

UNIVERSITY OF CALGARY

Nonparametric Change Point Detection for Univariate and Multivariate Non-Stationary
Time Series

by

Zixiang Guan

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

December, 2018

© Zixiang Guan 2018

Abstract

This thesis investigates the change point detection problem for non-stationary time series in a nonparametric way. Two topics, which are nonparametric change point detection method for univariate time series and multivariate time series, are studied respectively.

In the first topic, we consider a nonparametric method for detecting change points in non-stationary time series. The proposed method will divide the time series into several segments so that between two adjacent segments, the normalized spectral density functions are different. The theory is based on the assumption that within each segment, time series is linear process, which means that our method works not only for causal and invertible ARMA process, but also can be applied to non-invertible Moving Average process. We show that our estimations for change points are consistent. Also, a Bayesian information criterion is applied to estimate the number of change points consistently. Simulation results as well as empirical results will be presented.

A nonparametric method for detecting change points in multivariate non-stationary time series is our second topic. Under the assumption that non-stationary time series consists of several stationary ones, samples will be segmented into several stationary parts so that between two adjacent time series, the normalized eigenvalues of spectral density matrices are different. Also, we assume that stationary time series are multivariate linear processes, which means that our method works for several classic multivariate time series model, e.g., ARMA process, non-invertible moving average process, which is not considered much in literature. We show that our estimations for change points are consistent. Also, a Bayesian information criterion is applied to estimate the number of change points consistently, which is similar to the univariate case.

Acknowledgements

The thesis will not be possible without the help and encouragement from many people. Here I would like to express my gratefulness to them.

I would show my deepest appreciation to my supervisor, Dr. Gemai Chen, for his consistent patience. His door is always open for me whenever I run into trouble. He guided me into the right direction whenever it is necessary. He encourages me persistently whenever I was no longer determined.

I want to thank my supervisory committee members, Dr. Xuewen Lu, Dr. Alexander De Leon, for their insightful comments and suggestions. I appreciate the comments of Dr. Tak Fung and Dr. Roman Viveros-Aguilera. Also, I want to thank Dr. Ying Yan and Dr. Tak Fung for their helpful comments during my oral candidacy exam.

I am grateful to Ms. Yanmei Fei for her lots of suggestions on the life in Calgary. I also appreciate Mr. Freddie Yau for his help on IT issues. I would like to thank Professor Jim Stallard for advice on teaching assistant.

I want to express my gratitude towards several classmates and friends, Qicheng Zhang, Xiaofan Zhou, Wenyan Zhong, Lijun Shao, Kaida Cai, Longlong Huang, Yuan Dong, Kunlin Hao, Jixian Li, Jialin He, Hongyan Wang, Lin Zuo, for their help beyond the study in Calgary.

Last but not the least, I would like to express my sincere gratitude to my parents, Qiuming Guan, Kuanling Suo, for their persistent support and invaluable love during my stay in Calgary.

The financial support from the Department of Mathematics and Statistics is also appreciated.

Table of Contents

Abstract	ii
Acknowledgements	iii
Table of Contents	v
List of Figures and Illustrations	vii
List of Tables	viii
List of Symbols	ix
1 Introduction	1
2 Spectrums and Linear Process	5
2.1 Motivation	5
2.2 Estimation	8
2.3 Multivariate Time Series, Spectral Density Matrix and Its Estimation	10
2.4 Kullback-Leibler Divergence	13
2.5 Change Point Model and Algorithm	13
2.6 Non-Stationary Time Series with Change Points	15
2.7 Some Preliminary Simulations	16
3 Univariate Nonparametric Change Point Detection for Linear Process	21
3.1 Model and Methodology	22
3.2 Asymptotic Theory	24
3.3 Algorithm	26
3.4 Simulation	29
3.4.1 Autoregressive Process	30
3.4.2 ARMA Process and Invertible Moving Average Process	32
3.4.3 Non-invertible MA Process	34
3.4.4 Random Noise without the Existence of Higher Moments	35
3.4.5 Investigation of Smaller Sample Size	36
3.4.6 Further Investigation of BIC Criterion	37
3.4.7 Influence of Change Point Searching Unit	44
3.5 Case Study	45

3.5.1	Simulated Data	45
3.5.2	Infant Electrocardiogram Data (ECG)	47
3.5.3	Electroencephalography Recordings for Seizure	49
4	Multivariate Nonparametric Change Point Detection	51
4.1	Model and Methodology	53
4.2	Asymptotic Theory	55
4.3	Algorithm	56
4.4	Simulation	58
4.4.1	Autoregressive Process	59
4.4.2	ARMA Process and Invertible Moving Average Process	60
4.4.3	Non-invertible Moving Average Process	62
4.4.4	White Noise without the Existence of Higher Moments	63
4.4.5	Investigation of Smaller Sample Size	63
4.4.6	Further Investigation of BIC criterion	64
4.5	Case Study	68
4.5.1	Simulated Data	68
4.5.2	A Study on EEG Data	70
5	Conclusions and Discussions	72
	Bibliography	74
A	Proofs in the Thesis	78
A.1	Proofs in Chapter 3	78
A.2	Proofs in Chapter 4	94

List of Figures and Illustrations

2.1	Spectral Density Function of ARMA(2,1) Process: $X_t - X_{t-1} + 0.25X_{t-2} = \xi_t + 0.8\xi_{t-1}$,	5
2.2	Periodogram of an ARMA(2,1) Sample of Size 1800,	8
2.3	Smoothed Spectral Density Function of ARMA(2,1),	10
2.4	Estimation Accuracy of Spectral Density Function of ARMA(2,1) with Different Bandwidths	17
2.5	Estimation Accuracy of Spectral Density Function of ARMA(2,1) with Different Sample Sizes	18
2.6	Estimation Accuracy of the Maximum Eigenvalue of Spectral Density Matrix of Multivariate ARMA(1,1) with Different Bandwidths	19
2.7	Estimation Accuracy of the Maximum Eigenvalue of Spectral Density Matrix of Multivariate ARMA(1,1) with Different Sample Sizes	20
3.1	Normalized Spectral Density Functions in Davis <i>et al.</i> (2006)	31
3.2	Influence of m and ml on BIC Criterion with $m = N^{1/4}$ and $ml = 350$	38
3.3	Influence of m and ml on BIC Criterion with $m = N^{1/3}$ and $ml = 350$	39
3.4	Influence of m and ml on BIC Criterion with $m = N^{1/3}$ and $ml = 300$	40
3.5	Influence of m and ml on BIC Criterion with $m = N^{1/4}$ and $ml = 300$	41
3.6	Influence of m and ml on BIC Criterion with $m = N^{1/3}$ and $ml = 250$	42
3.7	Influence of m and ml on BIC Criterion with $m = N^{1/4}$ and $ml = 250$	43
3.8	Estimated Change Points by NSCD for Different Cases	46
3.9	Estimated Spectrums by NSCD for Different Cases	47
3.10	Infant ECG Data and the Analysis	48
3.11	Analysis of Different Channels of EEG Data	50
3.12	Estimated Spectrums of Different Channels of EEG Data	50
4.1	Influence of m on BIC Criterion with $m = N^{1/4}$	65
4.2	Influence of ml on BIC Criterion with $ml = 150$	66
4.3	Influence of ml on BIC Criterion with $ml = 100$	67
4.4	Estimated Change Points by MNSCD for Different Cases	69
4.5	Analysis of Different Channels of EEG Data by MNSCD	71

List of Tables

3.1	Performances of NSCD for Case 1 with Different Baseline Functions and Bandwidths When K Is Known	31
3.2	BIC criterion of NSCD for AR Processes with Comparison to AutoPARM, WBS, BS, and NMCD	32
3.3	Performance of NSCD for ARMA and Invertible MA Processes with Different Baseline Functions and Bandwidths When K Is Known	33
3.4	Performance of BIC Criterion of NSCD for ARMA and Invertible MA Processes with Comparison to AutoPARM, WBS, BS, NMCD	33
3.5	Performance of NSCD for Non-Invertible MA Processes with Different Baseline Functions and Bandwidths When K Is Known	34
3.6	Performance of BIC Criterion of NSCD for Non-Invertible MA Processes with Comparison to AutoPARM, WBS, BS, and NMCD of	35
3.7	Performance of NSCD for $t(4)$ Random Noise	35
3.8	Performance of NSCD with Different Bandwidths While The Sample Size Is Reduced by Half	36
3.9	Performance of BIC Criterion of NSCD with Different Bandwidths While The Sample Size Is Reduced by Half	37
3.10	Performance of NSCD with Searching Unit When K Is Known	44
3.11	Performance of BIC Criterion of NSCD with Searching Unit	44
4.1	Performance of MNSCD for Multivariate AR Processes When K Is Known	60
4.2	BIC criterion of MNSCD with comparison to AutoPARM and ECP for Multivariate AR Processes	60
4.3	Performance of MNSCD for Multivariate ARMA and Invertible MA Processes When K Is Known	61
4.4	BIC of MNSCD with Comparison to AutoPARM and MNSCD for Multivariate ARMA and Invertible MA Processes	61
4.5	Performance of MNSCD for Multivariate Non-Invertible MA Processes When K Is Known	62
4.6	BIC of MNSCD with Comparison to AutoPARM and ECP for Multivariate Non-Invertible MA Processes	62
4.7	MNSCD for Multivariate $t(4)$ White Noise	63
4.8	Performance of MNSCD While The Sample Size Is Reduced by Half	64

List of Symbols

Symbol or abbreviation

$\{X_t\}$

$\{\xi_t\}$

μ

σ^2

$\gamma(t)$

N

$f(u)$

$I_N(u)$

$w(v)$

$W(u)$

m

$\{X(t)\}$

$\xi(t)$

μ

Σ

$\mathbf{R}(t)$

$R_{ij}(t)$

$\mathbf{f}(\lambda)$

$f_{ij}(\lambda)$

$I_{XX}(v)$

K

$\tau_1^0, \dots, \tau_{K+1}^0$

$\hat{\tau}_1, \dots, \hat{\tau}_{K+1}$

$\kappa_1^0, \dots, \kappa_K^0$

$\hat{\kappa}_1^0, \dots, \hat{\kappa}_K^0$

$[x]$

Λ

ml

K_{\max}

Definition

A sequence of univariate time series

A sequence of univariate independent and identically distributed random variables

Expectation of X_t

Expectation of ξ_t

Covariance function at lag t

Sample Size

Spectral density function at frequency u

Periodogram at frequency u

Lag window

Spectral window

Bandwidth of spectral window

A sequence of multivariate time series

A sequence of multivariate independent and identically distributed random variables

Expectation of $X(t)$

Expectation of $\xi(t)$

Covariance matrix of $X(t)$

The (i, j) component of $\mathbf{R}(t)$

Spectral density matrix of $X(t)$

The (i, j) component of $\mathbf{f}(\lambda)$

Periodogram of X_t at frequency v

The true number of change points

True change points

Estimated change points

The proportions of true change points with respect to N

The proportions of estimated change points with respect to N

The largest integer no greater than x

A grid of frequencies

The minimal distance between two adjacent change points

A constant which is larger than K

Chapter 1

Introduction

Time series analysis is a well-developed branch of statistics with a wide range of applications in engineering, economics, biology and so on. In engineering, engineers apply Fourier analysis to the data to estimate the spectrum so that the information about power and energy of frequencies can be acquired, which is called frequency domain approach. For statisticians, some appropriate statistical models will be built with estimated parameters. Then reasonable explanations and predictions can be given, which is called time domain approach. There are some famous time series models, such as ARMA, ARCH-GARCH model (Brockwell and Davis 1991; Fan and Yao 2003). These models share the same assumption that time series is stationary. For example, ARMA assumes that data is second-order stationary, while ARCH-GARCH is strictly stationary.

For stationary time series, a general assumption is that it is a linear process. In frequency domain, properties of linear process, such as periodogram, spectral density function, have been well studied in extensive literatures (Priestly 1981; Barlett 1955; Brillinger 1965; Grenander and Rosenblatt 1966). ARMA model is an example of linear process. Under the assumption of linear process, spectral density function can be easily calculated analytically. Priestly (1981) is a good reference.

However, in real data analysis stationarity is often violated. Some classical methodologies

are proposed to accommodate this situation. ARIMA model assumes that non-stationary time series can be reduced to ARMA after finite differencing. Another way to handle non-stationarity is to divide time series into several segments so that within each segment, stationarity is preserved (Davis et al. 2006; Ombao et al. 2001). However, since the locations of change points, where non-stationary time series are divided, are unknown, we have to estimate them. Here change point detection problem is involved.

There are extensive literatures discussing change point detection (Yao and Au 1988; Zou et al. 2014; Harchaoui and Lévy-Leduc 2010). In general, change point detection focuses on how to find the locations where statistical property jumps from one level to another in a sequence of posteriori time series. In Yao and Au (1988), a change point estimation method is proposed to detect the changes for mean based on least square. Meanwhile, a BIC criterion method is proposed to estimate the unknown number of change points. Zou *et al.* (2014) adopts a nonparametric approach to give estimations when the probability distributions are not the same across the whole time series. Harchaoui and Lévy-Leduc (2010) applies dimension reduction method to estimate the number and locations of change points simultaneously. Csörgö and Horváth (1997) discusses the limit theory in change point detection.

About change point detection in univariate time series from the perspective of time domain, we can assume that each stationary segment can be modeled by appropriate statistical models while the structure varies across different segments. Davis *et al.* (2006) divided non-stationary time series into several different autoregressive processes using Minimal Description Length. Kitagawa and Akaike (1978) detected change points by AIC criterion. In frequency domain, Ombao, Raz, Von Sachs, and Malow (2001) used a family of orthogonal wavelet called SLEX to partition non-stationary process into stationary ones. Since it is not computationally feasible to calculate all the possible combinations, they assume that time series follows a dyadic structure. Lavielle and Ludeña (2000) estimated change points using Whittle log-likelihood when time series is parametric. In Korkas and Fryzlewicz (2017),

Locally Stationary Wavelets was applied to estimate the second-order structure, then Wild Binary Segmentation was imposed to divide the time series into several segments based on CUSUM statistics.

There are many literatures considering the correlation structure of multivariate time series. Lavielle and Teyssière (2006) used Gaussian log-likelihood function to define the contrast function and detect change points when covariance matrices changed. Aue, Hörmann, Horváth, and Reimherr (2009) used binary segmentation to detect the abrupt change in covariance matrix by vectorizing the lower diagonal of sample covariance matrix. Galeano and Peña (2007) applied a likelihood ratio test statistic to detect changes in covariance matrix for large sample, while using a CUSUM statistic for small sample size. Kirch, Mhalsal, and Ombao (2015) first assumed that samples were multivariate autoregressive process, then constructed the test statistics for the existence of change points using innovations for each individual components of multivariate data. Matteson and James (2014) proposed a nonparametric method for detecting changes in distribution for independent multivariate samples. In this thesis, we will develop our method by applying Kullback-Leibler divergence. The plan of this thesis is as follows:

Chapter 2 gives an introduction to basic concepts of time series in frequency domain. Linear process, spectral density function, periodogram, spectral windows, for both univariate and multivariate time series will be introduced. Also, change point model and dynamic programming will be discussed.

Chapter 3 will introduce the existing methods of change point detection for both online and offline univariate time series. For online time series, which also involves quality control, the methods often focus on the residuals (Apley and Shi 1999). For offline time series, some literatures assume that samples follow some particular parametric models. However, parametric method may not be sufficient enough to model the data. So we study linear process, which is more general, and estimate change point using spectral density function. Since using periodogram and spectral window does not assume any statistical models, our

method is nonparametric.

In Chapter 4, we will study change point detection for multivariate time series. Davis *et al.* (2006) extends their AutoPARM method to multivariate time series for detecting change points in EEG data. While they still assume that data follows a multivariate AR process structure, we extend our univariate method to multivariate based on multivariate linear process assumption. In frequency domain study of multivariate time series, spectral density matrix, which is the natural extension of spectral density function, is not real valued any more. However, it is easy to see that it is a positive semi-definite Hermitian matrix, which preserves nonnegative real valued eigenvalues. So, we extend our Kullback-Leibler divergence based method to eigenvalues of spectral density matrix. Some theoretical properties will be derived as well as simulations and real case study.

In Chapter 5, we draw some conclusions and point out some drawbacks of our theories as well as offer some discussions for future works.

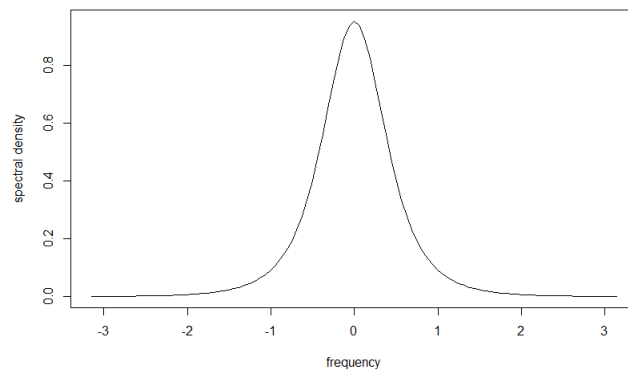
Chapter 2

Spectrums and Linear Process

2.1 Motivation

Spectrum analysis has a wide range of applications. For example, in engineering, engineers need to analyze the power of frequencies in vibration, especially those which could cause resonance. Spectral density function measures the density of energy contributed by some particular frequencies (Priestly 1981). A spectral density function of ARMA(2,1) is shown in Figure 2.1.

Figure 2.1: Spectral Density Function of ARMA(2,1) Process: $X_t - X_{t-1} + 0.25X_{t-2} = \xi_t + 0.8\xi_{t-1}$,



Generally speaking, it may not be easy to get the theoretical density of a given time series.

However, under the second-order stationarity assumption (defined below), the relationship between spectral density function and covariance function can be shown analytically. Let $\{X_t\}$ be a sequence of random variables defined on a common probability space, where $t = 0, \pm 1, \pm 2, \dots$

Definition 2.1. We say $\{X_t\}$ is second-order stationary if

- 1: $E(X_t) = \mu_t = \mu_{t+s}$, for all $s, t = 0, \pm 1, \pm 2, \dots$, and
- 2: $E(X_t - \mu_t)(X_s - \mu_s) = \gamma(t, s) = \gamma(t - s, 0)$, $\forall s, t = 0, \pm 1, \pm 2, \dots$, so the notation is abbreviated and written as $\gamma(t - s, 0) = \gamma(t - s)$.

We may suppress second-order later in the thesis without ambiguity. From definition of stationarity, it is easy to see that covariance function $\gamma(t)$ is non-negative definite in the sense that $\forall (a_1, \dots, a_n) \in \mathbb{C}^n$, $\sum_{i,j=1}^n \bar{a}_i \gamma(i - j) a_j \geq 0$, where \bar{a}_i is the complex conjugate of a_i . Herglotz's theorem (Brockwell and Davis 1991) describes the relationship between covariance function of a stationary time series and its spectrum.

Theorem 2.2 (Herglotz). *A (complex-valued) function $\gamma(\cdot)$ defined on the integers is non-negative definite if and only if*

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\nu} dF(\nu), \text{ for all } h = 0, \pm 1, \dots$$

is non-decreasing, bounded on $[-\pi, \pi]$ and $F(-\pi) = 0$, where F is a right-continuous function on $[-\pi, \pi]$. The function F is called the spectral distribution function of γ and if $F(\lambda) = \int_{-\pi}^{\lambda} f(u) du$, then f is called a spectral density of γ .

Among those stationary time series, linear process is frequently mentioned in the literature. Let us start with the simplest scenario, called white noise.

Definition 2.3. ξ_t is said to be white noise, denoted by $\xi_t \sim WN(\mu, \sigma^2)$ if it is second-order stationary with

- $E\xi_t = \mu$,

- $E(\xi_t - \mu)(\xi_s - \mu) = \begin{cases} \sigma^2 & r = s, \\ 0 & r \neq s. \end{cases}$

We can see from the definition above that white noise is a sequence of uncorrelated random variables. Obviously, a special case of white noise is that $\xi_t \stackrel{\text{i.i.d}}{\sim} (\mu, \sigma^2)$. In this thesis, we refer to this special case when we say that $\{\xi_t\}$ is a sequence of white noise, which is still denoted by $\xi_t \sim \text{WN}(\mu, \sigma^2)$. Now linear process is defined as below (Woodroffe and Van Ness 2003).

Definition 2.4 (Linear process). $\{X_t\}$ is called a univariate linear process if

$$X_t = \sum_{j=-\infty}^{+\infty} a_j \xi_{t-j},$$

$$\{\xi_t\} \sim \text{WN}(0, \sigma^2), \text{ and } \sum_{j=-\infty}^{+\infty} |a_j| < +\infty.$$

The convergence of absolute summation of a_j guarantees that the variance of X_t is finite. Also, $\sum_{j=-\infty}^{+\infty} |\gamma(j)| < +\infty$. Linear process includes some classical time series models, such as Autoregressive process, Moving Average process, ARMA process. Notice that this definition is different from the general definition in some books, such as Priestley (1981) and Fan and Yao (2003), in which linear process is the infinite summation of white noise ξ_t starting from current time t to the past. That is,

$$X_t = \xi_t + \sum_{j=1}^{\infty} a_j \xi_{t-j}.$$

By Herglotz's theorem, we know that there exist spectral distribution functions for linear processes, which can be calculated as follows (Fan and Yao 2003).

Theorem 2.5. Suppose $\gamma(j)$ is the autocovariance function satisfying $\sum_{j=-\infty}^{+\infty} |\gamma(j)| < +\infty$. Then the spectral density function is

$$f(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j) e^{-ij\omega},$$

where σ^2 is the variance of ξ_j .

2.2 Estimation

Now we need a method to estimate the distribution of energy. A classical way is called periodogram, which is defined as follows (Fan and Yao 2003):

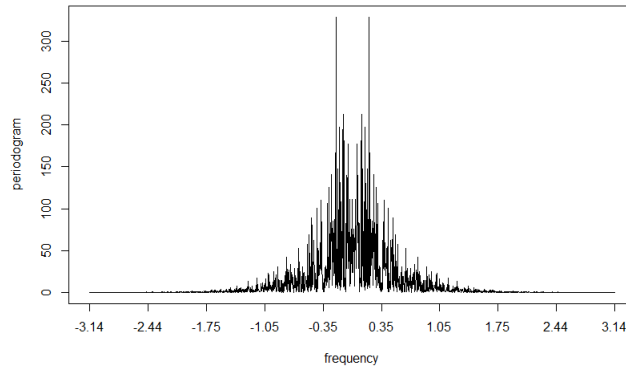
Definition 2.6 (Periodogram). Given a univariate time series $\{X_1, \dots, X_N\}$. The periodogram of the series is defined as

$$I_N(\omega_k) = \frac{1}{N} \left| \sum_{j=1}^N X_j e^{-ij\omega_k} \right|^2, \quad k = -\lfloor \frac{N-1}{2} \rfloor, \dots, -1, 0, 1, \dots, \lfloor \frac{N}{2} \rfloor.$$

where $\omega_k = 2\pi k/N$ is the Fourier frequency, and $\lfloor x \rfloor$ denotes the largest integer which is not greater than number x .

A periodogram of AR(2,1) process is shown in Figure 2.2.

Figure 2.2: Periodogram of an ARMA(2,1) Sample of Size 1800,



From the graph, we can see that periodogram estimation on different frequencies varies dramatically. In fact, based on the next theorem, we will see that periodogram is an inconsistent estimator (Fan and Yao 2003).

Theorem 2.7. For univariate linear process $\{X_1, \dots, X_N\}$, define

$$\varepsilon_{2k-1} = \frac{\sqrt{2}}{\sqrt{N}\sigma} \sum_{j=1}^N \xi_j \cos(\omega_k j), \quad \varepsilon_{2k} = \frac{\sqrt{2}}{\sqrt{N}\sigma} \sum_{j=1}^N \xi_j \sin(\omega_k j).$$

(1) As $N \rightarrow \infty$, $\{\varepsilon_k, k = 1, \dots, 2N\}$ is a sequence of asymptotically independent and standard normal random variables in the sense that for any fixed $c_1, \dots, c_r \in R$ and $r \geq 1$, $\sum_{j=1}^r c_j \varepsilon_{k_j} \xrightarrow{d} N(0, \sum_{j=1}^r c_j^2)$ for any $1 \leq k_1 \leq \dots \leq k_r \leq 2N$.

(2) For $k = 1, \dots, N$,

$$I_N(\omega_k) = 2\pi f(\omega_k) \frac{\varepsilon_{2k-1}^2 + \varepsilon_{2k}^2}{2} + R_N(\omega_k),$$

where f is the spectral density of $\{X_j\}$ and $\max_{1 \leq k \leq N} E|R_N(\omega_k)| \rightarrow 0$ as $N \rightarrow \infty$.

As it has been shown above, periodogram is asymptotically unbiased, but the variance does not tend to zero as N tends to infinity. To overcome this drawback, tapering is often applied on the periodogram to estimate the spectral density function. Let $w(u)$, which is called lag window, satisfy the following conditions (Fan and Yao 2003)

- $w(0) = 1$,
- $|w(u)| \leq 1, |u| \leq 1$,
- $w(u) = 0, |u| > 1$.

Then $\hat{f}(\omega)$, which is the estimate of spectral density function, can be achieved as follows

$$\hat{f}(\omega) = \sum_{k=-m}^m w(|k|/m) \hat{\gamma}(k) e^{-ik\omega},$$

where $\hat{\gamma}$ is the sample covariance function, m is the bandwidth which determines where to truncate the covariance function, so $m \leq N$. Set $W(u) = \frac{1}{2\pi} \int_{-\pi}^{\pi} w(v) e^{-iuv} dv$, which is called spectral window, then by Fourier transformation theory,

$$\hat{f}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\omega - u) I_N(u) du.$$

After smoothing, \hat{f} is still asymptotically unbiased while its variance tends to zero, as N goes to infinity, which is stated as follows (Brockwell and Davis 1991):

Theorem 2.8. Let $\{X_t\}$ be a linear process satisfying $\sum_{j=-\infty}^{+\infty} |a_j|j^{1/2} < +\infty$, and $E\xi_j^4 < +\infty$.

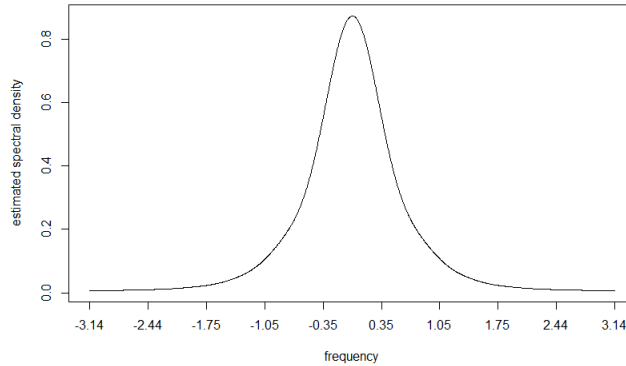
Then for \hat{f} , we have

$$(1) \lim_{N \rightarrow +\infty} E\hat{f}(\omega) = f(\omega)$$

$$(2) \lim_{N \rightarrow +\infty} \left(\sum_{|j| \leq m} W(j/m)^2 \right)^{-1} \text{Cov}(\hat{f}(\omega_1), \hat{f}(\omega_2)) = \begin{cases} 2f^2(\omega_1) & \omega_1 = \omega_2 = 0 \text{ or } \pi \\ f^2(\omega_1) & 0 < \omega_1 = \omega_2 < \pi \\ 0 & \omega_1 \neq \omega_2 \end{cases}$$

A smoothed spectral density estimation of AR(2) after smoothing the periodogram in Figure 2.2 is shown in Figure 2.3.

Figure 2.3: Smoothed Spectral Density Function of ARMA(2,1),



2.3 Multivariate Time Series, Spectral Density Matrix and Its Estimation

In the previous sections, we have discussed the second-order stationarity for univariate time series. In applications, some time series data are multivariate. In engineering, we may want to study the variation of pressure, temperature, and other indices over time. Similar to the univariate scenario,

Definition 2.9 (Multivariate Second-Order Stationary Time Series). Let $\{X(t)\}$ be a sequence of $p_x \times 1$ multivariate random vectors defined on a common probability space, where

$t = 0, \pm 1, \pm 2, \dots$. $\{X(t)\}$ is a second-order multivariate stationary process if,

- $EX(t) = \mu(t) = \mu(t + s), \forall t, s = 0, \pm 1, \pm 2, \dots$.
- $E(X(t) - \mu(t))(X(s) - \mu(s))^T$ depends on $|s - t|$ only, for s and $t = 0, \pm 1, \pm 2, \dots$. We denote this by $\mathbf{R}(s - t)$, and its (i, j) entry is denoted by $R_{ij}(s - t)$,

where T denotes the transpose.

Like univariate time series, a general class of stationary time series is the multivariate linear process.

Definition 2.10 (Multivariate White Noise). A sequence of random vectors $\{\boldsymbol{\xi}(t)\}$ is said to be white noise denoted by $\boldsymbol{\xi}(t) \sim WN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if

- $E\boldsymbol{\xi}(t) = \boldsymbol{\mu}$,
- $E(\boldsymbol{\xi}(t) - \boldsymbol{\mu})(\boldsymbol{\xi}(s) - \boldsymbol{\mu})^T = \begin{cases} \boldsymbol{\Sigma}, & s = t, \\ \mathbf{0} & \text{elsewhere.} \end{cases}$

Definition 2.11 (Multivariate Linear Process). $X(t)$ is a $p_x \times 1$ multivariate linear process, if

$$X(t) = \sum_{j=-\infty}^{+\infty} G(j)\boldsymbol{\xi}(j),$$

$\boldsymbol{\xi}(j)$ is a sequence of $p_\xi \times 1$ random vectors, and $\boldsymbol{\xi}(j) \sim WN(0, \boldsymbol{\Sigma})$, $G(j)$ is a sequence of $p_x \times p_\xi$ matrices satisfying $\sum_{j=-\infty}^{+\infty} G(j)G(j)^T < +\infty$.

For multivariate time series, we are more interested in the correlations between different dimensions. So the cross spectrum is defined as follows.

Definition 2.12. Suppose $\sum_{t=-\infty}^{+\infty} |R_{ij}(t)| < +\infty$, then the cross spectrum, or cross spectral density, of $\{X_{ti}\}$ and $\{X_{tj}\}$, is $f_{ij}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} e^{-ih\lambda} R_{ij}(h)$, $\lambda \in [-\pi, \pi]$. In matrix notation,

$$\mathbf{f}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} e^{-ih\lambda} \mathbf{R}(h) = \begin{pmatrix} f_{11}(\lambda) & \cdots & f_{1p_x}(\lambda) \\ \vdots & \ddots & \vdots \\ f_{p_x 1}(\lambda) & \cdots & f_{p_x p_x}(\lambda) \end{pmatrix}.$$

From the definition, we can see that f_{jj} 's are the spectral density functions of each component of $X(t)$, so they are real valued functions. Generally speaking, since $R_{ij}(h) \neq R_{ij}(-h)$, f_{ij} are complex valued functions, if $i \neq j$. Similar to the correlation coefficient, which would measure the linear dependence between random variables, coherency is introduced to measure the dependency between two time series in the frequency domain. The definition is as follows.

Definition 2.13 (Coherency). For multivariate stationary process \mathbf{X}_t and its spectral density matrix $\mathbf{f}(\lambda)$, $-\pi \leq \lambda \leq \pi$, the coherency between X_{ti} and X_{tj} is

$$\text{coh}(\lambda) = \frac{|f_{ij}(\lambda)|}{(f_{ii}(\lambda)f_{jj}(\lambda))^{1/2}}, \quad -\pi \leq \lambda \leq \pi.$$

Here $|f_{ij}(\lambda)|$ is the modulus of $f_{ij}(\lambda)$. The coherency remains invariant under linear transformations of X_{ti} and X_{tj} , so, roughly speaking, the coherency is a correlation coefficient in the frequency domain (Priestley 1981).

Similar to univariate time series, if we want to estimate spectral density matrix, first we will define a multivariate periodogram, then smooth its every entry with a selected spectral window.

