SOME ASPECTS OF CHANGE POINT ANALYSIS

DONGWEI WEI

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

March 2019

Abstract

Change point analysis is the process of detecting model changes within time-ordered observations. Research on change point problems started in Page (1955, 1957) and have flourished especially since the 1980s. The change point analysis has been extensively applied in quality control, finance, epidemiology, electrocardiograms and meteorology, etc.. The reason that why change point analysis is very important is that if there exists a change point, it is harmful to make a statistical analysis without any consideration of the existence of this change point and the results derived from such an analysis may be misleading.

In the first part of the dissertation, we propose two tests with the purpose of detecting change point in a sequence of independent random variables. Both the consistency and rate of convergence of the estimated change point are established. We then extend the application of the proposed test in the field of multiple change points detection problem. Simulation studies and real data analysis are given to examine the performance of our proposed methods.

In the second part of the dissertation, we propose a procedure for detecting multiple change points in a mean-shift model. We firstly convert the change point problem into a variable selection problem by partitioning the data sequence into several segments. Then, we apply a modified variance inflation factor regression algorithm to each segment in sequential order. When a segment that is suspected of containing a change point is found, we use a weighted cumulative sum to test if there is indeed a change-point in this segment. The proposed procedure is implemented in an algorithm which, compared to two popular methods via simulation studies, demonstrates satisfactory performance in terms of accuracy, stability and computational complexity. Finally, we apply our new algorithm to analyze two real data examples.

In the third part of the dissertation, our research is motivated by HIV viral dynamic studies, which have been popular in AIDS research in recent years. We jointly model HIV viral dynamics, CD4 process with measurement errors and change point model, and estimate the model parameters simultaneously via the Monte Carlo EM (MCEM) approach and hierarchical likelihood approximation approach. These approaches are illustrated in a real data example. Simulation results show that both of these two methods perform well and are much better than the commonly used naive method.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors Prof. Yuehua Wu and Prof. Wei Liu, who have provided me with many excellent ideas on change point analysis and longitudinal data analysis. Some of which are presented in this dissertation. Their helpful suggestions, important advices and constant encouragements make me feel confident to solve some novel and interesting problems. It is my pleasure to be their student.

I also wish to express my appreciation to Professor Yuejiao Fu and Professor Dong Liang as members of my supervisory committee. A lot of thanks to Prof. Hanfeng Chen and Prof. Jianguo Wang for attending my thesis oral exam and proposing some comments to promote my thesis. My appreciation also goes to all the members at Department of Mathematics and Statistics of York University for their help and assistance during my study for the PhD degree.

I place on record, my sincere gratitude to Prof. Xiaoping Shi, department of Mathematics and Statistics, Thompson Rivers University and Prof. Xiangsheng

Wang, Mathematics, University of Louisiana at Lafayette for their contribution to this thesis. I take the opportunity to record my grateful thanks to my master degree supervisors Prof. Baiqi Miao and Prof. Baisuo Jin, Statistics and Finance, University of Science and Technology of China. Thanks for their encouragement to take on this PhD study.

Special thanks to my wife Wenzhe Li, my lovely daughter Mia Wei, my parents, my mother by law and my father by law for their constant source of emotional support. Finally, my thanks also go to those who, directly or indirectly, helped me to finish my dissertation.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Change point

1.1.1 Definition of a Change Point

A change point refers to a location before and after which the observations follow two different models. The change point problem was originally stated by Page (1955, 1957) with the following test hypothesis:

- H_0 : Sample $x_1, \cdots, x_n$ have the same distribution function $F(x|\theta)$.
- H_1 : $x_1, \cdots, x_{k_0}$ come from $F(x|\theta)$ and $x_{k_0+1}, \cdots, x_n$ come from $F(x|\theta')$,

where $\theta' \neq \theta$ and k_0 is an unknown change point. For example, θ represents the mean or variance of a distribution.

Since a statistical model is not homogeneous when there is a change point, detecting all change points is very important in statistical applications. If there exists a change point, it is misleading to make a statistical analysis without any consideration

of the existence of this change point and the results derived from such an analysis may be incorrect. Change point analysis is widely used in the field of quality control, medicine, finance, environmetrics, geographics, etc.

Different change point models should be implemented for various datasets. In the literature, two types of change point models are quite popular: on one hand so called changes in distribution, referring to a genetic change of the distribution of observations before and after the change point, on the other hand changes in regression coefficient which includes the change points in mean as its special case.

As commented in Qian et al. (2014), the essential difference between the model with change points and the piecewise model is that the points of changes in the latter are specified while in the former they are unknown and need to be estimated. In addition, when fitting a data sequence by a change point model, it is even unknown whether or not change points exist, and how many there are when they exist. This uncertainty increases the difficulty and complexity in analyzing a change point model. Therefore, how to detect all of the change points has become an important task.

In Figure 1.1(a), the distribution of the observations changes from $N(0,1)$ to χ^2_3 at the location 100. We can find that the structure of observations before and after the change point are different. In Figure 1.1(b), there exist a change point at the location of 100, while observations in the left and right sides follow $N(0,1)$

and $N(1.5, 1)$, respectively. The mean of each segment is denoted by the real line. Obviously, there exists a mean shift at the location 100.

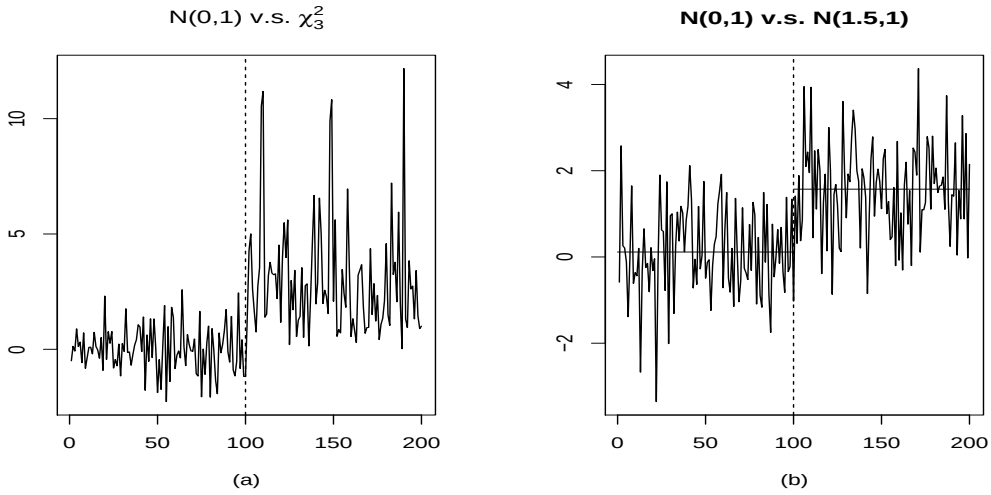

Figure 1.1: Examples of data sequence with change point. Data sequence in the left panel contains a change in distribution while the right one contains a change point in mean. The dotted lines denote the location of change points and the real lines represent the mean of the segment.

1.1.2 A Review of Some Relevant Literature

From the literature, the process of studying a change point problem is summarized as follows:

- propose a hypothesis test to test the existence of a change point while the null

hypothesis is there is no change point and the alternative hypothesis depends on the problem;

- propose a test statistic;
- explore the asymptotic distribution of the test statistic under null hypothesis and then determine the critical value;
- determine whether or not there exist a change point by comparing the value of test statistic and the critical value.

Also it is important to study the asymptotic behaviour of a change point estimator, which includes its consistency, its convergence rate as well as its asymptotic distribution.

There are quite a few methods that could be used in change point detection tests. Least-square test, Bayesian analysis test, maximum likelihood ratio test, and nonparametric test are the most widely used among them.

In Page (1957), it was assumed that the samples were generated from same distribution but with different parameters. The estimated location of change point is the one that maximizes the likelihood function of the hypothesis. And Page (1957) firstly introduced the CUSUM algorithm in change point detection problem. Basseville (1981) proposed filtered derivative algorithm and the idea behind the algo-

rithm is sample: if there is no noise, then changes in the mean translate into sharp jumps in the absolute value of the discrete derivatives of the signal.

Fisher (1958) is the first to apply the least-squares criterion for a change point problem to the best of our knowledge note that his approach does not come from likelihood maximization but rather from variance minimization. Yao and Au (1989) prove that estimated change point is consistent in probability under mild assumptions namely the continuity of the cumulative distribution function of the observations and a moment hypothesis. These assumptions are weakened further in Bai and Perron (1998) and the minimax convergence rate of $1/n$ is obtained, here n is the sample size. The least-squares estimation procedure was also shown to be consistent in the case of dependent processes (ARMA) with a single change point in Bai (1994), a work later extended for weak dependent disturbance processes (mixingales) by Bai and Perron (1998). Regarding multiple change points, Lavielle (1999); Lavielle and Moulines (2000) show the consistency of the least-squares estimate when the number of change points is known for a large class of dependent processes.

Chernoff and Zacks (1964) estimated the current mean of a normal distribution which was subjected to changes in time. The technique of using Bayesian inference was applied as a technical device to yield insight leading to simple robust procedures. A quadratic loss function was used to derive a Bayesian estimator of the current mean

for a priori probability distribution on the entire real line.

Besides the mentioned classical methods for change point detection, the wide variety of applications of change point analysis gives rise to the need of multifarious approaches to the change point problems. In industrial and health care applications, for example, the analysis usually proceeds sequentially, typically using control charts or stopping rules to perform real-time monitoring. Lai (2001) gave a review of problems in sequential analysis and its their applications to biomedicie, economics and engineering. Lai (2001) mentioned that the sequential analysis is still a vibrant subject after decades of continual development, with new ideas brought in from various fields of application.

To deal with the practical problems, different change point models were proposed and various tests were built to determine the existence of change point. For the single change point detection problem, Hinkley (1970) firstly proposed a likelihood ratio statistic to detect a change point and explored the asymptotic properties of the test statistic. The likelihood based approach was extended to the model with a change in variance within normally distributed observations by Gupta and Tang (1987). Bai (1994) and Shi et al. (2009) considered the mean shift problem and studied the convergence rate of the change point estimator. Dong et al. (2015) studied the change point in variance of measurement error and explored its convergence rate.

For the change in distributions, Hušková and Meintanis (2006a) considered a test statistic based on empirical characteristic function, and investigated the probability of type I error and the power of the test by some simulation studies. Zou et al. (2014) proposed a nonparametric maximum likelihood approach to detect multiple change points without any parametric assumption on the underlying distributions of the dataset. Thus, it is suitable for detection of any changes in the distributions.

With the increased size of dataset, there is a growing need to efficiently and accurately estimate the locations of multiple change points. Scott and Knott (1974) firstly proposed the binary segmentation which is one of the most widely used change point detection method. Another popular approach is the segment neighborhood algorithm (Auger and Lawrence, 1989), which was further explored by Bai and Perron (1998), Rigaill (2010), Hocking et al. (2017). Among these approachs, Rigaill (2010) proposeded a functional technique with $O(n \log n)$ average time complexity to prune the set of candidate change points. Here, n is the number of observations. Killick et al. (2012) developed an inequality pruning technique, which results in an efficient PELT algorithm which could reach the speed of $O(n)$. Maidstone et al. (2016) provided a clear discussion on the differences between the two pruning techniques. Moreover, many of the change point detection algorithms have their own R package publicly available.

For the change in regression function, recent works related to change point analysis include Muller (1992) and Loader (1996), who used kernel smoothers. Wang (1995) and Raimondo (1998) used empirical wavelet coefficients. Also, parameter change in an autoregressive model was considered by Davis *et al.* (1995) and Huskova *et al.* (2007). In addition, Bayesian method can also be used to estimate the number of change points (see Lee (1998)).

1.1.3 List of Problems About Change Point Analysis

There are mainly two challenges in detecting change points. Firstly, it is difficult to find the asymptotic distribution of the test statistics proposed in the literature (Shao and Zhang (2010), Hušková and Meintanis (2006a)) because it often involves the use of extreme-value type distributions or Brownian bridges (Csörgő and Horváth(1997)). Some other methods were developed to determine the critical value, such as bootstrap (Hušková and Meintanis (2006a)), simulation study (Duggins 2010), block permutation (Kirch 2007) and so on. Another challenge is that the inadequate approximation in the asymptotic distribution of the statistics is likely to result in large and uncontrolled difference between the actual type I error probability and the nominal one.

Also, some of the current existing change point detection methods fail to handle

the abnormal time series. Even though there are rich literature on how to detect a change point in a time series, these time series are usually assumed in a standard form with white noise. In practice, a time series may have a complicated structure and can not be modeled well by a standard time series model. It is then difficult to use the existing methods to detect a change point directly in such a time series. To detect a change point in such time series becomes a challenge problem.It is common that there may contains outliers in the data sequence. Some of the proposed test statistics are sensitive to outliers and perform badly when dealing with datasets which contain irregular observations.

1.2 Application of Change Pionts Analysis in Longitudinal Data

Longitudinal studies are increasingly common in many areas of research including medicine, public health, and the social sciences. Data are longitudinal if they track the same type of information on the same subjects at multiple time points. For example, HIV patients may be followed over time and monthly measures, such as CD4 cell counts and viral load, are collected to characterize immune status and disease burden, respectively. Longitudinal data include two types of variation: the intra-individual and the inter-individual variation. Exploring the intra-individual

variation allows one to study the change over time in longitudinal studies, while modelling the inter-individual variation helps one to understand the difference between individuals. In many longitudinal studies, the inter-individual variation may be partially explained by time-varying covariates. However, some covariates may be measured with errors and may contain missing data as well. Ignoring the measurement errors and missing data in covariates may lead to bias results. For example, in HIV studies, the CD4 cell count is a very important factor to reflect the efficacy of the anti-HIV therapy, and it is measured repeatedly on the same patient in a study. It is well known that CD4 cell count is often measured with substantial errors.

Moreover, it is quite common that the viral load of some patients may rebound during the treatment. Such rebound part in one patient's trajectory may be an important indicator to help quantify treatment effect and improve management of patient care, and the model may become a challenge if the response contains rebound part. To overcome this challenge, change point models should be introduced and simultaneously addressed for the response model. Thus, it is important to simultaneously address measurement errors, missing data in covariates and change points in longitudinal studies. One can refer Chapter 4 for more details.

1.3 Objective and Outline of the Dissertation

The primary objective of this dissertation is to develop new methods for detecting the potential change points in univariate data sequence. We implement the use of change point in the HIV viral dynamic studies. We will conduct simulation studies to illustrate the performance of our proposed methods. To explain how to implement our methods in applications, we will give some examples which include analyzing the financial data, genetic data, longitudinal data as well as image de-noising.

In Chapter 2, we aim to test the change point in distribution and estimate its location if the change point exists. Two tests with test statistics based on empirical characteristic function are proposed to detect a change point in a data sequence. Then we extend our methods to multiple change point problems by using iterated cumulative sums of squares (ICSS) algorithm. The consistency and the rate of convergence for the estimated change point are established. Some simulation studies as well as real data analysis are given to illustrate the effective and efficiency of these methods.

In Chapter 3, we propose a procedure for detecting multiple change points in a mean-shift model, where the number of change points is allowed to increase with the sample size. A theoretic justification for our new method is also given. We first convert the change point problem into a variable selection problem by partitioning

the data sequence into several segments. Then, we apply a modified variance inflation factor regression algorithm to each segment in sequential order. When a segment that is suspected of containing a change point is found, we use a weighted cumulative sum to test if there is indeed a change point in this segment. The proposed procedure is implemented in an algorithm which, compared to two popular methods via simulation studies, demonstrates satisfactory performance in terms of accuracy, stability and computation time. Finally, we apply our new algorithm to analyze two real data examples.

In Chapter 4, we propose a semiparametric nonlinear mixed-effects response model incorporating measurement errors and missing data in time-varying covariates and change points. The covariate measurement error models and models for the times of change points on response trajectories are introduced for joint likeliood inference. We propose two approaches to obtain approximate maximum likelihood estimates of the joint model parameters simultaneously. We illustrate the proposed approaches to analyze a real dataset. A simulation study is conducted to evaluate these proposed approaches.

In Chapter 5, we summarize this dissertation and discuss future research.

2 Change Point Detection Based on Empirical Characteristic Function

2.1 Introduction

Detection of possible change points is of interest in many fields, such as signal recognition, graphics analysis, finance and so on. In graphical analysis, each image contain a great deal of pixels and can be transformed to be a matrix. We can regard that each row be a series on which the change point detection method based. It is necessary to develop an effective and efficient detection method to solve this kind of problem because even a small sized image can be transformed to a matrix of high dimension.

Empirical characteristic functions (ECF) have been proved to be a useful tool in statistical inference. Some works on the ECF include, among others, Jiménez-Gamero et al.(2016), Henze et al.(2014), Tenreiro (2011), Hušková and Meintanis

(2009), etc. One can refer to Csörgő (1984) and Ushakov (1999) for review articles. Actually, empirical characteristic function can also be used in detecting change points. Hušková and Meintanis (2006a) presented the procedure to detect single change point in a sequence of independent observations based on empirical characteristic functions. For more related literature, one can refer to Hušková and Meintanis (2006b), Hušková and Meintanis (2008) and Hlávka et. al (2012).

In this section, we will follow the similar model setting as Hušková and Meintanis (2006a). Let $X_1, \cdots, X_n$ be independent random variables following the distribution of $F_i, i = 1, 2, \cdots, n$, respectively. We want to test the following null hypothesis

$$H_0 : F_1 = \cdots = F_n \tag{2.1}$$

against

$$H_1 : F_1 = \cdots = F_{k^*} \neq F_{k^*+1} = \cdots = F_n \tag{2.2}$$

where k^*, F_1 and F_n are unknown. k^* is called the change point and for the sake of convenience, we assume that there exist τ_1, τ_2 satisfying $1 < n\tau_1 < k^* < n\tau_2 < n$ (Csörgő and Horváth, 1997). We denote $\tau_0 = k^*/n$. Our aim in this chapter is to propose tests to determine the existence of a change point in the sequence and then estimate its location if it exists.

In Hušková and Meintanis (2006a), the following test statistic was proposed:

$$T_{n,\gamma}(\omega) = \max_{1\le k<n} \Big(\frac{k(n-k)}{n^2}\Big)^{\gamma} \frac{k(n-k)}{n} \int_{-\infty}^{\infty} |\phi_k(t) - \phi_k^0(t)|^2 \omega(t) dt \qquad (2.3)$$

where $\omega(\cdot)$ is a nonnegative weight function, $\phi_k(t)$ and $\phi_k^0(t)$ are empirical characteristic functions based on $X_1, \cdots, X_k$ and $X_{k+1}, \cdots, X_n$, respectively, i.e.

$$\phi_k(t) = \frac{1}{k} \sum_{j=1}^{k} \exp\{itX_j\}, k = 1, \cdots, n, \qquad (2.4)$$

$$\phi_k^0(t) = \frac{1}{n-k} \sum_{j=k+1}^{n} \exp\{itX_j\}, k = 1, \cdots, n-1. \qquad (2.5)$$

We rename this test statistic and its related testing method by "ECF" in the current chapter.

The choice of the weight function ω and tuning parameter γ will influence the limit behavior of this test statistic. For ω, we follow the choice of Hušková and Meintanis (2006a) and set $\omega(t; a) = \frac{1}{2a} \exp\{-a|t|\}$, where $t \in \mathbb{R}^1, a > 0$. Here the role of the weighted parameter a is to control the rate of decay of the weight function. Hušková and Meintanis (2006a) presented some simulation studies for different γ and simulation results are quite similar among chosen γ. We choose $\gamma = 1$ in this chapter.

The limit distribution of this test statistic is neither exactly nor asymptotically distribution free under H_0. This is an unpleasant property. The calculation of this test statistic is also time-consuming and not sufficient enough. We will propose new test statistics to overcome the mentioned disadvantages.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the details of the two proposed statistics as well as its asymptotic properties. In Section 2.3, we implement our proposed tests to detect multiple change points in a data sequence. In Section 2.4, we present the performance of our proposed procedures by simulation studies and the real data analysis. We conclude this chapter in Section 2.5.

2.2 The Change Point Estimator Based on Empirical Characteristic Function

2.2.1 COS Method and its Asymptotic Properties

To achive a fast method for change point detection, we firstly propose the following statistic that is based on the real part of empirical characteristic function combining with the traditional cumulative sum chart (CUSUM) method.

$$C_{1k}(t) = \sqrt{\frac{k(n-k)}{n}}\Big(\frac{1}{k}\sum_{j=1}^{k}\cos(tX_j) - \frac{1}{n-k}\sum_{j=k+1}^{n}\cos(tX_j)\Big) \tag{2.6}$$

here t is unknown and may be different for different dataset. By simple calculation, the value of $|C_{1k}(t)|$ should be the largest at the location of k^* (refer Lemma 2.2.1).

We give an example to illustrate the property of $|C_{1k}(t)|$ while the plot of $|C_{1k}(t)|$ is given in Figure 2.1, here we set $t = 0.4$. Figure 2.1(a) shows a series of $N(0,1)$

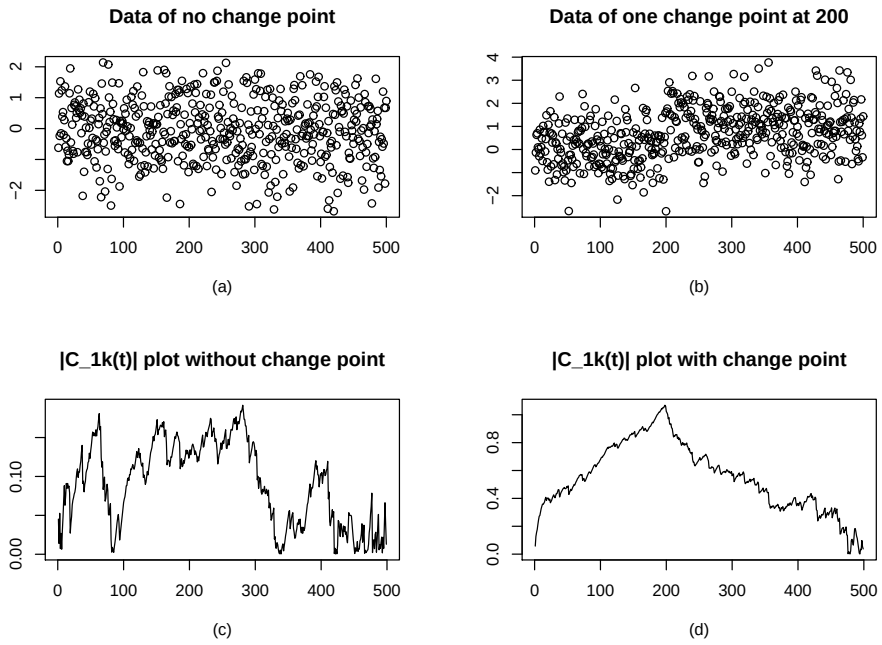

Figure 2.1: Example of $|C_{1k}(t)|$ plot. (a) and (b) are scatter plots of two datasets that without change point and with one change point, respectively. (c) and (d) are corresponding plots of $|C_{1k}(t)|$.

white noise with no change point; Figure 2.1(b) present the same length series with one change in mean at the place of 200 while the left part follows $N(0, 1)$ and the right part follows $N(1, 1)$. Figure 2.1(c) and 2.1(d) illustrate the function of $|C_{1k}(t)|$ respectively. From Figure 2.1(c), when there is no change point in the sequence, we can find there is no much difference in the values of $|C_{1k}(t)|$ because its range is about 0.2. What is more, there is no clear pattern for the curve of $|C_{1k}(t)|$. To

the contrary, when there exists a change point in the sequence, the values of $|C_{1k}(t)|$ firstly increase to the maximum and then decrease. The corresponding curve is given in Figure 2.1(d).

We follow the idea of Shao and Zhang (2010) and employ the self-normarlization (SN) method (Lobato 2001; Shao 2010) to the change point testing problem. In Shao and Zhang (2010), they aimed to test a change point in the mean of a univariate time series and under appropriate conditions, $\sqrt{n}(\bar{X}_n - \mu)$ converges to $N(0, \sigma^2)$ in distribution. To construct a confidence interval for μ, the traditional approach replaces the unknown variance σ^2 by its consistent estimate $\hat{\sigma}_n^2$. A commonly used estimate for σ^2 is as following

$$\hat{\sigma}_n^2 = \sum_{k=-l_n}^{l_n} \hat{\gamma}(k) K(k/l_n),$$

where $\hat{\gamma}(k) = n^{-1} \sum_{j=1}^{n-|k|} (X_j - \bar{X}_n)(X_{j+|k|-\bar{X}_n})$ is the sample autocovariance estimate at lag k, $K(\cdot)$ is a kernel function and $l = l_n$ is a bandwidth parameter. Then, the confidence interval for μ is constructed by using critical values from the $\chi^2(1)$ distribution because $n(\bar{X}_n - \mu)^2/\hat{\sigma}_n^2$ converges to $\chi^2(1)$ in distribution. However, for the traditional approach, the major difficulty is the choice of l_n. To avoid the the selection of l_n, Lobato (2001) proposed the SN approach as a good alternative to the traditional approach. Let $D_n^2 = n^{-2} \sum_{t=1}^{n} \{\sum_{j=1}^{t} (X_j - \bar{X}_n)\}^2$, then the continuous

mapping theorem implies that

$$n(\bar{X}_n - \mu)^2/D_n^2 \to \frac{B(1)^2}{\int_0^1 \{B(r) - rB(1)\}^2 dr}.$$

The corresponding critical values have been tabulated by Lobato (2011).

Following the idea of Shao and Zhang (2010), we propose the following test statistic and the related testing method is named "COS".

$$T_{1k}(t) = \frac{C_{1k}(t)}{D_{1k}(t)} \tag{2.7}$$

One example of D_{1k} (formula 1.4.25, Csörgő and Horváth 1997) is given in the following way

$$\begin{aligned} D_{1k}^2(t) &= \frac{1}{n}\Big\{ \sum_{i=1}^{k} \Big(\cos(tX_i) - \frac{1}{k}\sum_{j=1}^{k} \cos(tX_j)\Big)^2 \\ &\quad + \sum_{i=k+1}^{n} \Big(\cos(tX_i) - \frac{1}{n-k}\sum_{j=k+1}^{n} \cos(tX_j)\Big)^2 \Big\} \end{aligned} \tag{2.8}$$

An illustration of $D_{1k}(t)$ and T_{1k} is plotted in Figure 2.2. Here the data is the same as that of in Figure 2.1. From Figure 2.2(b) , it is easy to find out that $D_{1k}(t)$ reaches its minimum value at the location of the true change point. Thus, comparing with C_{1k}, $T_{1k}(t)$ is more effective in determining the change point.

For the test statistic T_{1k}, we have the following proposition that is useful for our hypothesis testing.

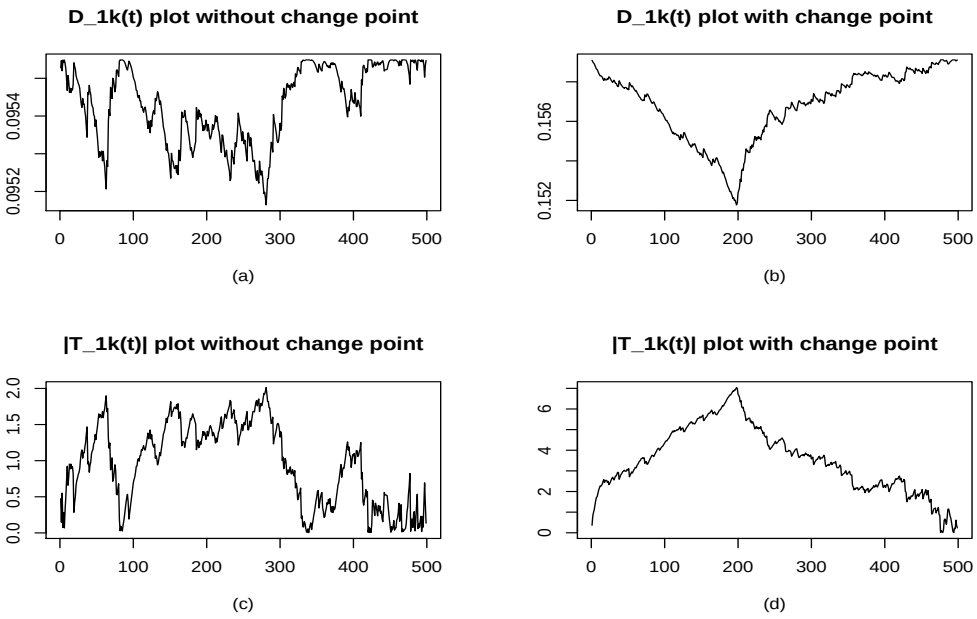

Figure 2.2: Plots of $D_{1k}(t)$ and $|T_{1k}|$ for two datasets without change point and with change point, respectively.

Proposition 2.2.1 *If $X_1, X_2, \cdots, X_n$ are independent identically distributed random variables, then under the null hypothesis, we have for any $0 < t < \infty$*

$$\lim_{n\to\infty} P\{A(\log n) \max_{1\le k\le n} |T_{1k}(t)| \le u + B(\log n)\} = \exp(-2e^{-u}) \tag{2.9}$$

where $A(x) = (2\log x)^{1/2}$ and $B(x) = 2\log x + \frac{1}{2}\log\log x - \frac{1}{2}\log\pi$.

Remark 2.2.1 *It is easy to prove the proposition referring to Csörgő M, Horváth L (1997) (Theorem 1.4.1).*

Before presenting our theorems to illustrate the consistency properties and con-

vergence rate, we need to state the following lemmas. For convenience, we denote $\cos(tX_1), \cdots, \cos(tX_n)$ by $Y_1, \cdots, Y_n$ and rename $\sin(tX_1), \cdots, \sin(tX_n)$ by $Z_1, \cdots, Z_n$, then

$$\begin{aligned}
C_{1k}(t) &= \sqrt{\frac{k(n-k)}{n}} \left(\frac{1}{k} \sum_{i=1}^{k} Y_i - \frac{1}{n-k} \sum_{i=k+1}^{n} Y_i \right), \\
D_{1k}^2(t) &= \frac{1}{n} \left\{ \sum_{i=1}^{k} \left(Y_i - \frac{1}{k} \sum_{j=1}^{k} Y_i \right)^2 + \sum_{i=k+1}^{n} \left(Y_i - \frac{1}{n-k} \sum_{j=k+1}^{n} Y_i \right)^2 \right\}, \\
C_{3k}(t) &\triangleq \sqrt{\frac{k(n-k)}{n}} \left(\frac{1}{k} \sum_{i=1}^{k} Z_i - \frac{1}{n-k} \sum_{i=k+1}^{n} Z_i \right), \\
D_{3k}^2(t) &\triangleq \frac{1}{n} \left\{ \sum_{i=1}^{k} \left(Z_i - \frac{1}{k} \sum_{j=1}^{k} Z_i \right)^2 + \sum_{i=k+1}^{n} \left(Z_i - \frac{1}{n-k} \sum_{j=k+1}^{n} Z_i \right)^2 \right\}.
\end{aligned}$$

After simple calculation, we have $C_{2k}(t) = C_{1k}(t) + iC_{3k}(t)$ and $D_{2k}^2(t) = D_{1k}^2(t) + D_{3k}^2(t)$. We assume that Y_1, Y_n, Z_1 and Z_n have the mean μ_1, μ_2, μ_3 and μ_4 and variance $\sigma_1^2, \sigma_2^2, \sigma_3^2$ and σ_4^2, respectively.

For convenience, we use $C_k(t)$ to denote $C_{1k}(t)$ or $C_{3k}(t)$, and similarly denote $D_k^2(t)$ as $D_{1k}^2(t)$ or $D_{3k}^2(t)$. First we give the following lemmas that are needed to prove our theorems.

Lemma 2.2.1 *$|EC_k(t)|$ obtains its maximum and $ED_k^2(t)$ obtains its minimum at the location of the true change point k^*.*

Proof of Lemma 2.2.1 After simple calculation, we have,

$$|EC_k(t)| = \begin{cases} \sqrt{\frac{(n-k)}{nk}} \cdot k^* |\mu_1 - \mu_2|, & 1 < k^* \leq k < n \\ \sqrt{\frac{k}{n(n-k)}} \cdot (n-k^*)|\mu_1 - \mu_2|, & 1 < k < k^* < n \end{cases}$$

and

$$E(D_k^2(t)) = \begin{cases} \frac{k^*(k-1)}{nk}(\sigma_1^2 - \sigma_2^2) + \frac{n-2}{n}\sigma_2^2 + \frac{k^*(k-k^*)}{nk}(\mu_1 - \mu_2)^2, & k \geq k^*, \\ \frac{(n-k^*)(n-k-1)}{n(n-k)}(\sigma_2^2 - \sigma_1^2) + \frac{n-2}{n}\sigma_1^2 + \frac{(n-k^*)(k^*-k)}{n(n-k)}(\mu_1 - \mu_2)^2, & k < k^*. \end{cases} \tag{2.10}$$

As $|EC_k(t)|$ is increasing for $k < k^*$ and decreasing for $k > k^*$, it is easy to conclude that it obtains its maximum when $k = k^*$.

