Master's Thesis

# Determining Changes in the Covariance Structure of Gaussian Processes

Jiyeon Han

Department of Computer Science and Engineering

Graduate School of UNIST

2020

# Determining Changes in the Covariance Structure of Gaussian Processes

Jiyeon Han

Department of Computer Science and Engineering

Graduate School of UNIST

# Determining Changes in the Covariance Structure of Gaussian Processes

A thesis
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Master of Science

Jiyeon Han

01/06/2020 of submission
Approved by

_____

Advisor
Kwang In Kim

# Determining Changes in the Covariance Structure of Gaussian Processes

Jiyeon Han

This certifies that the thesis of Jiyeon Han is approved.

01/06/2020 of submission

_____

Advisor: Kwang In Kim

_____

Committee Member: Jaesik Choi

_____

Committee Member: Se Young Chun

# Abstract

Time series data is everywhere, such as stock data in finance market, the sensor data in factories, or the temperature data in everyday life. Time series data have been studied for a long time to analyze and predict the future behavior. While it is tractable when the sequential data behave as stationary, it becomes difficult to model and predict non-stationary time series. Change point detection is a problem that identify non-stationarity which has been investigated for decades in many different names. Change point detection is a challenging problem because defining a change point decisively and objectively is difficult in nature. In this thesis we are trying to define and find a change point using hypothesis tests based on statistics. Specifically we focus on structural breaks in the covariance structure of Gaussian Processes. Further we propose an online change point detection algorithm, called Confirmatory Bayesian Online Change Point Detection, by leveraging the devised hypothesis tests into the conventional Bayesian online change point detection algorithm.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**BOCPD** Bayesian Online Change Point Detection. 1, 2, 5, 6, 10, 12, 20–26, 30

**CBOCPD** Confirmatory Bayesian Online Change Point Detection. 5, 20, 22–24

**CPD** Change Point Detection. 1, 2, 7, 10, 30

**GP** Gaussian Process. 2–4, 7, 9, 10, 13–16, 30

# Chapter I

# Introduction

Time series data is everywhere, such as stock data in finance market, the sensor data in factories, or the temperature data in everyday life. We model the time series data to analyze past events and predict the future behavior. When sequential data is stationary, i.e., the parameters in the underlying distribution does not change, predicting future events is feasible. Nonetheless, the stationarity assumption cannot be guaranteed in practice. The *change point detection* (CPD) problem, which intends to detect existence of changes in sequential data, is one of the fundamental problems in time series analysis that can be a key factor in improving the prediction of future events.

A *change point* is defined as a particular position in sequential data where the underlying distribution changes. Change points take critical roles in plentiful real-world applications, including image analysis [1], speech recognition [2], climate modeling [3], and human activity recognition [4]. In econometrics, CPD has been studied for decades in the name of *structural breaks*, that essentially employ CPD on regression models to identify stability in the structure of time series [5, 6]. As well, trend filtering determines change points with the assumption on piece-wise linearity in the sequential data [7]. Furthermore the domain adaptation problem, or the covariate shift in the other name, is also another field of study where CPD takes an important role, because changes in the distribution of the test data and the training data affect to the performance of the many machine learning models [8].

One can categorize existing CPD into hypothesis-test based-approaches or Bayesian-inference-based approaches. Hypothesis-test-based approaches apply statistical tests to determine the presence of changes, where the error probability is naturally defined in the framework. Hypothesis tests are applied in various ways including the kernel methods such as kernel Fisher discriminant ratio [9], two-sample tests based on the maximum mean discrepancy [10, 11], cumulative sum (CUSUM) test [12, 13], and likelihood ratio test [14, 15].

Bayesian inference methods [16, 17] adapt the Bayesian framework to compute the distribution of a possible change based on a prior belief about the occurrence of a change and the observed sequential data. BOCPD algorithms [18–21] detect change points in an online manner

considering the interval called *run length* between change points. However, with probabilistic methods we cannot define change points decisively nor guarantee statistical error bound of a change, which often affect the reliability of the algorithm. Furthermore aforementioned Bayesian algorithms are highly sensitive to selected hyperparameters of underlying predictive model.

While GPs are widely used to model time series data, there is a limitation that conventional GPs work on globally smooth functions. On the other hand GPs with change points can model locally smooth functions [20], where many of the real-world time series data yield only locally smooth functions rather than globally smooth functions. A number of studies have been conducted on CPD in GPs. In a previous research, a likelihood ratio test is proposed based on the null hypothesis of stationary GPs [14]. However, there is a limitation that even though the null hypothesis is rejected one cannot assure a change if the null distribution is not legitimate. Another work has studied to detect the changes in the mean function of GPs by proposing likelihood ratio tests [15].

Our goal is twofold. First, we propose novel likelihood ratio tests which detect structural breaks in the covariance of GPs. Secondly, we propose an online CPD algorithm by cooperating the proposed likelihood ratio test which takes advantages both from the statistical test and Bayesian framework.

The rest chapters are constructed as follows. In Chapter II, the basic concepts and related work are explained briefly to help understanding the following chapters. In Chapter III we propose likelihood ratio tests to define change points of covariance structural break and show that the proposed tests theoretically guarantee designated test error bound under proper conditions. In Chapter IV we propose an online CPD algorithm, Confirmatory BOCPD, that detects change points with an acceptable time delay. We further show that the proposed algorithm guarantees improved prediction performance compared to conventional algorithm. Moreover, we provide examples to show the proposed algorithm properly conforms the parameter of BOCPD algorithm to reduce missed detections and false alarms. We conclude with a summary of research findings and suggestions of future work in Chapter VI.

# Chapter II

# Background

## 2.1 CPD in the Mean of Gaussian Processes

### 2.1.1 Gaussian Processes

GP is a random process that is formed with a set of random variables where any finite subset of random variables follows a multivariate Gaussian distribution. It can also be considered as a distribution of functions that the marginal distribution of $(f(t_1), f(t_2), ..., f(t_n))$ given any finite set of inputs $(t_1, t_2, ..., t_n)$ forms a multivariate Gaussian distribution or a multivariate Normal distribution. We can fully describe a GP with the mean function along with the covariance function. The mean function $\mu(\cdot) = \mathbb{E}[f(\cdot)]$ describes the trend of a time series data. The covariance function, on the other hand, describes how much a data point would affect with another data point. In GPs, kernel functions are used for the covariance function as $\mathrm{Cov}(f(t_i), f(t_j)) = k(t_i, t_j)$. A kernel function calculates the covariance between two data points using only input aspects. For example, today's stock price is affected by a stock data of yesterday more compare to a stock data of a month ago. Both 'kernel function' and 'covariance function' will be used interchangeably throughout this paper. The kernel function represents distinctive aspects of the time series data, such as periodicity, length scales, and variance. As an instance, there is a Radial Basis Function kernel which is defined as $K(t, t') = \sigma^2 \exp(-\frac{(t-t')^2}{2l^2})$. As the length scale hyperparameter $l$ increases, the function becomes smoother as the penalty from the distance is relaxed with large $l$. Table 2.1 lists some commonly used kernel functions.

In this paper we consider sequential index $t$ as input with the assumption that the sequence is equi-interval. Then the data value $X_t$ corresponding to the input $t$ is modeled as $X_t \sim N(f(t), \sigma_n^2)$ with the noise variance $\sigma_n^2$. Given GP hyperparameters $\theta_m$, the log likelihood of the GP regression over the observed samples $\mathbf{X}$ can be computed as the likelihood of the multivariate Gaussian distribution as follows,

$$\log p(\mathbf{X}|\theta_m, \sigma_n) = -\frac{1}{2}(\mathbf{X}-\mu)(\Sigma + \sigma_n^2 I)^{-1}(\mathbf{X}-\mu) - \frac{1}{2}\log|\Sigma + \sigma_n^2 I| - \frac{n}{2}\log 2\pi$$

where $\mu$ and $\Sigma$ denote the mean function and the covariance matrix respectively and $n$ denotes

| Kernel | Formula |
|---|---|
| Linear | $k_{lin}(x, x') = \sigma_b^2 + \sigma_v^2(x - c)(x' - c)$ |
| Periodic | $k_{per}(x, x') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi\|x-x'\|/p)}{l^2}\right)$ |
| Squared Exponential | $k_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right)$ |

Table 2.1: Various kernels and its formula.

the number of observed data.

### 2.1.2 Hypothesis Tests for Structural Break

Structural break refers to the abrupt change of the parameters in the underlying regression model of time series, that can lead a large prediction error and therefore reduces reliability of the static regression model. To check the structural stability, i.e., the time-invariance of the regression coefficients, there have been investigated and used many methods most of which are based on hypothesis tests.

Two hypotheses are used for the hypothesis test with the observed sequential data. One hypothesis is the null hypothesis, $\mathbb{H}_0$, which assumes there is a structural break. The other hypothesis is called the alternative hypothesis, $\mathbb{H}_1$, which insists there exists a structural break. The hypothesis test is typically constructed as follows. Given the two hypotheses, the test statistic $Y$ is defined according to the regression model. Then test if the test statistic lies in a probable region with a threshold $\mathfrak{R}$ as

$$\mathfrak{T} = \mathbb{I}\left(Y \geq \mathfrak{R}\right)$$

with an indicator function $\mathbb{I}(\cdot)$. We reject $\mathbb{H}_0$ when $\mathfrak{T}_{GLRT}$=1, otherwise, we fail to reject $\mathbb{H}_0$. $\varphi_n(\mathfrak{T})$, named as the conditional detection error probability, is defined as

$$\varphi_n(\mathfrak{T}) = \mathbb{P}(\mathfrak{T} = 1|\mathbb{H}_0) + \mathbb{P}(\mathfrak{T} = 0|\mathbb{H}_1). \tag{2.1}$$

We call $\mathbb{P}(\mathfrak{T} = 1|\mathbb{H}_0)$ as Type $I$ error or false alarm rate. $\mathbb{P}(\mathfrak{T} = 0|\mathbb{H}_1)$ is called Type $II$ error or missing detection rate. Here we introduce examples of hypothesis tests for testing structural breaks.

#### Chow Test

Chow test is used for testing structural break in linear regression models. We introduce how it works in this subsection. First, assume we have n observed samples and try to model with a

normal linear regression.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Suppose we have $m$ additional observations. We want to determine whether they come from the same regression model as the previously observed $n$ samples.

Let's assume that the size $m$ of the second sample is larger than $p$. Then, the model of general linear hypotheses takes the form

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

Under the null hypothesis ($H_0 : \beta_1 = \beta_2\beta$), we can rewrite the model as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

$n$ observations and $m$ observations come from the same regression model, we can find the least-square estimator of $\beta$ as follows.

$$b_0 = \left[ \begin{pmatrix} X_1' & X_2' \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right] \begin{bmatrix} X_1' & X_2' \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$= \left[ X_1'X_1 + X_2'X_2 \right]^{-1} \begin{bmatrix} X_1' & X_2' \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \beta + \left[ X_1'X_1 + X_2'X_2 \right]^{-1} \begin{bmatrix} X_1' & X_2' \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

The summation of the squares of the residuals of true data and the estimated values under the null hypothesis $H_0$ ($\beta_1 = \beta_2 = \beta$) becomes

$$\left\| \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} b_0 \right\|^2 = \left[ \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} b_0 \right]' \left[ \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} b_0 \right]$$

$$= \begin{bmatrix} \epsilon_1' & \epsilon_2' \end{bmatrix} \left[ I - \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \left( X_1'X_1 + X_2'X_2 \right)^{-1} \begin{pmatrix} X_1' & X_2' \end{pmatrix} \right] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

In hear, the quadratic form has rank $n+m-p$. The summation over the squares of the residuals under the alternative hypothesis $H_\alpha$ ($\beta_1 \neq \beta_2$)

$$\left\| \begin{matrix} y_1 - X_1 b_1 \\ y2 - X_2 b_2 \end{matrix} \right\|^2 = ||y_1 - X_1 b_1||^2 + ||y_2 - X_2 b_2||^2$$

$$= \epsilon_1' \left[ I - X_1 (X_1'X_1)^{-1} X_1' \right] \epsilon_1 + \epsilon_2' \left[ I - X_2 (X_2'X_2)^{-1} X_2' \right] \epsilon_2$$