Definition 2.14 (Multivariate Periodogram). Suppose we have a sequence of random vectors $\{X(t)\}_{t=1}^N$. Then I_{XX} is called multivariate periodogram, if

$$I_{XX}(\omega_k) = \frac{1}{N} \left(\sum_{j=1}^N X(j)e^{-ij\omega_k} \right) \left(\sum_{j=1}^N X(j)e^{-ij\omega_k} \right)^*, \quad k = -[\frac{N-1}{2}], \dots, -1, 0, 1, \dots, [\frac{N}{2}],$$

where $\omega_k = 2\pi k/N$ is the Fourier frequency, $[x]$ denotes the largest integer which is not greater than number x , and $*$ denotes the conjugate transpose.

For (i, j) entry of I_{XX} , we still use the spectral window $w(u)$ with its Fourier transformation $W(u)$, to smooth each entry of multivariate periodogram. Denote $\hat{\mathbf{f}}$ as the estimated spectral density matrix, then $\hat{f}_{ij}(\omega) = \frac{1}{2\pi} \int W(\omega - u) I_{XX, i, j}(u) du$, where $I_{XX, i, j}$ is the (i, j) entry of I_{XX} .

2.4 Kullback-Leibler Divergence

In Kullback and Leibler (1951), a divergence function was introduced to measure the discrimination information between two distribution functions. The original definition is as follows. Suppose there are two probability distributions, $Z_1(x)$ and $Z_2(x)$. $z_1(x)$ and $z_2(x)$ denote the probability density functions of Z_1 and Z_2 , respectively. Then Kullback-Leibler (K-L) divergence from Z_2 to Z_1 is

$$D_{\text{KL}}(Z_1\|Z_2) = \int z_1(x) \log \left(\frac{z_1(x)}{z_2(x)} \right) dx.$$

K-L divergence is also called relative entropy of Z_1 with respect to Z_2 . In information theory, K-L divergence defines the information gain if Z_2 is used instead of Z_1 . K-L divergence preserves several good properties, such as non-negativity, additivity. As we can see, K-L divergence is defined for probability density function, which is non-negative. Since spectral density function is also non-negative, we will generalize K-L divergence so that it is applicable to non-negative functions, which is still called Kullback-Leibler divergence in this thesis and will be applied later in the forthcoming chapters. The definition is as follows.

Definition 2.15 (Kullback-Leibler Divergence). Suppose we have two non-negative functions, f_1 and f_2 , defined on a common support. Then the Kullback-Leibler divergence of $f_1(x)$ with respect to $f_2(x)$ is

$$D_{\text{KL}}(f_1\|f_2) = \int f_1(x) \log \frac{st(f_1)}{st(f_2)} dx,$$

where $st(f_1) = \frac{f_1(x)}{F_1}$, $st(f_2) = \frac{f_2(x)}{F_2}$, and $F_1 = \int f_1(x)dx$, $F_2 = \int f_2(x)dx$.

By Gibbs's inequality, this new K-L divergence is still non-negative and is equal to 0 if and only if $f_1 = c * f_2$ almost everywhere, where c is an appropriate constant.

2.5 Change Point Model and Algorithm

The change point model can be applied to describe a sequence of data which can be divided into several pieces. In each segment, the samples come from a statistical model while different

segments have different models. For example, the mean, variance, or distribution function could be different across all segments. The boundaries of those segments are called change points, which should be estimated. To detect the change points, an appropriate objective function will be given. Let us take a simple example to illustrate the main idea.

Suppose a sequence of observations $\{Y_t\}_{t=1}^n$ are independent random variables following normal distribution. A sequence of integers $\{\tau_0^0, \tau_1^0, \dots, \tau_K^0, \tau_{K+1}^0\}$ with convention $\tau_0^0 = 0$ and $\tau_{K+1}^0 = n$ satisfy $Y_j \stackrel{\text{i.i.d}}{\sim} N(\mu_i, \sigma^2)$, $\tau_{i-1}^0 < j \leq \tau_i^0$, $i = 1, \dots, K+1$ while $\mu_i \neq \mu_{i+1}$, $\forall i$, that is, the mean will change across different segments with common variance for each observation. To further simplify our model, we assume that K , the number of change points, is known. To estimate τ_1, \dots, τ_K , a natural approach is via maximum likelihood. After omitting some constants, the log likelihood function, which is also the objective function, is

$$R(\tau_1, \dots, \tau_K, \mu_1, \dots, \mu_{K+1}) = \sum_{i=1}^{K+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} (Y_j - \mu_i)^2.$$

The maximizers of the objective function are the estimators of those parameters. Directly maximizing the objective function is difficult due to the numerous combinations of τ_1, \dots, τ_K . If we take a close look at R , it can be seen that maximizing R can be divided into two steps. First, we fix τ_1, \dots, τ_K , and estimate μ_1, \dots, μ_{K+1} . The estimators are denoted by $\hat{\mu}_1, \dots, \hat{\mu}_{K+1}$. Notice that the means remain unchanged within $Y_{\tau_{i-1}}, \dots, Y_{\tau_i}$, $i = 1, \dots, K+1$. So it is easy to have $\hat{\mu}_i = \frac{1}{\tau_i - \tau_{i-1}} \sum_{j=\tau_{i-1}+1}^{\tau_i} Y_j$, $i = 1, \dots, K+1$. Set $Q(\tau_{i-1}, \tau_i) = \sum_{j=\tau_{i-1}+1}^{\tau_i} (Y_j - \hat{\mu}_i)^2$, then the maximization can be simplified as follows:

$$\max_{\tau_1, \dots, \tau_K} \max_{\mu_1, \dots, \mu_{K+1}} R(\tau_1, \dots, \tau_K, \mu_1, \dots, \mu_{K+1}) = \max_{\tau_1, \dots, \tau_K} \sum_{i=1}^{K+1} Q(\tau_{i-1}, \tau_i) = \max_{\tau_1, \dots, \tau_K} R(\tau_1, \dots, \tau_K).$$

We can see that $R(\tau_1, \dots, \tau_K)$ is separable, so the change points can be estimated recursively. That is, τ_K is estimated first, denoted by $\hat{\tau}_K$. Then $\tau_1, \dots, \tau_{K-1}$ are estimated based on observations $Y_1, \dots, Y_{\hat{\tau}_K}$. This separability is called “the principle of optimality” (Bellman and Dreyfus 1962, see also Hawkins 2001). That is, an optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision (Bellman and

Dreyfus 1962). So the dynamic programming algorithm can be applied here, which will solve an optimization problem by dividing it into several sub-problems in a recursive way. Specifically, the dynamic programming algorithm for the change point detection problem shown above can be written as follows (Hawkins 2001):

- 1: Set $F(r, m)$ to be the maximized log likelihood function for observations Y_1, \dots, Y_m . What is more, it is assumed that there are $r - 1$ change points in this segment. So it is easy to see that $F(K + 1, n) = \max_{\tau_1, \dots, \tau_K} R(\tau_1, \dots, \tau_K)$.
- 2: $F(1, m) = Q(1, m)$, for $m = 1, \dots, n$.
- 3: For each $r = 2, \dots, K + 1$, $m = 1, \dots, n$, calculate $F(r, m) = \max_h F(r - 1, h) + Q(h + 1, m)$.
- 4: Starting with $\hat{\tau}_{K+1} = n$, $\hat{\tau}_i = \arg \max_h F(i, h) + Q(h + 1, \hat{\tau}_{i+1})$, $i = K, K - 1, \dots, 1$.

One drawback of dynamic algorithm is that the computation complexity is $O(Kn^2)$. Some articles proposed new method so that the complexity can be reduced to $O(Kn)$. Check Killick *et al.* (2012) for further details.

2.6 Non-Stationary Time Series with Change Points

The non-stationary time series model involved in this thesis is as follows. Let $\{X_t\}_{t=1}^n$ be a sequence of observations. For a sequence of integers $0 = \tau_0^0 \leq \tau_1^0 \leq \dots \tau_K^0 \leq \tau_{K+1}^0$, we assume

$$X_t = \sum_{j=-\infty}^{\infty} a_j(i) \xi_{t-j}, \text{ for } j = \tau_{i-1}^0 + 1, \dots, \tau_i^0, i = 1, \dots, K + 1,$$

where $\{\xi_t\}_{t=-\infty}^{+\infty}$ is a set of independent and identically distributed random variables with mean 0 and variance σ^2 . In each segment, $\{X_t, \tau_{i-1}^0 + 1 \leq t \leq \tau_i^0\}$ is stationary with spectral density function $f_i(v)$, $v \in [-\pi, \pi]$. Furthermore, we assume that within each segment, the stationary time series is a linear process. To be more specific, $X_t = \sum_{j=-\infty}^{\infty} a_j(i) \xi_{t-j}$,

$i = 1, \dots, K + 1$, and $\sum_{j=-\infty}^{\infty} |a_j(i)| < \infty$ so that the spectral density function exists. Here we assume that only $\{a_j(i)\}$ change to $\{a_j(i + 1)\}$, while $\{\xi_j\}$, which is the basis of every segment, remains the same. For example, by our setting, $X_{\tau_i^0} = \sum_{j=-\infty}^{\infty} a_j(i)\xi_{\tau_i^0-j}$, so the next observation is $X_{\tau_i^0+1} = \sum_{j=-\infty}^{\infty} a_j(i)\xi_{\tau_i^0+1-j}$. Since τ_i^0 is a change point, $X_{\tau_i^0+1} = \sum_{j=-\infty}^{\infty} a_j(i + 1)\xi_{\tau_i^0+1-j}$. From the above settings, we can see that a common statistical structure exists within each segment (spectral density function) while it varies across different segments. So we can adopt change point detection model here with well-defined objective functions, which is the main subject of the forthcoming chapters.

2.7 Some Preliminary Simulations

From Theorem 2.7, we can see that periodogram is an inconsistent estimator of spectral density function. We have to smooth it using spectral window. In this section, some simulation results will be shown to demonstrate the performances of smoothed spectral density function under the influence of two parameters: sample size N and the bandwidth of spectral window m .

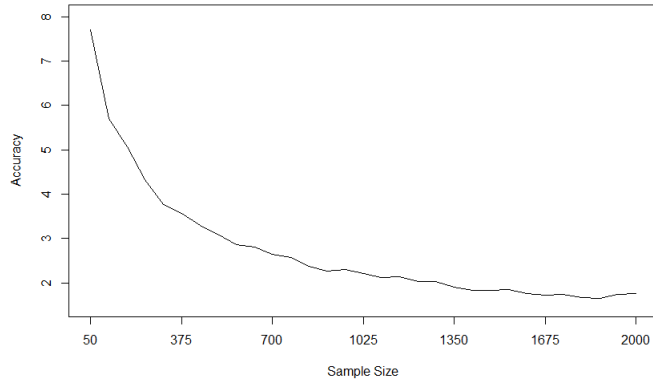
The first example is $X_t - X_{t-1} + 0.25X_{t-2} = \xi_t + 0.8\xi_{t-1}$, $\xi_t \stackrel{i.i.d}{\sim} N(0, 1)$. Here the sample size ranges from 50 to 2000. We fix the values of frequencies other than letting it change with N . So, for each randomly generated time series, Fourier transformation is applied first, then we estimate the values of spectral density function at those pre-fixed frequencies. Here we set those frequencies by dividing $[-\pi, \pi]$ equally into $N_v = 1800$ pieces. What is more, $m = N_v^c$ where c is a positive constant and we let c change from $1/5$ to $1/2$. We use $\max_{-\pi \leq v \leq \pi} |\hat{f}(v) - f(v)|$ to measure the accuracy of estimation. All simulations are obtained with 1000 replications.

Figure 2.4a to Figure 2.4c show the performances when N changes from 50 to 2000, with $c = 1/2, 1/3, 1/4$ under each circumstance. We can see that the bias will decrease as N increases. In Figure 2.5a to 2.5d, N is fixed at 245, 700, 1350, 2000, respectively, while m

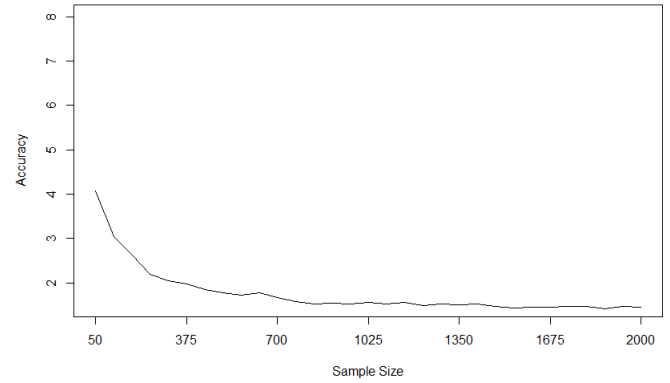
ranges from $N_v^{1/5}$ to $N_v^{1/2}$. As it is shown, the bias reaches its minimum at different locations for different N , which means that m should be customized.

Figure 2.4: Estimation Accuracy of Spectral Density Function of ARMA(2,1) with Different Bandwidths

(a) Bandwidth $m = N_v^{1/2}$



(b) Bandwidth $m = N_v^{1/3}$



(c) Bandwidth $m = N_v^{1/4}$

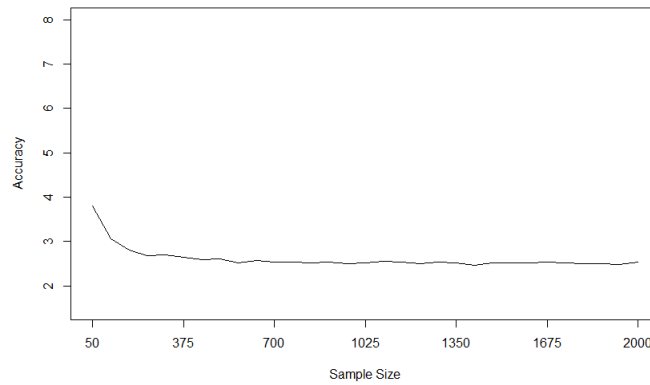
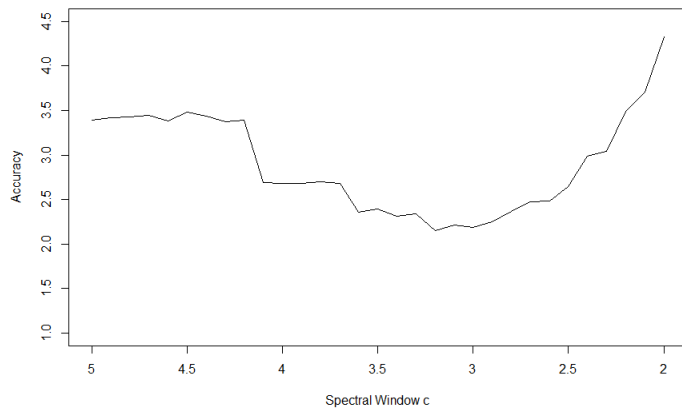
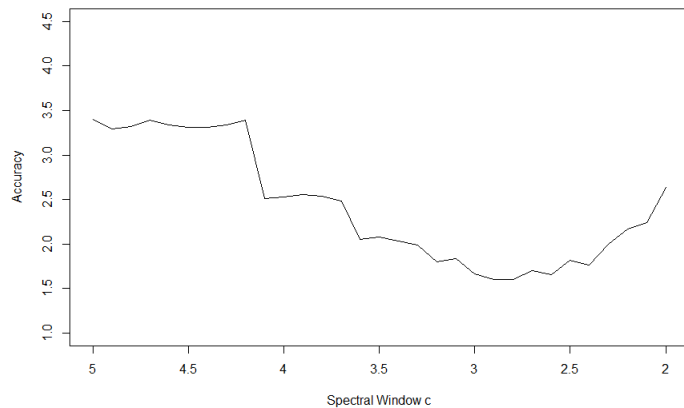


Figure 2.5: Estimation Accuracy of Spectral Density Function of ARMA(2,1) with Different Sample Sizes

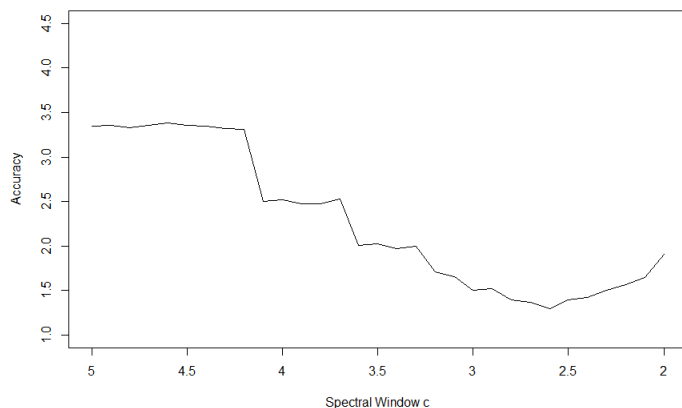
(a) Sample Size $N = 245$



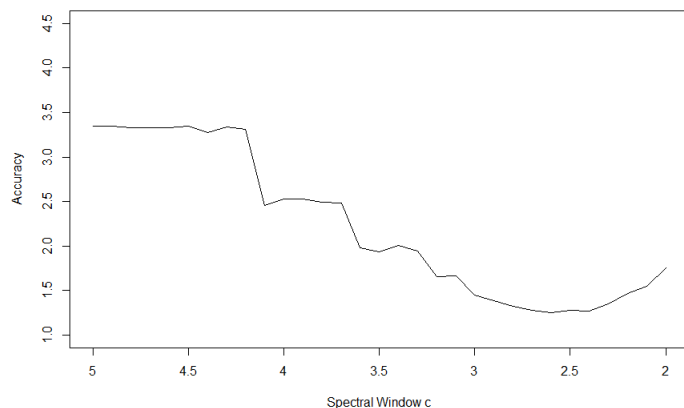
(b) Sample Size $N = 700$



(c) Sample Size $N = 1350$



(d) Sample Size $N = 2000$



Similar to the univariate circumstance, we generate multivariate ARMA(1,1):

$$X(t) - \Phi X(t-1) = \boldsymbol{\xi}_t + \Theta \boldsymbol{\xi}_{t-1},$$

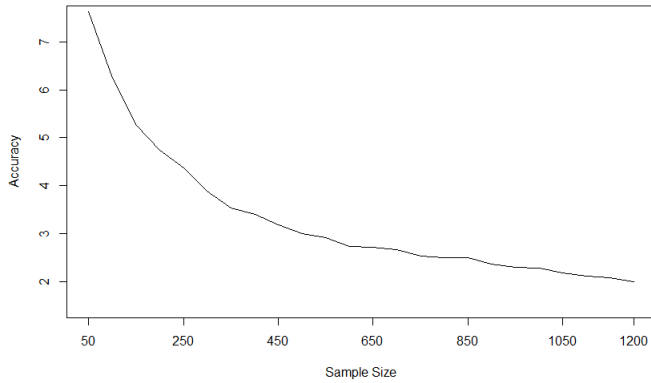
where $\Phi = \begin{pmatrix} 0.5 & 0.5 \\ 0 & 0.5 \end{pmatrix}$, $\Theta = \Phi^T$, and $\boldsymbol{\xi}_t \stackrel{i.i.d.}{\sim} N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$. Here we fix the frequencies by dividing $[-\pi, \pi]$ equally into $N_v = 1200$ pieces. N ranges from 50 to 1200, and $m = N_v^c$ with c changing from $1/5$ to $1/2$. What is different from univariate scenario

is that we estimate the largest eigenvalue λ_{\max} of spectral density matrix, and measure the performance by $\max_{-\pi \leq v \leq \pi} |\hat{\lambda}_{\max}(v) - \lambda_{\max}(v)|$.

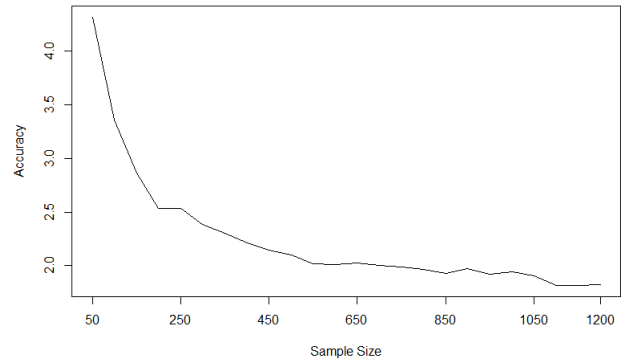
Figure 2.6a to Figure 2.6c show the results when $c = 1/2, 1/3, 1/4$ while N is changing. In Figure 2.7a to 2.7d, N is fixed at 200, 700, 900, 1200, respectively, and let c range from $1/5$ to $1/2$. We can still observe the similar performances. As N increases, the bias will shrink to zero, and the best estimation is achieved at different m for different sample size.

Figure 2.6: Estimation Accuracy of the Maximum Eigenvalue of Spectral Density Matrix of Multivariate ARMA(1,1) with Different Bandwidths

(a) Bandwidth $m = N_v^{1/2}$



(b) Bandwidth $m = N_v^{1/3}$



(c) Bandwidth $m = N_v^{1/4}$

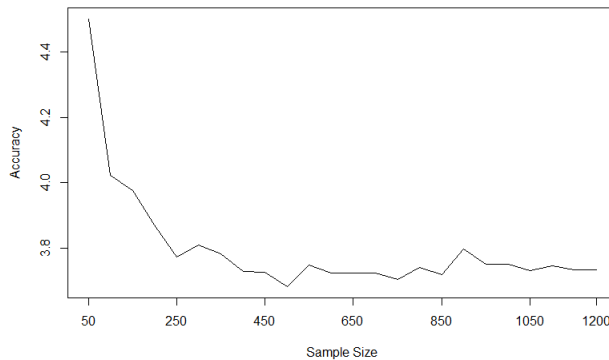
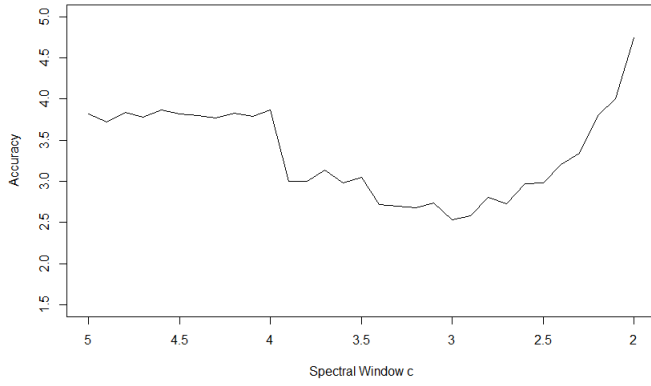
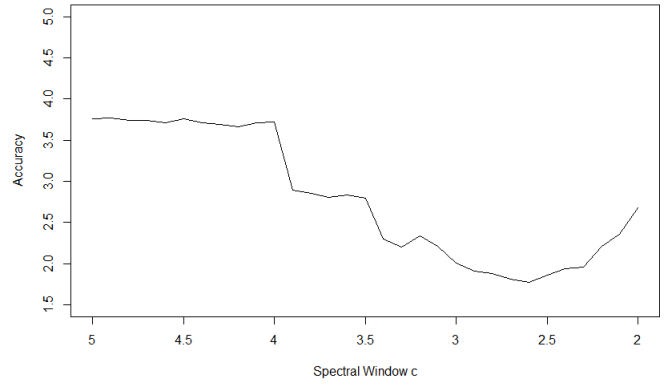


Figure 2.7: Estimation Accuracy of the Maximum Eigenvalue of Spectral Density Matrix of Multivariate ARMA(1,1) with Different Sample Sizes

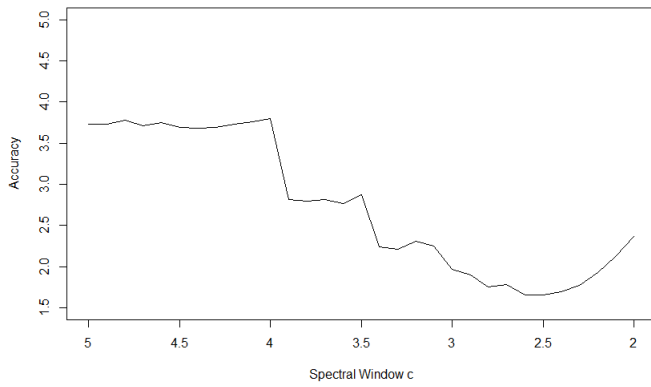
(a) Sample Size $N = 200$



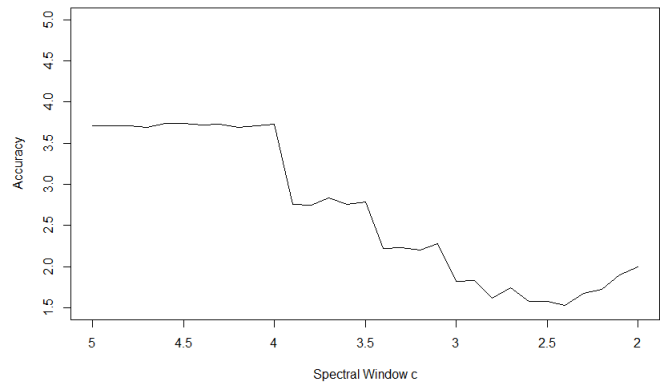
(b) Sample Size $N = 700$



(c) Sample Size $N = 900$



(d) Sample Size $N = 1200$



Chapter 3

Univariate Nonparametric Change Point Detection for Linear Process

In this chapter, we study univariate non-stationary time series. Consider a sequence of data $\{X_1, \dots, X_N\}$ and let $\tau_0^0, \tau_1^0, \tau_2^0, \dots, \tau_K^0, \tau_{K+1}^0$ be nonnegative integers satisfying $0 = \tau_0^0 < \tau_1^0 < \tau_2^0 < \dots < \tau_K^0 < \tau_{K+1}^0 = N$. We assume that within the j th segment of data, i.e., $\tau_{j-1}^0 + 1 \leq t \leq \tau_j^0$, $1 \leq j \leq K + 1$, X_t is stationary. τ_j^0 are called structural breaks, or change points, which are unknown. K is the true number of change points.

When autocovariance function is absolutely summable, it is well known that stationary process has spectral density function, which is the Fourier transformation applied to autocovariance function (Brockwell and Davis 1991). A good property that spectral density function preserves is that after being properly normalized, it becomes a well-defined probability density function, although it is not defined for any particular random variable (Priestley 1981). In probability theory, there are some existing functions to measure the difference between probability density functions. Kullback-Leibler divergence (K-L divergence) is one of them. In this thesis, we apply K-L divergence to normalized spectral density functions to define our objective function. Then we estimate change points by maximizing the objective function. That is, we find the locations where discrepancy between different

spectral density functions reaches its maximum. When estimating spectral density function, we adopt the classical method, that is, we first calculate periodogram then smooth it by choosing appropriate spectral window. Although we assume that each stationary segment is a linear process, which is a class of stationary time series more general than autoregressive process and ARMA model, we calculate spectral density function without estimating any parameters. So our change point detection method is nonparametric, we call it nonparametric spectral change-point detection (NSCD). In application, we do not know the number of change points, so a BIC criterion similar to Yao (1988) and Zou *et al.* (2014) is proposed. The consistency of both change point estimation and BIC criterion can be established under some conditions. The theoretical result of change point estimation is quite similar to the one shown in Davis *et al.* (2006), but slightly weaker. Dynamic programming algorithm (Hawkins 2001) is used, but due to its computational complexity, we adopt the screening algorithm (Zou *et al.* 2014). Also, Pruned Exact Linear Time (PELT) by Killick *et al.* (2012) can also boost the speed of algorithm when estimating the number and locations of change points simultaneously.

The rest of this chapter is organized as follows. In Section 3.1, we describe our objective functions. Asymptotic properties are presented in Section 3.2. Implementation and further details of our algorithm will be given in Section 3.3. In Section 3.4, numerical simulations as well as comparison with Davis *et al.* (2006) are shown. The analysis of EEG data are presented in Section 3.5.

3.1 Model and Methodology

Consider non-stationary time series $X_t = \sum_{j=-\infty}^{+\infty} a_j(k)\xi_{t-j}$, $\xi_j \stackrel{i.i.d.}{\sim} (0, \sigma^2)$, $\tau_{k-1}^0 < t \leq \tau_k^0$, $k = 1, \dots, K + 1$ with $\tau_0^0 = 0$, $\tau_{K+1}^0 = N$, $\tau_k^0 - \tau_{k-1}^0 = N_k$. Here we assume that within the k th segment, ξ_t are independent and identically distributed, and all ξ_t are all independent. $\{a_j(k)\}$ satisfy $\sum_j |a_j(k)| < +\infty$, $\forall k$, so that when $\tau_{k-1}^0 < t \leq \tau_k^0$, spectral density function

exists, denoted by f_k (Brockwell and Davis 1991). If $EX_t \neq 0$, we can subtract mean from observations by $X_t - \mu$. Otherwise, we assume that $EX_t = 0$ for all t . Our goal is to detect τ_k when f_k and f_{k+1} are different. If we could find a spectral density function f that it is different from all f_k , then a discriminant function can be applied so that it reaches its maximum at τ_k . Here we adopt Kullback-Leibler divergence and define our objective function as below:

$$R(\tau_1, \dots, \tau_{K+1}) = \sum_{k=1}^{K+1} (\tau_k - \tau_{k-1}) \int_{-\pi}^{\pi} \hat{f}_k(u) \log \frac{st(\hat{f}_k)}{st(f)} du.$$

Here $\hat{F}_k = \int_{-\pi}^{\pi} \hat{f}_k(u) du$, $F = \int_{-\pi}^{\pi} f(u) du$ so that $st(\hat{f}_k)$ and $st(f)$ become probability density functions, and $0 < \tau_1 < \tau_2 < \dots < \tau_K < N$ is a possible partition of the time series. There are two ways to find f . We can give an estimation of the spectral density function \hat{f}_0 , based on the whole time series. Although the whole time series is non-stationary, the estimated spectral density function still converges to a well-defined function, which is the weighted sum of f_k (See Lemma A.2 for details). If we know that every part of data is white noise, then $st(f) = \frac{1}{2\pi}$.

There are literatures concerning the application of K-L divergence in time series. Parzen (1982) used it to estimate the parameters of autoregressive process. Parzen (1983) extended it to estimate ARMA process. Shore (1981) calculated the minimum K-L divergence to estimate spectrum given its priori spectrum has an exponential form. See Rao (1993) for more reviews. From information point of view, K-L divergence measures the information gain when a new probability density function is used instead of the old one. So our objective function will detect the locations where we maximize the information gain when a new spectral density function f comes in.

To estimate the spectral density function, the classical methodology is adopted here. First, periodogram will be calculated as follows:

$$I_{\tau_{k-1}+1, \tau_k}(\lambda) = \frac{1}{\tau_k - \tau_{k-1}} \left| \sum_{j=\tau_{k-1}+1}^{\tau_k} X_j e^{-ij\lambda} \right|^2$$

Since periodogram is not consistent, we choose a spectral window, $W(u)$, to smooth periodogram, then

$$\hat{f}_k(\lambda) = m \int_{-\pi}^{\pi} W(m(\lambda - u)) I_{\tau_{k-1}+1, \tau_k}(u) du, \lambda \in [-\pi, \pi].$$

There are several choices for $W(u)$ (Priestley 1981). Since we want the normalized \hat{f}_k to be a probability density function, i.e., $\hat{f}_k \geq 0$, Bartlett kernel is chosen, which is $mW(mu) = \sin^2(mu/2)/(2\pi m \sin^2(u/2))$, where m is the bandwidth. One usually applies Fourier transformation to achieve an estimator on a grid of frequencies, denoted by $\Lambda = \{\lambda_1, \dots, \lambda_{N_\lambda}\}$, where N_λ is the cardinality of Λ , which could tend to infinity as N goes to infinity. Here we still use integral to denote the summation of estimators. We may set $N_\lambda = N$ so N_λ could tend to infinity, or Λ could be Nyquist frequency. Based on this grid of frequency, K-L divergence applied to normalized spectral density function is still non-negative, and equals to zero if and only if two normalized spectral density functions are equal on $[-\pi, \pi]$.

When estimating change points, we have no idea about the number of change points in the data, so K should be estimated. Yao and Au (1989), Zou *et al.* (2014) gave BIC criterion and showed its consistency. Here we also propose a BIC criterion as follows:

$$BIC_L = -\max_{\tau'_1, \dots, \tau'_L} R(\tau_1, \dots, \tau_L) + LC_N,$$

where L is the number of change points. Here C_N is an appropriate constant which will be illustrated later. So our estimator \hat{K} is chosen by minimizing the criterion above with respect to L .