For $|E(D_k^2(t))|$,

$$\begin{aligned} & E(D_{k+1}^2(t)) - E(D_k^2(t)) \\ = & \begin{cases} \frac{k^*}{nk(k+1)}\left[(\sigma_1^2 - \sigma_2^2) + k^*(\mu_1 - \mu_2)^2\right], & k > k^*, \\ -\frac{n-k^*}{n(n-k)(n-k-1)}\left[(\sigma_1^2 - \sigma_2^2) + (n-k^*)(\mu_1 - \mu_2)^2\right], & k < k^*. \end{cases} \end{aligned}$$

As k^* and $n-k^*$ are in the order of $O(n)$, it is easy to show that $[(\sigma_1^2 - \sigma_2^2) + k^*(\mu_1 - \mu_2)^2] > 0$ and $[(\sigma_1^2 - \sigma_2^2) + (n-k^*)(\mu_1 - \mu_2)^2] > 0$. Thus we can conclude that $E(D_k(t))$ is decreasing when $k < k^*$ and increasing when $k > k^*$ and obtains its minimum $[(k^*-1)\sigma_1^2 + (n-k^*-1)\sigma_2^2]/n$ when $k = k^*$.

Lemma 2.2.2 *We denote* $U_k = \left(\frac{k(n-k)}{n}\right)^{1-\alpha}\left(\frac{1}{k}\sum_{i=1}^{k} Y_i - \frac{1}{n-k}\sum_{i=k+1}^{n} Y_i\right)$ *and* $0 \le \alpha < 1$, *then we have* $n^{\alpha-1}\max_{1\le k<n}|U_k - EU_k| \to 0, a.s., as\ n \to +\infty$

Remark 2.2.2 *One can refer to Shi et al. (2008) for the proof of the above lemma.*

Lemma 2.2.3 *For* $C_{1k}(t)$, D_{1k}^2 *and* $C_{3k}(t)$, D_{3k}^2, *the following formulas exist.*

$$nD_{1k}^2 + C_{1k}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{2.11}$$

$$nD_{3k}^2 + C_{3k}^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2, \tag{2.12}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

Proof of Lemma 2.2.3 We only need to prove formula (2.11), and then obtain formula (2.12) similarly. Firstly, we denote $\bar{Y}_k = \frac{1}{k}\sum_{i=1}^{k} Y_i$ and $\bar{Y}_{n-k} = \frac{1}{n-k}Y_i$.

$$\begin{aligned}
nD_{1k}^2(t) + C_{1k}^2(t) &= \sum_{i=1}^{n} Y_i^2 - k\bar{Y}_k^2 - (n-k)\bar{Y}_{n-k}^2 + \frac{k(n-k)}{n}(\bar{Y}_k - \bar{Y}_{n-k})^2 \\
&= \sum_{i=1}^{n} Y_i^2 - k\bar{Y}_k^2 - (n-k)\bar{Y}_{n-k}^2 + \frac{k(n-k)}{n}\,(\bar{Y}_k^2 - 2\bar{Y}_k\bar{Y}_{n-k} + \bar{Y}_{n-k}^2) \\
&= \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\big(k\bar{Y}_k + (n-k)\bar{Y}_{n-k}\big)^2 \\
&= \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\big(\sum_{i=1}^{n} Y_i\big)^2 \\
&= \sum_{i=1}^{n}\big(Y_i - \bar{Y}\big)^2. \qquad \sharp
\end{aligned}$$

Next, we will establish the theorems to illustrate the asymptotic results of our proposed test statistic. For convenience, we denote $\hat{\tau}_1(t) = \hat{k}_1(t)/n$, where $\hat{k}_1(t) = \arg\max_{1<k<n} |T_{1k}(t)|$. And we can establish the consistency, the rate of convergence of $\hat{\tau}_1(t)$ in the following two theorems.

Theorem 2.2.1 *Under alternative hypothesis, we assume that there exists a* $t_0 \in (0, \infty)$ *such that* $E\{\cos(t_0 X_1)\} \neq E\{\cos(t_0 X_n)\}$ *and then we have*

$$\hat{\tau}_1(t_0) - \tau_0 \to 0, a.s.$$

Remark 2.2.3 *The theorem shows that the estimated location of change point by "COS" method is consistent for* τ_0.

Proof of Theorem 2.2.1 We denote $T_{1k}(t) = \frac{C_{1k}(t)}{D_{1k}(t)}$ and $\bar{T}_{1k}(t) = \frac{C_{1k}(t)}{\sqrt{ED^2_{1k}(t)}}$. We just consider the case that $1 < k^* < k < n$. For the situation that $1 < k < k^*$ we can prove in the same way.

$$\begin{aligned} |E\bar{T}_{1k^*}(t)| - |E\bar{T}_{1k}(t)| &= |E\bar{T}_{1k^*}(t)|\Big(1 - \Big|\frac{\sqrt{ED^2_{1k^*}(t)}}{\sqrt{ED^2_{1k}(t)}}\Big|\Big|\frac{EC_{1k}(t)}{EC_{1k^*}(t)}\Big|\Big) \\ &\geq |E\bar{T}_{1k^*}(t)|\Big(1 - \Big|\frac{EC_{1k}(t)}{EC_{1k^*}(t)}\Big|\Big) \\ &\geq |E\bar{T}_{1k^*}(t)|\left[1 - \left(\frac{EC_{1k}(t)}{EC_{1k^*}(t)}\right)^2\right]\Big/2. \end{aligned} \tag{2.13}$$

The second inequality holds because for any x, we have $1 - x^2 < 2(1 - |x|)$.

For those two terms in formula (2.13), we have

$$\begin{aligned} |E\bar{T}_{1k^*}(t)| &= \frac{|\mu_1-\mu_2|}{\sqrt{\frac{1}{n}[(k^*-1)\sigma_1^2+(n-k^*-1)\sigma_2^2]}}\sqrt{\frac{k^*(n-k^*)}{n}} \\ &\geq \frac{|\mu_1-\mu_2|}{\sqrt{\sigma_1^2+\sigma_2^2}}\sqrt{\frac{k^*(n-k^*)}{n}} \end{aligned}$$

and

$$1-\left(\frac{EC_{1k}(t)}{EC_{1k^*}(t)}\right)^2 = 1-\frac{(n-k)k^*}{(n-k^*)k} = \frac{n(k-k^*)}{(n-k^*)k} \geq \frac{k-k^*}{n-k^*},$$

thus, it is easy to know

$$|E\bar{T}_{k^*}(t)| - |E\bar{T}_k(t)| \geq \frac{|\mu_1-\mu_2|}{2\sqrt{\sigma_1^2+\sigma_2^2}}\sqrt{\frac{k^*}{n(n-k^*)}}\cdot(k-k^*)$$

It is the assumption that there exist τ' and τ'' satisfying $0<\tau'<\tau''<1$, s.t. $n\tau'<k^*<n\tau''$, so $k^*/(n-k^*)>\tau'/(1-\tau'')$. We obtain the conclusion that there exist a constant $M_1=\frac{|\mu_1-\mu_2|}{2\sqrt{\sigma_1^2+\sigma_2^2}}\cdot\sqrt{\frac{\tau'}{1-\tau''}}$, s.t.

$$|E\bar{T}_{1k^*}(t)| - |E\bar{T}_{1k}(t)| > M_1|k-k^*|/\sqrt{n}. \tag{2.14}$$

On the other hand,

$$\begin{aligned} |E\bar{T}_{1k^*}(t)| - |E\bar{T}_{1k}(t)| &\leq |E\bar{T}_{1k^*}(t)-\bar{T}_{1k^*}(t)| + |\bar{T}_{1k^*}(t)-T_{1k^*}(t)| + |T_{1k^*}(t)| \\ &\quad -|T_{1k}(t)| + |E\bar{T}_{1k}(t)-\bar{T}_{1k}(t)| + |\bar{T}_{1k}(t)-T_{1k}(t)| \\ &\leq 2\max_k|\bar{T}_{1k}(t)-E\bar{T}_{1k}(t)| + 2\max_k|T_{1k}(t)-\bar{T}_{1k}(t)| \\ &\quad +|T_{1k^*}(t)|-|T_{1k}(t)| \end{aligned} \tag{2.15}$$

From formula (2.14) and (2.15), we have

$$\begin{aligned} |k-k^*|/n &\leq \frac{1}{\sqrt{n}C_1}(|E\bar{T}_{1k^*}(t)| - |E\bar{T}_{1k}(t)|) \\ &\leq \frac{2}{\sqrt{n}M_1}\Big(\max_k |\bar{T}_{1k}(t) - E\bar{T}_{1k}(t)| + \max_k |T_{1k}(t) - \bar{T}_{1k}(t)| \\ &\qquad +|T_{1k^*}(t)| - |T_{1k}(t)|\Big) \end{aligned}$$

Because $\hat{k}_1$ is defined as $\arg\max_k |T_{1k}(t)|$, we can obtain the following formula after replacing k by $\hat{k}_1$

$$\begin{aligned} |\hat{k}_1-k^*|/n &\leq \frac{2}{\sqrt{n}M_1}\Big(\max_k |\bar{T}_{1k}(t) - E\bar{T}_{1k}(t)| + \max_k |T_{1k}(t) - \bar{T}_{1k}(t)|\Big) \\ &\triangleq I + II \end{aligned}$$

We can finish the proof if we can show that I and II converge to zero almost surely. For I, we have

$$\begin{aligned} I &= \frac{2}{\sqrt{n}M_1}\max_k |\bar{T}_{1k}(t) - E\bar{T}_{1k(t)}| \\ &= \frac{2}{\sqrt{n}M_1}\max_k \left|\frac{C_{1k}(t) - EC_k(t)}{\sqrt{ED^2_{1k}(t)}}\right| \\ &\leq \frac{2}{\sqrt{n}M_1}\frac{1}{\sqrt{[(k^*-1)\sigma_1^2 + (n-k^*-1)\sigma_2^2]/n}}\max_k |C_{1k}(t) - EC_{1k}(t)|. \end{aligned} \quad (2.16)$$

Thus, by Lemma 2.2.2, we know $I \to 0$ almost surely.

Secondly, for II, we have

$$\begin{aligned} II &= \frac{2}{\sqrt{n}M_1}\max_k|\bar{T}_{1k}(t)-\bar{T}_{1k}(t)| \\ &= \frac{2}{\sqrt{n}M_1}\max_k\Big|\frac{C_{1k}(t)}{D_{1k}(t)}-\frac{C_{1k}(t)}{\sqrt{ED^2_{1k}(t)}}\Big| \\ &\leq \frac{2}{\sqrt{n}M_1}\max_k|C_{1k}(t)|\cdot\max_k\left|\frac{1}{D_{1k}(t)}-\frac{1}{\sqrt{ED^2_{1k}(t)}}\right|. \end{aligned} \tag{2.17}$$

For the first part of formula (2.17), the following inequation exists:

$$\frac{2}{\sqrt{n}M_1}\max_k|C_{1k}(t)| \leq \frac{2}{\sqrt{n}M_1}\max_k|C_{1k}(t)-EC_{1k}(t)| + \frac{2}{\sqrt{n}M_1}\max_k|EC_{1k}(t)|.$$

By Lemma 2.2.2, it is easy to know $\frac{2}{\sqrt{n}M_1}\max_k|C_{1k}(t)-EC_{1k}(t)| \to 0, a.s.$ and we can obtain

$$\begin{aligned} \frac{2}{\sqrt{n}M_1}\max_k|EC_{1k}(t)| &= \frac{2}{\sqrt{n}M_1}\sqrt{\frac{(n-k^*)k^*}{n}}|\mu_1-\mu_2| \\ &\leq \frac{2}{M_1}\cdot\frac{1}{2}|\mu_1-\mu_2| = \frac{|\mu_1-\mu_2|}{M_1}, \end{aligned}$$

thus, we know $\frac{2}{\sqrt{n}M_1}\max_k|C_{1k}(t)| \leq \frac{|\mu_1-\mu_2|}{M_1}, a.s.$

For the second part of formula (2.17),

$$\begin{aligned} &\max_k\Big|\frac{1}{D_{1k}(t)}-\frac{1}{\sqrt{ED^2_{1k}(t)}}\Big| \\ =\ &\max_k\Big|\frac{D^2_{1k}(t)-ED^2_{1k}(t)}{D_{1k}(t)\sqrt{ED^2_{1k}(t)}\big(D_{1k}(t)+\sqrt{ED^2_{1k}(t)}\big)}\Big| \qquad (2.18) \\ \leq\ &\frac{\max_k|D^2_{1k}(t)-ED^2_{1k}(t)|}{\min_k\sqrt{ED^2_{1k}(t)}\cdot\min_k\left[D_{1k}(t)\left(D_{1k}(t)+\sqrt{ED^2_{1k}(t)}\right)\right]} \qquad (2.19) \end{aligned}$$

Denote $\bar{Y}_k = \sum_{j=1}^{k} Y_j \Big/ k$ and $\bar{Y}_{n-k} = \sum_{j=k+1}^{n} Y_j \Big/ (n-k)$. Then

$$\begin{aligned}
\max_k |D_{1k}^2(t) - ED_{1k}^2(t)| &= \frac{1}{n} \max_k \Big| \sum_{j=1}^{n} (Y_i^2 - EY_j^2) - k(\bar{Y_k}^2 - E\bar{Y}_k^2) \\
&\qquad -(n-k)(\bar{Y}_{n-k}^2 - E\bar{Y}_{n-k}^2) \Big| \\
&\leq \Big| \frac{1}{n} \sum_{j=1}^{n} (Y_j^2 - EY_j^2) \Big| + \max_k \Big| \frac{k}{n} (\bar{Y_k}^2 - E\bar{Y}_k^2) \Big| \\
&\qquad + \max_k \Big| \frac{n-k}{n} (\bar{Y}_{n-k}^2 - E\bar{Y}_{n-k}^2) \Big| \\
&\triangleq III + IV + V.
\end{aligned}$$

It is obviously that $III = |\frac{1}{n} \sum_{j=1}^{n} (Y_j^2 - EY_j^2)| \to 0, a.s.$ by Lemma 2.2.2. For IV, we have

$$\begin{aligned}
IV &= \max_k \left| \frac{k}{n} (\bar{Y_k}^2 - E\bar{Y}_k^2) \right| \\
&= \frac{1}{n} \max_k k |(\bar{Y}_k - E\bar{Y}_k)^2 + 2E\bar{Y}_k(\bar{Y}_k - E\bar{Y}_k) - var\bar{Y}_k| \\
&\leq \frac{1}{n} \max_k \frac{1}{k} \Big\{ \sum_{j=1}^{k} (Y_j - EY_j) \Big\}^2 + \frac{2}{n} \max_k |\sum_{j=1}^{k} (Y_j - EY_j)| \cdot \max_k |E\bar{Y}_k| \\
&\qquad + \frac{1}{n} \max_k k \cdot var\bar{Y}_k \\
&\leq \Big\{ \frac{1}{\sqrt{n}} \max_k |\frac{1}{\sqrt{k}} \sum_{j=1}^{k} (Y_j - EY_j)| \Big\}^2 + \frac{2}{n} \max_k \left| \sum_{j=1}^{k} (Y_j - EY_j) \right| \cdot \max(|\mu_1|, |\mu_2|) \\
&\qquad + \frac{1}{n} \max(\sigma_1^2, \sigma_2^2). \qquad (2.20)
\end{aligned}$$

As $\frac{1}{\sqrt{n}} \max_k |\frac{1}{\sqrt{k}} \sum_{j=1}^{k} (Y_j - EY_j)| \to 0, a.s.$ and $\max_k |\sum_{j=1}^{k} (Y_j - EY_j)| \to 0, a.s.$, it is straightforward to obtain $IV \to 0, a.s.$. Similarly, we have $V \to 0, a.s.$. Then, we

can conclude that

$$\max_k |D_{1k}^2(t) - ED_{1k}^2(t)| \to 0, \qquad a.s. \tag{2.21}$$

Next, we will prove $\min_k\{D_{1k}(t)[D_{1k}(t)+\sqrt{ED_{1k}^2(t)}\ ]\}$ is lower bounded. By Lemma (2.2.1), we know that there exists a constant M_2, s.t. $ED_{1k}^2(t) > M_2$, thus we have

$$\begin{aligned}
&\Big(D_{1k}(t)[D_{1k}(t) + \sqrt{ED_{1k}^2(t)}\ ]\Big)^2 = D_{1k}^2(t)\Big(D_{1k}(t) + \sqrt{ED_{1k}^2(t)}\Big)^2 \\
\geq\ & (ED_{1k}^2(t) - \max_k |D_{1k}^2(t) - ED_{1k}^2(t)|)\Big(\sqrt{ED_{1k}^2(t) - \max_k |D_{1k}^2(t) - ED_{1k}^2(t)|} \\
&\qquad + \sqrt{ED_{1k}^2(t)}\Big)^2 \\
\geq\ & (M_2 - o(1))(\sqrt{M_2 - o(1)} + M_2).
\end{aligned}$$

It means that there exists M_3, s.t.

$$\min_k\{D_{1k}(t)[D_{1k}(t) + \sqrt{ED_{1k}^2(t)}\ ]\} > M_3. \tag{2.22}$$

From (2.19), (2.21) and (2.22), we have $\max_k \left|\frac{1}{D_{1k}(t)} - \frac{1}{\sqrt{ED_{1k}^2(t)}}\right| \to 0, a.s.$ Then $II \to 0, a.s.$ and we finish the proof of Theorem 2.2.1

Theorem 2.2.2 *Under the alternative hypothesis, we assume that there exists a* $t_0 \in (0,\infty)$ *such that* $E\{\cos(t_0X_1)\} \neq E\{\cos(t_0X_n)\}$ *and then we have*

$$\hat{\tau}_1(t_0) - \tau_0 = O_p\left(\frac{1}{n}\right)$$

Proof of Theorem 2.2.2 By model setting, there exist τ_1, τ_2 satisfing $0 < \tau_1 < \tau_0 < \tau_2 < 1$. Therefore, there is a $\delta > 0$ such that $\tau_0 \in (\delta, 1-\delta)$. Since $\hat{k}_1/n$ is consistent for τ_0, for every $\epsilon > 0$, $P(\hat{k}_1/n \notin (\delta, 1-\delta)) < \epsilon$ when n is large. To prove Theorem 2.2.2, we will prove $P(|\hat{\tau}_1(t_0) - \tau_0| > M/n)$ is small when n and M are both large. For every $M > 0$, define $W_{n,M} = \{k; n\delta \le k \le n(1-\delta), |k - k_0| > M\}$. From Lemma 2.2.3, we have $nD_{1k}^2 + C_{1k}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \triangleq \Lambda_1$. Then

$$
\begin{aligned}
& P(|\hat{\tau}_1 - \tau_0| > M/n) \\
\le\ & P(\hat{\tau}_1(t_0) \notin (\delta, 1-\delta)) + P(|\hat{\tau}_1(t_0) - \tau_0| > M/n, \hat{\tau} \in (\delta, 1-\delta))) \\
\le\ & \epsilon + P(\max_{k \in W_{n,M}} |T_{1k}| \ge |T_{1k^*}|) \\
=\ & \epsilon + P(\max_{k \in W_{n,M}} T_{1k}^2 \ge T_{1k^*}^2) \\
\le\ & \epsilon + P(\max_{k \in W_{n,M}} T_{1k}^2 - T_{1k^*}^2 \ge 0) \\
=\ & \epsilon + P(\max_{k \in W_{n,M}} \frac{C_{1k}^2}{D_{1k}^2} - \frac{C_{1k^*}^2}{D_{1k^*}^2} \ge 0) \\
=\ & \epsilon + P(\max_{k \in W_{n,M}} [C_{1k}^2 D_{1k^*}^2 - C_{1k^*}^2 D_{1k}^2] \ge 0) \\
=\ & \epsilon + P\left(\max_{k \in W_{n,M}} \left[C_{1k}^2(\Lambda_1 - C_{1k^*}^2) - C_{1k^*}^2(\Lambda_1 - C_{1k}^2)\right] \ge 0\right) \\
=\ & \epsilon + P\left(\max_{k \in W_{n,M}} \Lambda_1(C_{1k}^2 - C_{1k^*}^2) \ge 0\right) \\
=\ & \epsilon + P\left(\max_{k \in W_{n,M}} C_{1k}^2 - C_{1k^*}^2 \ge 0\right) \triangleq \epsilon + P_1.
\end{aligned}
$$

By Bai(1993), P_1 converges to zero as n tends to infinity which concludes the proof.

Remark 2.2.4 *We derive the asymptotic properies of test statistic T_{1k} in Proposi-*

tion 2.2.1, Theorem 2.2.1 and Theorem 2.2.2. We can derive the critical value for test by Proposition 2.2.1. By Theorem 2.2.1, change point estimates convergent to the true change point with convergence rate $O(n)$, as shown in Theorem 2.2.2.

2.2.2 EXP Method and its Asymptotic Properties

In this section, we will estiablish another test with test statistic based on the empirical characteristic function. Similar as the statistic in section 2.2.1, we denote the numerator as $C_{2k}(t)$ and denominator as $D_{2k}(t)$ with the expression as follows:

$$\begin{aligned} C_{2k}(t) &= \sqrt{\frac{k(n-k)}{n}}\Big(\frac{1}{k}\sum_{j=1}^{k}\exp(itX_j) - \frac{1}{n-k}\sum_{j=k+1}^{n}\exp(itX_j)\Big) \\ D^2_{2k}(t) &= \frac{1}{n}\Big\{\sum_{j=1}^{k}\Big|\exp(itX_j) - \frac{1}{k}\sum_{m=1}^{k}\exp(itX_m)\Big|^2 + \\ &\qquad \sum_{j=k+1}^{n}\Big|\exp(itX_j) - \frac{1}{n-k}\sum_{m=k+1}^{n}\exp(itX_m)\Big|^2\Big\}. \end{aligned}$$

The plots of $C_{2k}(t)$ and $D_{2k}(t)$ are shown in Figure 2.3, where the data is the same as that of in Figure (2.1). We define $T_{2k} = C_{2k}/D_{2k}$ as the statistic we focus on in this section. From the plot, we find that the proposed test can detect the true change point when the value of $|T_{2k}|$ approaches its maximum. We regard this method as "EXP" method in the current dissertation.

For the statistic T_{2k}, we have the similar proposition as Prop.(2.2.1) that could be used to determine the critical value when testing the hypothesis.

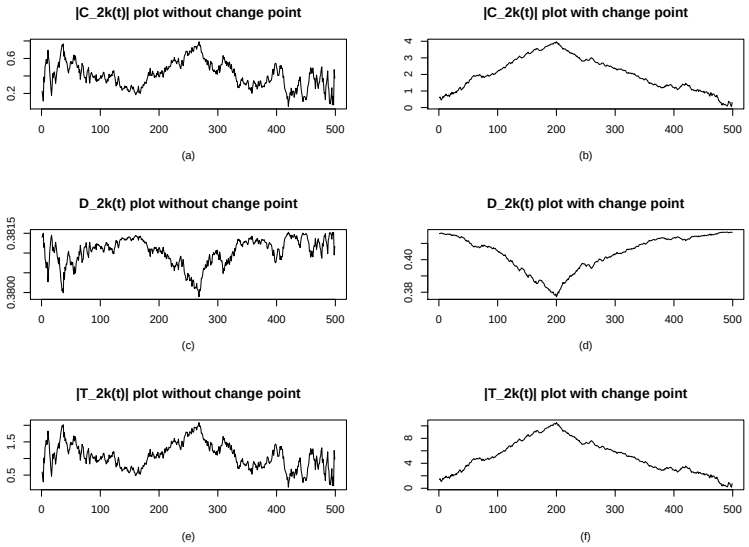

Figure 2.3: Example of $|C_{2k}(t)|$ plot. (a) and (b) are scatter plots of two datasets that without change point and with one change point, respectively. (c) and (d) are corresponding plots of $|C_{2k}(t)|$.

Proposition 2.2.2 *If $X_1, X_2, \cdots, X_n$ are independent and identically distributed random variables, then under the null hypothesis, we have*

$$\lim_{n\to\infty} P\{A(\log n) \max_{1\le k\le n} |T_{2k}| \le u + B(\log n)\} = \exp(-2e^{-u}), \tag{2.23}$$

where $A(x) = (2\log x)^{1/2}$ and $B(x) = 2\log x + \frac{1}{2}\log\log x - \frac{1}{2}\log\pi$.

Remark 2.2.5 *Similarly, we can prove the proposition referring to Csörgő and Horváth (1997) (Theorem 1.4.1).*

Define $\hat{\tau}_2(t) = \hat{k}_2(t)/n$ and $\hat{k}_2(t) = \arg\max_{1<k<n} |T_{2k}(t)|$. We have the following

theorems regarding the consistency and rate of convergence of $\hat{\tau}_2(t)$.

Theorem 2.2.3 *Under the alternative hypothesis, we assume that there exists a* $t_0 \in (0, \infty)$ *such that* $E\{\exp(it_0X_1)\} \neq E\{\exp(it_0X_n)\}$ *and then we have*

$$\hat{\tau}_2(t_0) - \tau_0 \to 0, a.s.$$

Proof of Theorem 2.2.3 Actually, the proof of Theorem 2.2.3 is similar to the proof of Theorem 2.2.1. Denote $\bar{T}_{2k}(t) = \frac{C_{2k}(t)}{\sqrt{ED^2_{2k}(t)}}$. We only consider the case that $1 < k^* < k < n$ while for the situation that $1 < k < k^* < n$, we can obtain the same conclusion in the similar way.

Similar to formula (2.13), we have

$$|E\bar{T}_{2k^*}(t)| - |E\bar{T}_{2k}(t)| \geq |E\bar{T}_{2k^*}(t)| \left[1 - \left(\frac{EC_{2k}(t)}{EC_{2k^*}(t)}\right)^2\right] \Big/ 2. \tag{2.24}$$

After simple calculation, we have

$$\begin{aligned} |E\bar{T}_{2k^*}(t)| &= \left|\frac{EC_{1k^*}(t) + iEC_{3k^*}(t)}{\sqrt{ED^2_{1k^*}(t) + ED^2_{3k^*}(t)}}\right| \\ &= \frac{\sqrt{(\mu_1 - \mu_2)^2 + (\mu_3 - \mu_4)^2}}{\sqrt{\frac{1}{n}\left[(k^* - 1)(\sigma_1^2 + \sigma_3^2) + (n - k^* - 1)(\sigma_2^2 + \sigma_4^2)\right]}} \sqrt{\frac{k^*(n - k^*)}{n}}, \end{aligned}$$

and

$$1 - \left(\frac{EC_{2k}(t)}{EC_{2k^*}(t)}\right)^2 \geq \frac{k - k^*}{n - k^*}. \tag{2.25}$$

We can then conclude that there exists a constant M_4, s.t.

$$|E\bar{T}_{2k^*}(t)| - |E\bar{T}_{2k}(t)| > M_4|k-k^*|\sqrt{n}. \tag{2.26}$$

Similar to formula (2.15), we have

$$\begin{aligned} |k-k^*|\sqrt{n} \;\leq\; & \frac{2}{\sqrt{n}M_3}\big(\max_k |\bar{T}_{2k}(t) - E\bar{T}_{2k}(t)| + \max_k |T_{2k}(t) - \bar{T}_{2k}(t)| \\ & + |T_{2k^*}(t)| - |T_{2k}(t)|\big). \end{aligned} \tag{2.27}$$

As $\hat{k}_2 = \arg\max_k |T_{2k}(t)|$, we obtain the following formula after replacing k by $\hat{k}_2$,

$$\begin{aligned} |\hat{k}-k^*|/n \;&\leq\; \frac{2}{\sqrt{n}M_4}\Big(\max_k |\bar{T}_{2k}(t) - E\bar{T}_{2k}(t)| + \max_k |T_{2k}(t) - \bar{T}_{2k}(t)|\Big) \\ &\triangleq\; \text{VI} + \text{VII}. \end{aligned}$$

We can finish the proof if we can prove that both VI and VII converge to 0 a.s. Firstly,

$$\begin{aligned} \text{VI} \;&=\; \frac{2}{\sqrt{n}M_4}\max_k |\bar{T}_{2k}(t) - E\bar{T}_{2k}(t)| \\ &=\; \frac{2}{\sqrt{n}M_4}\max_k \left|\frac{C_{1k}(t) - EC_{1k}(t) + i(C_{3k}(t) - EC_{3k}(t))}{\sqrt{ED_{1k}^2(t) + ED_{3k}^2(t)}}\right| \\ &\leq\; \frac{2}{\sqrt{n}M_4}\frac{\max_k |C_{1k}(t) - EC_{1k}(t) + i(C_{3k}(t) - EC_{3k}(t))|}{\sqrt{[(k^*-1)(\sigma_1^2+\sigma_3^2) + (n-k^*-1)(\sigma_2^2+\sigma_4^2)]/n}}. \end{aligned} \tag{2.28}$$

By Lemma 2.2.2, we know that $VI \to 0$ almost surely.