The first quadratic term has the rank $n-p$ and the second quadratic term has the rank $m-p$ respectively

$$= \begin{bmatrix} \epsilon_1' & \epsilon_2' \end{bmatrix} \begin{bmatrix} I - \begin{pmatrix} I - X_1(X_1'X_1)^{-1}X_1' \\ I - X_2(X_2'X_2)^{-1}X_2' \end{pmatrix} \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

Hence, the quadratic form has rank $n+m-2p$. We can decompose the sum of squares of the residuals under the null hypothesis $H_0$. First start from the identity,

$$\begin{bmatrix} y_1 - X_1 b_0 \\ y_2 - X_2 b_0 \end{bmatrix} = \begin{bmatrix} y_1 - X_1 b_1 \\ y_2 - X_2 b_2 \end{bmatrix} + \begin{bmatrix} X_1 b_1 - X_1 b_0 \\ X_2 b_2 - X_2 b_0 \end{bmatrix}$$

As norm of both sides are also equal, we square both sides to get

$$\left\| \begin{matrix} y_1 - X_1 b_0 \\ y_2 - X_2 b_0 \end{matrix} \right\|^2 = \left\| \begin{matrix} y_1 - X_1 b_1 \\ y_2 - X_2 b_2 \end{matrix} \right\|^2 + \left\| \begin{matrix} X_1 b_1 - X_1 b_0 \\ X_2 b_2 - X_2 b_0 \end{matrix} \right\|^2$$

Since $\begin{bmatrix} y_1 - X_1 b_1 \\ y_2 - X_2 b_2 \end{bmatrix}' \begin{bmatrix} X_1 b_1 - X_1 b_0 \\ X_2 b_2 - X_2 b_0 \end{bmatrix} = 0$ , we can write as

$$Q_1 = Q_2 + Q_3$$

Since

$$\left[ X_1'X_1 + X_2'X_2 \right] b_0 = X_1'y_1 + X_2'y_2 = X_1'X_1 b_1 + X_2'X_2 b_2$$

Following is satisfied

$$b_2 - b_0 = -(X_2'X_2)^{-1}X_1'X_1(b_1 - b_0)$$

We can rewrite $Q_3$ as

$$\left\| \begin{matrix} X_1 b_1 - X_1 b_0 \\ -X_2(X_2'X_2)^{-1}X_1'X_1(b_1 - b_0) \end{matrix} \right\|^2$$

$$= \left[ b_1' - b_0' \right] \begin{bmatrix} X_1' & -X_1'X_1(X_2'X_2)^{-1}X_2' \end{bmatrix} \begin{bmatrix} X_1 \\ -X_2(X_2'X_2)^{-1}X_1'X_1 \end{bmatrix} \left[ b_1 - b_0 \right]$$

$Q_3$ is a quadratic form in $b_1 - b_0$ and cannot have rank higher than $p$. We showed that the ranks of $Q_1$ and $Q_2$ are $n+m-p$ and $n+m-2p$ respectively. As the rank of $Q_1$ is smaller than or equal to the summation of the rank of $Q_2$ the rank of $Q_3$, the rank of $Q_3$ should be higher than or equal to the $n+m-p-(n+m-2p) = p$. Combining with the condition that the rank of $Q_3$ cannot be higher than $p$, we can say that the rank of $Q_3$ is equal to $p$. Under the null hypothesis, $Q_2$ and $Q_3$ independently follow $\chi^2(m+n-2p)\sigma^2$ and $\chi^2(p)\sigma^2$, respectively. The distribution of $Q_3$ will only be affected when $H_0$ is not the case. Thus we can test if $H_0$ holds or not using the ratio

$$F(p, m+n-2p) = \frac{Q_3/p}{Q_2/(m+n-2p)}$$

$$= \frac{||X_1 b_1 - X_1 b_0||^2 + ||X_2 b_2 - X_2 b_0||^2}{||y_1 - X_1 b_1||^2 + ||y_2 - X_2 b_2||^2} \cdot \frac{(m+n-2p)}{p}$$

**CUSUM**

CUSUM, standing for cumulative sum, tests if there exists a shift in a parameter of the probability distribution. CUSUM test accumulates the residuals from the target value of the parameter and detect a shift when the summation exceeds some threshold. Specifically for detecting a mean shift in sequential data $\{x_i\}$, we define

$$\begin{aligned} S_0 &= 0 \\ S_n^+ &= \max(0, S_{n-1} + x_n - \mu_0 - K/2) \\ S_n^- &= \max(0, S_{n-1} + x_n - \mu_0 + K/2) \end{aligned}$$

with the original mean value $\mu_0$ and the minimum jump size $K$. The increment in the mean can be detected with

$$S_n^+ - \min_{i<n}(S_i) \geq R$$

and similarly the decrement in the mean can be detected with

$$S_n^- - \max_{i<n}(S_i) \geq R.$$

When there is no shift in the mean, $S_n^{\pm}$ will keep decreasing or increasing by about $K/2$. When there is a shift in the mean, the cumulative residual from the original mean value will start to increase in case of positive jump and the difference from the minimum cumulative sum will exceed the threshold. Similar procedure is applied for the case of negative jump.

### 2.1.3 CPD in the Mean of Gaussian Processes

When modeling a time series, there are cases that mean jumps at a point. For such cases we need to verify if there exists such jump in mean and if there is, how much mean jumps. But even if we say there exists a jump, how sure we can be is another question. From structural break approach, [15] have proposed an optimal likelihood ratio test for detecting a single change point in the mean function of a GP. In this section we briefly review some of the results in the paper on the detection of a single jump in the mean function of a GP. We introduce a formal statistical test to find a sudden jump in the mean function of a GP and how to measure the certainty of the test.

We write $n$ samples of time series data as $X = \{X_t\}_{t=1}^n$. Let $t \in \mathcal{C}_n \subseteq \{1, ..., n\}$ represents the point of sudden change where $\mathcal{C}_n$ denotes the set of possible change point candidates. We set two hypotheses for the likelihood ratio test given observed sequential data. The first hypothesis is the null hypothesis, $\mathbb{H}_0$, which assumes there is no change of the parameters in the underlying GP model. The alternative hypothesis, $\mathbb{H}_1$, assumes there is at least one change in the parameters. The likelihood ratio test is constructed as follows. With previously defined hypotheses, the likelihood ratio is calculated as

$$2\mathfrak{L} = 2(\sup_{\theta \in \Theta_1} \ell(\theta_1) - \sup_{\theta \in \Theta_0} \ell(\theta_0))$$

where $\ell$ is the log likelihood function and $\Theta_0$ and $\Theta_1$ are the parameter spaces of $\mathbb{H}_0$ and $\mathbb{H}_1$ respectively. The generalized likelihood ratio test (GLRT) is then defined as

$$\mathfrak{T}_{GLRT} = \mathbb{I}\left(2\mathfrak{L} \geq \mathfrak{R}_{n,\delta}\right)$$

where $\mathfrak{R}_{n,\delta}$ is the threshold with $n$, the number of data points and $\delta$, the upper bound of the conditional detection error, $\varphi_n(\mathfrak{T})$, as in Equation (2.1). We reject $\mathbb{H}_0$ when $\mathfrak{T}_{GLRT}$=1, otherwise, we fail to reject $\mathbb{H}_0$. In the mean change detection problem in a GP, the null hypothesis assumes that the samples follow a GP of zero mean, which can be written as

$$\mathbb{H}_0 : \mathbb{E}X = \mathbf{0}.$$

The associative alternative hypothesis corresponding to a specific time $t$, in contrast, assumes that there exists a change point of jump size $b$ in the mean function at time $t$ which can be written as

$$\mathbb{H}_{1,t} : \exists\ b \neq 0,\ \mathbb{E}X = \frac{b}{2}\zeta_t.$$

Here $\zeta_t \in \mathbb{R}^n$ is defined with $\zeta_t(k) := sign(k-t)$ for $t \in \mathcal{C}_n$. For example, with $n = 5$, $\zeta_3 = [-1, -1, 1, 1, 1]$. Unifying over the set of change point candidates, the alternative hypothesis states there exists more than or equal to one change point with a jump size $b$ as written below,

$$\mathbb{H}_1 : \bigcup_{t\in\mathcal{C}_n} \mathbb{H}_{1,t}.$$

With the above hypotheses, we rewrite $2\mathfrak{L}$ with

$$\begin{aligned} 2\mathfrak{L} &= X^T\Sigma_n^{-1}X - \min_{t\in\mathcal{C}_n}\min_{b\neq 0}\left[\left(X-\frac{b}{2}\zeta_t\right)^T\Sigma_n^{-1}\left(X-\frac{b}{2}\zeta_t\right)\right] \\ &= \max_{t\in\mathcal{C}_n}\max_{b\neq 0}\left(-\frac{\zeta_t^T(\Sigma_n)^{-1}}{4}b^2 + b\zeta_t^T(\Sigma_n)^{-1}X\right) \qquad (2.2) \\ &= \max_{t\in\mathcal{C}_n}\left|\frac{(\zeta_t^T(\Sigma_n)^{-1}X)}{\sqrt{\zeta_t^T(\Sigma_n)^{-1}\zeta_t}}\right|^2. \qquad (2.3) \end{aligned}$$

in which $\Sigma_n$ represents the covariance matrix of $X$. From Equation (2.2) to Equation (2.3), we take derivative in terms of b to get the maximum value. Plugging $b$ that maximizes Equation (2.2) in the test and rearranging it, we get the following formulation.

$$\mathfrak{T}_{GLRT} = \mathbb{I}\left(\max_{t\in\mathcal{C}_n}\left|\frac{\zeta_t^T\Sigma^{-1}X}{\sqrt{\zeta_t^T\Sigma^{-1}\zeta_t}}\right|^2 \geq \mathfrak{R}_{n,\delta}\right)$$

When we find a proper threshold $\mathfrak{R}_{n,\delta}$, we can bound the conditional error probability by $\delta$ $\varphi_n(\mathfrak{T}_{GLRT}) \leq \delta$ under the sufficient condition on $b$ [15]. One choice of $\mathfrak{R}_{n,\delta}$ that would work is

$$\mathfrak{R}_{n,\delta} = 1 + 2\left[log\left(\frac{2n}{\delta}\right) + \sqrt{log\left(\frac{2n}{\delta}\right)}\right].$$