3.2 Asymptotic Theory

From the assumptions, we can see that K is a constant, and τ_i^0 , $i = 1, 2, \dots, K$, changes with N . So there is a sequence of constants, $0 < \kappa_1^0 < \kappa_2^0 < \dots < \kappa_K^0 < N$, such that $\{X_t, t = 1, 2, \dots, 1\}$ is a realization of non-stationary time series in Section 3.1 with $\tau_k^0 = [\kappa_k^0 N]$, $k = 1, 2, \dots, K$, where $[x]$ denotes the largest integer which is not greater than x .

Their estimators are denoted by $\hat{\kappa}_k$, $k = 1, \dots, K$. We can estimate τ_k^0 first, denoted by $\hat{\tau}_k$, by maximizing the objective function $R(\tau_1, \dots, \tau_K)$, then $\hat{\kappa}_k = \hat{\tau}_k/N$. In literature, Yao and Au (1989) achieved a consistent estimation of $O_p(1)$ for $\hat{\tau}_k$, which means that the difference between estimators and true change points is no bigger than a constant. Zou *et al.* (2014) also drew the same conclusion when the number of change points was constant. In Davis *et al.* (2006), the consistency was attained in the sense that $|\hat{\kappa}_k - \kappa_k| < \epsilon$ in probability 1, where ϵ is some constant. The reason that estimators cannot converge as fast as those in Yao and Au (1989) and Zou *et al.* (2014) is that estimating spectral density functions needs a sufficient number of samples. When estimating change points, we always find the next change point which is ml away from the previous one. For example, after giving $\hat{\tau}_j$, we look for the next change point starting from $\hat{\tau}_j + ml$. Our results are similar to Davis *et al.* (2006) and based on a set of discrete frequencies Λ . The following assumptions are needed to obtain the consistency:

A1: $E\xi_t^{8q} < \infty$ where q is some integer satisfying $q \geq 3$, $\{a_j(k)\}$ converge absolutely, $\forall k$.

A2: $W(v)$ is a non-negative, even, bounded, integrable function, $\int_{-\pi}^{\pi} W(v)dv = 1$,
 $\int_{-\pi}^{\pi} (W(v))^{1-\frac{1}{2q}} dv < \infty$.

A3: $N_{\min}/N \rightarrow c_{\min} > 0$, as $N \rightarrow \infty$, and $N_{\min} > m$, where $N_{\min} = \min_{1 \leq k \leq K} (\tau_k^0 - \tau_{k-1}^0)$.

A4: $\forall k$, f_k is everywhere positive and satisfies uniform Lipschitz condition:

$$|f_k(u_1) - f_k(u_2)| \leq B_{f_k}|u_1 - u_2|$$

$\forall u_1 \in [-\pi, \pi]$, $u_2 \in [-\pi, \pi]$, where B_{f_k} is a constant.

A5: $m = O(N^\alpha)$, where $\frac{1}{4} \leq \alpha < \alpha + \frac{3}{2q} < \frac{1}{2}$.

A6: $w(0) = 1$, and $w(v)$ has continuous derivatives to the order of $2q$.

A7: $\{f_k(u)\}$ are linearly independent for $u \in [-\pi, \pi]$.

A8: $ml = c_{ml}N$, where $c_{ml} > 0$. $ml < N_{\min}$

In A8, ml is the minimal length of time series when estimating spectral density functions, since a sufficient number of observations is always necessary, especially when the convergence rate of estimators is slow. Also, $ml < N_{\min}$ so that all change points are distinguishable. Assumption 7 is given to guarantee that any linear combination is not equal to f_j , $\forall j = 1, \dots, K$. Assumption 1-6 are similar to those in Woodroffe and Van Ness (1967) so that the $\max_{\lambda_j \in \Lambda} \frac{\hat{f}_k(\lambda_j)}{f_k(\lambda_j)}$ can be bounded in probability. Assumption 4 can be stronger so that in Assumption 5, α can be less than $\frac{1}{4}$ (see Woodroffe and Van Ness 1967 for further details). Theorem 3.1 gives the consistency of change point estimation.

Theorem 3.1. *When K is known, $\hat{\kappa}_j - \kappa_j \xrightarrow{p} 0$, $\forall j = 1, \dots, K$.*

To estimate the number of change points, a pre-specified upper bound K_{\max} satisfying $K < K_{\max}$ will be given. Then BIC values for each $1 \leq L \leq K_{\max}$ are calculated and \hat{K} will be the number where BIC reaches its minimum. The following theorem establishes the consistency of estimation of BIC criterion.

Theorem 3.2. *Suppose we choose K_{\max} such that $K \leq K_{\max}$. If $C_N/N \rightarrow 0$, $C_N/N^{\frac{q+3+2q\alpha}{2q}} \rightarrow \infty$, we have $P(\hat{K} = K) \rightarrow 1$.*

Proofs of Theorem 3.1 and 3.2 are given in Appendix A.

3.3 Algorithm

Similar to Zou *et al.* (2014), our objective function is separable. For change point detection, a commonly adopted algorithm, dynamic programming (Hawkins 2001), can be applied. The main idea of dynamic programming is that estimation of $\hat{\tau}_K$ is computed first, which is the rightmost change point. Then the time series data from 1 to $\hat{\tau}_K$ will be divided into $K - 1$ parts, and we will estimate τ_{K-1} recursively. However, the computation complexity is $O(KN^2)$, and taking Discrete Fourier transformation and spectrum smoothing into consideration, it is time-consuming.

To reduce the computation complexity, Zou *et al.* (2014) proposed a screening algorithm. For X_j, \dots, X_{j+l} , where l is some constant integer, calculate the location where the function below reaches its maximum.

$$r_j = \arg \max_r \int_{-\pi}^{\pi} \hat{f}_{j,j+r}(u) \log \frac{st(\hat{f}_{j,j+r})}{st(f)} du + (l-r) \int_{-\pi}^{\pi} \hat{f}_{j+r+1,j+l}(u) \log \frac{st(\hat{f}_{j+r+1,j+l})}{st(f)} du.$$

Here $\hat{f}_{j,j+r}$ and $\hat{f}_{j+r+1,j+l}$ denote the estimated spectral density functions based on observations from j to $j+r$, and $j+r+1$ to $j+l$. Let j change from 1 to $N-l$, then we have a set A_{sc} containing all r_j , then apply dynamic programming on A_{sc} . The main idea is that if a change point is included in X_j, \dots, X_{j+l} , then the equation above should reach its maximum at this true change point. That is, A_{sc} contains true change points. Here we should choose $l < N_{\min}$ so that X_j, \dots, X_{j+l} contain only one change point.

When calculating the spectrums within each segment of time series, the most widely used method is Fast Fourier Transformation (FFT). However, in FFT, a problem is that spectral density function is estimated on Nyquist frequency. If so, time series with different length is estimated on different set of Fourier frequency. So when calculating the integral in $R(\tau_1, \dots, \tau_K)$, we cannot align the frequencies where spectral density functions f_k and f are estimated. Our method is that we first choose a set of frequencies, denoted by Λ , then apply Discrete Fourier Transformation on Λ for every subset of samples, which will solve the alignment problem naturally.

For BIC criterion, usually dynamic programming will be applied first, then calculate values of BIC criterion for all $L \leq K_{\max}$. Obviously, this will increase complexity. Killick *et al.* (2012) proposed a method called Pruned Exact Linear Time (PELT), which would significantly reduce computation complexity. The main idea is that when a new sample is included, check all the remaining locations before new sample, if the objective function decreases to the extent that is larger than the penalty term C_N , remove those locations which do not satisfy the condition, and add the next sample into our calculation until the end. The cardinality of the set of all remaining locations is the estimated number of change points and the elements in that set will be the estimated change points. Under some assumptions,

the computation complexity is linear with respect to sample size.

In BIC criterion, another problem is how to choose C_N . Although we can set C_N to satisfy conditions in Theorem 3.2, this choice may be too large which leads to underestimation of K . To overcome this difficulty, we first choose a length, which equals ml , then compute the median, denoted by me_{BIC} , for all values of the function below:

$$\int_{-\pi}^{\pi} \hat{f}_{j,j+ml}(u) \log \frac{st(\hat{f}_{j,j+ml})}{st(f)} du, \forall j = 1, \dots, K.$$

$\hat{f}_{j,j+ml}$ is the estimated spectral density function from X_j, \dots, X_{j+ml} , $\hat{F}_{j,j+ml}$ is its integral. Finally, $C_N = me_{\text{BIC}} \times N^c$. Our simulation shows that an appropriate choice for c is 0.73 regardless of m .

The selection of spectral windows is an important topic in spectral estimation. In Priestley (1981), bandwidth is selected as follows:

$$B_W = 2\sqrt{6} \left(\frac{1}{m^r} k^{(r)} \right)^{1/r},$$

where $k^{(r)} = \lim_{u \rightarrow 0} \frac{1 - w(u)}{|u|^r}$, where $w(u)$ is the inverse Fourier transformation of $W(u)$, and r is the largest integer so that the limit aforementioned exists and is non-zero. m is the scale parameter in spectral windows. By the proofs of Theorem 3.1, we can see that estimators of change points reach consistency because the remaining term tends to infinity slower than N , and the convergence rate is dominated by m . So, bandwidth selection does not matter too much, which is verified by our simulations.

In application, sometimes sample size is large. Although we can apply some methods, such as screening, PELT, to boost the calculation, the computation complexity is still intolerable. Here, we set a ‘‘change point searching unit’’, denoted by n_{su} , which is mentioned in Hawkins (2001). That is, when searching for change points, we add a unit of observations into our calculation each time, not just one observation. This unit is different from ml , which is used to calculate spectrums since estimating spectral density function usually needs sufficient amount of observations. By setting this unit, change point can only be estimated at $n_{\text{su}}, 2n_{\text{su}}, 3n_{\text{su}}$, and so on. If we set this unit equal to 1, the algorithm degenerates to the

general scenario when no searching unit is given. This will dramatically increase the speed of our algorithm. The computation complexity of dynamic programming is $O(N^2)$ (Zou et al. 2014), so by setting a searching unit n_{su} , the complexity will drop to $O(N^2/n_{\text{su}}^2)$. Apparently the estimation accuracy will be sacrificed, since the true change points and the maximizers of objective function, may not lie on the grid. We suggest the choice of n_{su} by choosing the desired estimation accuracy first. Intuitively speaking, if we want the estimates having the accuracy of 1% of the total sample size, then we can set $n_{\text{su}} = 0.01N$.

Our method can cooperate with user's experience. Since Fourier transformation is done on a grid of frequencies Λ , so we can customize Λ . For example, if we know that the power of some particular high frequencies may change while lower frequencies remain stable, Fourier transformation can be carries out within high frequencies and lower frequencies will be ignored. This will not only reduce the computation complexity, but also remove the disturbance of irrelevant frequencies.

3.4 Simulation

In this section we show the finite sample properties of our method and compare it with AutoPARM, Wild Binary Segmentation (WBS), Binary Segmentation (BS), and NMCD under several cases. WBS and BS are in Korkas and Fryzlewicz (2017). AutoPARM (Davis et al. 2006) divided non-stationary time series into several different autoregressive processes by Minimal Description Length. In Korkas and Fryzlewicz (2017), Locally Stationary Wavelets was applied to estimate the second-order structure, then Wild Binary Segmentation or Binary Segmentation, was imposed to divide the time series into several segments based on CUSUM statistics. NMCD (Zou *et al.* 2014) detected the change of probability distribution functions for a sequence of independent random variables based on a nonparametric maximum likelihood approach.

Following the measure of two sets in Zou et al. (2014), we calculate the distance between

two sets G and \hat{G} , which are the true change point set and the estimated set respectively, by

$$\varrho(\hat{G}||G) = \sup_{b \in \hat{G}} \inf_{a \in G} |a - b|, \text{ and } \varrho(G||\hat{G}) = \sup_{a \in G} \inf_{b \in \hat{G}} |a - b|.$$

The first measurement shows if there is an estimator close enough to every true change point, while the second measurement reveals whether some estimators are far away from true change points. When K is known, both measures should give good performances in the sense that $\varrho(\hat{G}||G)$ and $\varrho(G||\hat{G})$ are small. For all the tables, $\varrho(\hat{G}||G)$ and $\varrho(G||\hat{G})$ are shown outside the parentheses while $\varrho(\hat{G}||G)/N$ and $\varrho(G||\hat{G})/N$ are shown inside the parentheses, which can measure the estimation accuracy of $\hat{\kappa}$. In the following subsections, $\xi_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$, if it is not mentioned. As mentioned earlier, to guarantee the estimation accuracy of spectral density function, we should set the minimal length of a segment, which is denoted by ml . In this simulation, we set $ml = 350$ if it is not mentioned. By setting ml , we also assume that the distance of two adjacent change points will not be closer than ml , so we set $K_{\max} = 6$ for each case. If K should be estimated, we report the percentage when K is accurately detected. All simulations are obtained with 1000 replications.

3.4.1 Autoregressive Process

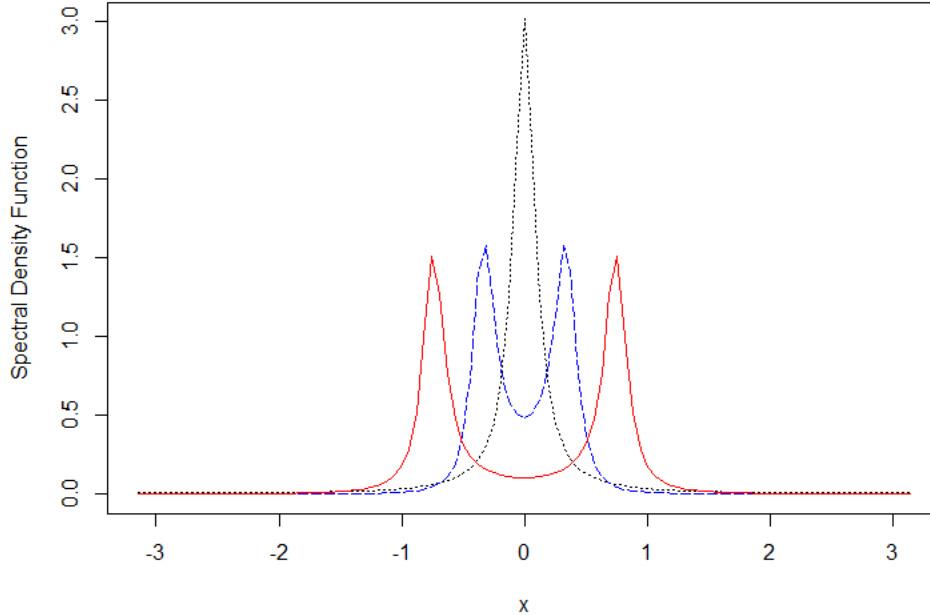
Following the examples in Davis *et al.* (2006) (Case 1), we generate the non-stationary time series from the following.

Autoregressive Processes (Case 1):

- 1 $X_t - 0.9X_{t-1} = \xi_t, 1 \leq t \leq 1024,$
- 2 $X_t - 1.69X_{t-1} + 0.81X_{t-2} = \xi_t, 1025 \leq t \leq 1536,$
- 3 $X_t - 1.32X_{t-1} + 0.81X_{t-2} = \xi_t, 1537 \leq t \leq 2048.$

and their normalized spectral density functions are shown in Figure 3.1.

Figure 3.1: Normalized Spectral Density Functions in Davis *et al.* (2006)



AR(1) is in black dotted line, the second AR is in blue dashed line, the last AR is in red solid line.

In Table 3.1, we show the results of our method when K is known with the spectral density function of total samples and white noise as the baseline functions, respectively, and different choices of bandwidths. In Table 3.2, the simulations are conducted when K is unknown. We adopt Bartlett window since it guarantees the non-negativity of estimated spectral density function. To reduce computation complexity, screening algorithm is applied.

Table 3.1: Performances of NSCD for Case 1 with Different Baseline Functions and Bandwidths When K Is Known

		$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
$m = N^{1/3}$	\hat{f}_0	22.99 (0.011)	22.99 (0.011)
	$\frac{1}{2\pi}$	23.55 (0.012)	23.55 (0.012)
$m = N^{1/4}$	\hat{f}_0	22.61 (0.011)	22.61 (0.011)
	$\frac{1}{2\pi}$	26.47 (0.013)	26.47 (0.013)

Table 3.2: BIC criterion of NSCD for AR Processes with Comparison to AutoPARM, WBS, BS, and NMCD

	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
AutoPARM	93.3%	7.61 (0.004)	8.05 (0.004)
NSCD	98.83%	25.00 (0.012)	37.33 (0.018)
WBS	37.9%	66.65 (0.033)	183.184 (0.089)
BS	62.7%	84.875 (0.041)	122.058 (0.060)
NMCD	0%	147.4365 (0.072)	554.9088 (0.271)

From Table 3.1, the choice of baseline function does not make much difference. Also, under different bandwidths, results are quite similar. We may suggest to choose larger bandwidth for large sample size since the bias of estimation is small. From Table 3.2, it is not surprising that AutoPARM outperforms NSCD, because AutoPARM knows the exact structure of data. WBS and BS perform weaker than NSCD, and can not estimate the number of change points accurately. NMCD fails to capture the true change points, since it is designed for independent random variables. In fact, it prefers to overestimate the number of change points, which is 3.86.

3.4.2 ARMA Process and Invertible Moving Average Process

Next, we simulate from ARMA and invertible MA processes. Table 3.3 and Table 3.4 contain the results.

ARMA (Case 2):

$$1: X_t - X_{t-1} + 0.25X_{t-2} = \xi_t + 0.8\xi_{t-1}, 1 \leq t \leq 500,$$

$$2: X_t - 0.5X_{t-1} = \xi_t, 501 \leq t \leq 1100,$$

$$3: X_t - 1.7X_{t-1} + 0.9X_{t-2} - 0.168X_{t-3} = \xi_t - 1.6\xi_{t-1} + 0.79\xi_{t-2} - 0.12\xi_{t-3}, 1101 \leq t \leq 1800.$$

Invertible MA (Case 3):

1: $X_t = (3 + B)(2 - B)\xi_t, 1 \leq t \leq 500,$

2: $X_t = (3 - B)(2 - B)\xi_t, 501 \leq t \leq 1100,$

3: $X_t = (3 + B)(2 - B)\xi_t, 1101 \leq t \leq 1800.$

Table 3.3: Performance of NSCD for ARMA and Invertible MA Processes with Different Baseline Functions and Bandwidths When K Is Known

		ARMA		MA	
		$\varrho(\hat{G} G)$		$\varrho(G \hat{G})$	
$m = N^{1/3}$	\hat{f}_0	35.29 (0.020)	34.74 (0.020)	13.09 (0.007)	13.09 (0.007)
	$\frac{1}{2\pi}$	32.74 (0.018)	32.07 (0.018)	13.11 (0.007)	13.11 (0.007)
$m = N^{1/4}$	\hat{f}_0	31.12 (0.017)	31.01 (0.017)	12.84 (0.007)	12.84 (0.007)
	$\frac{1}{2\pi}$	26.39 (0.015)	26.38 (0.015)	13.71 (0.008)	13.71 (0.008)

Table 3.4: Performance of BIC Criterion of NSCD for ARMA and Invertible MA Processes with Comparison to AutoPARM, WBS, BS, NMCD

	ARMA			MA		
	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
AutoPARM	95.3%	48.37 (0.027)	27.62 (0.015)	99%	11.42(0.006)	11.42 (0.006)
NSCD	99%	31.42 (0.017)	33.26 (0.018)	99.6%	14.59 (0.008)	15.69 (0.009)
WBS	74.9%	184.393 (0.102)	66.288 (0.037)	91.8%	33.586 (0.019)	45.362 (0.025)
BS	77.7%	188.955 (0.105)	69.194 (0.038)	96.3%	29.673 (0.016)	36.611 (0.020)
NMCD	0%	188.64 (0.105)	324.31 (0.180)	0.1%	160.4434 (0.089)	326.4282 (0.181)

We expect that AutoPARM still works since it is well known that ARMA and invertible MA processes can be approximated by causal AR processes. We see that under ARMA setting, the performance of NSCD is comparable to AutoPARM, while AutoPARM performs better under MA settings. Both NSCD and AutoPARM perform better than WBS and BS. NMCD

fails in both cases as it does in the previous section. The number of change points is overestimated by NMCD again, which is 3.005 and 3.00, respectively.

3.4.3 Non-invertible MA Process

We will show simulation results when samples are generated from non-invertible MA process. In theory causal AR cannot approximate non-invertible MA process. Results are given in Table 3.5 and 3.6.

Non-invertible MA (Case 4):

- $X_t = (1 + 2B + B^2 + 5B^3)\xi_t, 1 \leq t \leq 500,$
- $X_t = (1 - 2B + 2B^2 - 5B^3)\xi_t, 501 \leq t \leq 1100,$
- $X_t = (1 + 2B - B^2 + 5B^3)\xi_t, 1101 \leq t \leq 1800.$

Table 3.5: Performance of NSCD for Non-Invertible MA Processes with Different Baseline Functions and Bandwidths When K Is Known

		MA	
		$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
$m = N^{1/3}$	\hat{f}_0	33.60 (0.019)	33.60 (0.019)
	$\frac{1}{2\pi}$	36.00 (0.020)	36.00 (0.020)
$m = N^{1/4}$	\hat{f}_0	40.11 (0.022)	40.11 (0.022)
	$\frac{1}{2\pi}$	41.00 (0.023)	41.00 (0.023)

Table 3.6: Performance of BIC Criterion of NSCD for Non-Invertible MA Processes with Comparison to AutoPARM, WBS, BS, and NMCD of

	MA		
	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
AutoPARM	36.7%	408.60 (0.227)	27.24 (0.015)
NSCD	99.7%	37.49 (0.021)	38.02 (0.021)
WBS	18.2%	488.674 (0.271)	90.493 (0.050)
BS	8.6%	576.462 (0.320)	56.975 (0.032)
NMCD	0%	165.54 (0.092)	326.701 (0.182)

From Table 3.6, we can see that NSCD works for Case 4, while AutoPARM fails under the second one since the number of change points are all underestimated. The possible reason is that the marginal variance of Case 4 does not vary much. All the other three methods fail to capture true change points. WBS and BS underestimate K , which are 1.268 and 1.088, respectively. NMCD overestimates K again, which is 3.015.

3.4.4 Random Noise without the Existence of Higher Moments

Next, we investigate the results when ξ does not have higher-order expectations. Here $\xi \sim \frac{1}{\sqrt{2}}t(4)$ so that the variance of ξ is still 1. The results for four cases are shown in Table 3.7, when K is known.

Table 3.7: Performance of NSCD for $t(4)$ Random Noise

	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
Case 1	24.24 (0.012)	24.37 (0.012)
Case 2	29.67 (0.017)	28.91 (0.016)
Case 3	11.74 (0.007)	11.74 (0.007)
Case 4	36.36 (0.020)	36.36 (0.020)

Apparently, from Table 3.7, it is clear that Assumption 1 can be slightly violated in application, while still achieving good performance.

3.4.5 Investigation of Smaller Sample Size

In the previous sections, we discuss the performances when sample size is about 2000, which is large. Here we shrink the sample size by half, and investigate the estimate accuracy as well as the choice of bandwidth. We set $ml = 200$, and $K_{\max} = 5$. Baseline function is chosen to be \hat{f}_0 . Results are shown in Table 3.8.

Table 3.8: Performance of NSCD with Different Bandwidths While The Sample Size Is Reduced by Half

	bandwidth	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
Case 1	$m = N^{1/3}$	23.195 (0.023)	23.263 (0.023)
	$m = N^{1/4}$	27.716 (0.027)	28.075 (0.027)
Case 2	$m = N^{1/3}$	31.72 (0.035)	30.718 (0.035)
	$m = N^{1/4}$	28.104 (0.031)	27.734 (0.030)
Case 3	$m = N^{1/3}$	14.664 (0.016)	14.158 (0.016)
	$m = N^{1/4}$	12.652 (0.014)	12.652 (0.014)
Case 4	$m = N^{1/3}$	37.28 (0.041)	37.272 (0.041)
	$m = N^{1/4}$	44.3 (0.05)	43.877 (0.049)

Table 3.9: Performance of BIC Criterion of NSCD with Different Bandwidths While The Sample Size Is Reduced by Half

	bandwidth	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
Case 1	$m = N^{1/3}$	96%	24.972 (0.024)	30.44 (0.030)
	$m = N^{1/4}$	90.4%	35.268 (0.040)	39.494 (0.044)
Case 2	$m = N^{1/3}$	96.2%	38.326 (0.043)	28.1 (0.031)
	$m = N^{1/4}$	98.7%	27.494 (0.031)	26.384 (0.029)
Case 3	$m = N^{1/3}$	96.5%	31.458 (0.035)	12.73 (0.014)
	$m = N^{1/4}$	96%	32.484 (0.036)	11.436 (0.013)
Case 4	$m = N^{1/3}$	95%	41.228 (0.046)	40.246 (0.045)
	$m = N^{1/4}$	92%	47.67 (0.053)	53.254 (0.059)

As shown in the table, the estimate accuracy is not affected by the choice of bandwidth when sample size is relatively small. The estimation accuracy of κ is worse than the results in Section 3.4.1 to 3.4.3, since the sample size is decreased by half.

3.4.6 Further Investigation of BIC Criterion

In this section, we investigate the performance with different choice of N^c . c ranges from 0.1 to 0.9 and \hat{K} will be plotted in all the cases mentioned above with two choices of bandwidth. All the results are shown from Figures 3.2a to 3.3d with baseline function \hat{f}_0 . And it is obvious that $c = 0.73$ is a good choice for all the four cases.

Next, we are going to simulate the performances of BIC when ml changes. Since from the previous section, we can see that the choice of C_N depends on ml . Figures 3.4a to 3.5d show the results when $ml = 300$, while Figures 3.6a to 3.7d give the performances when $ml = 250$.

As we can see from the figures, $c = 0.73$ is not a good choice in general, and it will overestimate the number of change points. Although overestimating is tolerable since we do

not want to miss the true change points, we still suggest that a sufficiently large choice for ml is necessary. Since the maximum possible number of change points is $\lceil N/ml \rceil$, $K_{\max} * ml \leq N$. We suggest that one should choose a reasonable K_{\max} depending on some prior information, then choose a ml satisfying $ml \leq N/K_{\max}$.

Figure 3.2: Influence of m and ml on BIC Criterion with $m = N^{1/4}$ and $ml = 350$

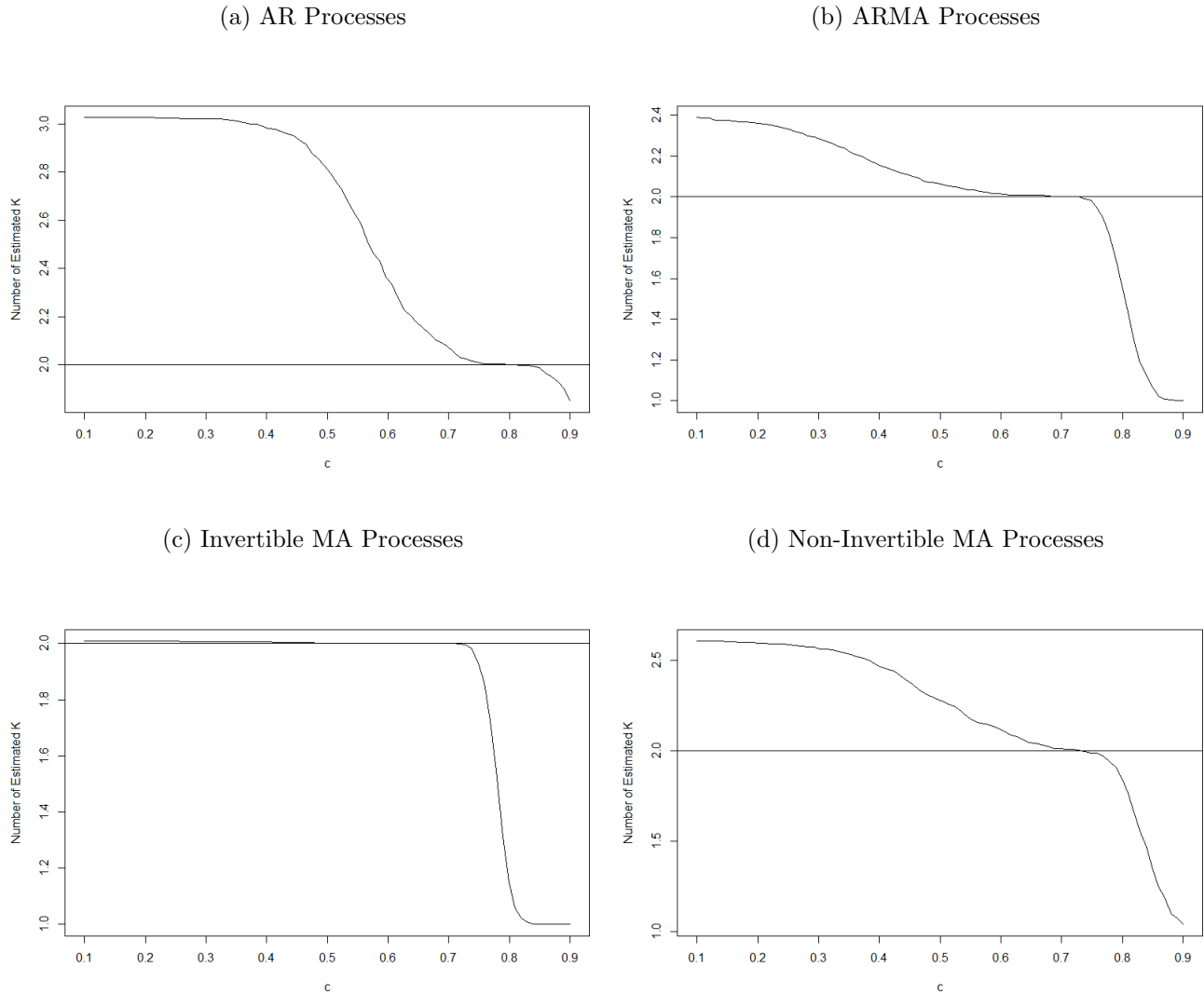
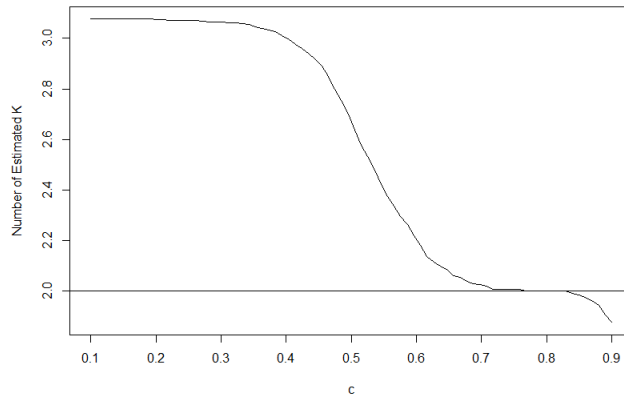
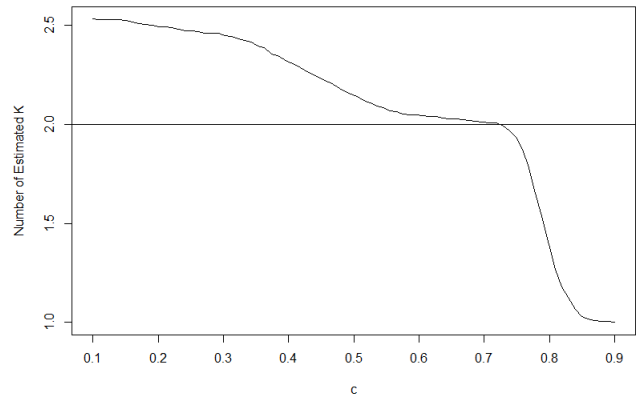


Figure 3.3: Influence of m and ml on BIC Criterion with $m = N^{1/3}$ and $ml = 350$

