Sigma_1$  to use for the calculation of the Mahalanobis distance. For Algorithm 2.2, the construct of the sequence  $z_{(1,k)}$  from  $\mathbf{y}_{(1,N)}$  is done as

$$z_i = \begin{cases} 1, & \text{if } \mathbf{y}'_i \Sigma_1^{-1} \mathbf{y}_i > c \\ 0, & \text{otherwise,} \end{cases}$$



**Figure 3.3:** Probability for a 5%-observation from a bivariate normal distribution when the  $\rho$  parameter changes between  $-0.99$  and  $0.99$ , see Example 3.4.

where  $c = \chi_{\alpha}^2(p)$ . Those changes make the algorithm possible to use for our proposed method in the multivariate case. In Section 3.2.4 we are making a simulation study comprising the multivariate case.

### 3.2.2 Eigenvalues of sample covariance matrix

Eigenvalues of a matrix are often sensitive to changes in the matrix. Therefore looking for changes in the eigenvalues of the sample covariance matrix can be a way of identifying a change. If the largest eigenvalue has changed more than we expect then we can assume that there has been a change in the covariance matrix. Denote the eigenvalues of  $\Sigma$  as  $\lambda_i(\Sigma)$ ,  $i = 1, \dots, p$ , where for simplicity it is assumed that  $\lambda_1(\Sigma) > \dots > \lambda_p(\Sigma)$ . If it is obvious which matrix then we will just write  $\lambda_i$ . The hypothesis for testing, if the largest eigenvalue has changed from  $\lambda$ , can be formulated as

$$H_0 : \lambda_1(\Sigma) = \lambda \text{ versus } H_1 : \lambda_1(\Sigma) \neq \lambda. \tag{3.6}$$

The exact distribution of the largest eigenvalue is given in Muirhead (1982) Theorem 9.7.1. It is stated that the distribution function of the largest eigenvalue,  $\lambda_1(k\mathbf{S})$ , the sample covariance matrix,  $\mathbf{S}$  of  $\Sigma$  times the number of observations  $k$ ,

equals

$$F_{\lambda_1(k\mathbf{S})}(x) = P(\lambda_1(k\mathbf{S}) < x) = \frac{\Gamma_p[\frac{1}{2}(p+1)]}{\Gamma_p[\frac{1}{2}(k+p+1)]} \left| \frac{1}{2}kx\boldsymbol{\Sigma}^{-1} \right|^{n/2} \times {}_1F_1\left(\frac{1}{2}k; \frac{1}{2}(k+p+1); -\frac{1}{2}kx\boldsymbol{\Sigma}^{-1}\right), \quad (3.7)$$

where  ${}_1F_1$  is a hypergeometric function with matrix argument. For the definition of  ${}_1F_1$  see (Muirhead, 1982, Chapter 7.3).

### Theorem 3.5

Let  $\mathbf{y}_{(1,k)}$  be a SIO from  $\mathbf{X}_2 \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_2)$  where  $\mathbf{S}_2$  is the sample covariance matrix of  $\boldsymbol{\Sigma}_2$ , and  $\lambda_1(k\mathbf{S}_2)$  the largest eigenvalue of  $k\mathbf{S}_2$ . Furthermore, let  $\boldsymbol{\Sigma}_1$  be a known covariance matrix with a sample covariance matrix  $\mathbf{S}_1$ .

Then the test of a change with significance level  $\alpha_{test}$  of the null hypothesis  $H_0 : \lambda_1(\boldsymbol{\Sigma}_1) = \lambda_1(\boldsymbol{\Sigma}_2)$ , rejects  $H_0$ , if  $\lambda(k\mathbf{S}_2) > c$ , where  $F_{\lambda_1(k\mathbf{S})}(c) > 1 - \alpha_{test}$ .

To use Theorem 3.5 we need to know the covariance matrix before the change  $\boldsymbol{\Sigma}_1$ . To avoid this problem we could use Roy's largest root test, which is presented in Muirhead (1982) Theorem 10.6.8. The theorem gives the distribution function of the largest eigenvalue of  $\frac{1}{k}\mathbf{S}_1^{-1}k\mathbf{S}_2$ , where  $\mathbf{S}_1$  and  $\mathbf{S}_2$  are the sample covariance matrices of  $\boldsymbol{\Sigma}_1$  and  $\boldsymbol{\Sigma}_2$ . We have that  $k\mathbf{S}_1$  and  $k\mathbf{S}_2$  are independent and central Wishart distributed with  $k$  degrees of freedom, respectively. Since they are central Wishart distributed we can use Corollary 10.6.9 in Muirhead (1982). There it states that the distribution of the largest eigenvalue of  $\mathbf{S}_1^{-1}\mathbf{S}_2$  is

$$F_{\lambda_1(\mathbf{S}_1^{-1}\mathbf{S}_2)}^{Roy}(x) = P(\lambda_1(\mathbf{S}_1^{-1}\mathbf{S}_2) \leq x) = \left( \frac{x}{1+x} \right)^{pk/2} \sum_{r=0}^{pt} \sum_{\kappa}^* \frac{\left(\frac{1}{2}k\right)_{\kappa} C_{\kappa}(\mathcal{I})}{(1+x)^r r!}, \quad (3.8)$$

where  $\kappa$  is a partition of  $(r_1, \dots, r_p)$ ,  $r_1 \geq \dots \geq r_p \geq 0$  of  $r = \sum_{i=1}^p r_i$  with  $\sum_{\kappa}^*$  denotes the sum where  $r_1 < t = (k-p-1)/2$ , and  $C_{\kappa}(\mathcal{I})$  is the zonal polynomial of the identity matrix, see (Muirhead, 1982, Chapter 7.2), and  $(r)_{\kappa}$  is the generalized hypergeometric coefficient, see (Muirhead, 1982, p. 248). Under the null hypothesis, the largest eigenvalue of  $\mathbf{S}_1^{-1}\mathbf{S}_2$  should be close to 1. By examining the characteristic polynomial of  $\mathbf{S}_1^{-1}\mathbf{S}_2$  it is possible to see that it is the same as for  $\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}$ , which gives us that  $\mathbf{S}_1^{-1}\mathbf{S}_2$  and  $\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}$  has the same eigenvalues.