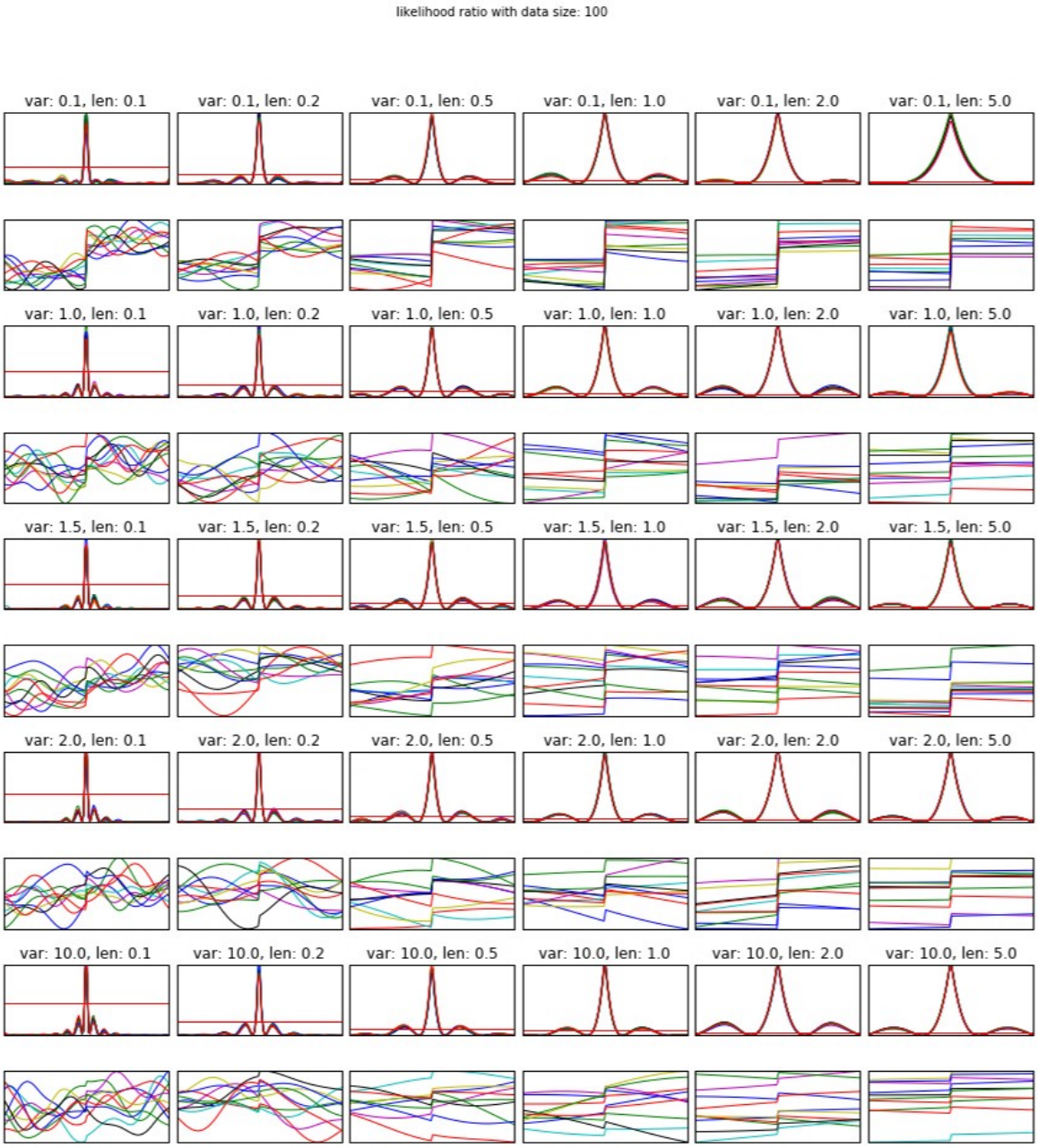



Figure 2.1: Synthetic data with a change in the mean function of a GP at the middle of 100 points of sequential data and its likelihood ratio varying variance and lengthscale hyperparameters. For every two rows, the upper row shows the likelihood ratio for each possible change point and the lower row shows the corresponding synthetic data.

Figure 2.1 shows various synthetic data of mean change of GPs sampled from Squared Exponential kernel with varying hyperparameters. The synthetic data is generated to have a change point at the middle of sequential data. The proposed threshold is ploted as a red horizontal line in the plot likelihood ratio, which is the upper row of the every two rows. We can see that in every cases, the likelihood ratio is maximized at the true change point, and the bound works correctly to test changes. However, the proposed bound is too general that it only counts for the total number of data and the desired detection error, but not the scale of the data nor the scale of the covariance values which can affect the scale of the test statistics.

## 2.2 Bayesian Online CPD

In many real world applications of change point detection, an online approach is necessary to instantly detect and utilize the change points. BOCPD [18] is one of the online change point detection algorithm that uses Bayesian inference to update the probabilistic distribution of the upcoming data point using the information of the change point distribution. We define a specific term for this that counts the number of time points from the last change point, called run length. One assumption behind the BOCPD framework is that the partitions, divided by the change points, are independent to each other while data in a single partition are correlated one another. The BOCPD works as follows. First let $x_t$ be the observed data at time t. The run length at time $t$ is written as $r_t$. Then ${x_t}^{(r_t)}$ is used to denote the observed data after the last change point. Our goal is to compute the probabilistic distribution of the future data $x_{t+1}$ given the observed data upto $t$. Write it in cumulative form on $r_t$,

$$P(x_{t+1}|x_{1:t}) = \sum_{r_t} P(x_{t+1}, r_t|x_{1:t}) \tag{2.4}$$

$$= \sum_{r_t} P(x_{t+1}|r_t, x_{1:t})P(r_t|x_{1:t}) \tag{2.5}$$

$$= \sum_{r_t} P(x_{t+1}|r_t, x_t^{(r_t)})P(r_t|x_{1:t}) \tag{2.6}$$

$$= \sum_{r_t} P(x_{t+1}|x_t^{(r_t)})P(r_t|x_{1:t}). \tag{2.7}$$

From (2.4) to (2.5), the Bayes theorem $P(A, B|C) = P(A|B, C)P(B|C)$ is used and from (2.5) to (2.6) and from (2.6) to (2.7), we use the assumption that $x_{t+1}$ only depends on the last $r_t$ number of data. Now expanding the second term of (2.7), which represents the posterior distribution of the run length conditioned on the previous data which can be written as, $P(r_t|x_{1:t}) = P(r_t, x_{1:t})/P(x_{1:t})$.

Then the numerator of right hand side can be expressed as marginal distribution over $r_{t-1}$,

$$P(r_t, x_{1:t}) = \sum_{r_{t-1}} P(r_t, r_{t-1}, x_{1:t}) \tag{2.8}$$

$$= \sum_{r_{t-1}} P(r_t, x_t | r_{t-1}, x_{1:t-1}) P(r_{t-1}, x_{1:t-1}) \tag{2.9}$$

$$= \sum_{r_{t-1}} \big[ P(r_t | r_{t-1}, x_{1:t}) P(x_t | r_{t-1}, x_{1:t-1}) \cdot P(r_{t-1}, x_{1:t-1}) \big] \tag{2.10}$$

$$= \sum_{r_{t-1}} \big[ P(r_t | r_{t-1}) P(x_t | r_{t-1}, x_{t-1}^{(r_{t-1})}) \cdot P(r_{t-1}, x_{1:t-1}) \big]. \tag{2.11}$$

Here, Bayes' theorem is used from Equation (2.8) to Equation (2.9) and from Equation (2.9) to Equation (2.10). From Equation (2.10) to Equation (2.11), $P(r_t|r_{t-1}, x_{1:t})$ is simplified to $P(r_t|r_{t-1})$ with assuming that the current run length is only depending on the run length from the previous time step, but not the data. Watching Equation (2.8) and Equation (2.11), we can see $P(r_t, x_{1:t})$ is in recursive form with time. So if we know first term and second term of Equation (2.11), which represent prior distribution of $r_t$ given $r_{t-1}$ and the predictive distribution of $x_t$ given the data after the last change point, then we can compute the joint distribution of run length and data recursively. Here the prior distribution of $r_t$ can be easily get as:

$$P(r_t | r_{t-1}) = \begin{cases} H(r_{t-1} + 1) & r_t = 0 \\ 1 - H(r_{t-1} + 1) & r_t = r_{t-1} + 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

with

$$H(\tau) = \frac{P_{gap}(g = \tau)}{\sum_{t=\tau}^{\infty} P_{gap}(g = t)}. \tag{2.13}$$

Here $H(\tau)$ is called the hazard function with $P_{gap}(g)$ denoting the a priori distribution of the gap between two change points. Especially, the hazard function is simplified to be a constant function as $H(\tau) = 1/\lambda$ when the geometric distribution is employed for $P_{gap}(g)$ with the timescale parameter $\lambda$. Then *lambda* is large, the change point is less likely to be happen and vice versa. The second term in Equation (2.11) is computed through a GP regression. The overall conditional predictive distribution $P(x_{t+1}|x_{1:t})$ can be updated by recursive message passing of $P(r_t, x_{1:t})$.

The BOCPD framework efficiently works to find changes when modeling time series wiht GPs. However we can enhance the algorithm by relaxing some assumptions behind the BOCPD framework. One such assumption is the constant hazard function which affects to the frequency of change points. Another assumption is that the kernel functions or the kernel parameters are

fixed during the algorithm works. These assumptions make BOCPD framework to be vulnerable to the parameters. In the following chapters, we propose an enhanced BOCPD algorithm which overcomes the aforementioned drawbacks.

# Chapter III

# Defining Changes in the Covariance Structure of Gaussian Processes

This section presents our new hypothesis tests to detect change points in the covariance structure of a GP.

## 3.1 Motivational Examples

While there have been many works concentrated on the mean change in the GPs, changes in the covariance structure of GPs have not yet been fully investigated. In this subsection, we address the motivational examples of covariance changes in time series analysis.

#### Covariance can represent various changes in time series

As we can see in Figure 3.1, covariance matrix can represent various changes in synthetic data. Figure 3.1a shows samples from a covariance matrix of

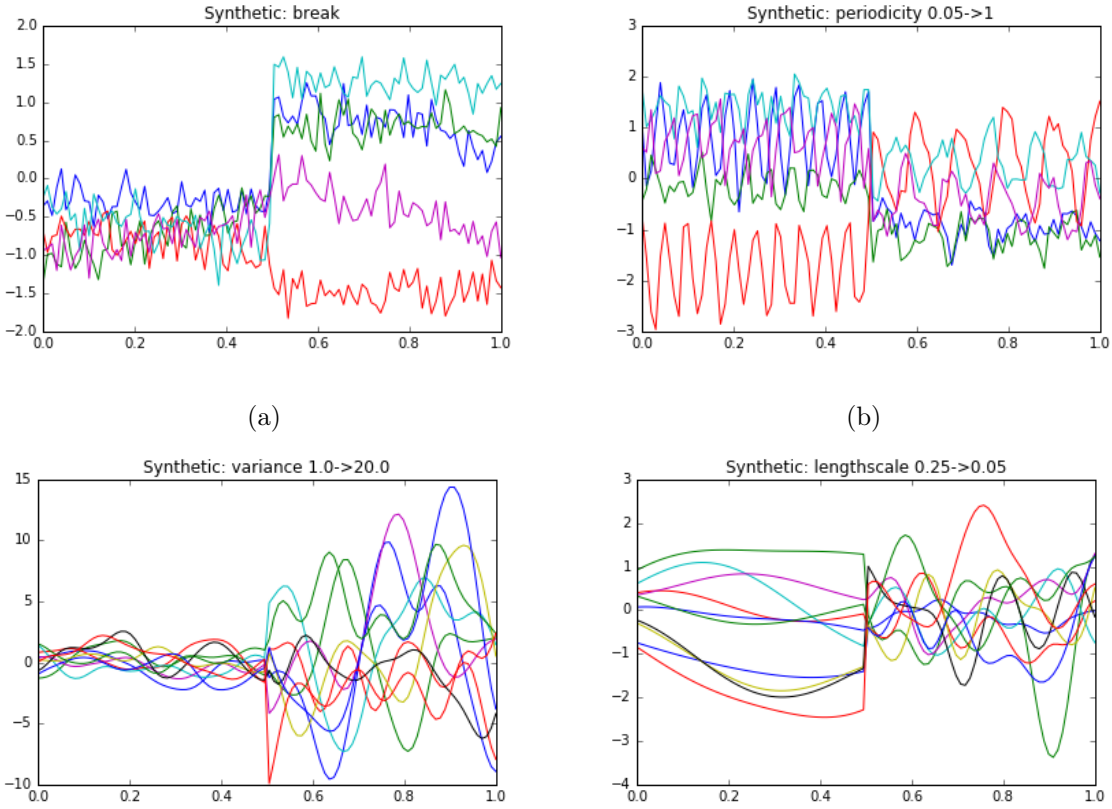
$$\Sigma_{break} = \begin{bmatrix} K & 0 \\ 0 & K \end{bmatrix}$$

where $K_{ij} = k(t_i, t_j)$ for some covariance function $k$. $\Sigma_{break}$ represents a structural break in the covariance matrix where the process before and after the change point are independent. Especially the change type of Figure 3.1a is similar to mean shift. In other words, changes such like mean shift can also be represented with covariance structural breaks. Figure 3.1b shows change of periodicity $p$ in the *Periodic* covariance function

$$k_{per}(t_i, t_j) = \sigma^2 \exp\left(-\frac{2\sin^2(\pi|t_i - t_j|/p)}{l^2}\right).$$

Figure 5.1b and 5.1b show changes of variance $\sigma$ and lengthscale $l$ in the *SquaredExponential* covariance function

$$k_{SE}(t_i, t_j) = \sigma^2 \exp\left(-\frac{(t_i - t_j)^2}{2l^2}\right).$$




(a) (b)

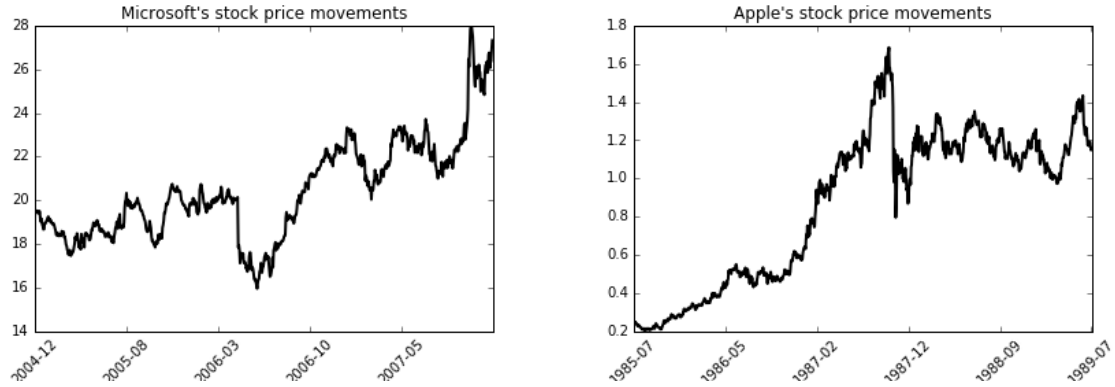
(c) (d)

Figure 3.1: Synthetic data with changes in the covariance structure. Figure 3.1a shows samples generated from the covariance matrix with a structural break at the middle. Figure 3.1b shows the periodicity change in the periodic covariance function. Figure 5.1b and 5.1a show the variance and lengthscale changes in the squared exponential covariance function, respectively.