Secondly, for formula VII, we have

$$\begin{aligned}
\text{VII} &= \frac{2}{\sqrt{n}M_4}\max_k|\bar{T}_{2k}(t)-\bar{T}_{2k}(t)| \\
&= \frac{2}{\sqrt{n}M_4}\max_k\left|\frac{C_{2k}(t)}{D_{2k}(t)}-\frac{C_{2k}(t)}{\sqrt{ED^2_{2k}(t)}}\right| \\
&\leq \frac{2}{\sqrt{n}M_4}\max_k|C_{2k}(t)|\cdot\max_k\left|\frac{1}{D_{2k}(t)}-\frac{1}{\sqrt{ED^2_{2k}(t)}}\right|. \qquad (2.29)
\end{aligned}$$

As we know $\frac{2}{\sqrt{n}}\max_k|C_{1k}(t)|\leq|\mu_1-\mu_2|$ and $\frac{2}{\sqrt{n}}\max_k|C_{3k}(t)|\leq|\mu_3-\mu_4|$, thus

$$\begin{aligned}
\frac{2}{\sqrt{n}M_4}\max_k|C_{2k}(t)| &= \frac{2}{\sqrt{n}M_3}\max_k|C_{1k}(t)+iC_{3k}| \\
&\leq \frac{2}{\sqrt{n}M_4}\Big(\max_k|C_{1k}(t)|+\max_k|C_{3k}(t)|\Big) \\
&\leq \frac{|\mu_1-\mu_2|+|\mu_3-\mu_4|}{M_4}.
\end{aligned}$$

For the second part of formula (2.29),

$$\begin{aligned}
&\max_k\Big|\frac{1}{D_{2k}(t)}-\frac{1}{\sqrt{ED^2_{2k}(t)}}\Big| \\
=\ &\max_k\Big|\frac{D^2_{2k}(t)-ED^2_{2k}(t)}{D_{2k}(t)\sqrt{ED^2_{2k}(t)}\big(D_{2k}(t)+\sqrt{ED^2_{2k}(t)}\big)}\Big| \\
\leq\ &\frac{\max_k|D^2_{2k}(t)-ED^2_{2k}(t)|}{\min_k\sqrt{ED^2_{2k}(t)}\cdot\min_k\Big[D_{2k}(t)\Big(D_{2k}(t)+\sqrt{ED^2_{2k}(t)}\Big)\Big]}. \qquad (2.30)
\end{aligned}$$

For the numerator of formula (2.30),

$$\begin{aligned}
\max_k|D^2_{2k}(t)-ED^2_{2k}(t)| &= \max_k|D^2_{1k}(t)-ED^2_{1k}(t)+D^2_{3k}(t)-ED^2_{3k}(t)| \\
&\leq \max_k|D^2_{1k}(t)-ED^2_{1k}(t)|+\max_k|D^2_{3k}(t)-ED^2_{3k}(t)| \\
&\to 0. \quad a.s.
\end{aligned}$$

It is easy to prove there exists a constant M_5, s.t. $ED_{2k}^2(t) > M_5$,

$$\begin{aligned}
&\left(D_{2k}(t)\left[D_{2k}(t)+\sqrt{ED_{2k}^2(t)}\right]\right)^2 = D_{2k}^2(t)\left(D_{2k}(t)+\sqrt{ED_{2k}^2(t)}\right)^2 \\
\geq\ & (ED_{2k}^2(t)-\max_k|D_{2k}^2(t)-ED_{2k}^2(t)|)\Big(\sqrt{ED_{2k}^2(t)-\max_k|D_{2k}^2(t)-ED_{2k}^2(t)|} \\
&\quad +\sqrt{ED_{2k}^2(t)}\Big)^2 \\
\geq\ & (M_5-o(1))(\sqrt{M_5-o(1)}+M_5).
\end{aligned}$$

So there exists a constant M_6, s.t. $\min_k\{D_{2k}(t)[D_{2k}(t)+\sqrt{ED_{2k}^2(t)}]\} > M_6, \quad a.s.$

From the above formulas, we can conclude that $\max_k\left|\frac{1}{D_{2k}(t)}-\frac{1}{\sqrt{ED_{2k}^2(t)}}\right| \to 0$, a.s. Then $VII \to 0, a.s.$ and we finish the proof of Theorem 2.2.3

Theorem 2.2.4 *Under alternative hypothesis, we assume that there exists a* $t_0 \in (0,\infty)$ *such that* $E\{\exp(it_0X_1)\} \neq E\{\exp(it_0X_n)\}$ *and then we have*

$$\hat{\tau}_2(t_0)-\tau_0 = O_p\left(\frac{1}{n}\right).$$

Proof of Theorem 2.2.4 From the model setting, we know that there exist τ_1, τ_2 satisfing $0 < \tau_1 < \tau_0 < \tau_2 < 1$, thus a δ exists and satisfies two conditions: larger than 0 and $\tau_0 \in (\delta, 1-\delta)$. Since $\hat{k}_2/n$ is consistent for τ_0, for every $\epsilon > 0$, $P(\hat{k}_2/n \notin (\delta, 1-\delta)) < \epsilon$ when n is large. In order to prove Theorem 2.2.4, we will prove that $P(|\hat{\tau}_2(t)-\tau_0| > M'/n)$ is small when n and M' are both large. For every $M' > 0$, define $W_{n,M'} = \{k; n\delta \leq k \leq n(1-\delta), |k-k^*| > M'\}$. From Lemma 2.2.3, we have $nD_{1k}^2 + C_{1k}^2 = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 \triangleq \Lambda_1$ and $nD_{3k}^2 + C_{3k}^2 = \sum_{i=1}^{n}(Z_i-\bar{Z})^2 \triangleq \Lambda_2$. Then

$$
\begin{aligned}
& P(|\hat{\tau}_2(t) - \tau_0| > M'/n) \\
\leq\ & P(\hat{\tau}_2(t) \notin (\delta, 1-\delta)) + P(|\hat{\tau}_2(t) - \tau_0| > M'/n, \hat{\tau}_2(t) \in (\delta, 1-\delta))) \\
\leq\ & \epsilon + P(\max_{k \in W_{n,M'}} |T_{2k}| \geq |T_{2k^*}|) \\
=\ & \epsilon + P(\max_{k \in W_{n,M'}} |T_{2k}|^2 \geq |T_{2k^*}|^2) \\
\leq\ & \epsilon + P(\max_{k \in W_{n,M'}} |T_{2k}|^2 - |T_{2k^*}|^2 \geq 0) \\
=\ & \epsilon + P(\max_{k \in W_{n,M'}} \frac{|C_{2k}|^2}{D_{2k}^2} - \frac{|C_{2k^*}|^2}{D_{2k^*}^2} \geq 0) \\
=\ & \epsilon + P(\max_{k \in W_{n,M'}} [|C_{2k}|^2 D_{2k^*}^2 - |C_{2k^*}|^2 D_{2k}^2] \geq 0) \\
=\ & \epsilon + P(\max_{k \in W_{n,M'}} \Big[(C_{1k}^2 + C_{3k}^2)(\Lambda_1 + \Lambda_2 - C_{1k^*}^2 - C_{3k^*}^2) \\
& \qquad -(C_{1k^*}^2 + C_{3k^*}^2)(\Lambda_1 + \Lambda_2 - C_{1k}^2 - C_{3k}^2)\Big] \geq 0) \\
=\ & \epsilon + P\left(\max_{k \in W_{n,M'}} (\Lambda_1 + \Lambda_2)(C_{1k}^2 - C_{1k^*}^2 + C_{3k}^2 - C_{3k^*}^2) \geq 0\right) \\
=\ & \epsilon + P\left(\max_{k \in W_{n,M'}} \left(C_{1k}^2 - C_{1k^*}^2 + C_{3k}^2 - C_{3k^*}^2\right) \geq 0\right)
\end{aligned}
$$

Because $C_{1k}^2 - C_{1k^*}$ and $C_{3k}^2 - C_{3k^*}^2$ have the same sign, formula (2.31) equals to $\epsilon + P\big(\max_{k \in W_{n,M'}}(C_{1k}^2 - C_{1k^*}^2) \geq 0\big)$ and will converge to 0 (Bai,1993). Thus we finish the proof of Theorem 2.2.4.

Remark 2.2.6 *Similar as the COS method in previous section, we estabished the asymptotic properies of test statistic T_{2k} in Proposition 2.2.2, Theorem 2.2.3 and Theorem 2.2.4. We can derive the critical value for test by Proposition 2.2.2. By*

Theorem 2.2.3, change point estimates convergent to the true change point with convergence rate $O(n)$, as shown in Theorem 2.2.4.

2.2.3 CESCP algorithm

We present a CESCP algorithm to detect a single change point in sample $X_1, \cdots, X_n$. Let $T_k(t)$ be the test statistic of the proposed test. For COS method, $T_k(t) = T_{1k}(t)$ while for EXP method, $T_k(t) = T_{2k}(t)$. The algorithm is given as follows

1. Choose equally spaced values from a prechosen interval $[a, b]$.

2. Set $t_0 = \arg \max_{t \in [a,b]; k} |T_k(t)|$.

3. Calculate $T_k(t_0)$ based on the sample $X_1, \cdots, X_n$.

4. Determine the critical value by Propositions 2.2.1 and 2.2.2.

5. If $\max_k |T_k(t_0)|$ is smaller than the critical value, there is no change point existing, otherwise, there exists a change point in the sample and the change point estimate is given by $\hat{k} = \arg \max_k |T_k(t_0)|$.

2.2.3.1 Type I Error and Power of Tests

In order to finite sample performance of the test, we perform the following simulation studies. We generate 500 samples with sample size $n = 100, 200$ and 300 for

n	100			200			300		
method	COS	EXP	ECF	COS	EXP	ECF	COS	EXP	ECF
N(1,1)	0.064	0.016	0.052	0.056	0.004	0.062	0.046	0.014	0.076
$\Gamma(3,2)$	0.072	0.024	0.066	0.053	0.040	0.056	0.508	0.048	0.050

Table 2.1: Rejection rates under the null hypothesis corresponding to 5% significant level for COS, EXP and ECF methods.

analysis. Firstly the test statistics for COS, EXP methods are calculated based on the original sample $X_1, X_2, \cdots, X_n$. Then the critical value of the test is approximated by Proposition 2.2.1 and 2.2.2, respectively. We also include the results of ECF method for comparison. For ECF method, we randomly choose B permutations of $(1, 2, \cdots, n)$ from all $n!$ total number of permutations. For each permutation, the test statistic is calculated based on the permuted data and the critical value is determined by $(1-\alpha)100\%$ quantile of the permutation distribution. Here, we choose $B = 100$.

We perform the simulation study to calculate the type I error. We choose $N(1, 1)$ and $\Gamma(3, 2)$ and set $\alpha = 0.05$. The corresponding results are given in Table 2.1.

From the Table 2.1, it can be seen that ECF method performs well in these two scenarioes because it employs the permutation method to determine the critical value.

However, it has the weakness of spending much more time to calculate the critical value. For both COS and EXP methods, we use the asymptotic distributions of these two test statistics to determine the critical values and it will become more accurate with the increasing sample size. The results in Table 2.1 show that the type I errors of COS method are close to the significant level α, especially when the sample size n is large. For the first scenario when $X_1 \sim N(1,1)$, the EXP method has difficulty in converging to α and the performance of ECF method is getting worse with the increasing sample size. It is acceptable because none of the methods can perform well for all the model settings. After more simulation studies, we are confident that all of these three methods can determine the true critical value, especially when the sample size is large. For the second scenario, all of these three methods' rejection rates are close to 5%.

To explore the simulation results of the powers of our proposed tests, we follow the setting in Hušková and Meintanis (2006a) and then compare the powers of these three methods. Similarly, we set $F_n(x) = F_1[(x-\delta)/b]$ where $\delta = 0.7$ and $b = 1.1$. Here we consider two cases:

1. $X_1 \sim N(1,1)$, thus $F_1(x)$ is the CDF of $N(1,1)$ distribution;

2. $X_1 \sim \Gamma(3,2)$, thus $F_1(x)$ is the CDF of $\Gamma(3,2)$ distribution.

For both cases, the location of change point is set to be $\tau_0 = 0.5$. The simulation

n	100			200			300		
method	COS	EXP	ECF	COS	EXP	ECF	COS	EXP	ECF
case 1	0.746	0.684	0.830	0.976	0.986	0.992	1	1	1
case 2	0.962	0.994	0.998	1	1	1	1	1	1

Table 2.2: Rejection rate under alternative hypothesis corresponding to 5% significant level for COS, EXP and ECF methods.

results are presented in Table 2.2.

From Table 2.2, it can be seen that all of these three methods achieve high powers for both cases. These results show that our proposed methods are powerful in detecting change point, and it is consistent with the asymptotic results.

2.2.4 Simulation Study of Single Change Point Model

In this section, we will discuss the estimation of a change point and compare our proposed methods with ECF method. The sample size is set to be $n = 100, 200, 300$ while the true change point $\tau_0 = k^*/n$ is $0.3, 0.4, 0.5, 0.6$ and 0.7, respectively. We consider two pairs of distributions for F_1 and F_n: N(1,1) against N(2,1), χ^2_3 against $\Gamma(3, 2)$, in our simulation study. For ECF method, we set $\gamma = 1$ and $a = 1$ in formula (2.3). For each parameter setting, we simulate 500 times and present the

N(1,1) v.s. N(2,1)															
	τ_0=0.3			τ_0=0.4			τ_0=0.5			τ_0=0.6			τ_0=0.7		
n	COS	EXP	ECF	COS	EXP	ECF	COS	EXP	ECF	COS	EXP	ECF	COS	EXP	ECF
100	75.8	81.4	72.2	76.8	81.0	80.2	72.6	78.0	80.4	76.4	80.8	80.0	72.6	77.8	68.0
200	90.2	93.4	84.8	88.2	92.4	88.6	90.4	92.4	92.6	88.4	91.4	88.8	85.6	91.8	82.2
400	97.6	99.6	91.6	98.4	98.2	97.2	97.6	98.6	98.6	97.8	99.2	97.6	97.4	98.8	91.8
χ^2_3 v.s. $\Gamma(3,2)$															
	τ_0=0.3			τ_0=0.4			τ_0=0.5			τ_0=0.6			τ_0=0.7		
n	COS	EXP	ECF	COS	EXP	ECF	COS	EXP	ECF	COS	EXP	ECF	COS	EXP	ECF
100	60.2	64.0	61.8	62.4	63.2	66.0	63.0	67.2	70.4	61.2	63.8	66.0	58.8	62.8	55.6
200	73.0	85.0	71.2	81.2	83.4	82.8	71.0	81.4	83.0	69.8	79.6	77.6	73.0	84.6	72.8
400	81.8	91.0	82.0	89.8	89.2	90.2	83.4	90.8	91.2	85.6	91.6	86.4	83.0	91.2	78.4

Table 2.3: Percentage of successful dection of change point by using COS, EXP and ECF methods based on 500 simulations for different model setting.

percentage of "success" in Table 2.3, here "success" means it can test the existence of the change point and the estimated location of the change point is within the interval $[\tau_0 - n * 2.5\%, \tau_0 + n * 2.5\%]$.

From Table 2.3, we can see that by using the test statistics proposed in this chapter, we can obtain the results that are of about the similar accuracy as that are derived by ECF method. Comparing with ECF method, COS method has better

performance when the true change point is away from 0.5. EXP method has satisfying performance in detecting the location of change points in the simulation study. The reason that ECF method does well when $\tau_0 = 0.5$ is that the weight function in model (2.3) is symmetry and has larger value when t is close to 0.5. On the other side, when the true location of change point τ_0 is away from 0.5, our methods are better than ECF. What is more, we can conclude that the accuracies of these three estimators are getting better with the increasing sample size.

The test statistics proposed in this chapter have their own advantage that they are much more efficient than ECF method. We present the elapsed time for each method in Figure 2.4. From the graph, we can find that the time consumed by ECF method increases in exponential rate, while for our proposed methods, it increases linearly. Furthermore, from Figure 2.5, it is easy to find that COS method is more efficient than EXP method.

2.3 Detection of Multiple Change Points

In this section, we focus on the multiple change points detection problems based on sample $X_1, \cdots, X_n$, such that

$$X_i \sim F_k(x), \qquad k_{m-1} \leq i \leq k_m - 1, \quad m = 1, \cdots, K+1; \quad i = 1, \cdots, n, \quad (2.31)$$

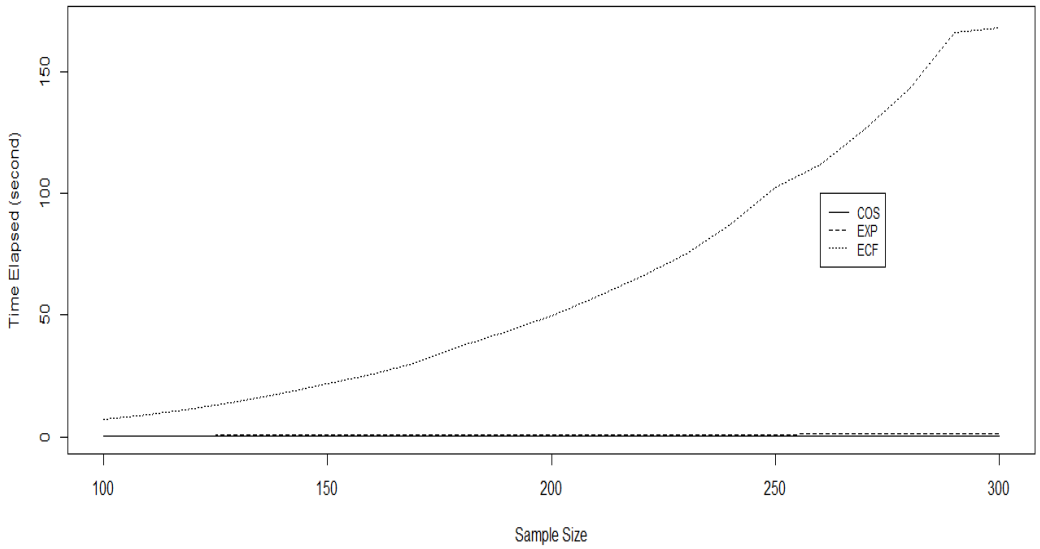

Figure 2.4: Solid line, dashed line and dotted line denote the time elapsed for 500 iterations of COS method, EXP method and ECF method, respectively.

where K is the true number of change points. k_m's are the locations of these change points with the convention of $k_0 = 1$ and $k_{K+1} = n+1$, and F_k is the cumulative distribution function satisfying $F_k \neq F_{k+1}$. In the following sections, we will introduce some existing methods to detect multiple change points in literature.

2.3.1 NMCD Method

Zou et al. (2014) proposed a nonparametric maximum likelihood approach to detect multiple change points in the data sequence. The idea is as following: if

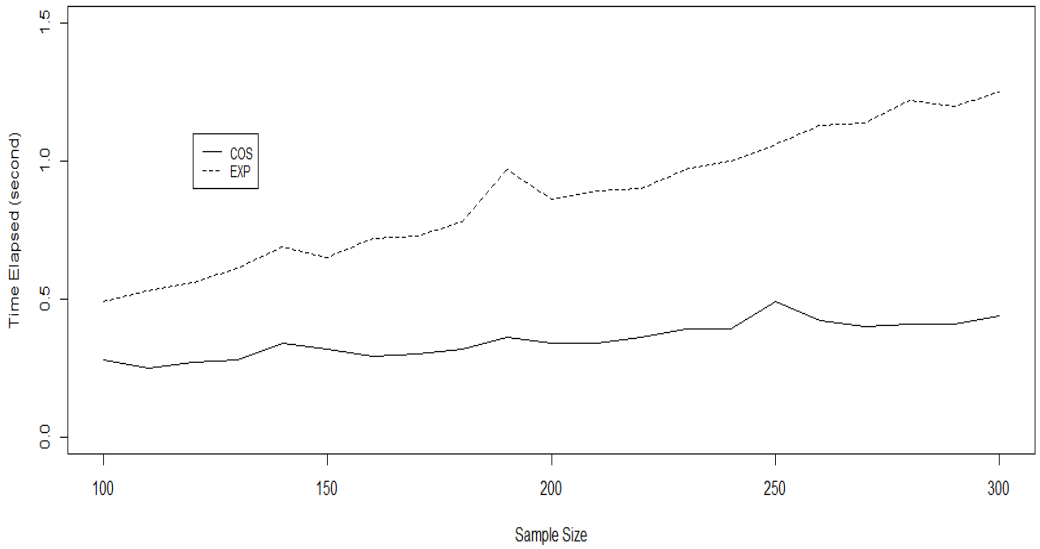

Figure 2.5: Solid line and dashed line denote the time elapsed for 500 iterations of COS method and EXP method, respectively

we assume that $X_1, \cdots, X_n$ are independently and identically distributed random variables following the distribution F_0, and let $\hat{F}_n$ denote the empirical CDF of the sample, then $n\hat{F}_n(\mu) \sim Binomial(n, F_0(\mu))$. Zou et al. (2014) regarded the sample as binary data with the probability of success $\hat{F}_n(\mu)$, the corresponding nonparametric maximum log-likelihood is

$$n\{\hat{F}_n(u)\log(\hat{F}_n(u)) + (1-\hat{F}_n(u))\log(1-\hat{F}_n(u))\}. \tag{2.32}$$

In the context of model (2.31), the joint log-likelihood for a candidate set of change points $(k'_1 < \cdots < k'_L)$ can be written as

$$\begin{aligned}\mathcal{L}_u(k'_1, \cdots, k'_L) &= \sum_{i=0}^{L}(k'_{i+1} - k'_i)\{\hat{F}_{k'_i}^{k'_{i+1}}(u)\log(\hat{F}_{k'_i}^{k'_{i+1}}(u)) \\ &\quad +(1 - \hat{F}_{k'_i}^{k'_{i+1}}(u))\log(1 - \hat{F}_{k'_i}^{k'_{i+1}}(u))\}, \qquad (2.33)\end{aligned}$$

where $\hat{F}_{k'_i}^{k'_{i+1}}(u)$ is the empirical CDF of the subsample $\{X_{k'_i}, \cdots, X_{k'_{i+1}}\}$ with $k'_0 = 1$ and $k'_{L+1} = n + 1$. To estimate the change points $1 < k'_1 < \cdots < k'_L \leq n$, Zou et al (2014) proposed the idea of maximizing formula (2.33) in an integrated form

$$R_n(k'_1, \cdots, k'_L) = \int_{-\infty}^{+\infty} \mathcal{L}_u(k'_1, \cdots, k'_L)d\omega(u),$$

where $\omega(\cdot)$ is some positive weight function so that $R_n(\cdot)$ is finite, and the integral is used to combine all the information across u. Zou et al. (2014) established the consistency of the NMCD method and propose the screening algorithm to reduce computational complexity. The performance of NMCD method is satisfactory in the real data analysis and the simulation studies.

2.3.2 E-Divisive Algorithm

Suppose that $\mathbf{X}$ and $\mathbf{Y}$ are $d-$dimensional random vectors and they follow distribution F and G, respectively. Suppose the characteristic functions of X and Y are $\phi_x(t)$ and $\phi_y(t)$, respectively. Szśekely and Rizzo (2010) introduced the following

divergence measure that can determine whether two independent random vectors are identically distributed or not.

$$\int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \omega(t) dt,$$

in which $\omega(t)$ is any positive weight function to make sure that the above integral is defined.

By choosing a suitable weight function, Matteson and James (2014) rewrote the divergence measure as

$$\mathcal{D}(\mathbf{X}, \mathbf{Y}; \alpha) = \int_{\mathbb{R}^{\mathbf{d}}} |\phi_{\mathbf{x}}(\mathbf{t}) - \phi_{\mathbf{y}}(\mathbf{t})|^{\mathbf{2}} \left(\frac{\mathbf{2}\pi^{\mathbf{d/2}}\mathbf{\Gamma}(\mathbf{1} - \alpha/\mathbf{2})}{\alpha \mathbf{2}^{\alpha}\mathbf{\Gamma}[(\mathbf{d} + \alpha)/\mathbf{2}]|\mathbf{t}|^{\mathbf{d}+\alpha}} \right)^{-\mathbf{1}} \mathbf{dt}.$$

for some fixed constant $\alpha \in (0, 2)$.

An alternative divergence measure based on Euclidean distances may be defined as follows

$$\mathcal{E}(\mathbf{X}, \mathbf{Y}; \alpha) = \mathbf{2E}|\mathbf{X} - \mathbf{Y}|^{\alpha} - \mathbf{E}|\mathbf{X} - \mathbf{X}'|^{\alpha} - \mathbf{E}|\mathbf{Y} - \mathbf{Y}'|^{\alpha}.$$

In the above equation, $\mathbf{X}'$ and $\mathbf{Y}'$ are independent copies of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Matteson and James (2014) presented the above E-Divisive method to perform hierarchical divisive estimation of multiple change points. As summarized by James and Matteson (2015), the way to estimate multiple change points for E-Divisive method is to iteratively apply a procedure for locating a single change point. The

progression of this method can be regarded as a binary tree because at each iteration, a new estimated change point will divides an existing segment. The root node of the tree corresponds to the case of no change points, and thus contains the entire time series. All other non-root nodes are either a copy of their parent, or correspond to one of the new segments created by the addition of a change point to their parent. Details on the estimation of change point locations can be found in Matteson and James (2014).

The time complexity of this method is $O(Kn^2)$, where K is the number of estimated change points, and n is the number of observations in the series. We may find the corresponding R function in the "ecp" package. In the current dissertation, we will employ E-Divisive algorithm for comparison. As our change point model is based on bivariate case, the E-Divisive will loss some of its advantages in computation efficient. However, it is still time consuming when comparing with our proposed methods. One can refer the simulation studies for more details.

2.3.3 E-Agglo Algorithm

We now present the E-Agglo method (Matteson and James 2014) which performs hierarchical agglomerative estimation of multiple change points. As concluded by James and Matteson (2015), the E-Agglo algorithm requires an initial segmen-

tation in order to reduce the computational complexity. It also allows to include a prior knowledge of possible change point locations. If no such assumptions are made, each observation can be assigned to its own segment. After the initial segmentation, neighboring segments are then sequentially merged to maximize a goodness-of-fit statistic. The estimated change points are determined by the iteration which maximizes the penalized goodness-of-fit statistic. When using the E-Agglo procedure it is assumed that there is at least one change point existing in the data sequence.

The goodness-of-fit statistic used in Matteson and James (2014) is the between-within distance (Székely and Rizzo 2005) among adjacent segments. Let $C = \{C_1, \cdots, C_K, C_{K+1}\}$ be a segmentation of the n observations into $K+1$ segments. The goodness-of-fit statistic is defined as

$$\hat{S}_{K+1}(\mathcal{C}; \alpha) = \sum_{i=1}^{K} \hat{\mathcal{Q}}(C_i, C_{i+1}; \alpha).$$

Since calculating the true maximum of the goodness-of-fit statistic for a given initial segmentation is computationally intensive, an advanced algorithm is used to find an approximate solution (James and Matteson (2015)). If overfitting is a concern, it is possible to penalize the sequence of goodness-of-fit statistics. Thus, the change point locations are estimated by maximizing

$$\tilde{S}_k = \hat{S}_k + \text{penalty}(\boldsymbol{\tau}(k)),$$

where $\boldsymbol{\tau}(k) = \{\tau_1, \tau_2, \cdots, \tau_k\}$ is the set of change points associated with the goodness-of-fit statistic $\hat{S}_k$. The E-Agglo method is quadratic in the number of observations with computational complexity $O(n^2)$. One can refer to James and Matteson (2014) for more information about the algorithm. The corresponding R function can be found in the "ecp" package.

2.3.4 ICSS Algorithm

Inclán and Tiao (1994) proposed a procedure to detect variance changes based on iterated cumulative sums of squares (ICSS) algorithm. Now we employ this algorithm and the test statistic T_k to detect the multiple change points in this section. The key of ICSS algorithm is the iterative scheme based on successive application of statistic to pieces of the series, dividing consecutively after a possible change point is found . We use $X[l_1 : l_2]$ to represent the piece of series $X_{l_1}, X_{l_1+1} \cdots, X_{l_2}$, here $l_1 < l_2$, and use the notation $T_k(X[l_1 : l_2])$ to indicate the test statistic $T_k(t)$ based on $X[l_1, l_2]$. $k^*(X[l_1 : l_2])$ is used to denote the point at which $\max_k T_k(X[l_1 : l_2])$ obtained and $M(X[l_1 : l_2])$ is the maximum value. Another useful value is the critical value which is denoted by $CV(X[l_1 : l_2])$. We can derive it by Theorem 2.2.1 and Theorem 2.2.3. The algorithm is given as follows

Stage A: find all possible change points.

1. Set $l_1 = 1, l_2 = n$, and calculate $M(X[l_1 : l_2]), k^*(X[l_1 : l_2])$; use $\mathcal{C}$ to store the change points.

2. while$(M(X[l_1 : l_2]) > CV(X[l_1 : l_2]))\{$

3. $\quad k_{first} = k_{last} = k^*(X[l_1 : l_2]);$

4. $\quad M_1 = M(X[l_1 : k_{first}]); k_1 = k^*(X[l_1 : k_{first}]);$

5. $\quad$ while$(M_1 > CV(X[l_1 : k_{first}]))\{$

6. $\quad\quad k_{first} = k_1; M_1 = M(X[l_1 : k_{first}]);$

7. $\quad\}$End 'while' in (5)

8. $\quad M_2 = M(X[k_{last} : l_2]); k_2 = k^*(X[k_{last} : l_2]);$

9. $\quad$ while$(M_2 > CV(X[k_{last} : l_2]))\{$

10. $\quad\quad k_{last} = k_2; M_2 = M(X[k_{last} : l_2]);$

11. $\quad\}$ End 'while' in (9)

12. $\quad$ if $(k_{first} == k_{last})$

13. $\quad\quad$ there is only one change in $[l_1 : l_2]$, save it in $\mathcal{C}$ and end the loop.

14. $\quad$ Else save the two candidate change points in $\mathcal{C}$ and continue

15. End 'if' in (12)

16. Reset $l_1 = k_{first}, l_2 = k_{last}$;

17. } End 'while' in (2)

Stage B: refine the change points (if there are two or more candidate change points).

18. Sort the locations of change points and denoted by $\mathbf{P}$ with the length of N, define two extreme values $\mathbf{P}_0 = 0$ and $\mathbf{P}_{N+1} = n$

19. Do {For $j = 1, \cdots, N$, check whether possible change points exist between $[\mathbf{P}_{j-1} + 1 : \mathbf{P}_{j+1}]$

20. If Yes, keep the point If Not, eliminate it.

21. } Until the number of change points doesn't change.

Remark 2.3.1 *In application, we set a new rule that the distance between the two nearby detected change points can't be smaller than* 10. *Thus, the Stage A of the algorithm will end when* $k_{last} - k_{first} < 10$.

2.4 Data Analysis

In this section, we provide the simulation studies and real data analysis to illustrate the performance of COS and EXP methods proposed in the current chapter.

2.4.1 Simulation Study

We will compare the simulation performance of the COS, EXP, NMCD (Zou et.al. 2014), E-Divisive (Matteson and James 2014) and E-Agglo (Matteson and James 2014) methods on various sequences. We will evaluate the performance of these methods by the following aspects:

1. the accuracy of successfully detecting each true change point;

2. the accuracy of successfully detecting all true change points under the condition that the number of true change points is correctly estimated;

3. elapsed running time in seconds.

$S1$: we set the sample size $n = 1000$ and number of change point $K = 1$. Observations before the change point follow $N(0, 0.6^2)$ while observations after the change point follow standard normal distribution. The location of the change point is $\tau_0 = 0.5$. We will simulate 500 times and then present the results in Table 2.4.

For the other two simulation studies, the sample size is set to be 1000 and there existing 5 change points in each data sequence. The locations of change points are $(162, 310, 511, 653, 805)$ and thus, the data sequence is divided into six pieces. The following is the model settings of these two simulation studies:

$S2$: for each segment, observations are sampled from $N(0, 0.6^2)$, $N(1.2, 1)$, $N(2.4, 1)$, $N(1.3, 1)$, $N(0, 0.7^2)$, and $N(1, 1)$, respectively. Simulated data for $S2$ is plotted in Figure 2.6.

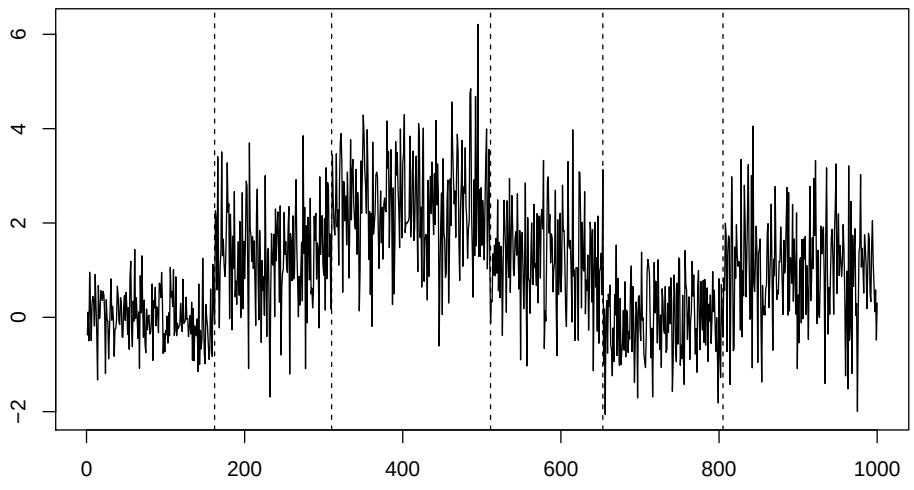

Figure 2.6: Simulated data for $S2$. The dashed lines denote the locations of true change points.

$S3$: For each segment, observations are sampled from $N(0,1)$, $\log N(0.8,1)$, $\Gamma(2,2)$, χ^2_3, $N(3.5,1)$, and noncentral t-distribution with 2 degrees of freedom and noncentrality parameter equals 2, respectively. Simulated data for $S3$ is plotted in Figure 2.7.

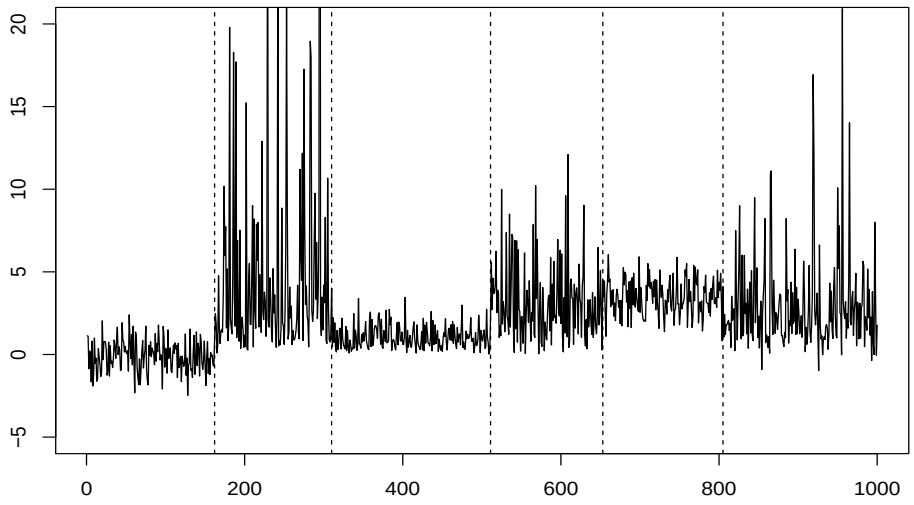

Figure 2.7: Simulated data for $S3$. The dashed lines denote the locations of true change points.

For each of the model settings $S1$, $S2$ and $S3$, we first generate a data sequence, and then apply all five methods COS, EXP, NMCD, E-Divisive and E-Agglo, to detect multiple change points in the dataset. We denote the set of estimated change

points in the mth simulation by $\mathcal{K}^{(m)}$ and define

$$A_{k_i}^{(m)} = \mathcal{K}^{(m)} \cap [k_i - 10, k_i + 10], \tag{2.34}$$

where $i = 1, \cdots, K$. Obviously, $K = 1$ for the first scenario and $K = 5$ for the other two scenarios. Actually, $A_{k_i}^{(m)}$ is the set of estimated change points, derived from the mth simulation, lying in the neighborhood of the ith true one. Here we regard 10 as a safe distance to determine whethere the estimated change point is a true one or not. Moreover, we define the function $J(x)$ as

$$J(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ 1 & \text{otherwise.} \end{cases} \tag{2.35}$$

Furthermore, we define $B^{(m)} = 1$ if $J(A_{k_i}^{(m)}) = 1$ for all $i = 1, \ldots, K$ and the size of $\mathcal{K}^{(m)}$ is exactly K; $B^{(m)} = 0$, otherwise. Note that $B^{(m)} = 1$ if and only if the mth simulation is successful in the sense that it detects exactly five change points and all of these five estimated change points are close to the corresponding exact change points. Here, "close" means the distance between these two locations is within 10. With the aid of $B^{(m)}$, we define

$$\text{ALLCP} = \sum_{m=1}^{M} B^{(m)}/M,$$

the successful simulations as a percentage of all simulations. Here, M is the number of simulations. In the current section, we choose $M = 500$.