Now we are ready to formulate a change point theorem with the help of Roy's largest eigenvalue criteria as follows:

### Theorem 3.6

Let  $\mathbf{x}_{(1,k)}$  be a SIO from  $\mathbf{X}_1 \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_1)$  and let  $\mathbf{y}_{(1,k)}$  be a SIO from  $\mathbf{X}_2 \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_2)$ . Furthermore,  $\boldsymbol{\Sigma}_1$  and  $\boldsymbol{\Sigma}_2$  are unknown and estimated with the sample covariance matrices  $\mathbf{S}_1$  and  $\mathbf{S}_2$ , respectively.

Then the test of the null hypothesis  $H_0 : \lambda_1(\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}) = 1$ , rejects  $H_0$  with significance level  $\alpha_{test}$ , if  $\lambda_1(\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}) > c$ , where  $F_{\lambda_1(\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2})}^{Roy}(c) > 1 - \alpha_{test}$ .

With Theorem 3.6 we do not need to know how the covariance matrix looks before the change to have something to compare with. Still, we need to calculate the largest eigenvalue of  $\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}$ , or  $\mathbf{S}_1^{-1}\mathbf{S}_2$ . Then with the help of (3.8) it is possible to find a threshold,  $c$ , to evaluate the largest eigenvalue of  $\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}$  for a significant change in the covariance matrix.

We use the test statistic from Theorem 3.6 to create the alert function

$$\delta(\mathbf{S}_1, \mathbf{S}_2) = \begin{cases} 1, & \text{if } \lambda_1(\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}) > c \\ 0, & \text{otherwise.} \end{cases} \quad (3.9)$$

To start a search algorithm we need to decide what is a reasonable window size  $k$ , calculate the threshold  $c$ , and create an estimate of  $\Sigma_1^{-1}$  from  $\mathbf{x}_{(1,n_0)}$ . The threshold can be calculated from (3.8) as discussed in Theorem 3.6. We can now modify Algorithm 2.2 to Algorithm 3.1.

---

### Algorithm 3.1

---

**Required:** thresholds  $c$ ; sample covariance matrix  $\mathbf{S}_1$ ; window size  $k$

**Data:**  $y_{(1,N)}$  an unknown SIO

**Steps:**

- 1) **set**  $i = 1$
  - 2) **estimate**  $\mathbf{S}_2$  from  $\mathbf{y}_{(i,i+k-1)}$
  - 3) **if**  $\delta(\mathbf{S}_1, \mathbf{S}_2) = 1$  **then**
    - **go to end**
  - else if**  $\delta(\mathbf{S}_1, \mathbf{S}_2) = 0$  **then**
    - **set**  $i = i + k$
    - **go to step 2**
  - 4) **end**
- 

As seen in Algorithm 3.1 we need the window size to run the algorithm. As for the method with  $\alpha$ -observations, this is a trade-off between how precise the method can be in detecting when a change occurred and the use of as few observations as possible after a change.

### 3.2.3 Divergence information measure

A metric on a set  $\mathcal{M}$  is a function  $D : \mathcal{M} \times \mathcal{M} \rightarrow \mathbf{R}^+$  that for  $x, y, z \in \mathcal{M}$  fulfills the following four criteria

- (i)  $D(x, y) > 0$  if  $x \neq y$ ,
- (ii)  $D(x, y) = 0$  if and only if  $x = y$ ,
- (iii)  $D(x, y) = D(y, x)$ ,

$$(iv) D(x, z) \leq D(x, y) + D(y, z).$$

A divergence is a generalization of the metric where only (i) and (ii) are fulfilled.

A short summary of algorithms using information measures in change point detection algorithms can be found in Aminikhanghahi and Cook (2017). The algorithms described are KLIEP (Kullback-Leibler Importance Estimation Procedure), uLSIF (unconstrained Least-Squares Importance Fitting), RuLSIF (Relative uLSIF), and SPLL (Semi-Parametric Log-Likelihood change detector) and in Aminikhanghahi et al. (2019) they describe SEP (Separation distance). In general, all these algorithms use a kernel-based estimate of the ratio of the density of the two windows. The estimate is then used to calculate the different divergences to decide if there has been a change or not. The algorithm KLIEP uses Kullback-Leibler, uLSIF, and RuLSIF using the Pearson divergence, SEP uses the separation distance and SPLL uses Kullback-Leibler as a base with a few extra assumptions.

We will look closer at the Kullback-Leibler divergence and the Bhattacharyya distance, see Table 3.2 for these definitions of the divergences. The hypothesis to

**Table 3.2:** Divergence definitions

Name	Definition
Kullback-Leibler divergence	$\int p(x) \ln \left( \frac{p(x)}{q(x)} \right) dx$
Bhattacharyya distance	$-\ln \left( \int \sqrt{p(x)q(x)} dx \right)$

check if there has been a change in the covariance matrix is

$$H_0 : \Sigma_1 = \Sigma_2 \quad \text{versus} \quad H_1 : \Sigma_1 \neq \Sigma_2. \quad (3.10)$$

The Kullback-Leibler divergence was introduced by Kullback and Leibler (1951), also called relative entropy, and used to see if there is a difference between two distributions. The following theorem shows how Kullback-Leibler divergence can be used to investigate if there has been a change in a window compared to a previous window.