These type of changes are not only theoretical examples but also happening in the real world. Figure 3.2 shows two real world examples of covariance changes. Figure 3.2a is Apple's stock movement for December 2004 to December 2007. We can see that the variance becomes larger after March 2006. Figure 3.2b is Microsoft's stock movement for July 1985 to July 1989. We can see that the lengthscale, i.e., the smoothness of the time series changes after September 1987.

**Detecting changes in the covariance structure is helpful for the prediction**

Figure 3.3 shows that covariance structural breaks can affect the performance of a GP regression. The left-most plot in Figure 3.3 presents the several samples generated from a GP with an embedded covariance change at the time step 5. The second plot shows a GP regression with GP hyperparameters learnt from the whole set of data at once. In contrast, the right-most plot presents a GP regression whose hyperparameter changes at the change point. Figure 3.3 advocates the necessity of the research on the covariance structural break as the figure clearly shows that nonstationary data is fitted better with the dynamic model than the time-invariant




(a) (b)

Figure 3.2: The figure 3.2a shows the Microsoft's stock price movements from December 2004 to December 2007 which represents variance change. The figure 3.2a shows the Apple's stock price from July 1985 to July 1989 which represents length scale change.

model. A GP becomes more expressive when it becomes to model a change in the covariance structure.

For the rest of this section, we first recall the notations to construct likelihood ratio test for general covariance change and then move on to our thesis interest, devising likelihood ratio tests for covariance structural break.

## 3.2 Problem Setting

In this section, we denote the sequential data as $X_i = f(t_i)$ with time sequence $\{t_i\}$ and $f \sim \mathcal{GP}$ as in Section 2.1.3. For possible set of change points, we denote with $\mathcal{C}_n$ where $n$ represents the size of the sequential data. We formulate a statistical test to detect the covariance structural changes. First we define the null hypothesis as

$$\mathbb{H}_0 : \mathrm{Cov}(X_i, X_j) = K(t_i, t_j)$$

and the alternative hypothesis as $\mathbb{H}_1 = \bigcup_{t \in \mathcal{C}_n} \mathbb{H}_{1,t}$,

with

$$\mathbb{H}_{1,t} : \mathrm{Cov}(X_i, X_j) = \begin{cases} K(t_i, t_j), & i, j < t \\ K'(t_i, t_j), & i, j \geq t \\ K''(t_i, t_j), & \text{otherwise.} \end{cases} \tag{3.1}$$

$K$, $K'$ and $K''$ can be arbitrary kernel functions. We write the covariance matrix for $\mathbb{H}_0$ as $\Sigma$ and the covariance matrices for $\mathbb{H}_{1,t}$ as $\Sigma'_t$, respectively. We can rewrite the likelihood ratio $2\mathfrak{L}$ as

$$\max_{t \in \mathcal{C}_n} \left[ X^T(\Sigma)^{-1}X - X^T(\Sigma'_t)^{-1}X + \ln\left(\frac{|\Sigma|}{|\Sigma'_t|}\right) \right]. \tag{3.2}$$

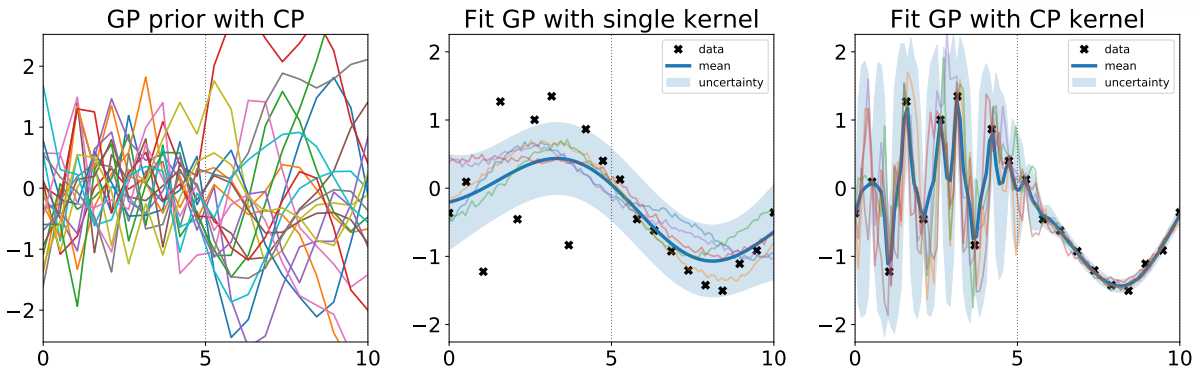



Figure 3.3: The left-most plot shows samples from a GP with a predefined covariance structural change. The middle plot shows the GP posterior from the GP regression with a static kernel. The right-most plot shows the GP posterior from the GP regression with two consecutive kernels representing the covariance structural break.

## 3.3 Tests for the Covariance Structural Break

In this subsection, we focus on the situation where the covariance structure breaks with two different kernels, i.e., $K''(i,j) = 0$ in Equation (3.1). Similarly we can write hypotheses as follows.

$$\mathbb{H}_0 : \mathrm{Cov}(X_i, X_j) = K(X_i, X_j), \quad v.s. \quad \mathbb{H}_1 = \bigcup_{t \in \mathcal{C}_n} \mathbb{H}_{1,t}$$

where specific alternative hypothesis with change point t, $\mathbb{H}_{1,t}$ is defined as

$$\mathbb{H}_{1,t} : \exists\ \alpha \neq 0, \mathrm{Cov}(X_i, X_j) = \begin{cases} K(X_i, X_j), & i,j < t \\ K'(X_i, X_j), & i,j \geq t \\ 0, & \text{Otherwise} \end{cases}$$

The covariance matrices under the null hypothesis and the alternative hypothesis can be written as

$$\Sigma = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{pmatrix}, \Sigma_t' = \begin{pmatrix} K_{aa} & 0 \\ 0 & K'_{bb} \end{pmatrix}.$$

Here, $K_{ab}$ is the covariance matrix between $X_a$ and $X_b$ where $X_a := X_{1:t}$ and $X_b := X_{t+1:n}$. $K_{aa}$, $K_{ba}$, $K_{bb}$ are similarly defined. The likelihood ratio test is formed as

$$\mathfrak{T}_{GLRT} = \mathbb{I}\left(2\mathfrak{L} \geq \mathfrak{R}_\delta\right)$$

.

**Lemma 3.3.1.** *Suppose a sequential data $X = X_{1:n}$ is bounded with $X_t \in [-V, V]$ for all t,*

$$\lambda_n n V^2 \leq X^T M X \leq \lambda_1 n V^2$$

*for symmetric matrix $M$ and its eigenvalues $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.*

*Proof.* As $M$ is a symmetric matrix, it can be decomposed as

$$M = Q\Lambda Q^T$$

where $Q$ is the orthogonal matrix whose columns are eigenvectors of $M$ and $\Lambda$ is a diagonal matrix with eigenvalues. Then

$$X^T M X = X^T Q \Lambda Q^T X = \sum_i \lambda_i < X, Q_i >^2 .$$

Using the fact $\lambda_n \leq \lambda_i \leq \lambda_1$ for any $i \in [1, n]$, it can be bouded as

$$\lambda_n \sum_i < X, Q_i >^2 \leq \sum_i \lambda_i < X, Q_i >^2 \leq \lambda_1 \sum_i < X, Q_i >^2 .$$

As $\sum_i < X, Q_i >^2 = X^T Q Q^T X = \|X\|^2$, we can conclude

$$\lambda_n n V^2 \leq X^T M X \leq \lambda_1 n V^2.$$

□

**Theorem 3.3.1** (Hoeffding bound)**.** *Let $Z$ be a random variables bounded by the interval $[a, b]$. Then*

$$P(Z - E[Z] \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{(b-a)^2}\right)$$
$$P(Z - E[Z] \leq -\epsilon) \leq \exp\left(-\frac{2\epsilon^2}{(b-a)^2}\right)$$

*for $\epsilon \geq 0$.*

From Lemma 3.3.1 and Hoeffding bound, we can find the thresholds for likelihood ratio tests.

**Lemma 3.3.2** (Type I Error)**.** *When the $n$ number of sequential data follow the null hypothesis as defined in Section 3.3 and bounded by $[-V, V]$, the error rate that the likelihood ratio test to wrongly detect a change is bounded as follows.*

$$\mathbb{P}(2\mathfrak{L} \geq \mathfrak{R}_{n,\delta,\mathbb{H}_0} | \mathbb{H}_0) \leq \delta/2,$$

*for*

$$\mathfrak{R}_{n,\delta,\mathbb{H}_0} = \left(n - Tr(\Sigma(\Sigma_t')^{-1}) + \ln\left(\frac{|\Sigma|}{|\Sigma_t'|}\right)\right) + (\lambda_{max} + \lambda'_{max} - \lambda_{min} - \lambda'_{min}) V^2 n \sqrt{0.5 \ln(2/\delta)}.$$

*Proof.* From the Hoeffding bound, for $2\mathfrak{L}$ bounded by $[a, b]$,

$$\mathbb{P}(2\mathfrak{L} - \mathbb{E}(2\mathfrak{L}|\mathbb{H}_0) \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{(b-a)^2}\right).$$

Letting right hand side as $\delta/2$ and rewriting $\epsilon$ with $\delta$,

$$\epsilon = (b-a)\sqrt{1/2\ln(2/\delta)}.$$

Thus

$$\mathbb{P}(2\mathfrak{L} \geq \mathbb{E}(2\mathfrak{L}|\mathbb{H}_0) + (b-a)\sqrt{1/2\ln(2/\delta)}) \leq \delta/2.$$

To calculate $\mathbb{E}(2\mathfrak{L}|\mathbb{H}_0)$, recall that $2\mathfrak{L} = X^T\Sigma^{-1}X - X^T\Sigma'^{-1}X + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right)$.

$$\begin{aligned}
&\mathbb{E}(X^T\Sigma^{-1}X - X^T\Sigma'^{-1}X + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right)|\mathbb{H}_0) \\
&= \mathbb{E}(X^T\Sigma^{-1}X|\mathbb{H}_0) - \mathbb{E}(X^T\Sigma'^{-1}X|\mathbb{H}_0) + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right) \\
&= \mathbb{E}(Tr(X^T\Sigma^{-1}X)|\mathbb{H}_0) + \mathbb{E}(Tr(X^T\Sigma^{-1}X)|\mathbb{H}_0) + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right) \\
&= \mathbb{E}(Tr(XX^T\Sigma^{-1})|\mathbb{H}_0) + \mathbb{E}(Tr(XX^T\Sigma^{-1})|\mathbb{H}_0) + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right) \\
&= Tr(\mathbb{E}(XX^T\Sigma^{-1}|\mathbb{H}_0)) + Tr(\mathbb{E}(XX^T\Sigma^{-1}|\mathbb{H}_0)) + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right)
\end{aligned}$$