Method	*COS*	*EXP*	*NMCD*	*E-Divisive*	*E-Agglo*
$\sum_{m=1}^{1000} J\big(A_{k_1}^{(m)}\big)^*$	380	370	331	353	478
cpnumber.R	495	496	417	472	0
ALLCP(%)	75.2	73.4	56.8	67.0	0
ERT.S	6.92	18.47	840.54	14625.74	311.86

$*$ $A_{k_i}^{(m)}$ is defined in formula (2.34) and $J(A)$ is given in formula (2.35).

Table 2.4: Simulation results of COS, EXP, NMCD, E-Divisive and E-Agglo based on 500 simulations for scenario $S1$.

The simulation results are reported respectively in Table 2.4 and 2.5. In the tables, "cpnumber.R" stands for the number of simulations in which the true number of change points is correctly estimated. "ALLCP" denotes the percentage of simulations in which exactly one change point is estimated, and the estimated change point is close to the corresponding exact change point. ERT.S means the total running time in seconds.

From Table 2.4, we can find E-Agglo performs best in detecting all of the change points. However, E-Agglo over-estimates the number of change points because its "cpnumber.R" is 0. COS and EXP methods perform better than NMCD and E-Divisive methods in terms of detecting the location and number of change points.

We present one of the simulated data and the estimated change points by COS and NMCD methods in Figure 2.8. From the figure, it can be seen that the NMCD method, which estimates three changes, over-estimate the number of change points in scenario $S1$.

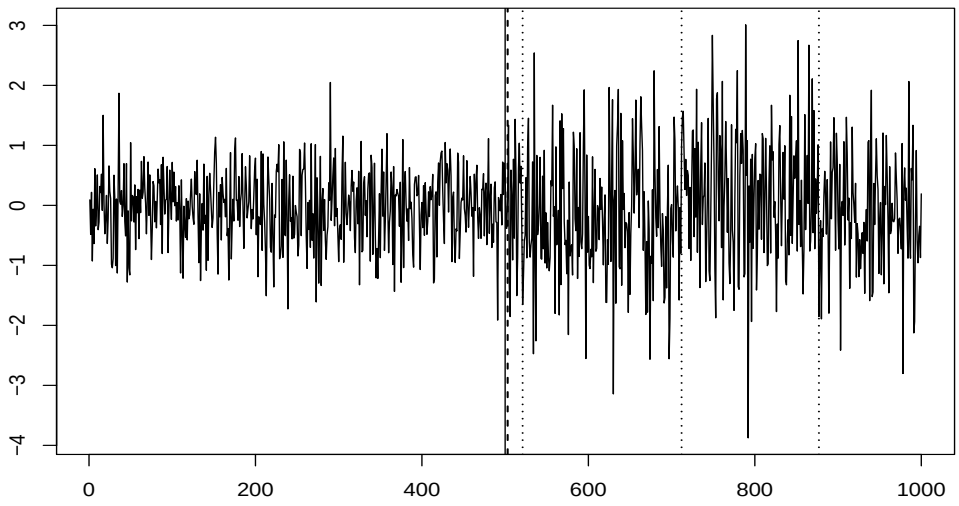

Figure 2.8: Simulated data for scenario $S1$. The solid line denote the location of true change point. The dashed line and dotted line denote the estimated change point by COS method and NMCD method, respectively.

What is more, from the value of "ALLCP", we can conclude that COS and EXP methods perform well in terms of estimating the number of change points and

Scenario	Scenario 2					Scenario 3				
Method	*COS*	*EXP*	*NMCD*	*E-Divisive*	*E-Agglo*	*COS*	*EXP*	*NMCD*	*E-Divisive*	*E-Agglo*
$\sum_{m=1}^{1000} J(A_{k_1}^{(m)})^*$	482	492	486	463	361	462	497	499	498	475
$\sum_{m=1}^{1000} J(A_{k_2}^{(m)})$	478	481	470	474	96	427	440	479	478	459
$\sum_{m=1}^{1000} J(A_{k_3}^{(m)})$	472	478	460	468	47	428	458	478	416	441
$\sum_{m=1}^{1000} J(A_{k_4}^{(m)})$	466	488	488	476	373	429	451	459	464	11
$\sum_{m=1}^{1000} J(A_{k_5}^{(m)})$	431	485	474	483	229	452	461	444	461	1
cpnumber.R	448	481	486	434	0	382	451	488	460	2
ALLCP(%)	69.0	83.8	76.0	66.6	0	57.2	68.4	73.6	64.2	0
ERT.S	20.07	54.77	1070.72	32037.44	390.50	25.02	61.79	1124.02	32657.34	410.09

$*$ $A_{k_i}^{(m)}$ is defined in formula (2.34) and $J(A)$ is given in formula (2.35).

Table 2.5: Simulation results of COS, EXP, NMCD, E-Divisive and E-Agglo based on 500 simulations for scenario $S2$ and $S3$.

the locations of change points simultaneously. The elapsed time of COS and EXP methods show that the proposed methods are more efficient when comparing with the other methods.

We observe from Table 2.5 that EXP method outperforms other methods in terms of accuracy for detecting change point locations. COS, EXP, NMCD and E-Divisive methods have similar performance and yield also good estimators of the true change points. E-Agglo performs unsatisfactory in scenario $S2$ and $S3$. There are two reasons

leading this. The fist one is that the change between two segments is not significant enough for E-Agglo method to detect and the second reason is that the performance of E-agglo method is highly influenced by the choice of initial segmentation.

If we compare these three methods in terms of ERT.S, we find that COS takes least time to estimete change points. EXP and E-Agglo methods take more time than COS method but can also be regard as efficient methods. NMCD is slow in detecting multiple change points and it takes approximate 2 seconds to analyze a sequence of size 1000. E-Divisive is the slowest method because it takes long time to perform the permutation in order to determine the p-value while testing the statistical significance of an estimated change point.

2.4.2 Real Data Analysis

As we mentioned at the beginning of Chapter 2, a picture can be transformed to a matrix and each row or column can be regarded as a data sequence. In this section, we will focus on the pattern recognition of letter "E" (Wang and Wang 2006). We add Gaussian white noise to the image to make the graphic more reasonable and practical. To de-noise, we first convert the noised image of the letter "E" to the image matrix of dimensions 542×719 and then apply COS and EXP methods to every row and column of this image matrix for image retrieval. The restored images

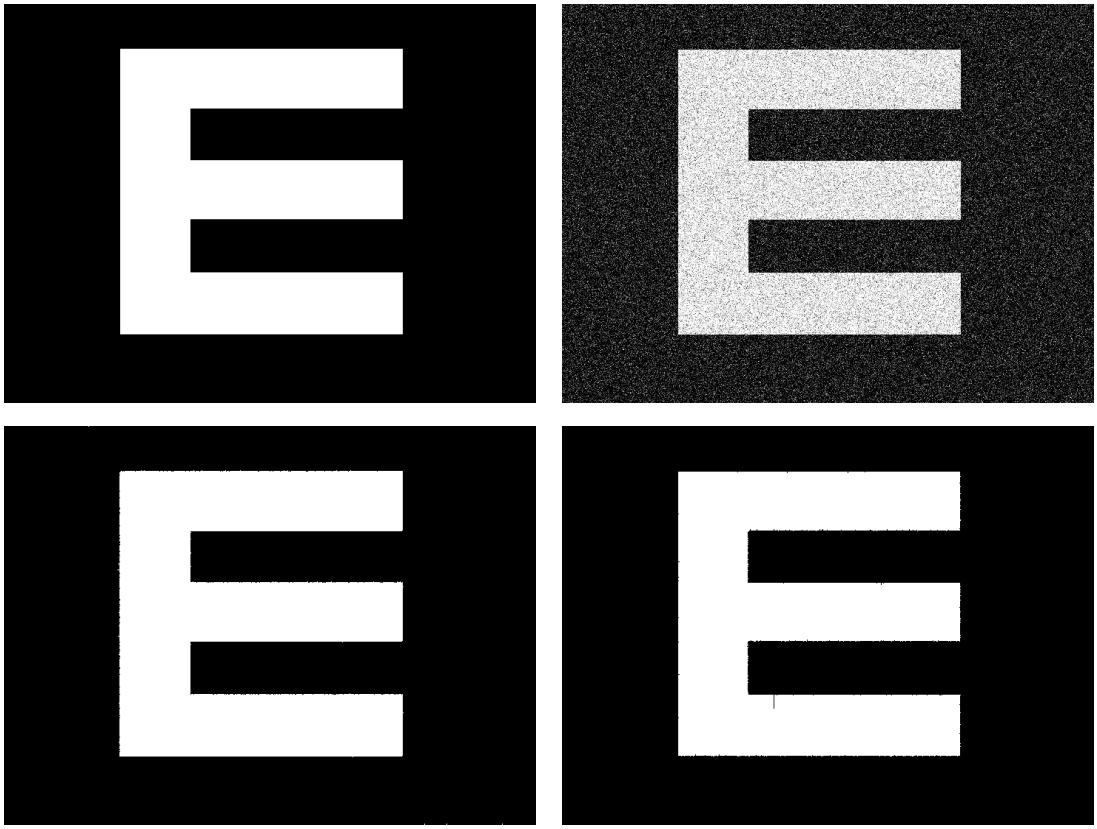

Figure 2.9: The upper left image is the original letter "E"; upper top image is the letter "E" with noise; the lower left image is processed by the COS method and the lower right one is processed by EXP method

after applying COS and EXP methods are displayed in Figure 2.9, which shows that the image of the letter "E" is retrieved successfully.

To compare with the other multiple change points detection methods, we present the de-noised image by NMCD and E-Divisive methods in Figure 2.10. We ignore

E-Agglo method because the performance of this method is poor and the de-noised image is mussy. From Figure 2.10, we can find E-Divisive performs well but NMCD has weak performance when determining the locations of change points. In practice, how to de-noise an image efficiently is another important aspect we should consider. In our example, we can find COS is the most efficient one because it takes only 13.89 second while EXP, NMCD and E-Divisive takes 40.31 seconds, 1023.65 seconds and 21624.59 seconds (approximate 6 hours), respectively.

Suppose the transformed matrix of "E" (without noise) is $\boldsymbol{X}_{n\times p}$, and the de-noised image by change point detection method is $\boldsymbol{Z}_{n\times p}$. We define $DIFF = \sum_{i=1}^{n}\sum_{j=1}^{p}|X_{ij}-Z_{ij}|$ with the purpose of comparing the difference between these two matrix. For these four methods, which including COS, EXP, NMCD and E-Divisive method, the corresponding $DIFF$ is $206, 257, 321$ and 169, respectively. After comparing the values of $DIFF$ for these four methods, we can find E-Divisive method performs best among these four methods. What is more, there are 389698 pixels in the image and most of the pixels are corrected estimated, so the results obtained by all of these methods are satisfactory.

Figure 2.10: The left image is processed by the NMCD method and the right one is processed by E-Divisive method.

2.5 Discussion

To detect the potential change point in a data sequence, we propose two test statistics based on empirical characteristic function and then establish COS and EXP methods for change point detection. From the simulation studies, we can conclude that these two methods are computationally efficient and are able to estimate the change point locations very well. Comparing with Hušková (2006a)'s method, these two methods' performance is satisfactory.

We extend our methods to multiple change points detection problems by the use of ICSS algorithm. In the simulation study, we compare our methods with both E-Divisive and E-Agglo methods and find that ours perform more effective and efficient.

In the real data analysis, we employ our methods to de-noise the "E" plot and find that the image can be retrieved successfully.

3 A Sequential Multiple Change Point Detection Procedure via VIF regression

3.1 Introduction

In this data-rich era, many data sequences have a very large size, and thus it is not surprising that multiple change points might occur in such a data sequence. It becomes desirable to find a fast and efficient method to detect the locations of these change points. Recent literature in this area includes Harchaoui and Lévy-Leduc (2008, 2010), Killick *et al.* (2012), Jin *et al.* (2013) among others. In this chapter, we will tackle the problem of multiple change point detection in a mean-shift model given below

$$y_i = \sum_{r=0}^{b} \mu_r I_{\{k_r,\ldots,k_{r+1}-1\}}(i) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{3.1}$$

where $I_A(\cdot)$ denotes the indicator function of the set A; $1 < k_1 < \cdots < k_b < n$ are the unknown locations of b change points satisfying $\lim_{n\to\infty} \min_r (k_r - k_{r-1})/n > 0$;

$\mu_0, \cdots, \mu_b$ are the means such that $\mu_r \neq \mu_{r+1}$ for $0 \leq r \leq b-1$; and $\varepsilon_1, \cdots, \varepsilon_n$ are random errors with zero mean. Here, we have used the convention that $k_0 = 1$ and $k_{b+1} = n + 1$. We denote the set of change points by $\mathcal{K} = \{k_1, \cdots, k_b\}$.

Let us illustrate the application of multiple change point detection by the following example. Consider the problem of recognizing a one-dimensional barcode that encodes 0123456789 in the top panel of Figure 3.1 (`http://barcode.tec-it.com/barcode-generator.aspx`). When the image is converted into matrix form, all of the values in the matrix lie between 0 (black pixel) and 1 (white pixel). It is noted that all rows in this matrix are identical, and $\min_r(k_r - k_{r-1})$ in any row is 40. The barcode recognition problem here can be converted into a multiple change point detection problem in a mean-shift model. Decontaminating the barcodes is equivalent to finding the set of change points $\mathcal{K}$. To simulate the scanned input, we add two levels of noise to each element of the matrix. The resulting data are left-truncated at 0 and right-truncated at 1, which yields two barcodes, shown respectively in panels 2-3 in Figure 3.1.

In addition to the barcode recognition problem, the detection of multiple change points has many applications in areas such as genetic data analysis (see, e.g., Barry and Hartigan 1992, 1993; Erdman and Emerson 2007, 2008) and signal processing (see, e.g., Qu and Tu, 2006).

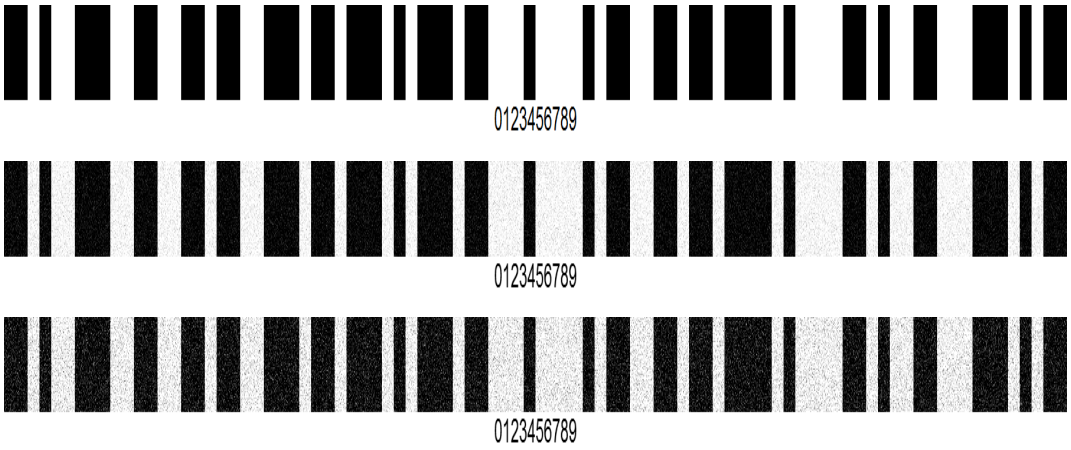

Figure 3.1: Top panel: the original barcode encoding 0123456789 without noise. Middle panel: the original barcode contaminated by the added Gaussian noise with mean zero and $\sigma = 0.1$. Bottom panel: the original barcode contaminated by the added Gaussian noise with mean zero and $\sigma = 0.2$.

There is a great need for efficient methods for detecting multiple change points. Barry and Hartigan (1993) proposed a Bayesian analysis for change point problems with the complexity of $O(n^3)$. This method was further improved by Erdman and Emerson (2008), who reduced the computation time to $O(n)$. While there are some other methods with computational complexity of order $O(n^2)$ (see e.g. Auger and Lawrence (1989), Jackson et al. (2005) and Rigaill (2010)), Scott and Knott (1974) proposed a faster binary segmentation algorithm with only $O(n \log n)$ computational complexity. The main feature of this algorithm is that it only considers a subset of

the 2^{n-1} possible solutions (Killick and Eckley, 2013).

The *Circular Binary Segmentation* (CBS) and the *Pruned Exact Linear Time* (PELT) are two popular methods for detecting multiple change points in a mean-shift model. CBS was proposed by Olshen *et.al.* (2004) to detect change points in the genomic data, and has been implemented in the R package **DNAcopy** (Seshan and Olshen, 2015). PELT was proposed in Killick *et al.* (2012), and has also been implemented in the R package **changepoint** (Killick *et al.* 2014). The main idea behind PELT is to consider the data sequentially, and record the optimal segmentation at each step, for the data up to that step (Killick *et al.* 2012). The computation time of PELT is of order $O(n)$, but its R package **changepoint** is not stable when there are outliers in the data. Throughout this chapter, we use CBS and PELT to stand for the R packages **DNAcopy** and **changepoint**, respectively.

It is noted that by properly segmenting a data sequence, the multiple change point detection problem above can be equivalently expressed as a linear regression variable selection problem, with a large number of regression coefficients (see Harchaoui and Lévy-Leduc 2008; Jin *et al.* 2013 among others). Thus a modern variable selection method can be utilized to obtain a rough estimation of multiple change points. Recently, Lin *et al.* (2011) proposed the variance inflation factor (VIF) regression algorithm for variable selection. This algorithm is much faster than many modern

variable selection methods including LASSO and SCAD. In this chapter, we modify the stagewise regression of the VIF regression algorithm, and perform the variable selection sequentially in segment order. Once the segment containing a possible change point is flagged, we adopt a weighted cumulative sum to justify and locate the change point in this segment. The proposed procedure is implemented by the algorithm VIFCP ("CP" stands for "change point"). We would like to remark that our new algorithm allows the number of change points to increase with the sample size, which makes our method applicable to various practical problems.

The rest of this chapter is organized as follows. In Section 3.2, the proposed procedure VIFCP is presented in detail and its theoretical justification is provided. In Section 3.3, we run simulation studies to examine the proposed procedure and to compare its performance with CBS and PELT. In Section 3.4, we give two real data examples. We conclude the chapter in Section 3.5.

The following notation is used throughout the rest of this chapter. Let $\{c_n\}$ be a sequence of nonnegative numbers and $\{d_n\}$ be a sequence of positive numbers. If the sequence $\{c_n/d_n\}$ is bounded, it is denoted as $c_n = O(d_n)$. If $c_n/d_n \to 0$ as $n \to \infty$, it is denoted as $c_n = o(d_n)$. If $c_n/d_n \to 1$ as $n \to \infty$, it is denoted as $c_n \sim d_n$. If a sequence of random variables $\{\xi_n\}$ tends to 0 in probability, it is denoted as $\xi_n = o_p(1)$. The symbol $\xrightarrow{d}$ denotes convergence in distribution. For convenience, we

denote the $m \times 1$ vectors $(1, \ldots, 1)^T$ and $(0, \ldots, 0)^T$ by $\mathbf{1}_m$ and $\mathbf{0}_m$, respectively, and write $\boldsymbol{\ell}_{m_1,m_2} = (\mathbf{0}_{m_1}^T, \mathbf{1}_{m_2}^T)^T$. In addition, I_m stands for an $m \times m$ identity matrix (the subscript m may be suppressed if there is no confusion), $\|\cdot\|$ stands for the Euclidean norm, $\lfloor c \rfloor$ the largest integer less than or equal to a real number c, and $\Phi(\cdot)$ the cumulative distribution function of the standard normal random variable.

3.2 The VIFCP Procedure and its Theoretical Justification

To establish a connection between the multiple change point detection and variable selection, we follow the ideas of Harchaoui and Lévy-Leduc (2008) and Jin *et al.* (2013) to reformulate the model (3.1) as follows:

$$\boldsymbol{y}_n = \sum_{r=0}^{b} \gamma_r \boldsymbol{\ell}_{k_r-1,n-(k_r-1)} + \boldsymbol{\varepsilon}_n, \tag{3.2}$$

where $\boldsymbol{y}_n = (y_1, \cdots, y_n)^T$ is a column vector of n observations, γ_r with $r = 1, \cdots, b$ are the differences between two successive means $\mu_r - \mu_{r-1}$, and $\gamma_0 = \mu_0$, and $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \cdots, \varepsilon_n)^T$. Thus we can consider detecting multiple change points for model (3.1) as carrying out variable selection for model (3.2). It is noted that this variable selection problem is different from the traditional one, since $\mathcal{K}$ is unknown in model (3.2). Nevertheless, the problem can be solved by applying the multiple change point detection procedure as given below. The main idea of our new procedure is to divide the data sequence into smaller segments and sample each segment in

sequential order. If no change point is detected in a segment, the next segment is added to the collective pool of other segments that have been labeled as such. If this segment exhibits potential for containing a change point, it is flagged and a weighted cumulative sum (CUSUM) is applied to test if there is a change point in this segment.

3.2.1 Modified VIF Regression Algorithm and its Justification

We first introduce an artificial partition $\mathcal{Q} = \{q_1, \cdots, q_a\}$ which divides the set $\{1, \cdots, n\}$ into $a+1$ segments, where $l = \lfloor n/(a+1) \rfloor$ is the length of each segment excluding the first one. We set $q_s = n - (a+1-s)l$ for each $s = 1, \cdots, a$. Without loss of generality, we may assume that n is a multiple of $a+1$, and hence $q_s = sl$ with $l = n/(a+1)$ being the length of all segments. By convention, we also set $q_0 = 0$.

Note that each artificial segment contains at most one change point by the setup of model (3.1) and Assumption A1 below.

To reflect the artificial partition in model (3.2), we rewrite it as

$$\boldsymbol{y}_n = \sum_{s=0}^{a} \beta_s \boldsymbol{\ell}_{q_s, n-q_s} - \boldsymbol{\eta}_n + \boldsymbol{\varepsilon}_n. \tag{3.3}$$

The regression coefficients β_s (with $s = 1, \cdots, a$) are zeros, except when the artificial segment $[q_s + 1, q_{s+1}]$ contains a change point, say k_r, and in this case, $\beta_s = \gamma_r$. By convention, we set $\beta_0 = \gamma_0$. The error vector $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is defined in the

same way as in (3.2). Thus, we have correction vector $\boldsymbol{\eta}_n = \sum_{s=0}^{a} \beta_s \boldsymbol{\tau}_n(q_s)$ with $\boldsymbol{\tau}_n(q_s)$ being the zero vector $\mathbf{0}_n$ if $\beta_s = 0$, that is, no change point exists in the segment $[q_s + 1, q_{s+1}]$, and

$$\boldsymbol{\tau}_n(q_s) = \boldsymbol{\ell}_{k_r-1,n-(k_r-1)} - \boldsymbol{\ell}_{q_s,n-q_s} = (\mathbf{0}_{q_s}^T, \mathbf{1}_{k_r-1-q_s}^T, \mathbf{0}_{n-(k_r-1)}^T)^T$$

if $\beta_s = \gamma_r$, i.e., the rth change point $k_r \in [q_s + 1, q_{s+1}]$. By convention, $\boldsymbol{\tau}_n(q_0) = \mathbf{0}_n$. It is readily seen that $\boldsymbol{\eta}_n$ is a sparse vector, because the change points are sparse and the length of each artificial segment is comparably small. We would like to remark that if the artificial partition has exactly n segments, then model (3.2) reduces to the one studied by Harchaoui and Lévy-Leduc (2008). An illustration of the artificial partition is plotted in Figure 3.2, where $n = 10$, $\boldsymbol{\varepsilon}_{10} = \mathbf{0}_{10}$, $b = 2$, $k_1 = 4$ and $k_2 = 7$. The model (3.2) is

$$\boldsymbol{y}_{10} = \gamma_0 \mathbf{1}_{10} + \gamma_1 \boldsymbol{\ell}_{3,7} + \gamma_2 \boldsymbol{\ell}_{6,4}.$$

Given an artificial partition $\mathcal{Q} = (2, 4, 6, 8)$, namely, $a = 4$, $l = 2$ and $q_s = 2s$, this model can be re-expressed as follows:

$$\boldsymbol{y}_{10} = \beta_0 \mathbf{1}_{10} + \beta_1 \boldsymbol{\ell}_{2,8} + \beta_2 \boldsymbol{\ell}_{4,6} + \beta_3 \boldsymbol{\ell}_{6,4} + \beta_4 \boldsymbol{\ell}_{8,2} - \boldsymbol{\eta}_{10},$$

where $\beta_0 = \gamma_0$, $\beta_1 = \gamma_1$, $\beta_2 = 0$, $\beta_3 = \gamma_2$, $\beta_4 = 0$, and the correction vector $\boldsymbol{\eta}_{10} = (0, 0, \gamma_1, 0, 0, 0, 0, 0, 0, 0)^T$, which is symbolically illustrated in Figure 3.2.

As mentioned previously, we will adopt the VIF regression algorithm (Lin *et al.*

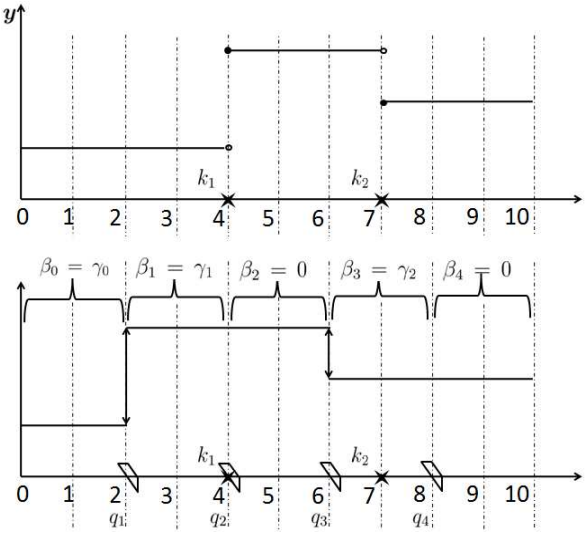

Figure 3.2: The upper plot is the observations $\boldsymbol{y}$ of size 10 without random errors; the one below is the symbolic illustration of a parametric transformation (without the correction vector) by an artificial partition. Here the signs 'star' and 'diagonal stripe' represent locations of change points and segments, respectively.

2011) because it is an extremely fast algorithm for variable selection with satisfactory accuracy. It consists of two steps: the search step and the evaluation step. The search step takes advantage of sparsity (i.e. the nonzero regression coefficients are sparse in the set of all regression coefficients). The evaluation step is similar to that of a variation of a stepwise regression, *forward stagewise regression*, which evaluates variables using only marginal correlations. A typical forward stagewise regression

can be used to test the following alternative model:

$$\boldsymbol{y} = \sum_{j=0}^{m} \beta_j \boldsymbol{x}_j + \beta_{\text{new}} \boldsymbol{x}_{\text{new}} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{x}_0, \cdots, \boldsymbol{x}_m$ are linearly independent predictors and $\boldsymbol{x}_{\text{new}}$ is a new predictor. Let $X = (\boldsymbol{x}_0, \cdots, \boldsymbol{x}_k)$. Now, we define $\boldsymbol{r}_y = \boldsymbol{y} - X(X^TX)^{-1}X^T\boldsymbol{y}$ and $\boldsymbol{r}_{\text{new}} = \boldsymbol{x}_{\text{new}} - X(X^TX)^{-1}X^T\boldsymbol{x}_{\text{new}}$ to be the residuals of $\boldsymbol{y}$ and $\boldsymbol{x}_{\text{new}}$, respectively. The least squares estimation of β_{new} is given by

$$\hat{\beta}_{\text{new}} = \boldsymbol{r}_{\text{new}}^T \boldsymbol{r}_y / \boldsymbol{r}_{\text{new}}^T \boldsymbol{r}_{\text{new}} = \boldsymbol{x}_{\text{new}}^T \boldsymbol{r}_y / \boldsymbol{x}_{\text{new}}^T \boldsymbol{r}_{\text{new}} = \boldsymbol{x}_{\text{new}}^T [I - X(X^TX)^{-1}X^T]\boldsymbol{y}/\rho^2, \quad (3.4)$$

where

$$\rho^2 = \boldsymbol{x}_{\text{new}}^T \boldsymbol{r}_{\text{new}} = \boldsymbol{x}_{\text{new}}^T [I - X(X^TX)^{-1}X^T]\boldsymbol{x}_{\text{new}}. \quad (3.5)$$

Since $I - X(X^TX)^{-1}X^T$ is an idempotent symmetric matrix (a fact which will be used frequently throughout this chapter), one can derive that the variance of $\hat{\beta}_{\text{new}}$ is $\rho^{-2}\sigma^2$. Lin *et al.* (2011) suggested constructing the t-statistic $\hat{t} = \hat{\beta}_{\text{new}}\rho/\hat{\sigma} = \boldsymbol{x}_{\text{new}}^T \boldsymbol{r}_y/(\hat{\sigma}\rho)$, where $\hat{\sigma} = \|\boldsymbol{r}_y\|/\sqrt{(n-k-2)}$, the corresponding root-mean-square error (RMSE) of the residual $\boldsymbol{r}_y$. If $\Phi(|\hat{t}|) > 1 - \alpha/2$ for significance level α, then the new predictor $\boldsymbol{x}_{\text{new}}$ is added to the model. This is the key to the algorithm given in Lin *et al.* (2011).

We remark that the VIF regression algorithm cannot be directly applied to our variable selection problem, because any two successive vectors $\boldsymbol{\ell}_{q_s, n-q_s}$ and $\boldsymbol{\ell}_{q_{s+1}, n-q_{s+1}}$

differ only by $o\left(n^{2/3}\right)$ number of elements under Assumption A1 below, and hence are asymptotically correlated. However, to overcome these obstacles, we can modify the stagewise regression of the VIF regression algorithm as follows.

Suppose the predictors $\boldsymbol{x}_{1,i}, \cdots, \boldsymbol{x}_{m,i}$ have been selected based on the first il rows of $\boldsymbol{y}_n$. Here $\boldsymbol{x}_{r,i} = \boldsymbol{\ell}_{s_r l,(i-s_r)l}$ and $s_1 < \ldots < s_m < i$. We now check whether $\boldsymbol{x}_{\text{new}}^{(i+1)} = \boldsymbol{\ell}_{il,l}$ should be included as a new predictor via the following model

$$\boldsymbol{y}^{(i+1)} = \sum_{j=0}^{m} \beta_{j,i+1} \boldsymbol{x}_{j,i+1} + \beta_{\text{new}}^{(i+1)} \boldsymbol{x}_{\text{new}}^{(i+1)} - \boldsymbol{\eta}^{(i+1)} + \boldsymbol{\varepsilon}^{(i+1)}, \tag{3.6}$$

where $\boldsymbol{y}^{(i+1)} = \boldsymbol{y}_{(i+1)l}$ contains the first $(i+1)l$ rows of $\boldsymbol{y}_n$ and $\boldsymbol{x}_{0,i+1} = \mathbf{1}_{(i+1)l}$. The error vector $\boldsymbol{\varepsilon}^{(i+1)}$ and correction vector $\boldsymbol{\eta}^{(i+1)}$ are the first $(i+1)l$ rows truncated from the original vectors $\boldsymbol{\varepsilon}_n$ and $\boldsymbol{\eta}_n$, respectively. Let $X^{(i+1)} = (\boldsymbol{x}_{0,i+1}, \cdots, \boldsymbol{x}_{m,i+1})$. $\beta_{\text{new}}^{(i+1)}$ is estimated by

$$\hat{\beta}_{\text{new}}^{(i+1)} = \rho_{i+1}^{-2} \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \left\{I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T\right\} \boldsymbol{y}^{(i+1)}, \tag{3.7}$$

where

$$\rho_{i+1}^2 = \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \left\{I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T\right\} \boldsymbol{x}_{\text{new}}^{(i+1)}. \tag{3.8}$$

Applying (3.6) and (3.8) to (3.7) gives

$$\hat{\beta}_{\text{new}}^{(i+1)} = \beta_{\text{new}}^{(i+1)} + \rho_{i+1}^{-2} \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \left\{I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T\right\} (\boldsymbol{\varepsilon}^{(i+1)} - \boldsymbol{\eta}^{(i+1)}) \tag{3.9}$$

Following Lin *et al.* (2011), let

$$\hat{t}_{i+1} = \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \boldsymbol{r}^{(i+1)} / (\hat{\sigma}_{i+1} \rho_{i+1}), \tag{3.10}$$

where $\boldsymbol{r}^{(i+1)} = [I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T]\boldsymbol{y}^{(i+1)}$ is the residual and $\hat{\sigma}_{i+1} = \|\boldsymbol{r}^{(i+1)}\| / \sqrt{(i+1)l - m - 2}$ is the corresponding RMSE. If $\Phi(|\hat{t}_{i+1}|) > 1 - \alpha/2$, we put $\boldsymbol{x}_{m+1,i+1} = \boldsymbol{x}_{\text{new}}^{(i+1)}$ and $s_{m+1} = i + 1$, and repeat the above process with i and m replaced respectively by $i + 1$ and $m + 1$. Otherwise, we repeat the above process by replacing i by $i + 1$.