**Theorem 3.7**

Let  $\mathbf{x}_{(1,k)}$  be a SIO from  $\mathbf{X}_1 \sim \mathcal{N}_p(\mathbf{0}, \Sigma_1)$ , and let  $\mathbf{y}_{(1,k)}$  be a SIO from  $\mathbf{X}_2 \sim \mathcal{N}_p(\mathbf{0}, \Sigma_2)$ . Then the Kullback-Leibler test statistic is given by

$$T_{KL}(\mathbf{x}_{(1,k)}, \mathbf{y}_{(1,k)}) = \frac{1}{2} \left( \ln |\mathbf{S}_1^{-1/2} \mathbf{S}_2 \mathbf{S}_1^{-1/2}| + \text{tr}((\mathbf{S}_1^{-1/2} \mathbf{S}_2 \mathbf{S}_1^{-1/2})^{-1}) - p \right) \quad (3.11)$$

Then the test of the null hypothesis in (3.10) rejects  $H_0$  with significance level  $\alpha_{test}$ , if  $T_{KL}(\mathbf{x}_{(1,k)}, \mathbf{y}_{(1,k)}) > c_{\alpha_{test}}$ , where  $c_{\alpha_{test}}$  is chosen so that the size of the test is  $\alpha_{test}$ .

**Proof:** Let  $\mathbf{X}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \Sigma_1)$  and  $\mathbf{X}_2 \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \Sigma_2)$  be two multivariate normal distributions. Then the Kullback-Leibler divergence measure between the distri-

butions, given in Table 3.2, has the following form

$$\begin{aligned} D_{KL}(\mathbf{X}_1, \mathbf{X}_2) &= \frac{1}{2} \left( \ln \left( \frac{|\Sigma_2|}{|\Sigma_1|} \right) - p + \text{tr}(\Sigma_1 \Sigma_2^{-1}) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)' \Sigma_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \right) \\ &= \frac{1}{2} \left( \ln |\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}| - p + \text{tr}((\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2})^{-1}) \right) \\ &\quad + \frac{1}{2} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)' \Sigma_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \end{aligned}$$

With our assumptions that  $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$  and replacing  $\Sigma_1$  and  $\Sigma_2$  with the sample covariance matrices  $k\mathbf{S}_1 = \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i'$  and  $k\mathbf{S}_2 = \sum_{i=1}^k \mathbf{y}_i \mathbf{y}_i'$ , respectively, the result follows.  $\square$

To be able to use the test statistic  $T_{KL}$  we need to find the threshold  $c_{\alpha_{test}}$ , which we have not succeeded in. The preferred approach involves deriving the distribution and subsequently calculating the threshold from it. Alternatively, it would also be possible to do a Monte Carlo simulation of the change in question. Even with a distribution of  $T_{KL}$ , it can be difficult to calculate the threshold, as it is for Roy's largest eigenvalue criteria.

In order to control the test's power, we must determine how the change influences both the threshold and window size. In the context of our proposed algorithm, controlling it is achieved through the manipulation of the window size. However, at this point, we do not possess a method to exercise such control and need to decide the window size in some other way.

With this in mind, we construct an alert function with the test statistic from Theorem 3.7 as follows

$$\delta(\mathbf{S}_1, \mathbf{S}_2) = \begin{cases} 1, & \text{if } T_{KL}(\mathbf{x}_{(1,n_0)}, \mathbf{y}_{(1,k)}) > c_{\alpha_{test}}, \\ 0, & \text{otherwise.} \end{cases} \quad (3.12)$$

The alert function (3.12) can then be used in Algorithm 3.1.

The Bhattacharyya distance was introduced by Bhattacharyya (1943) and has been used among other things in feature extraction and selection, see Reyes-Aldasoro and Bhalerao (2006). As for the Kullback-Leibler divergence, we formulate a theorem on how the Bhattacharyya distance can be used to identify a change.

### Theorem 3.8

Let  $\mathbf{x}_{(1,k)}$  be a SIO from  $\mathbf{X}_1 \sim \mathcal{N}_p(\mathbf{0}, \Sigma_1)$ , and let  $\mathbf{y}_{(1,k)}$  be a SIO from  $\mathbf{X}_2 \sim \mathcal{N}_p(\mathbf{0}, \Sigma_2)$ . Then the Bhattacharyya test statistic is given by

$$T_B(\mathbf{x}_{(1,k)}, \mathbf{y}_{(1,k)}) = \frac{1}{2} \ln \left( \frac{\left| \frac{\boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1/2}}{2} \right|}{\sqrt{|\boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1/2}|}} \right). \quad (3.13)$$

Then the test of the null hypothesis in (3.10) rejects  $H_0$  with significance level  $\alpha_{test}$ , if  $T_B(\mathbf{x}_{(1,k)}, \mathbf{y}_{(1,k)}) > c_{\alpha_{test}}$ , where  $c_{\alpha_{test}}$  is chosen so that the size of the test is  $\alpha_{test}$ .

**Proof:** Let  $\mathbf{X}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  and  $\mathbf{X}_2 \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$  be two multivariate normal distributions. Then the Bhattacharyya distance between them, given in Table 3.2, equals

$$\begin{aligned} D_B(\mathbf{X}_1, \mathbf{X}_2) &= \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \left( \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \left( \frac{|\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}|}{\sqrt{|\boldsymbol{\Sigma}_1| |\boldsymbol{\Sigma}_2|}} \right) \\ &= \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \left( \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \left( \frac{\left| \frac{\boldsymbol{\mathcal{I}} + \boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1/2}}{2} \right|}{\sqrt{|\boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1/2}|}} \right) \end{aligned}$$

With our assumptions that  $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$  and replacing  $\boldsymbol{\Sigma}_1$  and  $\boldsymbol{\Sigma}_2$  the sample covariance matrices  $k\mathbf{S}_1 = \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i'$  and  $k\mathbf{S}_2 = \sum_{i=1}^k \mathbf{y}_i \mathbf{y}_i'$ , respectively, the result follows.  $\square$

In (3.13) we see that for the Bhattacharyya test statistic as for the Kullback-Leibler test statistic, only depends on the estimate of  $\boldsymbol{\Psi}$  that is  $\boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1/2}$ . For  $T_B$ , it evaluates the change matrix with the determinant and an adjusted determinant of the change matrix, while  $T_{KL}$  evaluates it with the log of the determinant and the trace of its inverse. Hence, both are using similar functions depending on the eigenvalue of the change matrix so it is expected that they react in a similar manner to change.