Here we used the fact that the $X^T\Sigma^{-1}X$ is a scalar and that the trace has the cyclic property. As $X$ is under the null hypothesis, $\mathbb{E}(XX^T) = \Sigma$.

$$\begin{aligned}
&Tr(\mathbb{E}(XX^T\Sigma^{-1}|\mathbb{H}_0)) + Tr(\mathbb{E}(XX^T\Sigma^{-1}|\mathbb{H}_0)) + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right) \\
&= Tr(\Sigma\Sigma^{-1}) + Tr(\Sigma\Sigma^{-1}) + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right) \\
&= n + Tr(\Sigma\Sigma^{-1}) + \ln\left(\frac{|\Sigma|}{|\Sigma'|}\right)
\end{aligned}$$

Now, to find the interval $b-a$ for $2\mathfrak{L}$ we use Lemma 3.3.1. As the last term of $2\mathfrak{L}$ is constant over $X$, the upper bound of $2\mathfrak{L}$ can be calculated from the maximum of $X^T\Sigma^{-1}X$ minus the minimum of $X^T\Sigma'^{-1}X$. The lower bound of $2\mathfrak{L}$ can be calculated in a similar way. Writing the eigenvalues of $\Sigma^{-1}$ and $\Sigma'^{-1}$ as $\lambda$ and $\lambda'$ respectively, we can summarize the error bound as

$$\mathbb{P}(2\mathfrak{L} \geq \mathfrak{R}_{n,\delta,\mathbb{H}_0}|\mathbb{H}_0) \leq \delta/2,$$

for

$$\mathfrak{R}_{n,\delta,\mathbb{H}_0} = \left(n - Tr(\Sigma(\Sigma_t')^{-1}) + \ln\left(\frac{|\Sigma|}{|\Sigma_t'|}\right)\right) + (\lambda_{max} + \lambda'_{max} - \lambda_{min} - \lambda'_{min})V^2 n\sqrt{0.5\ln(2/\delta)}.$$

□

**Lemma 3.3.3** (Type II Error)**.** *When the n number of sequential data follow the alternative hypothesis as defined in Section 3.3 and bounded by $[-V, V]$, the error rate that the likelihood ratio test to wrongly detect a change is bounded as follows.*

$$\mathbb{P}(2\mathfrak{L} \leq \mathfrak{R}_{n,\delta,\mathbb{H}_1}|\mathbb{H}_1) \leq \delta/2,$$

*for*

$$\mathfrak{R}_{n,\delta,\mathbb{H}_0} = \left( Tr(\Sigma_t^{'}(\Sigma)^{-1}) - n - \ln\left(\frac{|\Sigma|}{|\Sigma_t^{'}|}\right)\right) + (\lambda_{max} + \lambda'_{max} - \lambda_{min} - \lambda'_{min})V^2 n\sqrt{0.5\ln(2/\delta)}.$$

*Proof.* The proof follows similar procedure as Lemma 3.3.2. □

When the specific condition is met, we can bound the conditional detection error probability as follows.

**Theorem 3.3.2.** *For $\mathfrak{R}_{n,\delta,\mathbb{H}_0}$, $\mathfrak{R}_{n,\delta,\mathbb{H}_1}$ in Lemmas 3.3.2 and 3.3.3, when $\mathfrak{R}_{n,\delta,\mathbb{H}_1} \geq \mathfrak{R}_{n,\delta,\mathbb{H}_0}$ and $\mathfrak{R}_{n,\delta,\mathbb{H}_0} \leq \mathfrak{R}_\delta \leq \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, the conditional detection error probability is bounded as*

$$\varphi_n(\mathfrak{T}) = \mathbb{P}(2\mathfrak{L} \geq \mathfrak{R}_\delta|\mathbb{H}_0) + \max_{t\in\mathcal{C}_n} \mathbb{P}(2\mathfrak{L} \leq \mathfrak{R}_\delta|\mathbb{H}_{1,t}) \leq \delta.$$

*Proof.* When we set the threshold to be $\mathfrak{R}_\delta \geq \mathfrak{R}_{n,\delta,\mathbb{H}_0}$, I error of the test can be guaranteed to be bounded from Lemma 3.3.2. Similarly, if we set the threshold to be $\mathfrak{R}_\delta \leq \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, type II error of the test can be guaranteed to be bounded from Lemma 3.3.3. The result directly follows. □

Using Theorem 3.3.2, the statistical error probability of the proposed test can be bounded at any rate for a covariance structural break of arbitrary kernels. In other words, the proposed test detects covariance structural breaks statistically correctly without specifying kernel types, if specific conditions are met.

There are three possible cases of inequalities between $\mathfrak{R}_{n,\delta,\mathbb{H}_0}$ and $\mathfrak{R}_{n,\delta,\mathbb{H}_1}$. In case $\mathfrak{R}_{n,\delta,\mathbb{H}_0} > \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, no threshold exists that satisfying the condition $\mathfrak{R}_{n,\delta,\mathbb{H}_0} \leq \mathfrak{R}_\delta \leq \mathfrak{R}_{n,\delta,\mathbb{H}_1}$. Thus it cannot be guaranteed either type I or type II errors. In case $\mathfrak{R}_{n,\delta,\mathbb{H}_0} = \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, only one threshold exists which guarantees both bounded type I and type II errors. Finally, in case $\mathfrak{R}_{n,\delta,\mathbb{H}_0} < \mathfrak{R}_{n,\delta,\mathbb{H}_1}$, the thresholds between $\mathfrak{R}_{n,\delta,\mathbb{H}_0}$ and $\mathfrak{R}_{n,\delta,\mathbb{H}_1}$ guarantee bounded type I and type II errors. The shaded area in Figure 3.4 shows the range of such thresholds.

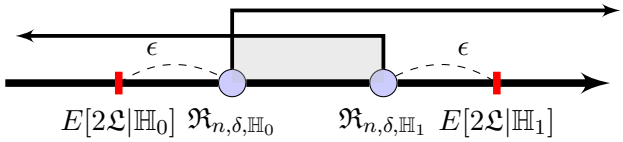
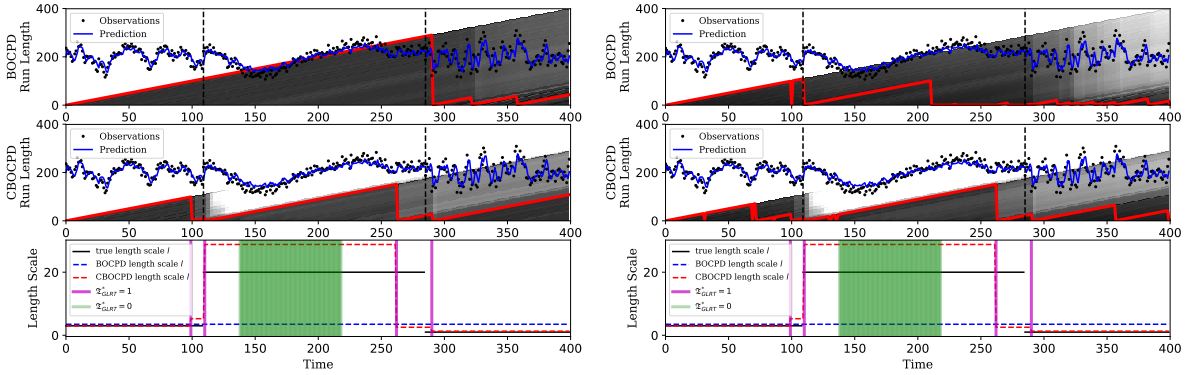



Figure 3.4: A horizontal line representing the range of thresholds guaranteeing the bounded type I error with the right-pointing arrow and the range of thresholds guaranteeing the bounded type II error under the alternative hypothesis. The shaded area is the range of threshold which can guarantee both bounded type I error and the bounded type II error.

# Chapter IV

# Confirmatory Bayesian Online Change Point Detection

In this section we propose an improved version of conventional BOCPD algorithm by leveraging the statistical hypothesis tests. We will explain how the algorithm works and further discuss the theoretical analysis of the algorithm.




(a) Timescale $\lambda = 200$ (b) Timescale $\lambda = 25$

Figure 4.1: Figures 4.1a and 4.1b represent the behaviors of the conventional BOCPD algorithm and the proposed CBOCPD algorithm. The top-most plot shows the run length distribution computed by the conventional BOCPD algorithm. And the middle plot shows the run length distribution of the CBOCPD algorithm. The bottom plot shows the results of statistical tests and the length scale hyperparameters where true hyperparameter is represented with black line and the trained hyperparameter is represented with the dashed black line.

## 4.1 Confirmatory BOCPD

We propose an online change detection algorithm, called CBOCPD in Algorithm 1. Based on BOCPD, CBOCPD compensates the distribution of change point by loosen the assumption that the run length solely depends on the run length of the previous time step, not the observed

**Algorithm 1** Confirmatory BOCPD

1: $m, \delta \leftarrow$ half window size, error bound for likelihood ratio test
2: $\mathbb{P}(X_0) \leftarrow \mathcal{N}(\mu_{prior}, \sigma^2_{prior})$
3: **for** $t \in [1, T]$ **do**
4: $\quad \mathfrak{R}_{\delta,\mathbb{H}_0}, \mathfrak{R}_{\delta,\mathbb{H}_1} \leftarrow$ thresholds for $\mathfrak{T}^0_{GLRT}, \mathfrak{T}^1_{GLRT}$
5: $\quad H \leftarrow H_{const}$
6: $\quad$ **if** $m < t < T - m$ **then**
7: $\quad\quad$ set window $W = X_{t-m:t+m}$
8: $\quad\quad 2\mathfrak{L}_\tau \leftarrow$ the likelihood ratio between $\mathbb{H}_0$ and $\mathbb{H}_{1,\tau}$ with data $W$
9: $\quad\quad \tau^*,\ 2\mathfrak{L} = \text{argmax}_{\tau \in \mathcal{C}_W} 2\mathfrak{L}_\tau,\ \max_{\tau \in \mathcal{C}_W} 2\mathfrak{L}_\tau$
10: $\quad\quad$ **if** $\mathfrak{T}^1_{GLRT} = 1$ and $\mathfrak{T}^0_{GLRT} = 1$ and $\tau^* = t$ **then**
11: $\quad\quad\quad H \leftarrow 1 - \delta$
12: $\quad\quad$ **else if** $\mathfrak{T}^1_{GLRT} = 0$ and $\mathfrak{T}^0_{GLRT} = 0$ **then**
13: $\quad\quad\quad H \leftarrow \delta$
14: $\quad\quad$ **end if**
15: $\quad$ **end if**
16: $\quad \pi^{(r)}_t \leftarrow \mathbb{P}(X_t | X^{(r)}_{t-1})$
17: $\quad \mathbb{P}(r_t = r_{t-1}+1, X_{1:t}) \leftarrow \mathbb{P}(r_{t-1}, X_{1:t-1}) \pi^{(r)}_t (1-H)$
18: $\quad \mathbb{P}(r_t = 0, X_{1:t-1}) \leftarrow \sum_{r_{t-1}} \mathbb{P}(r_{t-1}, X_{1:t-1}) \pi^{(r)}_t H$
19: $\quad \mathbb{P}(X_{1:t}) \leftarrow \sum_{r_t} \mathbb{P}(r_t, X_{1:t})$
20: $\quad \mathbb{P}(r_t | X_{1:t}) \leftarrow \mathbb{P}(r_t, X_{1:t}) / \mathbb{P}(X_{1:t})$
21: $\quad \mathbb{P}(X_{t+1} | X_{1:t}) \leftarrow \sum_{r_t} \mathbb{P}(X_{t+1} | X^{(r)}_t) \mathbb{P}(r_t | X_{1:t})$
22: **end for**

data. On the contrary, we claim that the run length is also affected by the observed data. We plug the proposed statistical test in here to adjust this constant hazard function. In Equation (2.10), $\mathbb{P}(r_t = 0 | r_{t-1}, X^{(r)}_{t-1})$ is assigned based on the likelihood ratio tests which are proposed in the prvious chapter. In Algorithm 1, the first two lines initialize the parameters $m$ and $\delta$, where $m$ indicates the half of the window size and $\delta$ indicates the designated error bound of the likelihood ratio test. In lines 3–13, Equation (2.11) is altered with

$$\mathbb{P}(r_t = 0 | r_{t-1}, X^{(r)}_{t-1}) = \begin{cases} 1 - \delta, & \mathfrak{T}^*_{GLRT} = 1 \text{ and } \tau^* = t \\ \delta, & \mathfrak{T}^*_{GLRT} = 0 \\ H_{const}, & \text{otherwise.} \end{cases}$$