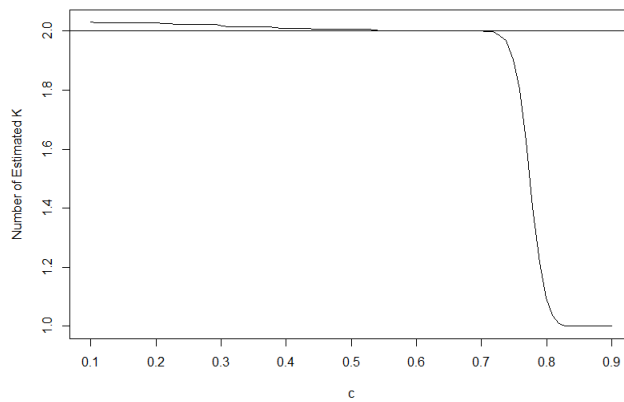
(a) AR Processes



(b) ARMA Processes



(c) Invertible MA Processes



(d) Non-Invertible MA Processes

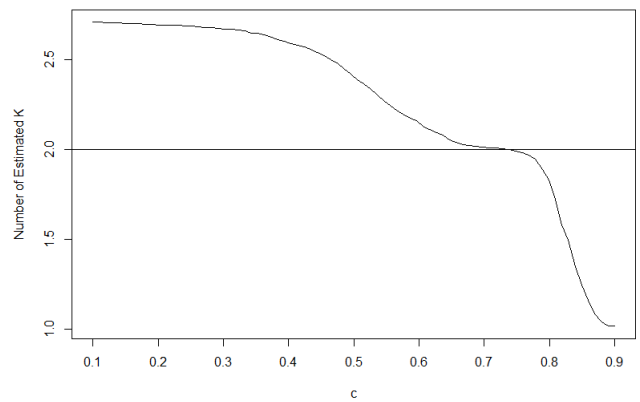
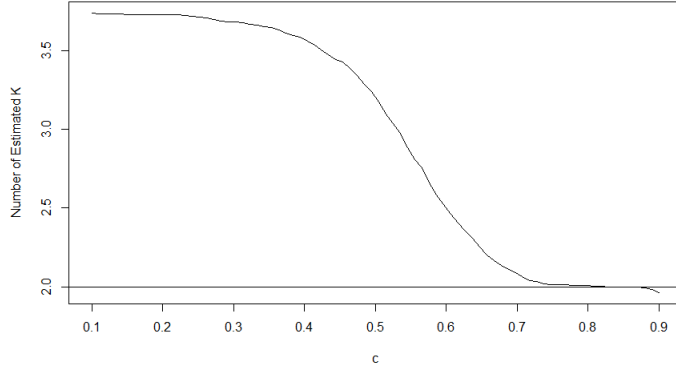
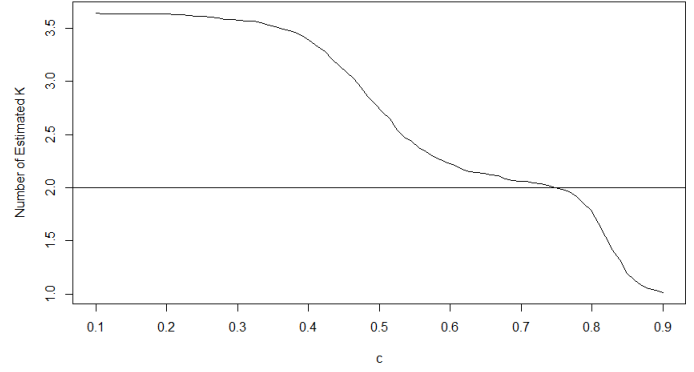


Figure 3.4: Influence of m and ml on BIC Criterion with $m = N^{1/3}$ and $ml = 300$

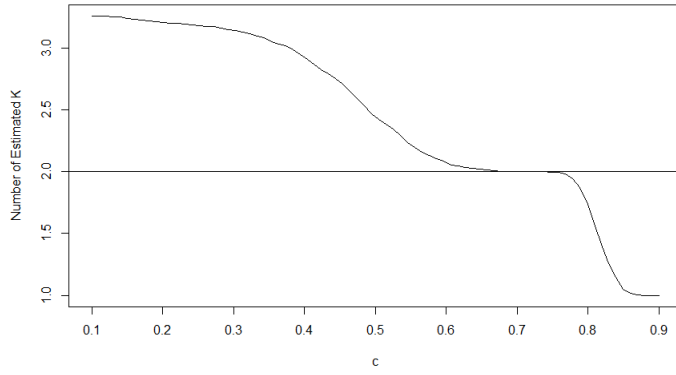
(a) AR Processes



(b) ARMA Processes



(c) Invertible MA Processes



(d) Non-Invertible MA Processes

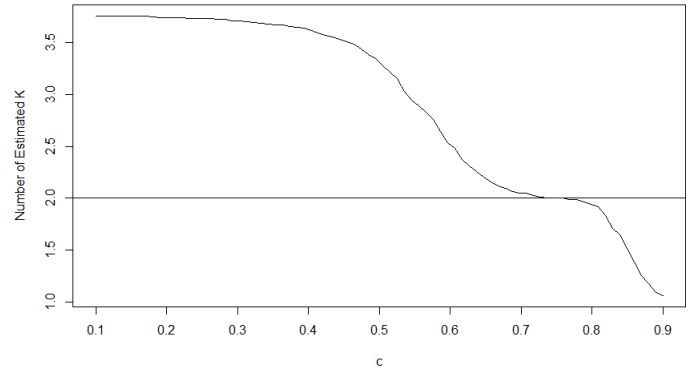
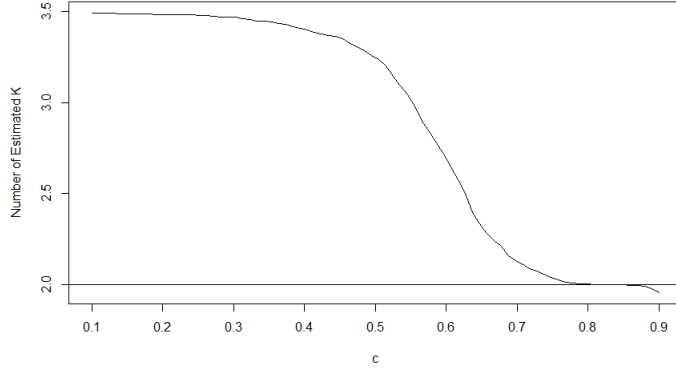
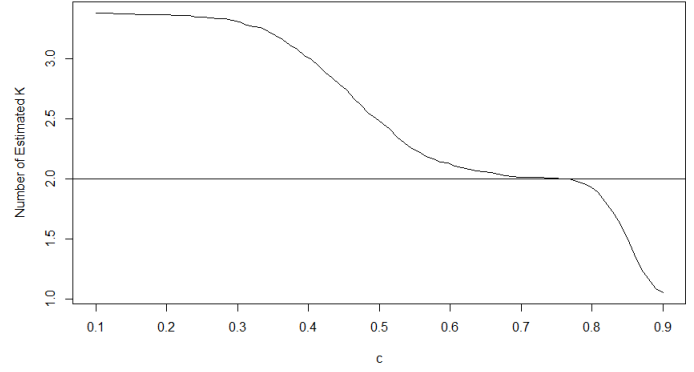


Figure 3.5: Influence of m and ml on BIC Criterion with $m = N^{1/4}$ and $ml = 300$

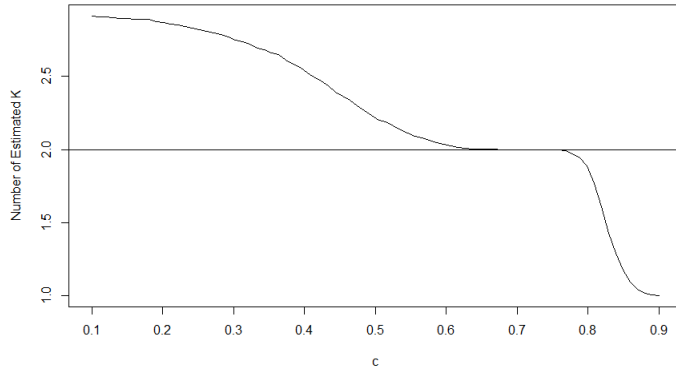
(a) AR Processes



(b) ARMA Processes



(c) Invertible MA Processes



(d) Non-Invertible MA Processes

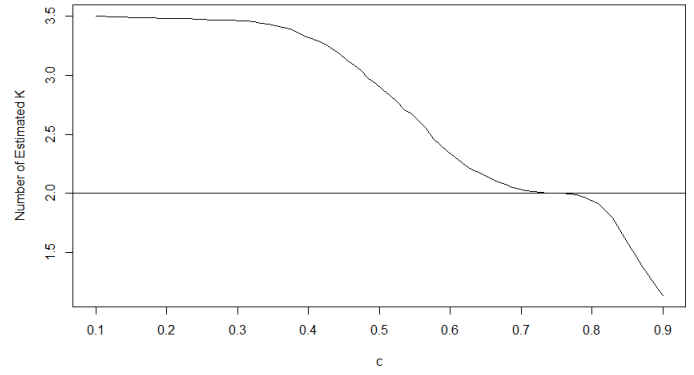
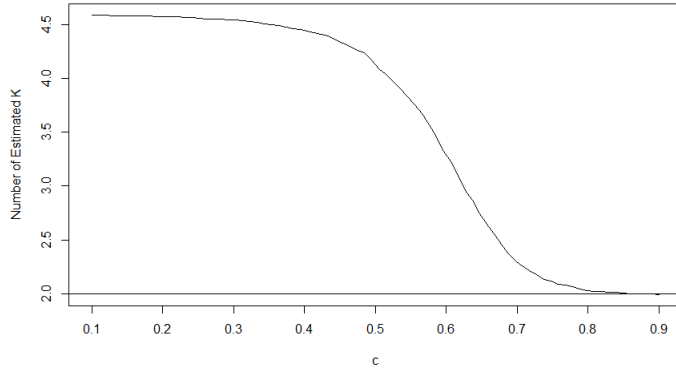
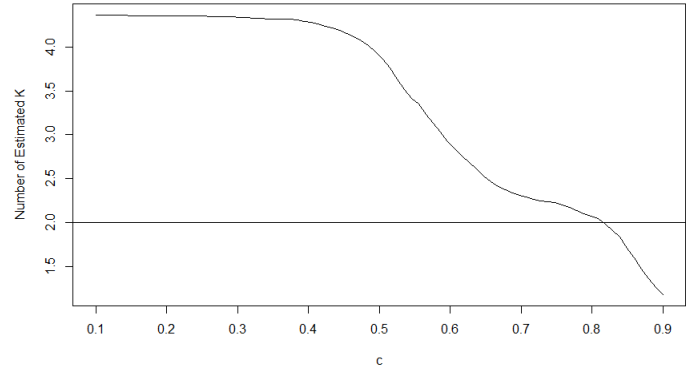


Figure 3.6: Influence of m and ml on BIC Criterion with $m = N^{1/3}$ and $ml = 250$

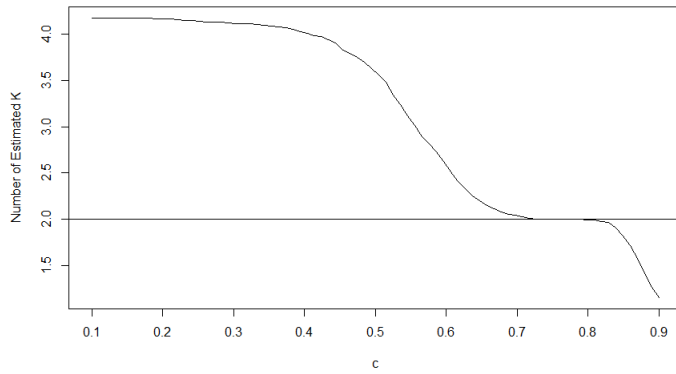
(a) AR Processes



(b) ARMA Processes



(c) Invertible MA Processes



(d) Non-Invertible MA Processes

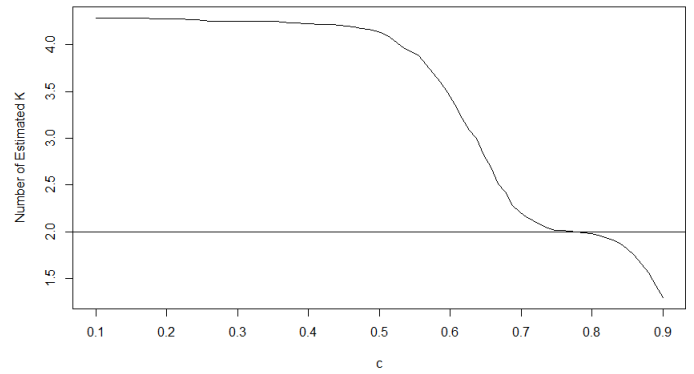
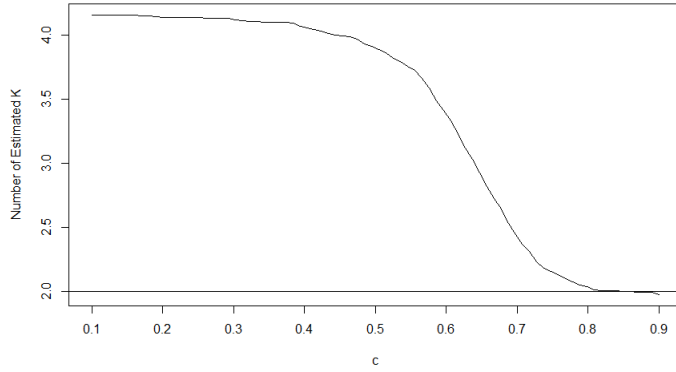
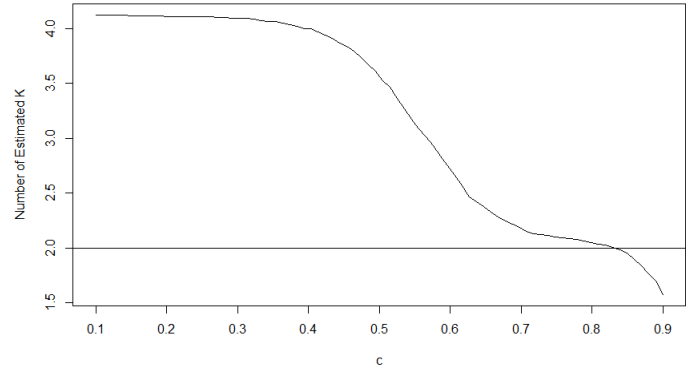


Figure 3.7: Influence of m and ml on BIC Criterion with $m = N^{1/4}$ and $ml = 250$

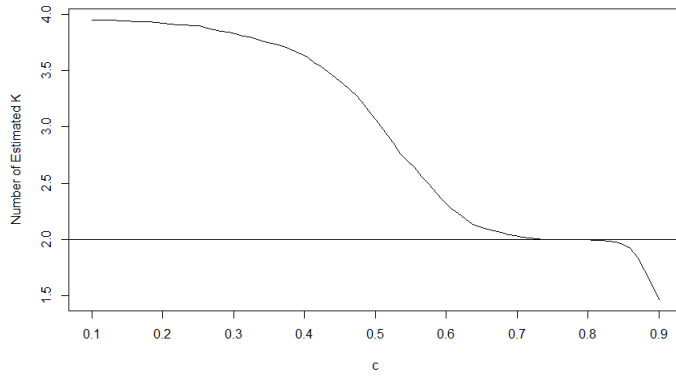
(a) AR Processes



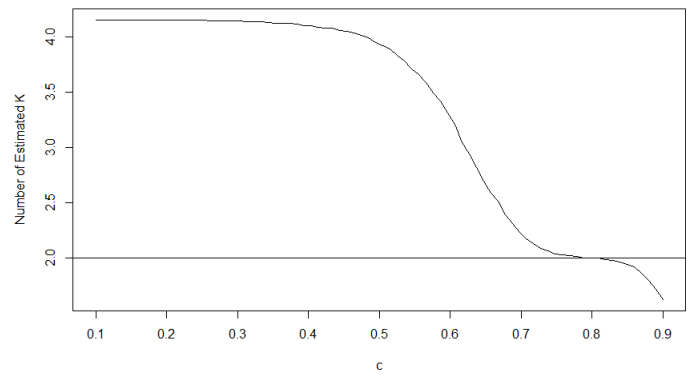
(b) ARMA Processes



(c) Invertible MA Processes



(d) Non-Invertible MA Processes



3.4.7 Influence of Change Point Searching Unit

In this section, we investigate the influence of n_{su} . For all the four cases, $n_{su} = 10$, which means that for Case 2-4, $\tau_1^0, \dots, \tau_K^0$ are all a multiple of n_{su} , while for Case 1, the true change points are not divisible by n_{su} . We set $m = N^{1/4}$ and baseline function $f = \hat{f}_0$ for four cases. The results are shown in Table 3.10 and 3.11.

Table 3.10: Performance of NSCD with Searching Unit When K Is Known

	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
Case 1	22.9 (0.011)	22.9 (0.011)
Case 2	29.26 (0.016)	29.26 (0.016)
Case 3	10.14 (0.006)	10.14 (0.006)
Case 4	40.52 (0.023)	40.52 (0.023)

Table 3.11: Performance of BIC Criterion of NSCD with Searching Unit

	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
Case 1	96.2%	22.696 (0.011)	39.72 (0.019)
Case 2	97.2%	42.02 (0.0233)	30.68 (0.017)
Case 3	99.6%	12.54 (0.007)	10.14 (0.006)
Case 4	98.6%	40.56 (0.023)	43.74 (0.024)

We can see that setting a n_{su} will not affect our results much even if the true change points are not a multiple of n_{su} in Case 1, so it is safe to apply this in application, which could boost the computation as well as give accurate estimations.

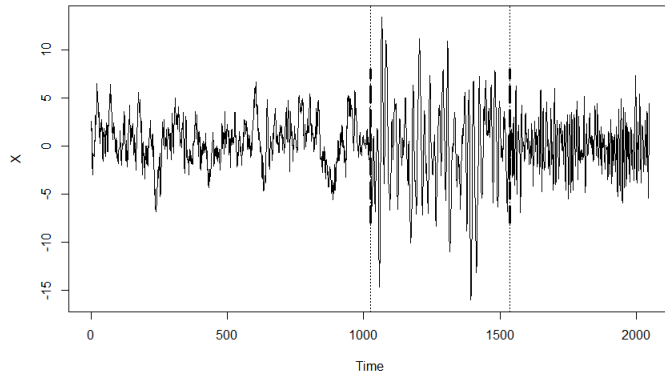
3.5 Case Study

3.5.1 Simulated Data

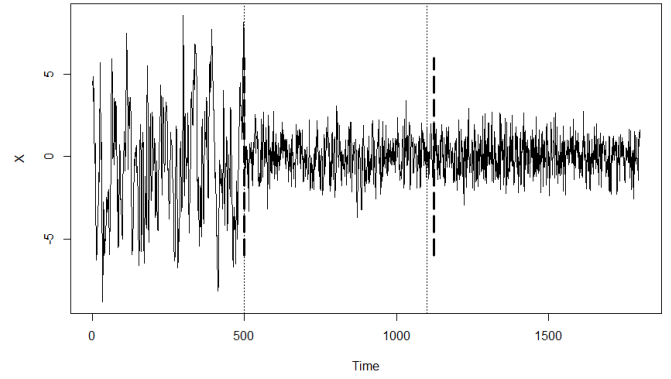
In the simulation section, we check the performances of our method based on 4 cases, and compare NMSD with other methods. Here we will investigate the change point detection of these 4 cases more directly. Figure 3.8a-3.8d are the realizations of Case 1 to Case 4. The long-dashed lines are the locations of estimated change points while the dotted lines represent the true change points. It is easy to see from Figure 3.8a that the existence of change points are obvious, since marginal variance of the second segment is bigger. In Case 2, the second change point is far less obvious than the first one which can be seen in Figure 3.8b. In Figure 3.8c, since the observations in the first two segments concentrate more around their mean compared to the third segment, the second change point may be captured by eyes. In Figure 3.8d, two change points are not apparent any more. However, from the perspective of spectrum, those change points are easily detected. Figure 3.9a-3.9d give the estimated spectral density functions. In Case 4, we can see that the power of spectrum of the first segment concentrates more at higher and lower frequencies, while the spectrum of the second part has more power at low frequency. The spectrum of the third part is similar to the first one, but it achieves more power in the low frequency range.

Figure 3.8: Estimated Change Points by NSCD for Different Cases

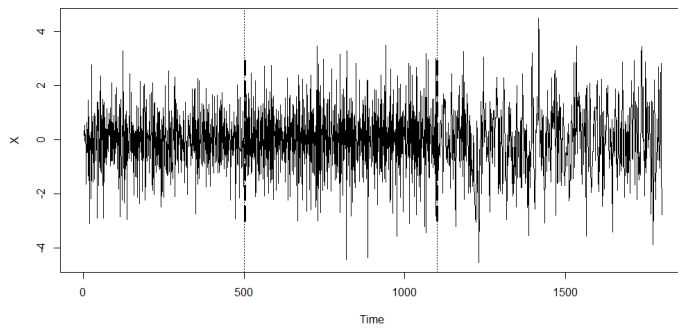
(a) AR Processes



(b) ARMA Processes



(c) Invertible MA Processes



(d) Non-Invertible MA Processes

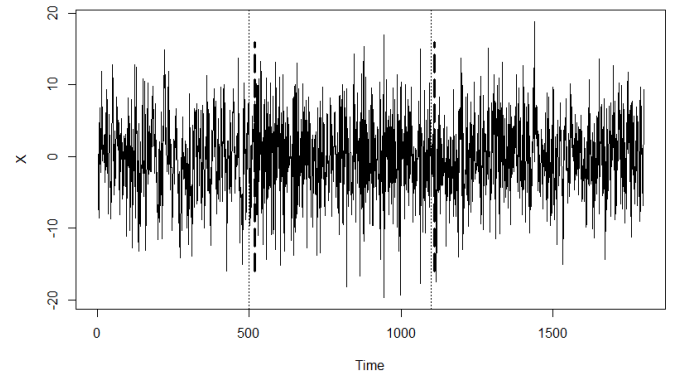
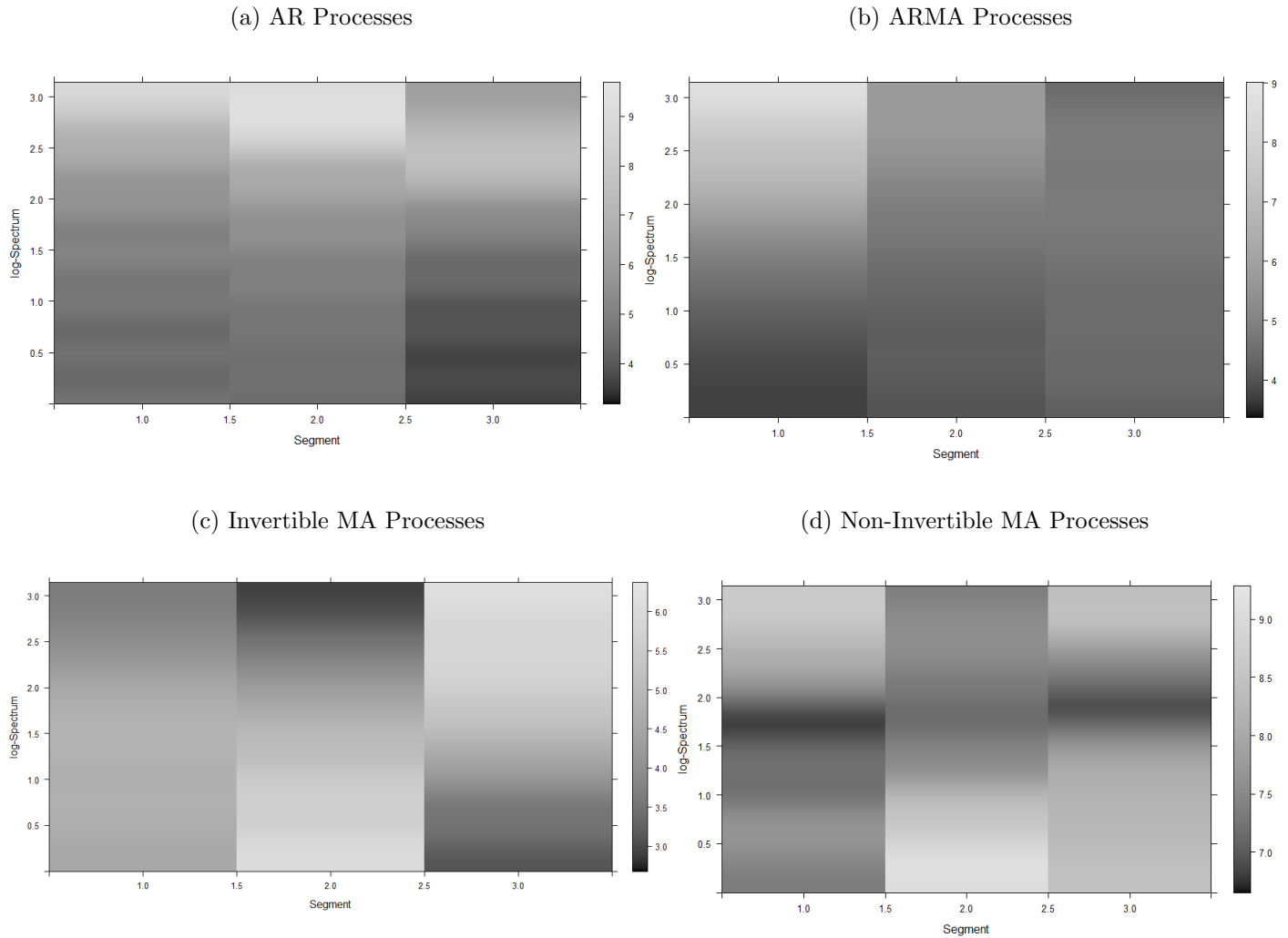


Figure 3.9: Estimated Spectrums by NSCD for Different Cases



3.5.2 Infant Electrocardiogram Data (ECG)

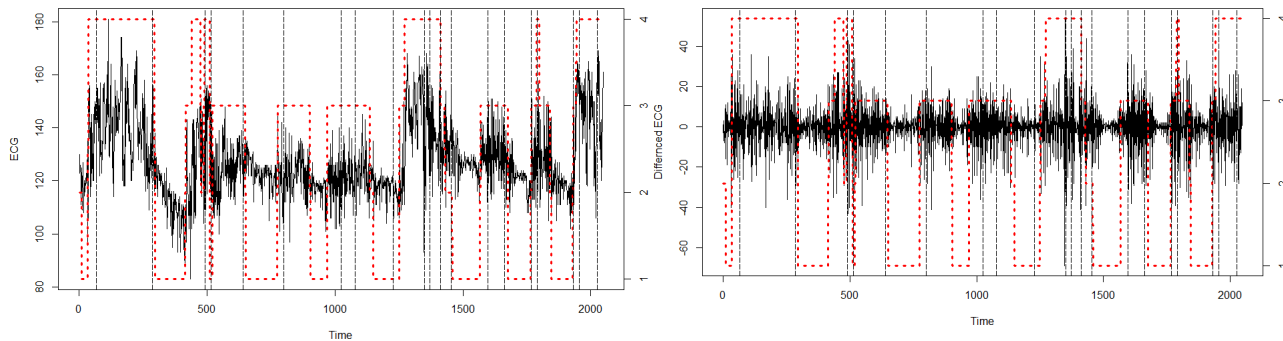
In this section, we will apply our method to the ECG data from the R package *wavethresh*. The ECG data is a record of an infant’s heart rate. The rate is sampled every 16 seconds from 21:17:59 to 06:27:18. The sample size is 2048. Meanwhile, a trained expert monitored EEG (brain) and EOG (eye-movement) to decide the sleep state of the infant. The sleep state codes are 1=quiet sleep, 2=between quiet and active sleep, 3=active sleep, 4=awake (See also, Nason 2016). The attachment of the recording sensors to the infants’ scalp (EEG) and face (EOG) may be distressing to both parents and infants, and may lead to artifacts

by interfering with the infants' sleep, and is not practicable in the home environment. Thus such recordings must be performed in the hospital which further adds to cost and potential distress (Nason et al. 2001). So the ECG data of infant heart rate is used to monitor the sleep state. By the descriptions above, the change points in the ECG data are where the sleep state changes. From Figure 3.10a, we can see that the record does not have zero mean. Furthermore, it can be observed that in some segments there are local means which means that we can not centralize each segment by subtracting mean of the whole sample. So like Korkas and Fryzlewicz (2017), we difference the samples once to remove the mean and then apply NSCD on the first difference of the ECG data to estimate the change points. Figure 3.10b shows the results, in which the horizontal dotted red line represents the sleep states, while the vertical dashed black line represents the estimated change points. As we can see, NSCD can detect most of the change points except around time 1000, where NSCD fails to detect two change points while estimating two change points with medium bias. Around time 1500, NSCD gives more estimations there due to the abrupt change of the difference of observations. Also, some changes happened at short segments can also be detected.

Figure 3.10: Infant ECG Data and the Analysis

(a) Infant ECG Data

(b) Analysis of Infant ECG Data

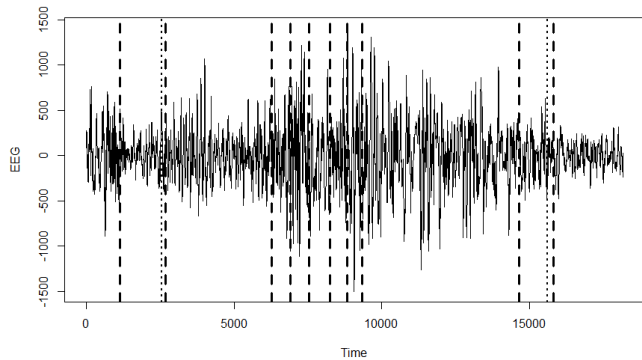


3.5.3 Electroencephalography Recordings for Seizure

In this section, we investigate the performance of NSCD when applied on EEG data for seizure (Goldberger et al. 2000). The data is retrieved from <https://www.physionet.org/pn6/chbmit/>. For each subject, the EEG signal was recorded into several data files, which was one-hour long. Here all subjects were monitored for several days to trace their states, so we only analyze the data file with seizures. The sampling rate is 256 Hz, so the number of observations in each recording is 9216000, which is huge. In the data file we choose (which is chb01_16), seizure happened only once with duration 51 seconds, so the number of change points is 2. The duration of seizure is very short compared with the total length of this recording, so we analyze a part of the data which begins at 10 seconds before the seizure and ends at the 10 seconds after the seizure, so the sample size is 18176, which is still large. So here we will set a change point searching unit equal to 64, which is 0.25 second, to ease the computation burden. What is more, the minimal length n_{\min} is 256 which is 1 second. There are totally 23 channels in EEG recording, we apply our method on channel “FP1-F3” and “FP1-F7”. The results are shown in Figure 3.11a and 3.11b. The vertical dotted lines represent the locations of the beginning and ending of seizure, while the vertical dashed lines are the estimated change points. To demonstrate the changes in spectrums, we plot the estimated spectrums in Figure F3s and F7s. As we can see, NSCD can successfully detect the true change points, while almost all the other estimates fall between the true change points. This is because during seizure, the EEG recordings change abruptly, which will bring more non-stationarity into the time series.

Figure 3.11: Analysis of Different Channels of EEG Data

(a) Channel FP1-F3



(b) Channel FP1-F7

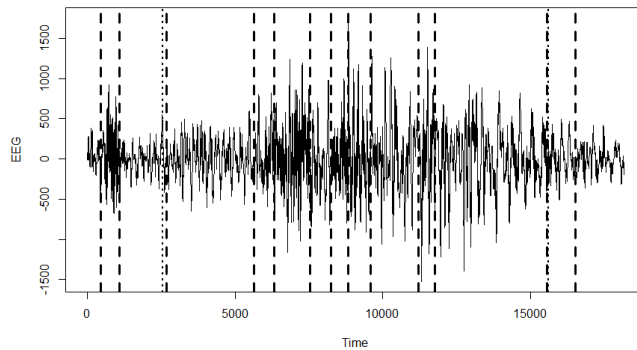
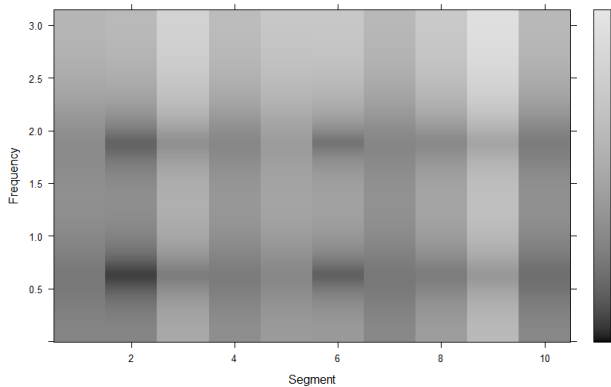
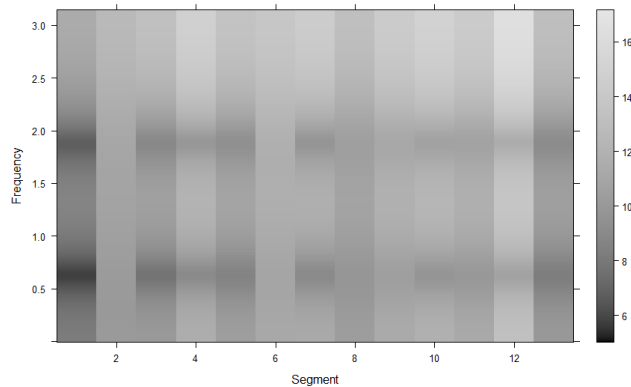


Figure 3.12: Estimated Spectrums of Different Channels of EEG Data

(a) Channel FP1-F3



(b) Channel FP1-F7



Chapter 4

Multivariate Nonparametric Change Point Detection

Univariate time series has been studied extensively. There are substantial literatures concerning univariate case from time domain and frequency domain. The classical approach usually assumes that time series is stationary. The definition of multivariate stationary time series is similar to the univariate one, which is that the expectation is constant and the covariance is dependent on time through the lag. In multivariate time series, researchers not only need to focus on the dependence in the same series, but also the relationship between different series. The existing univariate stationary models, such as autoregressive process, ARMA model, have been extended into multivariate scenario (Brockwell and Davis 1991). To simulate time series with high volatility in the scope of stationarity, some literatures developed multivariate ARCH-GARCH (Helmut 2006). Similar to univariate time series, multivariate spectrum analysis is well developed (Priestley 1981, Hannan 1970). However, for spectrum analysis of non-stationary time series, there are not many literatures, and this chapter will focus on this.