To use the Bhattacharyya distance in an algorithm, we have the same issue as for the Kullback-Leibler divergence. The threshold  $c_{\alpha_{test}}$  needs to be derived and a way to decide the window size. With that in mind, the alert function for the Bhattacharyya distance is constructed as

$$\delta(\mathbf{S}_1, \mathbf{S}_2) = \begin{cases} 1, & \text{if } T_B(\mathbf{x}_{(1, n_0)}, \mathbf{y}_{(1, k)}) > c_{\alpha_{test}}, \\ 0, & \text{otherwise.} \end{cases} \quad (3.14)$$

The alert function (3.14) can then be used in Algorithm 3.1.

### 3.2.4 Simulation and comparison

We will start by making a simulation study for testing for change points using the Mahalanobis distance on a SIO containing observations from two multivariate normal distributions with different covariance matrices. We will also compare the use of Mahalanobis distance with Roy's largest eigenvalue criteria, Kullback-Leibler divergence, and Bhattacharyya distance. This comparison will give an indication of how well the four methods can perform. Since we only have a threshold for the method using Mahalanobis distance we cannot quantify how well the others performed. There will just be an indication of how they could perform.

#### Setup

The test is set up to examine changes from three different covariance matrices, the identity matrix  $\boldsymbol{\mathcal{I}}$ , a random matrix  $\mathbf{M}$ , and a random matrix,  $\mathbf{A}$ , where the trace

is the same as for  $\mathcal{I}$ . The covariance matrices are then changed to a multiple of two. The matrices are constructed as

- i)  $\mathbf{M} = \widetilde{\mathbf{M}}' \widetilde{\mathbf{M}}$ , where entries of  $\widetilde{\mathbf{M}}$  are realisations of independent random variables distributed as  $\mathcal{N}(0, 1)$ ;
- ii)  $\mathbf{A} = \frac{p}{\text{tr}(\mathbf{A}^*)} \mathbf{A}^*$  where  $\mathbf{A}^* = \widetilde{\mathbf{A}}' \widetilde{\mathbf{A}}$  and entries of  $\widetilde{\mathbf{A}}$  are realisations of random variables distributed as  $\mathcal{N}(0, 1)$ .

The window size is calculated for the Mahalanobis method, see Algorithm 2.1. The parameters utilized are as follows:  $\alpha = \alpha_{FWER} = 1 - \beta = 0.05$ , and assume  $\Psi = 2\mathcal{I}$ . Additionally, an initial sequence  $\mathbf{x}_{(1, n_0)}$  consisting of either  $n_0 = 100$  or  $n_0 = 5000$  observations is employed to estimate  $\Sigma_1$ . The size of the sequences that are tested is configured to contain at least 10 windows and have at least 200 observations. The change point is selected to be in the middle of the SIO and we set the expected maximum observations before a change,  $N_{max}$  to contain the first half, the part without any change. In other words, we are expecting about  $\alpha_{FWER} N_{max}$  number of SIOs with a type I error. To calculate the probability for an  $\alpha$ -observation with the assumed change  $\Psi = 2\mathcal{I}$ , the gamma parameters  $a$  and  $b$  are needed and according to Theorem 3.4

$$a = \frac{(\text{tr}(\Psi))^2}{2\text{tr}(\Psi^2)} = \frac{p}{2}, \quad b = \frac{\text{tr}(\Psi)}{2\text{tr}(\Psi^2)} = \frac{1}{4}.$$

With all these choices and conditions the setup of the SIOs are

- $p = 2$  and  $\mathcal{N}_{max} = 250$  gives  $k = 56$  and  $N = 560$  with 10 windows and a threshold of  $m_k = 7$ ,
- $p = 10$  and  $\mathcal{N}_{max} = 100$  gives  $k = 13$  and  $N = 208$  with 16 windows and a threshold of  $m_k = 3$ .

The simulation of the method when using the Mahalanobis distance is repeated 10,000 times for both  $p = 2$  and  $p = 10$ .

### Simulation results

The results for the simulation when  $p = 2$  can be seen in Tables 3.4 and 3.5. The first column in Table 3.4 shows the amount of the type I error of the test, i.e., the number of times when there was an alert for a change before it had happened. The second column counts the number of SIOs where the alert came after a change had occurred. The third column contains the total number of type I error for the test of each window. The fourth column is the number of windows that were successfully identified to differ from the estimate done from  $\mathbf{x}_{(1, n_0)}$ . Table 3.5 shows in which window the alert of the change occurred. There is only one alert per SIO.

Let us first notice that we cannot see any difference between the different starting matrices,  $\mathcal{I}$ ,  $\mathbf{M}$  and  $\mathbf{A}$ , in the results given in Table 3.4, 3.5, 3.6 and 3.7. In other words, the Mahalanobis method seems to be equally good at noticing the change of  $\Psi = 2\mathcal{I}$  independently of the structure of  $\Sigma_1$ .

To analyze the result we need to point out that the significance level for a window has the upper limit  $\alpha_W(k)$  defined in (2.2). The actual significance level is calculated with the help of the window threshold as  $\tilde{\alpha}_W(k) = P(W > m_k)$  where  $W \sim Bin(k, \alpha)$ . For our setup with  $p = 2$ , we have that  $\alpha_W(k) = 0.0112$  and  $m_k = 7$ , hence  $\tilde{\alpha}_W(k) \approx 0.0065$  is our expected type I error for the window tests. When the estimate of  $\Sigma_1$  is done with  $n_0 = 100$  observations, the result is shown in the upper part of Table 3.4, we have a type I error for the windows of about  $1300/(10000 * 5) \approx 0.026$  which is larger than 0.0065. The "type I error" for the SIO is about  $1100/10000 = 0.11$  which is far above our expectation of  $\alpha_{FWER} = 0.05$ . This information is summarized in Table 3.3.

**Table 3.3:** The "type I error (window)" should be compared with the significance level of the window,  $\tilde{\alpha}_W$ . The "type I error (SIO)" should be compared with  $\alpha_{FWER}(N/2)/N_{max} \approx 0.05$ .