By default, we set $\mathbb{P}(r_t = 0 | r_{t-1}, X^{(r)}_{t-1})$ with $H_{const}$ as in conventional BOCPD. For simplicity we denote $\mathbb{P}(r_t = 0 | r_{t-1}, X^{(r)}_{t-1})$ as $H$ in the algorithm. We use two likelihood ratio tests with thresholds defined in line 4, $\mathfrak{T}^0_{GLRT} = \mathbb{I}\left(2\mathfrak{L} \geq \tilde{\mathfrak{R}}_{\delta,\mathbb{H}_0}\right)$ and $\mathfrak{T}^1_{GLRT} = \mathbb{I}\left(2\mathfrak{L} \geq \tilde{\mathfrak{R}}_{\delta,\mathbb{H}_1}\right)$. With the first test, we test to reject or fail to reject the null hypothesis. In contrary, if the result of the second test is 0, we reject the alternative hypothesis and we fail to reject the alternative

hypothesis if the result is 1. The reason why we use two tests is that, we only know when to 'reject' a hypothesis but not when to 'accept' a hypothesis as 'fail to reject' does not always infer acceptance. Thus it makes our decision stronger if the results of two tests meet. For example if we fail to reject the null hypothesis and we reject the alternative hypothesis, we have more confidence to say there is no change. We simplify the results by unifying $\mathfrak{T}^*_{GLRT} = 1$ if $\mathfrak{T}^1_{GLRT} = 1$ and $\mathfrak{T}^0_{GLRT} = 1$, and similarly $\mathfrak{T}^*_{GLRT} = 0$ if $\mathfrak{T}^1_{GLRT} = 0$ and $\mathfrak{T}^0_{GLRT} = 0$. Theoretically, the thresholds can be computed as provided in Section 3.2. However the empirical thresholds are used in this section since the theoretically calculated thresholds in Lemmas 3.3.2 and 3.3.3 are not tight enough to use in practice. We use the window-based approach at $t$ with $W = X_{t-m:t+m}$ when applying the statistical likelihood ratio tests to reduce the computational complexity and to be worked in an online manner. For every possible candidate of change points, we compute the likelihood ratio given the candidate as in line 8. In line 9, we name $\tau^*$ as the time step which maximizes the likelihood ratio among the possible candidates in the window, $\tau^* = \text{argmax}_{\tau \in \mathcal{C}_W} 2\mathfrak{L}_\tau$. Then $2\mathfrak{L}$ is the likelihood ratio at $\tau^*$ which is used for the test statistic. Here $\mathcal{C}_W \subseteq \{t-m, ..., t+m\}$ denotes a set of possible candidates of change points in the window. Based on the test results we modify $H$. If the likelihood ratio is maximized at the middle of the window and if the likelihood ratio at that point passes the tests, we decide that the time point $t$ is a change point and set $\mathbb{P}(r_t = 0 | r_{t-1}, X^{(r)}_{t-1}) = 1 - \delta$, which amplifies the possibility of changes in the BOCPD framework. On the contrary, when both tests does not passes, i.e., $\mathfrak{T}^*_{GLRT} = 0$, we take there is no change and reduce the prior of the occurrence of changes in the BOCPD algorithm. This is the reason this algorithm is named as *Confirmatory* BOCPD. There is one more condition $\tau^* = t$, that the likelihood is maximized at the middle of the window, is set to prevent situations in which the same time point is detected in multiple consecutive windows. The rest of the algorithm follows similar to the conventional BOCPD framework [18].

## 4.2 Theoretical Analysis of CBOCPD

In this section we discuss on the sufficient conditions for CBOCPD to provide the lower prediction error than the conditional BOCPD. The prediction error in this section is defined as the expectation of the absolute difference between the true predictive mean and the predictive mean from BOCPD or CBOCPD algorithm at the detected change point $t$. We focus the prediction performance at the change point as how to handle the change point highly affects to the overall performance. We will write the expected value of $X_t$ under BOCPD and CBOCPD as $\mathbb{E}_{BO}[X_t | X_{1:t-1}]$ and $\mathbb{E}_{CBO}[X_t | X_{1:t-1}]$, respectively. Further we define $\alpha_i = \mathbb{P}_{BO}(r_{t-1} = i | X_{1:t-1})$ under BOCPD and $\beta_i = 1 - \mathbb{P}_{CBO}(r_{t-1} = i | X_{1:t-1})$ under CBOCPD.

#### Under the Existence of a Change

In this section we examine a non-stationary situation in which there exist a change. We will find conditions where CBOCPD performs at least equal to BOCPD.

**Theorem 4.2.1.** *Consider BOCPD algorithm in Section 2.2 and CBOCPD algorithms in Section 4.1 where two statistical tests $\mathfrak{T}^0_{GLRT}$ and $\mathfrak{T}^1_{GLRT}$ are used. The type II error of $\mathfrak{T}^0_{GLRT}$ is denoted with $\delta_0^{II}$ and the type II error of $\mathfrak{T}^1_{GLRT}$ is $\delta_1^{II}$. Suppose there is a change point at time $t$ with the prior mean of $\mu_1$ and suppose it satisfies*

$$\forall i \in [0, t-1], 0 \leq |\mathbb{E}[X_t|\emptyset] - \mathbb{E}[X_t|X_{i:t-1}]| \leq \epsilon_U$$

*and*

$$\exists i \in [0, t-1], |\mathbb{E}[X_t|\emptyset] - \mathbb{E}[X_t|X_{i:t-1}]|\alpha_i - \sum_{i \neq j} |\mathbb{E}[X_t|\emptyset] - \mathbb{E}[X_t|X_{j:t-1}]\alpha_j| = \epsilon_L \geq 0$$

*where $\mathbb{E}[X_t|\emptyset] = \mu_1$ denotes the expecation of $X_t$ given no observed data which is equal to the prior mean. If the following inequality is satisfied*

$$\frac{\epsilon_U}{\epsilon_L} \leq \alpha_0 \left(1 + \frac{(1-\delta_0^{II})(1-\delta_1^{II})}{\delta_0^{II}\delta_1^{II}}\right)$$

*with $\alpha_0$, the probability that the run length not equal to zero, then the expected absolute error of CBOCPD at $t$ is less than or equal to the expected absolute error of BOCPD as stated below*

$$\mathbb{E}[|\mu_1 - \mathbb{E}_{BO}[X_t|X_{1:t-1}]|] \geq \mathbb{E}[|\mu_1 - \mathbb{E}_{CBO}[X_t|X_{1:t-1}]|].$$

The Theorem 4.2.1 mainly refers that the performance of CBOCPD is no worse than BOCPD with several conditions. The first condition indicates that the the absolute prediction error should be bounded. It is acquired from the assumption that the absolute prediction error with an incorrect run length should be greater than 0. The second condition refers that one conditional prediction error is greater than the weighted sum of the other conditional prediction error over the possible run lengths. This condition is set to guarantee that the lower bound of the prediction error of BOCPD to be strictly greater than the prediction error of CBOCPD. Lastly, if the ratio between the upper bound and the lower bound of the absolute prediction error in the first condition is upper bounded, then the expectation of the absolute error of CBOCPD is less than or equal to the BOCPD.

### Under the Absence of a Change

Here, we explore the case in which there does not exist a change.

**Theorem 4.2.2.** *Consider BOCPD algorithm in Section 2.2 and CBOCPD algorithms in Section 4.1 where two statistical tests $\mathfrak{T}^0_{GLRT}$ and $\mathfrak{T}^1_{GLRT}$ are used. The type I error of $\mathfrak{T}^0_{GLRT}$ is denoted with $\delta_0^I$ and the type I error of $\mathfrak{T}^1_{GLRT}$ is $\delta_1^I$. Suppose there is a statistically justified non change at time $t$ with the prior mean of $\mu_2$ and suppose it satisfies*

$$\forall i \in [0, t-1], \epsilon_L \leq |\mathbb{E}[X_t|X_{1:t-1}] - \mathbb{E}[X_t|X_{i:t-1}]| \leq \epsilon_U,$$

*and*

$$\exists i \in [0, t-2], |\mathbb{E}[X_t|X_{1:t-1}] - \mathbb{E}[X_t|X_{i:t-1}]|\alpha_i - \sum_{i \neq j} |\mathbb{E}[X_t|X_{1:t-1}] - \mathbb{E}[X_t|X_{j:t-1}]\alpha_j| = \epsilon_L \geq 0$$

*If the following inequality is satisfied*

$$\frac{\epsilon_U}{\epsilon_L} \leq \frac{(1-\delta_0^I)(1-\delta_1^I) + \delta_0^I\delta_1^I}{\beta_{t-1}(1-\delta_0^I)(1-\delta_1^I) + \delta_0^I\delta_1^I}$$

*with $\alpha_{t-1}$, the probability that the run length in BOCPD not equal to $t-1$, and $\beta_{t-1}$, the probability that the run length in CBOCPD not equal to $t-1$, then the expected absolute error of CBOCPD at $t$ is less than or equal to the expected absolute error of BOCPD as stated below*

$$\mathbb{E}[|\mu_2 - \mathbb{E}_{BO}[X_t|X_{1:t-1}]|] \geq \mathbb{E}[|\mu_2 - \mathbb{E}_{CBO}[X_t|X_{1:t-1}]|].$$

Similar to the Thereom 4.2.1, the Theorem 4.2.2 refers that the performance of CBOCPD is no worse than BOCPD under the situation where there is no change with several conditions. The first condition indicates that the the absolute prediction error should be bounded. Here the difference from condition of the Theorem 4.2.1 is that the prediction error is computed compared to the predictive mean given all the previously observed data. This is because the predictive mean becomes more accurate as we observe more and more data. The second condition is similarly defined as the second condition of Theorem 4.2.1. From the last condition, if the ratio between the upper bound and the lower bound of the absolute prediction error in the first condition is upper bounded, then the expectation of the absolute error of CBOCPD is less than or equal to the BOCPD.

From Theorems 4.2.1 and 4.2.2, we can say that the proposed CBOCPD algorithm performs as well as the BOCPD algorithm in either cases where there is a change or no change.

# Chapter V

# Experimental Results

## 5.1 Synthetic data

| | LEN-CHANGE | | VAR-CHANGE | |
|---|---|---|---|---|
| Method | NLL | MSE | NLL | MSE |
| BOCPD | 1.04±0.36 | 0.40±0.16 | 2.18±0.99 | 0.83±0.37 |
| CBOCPD | **0.51±0.19** | 0.44±0.21 | **0.71±0.16** | **0.43±0.11** |

Table 5.1: Comparison of BOCPD,and CBOCPD over NLL and MSE on synthetic datasets.

In this section we show experimental results on synthetic data. We generate synthetic data with intentional change points in times series data. Typically we used total size of the sequential data as $T = 200$, with two randomly generated change points. The first change point is randomly selected from the interval [50,80] and the second change point is randomly selected from the interval [150, 180]. We changed two hyperparameters of Squared Exponential kernel, variance and lengthscale respectively. Variance indicates the amplitude of variation of the regression function, and lengthscale indicates the smoothness of the function. For the variance, we changed from 1.0 to 4.0 then changed from 4.0 to 0.3. For the lengthscale, it is changed from 3.0 to 20.0 after the first change point then reduced to 1.0 after the second change point. Figure 5.1 shows the two examples of lengthscale change and variance change. We can see that CBOCPD algorithm catches true change points which BOCPD algorithm missed in both cases. We repeated the experiment for 10 times and computed the average of negative log likelihood and the mean squared error. The result is summarized in Table 5.1.