Consider a sequence of non-stationary time series $\{X(1), \dots, X(N)\}$ and let $\tau_0^0, \tau_1^0, \tau_2^0, \dots, \tau_K^0, \tau_{K+1}^0$ be nonnegative integers satisfying $0 = \tau_0^0 < \tau_1^0 < \tau_2^0 < \dots < \tau_K^0 < \tau_{K+1}^0 = N$. Here

$X(t)$ are all $p_x \times 1$ random vectors. Within the j th segment of the series, $\tau_{j-1}^0 < t \leq \tau_j^0$, $X(t)$ is stationary. τ_j^0 are called structural breaks, or change points, which is unknown. K is the number of change points. Set $EX(t) = \boldsymbol{\mu}$, $\mathbf{R}(d) = E(X(t+d) - \boldsymbol{\mu})(X(t) - \boldsymbol{\mu})^T$, and $R_{ls}(d)$ is the (l, s) entry of $\mathbf{R}(d)$. In multivariate spectrum analysis, besides considering spectral density functions $f_{kk}(v)$ for the k th time series, when $\sum_{d=-\infty}^{+\infty} |R_{kk}(d)| < +\infty$, the cross spectrum, f_{ls} between l and s time series, exists when $\{R_{ls}(d)\}_d$ is absolutely summable (Priestley 1981), and

$$f_{ls}(v) = \frac{1}{2\pi} \sum_{d=-\infty}^{+\infty} e^{-idv} R_{ls}(d), v \in [-\pi, \pi].$$

So instead of considering only one spectral density function in the univariate scenario, we should focus on spectral density matrix $\mathbf{f}(v) = (f_{ls}(v))_{1 \leq l, s \leq p_x}$. Since $R_{ls}(h)$ is no longer even function when $l \neq s$, in general, $\mathbf{f}(v)$ is a complex valued matrix. However, by $R_{ls}(h) = R_{sl}(-h)$, we have

$$f_{ls}(v) = \bar{f}_{sl}(v).$$

Here \bar{f}_{sl} denotes the conjugate of f_{sl} . So $\mathbf{f}(v)$ is a Hermitian matrices, $\forall v$, that is

$$\mathbf{f}(v) = \mathbf{f}^*(v),$$

here $*$ denotes conjugate transpose. By the property of Hermitian matrix, $\mathbf{f}(v)$ has non-negative real eigenvalues $\lambda_1(v) \geq \lambda_2(v) \geq \dots \geq \lambda_{p_x}(v)$, $\forall v$, which is also the basis of our method.

Since $\lambda_1(v)$ is a nonnegative function defined on $[-\pi, \pi]$, by appropriate scale, $\lambda_1(v)$ could become a probability density function, then the measurements for probability density function can be used here to measure the difference between eigenvalues. In this chapter, we use Kullback-Leibler divergence. We estimate change points by maximizing the discrepancy functions. When estimating spectral density matrix, we adopt classical nonparametric method, so we call our method multivariate nonparametric spectral change-point estimation (MNSCD). In application, usually we need to estimate the number of change points. So similar to Yao (1988), Zou, Yin, Feng, and Wang (2014), we propose a BIC criterion.

The rest of this chapter is organized as follows. In Section 4.1, we give the details of settings and method. Asymptotic theory is shown in Section 4.2. Some implementations and further details of our algorithm are given in Section 4.3. In Section 4.4, simulation results with comparison to other methods will be shown. In Section 4.5, we study the performance when applied to real data.

4.1 Model and Methodology

Consider a sequence of non-stationary multivariate time series $\{X(1), \dots, X(N)\}$. Here $X(t) = \sum_{j=-\infty}^{+\infty} G(k, j)\boldsymbol{\xi}(t-j)$, and $\boldsymbol{\xi}(t)$ is a $p_\xi \times 1$ random vector, $\boldsymbol{\xi}(t) \stackrel{\text{i.i.d.}}{\sim} (0, \boldsymbol{\Sigma})$, $\tau_{k-1}^0 < t \leq \tau_k^0$, $k = 1, \dots, K+1$ with $\tau_0^0 = 0$, $\tau_{K+1}^0 = N$, $\tau_k^0 - \tau_{k-1}^0 = N_k$, $G(k, j) = (g_{rs}(k, j))_{1 \leq r \leq p_x, 1 \leq s \leq p_\xi}$ is $p_x \times p_\xi$ matrix, which is called impulse response matrix (Priestley 1981), satisfying $\sum_{j=-\infty}^{+\infty} |g_{rs}(k, j)| < +\infty$, $\forall 1 \leq r \leq p_x, 1 \leq s \leq p_\xi$. We can see that this condition is stronger than the one in the definition of multivariate linear process (See Section 2.3). So basically, we assume that our non-stationary time series consists of several stationary multivariate linear processes. By Priestley (1981), we know that for the k th segment, the spectral density matrix $\mathbf{f}_k(v)$ exists and equals

$$\mathbf{f}_k(v) = \frac{1}{2\pi} \Gamma(v) \boldsymbol{\Sigma} \Gamma^*(v),$$

where $\Gamma(v) = \sum_{j=-\infty}^{+\infty} G(k, j)e^{-ijv}$. Without loss of generality, we assume that the expectation of $X(t)$ is zero. Our goal is to detect change points for multivariate time series. By the discussion above, $\mathbf{f}_k(v)$ is Hermitian, so its eigenvalues $\lambda_1(k, v) \geq \dots \geq \lambda_{p_x}(k, v)$ are non-negative. Here we only choose $\lambda_1(k, v)$ to detect change points. Although we can use all the eigenvalues, estimation of small eigenvalues would lead to huge bias. So in the remaining of this chapter, we suppress subscript “1”, and denote the largest eigenvalue of the k th segment by $\lambda_k(v)$ without ambiguity. Similar to univariate case, we scale $\lambda_k(v)$ by its integral $\Lambda_k = \int_{-\pi}^{\pi} \lambda_k(v) dv$. For the sake of conciseness, we define $st(f) = \frac{f(u)}{F}$, for any function f defined on $[-\pi, \pi]$, where $F = \int_{-\pi}^{\pi} f(u) du$. Since $st(\lambda_k)$ is a proper probability density

function, we define our objective function by Kullback-Leibler divergence:

$$R(\tau_1, \dots, \tau_K) = \sum_{k=1}^{K+1} (\tau_k - \tau_{k-1}) \int \hat{\lambda}_k(v) \log \frac{st(\hat{\lambda}_k)}{st(\lambda)} dv.$$

Here $\hat{\lambda}_k(v)$ is the estimated maximum eigenvalue of $\hat{\mathbf{f}}_{\tau_{k-1}, \tau_k}(v)$, which is the estimated spectral density matrix for $\{X(\tau_{k-1}), \dots, X(\tau_k)\}$, and $\Lambda = \int_{-\pi}^{\pi} \lambda(v) dv$, $\lambda(v)$ is chosen to be different from $\lambda_k(v)$, $\forall k$. Here we adopt the classical method to estimate $f_{rs}(\tau_{k-1}, \tau_k, v)$. First, periodogram is calculated as follows:

$$I_{XX}(\tau_{k-1}, \tau_k, v) = \frac{1}{\tau_k - \tau_{k-1}} \left(\sum_{t_1=\tau_{k-1}+1}^{\tau_k} X(t_1) e^{-it_1 v} \right) \left(\sum_{t_2=\tau_{k-1}+1}^{\tau_k} X(t_2) e^{-it_2 v} \right)^*.$$

Then for (r, s) entry of $I_{XX}(\tau_{k-1}, \tau_k, v)$, we smooth periodogram by applying window function $W(u)$ to get $\hat{f}_{rs}(\tau_{k-1}, \tau_k, v)$, that is

$$\hat{f}_{rs}(\tau_{k-1}, \tau_k, v) = m \int W(m(v-u)) I_{XX,rs}(\tau_{k-1}, \tau_k, u) du.$$

Since we still want $\hat{f}_{rr}(\tau_{k-1}, \tau_k, v)$ to be non-negative so that $\hat{\mathbf{f}}(\tau_{k-1}, \tau_k, v)$ is Hermitian, we adopt Bartlett kernel $mW(mu) = \sin^2(mu/2)/(2\pi m \sin^2(u/2))$, where m is the bandwidth. Of course, different $W(u)$ should be chosen to accommodate different spectrums. However, it is hard to choose appropriate smooth function due to unknown structural breaks. So Bartlett kernel is applied to every entry of $\hat{\mathbf{f}}(v)$. In application, one usually applies discrete Fourier transformation on time series data, which is denoted by $V = \{v_1, \dots, v_{N_v}\}$. We may allow $N_v = N$ so N_v can tend to infinity. Our objective function is based on this setting.

In real application, we need to estimate the number of change points K . Here we propose a BIC criterion

$$BIC_L = - \max_{\tau_1, \dots, \tau_L} R(\tau_1, \dots, \tau_L) + LC_N,$$

where C_N is constant which will be illustrated later, L is the number of change points. \hat{K} is chosen where BIC reaches it maximum.

4.2 Asymptotic Theory

Denote $\kappa_k = \tau_k^0/N$, $\hat{\kappa}_k = \hat{\tau}_k/N$ for each k . Our assumptions are as follows:

A1: $E\xi_{t,r_1} \cdots \xi_{t,r_{8q}} < +\infty$, $1 \leq r_1, \dots, r_{8q} \leq p_\xi$, where q is some integer satisfying $q \geq 3$, $\{g_{rs}(k, j)\}$ is absolutely summable in terms of j , $\forall k$.

A2: $W(v)$ is a non-negative, even, bounded, integrable function, $\int_{-\pi}^{\pi} W(v)dv = 1$,
 $\int_{-\pi}^{\pi} (W(v))^{1-\frac{1}{2q}} dv < +\infty$.

A3: $N_{\min}/N \rightarrow c_{\min} > 0$, as $N \rightarrow \infty$, and $N_{\min} > m$, where $N_{\min} = \min_{1 \leq k \leq K} (\tau_k^0 - \tau_{k-1}^0)$.

A4: $\forall k$, $\mathbf{f}_k(v)$ is positive semidefinite, and at least one of the eigenvalues is greater than zero for each v . What is more, $f_{k,r,s}(v)$, which is the (r, s) entry of $\mathbf{f}_k(v)$, satisfies uniform Lipschitz condition:

$$|f_{k,r,s}(u_1) - f_{k,r,s}(u_2)| \leq B_{k,r,s}|u_1 - u_2|$$

$\forall u_1 \in [-\pi, \pi]$, $u_2 \in [-\pi, \pi]$, and $B_{k,r,s}$ is some constant.

A5: $m = O(N^\alpha)$, where $\frac{1}{4} \leq \alpha < \alpha + \frac{3}{2q} < \frac{1}{2}$.

A6: $w(0) = 1$, and $w(v)$ has continuous derivatives to the order of $2q$.

A7: $\{\lambda_k(u)\}_{k=1}^{K+1}$ are linearly independent for $u \in [-\pi, \pi]$.

A8: $ml = c_{ml}N$, where $c_{ml} > 0$. $ml < N_{\min}$.

In Assumption 8, ml is the minimal length of time series when estimating spectral density functions, since a sufficient number of sample points is always necessary, especially when the convergence rate of estimators is slow. Also, $ml < N_{\min}$ so that all change points are distinguishable. A sufficient condition for Assumption 7 is that $\{\mathbf{f}_k(v)\}$ are linearly independent in terms of k , and can not be diagonalized by one orthonormal matrix at the same time. Assumption 1-6 are similar to those in Woodroffe and Van Ness (1967) so that

the $\max_{v \in V} st(\hat{\lambda}_k)$ can be bounded in probability. Assumption 4 can be made stronger so that in Assumption 5, α can be less than $\frac{1}{4}$ (see Woodroffe and Van Ness 1967 for further details). Theorem 4.1 gives the consistency of change point estimation.

Theorem 4.1. *When K is known, $\hat{\kappa}_j \xrightarrow{P} \kappa_j, \forall j = 1, \dots, K$.*

Similar to the univariate scenario, to estimate the number of change points, a pre-specified upper bound K_{\max} satisfying $K < K_{\max}$ will be given. Then BIC values for each $1 \leq L \leq K_{\max}$ are calculated and \hat{K} will be the number where BIC reaches its minimum. The following theorem establishes the consistency of estimation of BIC criterion.

Theorem 4.2. *If $C_N/N \rightarrow 0, C_N/N^{\frac{q+3+2q\alpha}{2q}} \rightarrow \infty$, we have $P(\hat{K} = K) \rightarrow 1$.*

4.3 Algorithm

Although data is multivariate, utilizing eigenvalues allows us to calculate in the scope of univariate scenario. Since our objective function is separable, we could still use dynamic programming (Zou et al. 2014, Hawkins 2001).

To reduce the computation complexity, Zou *et al.* (2014) proposed a screening algorithm. For $X(j), \dots, X(j+l)$, where l is some constant, calculate the location where the function below reaches its maximum.

$$r_j = \arg \max_r r \int_{-\pi}^{\pi} \hat{\lambda}_{j,j+r}(u) \log \frac{st(\hat{\lambda}_{j,j+r})}{st(\lambda)} du + (l-r) \int_{-\pi}^{\pi} \hat{\lambda}_{j+r+1,j+l} \log \frac{st(\hat{\lambda}_{j+r+1,j+l})}{st(\lambda)} du.$$

Here $\hat{\lambda}_{j,j+r}$ and $\hat{\lambda}_{j+r+1,j+l}$ denote the estimated eigenvalues of spectral density matrix of samples from j to $j+r$, and $j+r+1$ to $j+l$. Let j change from 1 to $n-l$, then we have a set A_{sc} containing all r_j , then apply dynamic programming on A_{sc} . The main idea is that if a change point is included in $X(j), \dots, X(j+l)$, then the equation above should reach its maximum at this true change point. That is, A_{sc} contains true change points. Here we should choose $l < N_{\min}$ so that $X(j), \dots, X(j+l)$ contain only one change point. In fact, comparing

to the univariate case, multivariate method needs to carry out eigendecomposition, which means that the computation complexity will increase inevitably.

When calculating the spectrums within each segment of time series, the most widely used method is Fast Fourier Transformation (FFT). However, in FFT, a problem is that spectral density function is estimated on Nyquist frequency. If so, time series with different length is estimated on different set of Fourier frequency. So when calculating the integral in $R(\tau_1, \dots, \tau_K)$, we cannot align the frequencies where spectral density functions $\lambda_k(v)$ and $\lambda(v)$ are estimated. Our method is that we first choose a set of frequencies, denoted by V , then apply Discrete Fourier Transformation on V for every subset of samples, which will solve the alignment problem naturally.

For BIC criterion, usually dynamic programming will be applied first, then calculate values of BIC criterion for all $L \leq K_{\max}$. Obviously, this will increase complexity. Killick, Fearnhead and Eckley (2012) proposed a method called Pruned Exact Linear Time (PELT), which would significantly reduce computation complexity. The main idea is that when a new sample is included, check all the remaining locations before new sample. If the objective function decreases to the extent that is larger than the penalty term C_N , remove those locations which do not satisfy the condition, and add the next sample into our calculation until the end. The cardinality of the set of all remaining locations is the estimated number of change points and the elements in that set will be the estimated change points. Under some assumptions, the computation complexity is linear with respect to sample size.

In BIC criterion, another problem is how to choose C_N . Although we can set C_N to satisfy conditions in Theorem 4.2, this choice may be too large which leads to underestimation of K . To overcome this difficulty, we first choose a length, which equals ml , then compute the median, denoted by me_{BIC} , for all values of the function below.

$$\int \hat{\lambda}_{j,j+ml}(u) \log \frac{st(\hat{\lambda}_{j,j+ml})}{st(\lambda)} du, \forall j.$$

$\hat{\lambda}_{j,j+ml}$ is the estimated spectral density function from $X(j), \dots, X(j+ml)$, $\hat{\Lambda}_{j,j+ml}$ is its integral. Finally, $C_N = me_{\text{BIC}} \times N^c$. Our simulation shows that $c = 0.73$ is an appropriate

choice.

The selection of spectral windows is an important topic in spectral estimation. In Priestley (1981), bandwidth is selected as follows

$$B_W = 2\sqrt{6} \left(\frac{1}{m^r} k^{(r)} \right)^{1/r},$$

where $k^{(r)} = \lim_{u \rightarrow 0} \frac{1 - w(u)}{|u|^r}$, where $w(u)$ is the inverse Fourier transformation of $W(u)$, and r is the largest integer so that the limit aforementioned exists and is non-zero. m is the scale parameter in spectral windows. By the proofs of Theorem 4.1, we can see that estimators of change points reach consistency because the remaining term tends to infinity slower than N , and the convergence rate is dominated by m . So bandwidth selection does not matter too much.

4.4 Simulation

In this section we show the finite sample properties of our method and compare it to Davis *et al.* (2006) and ECP (Matteson and James 2014) in several cases. Davis *et al.* (2006) introduced AutoPARM, which divides non-stationary time series into several AR processes based on Minimal Description Length. ECP used the α absolute moment to detect change points for multivariate independent random variables. Following the measure of two sets in Zou *et al.* (2014), we calculate the distance between two sets G and \hat{G} , which are the true change point set and the estimated set respectively, by

$$\varrho(\hat{G}||G) = \sup_{b \in \hat{G}} \inf_{a \in G} |a - b|, \text{ and } \varrho(G||\hat{G}) = \sup_{a \in G} \inf_{b \in \hat{G}} |a - b|.$$

The first measurement shows if there is an estimator close enough to every true change point, while the second measurement reveals whether some estimators are far away from true change points. When K is known, both measures should give good performances in the sense that $\varrho(\hat{G}||G)$ and $\varrho(G||\hat{G})$ are small. For all the following simulations, we set $m = N^{1/4}$, the baseline function is chosen to be $\hat{\lambda}_0$. As mentioned earlier, to guarantee the estimation

accuracy of spectral density function, we should set the minimal length of a segment, which is denoted by ml . In this simulation, we set $ml = 200$ if it is not mentioned. By setting ml , we also assume that the distance of two adjacent change points will not be closer than ml , so we set $K_{\max} = 6$ for each case. To reduce computation complexity, screening algorithm is applied. All simulations are obtained with 1000 replications.

4.4.1 Autoregressive Process

We generate the multivariate non-stationary time series from the following.

AR processes (Case 1)

$$1: X(t) - \Phi_{11}X(t-1) - \Phi_{12}X(t-2) = \boldsymbol{\xi}(t), 1 \leq t \leq 300,$$

$$2: X(t) - \Phi_2X(t-1) = \boldsymbol{\xi}(t), 301 \leq t \leq 700,$$

$$3: X(t) - \Phi_{31}X(t-1) - \Phi_{32}X(t-2) = \boldsymbol{\xi}(t), 701 \leq t \leq 1200,$$

$$\text{where } \Phi_{11} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}, \Phi_{12} = \begin{pmatrix} 0 & 0 \\ 0.3 & 0 \end{pmatrix}, \Phi_2 = \begin{pmatrix} -0.5 & -0.3 \\ 0 & -0.5 \end{pmatrix}, \Phi_{31} = \begin{pmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{pmatrix},$$

$$\Phi_{32} = \begin{pmatrix} 0 & 0 \\ -0.3 & 0 \end{pmatrix}, \text{ and } \boldsymbol{\xi}(t) \stackrel{i.i.d.}{\sim} N(0, \Sigma) \text{ with } \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

For all the tables, $\varrho(\hat{G}||G)$ and $\varrho(G||\hat{G})$ are shown outside the parentheses while $\varrho(\hat{G}||G)/N$ and $\varrho(G||\hat{G})/N$ are shown inside the parentheses, which can measure the estimation accuracy of \hat{k} , like what we did in Chapter 3. In Table 4.1, we show the results of our method when K is known with the spectral density function of total samples and white noise as the baseline functions, respectively, and different choices of bandwidths. In Table 4.2, the simulations are conducted when K is unknown. We adopt Bartlett window since it guarantees the non-negativity of estimated spectral density function.

Table 4.1: Performance of MNSCD for Multivariate AR Processes When K Is Known

		$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
$m = N^{1/4}$	$\hat{\lambda}_0$	11.78 (0.010)	11.78 (0.010)

Table 4.2: BIC criterion of MNSCD with comparison to AutoPARM and ECP for Multivariate AR Processes

	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
AutoPARM	100%	6.3 (0.005)	6.3 (0.005)
MNSCD	99%	13.04 (0.011)	15.61 (0.013)
ECP	12%	97.974 (0.082)	395.981 (0.330)

From Table 4.1, we can see that MNSCD can estimate the true change points accurately. From Table 4.2, AutoPARM is better than MNSCD, which is reasonable because AutoPARM knows the exact structure of data. Also, ECP fails to detect the true change points because ECP is designed for independent samples, not correlated observations.

4.4.2 ARMA Process and Invertible Moving Average Process

In this section, we investigate the performance of MNSCD in another two cases: ARMA (Case 2) and invertible MA process (Case 3).

ARMA processes (Case 2)

- $X(t) - \Phi_{11}X(t-1) = \xi(t) + \Theta_{11}\xi(t-1)$, $1 \leq t \leq 300$,
- $X(t) - \Phi_{21}X(t-1) = \xi(t) + \Theta_{21}\xi(t-1)$, $301 \leq t \leq 700$,
- $X(t) - \Phi_{31}X(t-1) = \xi(t) + \Theta_{31}\xi(t-1)$, $701 \leq t \leq 1200$,

where $\Phi_{11} = \begin{pmatrix} 0.5 & 0.5 \\ 0 & 0.5 \end{pmatrix}$, $\Theta_{11} = \Phi_{11}^T$, $\Phi_{21} = \begin{pmatrix} -0.7 & 0.5 \\ 0 & -0.7 \end{pmatrix}$, $\Theta_{21} = \Phi_{21}^T$, where $\Phi_{31} =$

$$\begin{pmatrix} 0.2 & 0.5 \\ 0 & 0.2 \end{pmatrix}, \Theta_{31} = \Phi_{31}^T.$$

Invertible MA processes (Case 3):

- $X(t) = \xi(t) + \Theta_{11}\xi(t-1)$, $1 \leq t \leq 300$,
- $X(t) = \xi(t) + \Theta_{21}\xi(t-1)$, $301 \leq t \leq 700$,
- $X(t) = \xi(t) + \Theta_{31}\xi(t-1)$, $701 \leq t \leq 1200$,

$$\Theta_{11} = \begin{pmatrix} 0.5 & 0.4 \\ 0.1 & 0.5 \end{pmatrix}, \Theta_{12} = \begin{pmatrix} 0 & 0.3 \\ 0 & 0 \end{pmatrix}, \Theta_2 = \begin{pmatrix} -0.5 & 0 \\ -0.3 & -0.5 \end{pmatrix}, \Theta_{31} = \begin{pmatrix} 0.5 & 0.4 \\ 0.1 & 0.5 \end{pmatrix},$$

$$\Theta_{32} = \begin{pmatrix} 0 & -0.3 \\ 0 & 0 \end{pmatrix}.$$

The results are shown in Table 4.3 and 4.4.

Table 4.3: Performance of MNSCD for Multivariate ARMA and Invertible MA Processes When K Is Known

	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
ARMA	2.24 (0.002)	2.24 (0.002)
MA	30.5 (0.025)	30.5 (0.025)

Table 4.4: BIC of MNSCD with Comparison to AutoPARM and MNSCD for Multivariate ARMA and Invertible MA Processes

	ARMA			MA		
	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
AutoPARM	99%	1.88 (0.002)	1.96 (0.002)	99.9%	7.498(0.006)	7.534 (0.006)
MNSCD	99%	31.42 (0.026)	33.26 (0.028)	99%	14.59 (0.012)	15.69 (0.013)
ECP	37.3%	124.998 (0.104)	292.686 (0.244)	14.6%	91.001 (0.076)	365.828 (0.305)

From Table 4.4, we can see that AutoPARM is still better than other two methods, since

causal AR processes can approximate ARMA and invertible Moving Average process. ECP is worse than the other two again, which is not surprising.

4.4.3 Non-invertible Moving Average Process

In this section, we will simulate non-stationary time series consisting of several non-invertible MA processes.

Non-Invertible MA (Case 4):

- $X(t) = \Theta_1 \boldsymbol{\xi}(t) + \Theta_2 \boldsymbol{\xi}(t - 1)$, $1 \leq t \leq 300$,
- $X(t) = \Theta_2 \boldsymbol{\xi}(t) + \Theta_3 \boldsymbol{\xi}(t - 1)$, $301 \leq t \leq 700$,
- $X(t) = \Theta_3 \boldsymbol{\xi}(t) + \Theta_4 \boldsymbol{\xi}(t - 1)$, $701 \leq t \leq 1200$,

where Θ_1 to Θ_4 are all 2×4 matrices, and $\Theta_1 = (0.5^{i-j})_{1 \leq i \leq 2, 1 \leq j \leq 4}$, $\Theta_2 = ((-0.8)^{i-j})_{1 \leq i \leq 2, 1 \leq j \leq 4}$, $\Theta_3 = (-0.5^{i-j})_{1 \leq i \leq 2, 1 \leq j \leq 4}$, $\Theta_4 = (0.5^{i-j})_{1 \leq i \leq 2, 1 \leq j \leq 4}$. We can also see that dimension of $\boldsymbol{\xi}_t$ is different from X_t . Tables 4.5 and 4.6 show the results.

Table 4.5: Performance of MNSCD for Multivariate Non-Invertible MA Processes When K Is Known

		MA	
		$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
$m = N^{1/4}$	$\hat{\lambda}_0$	47.94 (0.040)	47.94 (0.040)

Table 4.6: BIC of MNSCD with Comparison to AutoPARM and ECP for Multivariate Non-Invertible MA Processes

		MA	
	\hat{K}	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
AutoPARM	73.8%	119.07 (0.100)	15.15 (0.013)
MNSCD	90%	53.86 (0.045)	67.74 (0.056)
ECP	31.3%	78.00 (0.065)	190.29 (0.159)

From Table 4.6, we can see that MNSCD is the most stable method, while AutoPARM underestimate the number of change points, and ECP cannot give accurate estimates.

4.4.4 White Noise without the Existence of Higher Moments

Similar to univariate scenario, for MNSCD, we still need the existence of higher moments for white noise ξ_t . So we show the results when ξ_t is multivariate t distribution with degree of freedom 4. Table 4.8 gives the results. We can see that our method can still detect the true change points accurately.

Table 4.7: MNSCD for Multivariate t(4) White Noise

	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
Case 1	20.45 (0.017)	20.45 (0.017)
Case 2	2.48 (0.002)	2.48 (0.002)
Case 3	6.25 (0.005)	6.25 (0.005)
Case 4	60.87 (0.051)	58.53 (0.049)

4.4.5 Investigation of Smaller Sample Size

In this section, we shrink the sample size from 1200 to 900 to show the performance of estimation. ml is set to be 200 and $K_{\max} = 5$. The results are shown in Table 4.8. From the table, we can see that the estimation accuracy is worse than the performances when $N = 1200$, since the bias will reduce when N increases.

Table 4.8: Performance of MNSCD While The Sample Size Is Reduced by Half

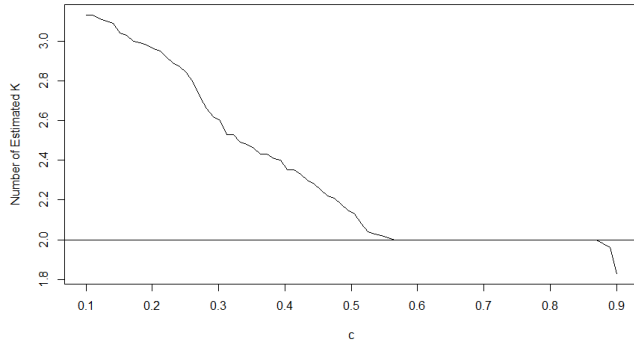
	$\varrho(\hat{G} G)$	$\varrho(G \hat{G})$
Case 1	12.32 (0.010)	12.32 (0.010)
Case 2	2.48 (0.002)	2.48 (0.002)
Case 3	33.4 (0.028)	33.4 (0.028)
Case 4	45.32 (0.038)	45.32 (0.038)

4.4.6 Further Investigation of BIC criterion

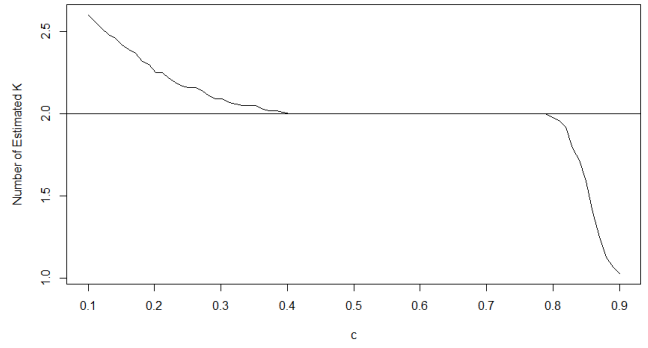
In this section, we will investigate the choice for c in BIC criterion. We let c range from 0.1 to 0.9 and plot the estimated number of change points. Figure 4.1a, Figure 4.1b, Figure 4.1c, and Figure 4.1d show the results for 4 cases mentioned above. The sample size and the bandwidth m are the same as above.

Figure 4.1: Influence of m on BIC Criterion with $m = N^{1/4}$

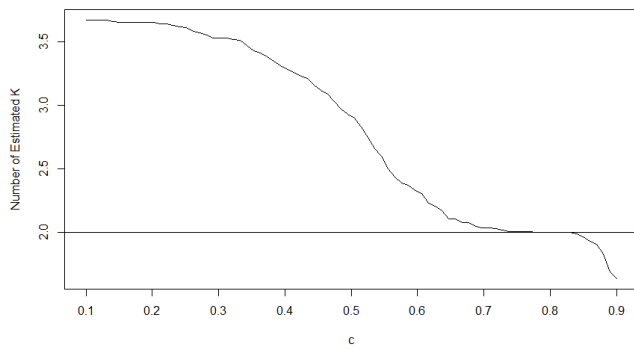
(a) Multivariate AR Processes



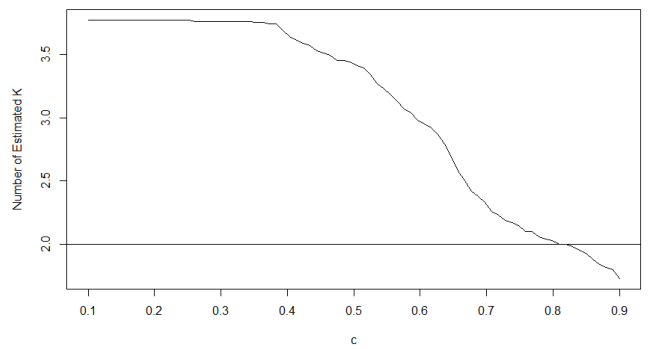
(b) Multivariate ARMA Processes



(c) Multivariate Invertible MA Processes



(d) Multivariate Non-Invertible MA Processes

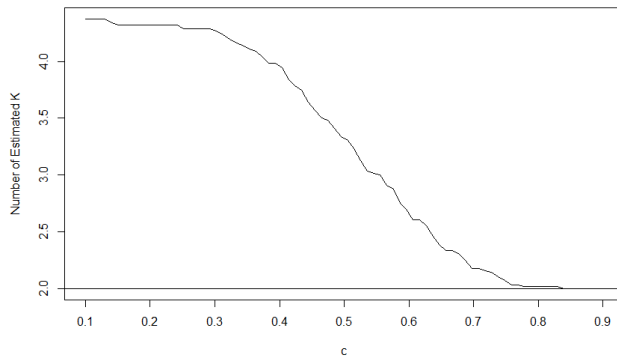


From the figures, we can see that $c = 0.73$ is a good choice for Case 1 to 3. For Case 4, we still choose 0.73 for c as a universal choice. Because the true change points should not be missed and the overestimated K can still detect the true change points well, overestimation will not hurt much.