Analysis of type I error								
$p$	$n_0$	$N_{max}$	$k$	$\alpha_W$	$m_k$	$\tilde{\alpha}_W$	Type I error Window	Type I error SIO
2	100	250	56	0.0112	7	0.0065	0.026	0.11
2	5000	250	56	0.0112	7	0.0065	0.0064	0.032
10	100	100	13	0.0065	3	0.0031	0.040	0.26
10	5000	100	13	0.0065	3	0.0031	0.0034	0.027

The window size is selected such that we should have found the change within one window, with a probability of  $\beta$ . For each window with observations from a random vector with covariance matrix  $\Sigma_2 \neq \Sigma_1$ . By looking in Table 3.5 with  $n_0 = 100$  we can see that out of the 10,000 SIOs about 8,900 did not have a type I error and could be used for calculating an estimate of the power. In 8,300 of them, the change was found in the first window. This gives us an estimation of the power to be about 0.94. The performance could also be analyzed by examining the number of windows with a different covariance matrix where the change was correctly identified. Then from Table 3.4 we have a success rate of identifying the change of about  $47,000/50,000 = 0.94$ . For the test with  $n_0 = 5000$ , the power for the full SIO is about 0.95, and for the windows about the same. With this said the power is almost as predicted.

For  $p = 10$ , the corresponding results of the simulation are presented in Tables 3.6, and 3.7, and the analysis of type I error can be seen in Table 3.3. Regarding the type I error when considering  $n_0 = 100$ , we observe that the type I error for the windows differs from the expected value with nearly a multiple of 10, while for the SIO, it's approximately 5 times higher. A closer examination of the power for the window demonstrates a high value of around 0.99. Consequently, it is apparent that the method tends to generate alerts too hastily. The situation improves when we increase  $n_0$  to 5000, resulting in a more accurate estimation of  $\Sigma_1$ . Subsequently, the type I error for the window aligns closely with the expected level, and for the SIO, it even drops below. Additionally, the power converges toward the expected level but still remains slightly elevated.

**Table 3.4:** Dimensions 2 - Out of 10,000 SIOs. The first two columns show the number of type I errors for the test of finding a change in a SIO and the count of successfully identified changes in a SIO. The last two columns contain the number of type I errors of the test for each window and how many windows following the change were successfully identified.

Estimation of $\Sigma_1$ with $n_0 = 100$ observations				
Change	SIO		Window	
	Type I error	Success	Type I error	Success (Power)
$\mathcal{I}$ to $2\mathcal{I}$	1074	8925	1283	47046 (0.941)
$\mathcal{M}$ to $2\mathcal{M}$	1113	8884	1357	47165 (0.943)
$\mathcal{A}$ to $2\mathcal{A}$	1127	8869	1367	47025 (0.941)

Estimation of $\Sigma_1$ with $n_0 = 5000$ observations				
Change	SIO		Window	
	Type I error	Success	Type I error	Success (Power)
$\mathcal{I}$ to $2\mathcal{I}$	314	9686	324	47555 (0.951)
$\mathcal{M}$ to $2\mathcal{M}$	323	9673	330	47617 (0.952)
$\mathcal{A}$ to $2\mathcal{A}$	310	9687	313	47635 (0.953)

## Comparison

We will use the same setup to compare the different methods as was used for the simulation of the Mahalanobis method. A sample of the results for  $p = 2$  is shown in Figure 3.4 and for  $p = 10$  it is presented in Figure 3.5. Each of the four methods examines the same SIO utilizing the same window size. In the case of  $p = 2$ , the window size is  $k = 56$ , while for  $p = 10$ , the window size is  $k = 13$ . The window sizes were selected from the process of the Mahalanobis method to use as few observations as possible after the change. The estimate of  $\Sigma_1$  was calculated from a SIO with  $n_0 = 100$  observations.

Figures 3.4 and 3.5 contain three different initial matrices where the change occurred at the sixth window for  $p = 2$  in Figure 3.4 and at the ninth window for  $p = 10$  in Figure 3.5. In the plot for the Mahalanobis method, we have the threshold printed, and for the Kullback-Leibler divergence and the Bhattacharyya distance, we have the exact value if the second window contains a change  $\Sigma_2 = 2\Sigma_1$ . The value for the Kullback-Leibler divergence between  $\mathbf{X}_1 \sim \mathcal{N}_p(\mathbf{0}, \Sigma_1)$  and  $\mathbf{X}_2 \sim \mathcal{N}_p(\mathbf{0}, \Sigma_2)$ , where  $\Sigma_2 = 2\Sigma_1$ , is calculated as

$$\begin{aligned}
 D_{KL}(\mathbf{X}_1, \mathbf{X}_2 \mid \Sigma_2 = 2\Sigma_1) &= \frac{1}{2} (\ln |\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}| + \text{tr}((\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2})^{-1}) - p) \\
 &= \frac{p}{2} \left( \ln 2 - \frac{1}{2} \right),
 \end{aligned}$$

**Table 3.5:** Dimensions 2 - Each of the 10,000 SIO is counted towards the window number where the threshold is first exceeded. The actual change in the SIOs happened in the sixth window.

Estimation of $\Sigma_1$ with $n_0 = 100$ observations						
Window number						
Change	1 - 5	6	7	8	9	10
$\mathcal{I}$ to $2\mathcal{I}$	1074	8358	488	62	14	3
$\mathcal{M}$ to $2\mathcal{M}$	1113	8355	446	68	13	2
$\mathcal{A}$ to $2\mathcal{A}$	1127	8310	479	68	7	5

Estimation of $\Sigma_1$ with $n_0 = 5000$ observations						
Window number						
Change	1 - 5	6	7	8	9	10
$\mathcal{I}$ to $2\mathcal{I}$	314	9203	464	17	2	0
$\mathcal{M}$ to $2\mathcal{M}$	323	9222	430	21	0	0
$\mathcal{A}$ to $2\mathcal{A}$	310	9205	463	17	2	0

and the Bhattacharyya distance as

$$\begin{aligned}
 D_B(\mathbf{X}_1, \mathbf{X}_2 \mid \Sigma_2 = 2\Sigma_1) &= \frac{1}{2} \left( \ln \left( \frac{\left| \frac{\mathcal{I} + \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}}{2} \right|}{\sqrt{|\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}|}} \right) \right) \\
 &= \frac{p}{2} \left( \ln \frac{3}{2} - \frac{1}{2} \ln 2 \right).
 \end{aligned}$$