For NLL, CBOCPD clearly shows better performance compared to BOCPD. BOCPD shows somewhat higher performance on mean squared error, while it is hard to see as significant considering the margin of error. Qualitatively, CBOCPD captures change points that can be missed by BOCPD but also corrects false change points as shown in Figure 5.1. Figure 5.1 shows how BOCPD and CBOCPD works with different settings of timescale parameter. The

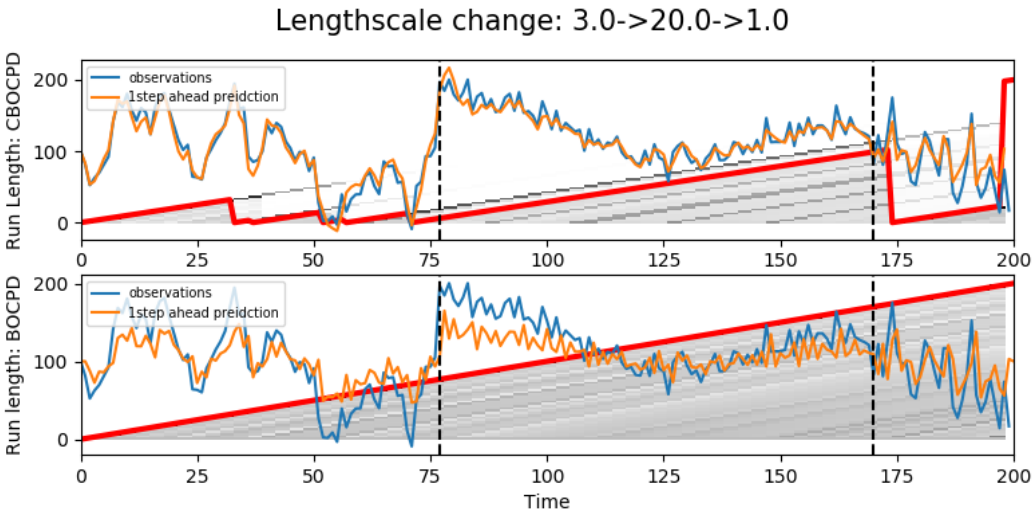
| | Gazebo:Env1 | | Gazebo:Env2 | | Gazebo:Env3 | |
|---|---|---|---|---|---|---|
| Method | NLL | MSE | NLL | MSE | NLL | MSE |
| BOCPD | 2.07±0.51 | 0.14±0.05 | 2.24±0.48 | 0.57±0.26 | 0.28±0.12 | 0.11±0.03 |
| CBOCPD | **-0.31±0.34** | **0.11±0.04** | **0.69±0.36** | **0.45±0.19** | **-0.99±0.47** | 0.10±0.04 |

Table 5.2: The NLL and MSE results of BOCPD, and CBOCPD on the Gazebo robot simulator with three change environments.

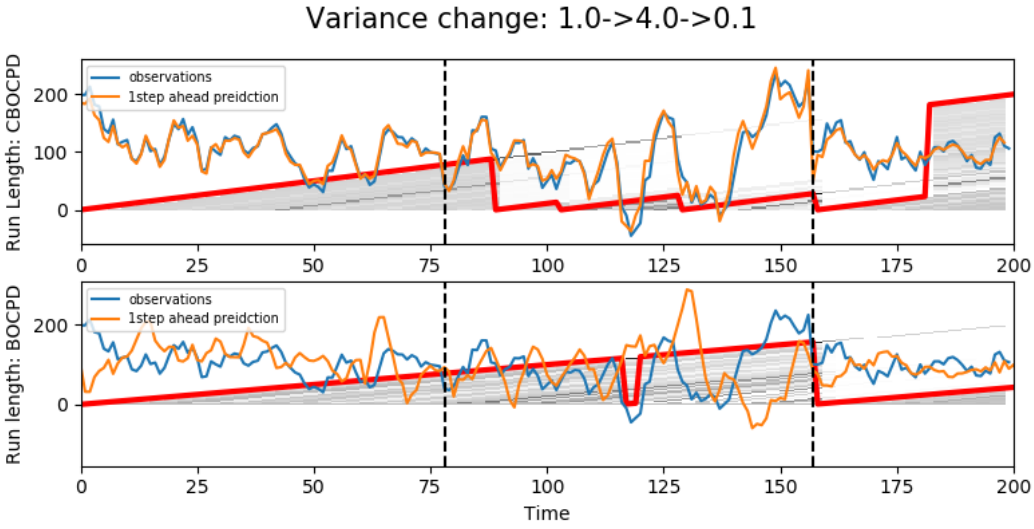
timescale parameter works as a prior of the probability of a change point. When timescale is small, it is more likely that change point occurs and vice versa. Conventional BOCPD algorithm is vulnerable to this parameter setting as we can see in the figure. When the timescale is set too large, BOCPD algorithm misses true change points. On the other hand, when the timescale is set too small, BOCPD algorithm alarms too much change points. CBOCPD algorithm helps to adjust BOCPD with the improper parameter setting. In Figure 4.1a, CBOCPD captures the first change point that BOCPD missed. In Figure 4.1a, CBOCPD reduces change points that are false alarmed by BOCPD which are statistically confirmed non-changes.

## 5.2 Robot Simulation Data

We further conduct experiments on the robot simulation data. For the experimental setting, we use Gazebo 8 for the robot simulator and Pioneer3AT robot is used through the experiments. The robot is moved in the environments where the properties of the ground changes. We then gathered the position data of the robot while the robot is moving. We prepared three types of environments of the ground by modifying the height-map of the simulator. Each type of environment is captured in Figure 5.3. In the first environment, the robot is moving from the plane ground to the bumpy ground with many bumps. Then the robot is moving from the bumpy ground to the more coarse bumpy ground with smaller but more dense bumps (Env2). In the final environment the robot is moving from the second bumpy ground back to the plane ground. For the experiments we use altitude of the robot, 'z'-axis data, to detect the change of the ground. Table 5.2 summarizes the results of the experiments on the Gazebo robot simulator. The table shows that CBOCPD outperforms the conventional BOCPD in all three environments with respect to the NLL (Negative Log Likelihood). CBOCPD also shows improved performance in MSE While it is not as significant in Env3 as in the other environments.

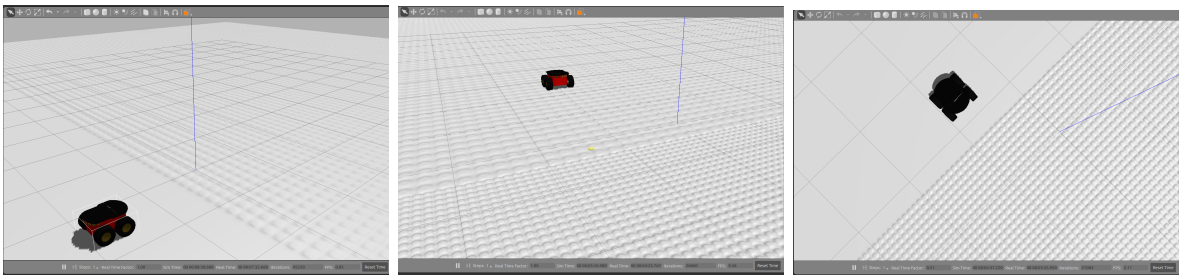



(a) Example of lengthscale change




(b) Example of variance change

Figure 5.1: The run length distribution from the CBOCPD and BOCPD algorithms on the synthetic datasets with two changes in hyperparameters. Figure 5.1a is the case where length scale is increasing from 3.0 to 20.0 and decreasing to 1.0. Figure 5.1b is the case where the variance is increasing from 1.0 to 4.0 and decreasing to 0.1. Dashed black line indicates the true change points and red line shows the most probable run length.

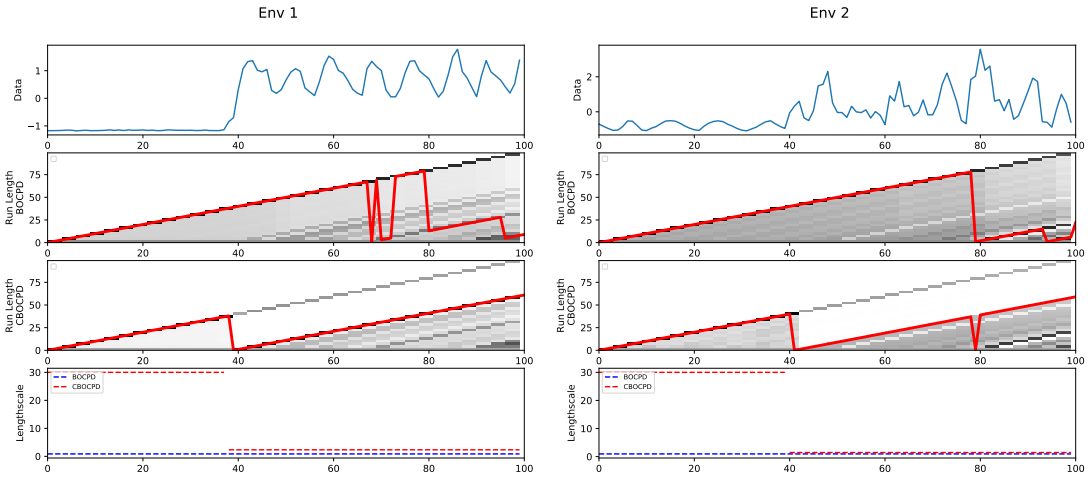


(a) Plane ground to Bumpy ground 1 (Env1)

(b) Bumpy ground 1 to Bumpy ground 2 (Env2)

(c) Bumpy ground 2 to Plane ground (Env3)

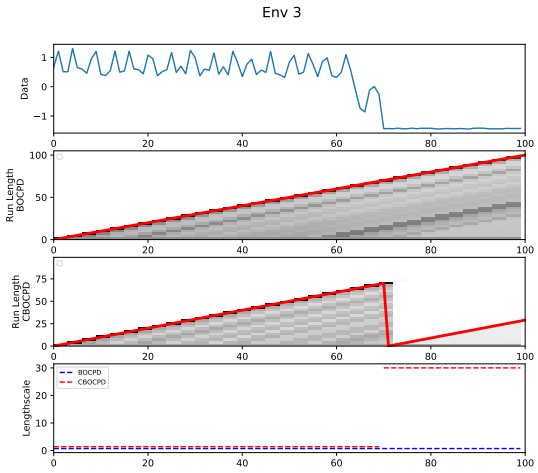
Figure 5.2: Gazebo robot simulation environments. In each environment the ground is changing. In the first environment (Env1) the ground is changing 'Plane ground' –> 'Bumpy ground 1'. The second plot shows the environment (Env2), where the ground is changing 'Bumpy ground 1' –> 'Bumpy ground 2'. In the right-most plot (Env3), the environment is changing 'Bumpy ground 2' –> 'Plane ground'.




(a) Env1




(b) Env2




(c) Env3

Figure 5.3: Results of BOCPD and CBOCPD on Gazebo robot simulation data. The top plot shows the z-directional data from each environment. The result of BOCPD is placed below the original data plot. The third plot shows the result of CBOCPD and the bottom plot shows estimated hyperparameters.

# Chapter VI

# Conclusion

We present a novel likelihood ratio tests for detecting covariance structural breaks in the covariance of the GP. We devise an online CPD algorithm, called Confirmatory BOCPD, which improves BOCPD by confirming changes or non-changes with statistical hypothesis tests. Although our work have shown theoretically correct threshold for tests, it is yet loose to use in practice. It could be further reduced with more conditions if needed. In this thesis we focused on abrupt change in the covariance structure of uni-variate time series. The future work could investigate on the likelihood ratio test to multi-variate time series or the likelihood ratio test for smoothly changing covariance.

# References

[1] A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher, "A critical synthesis of remotely sensed optical image change detection techniques," *Remote Sensing of Environment*, pp. 1–14, 2015.

[2] S. P. Panda and A. K. Nayak, "Automatic speech segmentation in syllable centric speech recognition system," *International Journal of Speech Technology*, pp. 9–18, 2016.

[3] G. Manogaran and D. Lopez, "Spatial cumulative sum algorithm with big data analytics for climate change detection," *Computers and Electrical Engineering*, pp. 207–221, 2018.

[4] I. Cleland, M. Han, C. Nugent, H. Lee, S. McClean, S. Zhang, and S. Lee, "Evaluation of prompted annotation of activity data recorded from a smart phone," *Sensors*, pp. 15 861–15 879, 2014.