Before giving a theoretical justification of the modified VIF regression algorithm, we make the following two assumptions.

A1. Assume that $l \to \infty$ and $bl^{3/2} \ll n$ as $n \to \infty$.

A2. Assume that the errors $\{\varepsilon_i\}$ in model (3.1) are independent and identically distributed (iid) zero-mean random variables with variance σ^2. Furthermore, $E|\varepsilon_i|^{2+\nu} < \infty$ for some positive constant $\nu > 0$.

Remark 1. Assumption A1 allows b to go to infinity in the order of $n/M(n)$, where $M(n) \to \infty$ as $n \to \infty$. Assumption A2 is a very basic assumption that is necessary for establishing asymptotic normality of the estimators of βs.

Remark 2. The choice of l should follow the rule that there is no more than one change point in one partition. Under this condition, we can expect a more accurate estimate of a change point with larger l.

The following theorem shows that under the assumptions A1-A2, the modified stagewise regression is warranted. Its proof is given in the appendix.

Theorem 3.2.1 *If the assumptions A1-A2 are satisfied, then as* $n \to \infty$,

$$\rho_{i+1}^{-1}(\boldsymbol{x}_{new}^{(i+1)})^T \left\{ I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T \right\} \boldsymbol{\varepsilon}^{(i+1)} \xrightarrow{d} N(0, \sigma^2) \quad (3.11)$$

$$[(X^{(i+1)})^T X^{(i+1)}]^{-1} = O(1/n), \quad \rho_{i+1}^2/l \to 1,$$

where $l = n/(a+1)$ *is the length of each artificial segment. Note that* l *is large but* $l^{3/2}/n$ *is small by Assumption A1. Furthermore, the following statements hold true:*

(a) If the null hypothesis is accepted, i.e., $\beta_{new}^{(i+1)} = 0$, *then the scaled estimate* $\rho_{i+1}\hat{\beta}_{new}^{(i+1)}$ *converges to* $N(0, \sigma^2)$ *in distribution as* $n \to \infty$.

(b) If the alternative hypothesis is accepted, i.e., $\beta_{new}^{(i+1)} \neq 0$, *then*

(i) $\hat{\beta}_{new}^{(i+1)} = \beta_{new}^{(i+1)}[1 - \rho_{i+1}^{-2}(k_m - il)] + o_p(1)$, *where* k_m *denotes the change point in the artificial segment* $[1 + il, (i+1)l]$.

(ii) Moreover, if the change point k_m *lies in the artificial segment* $[1+(i-1)l, il]$ *(i.e., the change point was not detected during the previous search), then* $\hat{\beta}_{new}^{(i)} = \beta_{new}^{(i)} + o_p(1)$.

Proof of Theorem 3.2.1 Since ε_i, $i = 1, 2, \ldots$, are iid zero-mean variables with variance σ^2, it follows from the definition of ρ_{i+1} in (3.8) and the idempotence of $I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T$ that the variance of

$$\rho_{i+1}^{-1}(\boldsymbol{x}_{\text{new}}^{(i+1)})^T \{ I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T \} \boldsymbol{\varepsilon}^{(i+1)}$$

is still σ^2. By the central limit theorem, we obtain that

$$\rho_{i+1}^{-1}(\boldsymbol{x}_{\text{new}}^{(i+1)})^T \left\{ I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T \right\} \boldsymbol{\varepsilon}^{(i+1)} \xrightarrow{d} N(0, \sigma^2).$$

Note that $(X^{(i+1)})^T X^{(i+1)}$ can be expressed as $(U^{(i+1)})^T \Lambda^{(i+1)} U^{(i+1)}$, where $U^{(i+1)}$ is the lower triangular matrix of order $k+1$ whose nonzero entries are all 1's, and $\Lambda^{(i+1)}$ is a diagonal matrix with diagonal entries being $k_1 - k_0,\ k_2 - k_1,\ \ldots,\ k_m - k_{m-1},\ 1 + (i+1)l - k_m$. Since the change points are well-separated, i.e., $k_r - k_{r-1} = O(n)$, $(\Lambda^{(i+1)})^{-1}$ is of order $O(1/n)$, we have that $[(X^{(i+1)})^T X^{(i+1)}]^{-1}$ is also of order $O(1/n)$.

Next, we prove that ρ_{i+1} defined in (3.8) is asymptotically equal to $\sqrt{l}$. Note that $\boldsymbol{x}_{\text{new}}^{(i+1)} = \boldsymbol{\ell}_{il,l}$ is the vector with only the last l elements being ones, and all other elements are zeros. It can be seen that $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T \boldsymbol{x}_{\text{new}}^{(i+1)} = l$ and $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T X^{(n+1)} = O(l)$. Therefore, as $n \to \infty$, it is readily seen from $[(X^{(i+1)})^T X^{(i+1)}]^{-1} = O(1/n)$ that

$$\begin{aligned}\rho_{i+1}^2 &= (\boldsymbol{x}_{\text{new}}^{(i+1)})^T \boldsymbol{x}_{\text{new}}^{(i+1)} - (\boldsymbol{x}_{\text{new}}^{(i+1)})^T \left\{ I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T \right\} \boldsymbol{x}_{\text{new}}^{(i+1)} \\ &= l - O(l^2/n) \sim l.\end{aligned}$$

Under the null hypothesis, there exists no change point in the interval $[1+il, (i+1)l]$. It can be shown that the last l elements of the correction vector $\boldsymbol{\eta}^{(i+1)}$ are zeros, which implies that $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T \boldsymbol{\eta}^{(i+1)} = 0$. Since $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T X^{(i+1)} = O(l)$, $(X^{(i+1)})^T \boldsymbol{\eta}^{(i+1)} = o_p(bl)$, $[(X^{(i+1)})^T X^{(i+1)}]^{-1} = O(1/n)$ and $\rho_{i+1}/\sqrt{l} \to 1$, by Assumption A1, it follows

that

$$\rho_{i+1}^{-1}(\boldsymbol{x}_{\text{new}}^{(i+1)})^T \left\{ I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T \right\} \boldsymbol{\eta}^{(i+1)} = o(1).$$

In view of the fact that $\beta_{\text{new}}^{(i+1)} = 0$, i.e., there is no change point in $[1+il, (i+1)l]$, and $\rho_{i+1} \to \infty$, by (3.7) and (3.9), we obtain that

$$\rho_{i+1}\hat{\beta}_{\text{new}}^{(i+1)} \xrightarrow{d} N(0, \sigma^2).$$

This proves Theorem 1(a).

Under the alternative hypothesis, there exists a change point, say k_m, in the segment $[1+il, (i+1)l]$. Moreover, $k_m - il$ many of the last l elements of the correction vector $\boldsymbol{\eta}^{(i+1)}$ are equal to $\beta_{\text{new}}^{(i+1)}$, and $\beta_{\text{new}}^{(i+1)} \neq 0$, which implies $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T \boldsymbol{\eta}^{(i+1)} = \beta_{\text{new}}^{(i+1)}(k_m - il)$.

Moreover, we have

$$\rho_{i+1}^{-2}(\boldsymbol{x}_{\text{new}}^{(i+1)})^T X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T \boldsymbol{\eta}^{(i+1)} = o_p(1)$$

from the proof of Theorem 1 (a). In view of (3.11), we obtain that

$$\rho_{i+1}^{-2}(\boldsymbol{x}_{\text{new}}^{(i+1)})^T \left\{ I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T \right\} \boldsymbol{\varepsilon}^{(i+1)} = o_p(1).$$

Applying these results to (3.7) yields

$$\hat{\beta}_{\text{new}}^{(i+1)} = \beta_{\text{new}}^{(i+1)}[1 - \rho_{i+1}^{-2}(k_m - il)] + o_p(1).$$

Furthermore, if the change point k_m is located in the artificial interval $[1+(i-1)l, il]$ (i.e., the change point was previously undetected), then the correction vector $\boldsymbol{\eta}^{(i+1)}$ has zero components in the last l rows, which implies that $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T\boldsymbol{\eta}^{(i+1)} = 0$. A similar argument as above yields that $\hat{\beta}_{\text{new}}^{(i+1)} = \beta_{\text{new}}^{(i+1)} + o_p(1)$. This ends the proof of Theorem 1 (b).

3.2.2 A CUSUM Test and its Justification

By Theorem 1 (b)(i), we may conclude that a change point exists in the artificial time segment $[il+1, (i+1)l]$ if $\hat{\beta}_{\text{new}}^{(i+1)} \neq 0$. The precise location of the change point, however, is unknown because the formula (3.7) does not fully reflect the information contained in the correction vector η_n. To locate a change point in the artificial segment $[il+1, (i+1)l]$, one may conduct a test for a single change point over this segment, which, jointly with Theorem 1 (b)(ii) suggests that the test only needs to be carried out over the segment $[1+(i-1)l, il+\lfloor l/2 \rfloor]$.

Consider a univariate sequence $\{Z_i\}$ for $i = 1, \cdots, n$ with variance σ^2. We intend to test the null hypothesis

$$H_0 : E(Z_1) = \cdots = E(Z_n)$$

versus the alternative hypothesis

$$H_a : E(Z_1) = \cdots = E(Z_{k^*}) \neq E(Z_{k^*+1}) = \cdots = E(Z_n)$$

for some $k^* \in (1, n)$. The change point k^* is unknown, and both k^*/n and $1 - k^*/n$ are assumed to be bounded away from zero as $n \to \infty$. Many single change point detection methods in the literature can be used to solve this problem. Here, we apply the following CUSUM

$$U_k = C_k / w_k \tag{3.12}$$

to perform the test, where

$$C_k = \left(\frac{n}{k(n-k)} \right)^{1/2} \left(\sum_{i=1}^{k} Z_i - \frac{k}{n} \sum_{i=1}^{n} Z_i \right), \tag{3.13}$$

and

$$w_k = \sqrt{\frac{1}{n} \sum_{i=1}^{k} \left(Z_i - \frac{1}{k} \sum_{j=1}^{k} Z_j \right)^2 + \frac{1}{n} \sum_{i=k+1}^{n} \left(Z_i - \frac{1}{n-k} \sum_{j=k+1}^{n} Z_j \right)^2}. \tag{3.14}$$

If $B(\log n) \max_{1 \le k < n} |U_k| \le -\log(-\frac{1}{2}\log(1-\alpha)) + D(\log n)$, where $B(x) = (2 \log x)^{1/2}$ and $D(x) = 2 \log x + (1/2) \log \log x - (1/2) \log \pi$, then there is no change point, otherwise the change point exists and is estimated by $\hat{k} = \arg\max |U_k|$. It is noted that the above CUSUM is also related to the quasi-likelihood ratio test statistic. See Csörgő, M. and Horváth, L. (1997) (Equation 1.4.25) for details.

3.2.3 The Algorithm

The proposed method is implemented by the algorithm VIFCP below. Here, we provide the pseudocode for the VIFCP algorithm:

1. INPUT $\boldsymbol{y}_n$ and l.

2. INITIALIZATION, $a = n/l - 1$, $w = 0.05$, $dw = 0.05$, flag$= 0$, $\widehat{\mathcal{K}} = \emptyset$, $i = 1$, $j = 1$.

3. LOOP{

4. SET $\alpha = w/(1 + i - \text{flag})$.

5. OBTAIN statistic $\hat{t}_{i+1}$ by (3.10).

6. IF $2\Phi(|\hat{t}_{i+1}|) > 2 - \alpha$

7. Test for a change point k^* in $[(i-1)l, il + \lfloor l/2 \rfloor]$ using the CUSUM.

8. IF the test is significant, obtain $\hat{k}_j$.

9. $\widehat{\mathcal{K}} \leftarrow \widehat{\mathcal{K}} \cup \{\hat{k}_j\}$, flag$\leftarrow i$, $w \leftarrow w + dw$, $j = j + 1$.

10. ELSE $w \leftarrow w - \alpha/(1 - \alpha)$.

11. END IF

12. ELSE $w \leftarrow w - \alpha/(1 - \alpha)$.

13. END IF

14. UPDATE $i \leftarrow i + 1$.

15. }UNTIL $i \geq a+1$ or $w \leq 0$.

16. RETURN $\widehat{\mathcal{K}}$.

The values w, dw and α represent the wealth, payout and significance level, respectively. The details are given in Lin *et al.* (2011). From the$3rd$ line to the $15th$ line, we use a loop to find all change points. In the 5th line, we calculate the statistic $\hat{t}_{i+1}$ using the formula (3.10); this is the first key part of our algorithm. If the test is significant, then there may exist a change point in the artificial segment $[il+1, (i+1)l]$. The $7th$ line is the second key part of our algorithm, where we apply the algorithm CUSUM defined in (3.12) to locate the change point k^* in the interval $[(i-1)l, il + \lfloor l/2 \rfloor]$. Here we set the significance level of CUSUM to be 0.05. After the loop, we obtain $\widehat{\mathcal{K}}$, the estimates of multiple change points. We remark that this algorithm has been implemented in the R package **VIFCP** (Shi *et al.*, 2015).

We use the example in Section 2 to provide a more thorough explanation of our algorithm. The true change points are located at 4 and 7. As the sample size is 10 and $l = 2$, firstly we set $i = 1$ and will find there is no change point in the interval [0, 3]; after setting $i = 2$, we will find a change point in the interval [2, 5] at 4. If we set $i = 3$, we will not detect any changes in the interval [4, 7]. The change point at 7 will be detected when we set $i = 4$ in the interval [6, 9]. No change point will be detected in [8, 10] upon setting $i = 5$.

Now, we study the computational complexity of the algorithm VIFCP. Under Assumption A1, the computation time for the variable selection is of order $O(n^2/l)$, and the computation time for performing all the single change point tests is of order $O(bl)$. Hence the complexity of the algorithm VIFCP is $O(n^2/l+bl)$. It is noted that for finite b, the complexity of the algorithm VIFCP can be as low as $O(n^{4/3}M(n))$ ($M(n)$ is defined in Remark 1), while for $b = o(n)$, the complexity is $o(n^2)$.

3.3 Simulation Studies

In this section, we present three simulation studies. A Dell server (two E5520 Xeon Processors, two 2.26GHz 8M Caches, 16GB Memory) is used to perform the simulation studies. We will compare the performance of the algorithm VIFCP with CBS and PELT in terms of the accuracy of successfully detecting each true change point, the accuracy of successfully detecting all true change points under the condition that the number of true change points is correctly estimated, and efficiency as determined by the elapsed running time in seconds (ERT).

In the simulation studies, we consider the following three model settings:

$S1$: $y_i = \sum_{r=0}^{5} \mu_r I_{\{k_r,\ldots,k_{r+1}-1\}}(i) + \varepsilon_i, \quad i = 1, \cdots, 2000$, where

(1) $\{k_0, k_1, k_2, k_3, k_4, k_5, k_6 - 1\} = \{1, 324, 620, 1102, 1386, 1610, 2000\}$,

(2) $(\mu_0, \mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (0, 0.3, 0.7, 0.2, -0.2, 0.3)$,

(3) ε_i, $1 \leq i \leq n$, are iid $\sim N(0, \sigma^2)$,

(4) $\sigma = 0.2, 0.3$ and 0.4.

Simulated data for $S1$ with different value of σ, are plotted in Figure 3.3.

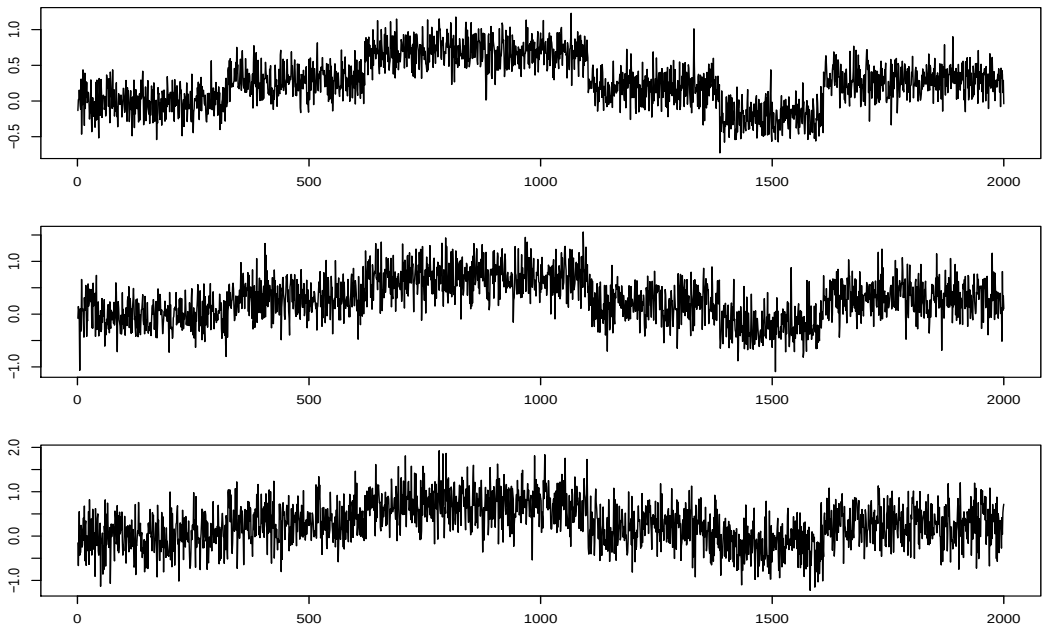

Figure 3.3: Simulated data for $S1$ with $\sigma = 0.2, 0.3$ and 0.4 (from the top to bottom panels).

$S2$: This setting is the same as the setting $S1$ with only the following exception: in each simulation, we randomly select 5 locations between 1 and n, and then add 5 to each value at these locations. The values at these 5 locations are considered as outliers.

$S3$: This setting is the same as the setting $S1$ with only the following exception: in each simulation, we randomly select 10 locations between 1 and n, and then add 5 to each value at these locations. The values at these 10 locations are considered as outliers.

For each of the model settings $S1$, $S2$, and $S3$, we first generate a data sequence, and then apply all three methods PELT, CBS, and VIFCP, to detect multiple change points in the dataset. We define $A_{k_i}^{(m)}$, where $m = 1, 2, \cdots, K$ and $J(A)$ in the same way as that in section 2.4.1 with the exception that $K = 5$. For each scenario, the number of simulation is set to be $M = 1000$.

The simulation results are reported respectively in Tables 3.1-3.3. Similar as the denotation in Chapter 2, "cpnumber.R" stands for the number of simulations in which the true number of change points is correctly estimated. "ALLCP" denotes the percentage of simulations in which exactly one change point is estimated, and the estimated change point is close to the corresponding exact change point. ERT.S means the total running time in seconds.

We observe from Tables 3.1 - 3.3 that VIFCP, PELT and CBS have similar performances in accuracy for $S1$. For $S2$, that differs from $S1$ by having 5 outliers, VIFCP and CBS have better accuracy in multiple change point detection than PELT, and hence, are more stable. However, for $S3$, the performance of PELT decreases

Method	*PELT*			*CBS*			*VIFCP*					
l							*100*			*80*		
σ	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
$\sum_{m=1}^{1000} J(A_{k_1}^{(m)})^*$	957	831	674	957	830	676	947	811	583	947	795	508
$\sum_{m=1}^{1000} J(A_{k_2}^{(m)})$	997	954	838	946	841	733	994	936	819	986	923	702
$\sum_{m=1}^{1000} J(A_{k_3}^{(m)})$	999	971	921	999	969	904	999	968	907	998	968	863
$\sum_{m=1}^{1000} J(A_{k_4}^{(m)})$	985	931	834	988	921	812	980	924	797	979	930	797
$\sum_{m=1}^{1000} J(A_{k_5}^{(m)})$	1000	982	915	997	975	905	994	972	904	997	977	899
cpnumber.R	1000	998	994	872	854	814	980	980	863	940	948	630
ALLCP(%)	93.8	69.9	41.0	89.2	59.5	32.3	91.3	65.4	34.2	90.6	65.6	33.0
ERT.S	9.312	9.502	9.269	201.196	202.099	205.639	0.422	0.503	0.420	0.402	0.417	0.376

$*$ $A_{k_i}^{(m)}$ is defined in formula (2.34) and $J(A)$ is given in (2.35).

Table 3.1: Simulation results of PELT, CBS, and VIFCP based on 1000 simulations for three different noise levels ($\sigma = 0.2$, $\sigma = 0.3$, and $\sigma = 0.4$) of scenario 1.

sharply with the increase of the number of outliers. Actually, PELT is very sensitive to the sudden change in observations, and detects outliers as change points in both $S2$ and $S3$.

If we compare these three methods in terms of ERT.S, we find that VIFCP is much faster than CBS and PELT in all three simulation studies. To examine whether or not the results obtained by using VIFCP are sensitive to the choice of l, we have varied the value of l. It can be seen from the three tables that the results for $l = 80$ and $l = 100$ are similar.

Method	*PELT*			*CBS*			*VIFCP*					
l							*100*			*80*		
σ	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
$\sum_{m=1}^{1000} J(A_{k_1}^{(m)})^*$	534	462	401	612	494	381	781	643	441	797	646	407
$\sum_{m=1}^{1000} J(A_{k_2}^{(m)})$	766	730	651	894	776	697	910	833	742	889	775	600
$\sum_{m=1}^{1000} J(A_{k_3}^{(m)})$	866	836	787	963	921	874	956	905	844	925	838	743
$\sum_{m=1}^{1000} J(A_{k_4}^{(m)})$	729	713	624	943	872	780	885	794	669	884	795	683
$\sum_{m=1}^{1000} J(A_{k_5}^{(m)})$	863	818	788	961	926	879	954	902	819	933	884	784
cpnumber.R	0	0	0	602	539	486	817	738	551	695	609	373
ALLCP(%)	0	0	0	69.1	45.3	27.8	64.6	43.2	28.3	67.9	45.3	22.5
ERT.S	7.005	6.940	6.982	117.529	118.142	128.291	0.411	0.423	0.451	0.425	0.374	0.358

$*$ $A_{k_i}^{(m)}$ is defined in formula (2.34) and $J(A)$ is given in (2.35).

Table 3.2: Simulation results of PELT, CBS, and VIFCP based on 1000 simulations for three different noise levels ($\sigma = 0.2$, $\sigma = 0.3$, and $\sigma = 0.4$) of scenario 2.

Method	*PELT*			*CBS*			*VIFCP*					
l							*100*			*80*		
σ	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
$\sum_{m=1}^{1000} J(A_{k_1}^{(m)})^*$	290	287	276	376	311	235	643	511	343	643	477	337
$\sum_{m=1}^{1000} J(A_{k_2}^{(m)})$	563	534	473	824	717	631	850	729	609	757	630	474
$\sum_{m=1}^{1000} J(A_{k_3}^{(m)})$	746	715	668	926	902	836	884	854	748	793	685	577
$\sum_{m=1}^{1000} J(A_{k_4}^{(m)})$	514	550	470	875	828	711	755	684	518	752	707	553
$\sum_{m=1}^{1000} J(A_{k_5}^{(m)})$	718	703	658	897	886	820	886	860	715	855	783	651
cpnumber.R	0	0	0	389	369	313	627	549	310	490	380	181
ALLCP	0	0	0	52.7	36.0	21.4	45.9	30.4	16.1	43.1	27.9	17.7
ERT.S	6.287	6.095	6.197	91.659	99.610	104.866	0.482	0.431	0.413	0.490	0.406	0.407

$*$ $A_{k_i}^{(m)}$ is defined in formula (2.34) and $J(A)$ is given in (2.35).

Table 3.3: Simulation results of PELT, CBS, and VIFCP based on 1000 simulations for three different noise levels ($\sigma = 0.2$, $\sigma = 0.3$, and $\sigma = 0.4$) of scenario 3.

3.4 Real Data Examples

In this section, we will analyze the following two real data examples.

3.4.1 Denoising a Barcode

The original barcode was given in the top panel of Figure 3.1. As explained in Section 1, all the values in the original image matrix range from 0 (black) to 1 (white). We now add Gaussian noises with mean 0, and standard deviation $\sigma = 0.1$ or 0.2, to each element of the original image matrix. Note that the resulting matrices may have elements smaller than 0 or larger than 1. To mimic an image matrix, we replace such elements by 0 or 1, i.e., we apply the transformation $xI_{[0,\,1]}(x) + I_{(1,\,\infty)}(x)$ to each element of the two noise-added matrices to make the noised grayscales range from 0 to 1. We name these two resulting matrices as Matrix 1 and Matrix 2, respectively.

One realization of the first row of each of Matrices 1-2 is plotted in Figure 3.4. The task is to reconstruct the original barcode, i.e., to find all the change points marked by vertical lines (obtained from the original image in the top panel of Figure 3.1). Here, the true number of change points is 48. We choose $l = 20$ for applying the VIFCP algorithm. For both datasets, VIFCP correctly detected all change points. In contrast, CBS and PELT failed to detect all change points. For the case when $\sigma = 0.1$, PELT detected 17 change points, while CBS correctly detected all of the

change points. When $\sigma = 0.2$, PELT still detected 17 change points, but CBS failed to detect two change points. Thus in terms of the multiple change point detection accuracy, even though CBS and PELT failed to compete with VIFCP, CBS outperformed PELT in this example.

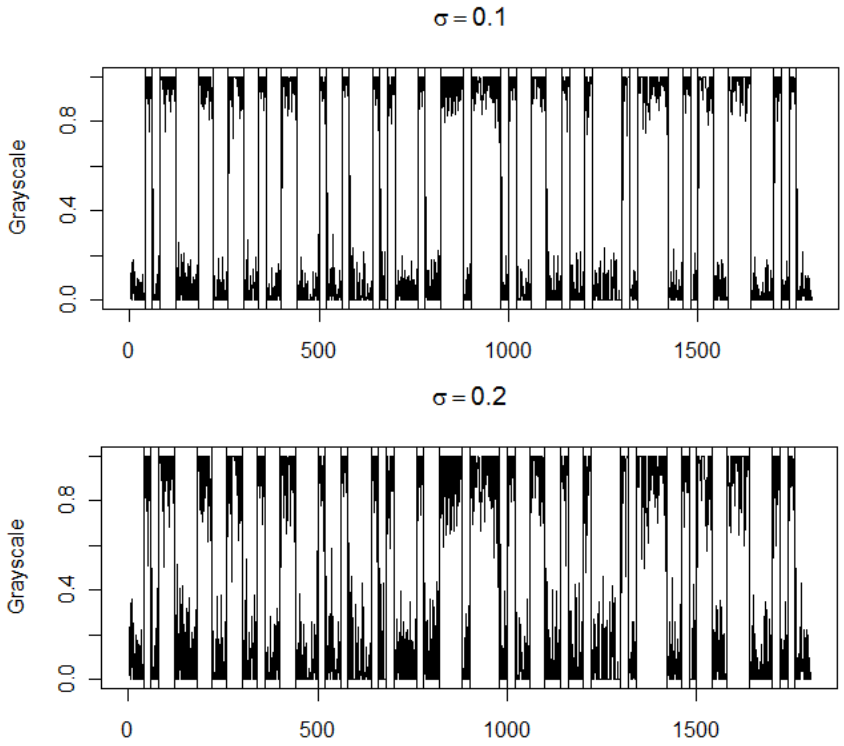

Figure 3.4: The data produced by a scanner through reading the first row of contaminated barcode image with different noise levels ($\sigma = 0.1$ and $\sigma = 0.2$). The true change points are marked by vertical lines.

3.4.2 Genetic Data

In this subsection, we consider a test using a genetic dataset involving 57 bladder tumor samples (Stransky. *et al.* 2006); see web page `http://microarrays.curie.fr/publications/oncologie_moleculaire/bladder_TCM/`. The problem is to find changes in the DNA copy number of an individual using array comparative genomic hybridization (CGH).

In order to perform multiple change point detection, we firstly deal with missing values in the dataset. Following Matteson and James (2013), we remove all series that had more than 7% of values missing, which left genome samples of 42 individuals for analysis. As in Matteson and James (2013), we also normalize the data so that the modal ratio is zero on a logarithmic scale. For each missing value, we find the 3 nearest neighbors using a Euclidean metric, and infer the missing value by averaging the values of its neighbors. As an illustration, we randomly choose two individuals' array CGH dataset for analysis. Here we choose the $11th$ and $13th$ individuals.

The choice of l is critical in the real data analysis. We will give a criterion for choosing l later in Section 5 (see the formula (3.15) for more details). We first limited the range of l to $\{5, 6, \ldots, 30\}$ before applying this criterion. The application of the criterion returned $l = 8$ and $l = 13$, respectively, for the $11th$ and $13th$ patient, as the optimal value of l.

For individual 11, VIFCP with $l = 8$ detects 22 change points, while PELT and CBS claim respectively 9 and 35 change points, which are displayed in Figure 3.5. From this figure, we observe that both VIFCP and CBS perform better than PELT. PELT fails to detect some change points. As a matter of fact, neither VIFCP or CBS is perfect in detecting the change points. There are three potential change points around 150, 500 and 600 but VIFCP fails to detect them. As for CBS, the minimum distance between two successive change points is 3, and in addition, the distance between adjacent change points in each of two pairs is 5. Thus CBS may overestimate the number of change points.

For individual 13, VIFCP with $l = 13$ detects 18 change points, while PELT and CBS report 6 and 44 change points, respectively. The result is shown in Figure 3.6. We conclude that CBS and VIFCP perform better than PELT because PELT fails to detect some potential change points. Moreover, VIFCP may fail to claim some change points while CBS obviously overestimates the number of change points.

3.5 Discussion

In this chapter, we propose a procedure, as well as its theoretical justification, for detecting multiple change points in the mean-shift model, where the number of change points is allowed to increase with the sample size. We first convert a

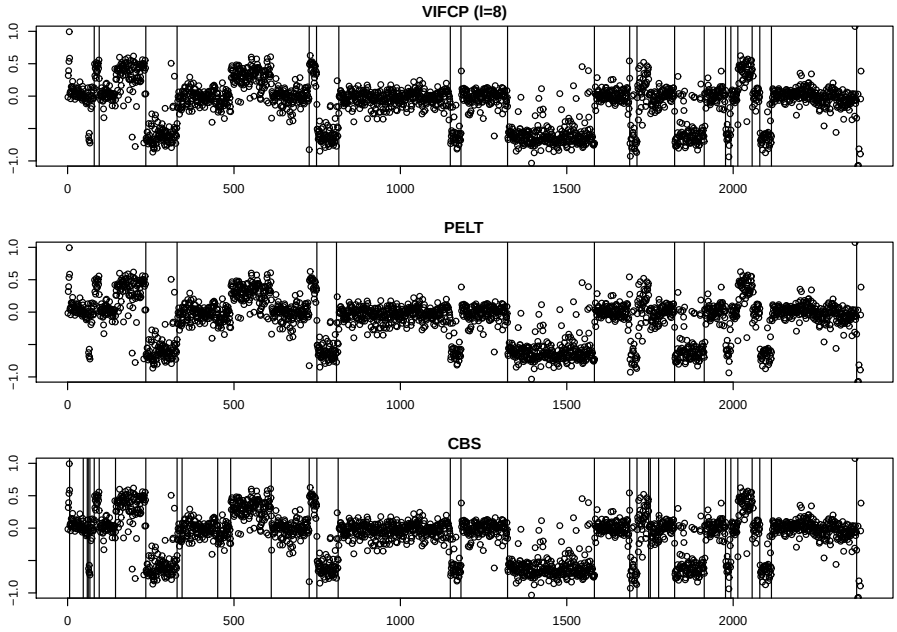

Figure 3.5: The normalized relative aCGH signal for the 11th individual with a bladder tumor. The change points detected by VIFCP with $l = 8$, PELT and CBS are indicated by the vertical lines.

change point detection problem into a variable selection problem by partitioning the data sequence. This allows us to apply a modified variance inflation factor regression algorithm to perform the variable selection sequentially in segment order. Once the segment containing a possible change point is flagged, a weighted CUSUM algorithm is applied to test if there is a change point in this segment. This procedure is implemented in the algorithm, named VIFCP. Simulation studies demonstrate

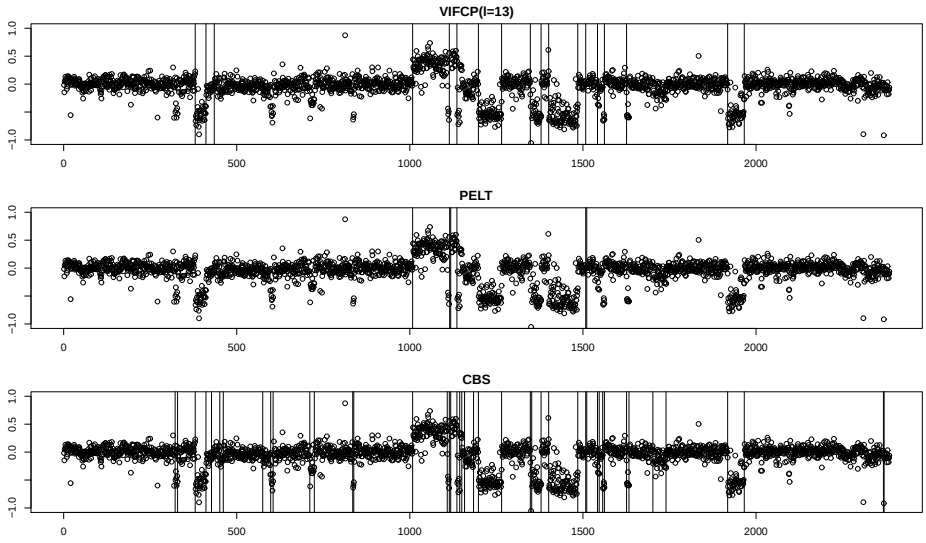

Figure 3.6: The normalized relative aCGH signal for the $13th$ individual with a bladder tumor. The change points detected by VIFCP with $l = 13$, PELT and CBS are indicated by the vertical lines.

that VIFCP, when compared with two popular algorithms, CBS and PELT, has a satisfactory performance in accuracy and computation time. It is also shown in the barcode example that VIFCP is better than CBS and PELT in terms of detection accuracy of multiple change points. In the second real-data analysis, VIFCP and CBS outperform PELT from the point-of-view of estimating change point locations.