This gives the following values for the different dimensions

$$\begin{aligned}
 p = 2 &\Rightarrow \begin{cases} D_{KL}(\mathbf{X}_1, \mathbf{X}_2 \mid \Sigma_2 = 2\Sigma_1) \approx 0.193, \\ D_B(\mathbf{X}_1, \mathbf{X}_2 \mid \Sigma_2 = 2\Sigma_1) \approx 0.059, \end{cases} \\
 p = 10 &\Rightarrow \begin{cases} D_{KL}(\mathbf{X}_1, \mathbf{X}_2 \mid \Sigma_2 = 2\Sigma_1) \approx 0.966, \\ D_B(\mathbf{X}_1, \mathbf{X}_2 \mid \Sigma_2 = 2\Sigma_1) \approx 0.29. \end{cases}
 \end{aligned}$$

In Figure 3.4 we can see that two of the three simulations for the Mahalanobis method would be counted as a success with the first alert at the first window with a change. The red sample gave a type I error with the alert at the fifth window. The increase at the sixth window can also be seen in the other methods. However, to be able to say how many examined SIOs were a success or not we need to find the threshold  $c_{\alpha_{test}}$  for Theorem 3.7 and Theorem 3.8. The exact value for the change marked in the plot gives a hint that at least the value for the windows after the change seems to have the right size, but it doesn't give how likely values of that size are. For Roy's largest eigenvalue criteria, it should be possible to use Theorem 3.6 to calculate the p-value of the test. However, since the calculation is quite demanding we have not managed to do an implementation

**Table 3.6:** Dimensions 10 - Out of 10,000 SIOs. The first two columns show the number of type I errors for the test of finding a change in a SIO and the count of successfully identified changes in a SIO. The last two columns contain the number of type I errors of the test for each window and how many windows following the change were successfully identified.

Estimation of $\Sigma_1$ with $n_0 = 100$ observations				
Change	SIO		Window	
	Type I error	Success	Type I error	Success(Power)
$\mathcal{I}$ to $2\mathcal{I}$	2652	7348	3244	79164 (0.990)
$\mathcal{M}$ to $2\mathcal{M}$	2541	7459	3151	79173 (0.990)
$\mathcal{A}$ to $2\mathcal{A}$	2605	7395	3182	79205 (0.990)
Estimation of $\Sigma_1$ with $n_0 = 5000$ observations				
Change	SIO		Window	
	Type I error	Success	Type I error	Success(Power)
$\mathcal{I}$ to $2\mathcal{I}$	270	9730	273	77324 (0.967)
$\mathcal{M}$ to $2\mathcal{M}$	251	9749	252	77263 (0.966)
$\mathcal{A}$ to $2\mathcal{A}$	281	9719	287	77146 (0.964)

of it, and the expected value of 1 before and 2 after is not correct since it varies with the dimension.

In Figure 3.5 we see three SIO subjected to the four methods. The method using Mahalanobis distance managed to successfully identify the change in the first window after the change for all three. It is also shown a noticeable difference of the largest eigenvalue of  $\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}$  in the ninth window. For the Kullback-Leibler divergence and the Bhattacharyya distance, we cannot see any difference before or after the change. Both of them give for the whole SIO values that are above their expected value,  $D_{KL}(\mathbf{X}_1, \mathbf{X}_2 \mid \Sigma_2 = 2\Sigma_1)$  and  $D_{KL}(\mathbf{X}_1, \mathbf{X}_2 \mid \Sigma_2 = 2\Sigma_1)$ , respectively. This is not an unexpected event since with only 13 observations to estimate a covariance matrix of 10 dimensions there is a large risk that the estimate would be ill-conditioned. The determinant is sensitive to ill-conditioned matrices which could give large numbers.

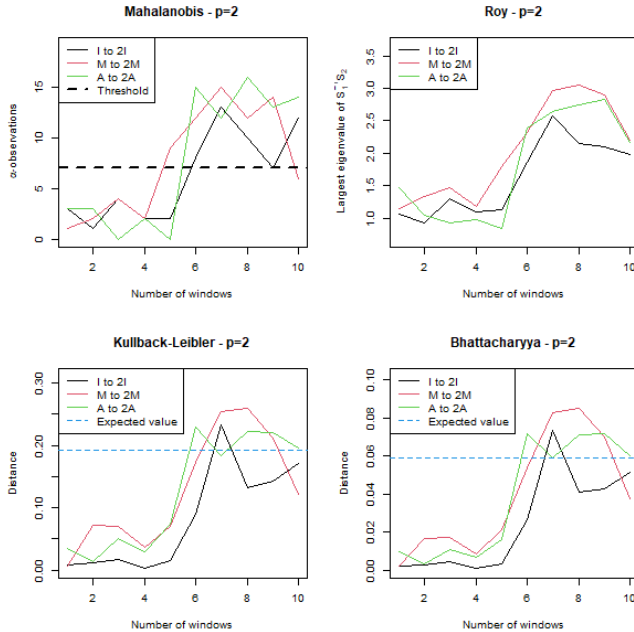
The shape of the graphs in Figure 3.4 for the Kullbak-Leibler divergence and the Bhattacharyya distance is very similar. In general, we can only see some small differences between them. Even in Figure 3.5 there are a lot of similarities, but the Kullback-Leibler divergence seems to give a more unstable result.

**Table 3.7:** Dimensions 10 - Each of the 10,000 SIO is counted towards the window number where the threshold is first exceeded. The actual change in the SIOs happened in the ninth window.