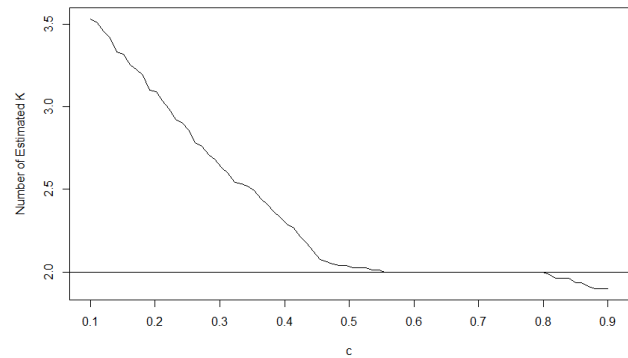
Next, we will investigate the performances of BIC when ml changes. In the previous section, we can see that the choice of C_N depends on ml . Figure 4.2a to 4.2d shows the results when $ml = 150$. Figure 4.3a to 4.3d give the performances when $ml = 100$. In all the figures, m is fixed to be $N^{1/4}$.

Figure 4.2: Influence of ml on BIC Criterion with $ml = 150$

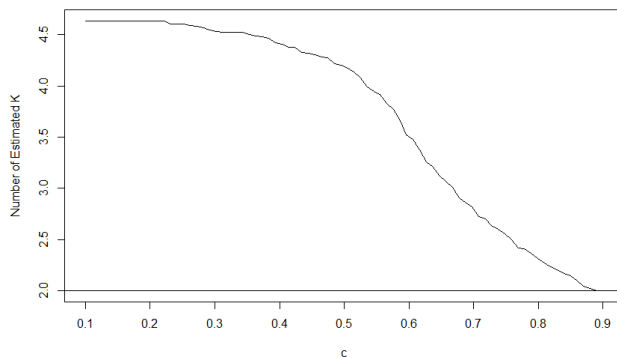
(a) Multivariate AR Processes



(b) Multivariate ARMA Processes



(c) Multivariate Invertible MA Processes



(d) Multivariate Non-Invertible MA Processes

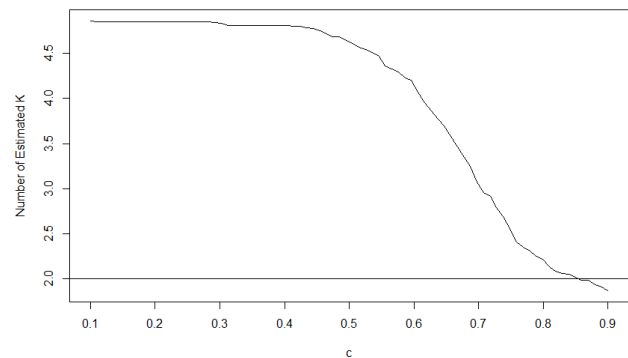
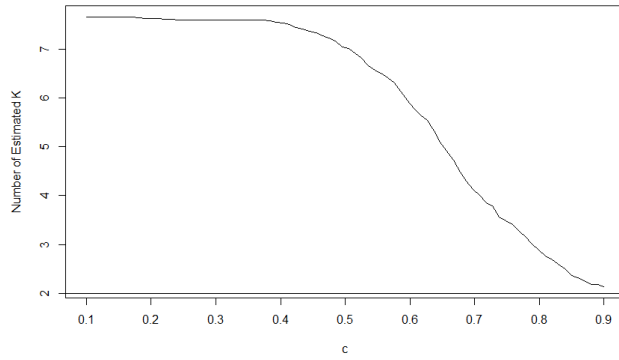
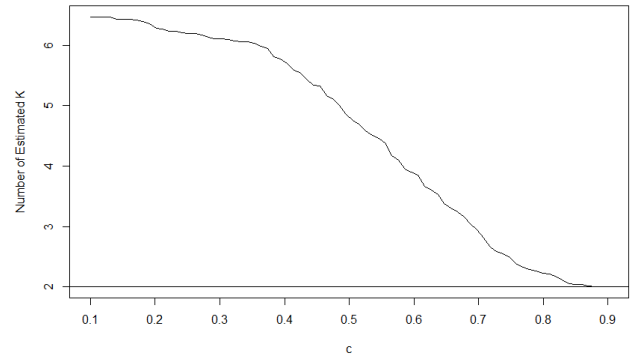


Figure 4.3: Influence of ml on BIC Criterion with $ml = 100$

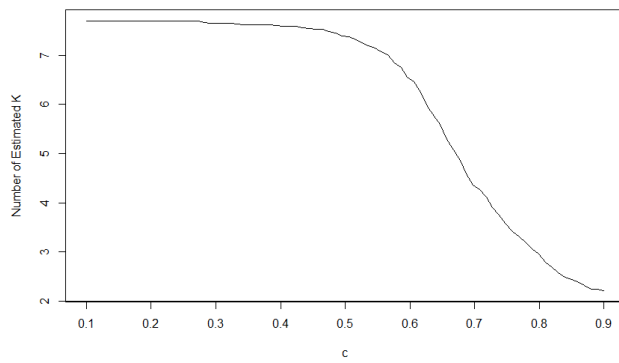
(a) Multivariate AR Processes



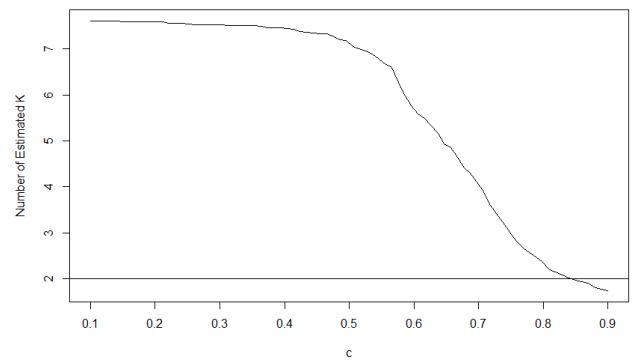
(b) Multivariate ARMA Processes



(c) Multivariate Invertible MA Processes



(d) Multivariate Non-Invertible MA Processes



From the figures, we can see that $c = 0.73$ is no longer a good choice for BIC criterion, which will lead to overestimating the number of change points. So similar to univariate scenario, we should cooperate the choice of ml with some prior information.

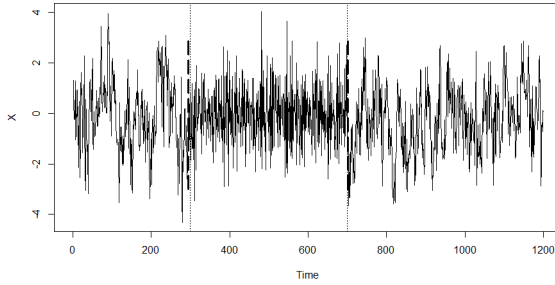
4.5 Case Study

4.5.1 Simulated Data

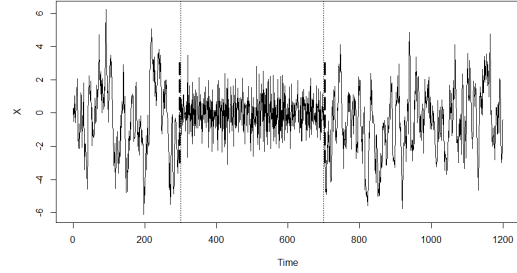
In this section, we show the results of change point detection method more directly. Figure 4.4a- 4.4h are the realizations for Case 1-4. Since time series are multivariate, we show both of the dimensions. In Figure 4.4a and 4.4b , we can see that different dimensions behave very differently. The existence of change points is more obvious in the second dimension for Case 1. In Figures 4.4c and 4.4d, we can see that the second part of observations concentrates more around its mean, while in Figure 4.4e- 4.4f , change points in Case 3 are not as apparent as they are in Case 2. As it is shown in Figures 4.4g and 4.4h, the change point detection for Case 4 is not as accurate as it is in the previous cases, and the existence is not obvious any more.

Figure 4.4: Estimated Change Points by MNSCD for Different Cases

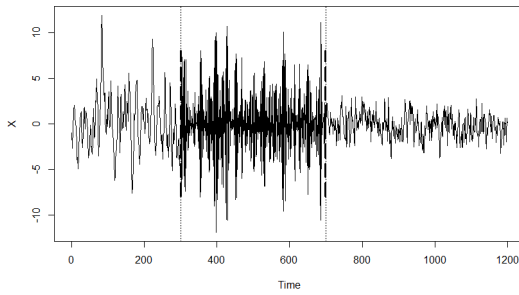
(a) AR Processes (1st Dimension)



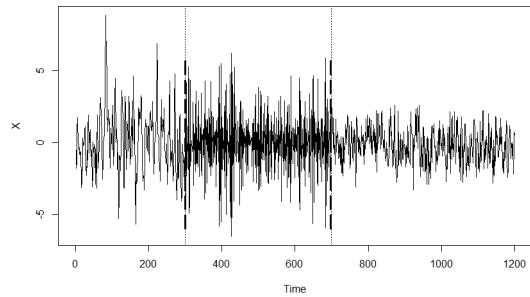
(b) AR Processes (2nd Dimension)



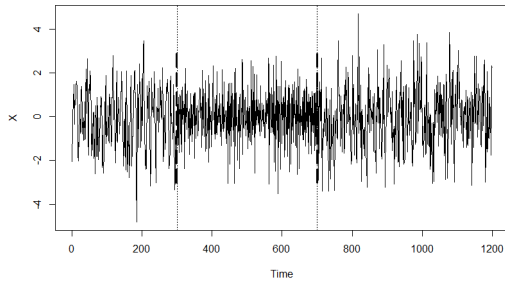
(c) ARMA Processes (1st Dimension)



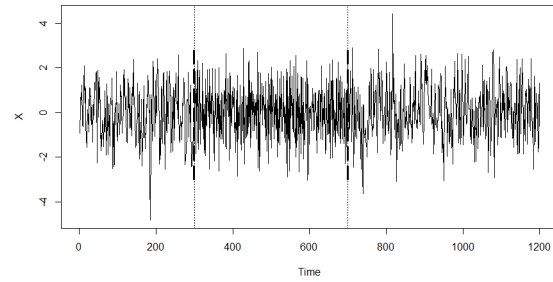
(d) ARMA Processes (2nd Dimension)



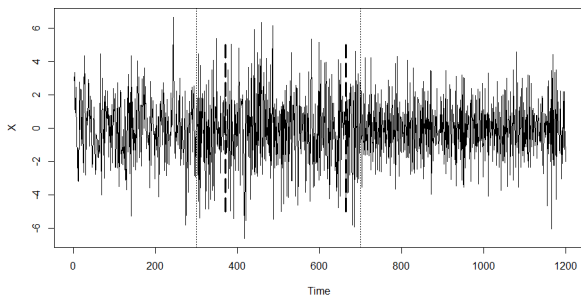
(e) Invertible MA Processes: (1st Dimension)



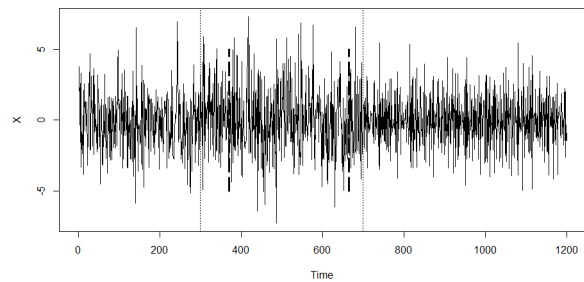
(f) Invertible MA Processes (2nd Dimension)



(g) Non-Invertible MA Processes (1st Dimension)



(h) Non-Invertible MA Processes (2nd Dimension)

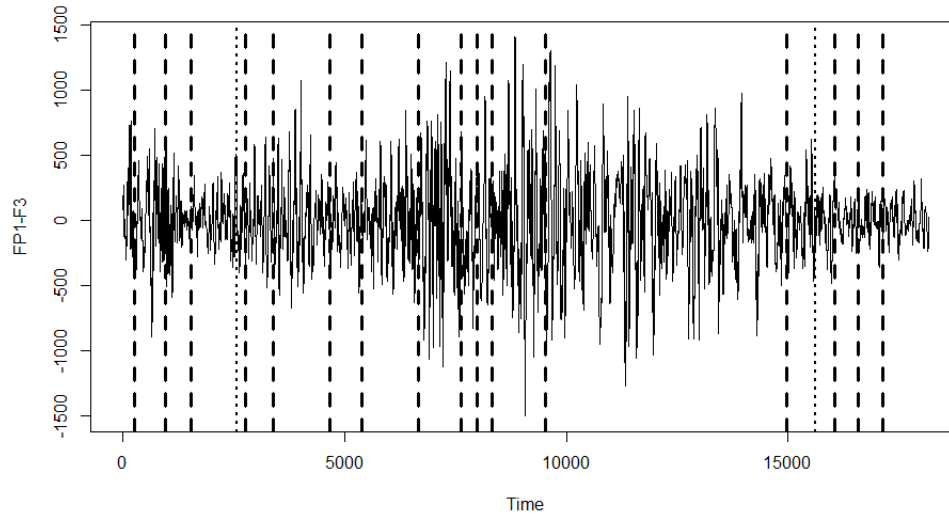


4.5.2 A Study on EEG Data

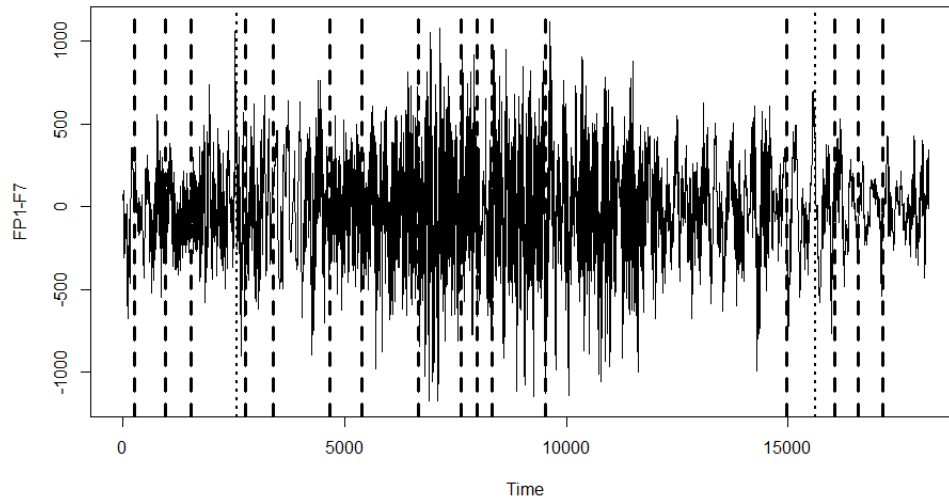
In this section, we apply MNSCD to Electrocardiogram data. Our data is from <https://www.physionet.org/pn6/chbmit/>, which is the same as the one used in Chapter 3. During the experiment, every subject was monitored for several days to record the Electrocardiogram signal from brain. There are totally 23 channels for this EEG data. In Chapter 3, we choose 2 channels to analyze the signal. Here we will apply MNSCD to all the channels to investigate the performance. Here the same data file is used as in Chapter 3, which means that MNSCD is applied to part of the data which begins at 10 seconds before seizure and ends at 10 seconds after seizure. Figures 4.5a and 4.5b show the results. Here we only plot two channels $FP1 - F3$ and $FP1 - F7$, which are used in Chapter 3. As we can see, our method gives more change points since MNSCD takes the correlations between different channels into consideration. MNSCD can capture the first change point while for the second change point, the estimation is a little biased (1.75 seconds). This is because the maximum eigenvalue could only include part of the correlation between channels.

Figure 4.5: Analysis of Different Channels of EEG Data by MNSCD

(a) Channel FP1-F3



(b) Channel FP1-F7



Chapter 5

Conclusions and Discussions

In this thesis, we discuss how to detect structural breaks when non-stationary time series can be divided into several stationary segments. In Chapter 3, we propose a change point detection method based on spectral density function for univariate non-stationary time series. We assume that non-stationary time series can be segmented into several linear processes. Then Kullback-Leibler divergence is applied to measure the discrepancy between different spectral density functions. Since our method is based on spectrum, we can apply our method on non-invertible moving average process. Also, the estimation of spectral density function is given in a nonparametric way, so the change point detection method is nonparametric. That is, we do not need to impose parametric assumption on the observations. Meanwhile, a BIC criterion is suggested to estimate the number of change points. Due to the separable structure of objective function, we use dynamic programming to find the estimators. We show the estimation consistency in theory and accuracy by simulations.

In Chapter 4, we show a change point detection method based on eigenvalues of spectral density matrices for multivariate non-stationary time series. We also assume that non-stationary time series can be segmented into several multivariate linear processes. Then Kullback-Leibler divergence is applied to measure the discrepancy between different eigenvalues of spectral density matrices due to the fact that they are Hermitian. Then we achieve

similar results compared to the univariate time series.

In this thesis, we can see that spectral density functions should be estimated before change point detection, and dynamic programming with the computation complexity of $O(N^2)$, is applied, so the whole computation process is time-consuming. We may need to study how to incorporate the estimation of spectral density function during the calculation of objective function to boost the computation. Meanwhile, we notice that the existence of higher moment is needed, while simulations show that it is not necessary, so the proofs need some improvements. Another assumption imposed on samples is that spectral density functions of different stationary segments are linearly independent. So it can be expected that the performance of our method will not be good, if different stationary segments of time series have the same structure, but only the variances are different. For example, the spectral density functions of white noise is a constant, so if the observations consists of several parts of white noise while only the variances of all the parts are different, then our method will fail. The structure of observations, which is that non-stationary time series consists of several stationary ones, restricts the range of applications. As we can see in the case study, it is hard to partition the observations into several stationary ones due to the abrupt changes of spectrums. In literature, some modified spectrums are proposed, for example, evolutionary spectrum (Priestley 1981). So we may apply our method directly on non-stationary time series.

Bibliography

- [1] Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009), “Break Detection in the Covariance Structure of Multivariate Time Series Models,” *The Annals of Statistics*, 37, 4046-4087.
- [2] Bellman, R.E., and Dreyfus, S.E. (1962) *Applied Dynamic Programming*, Princeton University Press, Princeton.
- [3] Brillinger, D. R. (1975), *Time Series: Data Analysis and Theory*, Holt, Rinehart and Winston, Inc.
- [4] Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods (2nd Edition)*, Springer-Verlag.
- [5] Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006), “Structural Break Estimation for Nonstationary Time Series Models,” *Journal of the American Statistical Association*, 101(473), 223-239.
- [6] Fan, J. and Yao, Q., (2003), *Nonlinear Time Series*, Springer.
- [7] Fejér, L. (1910), “Lebesguesche Konstanten und divergente Fourierreihen,” *Journal für die reine und angewandte Mathematik*, 138, 22-53.
- [8] Galeano, P., and Peña, D. (2007), “Covariance Changes Detection in Multivariate Time Series,” *Journal of Statistical Planning and Inference*, 137, 194-211.

- [9] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C-K., Stanley, H. E. (2000), “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals”, *Circulation* 101(23), e215-e220.
- [10] Hannan, E. J. (1970), *Multiple Time Series*, New York: Wiley.
- [11] Hawkins, D. M. (2001), “Fitting Multiple Change-point Models to Data,” *Computational Statistics & Data Analysis*, 37(3), 323-341.
- [12] Killick, R., Fearnhead, P., and Eckley, I. A. (2012), “Optimal Detection of Change-points With a Linear Computational Cost,” *Journals of the American Statistical Association*, 107(500), 1590- 1598.
- [13] Kirch, C., Muhsal, B., and Ombao, H. (2015), “Detection of Changes in Multivariate Time Series With Application to EEG Data,” *Journal of the American Statistical Association*, 110, 1197–1216.
- [14] Kitagawa, G., and Akaike, H. (1978), “A Procedure for the Modeling of Non-Stationary Time Series,” *Annals of the Institute of Statistical Mathematics*, 30, 351-363.
- [15] Korkas, K. K., and Fryzlewicz, P. (2017), “Multiple Change-Point Detection for Non-Stationary Time Series Using Wild Binary Segmentation,” *Statistica Sinica*, 27(1), 287-311.
- [16] Kullback, S., and Leibler, R.A. (1951) “On information and sufficiency,” *Annals of Mathematical Statistics*, 22(1), 79-86.
- [17] Lavielle, M., and Ludeña, C. (2000), “The Multiple Change-Points Problems for the Spectral Distribution,” *Bernoulli*, 6(5), 845-869.
- [18] Lavielle, M., and Teyssière, G. (2006), “Detection of Multiple Change-Points in Multivariate Time Series,” *Lithuanian Mathematical Journal*, 46, 287-306.

- [19] Lütkepohl, H. (2006), *New Introduction to Multiple Time Series Analysis*, Springer.
- [20] Matteson, D. S., and James, N. A. (2014), “A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data,” *Journal of the American Statistical Association*, 109, 334–345.
- [21] Nason, G. (2016), *Package “wavethresh”*, retrieved from <https://cran.r-project.org/web/packages/wavethresh/wavethresh.pdf>.
- [22] Nason, G., Sapatinas, T., and Sawczenko, Andrew. (2001), “Wavelet Packet Modelling of Infant Sleep State Using hHeart Rate Data,” *Sankhya B*, 63, 199–217.
- [23] Nason, G., Von Sachs, R., and Kroisandt, G. (2000), “Wavelet Processes and Adaptive Estimation of the Evolutionary Wavelet Spectrum”. *Journal of the Royal Statistical Society: Series B*, 62, 271-292.
- [24] Ombao, H. C., Raz, J. A., Von Sachs, R., and Malow, B. A. (2001), “Automatic Statistical Analysis of Bivariate Nonstationary Time Series,” *Journal of the American Statistical Association*, 96(454), 543-560.
- [25] Parzen, E. (1982), “Maximum Entropy Interpretation of Autoregressive Spectral Densities,” *Statistics and Probability*, 1, 2-6.
- [26] Parzen, E. (1983), “Time Series ARMA Model Identification by Estimating Information,” Technical Report, Texas A&M Research Foundation Project No. 4226T, N-37.
- [27] Priestley, M. B. (1981), *Spectral Analysis and Time Series*, Academic Press.
- [28] Rao, T. S. (1993), *Developments in Time Series Analysis*, Chapman and Hall/CRC.
- [29] Rosenblatt, M. (1974), *Random Processes*, Springer-Verlag New York.
- [30] Shore, J. E. (1981), “Minimum Cross-Entropy Spectral Analysis,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(2), 230-237.

- [31] Tao, T. (2012), *Topics in Random Matrix Theory*, American Mathematical Society.
- [32] Wilkinson, J. H. (1965), *The Algebraic Eigenvalue Problem*, Oxford: Oxford University Press.
- [33] Woodroffe, M. B., and Van Ness, J. W. (1967), “The Maximum Deviation of Sample Spectral Densities,” *The Annals of Mathematical Statistics*, 38(5), 1558-1569.
- [34] Yao, Y.-C. (1988), “Estimating the Number of Change-Points Via Schwarz’ Criterion,” *Statistics & Probability Letters*, 6(3), 181-189.
- [35] Yao, Y.-C., and Au, S. T. (1989), “Least-Squares Estimation of a Step Function,” *Sankhyā*, 51(3), 370-381.
- [36] Zou, C., Yin, G., Feng, L., and Wang, Z. (2014), Nonparametric Maximum Likelihood Approach to Multiple Change-Point Problems. *The Annals of Statistics*, 42(3), 970-1002.

Appendix A

Proofs in the Thesis

A.1 Proofs in Chapter 3

In the appendix, B_1 to B_8 are appropriate constant and $B_{7\epsilon}$, $B_{8\epsilon}$ are constant with respect to ϵ .

Lemma A.1. *Suppose $X_t = \sum_{j=-\infty}^{+\infty} a_j \xi_{t-j}$, $1 \leq t \leq N$, where $\sum_j |a_j| < +\infty$. $\xi_j \stackrel{\text{i.i.d}}{\sim} (0, \sigma^2)$. Denote $f(\lambda)$ as the spectral density function of X_t . \hat{f} is the estimated spectral density function. Then under Assumptions 1-6,*

$$\begin{aligned} \left| \frac{\hat{f}(\lambda) - f(\lambda)}{f(\lambda)} \right|^{2q} &= O_p\left(\frac{m^{2q}}{N^q}\right) \\ \max_{\lambda_i \in \Lambda} \left| \frac{\hat{f}(\lambda_i) - f(\lambda_i)}{f(\lambda_i)} \right|^{2q} &= O_p\left(\frac{m^{2q}}{N^{q-1}}\right) \end{aligned}$$

Proof: Following the method of Woodroffe and Van Ness (1967), we separate our proofs into 3 parts.

Part 1: we show that for given $\lambda \in [-\pi, \pi]$, $E|g_N(\lambda) - 1|^{2q} = O_p\left(\frac{m^{2q}}{N^q}\right)$. Here $g_N(\lambda)$ is the smoothed spectral density estimation for $\{\xi_1, \dots, \xi_N\}$, ξ

$$|g_N(\lambda) - 1|^{2q} = \left(m \int W(m(u - \lambda)) \frac{1}{N} \left| \sum_{t=1}^N \xi_t e^{-itu} \right|^2 du - \sigma^2 \right)^{2q}$$

$$\begin{aligned}
&= \frac{1}{N^{2q}} \left(\int W(u) \left(\left(\sum_{t=1}^N \xi_t e^{-it(um^{-1}+\lambda)} \right)^2 - N\sigma^2 \right) du \right)^{2q} \\
&= \frac{1}{N^{2q}} \left(2 \sum_{t=1}^N \sum_{v=1}^{N-t} \xi_t \xi_{t+v} \cos(v\lambda) w(vm^{-1}) + \sum_{t=1}^N \xi_t^2 - N\sigma^2 \right)^{2q} \\
&= \frac{1}{N^{2q}} (Z_N(\lambda) + r_N(\lambda) + r_N)^{2q},
\end{aligned}$$

where $Z_N(\lambda) = \sum_{t=1}^N Z_{N,t}(\lambda)$,

$$\begin{aligned}
Z_{N,t}(\lambda) &= 2 \sum_{v=1}^{m-1} \xi_t \xi_{t+v} w(vm^{-1}) \cos(v\lambda), \\
r_N(\lambda) &= 2 \sum_{t=N-m+2}^N \sum_{v=N-t+1}^{m-1} \xi_t \xi_{t+v} w(vm^{-1}) \cos(v\lambda), \\
r_N &= \sum_{t=1}^N (\xi_t^2 - \sigma^2).
\end{aligned}$$

Since for any real number b_1, \dots, b_p , $(b_1 + \dots + b_p)^{2q} \leq 2^{2pq}(b_1^{2q} + \dots + b_p^{2q})$. So we only need to prove that $(Z_N(\lambda))^{2q}$, $(r_N(\lambda))^{2q}$, $(r_N)^{2q}$ are $O_p(N^q m^{2q})$. By Markov's inequality, it suffices to show that $E(Z_N(\lambda))^{2q}$, $E(r_N(\lambda))^{2q}$, $E(r_N)^{2q}$ are $O(N^q m^{2q})$.

After expanding $(r_N(\lambda))^{2q}$, we have

$$\begin{aligned}
(r_N(\lambda))^{2q} &= \sum_{p=1}^{2q} \sum_{t_1+\dots+t_p=2q} \sum_{j_1 \dots j_p} \sum_{v_1 \dots v_p} (\xi_{j_1} \xi_{j_1+v_1})^{t_1} \dots (\xi_{j_p} \xi_{j_p+v_p})^{t_p} \\
&\quad w^{t_1}(v_1 m^{-1}) \dots w^{t_p}(v_p m^{-1}) \cos^{t_1}(v_1 \lambda) \dots \cos^{t_p}(v_p \lambda),
\end{aligned}$$

where $\{j_1, \dots, j_p\} \subset \{N-m+2, \dots, N\}$. If $t_k \geq 2$ for any k , then expectation of $(\xi_{j_1} \xi_{j_1+v_1})^{t_1} \dots (\xi_{j_p} \xi_{j_p+v_p})^{t_p}$ is not zero. If one of t_k is 1, for example, $t_{k_0} = 1$, since $j_k \leq N < j_k + v_k$ for any k , it is impossible to find $\xi_{j_{k_0}}$ in $\{\xi_{j_1+v_1}, \dots, \xi_{j_p+v_p}\}$, so the expectation would be zero. To sum up, for all non-zero terms in $E(r_N(\lambda))^{2q}$, t_k should be greater than 2. Since for any $s \leq 2q$, $\sum_{k=1}^s t_k = 2q$, $w^j(u) \cos^j(u)$, $E\xi_1^{t_1} \dots \xi_s^{t_s}$ can be bounded by a real number, denoted by B_1 , so we only need to count the number of non-zero terms. When $t_k = 2$ for any k , then $p = q$, and the number of non-zero terms are no more than $(C_{2q}^2 C_{2q-2}^2 \dots C_2^2 C_m^q)^2 = O(m^{2q})$. When at least one of $t_k > 2$, say $t_{k_0} = 3$, the number of

non-zero terms is no more than $(C_{2q}^2 C_{2q-2}^2 \cdots C_4^3 C_m^{q-1})^2 = O(m^{2q-2})$. So

$$E(r_N(\lambda))^{2q} \leq B_1 m^{2q}.$$

For

$$E(r_N)^{2q} = E\left(\sum_{t=1}^N (\xi_t^2 - \sigma^2)\right)^{2q} = \sum_{p=1}^{2q} \sum_{t_1+\dots+t_p=2q} \sum_{j_1 \cdots j_p} (\xi_{j_1}^2 - \sigma^2)^{t_1} \cdots (\xi_{j_p}^2 - \sigma^2)^{t_p} \leq B_1 m^{2q},$$

we can see that $t_k \geq 2$ for any k in all non-zero terms, so following the discussions above, we have

$$E\left(\sum_{t=1}^N (\xi_t^2 - \sigma^2)\right)^{2q} \leq B_2 N^q.$$

By far, we have proven that $(r_N(\lambda))^{2q}$ and r_N^{2q} are $O_p(N^q m^{2q})$.

$$\begin{aligned} (Z_N(\lambda))^{2q} &= \left(\sum_{t=1}^N Z_{N,t}(\lambda) \right)^{2q} \\ &= \left(2 \sum_{t=1}^N \sum_{v=1}^{m-1} \xi_t \xi_{t+v} w(vm^{-1}) \cos(v\lambda) \right)^{2q} \\ &= \sum_{p=1}^{2q} \sum_{t_1+\dots+t_p=2q} \sum_{j_1 \cdots j_p} Z_{N,j_1}^{t_1}(\lambda) \cdots Z_{N,j_p}^{t_p}(\lambda). \end{aligned}$$

First, when $t_k \geq 2$ for any k , $p \leq q$, the number of terms after expanding $Z_{N,j_1}^{t_1}$ is $(m-1)^{t_1}$. So for fixed t_1, \dots, t_p , $Z_{N,j_1}^{t_1} \cdots Z_{N,j_p}^{t_p}$ contains no more than $(m-1)^{t_1+\dots+t_p} = (m-1)^{2q}$ terms. What is more, for any fixed p , the number for all possible j_1, \dots, j_p , of which the power satisfy $t_1 + \dots + t_p = 2q$, is no more than $p! C_N^p$. So there are no more than $N^q m^{2q}$ terms, which means that $E \sum_{p=1}^q \sum_{t_1+\dots+t_p=2q} \sum_{j_1 \cdots j_p} Z_{N,j_1}^{t_1}(\lambda) \cdots Z_{N,j_p}^{t_p}(\lambda) \leq B_1 N^q m^{2q}$.