In the simulation studies, segment length l is set to be 100, 80 for $n = 2000$. In the barcode example, l is set as 20, to account for the barcode design. The choice

of l is a very important issue. We may make the optimal choice of l by applying a Bayesian information criterion as follows:

$$l_{\text{opt}} = \arg\min_{l}\{\log(n)(DF_l + 1) + n\log(RSS_l/n)\}, \tag{3.15}$$

where DF is the number of estimated change points and $RSS_l = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

In this chapter, it can be seen that the proposed procedure for a mean-shift model can be extended to detect multiple change points in other types of regression models, including generalized linear models. The algorithm for implementing such a procedure is also feasible, requiring only a straightforward extension of VIFCP.

4 A Semiparametric Nonlinear Mixed-Effects Model with Covariate Measurement Errors and Change Points

4.1 Introduction

With the rapid development of longitudinal studies, various statistical models have been proposed to analyze different kinds of longitudinal data. The mixed-effects model is a commonly used approach, and it assumes that the response is linked to a function of covariates with fixed and random regression coefficients. When the mechanism of the data is known, a parametric nonlinear mixed-effect (NLME) model is often used to fit the longitudinal data. In literature, there are mainly three types of NLME models: parametric NLME models, nonparametric NLME models, and semiparametric NLME models. The parametric NLME models have been widely used in many longitidunal studies, such as HIV viral dynamics and pharmacokinetic

analysis. However, the performance of parametric NLME models is less satisfactory, especially when the underlying mechanism which generates the data is complicated in practice. In these cases, semiparametric or nonparametric NLME models may be more flexible in modelling the complex longitudinal data (Ke and Wang, 2001; Wu and Zhang, 2002).

In many longitudinal studies, great attention is paid to the inter-patient variation. This variation may be partially explained by time-varying covariates. For example, in HIV viral dynamic studies, patients' CD4 cell counts are repeatedly measured during the treatment, and they may partially explain the inter-patient variation. However, some covariates may be measured with substantial errors and may contain missing values. Ignoring measurement errors and missing data in covariates may lead to biased results (Higgins et al., 1997; Wu, 2002).

It is a common practice to analyze complex longitudinal data using NLME models in literature, however, these models may bacome a challenge if the response contains rebound part, which occurs often in longitudinal studies. Such rebound part in one patient's trajectory may be an important indicator to help quantify treatment effect and improve management of patient care. To overcome this challenge, change point models are introduced and should be simultaneously addressed for the NLME model. To the best of our knowledge, there are limited studies under the framework of

change point NLME model for longitudinal data. Huang (2013) studied change point modeling methods to investigate segmental mixed-effects models and illustrate the proposed methodology by longitudinal data from HIV viral dynamic study. Huang et al. (2015) implemented change point methods to analyze piecewise linear mixed-effects model for longitudinal studies under a Bayesian framework.

To consider the above problems in the longitudinal studies, we will address covariate measurement errors and change points in semiparametric NLME models based on the likelihood method in this chapter.

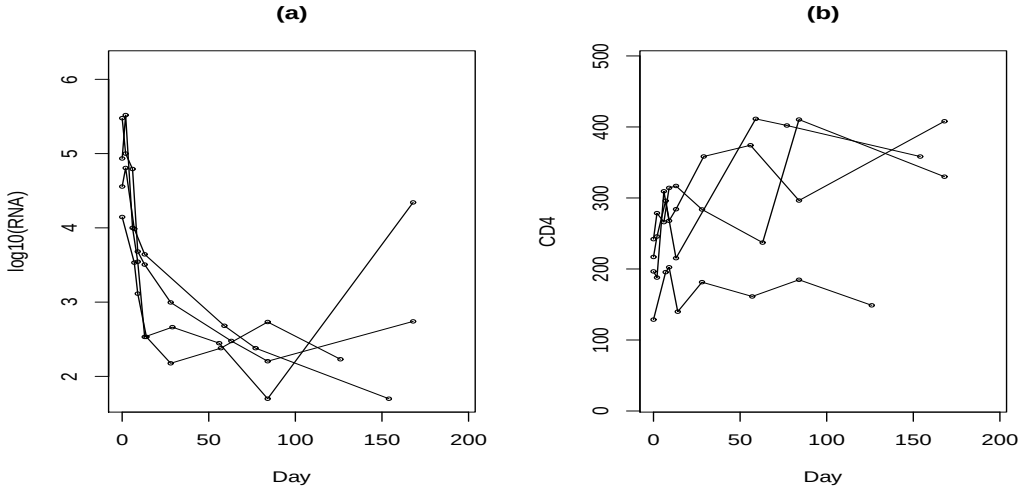

Figure 4.1: Viral loads and CD4 cell counts of four randomly selected HIV patients.

Our research in this chapter is motivated by HIV viral dynamic studies, and we are aiming to build suitable models to describe the viral load trajectories after the

start of an anti-HIV treatment. Figure 4.1(a) shows the viral load trajectories during an anti-HIV treatment for four randomly selected patients. We see that the viral loads of patients decrease sharply in the initial period after starting the treatment (a rapid initial decay, called first-phase viral decay). The pattern of the first-phase viral decay is clear and can be described parametrically. However, after the initial period of the treatment, the viral loads may continue to decrease with slower decay (called second-phase viral decay), and some of them may rebound at the late stage of the treatment. Thus, there may exist change points in the viral load trajectories. The viral load trajectories before the change points exhibit a clear pattern and may be modeled parametrically. On the other hand, the viral load trajectories after the change points can be quite complicated, so a parametric modeling may not be appropriate.

Based on some biological arguments, Wu and Ding (1999) derived the following two-exponential model with individual-specific parameters for short-term HIV viral dynamics (see also Wu, 2002)

$$\begin{aligned} y_{ij} &= \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + e_{ij}, \qquad (4.1) \\ log(P_{1i}) &= \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + b_{2i}, \\ log(P_{2i}) &= \beta_3 + b_{3i}, \qquad \lambda_{2ij} = \beta_4 + \beta_5 \mathrm{CD4}_{ij} + b_{4i}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, n_i, \end{aligned}$$

where y_{ij} is the $\log_{10}$-transformation of the viral load measurement for patient i at

time t_{ij}. P_{1i} and P_{2i} are baseline values and β_i's are fixed effects coefficients. λ_{1i} and λ_{2i} are the first (initial) and the second phases of viral decay rates respectively. We assume that $E(\lambda_{1ij}) > E(\lambda_{2ij})$ and $\beta_1 > \beta_3$. e_{ij}'s represent within-individual errors, and b_{ki}'s are random effects coefficients. We use a time-varying covariate $CD4$ cell count to partially explain the large inter-patient variation. Figure 4.1(b) shows the CD4 trajectories of four randomly selected patients. It is well known that CD4 cell count is often measured with substantial error, so it is reasonable to assume that the viral loads in model (4.1) are related to the true but unobserved CD4 values rather than the observed but mis-measured CD4. What is more, some CD4 values are not measured at the same time as the viral loads, which leads to missing data in CD4.

To avoid the truncation of the data and take change points and covariate measurement errors and missing data into account, we consider the following semiparametric NLME model for long-term HIV viral dynamics with covariate measurement errors and change points

$$\begin{aligned} y_{ij} &= \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + s_i(t_{ij})\max(t_{ij} - \tau_i,\ 0) + e_{ij}, \quad (4.2) \\ \log(P_{1i}) &= \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + b_{2i}, \\ \log(P_{2i}) &= \beta_3 + b_{3i}, \qquad \lambda_{2ij} = \beta_4 + \beta_5 \mathrm{CD4}^*_{ij}, \\ s_i(t_{ij}) &= \omega(t_{ij}) + h_i(t_{ij}), \end{aligned}$$

where $\omega(t_{ij})$ and $h_i(t_{ij})$ are nonparametric fixed and random smooth functions, re-

spectively. $CD4_{ij}^{*}$ is the true but unobserved CD4 values. τ_i denotes the change point for the ith patient, and the expression $\max(t_{ij} - \tau_i,\ 0)$ is used to maintain continuity of viral load trajectories at the change point. Compared with model (4.1), the model (4.2) still chooses the two-exponential models to fit the viral load trajectories before change points, and it has the advantage of taking the rebound parts of the trajectories into consideration.

Commonly used measurement errors models are reviewed in Carroll et al. (1995). For NLME models with covariate measurement errors, Higgins et al. (1997) proposed a two-step method and a bootstrap method, Wu (2002) considered consored response and covariate measurement errors based on a joint model, and Liu and Wu (2007) studied semiparametric NLME models with covariate measurement errors and missing responses.

Some authors have proposed random change point models for longitudinal data. For example, Carlin, Gelfand,and Smith (1992) proposed hierarchical Bayes models and applied to HIV/AIDS data. Morrell et al.(1995) used a nonlinear mixed-effects model to describe longitudinal changes in PSA in men before their prostate cancers were detected clinically. Hall et al. (2003) and Jacqmin-Gadda, Commenges, and Dartigues (2006) proposed change point models for cognitive function and dementia. However, there is little literature on simultaneously addressing covariate

measurement errors and change points for semiparametric NLME models. This is the objective of the current chapter.

In Section 4.2, we propose a general semiparametric NLME response model with covariate measurement errors and change point and then approximate it by a parametric NLME model, following Wu and Zhang (2002). We model the covariate process by using a mixed-effects model to incorporate measurement errors and missing data. Moreover, we consider survival models for the possibly censored times of change points. In Section 4.3, we simultaneously obtain maximum likelihood estimates (MLE) of all model parameters by using a Monte Carlo EM (MCEM) algorithm along with Gibbs sampler methods. We also employ the hierarchical likelihood (h-likelihood) approach, which is computationally much more efficient than MCEM approach, for an approximate MLE of those parameters. Both of the proposed approaches are illustrated in a real AIDS study in Section 4.4 and are evaluated via simulation in Section 4.5, respectively. We conclude this chapter in Section 4.6.

4.2 A General Semiparametric NLME Model with Covariate Measurement Errors and Change Points

4.2.1 A Semiparametric NLME Response Model with Change Points

Suppose there are n independent subjects in a study. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ be the observed response trajectory for individual i with the change point τ_i, where y_{ij} is the observed response value for individual i at time t_{ij}, $i = 1, \ldots, n, j = 1, \ldots, n_i$. During the study period, not all the τ_i's are observed since some individuals' trajectories may not experience their change points. Thus, the observed change point data for individual i are $T_i = \min(\tau_i, t_{in_i})$ and $c_i = I(\tau_i \leq t_{in_i})$, where $I(\cdot)$ is an indicator function. Let z_{ikl} be the observed value, possibly measured with error, and z^*_{ikl} be the corresponding true but unobservable value of covariate k for individual i at time u_{il}, $i = 1, \ldots, n, k = 1, \ldots, \nu, l = 1, \ldots, m_i$. Here, we allow the covariate measurement times u_{il} to differ from the response measurement times t_{ij}, i.e., we allow missing data in the covariates. We denote $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{im_i}^T)^T$, where $\boldsymbol{z}_{il} = (z_{i1l}, \ldots, z_{i\nu l})^T$, $l = 1, \ldots m_i$. The observed data are $\{(\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i),\ i = 1, \ldots, n\}$.

For the response process, we consider a general semiparametric NLME model which incorporates possibly censored change points and mis-measured time-varying

covariates

$$\begin{aligned} y_{ij} &= g_1(t_{ij}, \boldsymbol{\beta}^*_{ij}) + s_i(t_{ij})(t_{ij} - T_i)_+ + e_{ij}, \\ \boldsymbol{\beta}^*_{ij} &= \boldsymbol{d}(\boldsymbol{z}^*_{ij}, \boldsymbol{\beta}^*, \boldsymbol{b}^*_i), \\ s_i(t) &= g_2(\omega(t), h_i(t)), \quad i = 1, \ldots, n, j = 1, \ldots, n_i, \end{aligned} \tag{4.3}$$

where x_+ denotes $\max(x, 0)$ for a variable x, $g_1(\cdot), \boldsymbol{d}(\cdot)$ and $g_2(\cdot)$ are known parametric functions, $\omega(t)$ and $h_i(t)$ are unknown nonparametric smooth fixed-effects and random-effects functions, respectively. $\boldsymbol{\beta}^*_{ij}$ are individual-specific and time-dependent parameters, $\boldsymbol{\beta}^*$ are population parameters, e_{ij} are within-individual random errors, and $\boldsymbol{b}^*_i$ are random effects. Let $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{in_i})^T$ and we assume $\boldsymbol{e}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \delta^2 I_{n_i})$, where δ^2 is the unknown within-individual variance, and I_{n_i} is the $n_i \times n_i$ identity matrix. $\boldsymbol{b}^*_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, B^*)$ with B^* being an unstructured variance-covariance matrix, $h_i(t)$'s are $i.i.d.$ realizations of a zero-mean stochastic process, and $\boldsymbol{b}^*_i$ and $h_i(t)$ are independent of $\boldsymbol{e}_i$.

Note that in model (4.3), we assume that the individual-specific parameters $\boldsymbol{\beta}^*_{ij}$ depend on the true but unobservable covariates $\boldsymbol{z}^*_{ij}$ rather than the observed covariates $\boldsymbol{z}_{ij}$, which may be measured with substantial errors. In model (4.3), we incorporate change points in the response trajectories. We parametrically and nonparametrically model the response trajectories before and after the change points to account for the data mechanism before the change points and the trajectory com-

plexity after the change points, respectively.

For likelihood reference, we approximate the nonparametric functions $\omega(t)$ and $h_i(t)$ by linear combinations of basis functions $\Psi_p(t) = (\psi_0(t), \psi_1(t), \ldots, \psi_{p-1}(t))^T$ and $\Phi_q(t) = (\phi_0(t), \phi_1(t), \ldots, \phi_{q-1}(t))^T$ as follows (Rice and Wu 2001):

$$\begin{aligned} \omega(t) &\approx \omega_p(t) = \sum_{k=0}^{p-1} \mu_k \psi_k(t) = \Psi_p(t)^T \boldsymbol{\mu}_p, \\ h_i(t) &\approx h_{iq}(t) = \sum_{k=0}^{q-1} \xi_{ik} \phi_k(t) = \Phi_p(t)^T \boldsymbol{\xi}_{iq} \end{aligned} \tag{4.4}$$

where $\boldsymbol{\mu}_p$ and $\boldsymbol{\xi}_{iq}$ are unknown vectors of fixed and random coefficients respectively. We can regard $\boldsymbol{\xi}_{iq}$ as i.i.d. realizations of a zero-mean random vectors. We consider natural cubic spline bases with percentile-based knotes, and the number of knotes is determined by the AIC or BIC criteria. If we substitute $\omega(t)$ and $h_i(t)$ by their approximations $\omega_p(t)$ and $h_{iq}(t)$, then the semiparametric NLME model (4.3) can be approximated by the following parametric NLME model

$$\begin{aligned} y_{ij} &= g_1(t_{ij},\ \boldsymbol{d}(\boldsymbol{z}_{ij}^*, \boldsymbol{\beta}^*, \boldsymbol{b}_i^*)) + g_2(\Psi_p(t_{ij})^T \boldsymbol{\mu}_p, \Phi_p(t_{ij})^T \boldsymbol{\xi}_{iq})(t_{ij} - T_i)_+ + e_{ij} \\ &\equiv g(t_{ij}, \boldsymbol{z}_{ij}^*, \boldsymbol{\beta}, \boldsymbol{b}_i, T_i) + e_{ij} \end{aligned} \tag{4.5}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}^*, \boldsymbol{\mu}_p)$ are fixed effects, $\boldsymbol{b}_i = (\boldsymbol{b}_i^*, \boldsymbol{\xi}_{iq})$ are random effects, $\boldsymbol{b}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{B})$ with $\boldsymbol{B}$ being an unstructured variance-covariance matrix, and the function $g(\cdot)$ is known.

4.2.2 Measurement Errors and Missing Data in Covariates

In the NLME model (4.5), covariate value $\boldsymbol{z}_{ij}$ must be available at the response measurement times t_{ij}. However, due to different covariate measurement schedules or other problems, covariates may be missing at times t_{ij}. Because of the existance of measurement errors and missing data in the time-varying covariates, we need to model the covariate processes. We consider the following multivariate LME model (Shah, Laird, and Schoenfeld, 1997) to describe the covariate process

$$\boldsymbol{z}_{il} = U_{il}\boldsymbol{\alpha} + V_{il}\boldsymbol{a}_i + \boldsymbol{\epsilon}_{il} (\equiv \boldsymbol{z}^*_{il} + \boldsymbol{\epsilon}_{il}), \quad i = 1, \ldots, n, l = 1, \ldots, m_i, \tag{4.6}$$

where U_{il} and V_{il} are $\nu \times d$ and $\nu \times r$ design matrices, $\boldsymbol{\alpha}$ and $\boldsymbol{a}_i$ are unknown population (fixed-effects) and individual-specific (random-effects) parameter vectors, and $\boldsymbol{\epsilon}_{il}$ are the random measurement errors for individual i at time u_{il}. We assume that the true covariate values are $\boldsymbol{z}^*_{il} = U_{il}\boldsymbol{\alpha} + V_{il}\boldsymbol{a_i}$. Moreover, we assume $\boldsymbol{a}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, A)$, $\boldsymbol{\epsilon}_{il} \overset{i.i.d.}{\sim} N(\boldsymbol{0}, R)$, where A and R are unknown variance-covariance matrices. Let $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}^T_{i1}, \ldots, \boldsymbol{\epsilon}^T_{im_i})^T$, which is assumed to be indepedent with $\boldsymbol{a_i}$. We also assume that $\boldsymbol{a_i}$ and $\boldsymbol{\epsilon}_i$ are independent of $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ in the response model. Models such as (4.6) may be interpreted as a covariate measurement error model (Carroll et al., 1995).

4.2.3 Change Points Analysis

In this section, we will build a suitable model for the times of the change points in model (4.3). The time of the rebound is likely related to the longitudinal response and covariate process. We can specify the association by assuming the time of change point $\tau_i \sim f(t|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}, \eta)$ with unknown parameters $\boldsymbol{\gamma}$ and η, where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are random effects in the covariate and the response model, respectively. We may consider the following model

$$\log(\tau_i) = \gamma_0 + \boldsymbol{\gamma}_1^T \boldsymbol{a}_i + \boldsymbol{\gamma}_2^T \boldsymbol{b}_i + \zeta_i, \quad i = 1, \cdots, n, \tag{4.7}$$

where $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)^T$ are regression coefficients, and the random errors ζ_i's are *i.i.d.* and follow a parametric distribution with mean 0 and other parameter η, such as $\zeta_i \sim N(0, \eta^2)$. We assume that ζ_i's are independent of $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$. Model (4.7) may be a good choice when the change points are thought to depend on individual-specific longitudinal trajectories, such as initial slopes and intercepts, or summaries of the longitudinal trajectories, and it is closely related to so-called shared parameter models (Wu and Carroll, 1988; DeGruttola and Tu, 1994).

4.3 Joint Likelihood Inference

We consider likelihood inference for semiparametric NLME models with covariate measurement errors and change points based on the approximate parametric NLME response model (4.3), the covariate measurement error model (4.6), and the change point model (4.7).

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, R, A, B, \eta)$ be the collection of all unknown model parameters, and let $f(\cdot)$ denote a generic density function. The approximate log-likelihood function of $\boldsymbol{\theta}$ for the observed data $\{(\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i), i = 1, \cdots, n\}$ can be written as

$$\begin{aligned} l_o(\boldsymbol{\theta}) &= \sum_{i=1}^{n} l^{(i)}(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n} \log \int \int \Big[f_Y(\boldsymbol{y}_i | \boldsymbol{a}_i, \boldsymbol{b}_i, T_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) f_Z(\boldsymbol{z}_i | \boldsymbol{a}_i; \boldsymbol{\alpha}, R) \\ &\quad \times [f(T_i | \boldsymbol{a}_i, \boldsymbol{b}_i; \gamma, \eta^2)]^{c_i} [1 - F(T_i | \boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}, \boldsymbol{\eta})]^{1-c_i} \Big] f(\boldsymbol{a}_i; A) f(\boldsymbol{b}_i; B) d\boldsymbol{a}_i d\boldsymbol{b}_i \end{aligned} \tag{4.8}$$

where $F(T_i | \boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{\gamma}, \boldsymbol{\eta})$ is the cumulative distribution function. The approximate maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ can be obtained by directly maximizing the observed data log-likelihood $l_o(\boldsymbol{\theta})$. However, this is computationally infeasible since the function $l_o(\boldsymbol{\theta})$ may not have a closed-form expression with high-dimensional and intractable integrals. Therefore, we consider the following two alternative approaches to obtain the approximate MLE of $\boldsymbol{\theta}$.

4.3.1 A Monte Carlo Expectation Maximization Approach

We consider a Monte-Carlo Expectation Maximization (MCEM) algorithm to find the approximate MLE of $\boldsymbol{\theta}$. By treating the unobservable random effects $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ as additional "missing" data, we have the "complete data" $\{(\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i, \boldsymbol{a}_i, \boldsymbol{b}_i), i = 1, \cdots, n\}$. Thus the "complete data" log-likelihood function of $\boldsymbol{\theta}$ for all individuals can be expressed as

$$\begin{aligned} l_{com}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} l_{com}^{(i)}(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n} \big\{ \log f_Y(\boldsymbol{y}_i|\boldsymbol{a}_i, \boldsymbol{b}_i, T_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) + \log f_Z(\boldsymbol{z}_i|\boldsymbol{a}_i; \boldsymbol{\alpha}, R) \\ &\quad + c_i \log f(T_i|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}, \eta) + (1 - c_i) \log[1 - F(T_i|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}, \boldsymbol{\eta})] \qquad (4.9) \\ &\quad + \log f(\boldsymbol{a}_i; A) + \log f(\boldsymbol{b}_i; B) \big\}. \end{aligned}$$

Let $\boldsymbol{\theta}^{(t)}$ be the parameter estimate of $\boldsymbol{\theta}$ from the t-th EM iteration. The E-step for individual i at the $(t+1)$th EM iteration can be written as

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E(l_{com}^{(i)}(\boldsymbol{\theta})|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i; \boldsymbol{\theta}^{(t)}) = \int\int \Big[\log f_Y(\boldsymbol{y}_i|\boldsymbol{a}_i, \boldsymbol{b}_i, T_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) \\ &\quad + \log f_Z(\boldsymbol{z}_i|\boldsymbol{a}_i; \boldsymbol{\alpha}, R) + c_i \log f(T_i|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}, \eta) \\ &\quad + (1 - c_i) \log[1 - F(T_i|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}, \boldsymbol{\eta})] + \log f(\boldsymbol{a}_i; A) + \log f(\boldsymbol{b}_i; B) \Big] \\ &\quad \times f(\boldsymbol{a}_i, \boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i; \boldsymbol{\theta}^{(t)}) d\boldsymbol{a}_i d\boldsymbol{b}_i. \qquad (4.10) \end{aligned}$$

Generally there is no closed-form for the above integration, and numerically evaluation of the integral is usually infeasible. However, note that $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is an expectation with respect to the conditional distribution $f(\boldsymbol{a}_i, \boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i; \boldsymbol{\theta}^{(t)})$, so it can

be evaluated using MCEM algorithm (Wei and Tanner, 1990; Ibrahim et al. 2001). Specifically, we can use the Gibbs sampler (Gelfand and Smith, 1990) to generate samples of $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ from $f(\boldsymbol{a}_i, \boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i; \boldsymbol{\theta}^{(t)})$ by iteratively sampling from the full conditionals $f(\boldsymbol{a}_i|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i, \boldsymbol{b}_i; \boldsymbol{\theta}^{(t)})$ and $f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i, \boldsymbol{a}_i; \boldsymbol{\theta}^{(t)})$ as follows:

$$\begin{aligned} f(\boldsymbol{a}_i|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i, \boldsymbol{b}_i; \boldsymbol{\theta}^{(t)}) &\propto f(\boldsymbol{z}_i|\boldsymbol{a}_i; \boldsymbol{\alpha}^{(t)}) f(\boldsymbol{a}_i; A^{(t)}) f(\boldsymbol{y}_i|\boldsymbol{a}_i, \boldsymbol{b}_i, T_i; \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \delta^{(t)}) \\ &\quad \times [f(T_i|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}^{(t)}, \eta^{(t)})]^{c_i} [1 - F(T_i|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}^{(t)}, \eta^{(t)})]^{1-c_i}, \\ f(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i, \boldsymbol{a}_i; \boldsymbol{\theta}^{(t)}) &\propto f(\boldsymbol{b}_i; B^{(t)}) f(\boldsymbol{y}_i|\boldsymbol{a}_i, \boldsymbol{b}_i, T_i; \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \delta^{(t)}) [f(T_i|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}^{(t)}, \eta^{(t)})]^{c_i} \\ &\quad \times [1 - F(T_i|\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\gamma}^{(t)}, \eta^{(t)})]^{1-c_i}. \end{aligned}$$

After generating large random samples of $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ from $f(\boldsymbol{a}_i, \boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i; \boldsymbol{\theta}^{(t)})$, we can replace the "missing data" by the simulated values and then approximate the expectation $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ by its empirical mean. The M-step, which maximize $\sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, is like a complete-data maximization, so complete-data optimization procedures may be used to update the parameter estimates.

When the MCEM algorithm is convergent, we obtain the approximate MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, whose variance-covariance matrix can be calculated by the following formula (McLachlan and Krishnan, 1997):

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \left[\sum_{i=1}^{n} E(S_c^{(i)}|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i; \hat{\boldsymbol{\theta}}) E(S_c^{(i)}|\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, c_i; \hat{\boldsymbol{\theta}})^T \right]^{-1}, \tag{4.11}$$

where $S_c^{(i)} = \partial l_c^{(i)}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and the expectations can be approximated by Monte Carlo

methods.

4.3.2 An Approximation Approach Based on Hierarchical Likelihood

As the MCEM method involves sampling the random effects ($\boldsymbol{a}_i$ and $\boldsymbol{b}_i$), which may have a high dimension, it may be computationally intensive and sometimes has the problem of nonconvergence, even with the availability of modern computers. To overcome these disadvantages, in this section we consider an alternative approach called the hierarchical likelihood (h-likelihood) approach for approximate likelihood inference. The h-likelihood approach is computationally feasible and may be used to obtain good parameter starting values for the MCEM approach.

Let $\boldsymbol{\xi}$ denote the general "nuisance parameters" and $\boldsymbol{\theta}$ denote the parameters of interest. Lee and Nelder (1996) considered the following function

$$p_{\hat{\boldsymbol{\xi}}}(l(\boldsymbol{\theta},\boldsymbol{\xi})) = \left[l(\boldsymbol{\theta},\boldsymbol{\xi}) - \frac{1}{2}\log\left|\frac{1}{2\pi}D(l(\boldsymbol{\theta},\boldsymbol{\xi}),\boldsymbol{\xi})\right| \right]_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}}, \tag{4.12}$$

where $D(l(\boldsymbol{\theta},\boldsymbol{\xi}),\boldsymbol{\xi}) = -\partial^2 l(\boldsymbol{\theta},\boldsymbol{\xi})/\partial\boldsymbol{\xi}^2$, and $\hat{\boldsymbol{\xi}}$ is the solution to $\partial l(\boldsymbol{\theta},\boldsymbol{\xi})/\partial\boldsymbol{\xi} = \mathbf{0}$. Following Lee and Nelder (1996), the complete-data log-likelihood function $l_{com}(\boldsymbol{\theta})$ in (4.9) can be called the hierarchical log-likelihood function because it combines the two stages of mixed-effects models. Specifically, if we define $\boldsymbol{\omega}_i = (\boldsymbol{a}_i^T, \boldsymbol{b}_i^T)^T$ and $\boldsymbol{\omega} = (\boldsymbol{\omega}_1^T, \cdots, \boldsymbol{\omega}_n^T)^T$, the complete-data log-likelihood $l_{com}(\boldsymbol{\theta})$ of (4.9) can be denoted

as $l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega})$. Thus, the function $p_{\hat{\boldsymbol{\omega}}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}))$ can be written as

$$p_{\hat{\boldsymbol{\omega}}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega})) = \sum_{i=1}^{n} \left[l_{com}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\omega}) - \frac{1}{2} \log \left| \frac{1}{2\pi} D(l_{com}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\omega}_i) \right| \right]_{\boldsymbol{\omega}_i = \hat{\boldsymbol{\omega}}_i}. \quad (4.13)$$

where $\hat{\boldsymbol{\omega}} = (\hat{\boldsymbol{\omega}}_i^T, \cdots, \hat{\boldsymbol{\omega}}_n^T)^T$ are the solutions to the equations $\partial l_{com}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\omega}_i)/\partial \boldsymbol{\omega}_i = \mathbf{0}, i = 1, \cdots, n$. It can be shown that $p_{\hat{\boldsymbol{\omega}}}(l_h(\boldsymbol{\theta}))$ is the first-order Laplace approximation to the marginal log-likelihood $l_o(\boldsymbol{\theta})$ in (4.8) using the hierarchical log-likelihood function $l_{com}(\boldsymbol{\theta})$.

Recall that the first-order Laplace approximation to the intractable integral, which is in the form of $\int e^{k\rho(\boldsymbol{\nu})} d\nu$, can be written as

$$\int e^{k\rho(\boldsymbol{\nu})} d\boldsymbol{\nu} = \left(\frac{2\pi}{k} \right)^{d/2} \cdot \left| \frac{\partial^2 \rho(\hat{\boldsymbol{\nu}})}{\partial \boldsymbol{\nu}^2} \right|^{-\frac{1}{2}} \cdot e^{k\rho(\hat{\boldsymbol{\nu}})} + O(k^{-1}), \quad (4.14)$$

where ν is a d-dimension vector, $\hat{\boldsymbol{\nu}}$ maximizes $\rho(\boldsymbol{\nu})$, and $\frac{\partial^2 \rho(\hat{\boldsymbol{\nu}})}{\partial \boldsymbol{\nu}^2} = \partial^2 \rho(\boldsymbol{\nu})/\partial \boldsymbol{\nu}^2|_{\boldsymbol{\nu}=\hat{\boldsymbol{\nu}}}$.