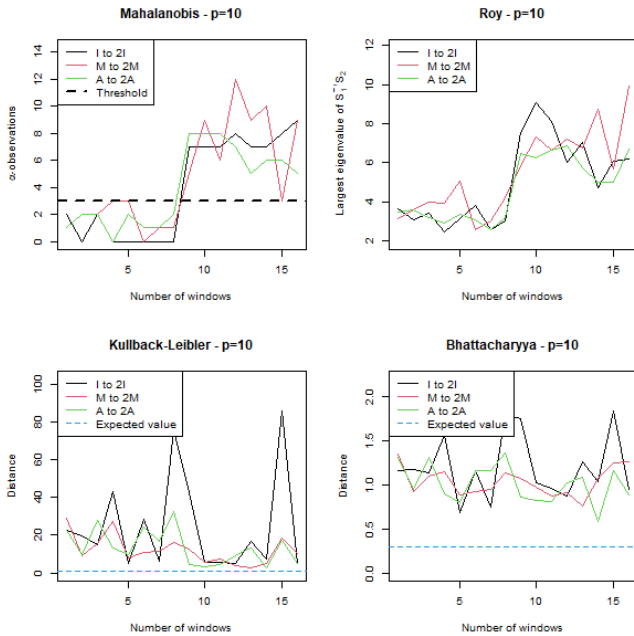
Estimation of $\Sigma_1$ with $n_0 = 100$ observations					
Window number					
Change	1 - 8	9	10	11	12
$\mathcal{I}$ to $2\mathcal{I}$	2652	7263	84	1	0
$M$ to $2M$	2541	7375	83	1	0
$A$ to $2A$	2605	7322	71	2	0

Estimation of $\Sigma_1$ with $n_0 = 5000$ observations					
Window number					
Change	1 - 8	9	10	11	12
$\mathcal{I}$ to $2\mathcal{I}$	270	9402	319	9	0
$M$ to $2M$	251	9417	321	11	0
$A$ to $2A$	281	9377	326	16	0



**Figure 3.4:** Calculation of four different methods on three different SIOs from  $\mathcal{N}_2(\mathbf{0}, \Sigma)$ . The change  $\Sigma = \Sigma_1$  to  $\Sigma = 2\Sigma_1$  occurs in the sixth window. The expected values  $D_{KL}(\mathbf{X}_1, \mathbf{X}_2 | \Sigma_2 = 2\Sigma_1)$  and  $D_{KL}(\mathbf{X}_1, \mathbf{X}_2 | \Sigma_2 = 2\Sigma_1)$  are presented as the blue dashed line.



**Figure 3.5:** Calculation of four different methods on a SIO from  $\mathcal{N}_{10}(\mathbf{0}, \Sigma)$ . The change from  $\Sigma = \Sigma_1$  to  $\Sigma = 2\Sigma_1$  occurs in the ninth window. The expected values  $D_{KL}(\mathbf{X}_1, \mathbf{X}_2 | \Sigma_2 = 2\Sigma_1)$  and  $D_{KL}(\mathbf{X}_1, \mathbf{X}_2 | \Sigma_2 = 2\Sigma_1)$  are presented as the blue dashed line.



# 4

---

## Concluding remarks

In this thesis, the main focus has been on constructing a simple algorithm to find a change in variance in a sequence of observations from independent normally distributed random variables. We also had a goal to see if it was possible to extend the method to the covariance matrix for multivariate observations.

### 4.1 Summary

In Chapter 2 we construct an algorithm to detect a change in variance. This is done by considering windows of observations in a sequence of independent observations (SIO) and counting the number of  $\alpha$ -observation that occur. In Section 3.1 we derived the expression of the probability of an  $\alpha$ -observation to occur depending on  $\psi$ , the relative change in variance. With the probability of the  $\alpha$ -observation in Theorem 3.1 and by treating the number of  $\alpha$ -observations in a window as a binomial random variable, it is possible to derive how many  $\alpha$ -observations are expected for different window sizes, which gave us the window threshold, see Definition 2.8. Furthermore, by specifying the power of the test  $\beta$ , we could derive the window size that uses the fewest number of observations after the change that fulfills our criteria, see Theorem 2.1.

In Section 3.2.1 we extend the algorithm to independent multivariate observations with a covariance matrix  $\Sigma$  and zero mean. Using the squared Mahalanobis distance to transform all observations into univariate values one by one, which under the null hypothesis will be  $\chi^2(p)$ -distributed. Hence, we have under the null hypothesis a sequence from a  $\chi^2(p)$ -distributed random variables, where we can find a threshold for the  $\alpha$ -observations.

To be able to calculate the power of the test, we derived an approximated distribution of the probability for the  $\alpha$ -observation. This was done by noticing that the Mahalanobis distance of an observation is a quadratic form which

has the same distribution as a weighted sum of  $\chi^2$ -variables. The weighted sum can be approximated to a gamma variable with parameters that only depends on  $\Psi = \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}$ . From here the probability for an  $\alpha$ -observation is straightforwardly calculated with the help of Corollary 3.1. Now we can use the same framework as for the univariate case to control the type I and type II errors, which is carried out by adjusting the threshold and the window size, respectively.

Moreover, some alternative approaches have been studied, i.e.,

- the largest eigenvalue of the sample covariance matrix through Roy's largest eigenvalue criteria,
- divergence information measure, such as
  - the Kullback-Leibler divergence,
  - the Bhattacharyya distance.

All of these methods use the sample covariance matrix from two windows to test if the observations from the different windows come from different distributions. Hence, the hypothesis for a change in the sequence for the multivariate case is formulated as

$$H_0 : \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} = \mathcal{I} \quad \text{versus} \quad H_1 : \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} \neq \mathcal{I}.$$

We have seen that all the methods are depending on the eigenvalues of the change matrix  $\Psi = \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}$ .

The simulations done in Section 3.2.4 show that the Mahalanobis method only works as expected if there is a very good estimate of  $\Sigma_1$ . For the alternative methods, we could not evaluate them since we did not have their thresholds. However, by inspecting the results Roy's largest eigenvalue and the divergence measures all seem to be good candidates for a change-point detection algorithm.

## 4.2 Conclusion

The univariate theory is developed with the assumption that the parameters before the change point are known, which also gives the threshold for the  $\alpha$ -observations. When the parameters are estimated then the threshold is a stochastic variable that needs to be taken into account. This will lead to the same issue that we saw during the simulations for the Mahalanobis method.