When $p \leq q$ and some of t_k are equal to 1, we assume that $t_{k_{s_1}} = \dots = t_{k_{s_c}} = 1$. For $t_{k_{s_1}}, j_{k_{s_1}}$ should satisfy $j_{k_{s_1}} - j_{k_{s_1}-1} < m$, that is, $\xi_{j_{k_{s_1}}} \in \{\xi_{j_{k_{s_1}-1}+1}, \dots, \xi_{j_{k_{s_1}-1}+m-1}\}$. If not, since $\xi_{j_{k_{s_1}}}$ is independent of $(Z_{N,j_1})^{t_1}, \dots, (Z_{N,j_{k_{s_1}-1}})^{t_{j_{k_{s_1}-1}}}$, $\left(\xi_{j_{k_{s_1}+1}} + \dots + \xi_{j_{k_{s_1}+m-1}}\right)$, $\left(Z_{N,j_{k_{s_1}+1}}\right)^{t_{j_{k_{s_1}+1}}}, \dots, (Z_{N,j_p})^{t_p}$, we have

$$E(Z_{N,j_1})^{t_1} \cdots (Z_{N,j_p})^{t_p} = 0.$$

So the number of all non-zero terms is no more than $C_N^{p-c}m^c$. For $(Z_{N,j_{k_{s_1}-1}})^{t_{j_{k_{s_1}-1}}}$, there are $(m-2)^{t_{j_{k_{s_1}-1}}}$ terms which contain $\xi_{j_{k_{s_1}}}$. So there are $(m-1)^{t_{j_{k_{s_1}-1}}} - (m-2)^{t_{j_{k_{s_1}-1}}} = O((m-2)^{t_{j_{k_{s_1}-1}}})$ terms which do not contain $\xi_{j_{k_{s_1}}}$. Then the number of non-zero terms is $O(N^{p-c}m^c m^{t_1+\dots+t_p-c}) \leq O(N^q m^{2q})$. For $\xi_{j_{k_{s_1}}}$, if $Z_{N,j_{k_1}}^{t_{k_1}}, \dots, Z_{N,j_{k_d}}^{t_{k_d}}$ contain $\xi_{j_{k_{s_1}}}$, then it is easy to see that there are totally no more than $m^{t_{k_1}+\dots+t_{k_d}-1}$ terms which do not contain $\xi_{j_{k_{s_1}}}$. However, now for all j_{k_1}, \dots, j_{k_d} , $|j_{k_v} - j_{k_{s_1}}| \leq m-1$. So the number of all possible non-zero terms is $O(N^{p-c-d+1}m^{c+d}m^{t_1+\dots+t_p-c}) \leq O(N^q m^{2q})$. Hence $E(Z_{N,j_1})^{t_1} \dots (Z_{N,j_p})^{t_p} \leq B_1 N^q m^{2q}$.

When $p > q$, there are at least $p-q$ of $t_j = 1$, so following the discussions above, $E(Z_{N,j_1})^{t_1} \dots (Z_{N,j_p})^{t_p} \leq B_1 N^q m^{2q}$. Therefore summarizing discussions above, $E(Z_N(\lambda))^{2q} = O_p(\frac{m^{2q}}{N^q})$.

Part 2, we prove that $\left| \frac{f(\lambda) - Ef_N(\lambda)}{f(\lambda)} \right|^{2q} = O(\frac{m^{2q}}{N^q})$.

$$\begin{aligned} \left| \frac{f(\lambda) - Ef_N(\lambda)}{f(\lambda)} \right|^{2q} &\leq B \left| f(\lambda) - m \int W(m(\lambda - u))f(u)du \right|^{2q} \\ &\quad + B \left| m \int W(m(\lambda - u))(f(u) - EI_N(u))du \right|^{2q} \end{aligned}$$

And

$$\begin{aligned} \left| f(\lambda) - m \int W(m(\lambda - u))f(u)du \right|^{2q} &= \left| \int (f(\lambda) - f(\lambda - um^{-1}))W(u)du \right|^{2q} \\ &\leq \left(\int |f(\lambda) - f(\lambda - um^{-1})| W(u)du \right)^{2q} \\ &\leq B_f m^{-2q} \left| \int |u|W(u)du \right|^{2q} \end{aligned}$$

Since $\alpha \geq \frac{1}{4}$, we have $\lim_{N \rightarrow \infty} \frac{1}{m^{2q}} / \frac{m^{2q}}{N^q} < +\infty$.

And

$$\left| m \int W(m(\lambda - u))(f(u) - EI_N(u))du \right|^{2q} = \left| \int W(u)(f(\lambda - um^{-1}) - EI_N(\lambda - um^{-1}))du \right|^{2q}$$

$$\leq \left(\int W(u) |f(\lambda - um^{-1}) - EI_N(\lambda - um^{-1})| du \right)^{2q}$$

Since by Woodrooffe and Van Ness (1967),

$$\max_{|\lambda| \leq \pi} |f(\lambda) - EI_N(\lambda)| \leq B_{pe} \log N/N,$$

where B_{pe} is a constant. So

$$\begin{aligned} \left| m \int W(m(\lambda - u))(f(u) - EI_N(u)) du \right|^{2q} &\leq \left| \int W(u) \frac{B_{pe} \log N}{N} du \right|^{2q} \\ &\leq B_{pe}^{2q} \frac{(\log N)^{2q}}{N^{2q}}. \end{aligned}$$

And $\lim_{N \rightarrow \infty} \frac{(\log N)^{2q}}{N^{2q}} / \frac{m^{2q}}{N^q} \rightarrow 0$. So we complete Part 2.

Part 3: we show that $E \left| \frac{f_N(\lambda) - Ef_N(\lambda)}{f(\lambda)} - (g_N(\lambda) - \sigma^2) \right|^{2q} = O_p\left(\frac{m^{2q}}{N^q}\right)$.

$$\begin{aligned} &\left| \frac{f_N(\lambda) - Ef_N(\lambda)}{f(\lambda)} - (g_N(\lambda) - \sigma^2) \right|^{2q} \\ &= \left| \frac{f_N(\lambda) - Ef_N(\lambda) - f(\lambda)(g_N(\lambda) - \sigma^2)}{f(\lambda)} \right|^{2q} \\ &\leq B_f |f_N(\lambda) - Ef_N(\lambda) - f(\lambda)(g_N(\lambda) - \sigma^2)|^{2q} \\ &= B_f \left| m \int W(m(\lambda - u)) (I_N(u) - EI_N(u) - f(u)(J_N(u) - EJ_N(u))) du \right. \\ &\quad \left. + m \int W(m(\lambda - u)) (f(\lambda) - f(u)) (J_N(u) - EJ_N(u)) du \right|^{2q} \\ &\leq B_3 \left| m \int W(m(\lambda - u)) (I_N(u) - EI_N(u) - f(u)(J_N(u) - EJ_N(u))) du \right|^{2q} \\ &\quad + B_3 \left| m \int W(m(\lambda - u)) (f(\lambda) - f(u)) (J_N(u) - EJ_N(u)) du \right|^{2q} \\ &= R_1(\lambda) + R_2(\lambda), \end{aligned}$$

where $J_N(u)$ is the periodogram for ξ_1, \dots, ξ_N , B_3 is a constant. After some manipulations

(see Woodroffe and Van Ness 1967, Grenander and Rosenblatt 1957).

$$\begin{aligned}
ER_1(\lambda) &= E \frac{1}{N^{2q}} \left| \sum_{r,s=-\infty}^{+\infty} a_r a_s d_{rs}(\lambda) \right|^{2q} \\
&\leq \frac{1}{N^{2q}} \sum_{p=1}^{2q} \sum_{t_1+\dots+t_p} \sum_{\substack{r_1, s_1 \\ \dots \\ r_p, s_p}} E \left(|a_{r_1} a_{s_1} d_{r_1 s_1}|^{t_1} \cdots |a_{r_p} a_{s_p} d_{r_p s_p}|^{t_p} \right) \\
&= \frac{1}{N^{2q}} \sum_{p=1}^{2q} \sum_{t_1+\dots+t_p} \sum_{\substack{r_1, s_1 \\ \dots \\ r_p, s_p}} |a_{r_1} a_{s_1}|^{t_1} \cdots |a_{r_p} a_{s_p}|^{t_p} E \left(|d_{r_1 s_1}|^{t_1} \cdots |d_{r_p s_p}|^{t_p} \right) \\
&\leq \frac{1}{N^{2q}} \sum_{p=1}^{2q} \sum_{t_1+\dots+t_p} \sum_{\substack{r_1, s_1 \\ \dots \\ r_p, s_p}} |a_{r_1} a_{s_1}|^{t_1} \cdots |a_{r_p} a_{s_p}|^{t_p} \left(E |d_{r_1 s_1}|^{2t_1} \cdots |d_{r_p s_p}|^{2t_p} \right)^{\frac{1}{2}}
\end{aligned}$$

Now let us focus on $|d_{rs}|^{4q}$, that is $r_1 = \dots = r_p$, $s_1 = \dots = s_p$, and denote $R_\xi(v)$ as the autocovariance function of ξ_j .

$$\begin{aligned}
|d_{rs}|^{4q} &= \left| \sum_{v_1=1-r}^{N-r} \sum_{v_2=1-s}^{N-s} w((v_1 - v_2 + r - s)m^{-1}) e^{-i(v_1 - v_2 + r - s)\lambda} (\xi_{v_1} \xi_{v_2} - R_\xi(v_1 - v_2)) \right. \\
&\quad \left. - \sum_{v_1=1}^N \sum_{v_2=1}^N w((v_1 - v_2 + r - s)m^{-1}) e^{-i(v_1 - v_2 + r - s)\lambda} (\xi_{v_1} \xi_{v_2} - R_\xi(v_1 - v_2)) \right|^{4q} \\
&= \left| \sum_{v_1=1-r}^{N-r} \sum_{v_2=1-s}^{N-s} w((v_1 - v_2 + r - s)m^{-1}) \cos((v_1 - v_2 + r - s)\lambda) (\xi_{v_1} \xi_{v_2} - R_\xi(v_1 - v_2)) \right. \\
&\quad + i \sum_{v_1=1-r}^{N-r} \sum_{v_2=1-s}^{N-s} w((v_1 - v_2 + r - s)m^{-1}) \sin((v_1 - v_2 + r - s)\lambda) (\xi_{v_1} \xi_{v_2} - R_\xi(v_1 - v_2)) \\
&\quad \left. - \sum_{v_1=1}^N \sum_{v_2=1}^N w((v_1 - v_2 + r - s)m^{-1}) \cos((v_1 - v_2 + r - s)\lambda) (\xi_{v_1} \xi_{v_2} - R_\xi(v_1 - v_2)) \right. \\
&\quad \left. + i \sum_{v_1=1}^N \sum_{v_2=1}^N w((v_1 - v_2 + r - s)m^{-1}) \sin((v_1 - v_2 + r - s)\lambda) (\xi_{v_1} \xi_{v_2} - R_\xi(v_1 - v_2)) \right|^{4q} \\
&= |Q_{rs} + iT_{rs}|^{4q} \leq 2^{2q} (Q_{rs}^{4q} + T_{rs}^{4q})
\end{aligned}$$

Again, we only need to prove that the expectation of Q_{rs}^{4q} and T_{rs}^{4q} are $O(N^{2q}m^{4q})$. And

$$\begin{aligned}
EQ_{rs}^{4q} &= E \left(\sum_{v_1=1-r}^{N-r} \sum_{v_2=1-s, v_2 \neq v_1}^{N-s} w((v_1 - v_2 + r - s)m^{-1}) \cos((v_1 - v_2 + r - s)\lambda) \xi_{v_1} \xi_{v_2} \right. \\
&\quad - \sum_{v_1=1}^N \sum_{v_2=1, v_2 \neq v_1}^N w((v_1 - v_2 + r - s)m^{-1}) \cos((v_1 - v_2 + r - s)\lambda) \xi_{v_1} \xi_{v_2} \\
&\quad \left. + \sum_{v_1=\max(1-r, 1-s)}^{\min(N-r, N-s)} (\xi_{v_1}^2 - 1) - \sum_{v_1=1}^N (\xi_{v_1}^2 - 1) \right)^{4q} \\
&\leq 2^{16q} \left(E \left(\sum_{v_1=1-r}^{N-r} \sum_{v_2=1-s, v_2 \neq v_1}^{N-s} w((v_1 - v_2 + r - s)m^{-1}) \cos((v_1 - v_2 + r - s)\lambda) \xi_{v_1} \xi_{v_2} \right) \right)^{4q} \\
&\quad + E \left(\sum_{v_1=\max(1-r, 1-s)}^{N-r, N-s} (\xi_{v_1}^2 - \sigma^2) \right)^{4q} \\
&\quad + E \left(\sum_{v_1=1}^N \sum_{v_2=1, v_2 \neq v_1}^N w((v_1 - v_2 + r - s)m^{-1}) \cos((v_1 - v_2 + r - s)\lambda) \xi_{v_1} \xi_{v_2} \right)^{4q} \\
&\quad + E \left(\sum_{v_1=1}^N (\xi_{v_1}^2 - \sigma^2) \right)^{4q}
\end{aligned}$$

For those four terms in the last inequality above, following the proofs of Part 1, we can show that each of them is $O(N^{2q}m^{4q})$. Similarly to the discussions above, $E(T_{rs}(\lambda))^{4q} = O(N^{2q}m^{4q})$. So we have $E|d_{rs}(\lambda)|^{4q} \leq O(N^{2q}m^{4q})$. When some of (r_k, s_k) are different from each other, for example, $(r_1, s_1) \neq (r_k, s_k)$ for $k \geq 2$, then in $E|d_{r_1 s_1}(\lambda)|^{2t_1} \cdots |d_{r_p s_p}(\lambda)|^{2t_p}$, the number of non-zero terms should be less than $E|d_{rs}(\lambda)|^{4q}$ because some ξ_j in $d_{r_1 s_1}(\lambda)$ are not in $d_{r_k s_k}(\lambda)$. And now, when the power of ξ_j is 1 in the expansion of $(d_{r_1 s_1}(\lambda))^{2t_1}$, we can not find any ξ_j in $(d_{r_k s_k}(\lambda))^{2t_k}$, so the expectations of these terms are 0. So,

$$ER_1(\lambda) \leq \frac{BN^q m^{2q}}{N^{2q}} \sum_{p=1}^{2q} \sum_{t_1 + \cdots + t_p} \sum_{\substack{r_1, s_1 \\ \vdots \\ r_p, s_p}} |a_{r_1} a_{s_1}|^{t_1} \cdots |a_{r_p} a_{s_p}|^{t_p}$$

Since $\sum_j |a_j| < +\infty$,

$$\sum_{\substack{r_1, s_1 \\ \dots \\ r_p, s_p}} |a_{r_1} a_{s_1}|^{t_1} \cdots |a_{r_p} a_{s_p}|^{t_p} < +\infty$$

for any given t_1, \dots, t_p . Since $t_1 + \dots + t_p = 2q$, $ER_1(\lambda) \leq B_1 \frac{N^q m^{2q}}{N^{2q}} = B_1 \frac{m^{2q}}{N^q}$.

$$\begin{aligned} R_2(\lambda) &= B_2 \left(\int W(u) (f(\lambda) - f(\lambda - um^{-1})) (J_N(\lambda - um^{-1}) - E(J_N(\lambda - um^{-1}))) du \right)^{2q} \\ &= B_3 \left(\int W(u)^{1-\frac{1}{2q}} W(u)^{\frac{1}{2q}} (f(\lambda) - f(\lambda - um^{-1})) \times \right. \\ &\quad \left. (J_N(\lambda - um^{-1}) - E(J_N(\lambda - um^{-1}))) du \right)^{2q} \\ &\leq B_4 \left(\int W(u)^{1-\frac{1}{2q}} W(u)^{\frac{1}{2q}} um^{-1} (J_N(\lambda - um^{-1}) - E(J_N(\lambda - um^{-1}))) du \right)^{2q} \\ &\leq B_4 \left(\int \left((W(u))^{1-\frac{1}{2q}} \right)^{\frac{2q}{2q-1}} du \right)^{2q \frac{2q-1}{2q}} \times \\ &\quad \left(\int u^{2q} m^{-2q} W(u) (J_N(\lambda - um^{-1}) - EJ_N(\lambda - um^{-1}))^{2q} du \right)^{2q \frac{1}{2q}} \\ &= B_4 m^{-2q} \left| \int u^{2q} W(u) (J_N(\lambda - um^{-1}) - EJ_N(\lambda - um^{-1}))^{2q} du \right| \\ &= B_4 m^{-2q} \left| i^{-2q} \int (iu)^{2q} W(u) (J_N(\lambda - um^{-1}) - EJ_N(\lambda - um^{-1}))^{2q} du \right| \\ &= B_4 m^{-2q} \left| \int \mathcal{F}(w^{(2q)})(u) (J_N(\lambda - um^{-1}) - EJ_N(\lambda - um^{-1}))^{2q} du \right| \\ &= B_4 m^{-2q} \int \mathcal{F}(w^{(2q)})(u) \left(\sum_{t=1}^N \sum_{v=1}^{N-t} \xi_t \xi_{t+v} e^{-i(\lambda - um^{-1})v} + \sum_{t=1}^N (\xi_t^2 - 1) \right)^{2q} du \end{aligned}$$

where the second inequality holds because of Hölder inequality, and $\mathcal{F}(w^{(2q)})(u)$ is the Fourier transformation of the $2q$ th derivative of $w(v)$. Since $w(v) = 0$ for $v \geq 1$, we have $w^{(2q)} = 0$ for $v \geq 1$, and $w^{(2q)}(v)$ is bounded. So similar to Part 1, we have $R_2(\lambda) \leq B_5 m^{-2q} \frac{m^{2q}}{N^q} = \frac{B_5}{N^q}$.

Here we complete Part 3.

Since

$$\begin{aligned} \left| \frac{f_N(v) - f(v)}{f(v)} \right|^{2q} &\leq \left| \frac{f_N(v) - Ef_N(\lambda) + Ef_N(\lambda) - f(\lambda)}{f(\lambda)} - (g_N(\lambda) - 1) + (g_N(\lambda) - 1) \right|^{2q} \\ &\leq 2^{6q} \left(\left| \frac{f_N(\lambda) - Ef_N(\lambda)}{f(\lambda)} - (g_N(\lambda) - \sigma^2) \right|^{2q} + \left| \frac{f(\lambda) - Ef_N(\lambda)}{f(\lambda)} \right|^{2q} \right) \end{aligned}$$

$$+ |g_N(\lambda) - \sigma^2|^{2q}),$$

we have

$$E \left| \frac{f_N(v) - f(v)}{f(v)} \right|^{2q} \leq B \frac{m^{2q}}{N^q}.$$

Then, by Markov inequality,

$$\begin{aligned} P \left(\left| \frac{\hat{f}(\lambda) - f(\lambda)}{f(\lambda)} \right| \geq \epsilon \frac{m}{N^{1/2}} \right) &\leq \frac{N^q}{m^{2q} \epsilon^{2q}} E \left| \frac{f_N(v) - f(v)}{f(v)} \right|^{2q} \\ &\leq \frac{N^q}{m^{2q} \epsilon^{2q}} B \frac{m^{2q}}{N^q} = \frac{B}{\epsilon^{2q}} \end{aligned}$$

$$\begin{aligned} P \left(\max_{\lambda_j} \left| \frac{f_N(\lambda_j) - f(\lambda_j)}{f(\lambda_j)} \right| \geq \epsilon \frac{m}{n^{(q-1)/(2q)}} \right) &\leq \sum_{j=1}^N P \left(\left| \frac{f_N(\lambda_j) - f(\lambda_j)}{f(\lambda_j)} \right| \geq \epsilon \frac{m}{N^{(q-1)/(2q)}} \right) \\ &\leq \sum_{j=1}^N \frac{N^{q-1}}{\epsilon^{2q} m^{2q}} B_5 \frac{m^{2q}}{N^q} = \frac{B_5}{\epsilon^{2q}}. \end{aligned}$$

So

$$P \left(\max_{1 \leq k < l \leq N} \max_{\lambda_i} \left| \frac{f_N(\lambda_i) - f(\lambda_i)}{f(\lambda_i)} \right| \geq \epsilon \frac{m}{N^{\frac{q-3}{2q}}} \right) \leq \sum_{k=1}^N \sum_{l=1}^N B_5 \frac{m^{2q} N^{q-3}}{N^{q-1} \epsilon^{2q}} = \frac{B_5 m^{2q}}{\epsilon^{2q}}.$$

Here we complete the proof of Lemma A.1.

Lemma A.2. Suppose $\forall k$, $X_t = \sum_{j=-\infty}^{+\infty} a_j(k) \epsilon_{t-j}$ satisfy Assumptions 1-8, then

$$f_0(\lambda) = \sum_{k=1}^{K+1} \frac{N_k}{N} f_k(\lambda) \tag{A.1}$$

$$\max_{\lambda_i \in \Lambda} \left| \frac{\hat{f}_0(\lambda_i) - f_0(\lambda_i)}{f_0(\lambda_i)} \right| = O_p \left(\frac{m}{N^{(q-1)/2q}} \right) \tag{A.2}$$

Proof:

$$\begin{aligned}
\hat{f}_0(\lambda) &= m \int_{-\infty}^{+\infty} W(m(u - \lambda)) \hat{I}_0(u) du \\
&= m \int_{-\infty}^{+\infty} W(m(u - \lambda)) \frac{1}{N} \left| \sum_{j=1}^N e^{-iju} X_j \right|^2 du \\
&= m \int_{-\infty}^{+\infty} W(m(u - \lambda)) \frac{1}{N} \left| \sum_{k=1}^{K+1} \sum_{j=\tau_{k-1}^0+1}^{\tau_k^0} e^{-iju} X_j \right|^2 du \\
&= m \int_{-\infty}^{+\infty} W(m(u - \lambda)) \frac{1}{N} \left(\sum_{k=1}^{K+1} \left| \sum_{j=\tau_{k-1}^0+1}^{\tau_k^0} e^{-iju} X_j \right|^2 \right. \\
&\quad \left. + \sum_{k_1 \neq k_2} \sum_{j=\tau_{k_1-1}^0+1}^{\tau_{k_1}^0} e^{-iju} X_j \sum_{j=\tau_{k_2-1}^0+1}^{\tau_{k_2}^0} e^{iju} X_j \right) du \\
&= m \int_{-\infty}^{+\infty} W(m(u - \lambda)) \frac{1}{N} \left(\sum_{k=1}^{K+1} \left| \sum_{j=1}^{N_k} e^{-iju} X_j \right|^2 \left| e^{i\tau_{k-1}^0 u} \right|^2 + \sum_{k_1 \neq k_2} \zeta_{k_1}(u) \bar{\zeta}_{k_2}(u) \right) du \\
&= \sum_{k=1}^{K+1} \hat{f}_k(\lambda) + \sum_{k_1 \neq k_2} \frac{m}{N} \int_{-\infty}^{\infty} W(m(u - \lambda)) \zeta_{k_1}(u) \bar{\zeta}_{k_2}(u) du,
\end{aligned}$$

where $\zeta_{k_1}(u) = \sum_{j=\tau_{k_1-1}^0+1}^{\tau_{k_1}^0} e^{-iju} X_j$. So

$$\begin{aligned}
&P \left(\left| \frac{\hat{f}_0(\lambda) - f_0(\lambda)}{f_0(\lambda)} \right| \geq \epsilon \frac{m}{N^{1/2}} \right) \\
&\leq P \left(B_f \left| \frac{\hat{f}_0(\lambda) - f_0(\lambda)}{f_0(\lambda)} \right| \geq \epsilon \frac{m}{N^{1/2}} \right) \\
&\leq \sum_{k=1}^{K+1} P \left(\frac{N_k}{N} \left| \hat{f}_k(\lambda) - f_k(\lambda) \right| \geq \epsilon \frac{m}{N^{1/2}(K+1)^2} \right) \\
&\quad + \sum_{k_1 \neq k_2} P \left(\left| \frac{m}{N} \int_{-\infty}^{+\infty} W(m(u - \lambda)) \zeta_{k_1}(u) \bar{\zeta}_{k_2}(u) du \right| \geq \epsilon \frac{m}{N^{1/2}(K+1)^2} \right).
\end{aligned}$$

By Lemma A.1, we only need to show that

$$\left| \frac{m}{N} \int_{-\infty}^{+\infty} W(m(u - \lambda)) \zeta_{k_1}(u) \bar{\zeta}_{k_2}(u) du \right| = O_p\left(\frac{m}{N^{1/2}}\right).$$

What is more, assume $k_1 < k_2$ without loss of generality, then we have

$$\begin{aligned} & \left| \frac{m}{N} \int_{-\infty}^{+\infty} W(m(u - \lambda)) \zeta_{k_1}(u) \bar{\zeta}_{k_2}(u) du \right| \\ &= \frac{1}{N} \left| \int W(u) (\zeta_{k_1}(um^{-1} + \lambda) \bar{\zeta}_{k_2}(um^{-1} + \lambda)) du \right| \\ &= \frac{1}{N} \left| \int W(u) \left(\sum_{j=\tau_{k_1-1}^0}^{\tau_{k_1}^0} e^{-ij(um^{-1}+\lambda)} X_j \sum_{j=\tau_{k_2-1}^0}^{\tau_{k_2}^0} e^{ij(um^{-1}+\lambda)} X_j \right) du \right| \\ &= \frac{1}{N} \left| \sum_{l=\tau_{k_2-1}^0+1-\tau_{k_1}^0}^{\tau_{k_2-1}^0+1-\tau_{k_1}^0+N_{k_1}} w(um^{-1}) e^{il\lambda} \sum_{j=\tau_{k_1}^0-(l-\tau_{k_2-1}^0+1-\tau_{k_1}^0)}^{\tau_{k_1}^0} X_j X_{j+l} \right. \\ & \quad + \sum_{l=\tau_{k_2-1}^0+1-\tau_{k_1}^0+N_{k_1}+1}^{\tau_{k_2}-\tau_{k_1}} w(um^{-1}) e^{il\lambda} \sum_{j=\tau_{k_1-1}}^{\tau_{k_1}} X_j X_{j+l} \\ & \quad \left. + \sum_{l=\tau_{k_2}-\tau_{k_1}+1}^{\tau_{k_2}-\tau_{k_1}-1} w(um^{-1}) e^{il\lambda} \sum_{j=\tau_{k_1-1}+1}^{\tau_{k_1}+1+N_{k_2}-N_{k_1}-(l-\tau_{k_2}-\tau_{k_1}+1)} X_j X_{j+l} \right|. \end{aligned}$$

Since $N_k > m$, $\forall k$, $w(u) = 0$ for $|u| \geq 1$, so if $k_2 - k_1 > 1$, $\zeta_{k_1}(u) \bar{\zeta}_{k_2}(u) = 0$. So without loss of generality, we set $k_1 = 1$, $k_2 = 1$. Next, we separate the proofs into 2 parts, as in Lemma A.1.

Part 1:

$$\begin{aligned} & \frac{1}{N} \left| m \int W(m(u - \lambda)) \zeta_1(u) \bar{\zeta}_2(u) \right| \\ &= \frac{1}{N} \left| \int W(u) \sum_{j=1}^{N_1} e^{-ij(um^{-1}+\lambda)} \xi_j \sum_{j=N_1+1}^{N_2} e^{ij(um^{-1}+\lambda)} \xi_j \right| \\ &= \frac{1}{N} \left| \sum_{l=1}^{N_1} w(lm^{-1}) e^{il\lambda} \sum_{j=N_1-l+1}^{N_1-m} \xi_j \xi_{j+l} \right|. \end{aligned}$$

So

$$\begin{aligned}
& E \frac{1}{N^{2q}} \left| \sum_{l=1}^{N_1} w(lm^{-1}) e^{il\lambda} \sum_{j=N_1-l+1}^{N_1-m} \xi_j \xi_{j+l} \right|^{2q} \\
& \leq E \left(\frac{1}{N^{2q}} 2^{2q} \left(\sum_{l=1}^m w(lm^{-1}) \cos(l\lambda) \sum_{j=N_1-l+1}^{N_1-m} \xi_j \xi_{j+l} \right)^{2q} \right. \\
& \quad \left. + \frac{1}{N^{2q}} 2^{2q} \left(\sum_{l=1}^m w(lm^{-1}) \sin(l\lambda) \sum_{j=N_1-l+1}^{N_1-m} \xi_j \xi_{j+l} \right)^{2q} \right).
\end{aligned}$$

So for two terms in the inequality above, following the proofs in Part 1 of Lemma A.1, we have

$$\begin{aligned}
& E \left(\frac{1}{N^{2q}} 2^{2q} \left(\sum_{l=1}^m w(lm^{-1}) \cos(l\lambda) \sum_{j=n_1-l+1}^{n_1-m} \xi_j \xi_{j+l} \right)^{2q} \right) = O \left(\frac{m^{3q}}{N^{2q}} \right), \\
& E \left(\frac{1}{N^{2q}} 2^{2q} \left(\sum_{l=1}^m w(lm^{-1}) \sin(l\lambda) \sum_{j=n_1-l+1}^{n_1-m} \xi_j \xi_{j+l} \right)^{2q} \right) = O \left(\frac{m^{3q}}{N^{2q}} \right).
\end{aligned}$$

Part 2: Set $f_{11}(u)$, $f_{22}(u)$ satisfying $f_1(u) = f_{11}(u)\bar{f}_{11}(u)$, $f_2(u) = f_{22}(u)\bar{f}_{22}(u)$, where \bar{f}_{22} denotes the conjugate of f_{22} . We show that $\left| \frac{m \int W(m(u-\lambda))\zeta_1(u)\bar{\zeta}_2(u)du}{f_{11}\bar{f}_{22}} - g_n(\lambda) \right|^{2q} = O_p\left(\frac{m^{2q}}{N^q}\right)$.

$$\begin{aligned}
& \left| \frac{m \int W(m(u-\lambda))\zeta_1(u)\bar{\zeta}_2(u)du}{f_{11}\bar{f}_{22}} - g_n(\lambda) \right|^{2q} \\
& \leq B_f \left| m \int W(m(u-\lambda))\zeta_1(u)\bar{\zeta}_2(u)du - f_{11}(\lambda)\bar{f}_{22}(\lambda)g_n(\lambda) \right|^{2q} \\
& \leq \left| m \int W(m(u-\lambda))(\zeta_1(u)\bar{\zeta}_2(u) - f_{11}(u)\bar{f}_{22}(u)) \right|^{2q} \\
& \quad + \left| m \int W(m(u-\lambda))(f_{11}(\lambda)\bar{f}_{22}(\lambda) - f_{11}(u)\bar{f}_{22}(u))J_1(u)J_2(u)du \right|^{2q} \\
& = R_1(\lambda) + R_2(\lambda).
\end{aligned}$$

Since $\forall k$, $f_k(\lambda)$ all satisfy uniform Lipschitz condition, then it is easy to see that $f_{11}(u)\bar{f}_{22}(u)$

also satisfies uniform Lipschitz condition. Following the proofs in Part 3 of Lemma A.1, we have

$$R_2(\lambda) = O_p\left(\frac{1}{N^q}\right).$$

Following the proofs in Part 3 of Lemma A.1 again, we have $R_1(\lambda) = \left| \sum_{r=-\infty, s=-\infty}^{+\infty} a_r(1)a_s(2)d_{rs} \right|^{2q}$, where

$$\begin{aligned} d_{rs} &= \sum_{v_1=1-r}^{\tau_1-r} \sum_{v_2=\tau_1+1-s}^{\tau_2-s} \xi_{v_1} \xi_{v_2} e^{i(v_1-v_2)(um^{-1}+\lambda)} w((v_2-v_1)m^{-1}) \\ &\quad - \sum_{v_1=1}^{\tau_1} \sum_{v_2=1}^{\tau_2} \xi_{v_1-r} \xi_{v_2-s} e^{i(v_2-v_1)(um^{-1}+\lambda)} w((v_2-v_1)m^{-1}). \end{aligned}$$

So following the proofs in Part 3 of Lemma A.1 again. we have

$$R_1(\lambda) = O_p\left(\frac{m^{2q}}{N^q}\right).$$

Here we complete the proof of Lemma A.2.