Letting $N_i = n_i + m_i$ and $N = \min_i N_i$, we assume that $N_i = O(N)$ uniformly for $i = 1, \cdots, n$. Taking $k = N_i$, $k\rho(\boldsymbol{\nu}) = l_{com}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\omega}_i)$, $d = dim(\boldsymbol{\omega}_i)$, and $\boldsymbol{\nu} = \boldsymbol{\omega}_i$ in the Laplace approximation (4.14), we can approximate the ith individual's contribution

$l^{(i)}(\boldsymbol{\theta})$ to the overall observed-data log-likelihood function $l_o(\boldsymbol{\theta})$ as

$$\begin{aligned}
l^{(i)}(\boldsymbol{\theta}) &= \log \int e^{l^{(i)}_{com}(\boldsymbol{\theta},\boldsymbol{\omega}_i)} d\boldsymbol{\omega}_i = \log \int e^{N_i \rho(\boldsymbol{\omega}_i)} d\boldsymbol{\omega}_i \\
&= \log \left\{ \left(\frac{2\pi}{N_i}\right)^{d/2} \left| D(\rho(\boldsymbol{\omega}_i), \boldsymbol{\omega}_i)|_{\boldsymbol{\omega}_i=\hat{\boldsymbol{\omega}}_i} \right|^{-1/2} e^{N_i \rho(\hat{\boldsymbol{\omega}}_i)} + O_p(N_i^{-1}) \right\} \\
&= \log \left\{ \left(\frac{2\pi}{N_i}\right)^{d/2} \left| \frac{1}{N_i} D(l_h^{(i)}(\boldsymbol{\theta}, \boldsymbol{\omega}_i), \boldsymbol{\omega}_i)\big|_{\boldsymbol{\omega}_i=\hat{\boldsymbol{\omega}}_i} \right|^{-1/2} e^{l^{(i)}_{com}(\boldsymbol{\theta},\hat{\boldsymbol{\omega}}_i)} + O_p(N_i^{-1}) \right\} \\
&= \log \left\{ \left| \frac{1}{2\pi} D(l^{(i)}_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}_i), \boldsymbol{\omega}_i)|_{\boldsymbol{\omega}_i=\hat{\boldsymbol{\omega}}_i} \right|^{-1/2} e^{l^{(i)}_{com}(\boldsymbol{\theta},\boldsymbol{\omega}_i)} + O_p(N_i^{-1}) \right\} \\
&= \log \left[\exp\{p_{\hat{\boldsymbol{\omega}}_i}(l^{(i)}_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}_i))\} + O_p(N_i^{-1}) \right] \\
&= p_{\hat{\boldsymbol{\omega}}_i}(l^{(i)}_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}_i)) + O(N_i^{-1}). \qquad (4.15)
\end{aligned}$$

Hence, the observed data log-likelihood function $l_o(\boldsymbol{\theta})$ in (4.8) can be approximated as

$$\begin{aligned}
l_o(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_o^{(i)}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} [p_{\hat{\boldsymbol{\omega}}_i}(l^{(i)}_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}_i)) + O(N_i^{-1})] \\
&= p_{\hat{\boldsymbol{\omega}}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega})) + \sum_{i=1}^{n} O(N_i^{-1}) \\
&= p_{\hat{\boldsymbol{\omega}}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega})) + \sum_{i=1}^{n} O(N^{-1}) \\
&= p_{\hat{\boldsymbol{\omega}}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega})) + nO(N^{-1})
\end{aligned}$$

As $N = \min_i N_i$ grows faster than n, the function $p_{\hat{\boldsymbol{\omega}}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}))$ approaches the observed-data log-likelihood function $l_o(\boldsymbol{\theta})$. Thus, the estimate of $\boldsymbol{\theta}$, which maximizes $p_{\hat{\boldsymbol{\omega}}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}))$, also maximizes $l_o(\boldsymbol{\theta})$. Therefore we propose the following algorithm to obtain an approximate MLE of $\boldsymbol{\theta}$ denoted by $\hat{\boldsymbol{\theta}}_h$:

Step 1. Initialize the estimate $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}, \delta^{(0)}, A^{(0)}, B^{(0)}, R^{(0)}, \eta^{(0)})$ of $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, A, B, R, \eta)$ based on a naive approach.

Step 2. Given the current parameter estimates $\boldsymbol{\theta}^{(t)}$, update $\boldsymbol{\omega}_i^{(t+1)}$ by maximizing $l_{com}^{(i)}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\omega}_i)$ with respect to $\boldsymbol{\omega}_i, i = 1, \cdots, n$.

Step 3. Given the random effects estimates $\boldsymbol{\omega}^{(t+1)}$, update the parameter estimate $\boldsymbol{\theta}^{(t+1)}$ by maximizing $p_{\boldsymbol{\omega}^{(t+1)}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}^{(t+1)}))$ with respect to $\boldsymbol{\theta}$.

Step 4. Iterate between Step 2 and Step 3 until convergence.

We can use Fisher information to obtain the following approximate formula for the variance-covariance matrix of the approximate MLE $\hat{\boldsymbol{\theta}}_h$.

$$Cov(\hat{\boldsymbol{\theta}}_h) = \left[-\frac{\partial^2 p_{\hat{\boldsymbol{\omega}}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]^{-1}_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_h}.$$

4.3.3 Asymptotic Properities of the Approximate MLE $\hat{\boldsymbol{\theta}}_h$

Firstly we give the following lemma which will be useful in the proof of the theorem.

Lemma 4.3.1 *Let Y_n be a sequence of random variables satisfying $Y_n = c + O_p(a_n)$ where $a_n = o(1)$. If $f(x)$ is a function with r continuous derivatives at $x = c$, then*

$$f(Y_n) = f(c) + f^{(1)}(c)(Y_n - c) + \cdots + [1/(r-1)!] f^{(r-1)}(c)(Y_n - c)^{r-1} + O_p(a_n^r),$$

where $f^{(k)}(c)$ is the kth derivative of f evaluated at c. In particular, $f(Y_n) = f(c) +$

$O_p(a_n)$. This result holds when $O_p(\cdot)$ is replaced everywhere by $o_p(\cdot)$ or when Y_n and c are replaced by a vector/matrix random variable $\mathbf{Y}_n$ and vector/matrix constant $\mathbf{c}$.

One can refer to Vonesh and Chinchilli (1997) for the proof of Lemma (4.3.1).

Theorem 4.3.1 *Suppose $l_o(\boldsymbol{\theta})$ in expression (4.8) has second continuous derivatives, and $\partial l_o^{(i)}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ are i.i.d. with finite entries covariance, we have*

$$(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0) = O_p\left[\max\left\{n^{-\frac{1}{2}}, \left(\min_i N_i\right)^{-1}\right\}\right], \tag{4.16}$$

where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$.

Proof. Let $\hat{\boldsymbol{\omega}}_i$ maximize $l_{com}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\omega}_i)$ with respect to $\boldsymbol{\omega}_i$ for fixed $\boldsymbol{\theta}$. We denote $N_i = n_i + m_i$ and $N = \min_i N_i$, and assume that $N_i = O(N)$ uniformly for $i = 1, \cdots, n$. The ith individual's contribution $l^{(i)}(\boldsymbol{\theta})$ to the overall observed-data log-likelihood may be approximated as

$$l^{(i)}(\boldsymbol{\theta}) = p_{\hat{\omega}_i}(l_{com}^{(i)}(\boldsymbol{\theta}, \boldsymbol{\omega}_i)) + O(N_i^{-1}). \tag{4.17}$$

Hence, the observed-data log-likelihood $l_o(\boldsymbol{\theta})$ can be written as

$$l_o(\boldsymbol{\theta}) = l^*(\boldsymbol{\theta}) + O(nN^{-1}), \tag{4.18}$$

where $l^*(\boldsymbol{\theta}) = p_{\hat{\omega}}(l_{com}(\boldsymbol{\theta}, \boldsymbol{\omega}_i)) = \sum_{i=1}^n p_{\hat{\omega}_i}(l_{com}^{(i)}(\boldsymbol{\theta}))$. Let $\boldsymbol{u}^*(\boldsymbol{\theta}) = \partial l^*(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and let $\hat{\boldsymbol{\theta}}_h$ be the approximate maximum likelihood estimate satisfying $\boldsymbol{u}^*(\hat{\boldsymbol{\theta}}_h) = 0$. As

$l_o(\boldsymbol{\theta})$ has second continuous derivatives and if we assume $\hat{\boldsymbol{\theta}}_h$ is an interior point in a neighborhood containing $\boldsymbol{\theta}_0$, by the Lagrange theorem, we can know that there exists a vector $\tilde{\boldsymbol{\theta}}$ on the line segment between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_h$ such that

$$n^{-1}\boldsymbol{u}(\hat{\boldsymbol{\theta}}_h) = n^{-1}\boldsymbol{u}(\boldsymbol{\theta}_0) + n^{-1}M(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0), \tag{4.19}$$

where $\boldsymbol{u}(\boldsymbol{\theta}) = \partial l_o(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$ and $M(\boldsymbol{\theta}) = \partial^2 l_o(\boldsymbol{\theta})/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T$ are the first and second order derivatives of $l_o(\boldsymbol{\theta})$.

The first term $n^{-1}\boldsymbol{u}(\boldsymbol{\theta}_0)$ can be rewritten as

$$\frac{1}{n}\boldsymbol{u}(\boldsymbol{\theta}_0) = \frac{1}{n}\frac{\partial l_o(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_o^{(i)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{4.20}$$

As $\partial l_o^{(i)}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ are i.i.d. and the covariance has finite entries, we can apply the Lindeberg Central Limit Theorem and then have

$$\frac{1}{\sqrt{n}}\boldsymbol{u}(\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \bar{I}(\boldsymbol{\theta}_0)), \tag{4.21}$$

where $\bar{I}(\boldsymbol{\theta}) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I_i(\boldsymbol{\theta})$ and $I_i(\boldsymbol{\theta})$ is the information matrix for individual i. Thus, we have

$$\frac{1}{\sqrt{n}}\boldsymbol{u}(\boldsymbol{\theta}_0) = O_p(1) \tag{4.22}$$

which is equivalent to $\frac{1}{n}\boldsymbol{u}(\boldsymbol{\theta}_0) = O_p(n^{-1/2})$.

Next, we will focus on the second term $n^{-1}M(\tilde{\boldsymbol{\theta}})$ on the right side in formula (4.19). By the Law of Large Numbers, we have

$$\frac{1}{n}M(\tilde{\boldsymbol{\theta}}) = \frac{1}{n}\frac{\partial^2 l_o(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \xrightarrow{p} -\bar{I}(\tilde{\boldsymbol{\theta}}). \tag{4.23}$$

As $\bar{I}(\tilde{\boldsymbol{\theta}})$ is invertible tends to 1, the probability that $n^{-1}M(\tilde{\boldsymbol{\theta}})$ is invertible tends to 1. We can rewrite formula (4.23) as $\frac{1}{n}M(\tilde{\boldsymbol{\theta}}) = -\bar{I}(\tilde{\boldsymbol{\theta}}) + o_p(1)$. As $l_o(\boldsymbol{\theta})$ has second continuous derivatives, we know $\boldsymbol{u}(\boldsymbol{\theta}_0)$ is derivativable and we can apply Lemma (4.3.1) to derive

$$[n^{-1}M(\tilde{\boldsymbol{\theta}})]^{-1} = -\bar{I}(\tilde{\boldsymbol{\theta}})^{-1} + o_p(1). \tag{4.24}$$

Similarly, as $l_o(\boldsymbol{\theta})$ has second continuous derivatives, we can apply the Lemma (4.3.1) to the partial derivative function in the expression (4.18) and then we have

$$n^{-1}\boldsymbol{u}(\hat{\boldsymbol{\theta}}_h) = n^{-1}\boldsymbol{u}^*(\hat{\boldsymbol{\theta}}_h) + O(N^{-1}). \tag{4.25}$$

From formula (4.19), we have

$$n^{-1}M(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0) = n^{-1}\boldsymbol{u}(\hat{\boldsymbol{\theta}}_h) - n^{-1}\boldsymbol{u}(\boldsymbol{\theta}_0).$$

Thus,

$$\begin{aligned}
(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0) &= [n^{-1}M(\tilde{\boldsymbol{\theta}})]^{-1}[n^{-1}\boldsymbol{u}(\hat{\boldsymbol{\theta}}_h) - n^{-1}\boldsymbol{u}(\boldsymbol{\theta}_0)] \\
&= (-\bar{I}(\tilde{\boldsymbol{\theta}})^{-1} + o_p(1))[n^{-1}\boldsymbol{u}(\hat{\boldsymbol{\theta}}_h) - n^{-1}\boldsymbol{u}(\boldsymbol{\theta}_0)] \\
&= (-\bar{I}(\tilde{\boldsymbol{\theta}})^{-1} + o_p(1))[n^{-1}\boldsymbol{u}^*(\hat{\boldsymbol{\theta}}_h) + O(N^{-1}) + O_p(n^{-1/2}))] \qquad (4.26) \\
&= -\bar{I}(\tilde{\boldsymbol{\theta}})^{-1}O_p\left[\max\left\{n^{-1/2}, N^{-1}\right\}\right] + o_p\left[\max\left\{n^{-1/2}, N^{-1}\right\}\right] \\
&= O_p\left[\max\left\{n^{-1/2}, \left(\min_i N_i\right)^{-1}\right\}\right].
\end{aligned}$$

Finally, we use $\hat{\boldsymbol{\theta}}_{ML}$ to denote the "exact" maximum likelihood estimate with $\boldsymbol{u}(\hat{\boldsymbol{\theta}}_{ML}) = \mathbf{0}$. Let $\min_i N_i = O(n^v)$ for $v > 1$ so that the accuracy of the Laplace approximation to the marginal log-likelihood is approximately $O(n^{1-v}) = o(1)$. Then under the same regularity conditions as before, by multiplying n on the both sides of equation (4.25) and noting that $\boldsymbol{u}(\hat{\theta}_{HL}) = \mathbf{0}$, we have

$$\boldsymbol{u}(\hat{\boldsymbol{\theta}}_h) = u^*(\hat{\boldsymbol{\theta}}_h) + o_p(1) = \mathbf{0} + o_p(1) \equiv \boldsymbol{u}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(1). \tag{4.27}$$

Thus $\boldsymbol{u}(\hat{\boldsymbol{\theta}}_h) - \boldsymbol{u}(\hat{\boldsymbol{\theta}}_{ML}) = o_p(1)$ and hence $\hat{\boldsymbol{\theta}}_h$ is asymtotically equivalent to the "exact" maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{ML}$.

Next, we explore the asymprotic normality of $\hat{\boldsymbol{\theta}}_h$ and have the following theorem.

Theorem 4.3.2 *If we suppose $N = O(n^v)$ for $v > \frac{1}{2}$, $l_o(\boldsymbol{\theta})$ has second continuous derivatives, and $\partial l_o^{(i)}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ are i.i.d. with finite entries covariance, the approximate MLE $\hat{\boldsymbol{\theta}}_h$ and the "exact" MLE $\hat{\boldsymbol{\theta}}_{ML}$ have the same asymptotic distribution.*

Proof. Noting that the approximate MLE $\hat{\boldsymbol{\theta}}_h$ satisfies a set of the equations $\boldsymbol{u}^*(\hat{\boldsymbol{\theta}}_h) = 0$, we can have the following formular after taking a first-order Taylor series expansion of $\boldsymbol{u}^*(\hat{\boldsymbol{\theta}}_h)$ around the true parameter $\boldsymbol{\theta}_0$

$$\mathbf{0} = \boldsymbol{u}^*(\hat{\boldsymbol{\theta}}_h) = \boldsymbol{u}^*(\boldsymbol{\theta}_0) + \frac{\partial \boldsymbol{u}^*(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T}(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0), \tag{4.28}$$

where $\boldsymbol{\theta}^*$ is on the line segment joining $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_h$, which implies

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0) &= \left[-\frac{1}{n}\frac{\partial \boldsymbol{u}^*(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^T}\right]^{-1}\left[\frac{1}{\sqrt{n}}\boldsymbol{u}^*(\boldsymbol{\theta}_0)\right] \qquad (4.29)\\ &= \left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 p_{\hat{\boldsymbol{\omega}}_i}(l_{com}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\omega}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1}\left[-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial p_{\hat{\boldsymbol{\omega}}_i}(l_{com}^{(i)}(\boldsymbol{\theta}_0,\boldsymbol{\omega}))}{\partial\boldsymbol{\theta}}\right].\end{aligned}$$

Now we apply Lemma (4.3.1) to the first and second partial derivative functions in the expression in formula (4.18), we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\boldsymbol{u}^*(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\boldsymbol{u}(\boldsymbol{\theta}) + O(n^{1/2}N^{-1}),\\ \iff\ &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial p_{\hat{\boldsymbol{\omega}}_i}(l_{com}^{(i)}(\boldsymbol{\theta},\boldsymbol{\omega}))}{\partial\boldsymbol{\theta}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l^{(i)}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + O(n^{1/2}N^{-1}), \qquad (4.30)\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}\frac{\partial \boldsymbol{u}^*(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^T} = \frac{1}{n}\frac{\partial \boldsymbol{u}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T} + O(N^{-1}),\\ \iff\ &\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 p_{\hat{\boldsymbol{\omega}}_i}(l_{com}^{(i)}(\boldsymbol{\theta}^*))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}\frac{\partial^2 l^{(i)}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} + O(N^{-1}). \qquad (4.31)\end{aligned}$$

If we suppose $N = O(n^v)$ for $v > \frac{1}{2}$, then it is easy to know $O(n^{1/2}N^{-1}) = O(n^{1/2-v}) = o(1)$. From (4.30) and (4.31), we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial p_{\hat{\boldsymbol{\omega}}_i}(l_{com}^{(i)}(\boldsymbol{\theta}_0,\boldsymbol{\omega}))}{\partial\boldsymbol{\theta}} &= \lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l^{(i)}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}},\\ \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 p_{\hat{\boldsymbol{\omega}}_i}(l_{com}^{(i)}(\boldsymbol{\theta}^*,\boldsymbol{\omega}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l^{(i)}(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}. \qquad (4.32)\end{aligned}$$

Note that $\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0 = O_p[\max\{n^{-\frac{1}{2}}, N^{-1}\}] = O_p(n^{-\frac{1}{2}})$. Since $\boldsymbol{\theta}^*$ is on the line segment joining $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_h$, $\boldsymbol{\theta}^* \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$. Under the condition that $\partial l_o^{(i)}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ are

i.i.d. and the covariance has finite entries, it follows from (4.21) and (4.23) that

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l_o^{(i)}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} &\xrightarrow{d} N(\mathbf{0}, \bar{I}(\boldsymbol{\theta}_0)), \\ -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_o^{(i)}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &\xrightarrow{p} \bar{I}(\boldsymbol{\theta}_0). \end{aligned} \tag{4.33}$$

Combining the results in (4.32) and (4.33) and using Slutsky's theorem, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \bar{I}(\boldsymbol{\theta}_0)^{-1}). \tag{4.34}$$

Thus, we can conclude that when $N = O(n^v)$ for $v > 1/2$, the approximate MLE $\hat{\boldsymbol{\theta}}_h$ and the "exact" MLE $\hat{\boldsymbol{\theta}}_{ML}$ have the same asymptotic distribution.

4.4 Real Data Analysis

In this section, we analyze the HIV dataset described in Section 4.1 to illustrate the proposed likelihood estimation methods. The study contained 45 HIV infected patients who were given an anti-HIV treatment. Viral load, CD4 cell counts, and other variables, were repeatedly measured over a period of 48 weeks and the number of observations for each patient varied from 4 to 10. The viral load has a detectable limit of 100 RNA copies/ML. For simplicity, we impute the censored viral load values by 50 RNA copies/ML. We apply the $\log_{10}$-transformation to viral load measurements to stabilize the variance and make the data more normally distributed. It is well known that CD4 cell counts are measured with substantial errors.

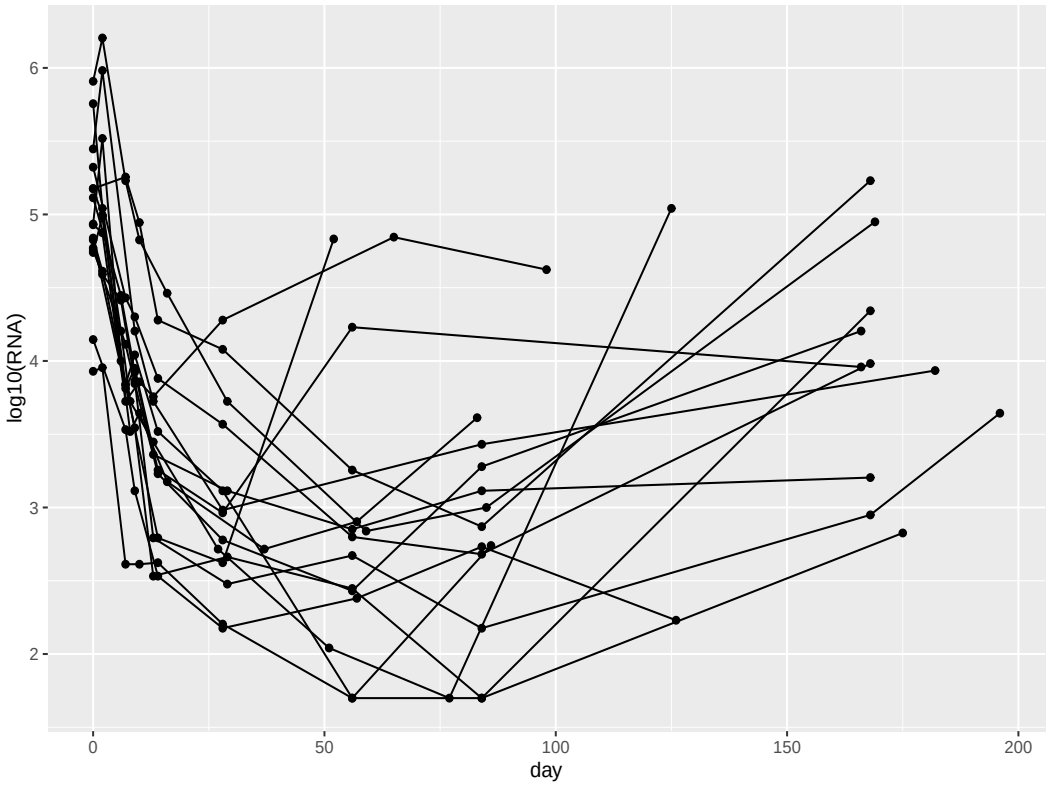

Figure 4.2: 15 viral load trajectories with change points in the study.

We consider there exists a rebount point (the lowest point) on a viral load trajectory if i). there is at least two observations after the minimum viral load is achieved or ii). there is only one observation after the minimum viral load, but the absolute value of the slope after the minimum value is larger than that just before it.

Figure 4.2 shows 15 trajectories with rebound points among the 45 patients in the study.

4.4.1 The Semiparametric NLME Response Model

We consider semiparametric NLME model (4.2) in Section 4.1 for long-term HIV viral dynamic with covariate measurement errors and change points. For completeness, we describe this model again here.

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + s_i(t_{ij})\max(t_{ij} - \tau_i, 0) + e_{ij}, \quad (4.35)$$

$$\log(P_{1i}) = \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + b_{2i}, \quad (4.36)$$

$$\log(P_{2i}) = \beta_3 + b_{3i}, \qquad \lambda_{2ij} = \beta_4 + \beta_5 CD4^*_{ij}, \quad (4.37)$$

$$s_i(t_{ij}) = \omega(t_{ij}) + h_i(t_{ij}), \quad (4.38)$$

where y_{ij} is the $\log_{10}$ transformation of the viral load measurement for the ith patient at time t_{ij}. P_{1i} and P_{2i} are baseline values, λ_{1ij} and λ_{2ij} are the first and the second phases of viral decay rates, respectively. $CD4^*_{ij}$ is the true but un-observed CD4 cell counts observed at time t_{ij}, and e_{ij} represents random errors. $\beta'_i s, i = 1, \cdots, 5$, are fix effects, and $b'_{ik}s$ are random effects which represent individual deviations. We assume that $E(\lambda_{1ij}) > E(\lambda_{2ij})$ and $\beta_1 > \beta_3$. $\omega(t_{ij})$ and $h_i(t_{ij})$ are nonparametric fixed and random smooth functions, respectively. τ_i denotes the time of the change point for the ith patient, and the expression $\max(t_{ij} - \tau_i,\ 0)$ is used to maintain continuity of viral load trajectories at the change point.

As discussed in Section 4.2, we approximate the nonparametric functions $\omega(t)$

p	q	AIC	BIC
3	3	498.70	614.61
3	2	498.18	590.91
3	1	**488.60**	**562.02**
2	2	494.56	583.43
2	1	498.39	567.94

Table 4.1: AIC and BIC values for the response model (4.35)-(4.38), with $1 \leq q \leq p \leq 3$.

and $h_i(t)$ by linear combinations of basis functions $\Psi_p(t)$ and $\Phi_q(t)$. Following Wu and Zhang (2002), we take the same natual cubic splines with $q \leq p$ in order to decrease the dimension of random effects. The AIC and BIC criteria are used to determine the values of p and q. Based on these AIC and BIC values in Table 4.1, the model with $p = 3$ and $q = 1$, i.e.,

$$s_i(t_{ij}) \approx \beta_6 + b_{4i} + \beta_7\psi_1(t_{ij}) + \beta_8\psi_2(t_{ij}) \tag{4.39}$$

seems to be the best, and thus it is selected for our analysis.

We denote $\boldsymbol{b}_i = (b_{1i}, \cdots, b_{4i})^T$ and assume $\boldsymbol{b}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, B)$. In order to reduce the number of nuisance parameters, we assume that the variance-covariance matrics B of the random effects is diagonal matrices. To avoid very small (large) estimates, we

standardize the observed but error-proned CD4 cell counts and rescale the original time t so that the new time scale is between 0 and 1.

4.4.2 The Covariate Model

It is well-known that the time-varying covariate CD4 cell counts is measured with substantial errors. If we ignore the covariate measurement errors, the statistical inference would be misleading. So we need to model the CD4 process to address the covariate measurement errors and missing values. In the absence of a theorical rationale, we employ the empirical polynomial linear mixed-effects (LME) model for the CD4 process to account for the large inter-patient variation.

Type	Random Effect	AIC	BIC
Linear	a_0	770.04	785.47
Linear	$a_0 a_1$	772.64	795.79
Linear	a_1	936.30	951.73
Quadratic	$a_0 a_1 a_2$	718.68	757.23
Quadratic	$a_0 a_1$	721.15	748.13
Quadratic	a_0	728.77	748.05

Table 4.2: AIC and BIC values for the linear and quadratic LME models

We choose the best fitted CD4 model based again on AIC and BIC criteria. Table 4.2 presents AIC and BIC values for these models. Specifically, we consider the following quadratic LME model for the CD4 process.

$$\begin{aligned} \text{CD4}_{ij} &= (\alpha_0 + a_{2i}) + (\alpha_1 + a_{2i})u_{il} + \alpha_2 u_{il}^2 + \epsilon_{il}, \quad (4.40) \\ \text{CD4}_{ij}^* &= (\alpha_0 + a_{1i}) + (\alpha_1 + a_{2i})u_{il} + \alpha_2 u_{il}^2, \quad (4.41) \end{aligned}$$

where u_{il} is the observed time, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ are the population parameters, and $\boldsymbol{a_i} = (a_{1i}, a_{2i})^T$ are the ransom effects. We assume $\boldsymbol{a_i} \overset{i.i.d.}{\sim} N(\mathbf{0}, A)$ and $\epsilon_{il} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

4.4.3 A Model for the Times of Change Points on Response Trajectories

For likelihood inference, we need to model the times of change points on response trajectories. We assume that these times of change points are related to the longitudinal process through the random effects in the semiparametric NLME response model and the covariate model. Specifically, we consider the following time-to-event model (Wu, Liu and Hu, 2010)

$$\log(\tau_i) = \gamma_0 + \gamma_1 a_{1i} + \gamma_2 a_{2i} + \gamma_3 b_{2i} + \gamma_4 b_{3i} + \zeta_i, \quad (4.42)$$

where γ_i's are parameters and the random error $\zeta_i \overset{i.i.d.}{\sim} N(0, \eta^2)$.

4.4.4 Estimation Methods and Computation Issues

We estimate the model parameters using the *naive* approach for comparison and the two proposed likelihood approaches. In the naive approach, we ignore the covariate measurement errors and the change points on the viral load trajectories. For the MCEM and h-likelihood approach, we use the estimated parameters obtained by the naive approach as the starting values of the algorithms.

For the naive approach, we use the software $\boldsymbol{R}$ function $nlme()$ to obtain parameter estimates and the standard errors. For the MCEM approach, the time series plots and the sample autocorrelation function plots are drawn to check the convergence of the Gibbs sampler. For example, in Figure 4.3 and 4.4 for patient 10, we plot the time series for the random effects $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$. The autocorrelacation function plot for b_1 associated with patient 15 is given in Figure 4.5. From these figures, we can see that the Gibbs sampler converges quickly and the autocorrelations between successive generated samples are negligible after lag 20.

Based on the findings in the time series plots and the sample autocorrelation function plots, we discard the first 500 samples as burn-in, and then choose one sample from every 20 simulated samples to obtain "independent" samples.

Convergence of the MCEM and the h-likelihood approaches are considered to be achieved when the maximum percentage change of all estimates is less than 5% in

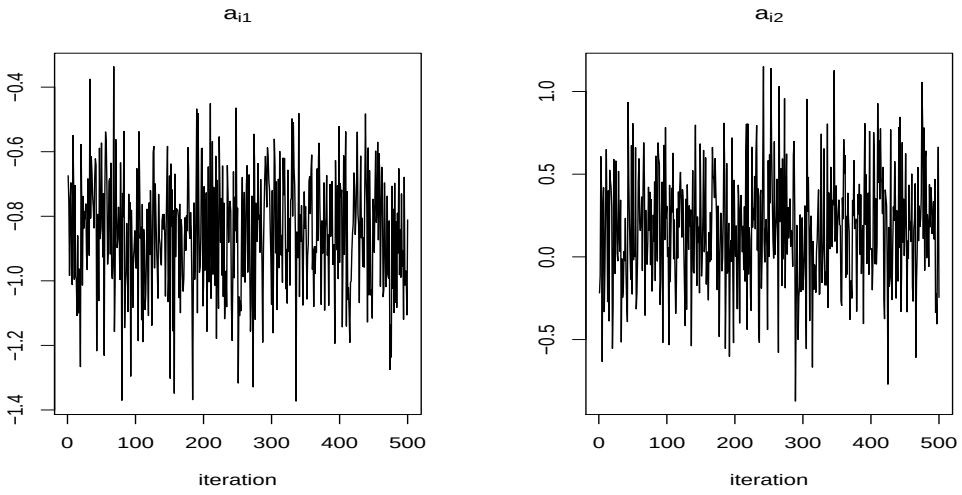

Figure 4.3: The time series plots of the sampled values of $\boldsymbol{a}_i$ for patient 10.

two consecutive iterations. In the real data analysis, the h-likelihood approach can significantly reduce the computationally time, and thus is more efficient than the MCEM approach.

4.4.5 Analysis Results

The MCEM approach and the h-likelihood (HL) approach are applied to simultaneously estimate all the model parameters in the three joint model for the viral load dynamics, the CD4 process, and the change points on the viral load trajectories. For comparison purpose, we use the naive approach for the parameter estimation, which ignoring both the covariate measurement errors and the change points. The result-

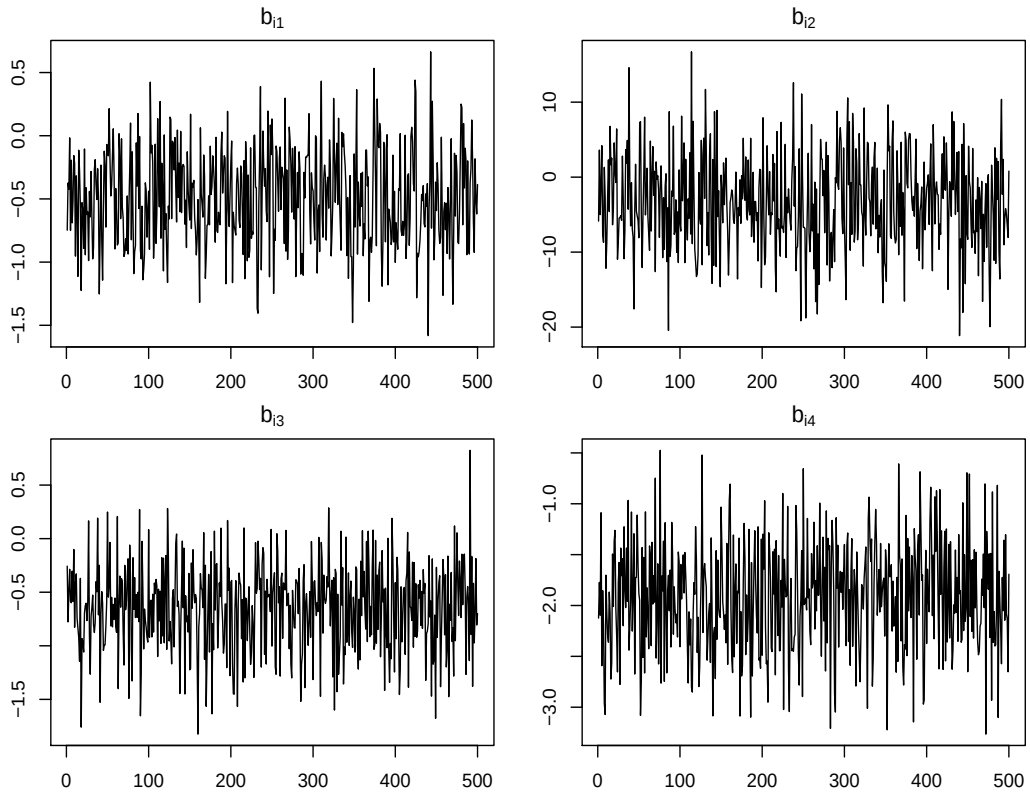

Figure 4.4: The time series plot of the sampled values of $\boldsymbol{b}_i$ for patient 10.

ing parameter estimates, together with their standard errors, are reported in Table 4.3. We can see that the naive approach may severely under-estimate the CD4 effect (i.e. β_5), and may poorly estimate some other parameters as well. The parameter estimates based on the MCEM and the h-likelihood approaches are similar and may be more reliable, while the h-likelihood approach takes much less time to implement.