During the simulation, we noticed that the result with the type I and II errors got closer to what the theory presented in Chapter 2 said when increasing  $n_0$ . Hence, with more observations used to estimate the covariance matrix,  $\Sigma_1$ , before the change the result followed the theory better. By looking at the theory we can see that we developed the method about  $\alpha$ -observations with a known  $\Sigma_1$ . By doing so, we ignored the uncertainty in estimating a covariance matrix. We would need to find a distribution for

$$\mathbf{Y}' \Sigma_1^{-1} \mathbf{Y},$$

where  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_2)$  with both  $\boldsymbol{\Sigma}_1$  and  $\boldsymbol{\Sigma}_2$  as unknown. Still, the simulation gave an indication that it could work. It was also possible to see that as the dimension increased so did the need for a good approximation of  $\boldsymbol{\Sigma}_1$ .

So far all the tests we have looked at have only been a function of the eigenvalues of the change matrix  $\boldsymbol{\Psi} = \boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1/2}$ . That is enough for the hypothesis of the largest eigenvalue, (3.6). A major thing that is missed in that hypothesis is that there could have been a change within the eigenvectors that is not noticed by the eigenvalues. All presented methods have had test statistics that just depend on the eigenvalues and would probably miss a change point that only affects the eigenvectors.

### 4.3 Future research

As commented in the simulation, to use the Kullback-Leibler divergence and the Bhattacharyya distance we need to find a threshold  $c_{\alpha_{test}}$ . This could be done by finding the distribution of (3.11) and (3.13).

The simulation showed that we need a good estimate of  $\boldsymbol{\Sigma}_1$  or the matrix itself for the result to be in line with the theory developed in Chapter 2 and 3. This was especially true for the type I error. We want to be able to use the method without needing to rely on the asymptotic result of the estimate of  $\boldsymbol{\Sigma}_1$ . In order to accomplish this, it becomes important to obtain the distribution of  $\mathbf{Y}'\mathbf{S}_1^{-1}\mathbf{Y}$ , whereby  $n_0\mathbf{S}_1$  follows a Wishart distribution characterized by the covariance matrix  $\boldsymbol{\Sigma}_1$  with  $n_0$  degrees of freedom. Here,  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_2)$ , with  $\boldsymbol{\Sigma}_2$  being an unknown parameter.

Since the Mahalanobis method does not estimate the covariance matrix of the new window, it implies that the window size could be smaller than the dimensions of the observations. This means it could be used as a 'semi high-dimensional' method, allowing the identification of changes with fewer observations than dimensions. As the dimensions increase, it becomes more important to have precise knowledge of  $\boldsymbol{\Sigma}_1$  in order to detect when a change occurs. Additionally, it would probably also work with knowing the distribution of  $\mathbf{Y}'\mathbf{S}_1^{-1}\mathbf{Y}$ , where  $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_2)$  with unknown  $\boldsymbol{\Sigma}_2$ .



---

## Bibliography

- Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367.
- Aminikhanghahi, S., Wang, T., and Cook, D. J. (2019). Real-time change point detection with application to smart home time series data. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):1010–1023.
- Avanesov, V. and Buzun, N. (2018). Change-point detection in high-dimensional covariance structure. *Electronic Journal of Statistics*, 12(2):3254–3294.
- Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109.
- Chen, J. and Gupta, A. K. (2012). *Parametric statistical change point analysis*. Birkhäuser, New York, NY, second edition.
- Dette, H. and Gösmann, J. (2018). Relevant change points in high dimensional time series. *Electronic Journal of Statistics*, 12(2):2578–2636.
- Dette, H. and Wied, D. (2016). Detecting relevant changes in time series models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 78(2):371–394.
- Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *Annals of Statistics*, 47(4):2051–2079.
- Feng, Y., Wen, M., Zhang, J., Ji, F., and Ning, G. X. (2016). Sum of arbitrarily correlated Gamma random variables with unequal parameters and its application in wireless communications. In *2016 International Conference on Computing, Networking and Communications, ICNC 2016*. IEEE.

- Gombay, E., Horváth, L., and Huskova, M. (1996). Estimators and tests for change in variances. *Statistics & Risk Modeling*, 14(2):145–160.
- Graybill, F. A. (1976). *Theory and application of the linear model*. Duxbury Press, North Scituate.
- Inclán, 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*, 89(427):913–923.
- Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, 43(6):2451–2483.
- Jitkrittum, W. (2013). Hadamard product to trace of matrix product. <http://wittawat.com/posts/hadamard-diag-trace.html>. Accessed: 2022-11-18.
- Keogh, E., Chu, S., Hart, D., and Pazzani, M. (2001). An online algorithm for segmenting time series. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 289–296.
- Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Miller, R. G. (1966). *Simultaneous statistical inference*. McGraw-Hill Book Co., New York.
- Moschopoulos, P. G. (1985). The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37(3):541–544.
- Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons, Inc., New York, NY.
- Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527.
- Reyes-Aldasoro, C. C. and Bhalerao, A. (2006). The Bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition*, 39(5):812–826.
- Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167.
- Wang, S. and Reynolds, M. R. (2013). A GLR control chart for monitoring the mean vector of a multivariate normal process. *Journal of Quality Technology*, 45(1):18–33.
- Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society Series B*, 80(1):57–83.

- 
- Wichern, D. W., Miller, R. B., and Hsu, D.-A. (1976). Changes of Variance in First-Order Autoregressive Time Series Models-With an Application. *Journal of the Royal Statistical Society*, 25(3):248–256.
- Zamba, K. D. and Hawkins, D. M. (2006). A multivariate change-point model for statistical process control. *Technometrics*, 48(4):539–549.



## **FACULTY OF SCIENCE AND ENGINEERING**

Linköping Studies in Science and Technology, Licentiate Thesis No. 1970, 2023  
Department of Mathematics

Linköping University  
SE-581 83 Linköping, Sweden

[www.liu.se](http://www.liu.se)