[5] G. C. Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, pp. 591–605, 1960.

[6] B. T. Ewing and F. Malik, "Volatility spillovers between oil prices and the stock market under structural breaks," *Global Finance Journal*, pp. 12–23, 2016.

[7] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "$l_1$ trend filtering," *SIAM review*, pp. 339–360, 2009.

[8] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," *Annals of the Institute of Statistical Mathematics*, pp. 699–746, 2008.

[9] M. Eric, F. R. Bach, and Z. Harchaoui, "Testing for homogeneity with kernel fisher discriminant analysis," in *Proceedings of the Neural Information Processing Systems*, 2008, pp. 609–616.

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, pp. 723–773, 2012.

[11] S. Li, Y. Xie, H. Dai, and L. Song, "M-statistic for kernel change-point detection," in *Proceedings of the Neural Information Processing Systems*, 2015, pp. 3366–3374.

[12] H. Chernoff and S. Zacks, "Estimating the current mean of a normal distribution which is subjected to changes in time," *The Annals of Mathematical Statistics*, pp. 999–1018, 1964.

[13] E. Gombay, L. Horváth, and M. Husková, "Estimators and tests for change in variances," *Statistics and Risk Modeling*, pp. 145–160, 1996.

[14] O. Isupova, "Machine learning methods for behaviour analysis and anomaly detection in video," Ph.D. dissertation, University of Sheffield, 2017.

[15] H. Keshavarz, C. Scott, and X. Nguyen, "Optimal change point detection in gaussian processes," *Journal of Statistical Planning and Inference*, pp. 151–178, 2018.

[16] D. Barry and J. A. Hartigan, "A bayesian analysis for change point problems," *Journal of the American Statistical Association*, pp. 309–319, 1993.

[17] X. Xuan and K. Murphy, "Modeling changing dependency structure in multivariate time series," in *Proceedings of the International Conference on Machine Learning*, 2007, pp. 1055–1062.

[18] R. P. Adams and D. J. MacKay, "Bayesian online changepoint detection," *arXiv preprint arXiv:0710.3742*, 2007.

[19] R. Garnett, M. A. Osborne, S. Reece, A. Rogers, and S. J. Roberts, "Sequential bayesian prediction in the presence of changepoints and faults," *The Computer Journal*, pp. 1430–1446, 2010.

[20] Y. Saatçi, R. D. Turner, and C. E. Rasmussen, "Gaussian process change point models," in *Proceedings of the International Conference on Machine Learning*, 2010, pp. 927–934.

[21] T. Kim and J. Choi, "Reading documents for bayesian online change point detection," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2015, pp. 1610–1619.

# Acknowledgements

First of all, I would like to express my special thanks of gratitude to my advisor Jaesik Choi who teaches me not only the knowledge in Machine Learning or how to write a paper but more importantly the way how to become a researcher. Secondly I would like to appreciate Kyowoon Lee, for all the countless efforts to build and develop the work with me besides the gentle discussions. I would also like to give many thanks to my colleagues in Statistical Artificial Intelligence Lab. who support me by making comments, having discussions, giving warm encourages in many other ways. I am really thankful to them. For the last but not the least, I would like to thank my parents and friends who helped me a lot emotionally and physically to finish my thesis.

# Appendix

## A Proofs for Chapter IV

*Proof of Theorem 4.2.1.* Let's define the gain of CBOCPD over BOCPD as

$$|\mathbb{E}[x_t|\emptyset] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| - |\mathbb{E}[x_t|\emptyset] - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]|.$$

In the case $\mathfrak{T}^*_{GLRT} = 1$, the gain is written as

$$\begin{aligned}
&|\mathbb{E}[x_t|\emptyset] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| - |\mathbb{E}[x_t|\emptyset] - \mathbb{E}[x_t|\emptyset]| \\
&\quad = |\mathbb{E}[x_t|\emptyset] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| \\
&\quad = \left| \sum_{r_{t-1}=0}^{t-1} \mathbb{E}[x_t|\emptyset]\mathbb{P}_{BO}(r_{t-1}|x_{1:t-1}) - \sum_{r_{t-1}=0}^{t-1} \mathbb{E}[x_t|x^{(r)}_{t-1}]\mathbb{P}_{BO}(r_{t-1}|x_{1:t-1}) \right| \\
&\quad = \left| \sum_{r_{t-1}=1}^{t-1} \left( \mathbb{E}[x_t|\emptyset] - \mathbb{E}[x_t|x^{(r)}_{t-1}] \right) \mathbb{P}_{BO}(r_{t-1}|x_{1:t-1}) \right| \\
&\quad \geq \max_i |\mathbb{E}[x_t|\emptyset] - \mathbb{E}[x_t|x^{(i)}_{t-1}]|\alpha_i - \sum_{j\neq i} |\mathbb{E}[x_t|\emptyset] - \mathbb{E}[x_t|x^{(j)}_{t-1}]|\alpha_j \\
&\quad \geq |\mathbb{E}[x_t|\emptyset] - \mathbb{E}[x_t|x^{(i^*)}_{t-1}]|\alpha_{i^*} - \sum_{j\neq i^*} |\mathbb{E}[x_t|\emptyset] - \mathbb{E}[x_t|x^{(j)}_{t-1}]|\alpha_j \\
&\quad = \epsilon_L > 0.
\end{aligned}$$

Here the last two lines are induced from the assumption that

$$\exists i \in [0, t-1], \|\mathbb{E}[X_t|\emptyset] - \mathbb{E}[X_t|X_{i:t-1}]|\alpha_i - \sum_{i\neq j} |\mathbb{E}[X_t|\emptyset] - \mathbb{E}[X_t|X_{j:t-1}]\alpha_j\| = \epsilon_L \geq 0.$$

In the case $\mathfrak{T}^*_{GLRT} = 0$, the loss of CBOCPD is written as

$$\begin{aligned}
&|\mathbb{E}[x_t|\emptyset] - \sum_{r_{t-1}=1}^{t-1} \mathbb{E}[x_t|x_{t-1}^{(r)}] \cdot \mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1})| \\
&= \left| \sum_{r_{t-1}=1}^{t-1} \mathbb{E}[x_t|\emptyset]\mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) - \sum_{r_{t-1}=1}^{t-1} \mathbb{E}[x_t|x_{t-1}^{(r)}]\mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) \right| \\
&= \left| \sum_{r_{t-1}=1}^{t-1} \left( \mathbb{E}[x_t|\emptyset] - \mathbb{E}[x_t|x_{t-1}^{(r)}] \right) \mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) \right| \\
&\leq \left| \epsilon_U \sum_{r_{t-1}=1}^{t-1} \mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) \right| = \epsilon_U.
\end{aligned}$$

The equation in the last line comes from the fact that

$\Sigma_{r_{t-1}}^{t-1} \mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) = 1$ under the CBOCPD when non-change is detected. Then, the gain is bounded as

$$|\mathbb{E}[x_t|\emptyset] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| - |\mathbb{E}[x_t|\emptyset] - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]| \geq \epsilon_L - \epsilon_U.$$

As $\mathbb{P}(\mathfrak{T}^*_{GLRT} = 1) = (1 - \delta_0^{\text{II}})(1 - \delta_1^{\text{II}})$ and $\mathbb{P}(\mathfrak{T}^*_{GLRT} = 0) = \delta_0^{\text{II}}\delta_1^{\text{II}}$ in non-stationary case, the expected gain is bounded from below as

$$\begin{aligned}
&\mathbb{E}(|\mathbb{E}[x_t|\emptyset] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| - |\mathbb{E}[x_t|\emptyset] - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]|) \\
&\quad \geq \epsilon_L(1 - \delta_0^{\text{II}})(1 - \delta_1^{\text{II}}) + (\epsilon_L - \epsilon_U)\delta_0^{\text{II}}\delta_1^{\text{II}} \geq 0
\end{aligned}$$

where the last inequality follows the assumption. Thus we can conclude that the expected gain is non-negative.

□

*Proof of Theorem 4.2.2.* Let's define the gain of CBOCPD over BOCPD as

$$|\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| - |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]|.$$

The loss of BOCPD is written as

$$
\begin{aligned}
&|\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| \\
&= \left| \sum_{r_{t-1}=0}^{t-1} \mathbb{E}[x_t|x_{1:t-1}]\mathbb{P}_{BO}(r_{t-1}|x_{1:t-1}) - \sum_{r_{t-1}=0}^{t-1} \mathbb{E}[x_t|x_{t-1}^{(r)}]\mathbb{P}_{BO}(r_{t-1}|x_{1:t-1}) \right| \\
&= \left| \sum_{r_{t-1}=0}^{t-2} \mathbb{E}[x_t|x_{1:t-1}]\mathbb{P}_{BO}(r_{t-1}|x_{1:t-1}) - \sum_{r_{t-1}=0}^{t-2} \mathbb{E}[x_t|x_{t-1}^{(r)}]\mathbb{P}_{BO}(r_{t-1}|x_{1:t-1}) \right| \\
&\geq \max_{i\in[1,t-2]} |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}[x_t|x_{t-1}^{(i)}]|\alpha_i - \sum_{j\neq i} |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}[x_t|x_{t-1}^{(j)}]|\alpha_j \\
&\geq |\mathbb{E}[x_t|x_{1:t-1} - \mathbb{E}[x_t|x_{t-1}^{(i^*)}]|\alpha_{i^*} - \sum_{j\neq i^*} |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}[x_t|x_{t-1}^{(j)}]|\alpha_j \\
&= \epsilon_L > 0.
\end{aligned}
$$

In the case $\mathfrak{T}^*_{GLRT} = 1$, the loss of CBOCPD is written as

$$|\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]| = |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}[x_t|\emptyset]| \leq \epsilon_U.$$

Then, the gain is bounded as

$$|\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| - |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]| \geq \epsilon_L - \epsilon_U.$$

In the case $\mathfrak{T}^*_{GLRT} = 0$, the loss of CBOCPD is written as

$$
\begin{aligned}
&|\mathbb{E}[x_t|x_{1:t-1}] - \sum_{r_{t-1}=1}^{t-1} \mathbb{E}[x_t|x_{t-1}^{(r)}] \cdot \mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1})| \\
&= \left| \sum_{r_{t-1}=1}^{t-1} \mathbb{E}[x_t|x_{1:t-1}]\mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) - \sum_{r_{t-1}=1}^{t-1} \mathbb{E}[x_t|x_{t-1}^{(r)}]\mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) \right| \\
&= \left| \sum_{r_{t-1}=1}^{t-2} \mathbb{E}[x_t|x_{1:t-1}]\mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) - \sum_{r_{t-1}=1}^{t-2} \mathbb{E}[x_t|x_{t-1}^{(r)}]\mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) \right| \\
&\leq \left| \epsilon_L \sum_{r_{t-1}=1}^{t-2} \mathbb{P}_{CBO}(r_{t-1}|x_{1:t-1}) \right| = \epsilon_U \cdot \beta_{t-1}.
\end{aligned}
$$

Then, the gain is bounded as

$$|\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| - |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]| \geq \epsilon_L - \epsilon_U \cdot \beta_{t-1}.$$

As $\mathbb{P}(\mathfrak{T}^*_{GLRT} = 1) = \delta_0^{\mathrm{I}}\delta_1^{\mathrm{I}}$ and $\mathbb{P}(\mathfrak{T}^*_{GLRT} = 0) = (1-\delta_0^{\mathrm{I}})(1-\delta_1^{\mathrm{I}})$ in stationary case, the expected gain is bounded from below as

$$
\begin{aligned}
&\mathbb{E}(|\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{BO}[x_t|x_{1:t-1}]| - |\mathbb{E}[x_t|x_{1:t-1}] - \mathbb{E}_{CBO}[x_t|x_{1:t-1}]|) \\
&\quad \geq (\epsilon_L - \epsilon_U)\delta_0^{\mathrm{I}}\delta_1^{\mathrm{I}} + (\epsilon_L - \epsilon_U\beta_{t-1})(1-\delta_0^{\mathrm{I}})(1-\delta_1^{\mathrm{I}}) \geq 0
\end{aligned}
$$

where the last inequality follows the assumption. Thus we can conclude that the expected gain is non-negative.

□