The left and the right panel of Figure 4.6 present the viral load trajectories

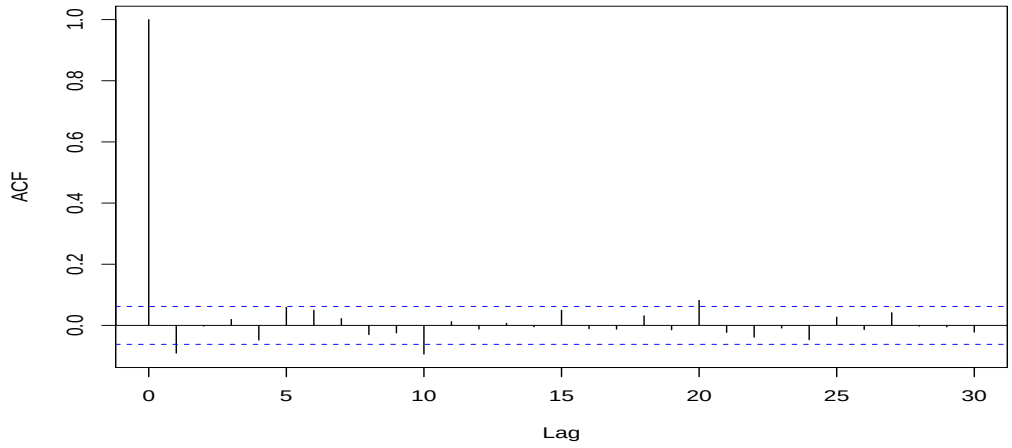

Figure 4.5: The autocorrelation function plot for b_1 associated with patient 15.

for three patients without and with change points, respectively. We can see that the naive approach fits the data poorly for all six HIV patients, especially after the change points, while both the MCEM approach and the h-likalihood approach perform very well in fitting the observed viral load trajectories with and without change points.

4.5 The Simulation Study

In the simulation study, we evaluate the proposed joint approaches (MCEM and HL), and compare them with the naive approach. We generate 100 datasets from

Method	α_0	α_1	α_2	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Naive				11.64	54.93	6.27	-0.75	0.97			
				(0.18)	(3.24)	(0.27)	(0.41)	(0.27)			
MCEM	-0.42	4.78	-4.60	11.73	61.99	6.87	1.81	1.58	17.14	-18.52	-12.53
	(0.15)	(0.49)	(0.58)	(0.13)	(4.05)	(0.19)	(0.51)	(0.24)	(2.99)	(4.82)	(2.68)
HL	-0.49	4.72	-4.20	11.70	62.77	6.86	2.08	1.53	16.87	-18.55	-12.75
	(0.05)	(0.42)	(0.53)	(0.13)	(4.08)	(0.19)	(0.47)	(0.27)	(2.87)	(4.63)	(2.58)
Method	γ_0	γ_1	γ_2	γ_3	γ_4	δ	σ	η			
Naive						0.49					
MCEM	0.15	0.48	0.30	0.03	0.27	0.33	0.51	1.16			
	(0.05)	(0.04)	(0.03)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)			
HL	0.19	0.53	0.31	0.02	0.26	0.29	0.49	1.22			
	(0.03)	(0.05)	(0.07)	(0.01)	(0.03)	(0.02)	(0.03)	(0.04)			

Table 4.3: Estimates (standard errors) of the parameters in the joint models in the example.

the following model, which corresponds to model (4.35)-(4.38),

$$y_{ij} = \log_{10}(P_{1i}e^{-\lambda_{1ij}t_{ij}} + P_{2i}e^{-\lambda_{2ij}t_{ij}}) + s_i(t_{ij})\max(t_{ij}-\tau_i, 0) + e_{ij}, \quad (4.43)$$

$$log(P_{1i}) = \beta_1 + b_{1i}, \qquad \lambda_{1ij} = \beta_2 + b_{2i}, \quad (4.44)$$

$$log(P_{2i}) = \beta_3 + b_{3i}, \qquad \lambda_{2ij} = \beta_4 + \beta_5 CD4^*_{ij}, \quad (4.45)$$

$$s_i(t_{ij}) = (4.8 + 0.1b_{4i})\sin(4.2 + 3.1t_{ij}), \quad (4.46)$$

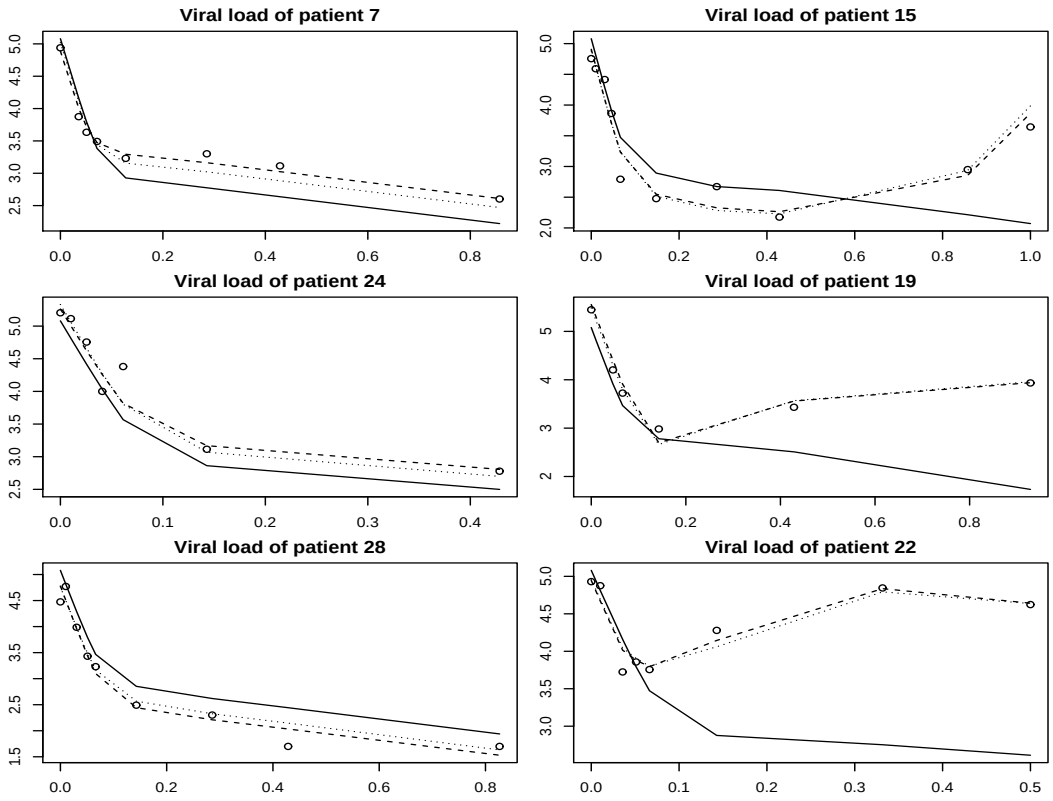

Figure 4.6: The observed (open-circle) and the fitted viral load trajectories for randomly selected three HIV patients without change points (left panel) and three patients with change points (right panel) based on the naive approach (solid line), the MCEM approach (dashed line), and the h-likelihood approach (dotted line).

where the nonparametric model (4.43)-(4.46) is carefully chosen to closely mimic the observed viral load trajectory after the change point in the real-data example in the previous section. The covariate model and the model for the times of change

points are the same as those in the real-data example. We set up $n = 50$ and $n_i = 15$, $i = 1, \cdots, n$, with the equal-spaced measurement time points between 0 and 1. The true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as well as $\boldsymbol{\gamma}$ and the precision parameters are presented in Tables (4.4) and (4.5), respectively. The true variance-covariance matrixes $\boldsymbol{A} = \text{diag}(0.7, 1.2)$ and $\boldsymbol{B} = \text{diag}(1, 9, 2, 4)$. In the simulation study, when we use the MCEM approach, we discard the first 1000 samples as burn-in, and then choose one sample from every 30 simulated samples to obtain "independent" samples.

Firstly, we evaluate the nonparametric modeling by studying the performance of AIC and BIC for selecting the numbers of knots, i.e. p and q. For the 100 simulated dataset from models (4.43)-(4.46), we find that most of the BIC values and AIC values lead to $p = 3$, $q = 1$. Furthermore, we simulate 100 datasets from models (4.35)-(4.38) with $(\beta_6, \beta_7, \beta_8) = (17, -18.5, -12.5)$, and then use AIC and BIC criteria to select the best model. The performance of AIC and BIC criteria is again excellent and similar to the foregoing results. Thus, we may conclude that AIC and BIC criteria perform well in the current modelsetting.

To compare different estimation approaches, we calculate averages of the resulting estimates (EST) and their standard errors, the percent relative biases (BIAS) defined by $(\hat{\beta}_j - \beta_j)/|\beta_j| \times 100\%$, and the percent relative root mean-square-errors (MSE) defined by $100 \times \sqrt{MSE_j}/|\beta_j|$ based on each of three approaches. BIAS and MSE

for the nonparametric fixed-effects function are defined by $\frac{1}{n}\int_{\tau_i}^{1}[\hat{\omega}_i(t)-\omega(t)]dt$ and $\frac{1}{n}\int_{\tau_i}^{1}[\hat{\omega}_(t)-\omega(t)]^2dt$, respectively, where $\hat{\omega}_i(t)$ and $\omega(t)$ are the ith patient's fitted and true trajectory, respectively. These simulations results are reported in Tables 4.4 and 4.5.

	Parameter	α_0	α_1	α_2	β_1	β_2	β_3	β_4	β_5	$\beta_6 \sim \beta_8$
	True value	-0.45	4.75	-4.5	11.7	62	7	1.8	1.5	$4.8\sin(4.2+3.1t)$
EST	Naive				9.72 (0.38)	33.29 (5.52)	3.93 (0.33)	-1.65 (0.34)	0.41 (0.09)	
	MCEM	-0.44 (0.15)	4.77 (0.31)	-4.48 (0.45)	11.68 (0.23)	61.85 (3.21)	6.96 (2.55)	1.76 (0.41)	1.51 (0.35)	
	HL	-0.44 (0.17)	4.71 (0.34)	-4.55 (0.47)	11.66 (0.20)	62.37 (3.56)	7.04 (2.49)	1.85 (0.38)	1.53 (0.31)	
BIAS	Naive				-17.51	-45.49	-42.13	-195.19	-78.24	50.41
	MCEM	-2.34	0.53	-0.44	-0.27	-0.20	-0.48	-2.47	0.84	-1.43
	HL	-2.47	-0.90	1.37	-0.54	0.79	0.54	2.62	1.75	2.48
MSE	Naive				30.41	45.98	46.14	253.19	110.43	57.05
	MCEM	3.87	1.17	0.74	0.52	0.30	1.12	4.18	1.54	2.42
	HL	4.05	1.75	2.15	0.86	1.63	1.34	5.37	2.45	4.68

Table 4.4: Simulation results for the estimates (standard errors) of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

From the simulation results in Tables 4.4 and 4.5, we can see that these two proposed joint model approaches (MCEM and HL) perform well in terms of both

BIAS and MSE. MCEM approach performs better than HL approach as expected, while HL approach also performs reasonally well and is computationally much more efficient. The naive approach may lead to severely biased estimates and large MSEs for some parameters. Therefore, it is important and necessary to take covariate measurement errors and change points into account when analyzing the longitudinal response data.

	Parameters	γ_0	γ_1	γ_2	γ_3	γ_4	δ	σ	η
	True value	0.15	0.48	0.3	0.028	0.27	0.33	0.51	1.16
Estimate	MCEM	0.15 (0.13)	0.47 (0.15)	0.3 (0.13)	0.03 (0.01)	0.26 (0.03)	0.33 (0.04)	0.52 (0.02)	1.19 (0.04)
	HL	0.15 (0.14)	0.50 (0.21)	0.31 (0.16)	0.03 (0.01)	0.28 (0.03)	0.32 (0.05)	0.48 (0.03)	1.21 (0.06)
Bias	MCEM	-0.78	-1.69	2.61	2.37	-3.16	-2.12	2.81	2.58
	HL	0.86	4.76	3.43	3.61	3.42	-4.28	-5.45	4.17
MSE	MCEM	1.95	3.65	6.53	5.34	6.13	4.35	6.35	4.98
	HL	2.15	7.69	7.67	7.33	6.54	7.19	8.76	7.45

Table 4.5: Simulation results for the estimates (standard errors) of $\boldsymbol{\gamma}$ and the precision parameters.

4.6 Conclusion

In this chapter, we simultaneously address the measurement errors in time-varying covariates and change points for semiparametric NLME models. To obtain the approximate MLE of joint model parameters, we implement two approaches: the MCEM and the h-likelihood approach. Both approaches can derive reliable results and the h-likelihood approach can be computationally efficient. Thus, the h-likelihood approach can be used as an alternative one to estimate the parameters in the joint models. The h-likelihood approach may also provide excellent parameter starting values for the MCEM approach. The simulation study shows that the two proposed approaches produce satisfactory results, while the naive approach, which ignores the measurement errors and change points, may perform poorly.

In the real data analysis, we impute the censored response values by half of the detection limit of response for simiplicity. It is reasonable to treat the response values below the detection limit as the left-censored data and include them in the likelihood inference. The proposed approaches may be extended to analyse this type of datasets.

5 Discussion

5.1 Summary

In Chatpter 2, we are aim to propose simultaneously efficient and efficient methods for change point detection. For the univariate dataset, we propose COS and EXP method to test the existence of the change point and then detect it if it exists. We explore the asymptotic results of our proposed test statistics and analyze the type I error as well as the power in this chapter. We employ ICSS algorithm to extend our methods to detect multiple change point in data sequences and achive satisfactory results. The main advantage of our test statistics is that we can detect the change points effectively and efficiently. The comparison of our methods with some other popular multiple change points detection methods is present in Section 2.4.

In Chapter 3, we proposed a procedure, as well as its theoretical justification, for detecting multiple change points in the mean-shift model, where the number of change points is allowed to increase with the sample size. We first convert a

change point detection problem into a variable selection problem by partitioning the data sequence. Once the segment containing a possible change point is flagged, a weighted CUSUM algorithm is applied to test if there is a change point in this segment. Simulation studies and real data analysis have provided solid ground to prove that our procedure is efficient and effective.

In Chapter 4, we simultaneously address the measurement errors in time-varying covariates and change points for semiparametric NLME models. To obtain the approximate MLEs of the joint model parameters, we have proposed the two approaches: the MCEM and the h-likelihood approach. We illustrate the proposed approaches by analyzing a real HIV dataset and evaluate their performance by conducting a simulation study. The simulation study shows both approaches may produce satisfactory results in estimating the joint model parameters, while the naive approach, which ignores the measurement errors and change points, may perform poorly.

5.2 Future Research

The proposed COS method performs well when there exists significant mean shift in the dataset after the cosine transformation. When the mean shift of $cos(X)$ is too small, the performance of COS method is not satisfactory.

With the rapid development of multiple change points detection methods, our proposed algorithm in Chapter 3 may play important role in detecting some other types of change points, such as changes in variance, changes in distribution. On the other hand, we could advance our method to detect change points in the multivariate dataset. Both of these two improvements will make our method more general and effective in practice.

There are several research topics that can be explored in the future. Firstly, it is reasonable and meaningful to treat the response values below the detection limit as the left-censored data and include them in the likelihood inference. Secondly, it is very common that some individuals may drop out of the study before the scheduled end for various reasons such as drug intolerance. The dropout may be informative in the sense that it may be related to the missing values. We may extend the proposed approaches to take the informative dropout into account. Thirdly, the h-likelihood approach is much more computationally efficient than the MCEM approach. We may explore other approximate likelihood approaches such as a first-order Taylor approximation to the nonlinear functions in the response model.

Bibliography

[1] Auger, I. E., Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, **51(1)**, 39-54.

[2] Bai, J. (1994). Least squares estimationi of a shift in linear processes, *Journal of Time Series Analysis*, **15**, 453-472.

[3] Bai, J., Perron, P. (1998). Estimating and testing linear models with multiple structural changes, *Econometrica*, **66(1)**, 47-78.

[4] Bleakley, K., Vert, J.-P. (2011). The group fused Lasso for multiple change-point detection, `http://hal.archives-ouvertes.fr/docs/00/60/21/21/PDF/techreport.pdf`.

[5] Bolton, R., and Hand, D. (2002). Statistical fraud detection: a review, *Statistical Science*, **17**, 235-255.

[6] Cadima, J., and Jolliffe, I. (1995). Loadings and correlations in the interpretation of principal components, *Journal of Applied Statistics*, **22**, 203-214.

[7] Candès, E., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis?, *Journal of the ACM (JACM)*, **58(3)**, 11.

[8] Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hier-archical Bayesian analysis of changepoint problems, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **41**, 389-405.

[9] Carroll, R.J., Ruppert, D., Stefanski, L.A. (1995). *Measurement error in nonlinear models*, Chapman and Hall, London.

[10] Chen, J., Gupta, A.K. (2012). *Parametric statistical change point analysis*, Birkhäuser.

[11] Chernoff, H., Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time, *Annals of Mathematical Statistics*, **35**, 999-1018.

[12] Croux, C., Filzmoser, P., and Fritz, H. (2013). Robust sparse principal component analysis, *Technometrics*, **55**, 202-214.

[13] Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures, *Biometrika*, **4**, 407-410.

[14] Csörgő, S. (1984). Testing by the empirical characteristic function: a survey. *Asymptotic Statistics*, **2**, 45-56.

[15] Csörgő, M., Horváth, L. (1997). *Limit theorems in change-point analysis*, Chichester: Wiley.

[16] Davis, R.A., Huang, D. and Yao Y.C. (1995). Testing for a change in the parameter values and order of an autoregressive model, *The Annals of Statistics*, **23**, 282-304.

[17] Dong, C., Miao, B., Tan, C., Wei, D., and Wu, Y. (2015). An estimate of a change point in variance of measurement errors and its convergence rate, *Communications in Statistics - Theory and Methods*, **44**, 790-797.

[18] Duggins, J. W. (2010). *Parametric resampling methods for retrospective change-point analysis*, Dissertation, Virginia Tech, Blacksburg, Virginia, USA.

[19] Filzmoser, P., Fritz, H., and Kalcher, K. (2016). pcaPP: robust PCA by projection pursuit, R package version 1.9-61. Available at `https://CRAN.R-project.org/package=pcaPP`.

[20] Gardner, L. A. (1969). On detecting change in the mean of normal variates, *Annals of Mathematical Statistics*, **40**, 116-126.

[21] Gelfand A. and Smith A. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398-409.

[22] Gosset, W. S. (1908). The probable error of a mean, *Biometrika*, **6**, 1-25.

[23] Gupta AK and Tang J. (1987). On testing homogeneity of variances for gaussian models, *Journal of Statistical Computation and Simulation*, 27(2), 155-173.

[24] Hall, C. B., Ying, J., Kuo, L., and Lipton, R. B. (2003). Bayesianand profile likelihood change point methods for modelingcognitive function over time. *Computational Statistics and Data Analysis*, **42**, 91-109.

[25] Hawkins, D. M. (1992). Detecting shifts in functions of multivariate locationand covariance parameters, *Journal of Statistical Planning and Inference*, **33**, 233-244.

[26] Hedeker, D., Gibbons, R. D. (2006). *Longitudinal data analysis*, Wiley Publications.

[27] Henderson, R., Diggle, P.J., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data, *Biostatistics*, **1**, 465-480.

[28] Heathcote, C. R. (1972). A test of goodness of fit for symmetric random variables, *Aust. J. Statist*, **14**, 172-182.

[29] Henze, N., Hlávka, Z., Meintanis, S.G. (2014). Testing for spherical symmetry via the empirical characteristic function, *Statistics*, **48**, 1282-1296.

[30] Higgins, D.M., Davidian, M., Giltinan, D.M. (1997). A two-step approach to measurement error in time-dependent covariates in nonlinear mixed-effects models, with application to IGF-I pharmacokinetics. *J. Amer. Statist. Assoc.* **92**, 436-448.

[31] Hinkley, D. (1970). Inference about the change-point in a sequence of random variables, *Biometrika*, **57**, 1-17.

[32] Hlávka, Z., Hušková, M., Kirch, C. and Meintanis, S.G. (2012). Monitoring change in the error distribution of autoregression models based on Fourier methods, *Test*, **21**, 605-634 .

[33] Huang Y. (2013). Segmental modeling of viral load changes for HIV longitudinal data with skewness and detection limits, *Stat Med.*, **32(2)**, 319-334.

[34] Huang Y., Chen J. (2016). Bayesian quantile regression-based nonlinear mixed-effects joint models for time-to-event and longitudinal data with multiple features, *Statistics in Medicine*, **35**, 5666-5685.

[35] Hušková, M. and Meintanis, S.G. (2006a). Change point analysis based on empirical characteristic function, *Metrika*, **63**, 145-168.

[36] Hušková, M. and Meintanis, S.G. (2006b). Change point analysis based on empirical characteristic function of ranks, *Sequential Analysis*, **25**, 421-436.

[37] Hušková, M. and Kirch, C. (2008). Bootstrapping confidence intervals for the change-point of time series, *J. of Time Series Analysis*, **29**, 947-972.

[38] Hušková, M. and Meintanis, S.G. (2009). Goodness-of-fit tests for parametric regression models based on empirical characteristic functions, *Kybernetika*, **45**, 960-971.

[39] Hušková, M., Pràškovà, Z. and Steinebach, J. (2007). On the detection of changes in autoregressive times series I. Asymptotics, *Journal of Statistical Planning and Inference*, **137**, 1243-1259.

[40] Ibrahim, J.G., Chen, M.H., Lipsitz, S.R., (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable, *Biometrika*, **88**, 551-564.

[41] Inclán, C. and Tiao, G. (1994). Use of cumulative sums of squares for retrospective detection of change of variance, *Journal of the American Statistical Association*, **89**, 913-923.

[42] Jackson B., Sargle J., Barnes D., Arabhi S., Alt A., Gioumousis P., Gwin E., Sangtrakulcharoen P., Tan L., Tsai, T.T. (2005). An algorithm for optimal partitioning of data on an interval, *IEEE Signal Processing Letters*, **12**, 105-108.

[43] Jacqmin-Gadda, H., Commenges, D., and Dartigues, JF. (2006). Random changepoint model for joint modeling of cognitivedecline and dementia, *Biometrics*, **62**, 254-260.

[44] James, N. A., Matteson, D. S. (2015). ecp: An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data, *Journal of Statistical Software*, **62**, 1-25.

[45] Jiménez-Gamero, M.D., Batsidis, A., Alba-Fernández, M.V. (2016). Fourier methods for model selection, *Ann. Inst. Statist. Math*, **68**, 105-133.

[46] Jin, B., Shi, X., and Wu, Y. (2013). A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models, *Statistics and Computing*, **23**, 221-231.

[47] Jolliffe, J. (1995). Rotation of principal components: choice of normalization constraints, *Journal of Applied Statistics*, **22**, 29-35.

[48] Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the LASSO, *Journal of Computational and Graphical Statistics*, **12**, 531-547.

[49] Ke, C., Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications (with discussions), *Journal of the American Statistical Association*, **96**, 1272-1298.

[50] Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost, *J. Am. Statist. Ass.*, **107**, 1590-1598.

[51] Kim, A., Marzban, C., Percival, D., and Stuetzie, W. (2009). Using labeled data to evaluate change detectors in a multivariate streaming environment, *Signal Processing*, **89(12)**, 2529-2536.

[52] Kirch C. (2007). Block permutation principles for the change analysis of dependent data, *J. Statist. Plann. Inference*, **137**, 2453-2474.

[53] Kiuchi, A. S., Hartigan, J. A., Holford, T. R., Rubinstein, P., and Stevens, C. E. (1995). Change points in the series of T4 countsprior to AIDS. *Biometrics*, **51**, 236-248.

[54] Koutrouvelis I.A. (1980). A goodness of fit test for simple hypothesis based on the empirical characteristic function, *Biometrika*, **67**, 238-240.

[55] Lai, T. (2001). Sequential analysis: some classical problems and new challenges, *Statistica Sinica*, **11(2)**, 303-408.

[56] Larid N.M., Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, **57**, 253-259.

[57] Lavielle, M. and Teyssière, G. (2006). Detection of multiple change-points in multivariate time series, *Lithuanian Mathematical Journal*, **46**, 287-306.

[58] Lee, C. (1998). Bayesian estimation of the number of change points, *Statistica Sinica*, **8**, 923-939.

[59] Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models, *Journal of the Royal Statistical Society B*, **58**, 619-678.

[60] Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized linear models with random effects: unified analysis via h-likelihood*, Chapman and Hall/CRC, London.

[61] Liang, H. (2007). Segmentail modeling of changing viral load to assess drug resistance in HIV infection, *Statistical Methods in Medical Research*, **17**, 365-373.

[62] Lin, X., Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error, *Journal of the American Statistical Association*, **95**, 520-534.

[63] Lin, X., Carroll R.J. (2001). Semiparametric regression for clustered data using generalized estimating equations, *Journal of the American Statistical Association*, **96**, 1045-1056.

[64] Lin, D., Ying, Z. (2001). Nonparametric tests for the gap time distributions of serial events based on censored data, *Biometrics*, **57**, 369-375.

[65] Liu, W. (2006). The theory and methods for measurement errors and missing data problems in semiparametric nonlinear mixed-effects models (T). Retrieved from `https://open.library.ubc.ca/collections/ubctheses/831/items/1.0092843`.

[66] Liu, W., Wu, L. (2007). Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses, *Biometrics*, **63**, 342-350.

[67] Liu, W., Wu, L. (2008). A semiparametric nonlinear mixed-effects model with non-ignorable missing data and measurement errors for HIV viral data, *Computational Statistics and Data Analysis*, **53**, 112-122.

[68] Loader, C. R. (1996). Change point estimation using nonparametric regression, *Ann. Statist*, **24**, 1667-1678.

[69] Madan, D.B., Seneta, E. (1987). Simulation of estimates using the empirical characteristic function, *International Statistical Review*, **55**, 153-161.

[70] Matteson, D. S., James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data, *J. Amer. Statist. Assoc.*, **109**, 334-345.

[71] McCabe, G. (1984). Principal variables, *Technometrics*, **26**, 137-144.

[72] Meintanis, S. (2004). Goodness-of-fit tests for the logistic distribution based on empirical transforms, *Sankhyā*, **66**, 306-326.

[73] Meintanis, S., James, A. (2014). A nonparametric approach for multiple change point analysis of multivariate data, *J. Amer. Statist. Assoc.*, **109**, 334-345.

[74] Meintanis, S.G., Swanepoel, J., Allison, J. (2014). The probability weighted characteristic function and goodnessoffit testing. *J. Statist. Plann. Inference*, **146**, 122-132.

[75] Morrell, C. H., Pearson, J. D., Carter, H. B., and Brant, L. J. (1995). Estimating unknown transition times using a piece-wise nonlinear mixed-effects model in

men with prostate can-cer, *Journal of the American Statistical Association*, **90**, 45-53.

[76] Moyeed, R.A., Diggle, P.J. (1994). Rates of convergence in semi-parametric modeling of longitudinal data, *Australian Journal of Statistics*, **36**, 75-93.

[77] Muggeo, V. M., and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements, *Bioinformatics*, **27**, 161 -166.

[78] Muller, H. (1992). Change points in nonparametric regression analysis, *Annals of Statistics*, **20**, 737-761.

[79] Ombao, H., Raz, J. A., von Sachs, R. and Malow, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series, *J. Am. Statist. Ass.*, **96**, 543-560.

[80] Page, E. S. (1955). A test for a chance in a parameter occurring at an unknown point, *Biometrika*, **42**, 523-527.

[81] Page, E.S. (1957). On problem in which a change in a parameter occurs at an unknown point, *Biometrika*, **44**, 248-252.

[82] Press, S.J. (1972). Estimation in univariate and multivariate stable distributions, *J. Amer. Statist. Assoc.*, **67**, 842-846.

[83] Ombao, H., von Sachs, R., and Guo, W. (2005). SLEX analysis of multivariate nonstationary time series, *Journal of the American Statistical Association*, **100**, 519-531.

[84] Qian G., Shi X., and Wu Y. (2014). A statistical test of change-point in mean that almost surely has zero error probabilities, *Aust. N. Z. J. Stat.*, **55**, 435-454.

[85] Raimondo, M. (1998). Minimax estimation of sharp change points, *Ann. Statist.* **26**, 1379-1397.

[86] Rice, J.A., Wu, C.O. (2001). Nonparametric mixed-effects models for unequally sampled noisy curves, *Biometrics*, **57**, 253-259.

[87] Rousseeuw, P., and Croux, C. (1993). Alternatives to the median absolute deviation, *Journal of the American Statistical Association*, **88**, 1273-1283.

[88] Scott, AJ, Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance, *Biometrics*, **30(3)**, 507-512.

[89] Sen, A., Srivastava, M.S. (1975). On tests for detecting change in mean, *The Annals of Statistics*, **3(1)**, 98-108.

[90] Shao, X. and Zhang, X. (2010). Testing for change points in time series, *Journal of the American Statistical Association*, **105**, 1228-1240.

[91] Shi, X., Wu, Y., and Miao, B. (2009). Strong convergence rate of estimators of change point and its application, *Computational Statistics and Data Analysis*, **53**, 990-998.

[92] Shi, X., Wang, X., Wei, D., and Wu. Y. (2016). A sequential multiple change-point detection procedure via VIF regression, *Computational Statistics*, **31(2)**, 671-691.

[93] Székely GJ, Rizzo ML. (2010). Disco analysis: a nonparametric extension of analysis of variance, *The Annals of Applied Statistics*, **4(2)**, 1034-1055.

[94] Tenreiro, C. (2011). An affine invariant multiple test procedure for assessing multivariate normality, *Comput. Statist. Data*, **55**, 1980-1992.

[95] Tsiatis, AA., Davidian M. (2004). An overview of joint modeling of longitudinal and time-to-event data, *Statistica Sinica*, **14**, 793-818.

[96] Vert, J., Bleakley, K. (2010). Fast detection of multiple change-points shared by many signals using group LARS, *Advances in Neural Information Processing Systems*, **23**, 2343-2351.

[97] Vines, S. (2000). Simple principal components, *Applied Statistics*, **49**, 441-451.

[98] Wang, Y. (1995). Jump and sharp cusp detection by wavelets, *Biometrika*, **82**, 385-397.

[99] Wang, G. and Wang, S. (2006). Recursive computation of tchebichef moment and its inverse transform, *Pattern Recognition*, **39**, 47-56.

[100] Wei. C. G., Tanner M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, **85**, 699-704.

[103] Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *J. Amer. Statist. Assoc.* **97**, 955-964.

[102] Wu, H. (2005). Statistical methods for HIV dynamic studies in AIDS clinical trails, *Statistical Methods in Medical Research*, **14**, 171-192.

[103] Wu, H., Zhang, J. (2002). The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models, *Statistical in Medicine*, **21**, 3655-3675.

[104] Wu, L., Hu, X., and Wu, H. (2008). Joint inference for nonlinear mixed-effects models and time to event at the presence of missing data, *Biostatistics*, **9(2)**, 308-320.

[105] Wu, L., Liu, W., Hu, X. J. (2010). Joint inference on HIV viral dynamics and immune suppression in presence of measurement errors, *Biometrics*, **66**, 327-335.

[106] Wu, M.C., Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics*, **44**, 175-188.

[107] Ushakov, N. G. (1999). *Selected topics in characteristic functions*, Utrecht: VSP.

[108] Zeger, S.L., Liang, K., and Self, S.G. (1985). The analysis of binary longitudinal data with time-independent covariates, *Biometrika*, **72(1)**, 31-38.

[109] Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, **15(2)**, 265-286.

[110] Zou, C., Yin, G., Feng, L. and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems, *The Annals of Statistics*, **42**, 970-1002.