Lemma A.3. Assume Assumption 1-8, $\forall s = 1, \dots, K+1$, $\max_{\tau_{s-1}^0 \leq k < l \leq \tau_s^0} \vartheta_{kl} \sim O_p(N^{-\frac{q-3-2q\alpha}{2q}})$, when $l - k = N_{kl} \geq ml$. Here

$$\vartheta_{kl} = \frac{N_{kl}}{N} \int_{-\pi}^{\pi} \hat{f}_k^l(v) \log \left(\frac{st(\hat{f}_k)}{st(f_s)} \right) dv$$

Proof: Set $m_{kl} = N_{kl}^\alpha$, and B_1 to B_6 are all constant. By Lemma A.1, we have

$$P \left(\max_{\lambda_i} \left| \frac{\hat{f}_k^l(\lambda_i) - f(\lambda_i)}{f(\lambda_i)} \right|^{2q} > \epsilon \frac{m^{2q}}{N^{(q-3)}} \right) \leq \frac{B_5}{N^2 \epsilon^{2q}},$$

So denote $A_{kl} = \left\{ \max_{\lambda_i} \left| \frac{\hat{f}_k^l(\lambda_i) - f(\lambda_i)}{f(\lambda_i)} \right| \leq \epsilon \frac{m}{N^{(q-3)/(2q)}} \right\}$, then on set A_{kl} ,

$$\left| \frac{\hat{f}_k^l(\lambda_i) - f(\lambda_i)}{f(\lambda_i)} \right| \leq \epsilon \frac{m}{N^{(q-3)/(2q)}} \Rightarrow \max(0, 1 - \epsilon \frac{B_5 m}{N^{(q-3)/(2q)}}) \leq \frac{\hat{f}_{kl}(\lambda_i)}{f_s(\lambda_i)} \leq 1 + \epsilon \frac{B_5 m}{N^{(q-3)/(2q)}},$$

$\forall k, l, \lambda_i.$

Since $\frac{\hat{f}_{kl}(\lambda_i)}{f_s(\lambda_i)} \geq 0$,

$$\max(0, 1 - \epsilon \frac{B_5 m}{N^{(q-3)/(2q)}}) \leq \frac{\hat{f}_{kl}(\lambda_i)}{f_s(\lambda_i)} \leq 1 + \epsilon \frac{B_5 m}{N^{(q-3)/(2q)}}.$$

So on set A_{kl} , we have

$$\begin{aligned} 0 &\leq \frac{\hat{f}_{kl}(\lambda_i)}{f_s(\lambda_i)} \leq 1 + \epsilon \frac{B_5 m}{N^{(q-3)/(2q)}} \\ \Rightarrow 0 &\leq \hat{f}_{kl}(\lambda_i) \leq f_s(\lambda_i) \left(1 + \epsilon \frac{B_5 m}{N^{(q-3)/(2q)}}\right) \\ \Rightarrow 0 &\leq \hat{F}_{kl}(\pi) \leq F_s(\pi) \left(1 + \epsilon \frac{B_6 m}{N^{(q-3)/(2q)}}\right). \end{aligned}$$

Then on $\bigcap_{kl} A_{kl}$,

$$\begin{aligned} \max_{kl} \vartheta_{kl} &= \max_{kl} \frac{N_{kl}}{N} \int_{-\pi}^{\pi} \hat{f}_k^l(u) \log \left(\frac{st(\hat{f}_k)}{st(f_s)} \right) du \\ &= \max_{kl} \frac{N_{kl}}{N} \int_{-\pi}^{\pi} \hat{f}_k^l(u) \log \left(\frac{F_s(\pi)}{\hat{F}_k(\pi)} \times \frac{\hat{f}_k(u)}{f_s(u)} \right) du \\ &\leq \max_{kl} \frac{N_{kl}}{N} \int_{-\pi}^{\pi} \max_u \hat{f}_k^l(u) \log \left(\frac{F_s(\pi)}{\hat{F}_k(\pi)} \times \max_u \frac{\hat{f}_k(u)}{f_s(u)} \right) du \\ &\leq \max_{kl} \frac{N_{kl}}{N} \int_{-\pi}^{\pi} \max_u f_s(u) \left(1 + \frac{B_5 m \epsilon}{N^{(q-3)/(2q)}}\right) \\ &\quad \log \left(\left(1 + \frac{B_6 m \epsilon}{N^{(q-3)/(2q)}}\right) \left(1 + \frac{B_5 m \epsilon}{N^{(q-3)/(2q)}}\right) \right) du \\ &\stackrel{\log(1+x) \leq x}{\leq} \max_{kl} \frac{N_{kl}}{N} \int_{-\pi}^{\pi} \max_u f_s(u) \left(1 + \frac{B_5 m \epsilon}{N^{(q-3)/(2q)}}\right) \left(\frac{B_6 m \epsilon}{N^{(q-3)/(2q)}} + \frac{B_5 m \epsilon}{N^{(q-3)/(2q)}} \right) du \\ &\leq B_{7\epsilon} \frac{m}{N^{(q-3)/(2q)}} = B_{7\epsilon} N^{-\frac{q-3-2q\alpha}{2q}}. \end{aligned}$$

Here $B_{7\epsilon}$ is some constant containing ϵ . So set $B_{8\epsilon} > B_{7\epsilon}$,

$$\begin{aligned} P \left(\max_{kl} \vartheta_{kl} \geq B_{8\epsilon} N^{-\frac{q-3-2q\alpha}{2q}} \right) &= P \left(\max_{kl} \vartheta_{kl} \geq B_{8\epsilon} N^{-\frac{q-3-2q\alpha}{2q}} \mid \bigcap_{kl} A_{kl} \right) P \left(\bigcap_{kl} A_{kl} \right) \\ &\quad + P \left(\max_{kl} \vartheta_{kl} \geq B_{8\epsilon} N^{-\frac{q-3-2q\alpha}{2q}} \mid \bigcup_{kl} A_{kl}^c \right) P \left(\bigcup_{kl} A_{kl}^c \right) \\ &\leq P \left(\max_{kl} \vartheta_{kl} \geq B_{8\epsilon} N^{-\frac{q-3-2q\alpha}{2q}} \mid \bigcap_{kl} A_{kl} \right) + P \left(\bigcup_{kl} A_{kl}^c \right) \end{aligned}$$

$$\leq 0 + \frac{B_5}{\epsilon^{2q}},$$

and the last inequality can be arbitrarily small by choosing sufficiently large ϵ . Here we complete proofs of Lemma A.3.

Lemma A.4. *Assume Assumption 1-8, $\forall s = 1, \dots, K+1$, $\max_{1 \leq k < l \leq N} \vartheta_{kl} \sim O_p(N^{-\frac{q-3-2q\alpha}{2q}})$, when $l - k = N_{kl} \geq ml$. Here*

$$\vartheta_{kl} = \frac{N_{kl}}{N} \int_{-\pi}^{\pi} \hat{f}_k^l(v) \log \left(\frac{st(\hat{f}_k)}{st(f_s)} \right) dv$$

Proof: Following the proofs of Lemma A.3, this lemma can be easily obtained.

Proofs of Theorem 3.1:

Denote $A_{kl} = \left\{ \omega : \max_{\lambda_i} \left| \frac{\hat{f}_k^l(\lambda_i) - f(\lambda_i)}{f(\lambda_i)} \right| \leq \epsilon \frac{m}{N^{(q-3)/(2q)}}, \forall j = 1, \dots, K \right\}$, where ω is the event in probability space (Ω, \mathcal{F}, P) . Then $\forall \omega \in \bigcap_{kl} A_{kl}$, we prove that $\hat{\kappa}_k \rightarrow \kappa_k^0, \forall j$.

For $\hat{\kappa}_k$, since $\hat{\kappa}_k(\omega)$ is bounded, $\forall k$, then there exists $\{n_s\}$ such that $\hat{\kappa}_{n_s}(\omega) \rightarrow \kappa_k^*$ on the subsequence. It follows from Lemma A.2 and A.3, that

$$\frac{1}{N} R(\kappa_1^*, \dots, \kappa_K^*) \leq \sum_{i=1}^{K+1} (\kappa_i^* - \kappa_{i-1}^*) \int f_{\kappa_{i-1}^* \kappa_i^*}(v) \log \frac{st(f_{\kappa_{i-1}^* \kappa_i^*})}{st(f)} dv + O(N^{-\frac{q-3-2q\alpha}{2q}}),$$

where $\lambda_0^* = 0, \lambda_{K+1}^* = 1$. If $\kappa_{i-1} \leq \kappa_{j-1}^* < \kappa_k < \dots < \kappa_{i+k} < \kappa_j^*$, then by Lemma A.3, we have

$$\begin{aligned} f_{\kappa_{j-1}^* \kappa_j^*} &= \frac{\kappa_i - \kappa_{j-1}^*}{\kappa_j^* - \kappa_{j-1}^*} f_i + \frac{\kappa_{i+1} - \kappa_i}{\kappa_j^* - \kappa_{j-1}^*} f_{i+1} + \dots + \frac{\kappa_j^* - \kappa_{j+k}}{\kappa_j^* - \kappa_{j-1}^*} f_{j+k+1} \\ F_{\kappa_{j-1}^* \kappa_j^*} &= \frac{\kappa_i - \kappa_{j-1}^*}{\kappa_j^* - \kappa_{j-1}^*} F_i + \frac{\kappa_{i+1} - \kappa_i}{\kappa_j^* - \kappa_{j-1}^*} F_{i+1} + \dots + \frac{\kappa_j^* - \kappa_{j+k}}{\kappa_j^* - \kappa_{j-1}^*} F_{j+k+1}. \end{aligned}$$

Since $\frac{1}{N} R(\kappa_1, \dots, \kappa_K) = \sum_{k=1}^{K+1} \int f_k(u) \log \frac{st(f_k)}{st(f)} du$.

So

$$\begin{aligned} & \frac{1}{N} R(\lambda_{j-1}^*, \lambda_j^*) - \frac{1}{N} R(\lambda_{j-1}^*, \lambda_i) - \dots - \frac{1}{N} R(\lambda_{i+k-1}^*, \lambda_{i+k}) \\ &= (\lambda_j^* - \lambda_i^0) \int f_i(u) \log \frac{st(f_{\lambda_{j-1}^* \lambda_j^*})}{st(f_i)} du + (\lambda_{i+1}^0 - \lambda_i^0) \int f_{i+1}(u) \log \frac{st(f_{\lambda_{j-1}^* \lambda_j^*})}{st(f_{i+1})} du \end{aligned}$$

$$+\dots + (\lambda_j^* - \lambda_{i+k}^0) \int f_{i+k}(u) \log \frac{st(f_{\lambda_{j-1}^* \lambda_j^*})}{st(f_{i+k})} du + O(N^{-\frac{q-3-2q\alpha}{2q}}) < 0$$

as $N \rightarrow \infty$, since every term above is always negative. If $\kappa_{j-1} < \kappa_k^* < \kappa_{k+1}^* < \kappa_j$, then

$$\frac{1}{N}R(\hat{\lambda}_{k-1}, \hat{\lambda}_k) = (\lambda_j^* - \lambda_{j-1}^*) \int f_j(v) \log \frac{st(f_j)}{st(f)} dv + O(N^{-\frac{q-3-2q\alpha}{2q}}).$$

So we have

$$\frac{1}{N}R(\hat{\kappa}_1, \dots, \hat{\kappa}_K) = \frac{1}{N}R(\kappa_1^*, \dots, \kappa_K^*) + O(N^{-\frac{q-3-2q\alpha}{2q}}) - \frac{1}{N}R(\kappa_1^0, \dots, \kappa_K^0) < 0$$

as $N \rightarrow \infty$. This is a contradiction because $\hat{\kappa}_k$ are the maximizers of $R(\kappa_1, \dots, \kappa_K)$. Since

$$P\left(\bigcap_{k,l} A_{kl}\right) = 1 - P\left(\bigcup_{k,l} A_{kl}^c\right) \geq 1 - \sum_{l-k \geq ml}^N P(A_{kl}^c) = 1 - \frac{B_1}{\epsilon^{2q}}.$$

so $\hat{\kappa}_j \xrightarrow{p} \kappa_j^0$.

Proofs of Theorem 3.2:

If $L < K$, then there should be a change point λ_j^0 that can not estimated consistently.

Then following the proof in Theorem 3.1,

$$\begin{aligned} & -\frac{1}{N}R(\hat{\kappa}_1, \dots, \hat{\kappa}_L) + LC_N/N + \frac{1}{N}R(\hat{\kappa}_1, \dots, \hat{\kappa}_K) - KC_N/N \\ & < -c + O_p(N^{-\frac{q-3-2q\alpha}{2q}}) - (K-L)C_N/N < 0 \end{aligned}$$

as $N \rightarrow \infty$.

If $\hat{K} > K$, then still every change point λ_i should be estimated consistently. So, there should be a change point $\kappa_{j-1} < \kappa_k^* < \kappa_j$. Then by Lemma A.2 and A.3, $R(\kappa_{j-1}, \kappa_k^*, \kappa_j) - R(\kappa_{j-1}, \kappa_j) = O_p(N^{-\frac{q-3-2q\alpha}{2q}})$. So $BIC_L - BIC_K = O_p(N^{-\frac{q-3-2q\alpha}{2q}}) + (K-L)C_N/N < 0$, as $N \rightarrow \infty$. Here we complete the proofs of Theorem 3.2.

A.2 Proofs in Chapter 4

We set B_1 to B_8 , and B_{ab} to be appropriate constants. $B_{7\epsilon}$ and $B_{8\epsilon}$ are constants related to ϵ .

Lemma A.5. *Under Assumption 1-6, $\left| \max_{v_i} \max_{1 \leq j \leq p_\xi} \hat{\lambda}_j(v_i) - \max_{v_i} \max_{1 \leq j \leq p_\xi} \lambda_j(v_i) \right| = \left| \max_{v_i} \hat{\lambda}_1(v_i) - \max_{v_i} \lambda_1(v_i) \right| = O_p\left(\frac{m}{N^{2q}}\right)$.*

Proof: Recall that Σ is the covariance matrix of ξ and

$$\Gamma(v) = \sum_{j=-\infty}^{+\infty} \begin{pmatrix} g_{11}(j) & \cdots & g_{1p_\xi}(j) \\ \vdots & \ddots & \vdots \\ g_{p_x 1} & \cdots & g_{p_x p_\xi}(j) \end{pmatrix} e^{-ijv}.$$

Then we have the spectral density matrix for X_t

$$\begin{aligned} \mathbf{f}_{xx}(v) &= \frac{1}{2\pi} \Gamma(v) \Sigma \Gamma^*(v) \\ &= \frac{1}{2\pi} \left(\sum_{j_{\xi_1}=1}^{p_\xi} \sum_{j_{\xi_2}=1}^{p_\xi} \sigma_{j_{\xi_1} j_{\xi_2}} \Gamma_{a j_{\xi_1}}(v) \Gamma_{b j_{\xi_2}}^*(v) \right)_{ab}, \end{aligned}$$

where $*$ denotes the conjugate transpose. Denote $f_{ab}(v) = \sum_{j_{\xi_1}=1}^{p_\xi} \sum_{j_{\xi_2}=1}^{p_\xi} \sigma_{j_{\xi_1} j_{\xi_2}} \Gamma_{a j_{\xi_1}}(v) \Gamma_{b j_{\xi_2}}^*(v)$.

$I_{xx}(v)$ denotes the periodogram matrix, and $\hat{\mathbf{f}}_{xx}(v)$ is the smoothed estimated spectral density matrix. that is,

$$\begin{aligned} I_{xx}(v) &= \frac{1}{N} \sum_{t_1=1}^N X(t_1) e^{-it_1 v} \left(\sum_{t_2=1}^N X(t_2) e^{-it_2 v} \right)^* \\ \hat{\mathbf{f}}_{xx}(v) &= m \int I_{xx}(u) W(m(v-u)) du, \end{aligned}$$

where $W(u)$ is the spectral window. Here without loss of generality, we assume that all entries in $\hat{\mathbf{f}}_{xx}$ is smoothed by the same spectral window. By Wielandt-Hoffman theorem (Brillinger 1975, Wilkinson 1965), we have

$$|\hat{\lambda}_1(v) - \lambda_1(v)|^{2q} \leq \left(\sum_{r=1}^{p_x} |\hat{\lambda}_r(v) - \lambda_r(v)|^2 \right)^q \leq \left(\sum_{r_1=1}^{p_x} \sum_{r_2=1}^{p_x} \left| \hat{f}_{r_1 r_2}(v) - f_{r_1 r_2}(v) \right|^2 \right)^q$$

$$\leq B_1 \sum_{r_1=1}^{p_x} \sum_{r_2=1}^{p_x} |\hat{f}_{r_1 r_2}(v) - f_{r_1 r_2}(v)|^{2q},$$

where B_1 is constant. So if $E \left| \hat{f}_{r_1 r_2}(v) - f_{r_1 r_2}(v) \right|^{2q} = O\left(\frac{m^{2q}}{N^q}\right)$ for every $v \in V$, then

$$\begin{aligned} P\left(\max_{v_i} \max_r |\hat{\lambda}_r(v_i) - \lambda_r(v_i)| \geq \frac{\epsilon m}{N^{\frac{q-1}{2q}}}\right) &\leq \sum_{j=1}^N P\left(\max_r |\hat{\lambda}_r(v_j) - \lambda_r(v_j)| \geq \frac{\epsilon m}{N^{\frac{q-1}{2q}}}\right) \\ &\leq \sum_{j=1}^N P\left(|\hat{\lambda}_1(v_j) - \lambda_1(v_j)| \geq \frac{\epsilon m}{N^{\frac{q-1}{2q}}}\right) \\ &\leq \sum_{j=1}^N \frac{E|\hat{\lambda}_1(v_j) - \lambda_1(v_j)|^{2q} N^{q-1}}{\epsilon^{2q} m^{2q}} \\ &\leq \sum_{j=1}^N \frac{B_1 \sum_{r_1=1}^{p_x} \sum_{r_2=1}^{p_x} E|\hat{f}_{r_1 r_2}(v) - f_{r_1 r_2}(v)|^{2q} N^{q-1}}{\epsilon^{2q} m^{2q}} \\ &\leq \frac{B_2}{\epsilon^{2q}}, \end{aligned}$$

and B_2 is some constant. So next we separate the proofs into 2 parts.

Part 1: Denote $\mathbf{h}(v)$ as the smoothed periodogram of $\{\boldsymbol{\xi}(1), \dots, \boldsymbol{\xi}(N)\}$. The (a, b) entry of $\mathbf{h}(v)$ is denoted by h_{ab} . Then

$$\begin{aligned} &|h_{ab} - \sigma_{ab}|^{2q} \\ &= \left(\frac{1}{N} m \int W(m(v-u)) \sum_{t_1=1}^N \xi_a(t_1) e^{-it_1 u} \sum_{t_2=1}^N \xi_b(t_2) e^{it_2 u} du - \sigma_{ab} \right)^{2q} \\ &= \frac{1}{N^{2q}} \left(\int W(u) \sum_{t_1=1}^N \sum_{t_2=1}^N e^{i(t_2-t_1)(um^{-1}+v)} \xi_a(t_1) \xi_b(t_2) du - N\sigma_{ab} \right)^{2q} \\ &= \frac{1}{N^{2q}} \left(\sum_{t_1=1}^m \sum_{\substack{t_2=1-t_1 \\ t_2 \neq 0}}^{m-1} \xi_a(t_1) \xi_b(t_2) w(t_2 m^{-1}) e^{-it_2 v} + \sum_{t_1=m+1}^{N-m+1} \sum_{\substack{t_2=1-m \\ t_2 \neq 0}}^{m-1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) e^{-it_2 v} \right. \\ &\quad \left. + \sum_{t_1=N-m+2}^N \sum_{t_2=1-m}^{N-t_1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) e^{-it_2 v} + \sum_{t=1}^N (\xi_a(t) \xi_b(t) - \sigma_{ab}) \right)^{2q} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{B_3}{N^{2q}} \left(\left| \sum_{t_1=1}^m \sum_{\substack{t_2=1-t_1 \\ t_2 \neq 0}}^{m-1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) e^{-it_2 v} \right|^{2q} \right. \\
&\quad + \left| \sum_{t_1=m+1}^{N-m+1} \sum_{\substack{t_2=1-m \\ t_2 \neq 0}}^{m-1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) e^{-it_2 v} \right|^{2q} \\
&\quad + \left| \sum_{t_1=N-m+2}^N \sum_{t_2=1-m}^{N-t_1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) e^{-it_2 v} \right|^{2q} + \left| \sum_{t=1}^N (\xi_a(t) \xi_b(t) - \sigma_{ab}) \right|^{2q} \Big) \\
&\leq \frac{B_4}{N^{2q}} \left(\left| \sum_{t_1=1}^m \sum_{\substack{t_1=1-t_1 \\ t_2 \neq 0}}^{m-1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) \cos(t_2 v) \right|^{2q} \right. \\
&\quad + \left| \sum_{t_1=1}^m \sum_{\substack{t_1=1-t_1 \\ t_2 \neq 0}}^{m-1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) \sin(t_2 v) \right|^{2q} \\
&\quad + \left| \sum_{t_1=m+1}^{N-m+1} \sum_{\substack{t_1=1-m \\ t_2 \neq 0}}^{m-1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) \cos(t_2 v) \right|^{2q} \\
&\quad + \left| \sum_{t_1=m+1}^{N-m+1} \sum_{\substack{t_1=1-m \\ t_2 \neq 0}}^{m-1} \xi_a(t_1) \xi_b(t_1+t_2) w(t_2 m^{-1}) \sin(t_2 v) \right|^{2q} \\
&\quad + \left| \sum_{N-m+2}^N \sum_{t_2=1-m}^{N-t_1} \xi_a(t_1) \xi_b(t_1+t_2) \cos(t_2 v) \right|^{2q} + \left| \sum_{N-m+2}^N \sum_{t_2=1-m}^{N-t_1} \xi_a(t_1) \xi_b(t_1+t_2) \sin(t_2 v) \right|^{2q} \\
&\quad + \left(\sum_{t=1}^N (\xi_a(t) \xi_b(t) - \sigma_{ab}) \right)^{2q} \Big) \\
&= \frac{B_5}{N^{2q}} (r_1(v) + r_2(v) + Z_1(v) + Z_2(v) + r_3(v) + r_4(v) + r_5(v)).
\end{aligned}$$

Let us focus on $Z_1(v)$ first.

$$E|Z_1(v)|^{2q} = E \left| \sum_{t=-m+1}^{N-m+1} \sum_{\substack{s=1-m \\ s \neq 0}}^{m-1} \xi_a(t) \xi_b(t+s) w(sm^{-1}) \cos(sv) \right|^{2q}$$

$$\begin{aligned}
&= E \left| \sum_{t=-m+1}^{N-m+1} Z_{N,t}(v) \right|^{2q} \\
&= E \sum_{r=1}^{2q} \sum_{t_1+\dots+t_r=2q} \sum_{j_1 \dots j_r} \left(\sum_{\substack{s=1-m \\ s \neq 0}}^{m-1} \xi_a(j_1) \xi_b(j_1+s) w(sm^{-1}) \cos(sv) \right)^{t_1} \cdots \\
&\quad \left(\sum_{\substack{s=1-m \\ s \neq 0}}^{m-1} \xi_a(j_r) \xi_b(j_r+s) w(sm^{-1}) \cos(sv) \right)^{t_r}
\end{aligned}$$

When $t_i \geq 2$, the non-zero terms in $E|Z_1(v)|^{2q}$ is no more than $C_{2q}^2 \cdots C_2^2 C_{2N}^r = O(N^q m^{2q})$.

When $t_{k_1} = \cdots = t_{k_c} = 1$ for some integer c , since

$$\sum_{\substack{s=1-m \\ s \neq 0}}^{m-1} \xi_a(j_{k_1}) \xi_b(j_{k_1}+s) w(sm^{-1}) \cos(sv) = \xi_a(j_{k_1}) \left(\sum_{\substack{s=1-m \\ s \neq 0}}^{m-1} \xi_b(j_{k_1}+s) w(sm^{-1}) \cos(sv) \right),$$

and $\xi_a(j)$ and $\xi_b(k)$ are independent for $j \neq k$, so $\xi_a(j_{k_1})$ can not be found in $Z_{N,j_1}, \dots, Z_{N,j_{k_1-1}}, Z_{N,j_{k_1+1}}, \dots, Z_{N,j_r}$. That is, the expectation for $t_{k_1} = \cdots = t_{k_c} = 1$ is zero. Since $\cos(v), w(v), E\xi_{t,r_1} \cdots \xi_{t,r_{8q}}$ are all bounded, $E|Z_1(v)|^{2q} = O(N^q m^{2q})$. Similarly, $E|Z_2(v)|^{2q} = O(N^q m^{2q})$.

For $r_1(v), r_2(v), r_3(v), r_4(v)$, they are similar to $Z_1(v)$, while the only difference is that they have smaller number of terms, so following the discussions above, we have $Er_1(v) = Er_2(v) = Er_3(v) = Er_4(v) = O(m^{2q})$. And,

$$\begin{aligned}
Er_5(v) &= E \left(\sum_{t=1}^N (\xi_a(t) \xi_b(t) - \sigma_{ab}) \right)^{2q} \\
&= E \sum_{r=1}^{2q} \sum_{t_1+\dots+t_r=r} \sum_{1 \leq j_1 \dots j_r \leq m} (\xi_a(j_1) \xi_b(j_1) - \sigma_{ab})^{t_1} \cdots (\xi_a(j_r) \xi_b(j_r) - \sigma_{ab})^{t_r}.
\end{aligned}$$

Following the discussions above, $Er_5(v) = O(N^q)$. So combining all the discussions above, we have proven that $|h_{ab}(v) - \sigma_{ab}|^{2q} = O(\frac{m^{2q}}{N^q})$.

Part 2:

$$\left| \hat{f}_{ab}(v) - f_{ab}(v) \right|^{2q} = \left| \hat{f}_{ab}(v) - E\hat{f}_{ab}(v) + E\hat{f}_{ab}(v) - f_{ab}(v) \right|^{2q}$$

$$\begin{aligned}
&\leq 2^{2q} \left(\left| f_{ab}(v) - E\hat{f}_{ab}(v) \right|^{2q} + \left| \hat{f}_{ab}(v) - E\hat{f}_{ab}(v) \right|^{2q} \right) \\
&= U_1(v) + U_2(v).
\end{aligned}$$

And

$$\begin{aligned}
U_2(v) &= \left| m \int W(m(v-u)) \frac{1}{N} \left(\sum_{t_1=1}^N X_a(t_1) e^{-it_1 u} \sum_{t_2=1}^N X_b(t_2) e^{it_2 u} \right) du \right. \\
&\quad \left. - m \int W(m(v-u)) \frac{1}{N} E \left(\sum_{t_1=1}^N X_a(t_1) e^{-it_1 u} \sum_{t_2=1}^N X_b(t_2) e^{it_2 u} \right) du \right|^{2q} \\
&= \left| m \int W(m(v-u)) \frac{1}{N} \left(\sum_{t_1=1}^N e^{-it_1 u} \sum_{j_1=-\infty}^{+\infty} \sum_{j_{\xi_1}=1}^{p_{\xi}} g_{aj_{\xi_1}}(j_1) \xi_{j_{\xi_1}}(t_1 - j_1) \right. \right. \\
&\quad \left. \left. \sum_{t_2=1}^N e^{it_2 u} \sum_{j_2=-\infty}^{+\infty} \sum_{j_{\xi_2}=1}^{p_{\xi}} g_{aj_{\xi_2}}(j_2) \xi_{j_{\xi_2}}(t_2 - j_2) du \right) \right. \\
&\quad \left. - m \int W(m(v-u)) \frac{1}{N} E \left(\sum_{t_1=1}^N e^{-it_1 u} \sum_{j_1=-\infty}^{+\infty} \sum_{j_{\xi_1}=1}^{p_{\xi}} g_{aj_{\xi_1}}(j_1) \xi_{j_{\xi_1}}(t_1 - j_1) \right. \right. \\
&\quad \left. \left. \sum_{t_2=1}^N e^{it_2 u} \sum_{j_2=-\infty}^{+\infty} \sum_{j_{\xi_2}=1}^{p_{\xi}} g_{aj_{\xi_2}}(j_2) \xi_{j_{\xi_2}}(t_2 - j_2) \right) du \right|^{2q} \\
&\leq \sum_{j_{\xi_1}=1}^{p_{\xi}} \sum_{j_{\xi_2}=1}^{p_{\xi}} \frac{B_6}{N^{2q}} \left| m \int W(m(v-u)) \sum_{j_1=-\infty}^{+\infty} \sum_{j_2=-\infty}^{+\infty} \sum_{t_1=1}^N \sum_{t_2=1}^N e^{i(t_2-t_1)u} \right. \\
&\quad \left. \left(g_{aj_{\xi_1}}(j_1) g_{bj_{\xi_2}}(j_2) \xi_{j_{\xi_1}} \xi_{j_{\xi_2}} - E g_{aj_{\xi_1}}(j_1) g_{bj_{\xi_2}}(j_2) \xi_{j_{\xi_1}} \xi_{j_{\xi_2}} \right) du \right|^{2q} \\
&= \sum_{j_{\xi_1}=1}^{p_{\xi}} \sum_{j_{\xi_2}=1}^{p_{\xi}} R_{j_{\xi_1} j_{\xi_2}}(v)
\end{aligned}$$

So, we only to prove $\forall j_{\xi_1}, j_{\xi_2}$,

$$\begin{aligned}
&\frac{1}{N^{2q}} \left| m \int W(m(v-u)) \sum_{j_1=-\infty}^{+\infty} \sum_{j_2=-\infty}^{+\infty} \sum_{t_1=1}^N \sum_{t_2=1}^N e^{i(t_2-t_1)u} \right. \\
&\quad \left. \left(g_{aj_{\xi_1}}(j_1) g_{bj_{\xi_2}}(j_2) \xi_{j_{\xi_1}} \xi_{j_{\xi_2}} - E g_{aj_{\xi_1}}(j_1) g_{bj_{\xi_2}}(j_2) \xi_{j_{\xi_1}} \xi_{j_{\xi_2}} \right) du \right|^{2q} = O_p\left(\frac{m^{2q}}{N^q}\right).
\end{aligned}$$

Take $j_{\xi_1} = c, j_{\xi_2} = d,$

$$\begin{aligned}
& \left| m \int W(m(v-u)) \frac{1}{N} \sum_{j_1=-\infty}^{+\infty} \sum_{j_2=-\infty}^{+\infty} \sum_{t_1=1}^N \sum_{t_2=1}^N e^{i(t_2-t_1)u} \right. \\
& \quad \left. \left(g_{aj_{\xi_1}}(j_1) g_{bj_{\xi_2}}(j_2) \xi_{j_{\xi_1}} \xi_{j_{\xi_2}} - E g_{aj_{\xi_1}}(j_1) g_{bj_{\xi_2}}(j_2) \xi_{j_{\xi_1}} \xi_{j_{\xi_2}} \right) du \right|^{2q} \\
& \leq \frac{1}{N^{2q}} \sum_{r=1}^{2q} \sum_{t_1+\dots+t_r=2q} \sum_{\substack{j_1 j_2 \\ \vdots \\ j_{1r} j_{2r}}} |g_{ac}(j_{11}) g_{cd}(j_{21})|^{t_1} \cdots |g_{ac}(j_{1r}) g_{cd}(j_{2r})|^{t_r} E |d_{j_{11}j_{21}}|^{t_1} \cdots |d_{j_{1r}j_{2r}}|^{t_r} \\
& \leq \frac{1}{N^{2q}} \sum_{r=1}^{2q} \sum_{t_1+\dots+t_r=2q} \sum_{\substack{j_1 j_2 \\ \vdots \\ j_{1r} j_{2r}}} |g_{ac}(j_{11}) g_{cd}(j_{21})|^{t_1} \cdots |g_{ac}(j_{1r}) g_{cd}(j_{2r})|^{t_r} (E |d_{j_{11}j_{21}}|^{2t_1} \cdots |d_{j_{1r}j_{2r}}|^{2t_r})^{\frac{1}{2}}.
\end{aligned}$$

Let us focus on $|d_{j_1 j_2}|^{4q},$

$$\begin{aligned}
|d_{j_1 j_2}|^{4q} &= \left| m \int W(m(v-u)) \sum_{t_1=1-j_1}^{N-j_1} \sum_{t_2=1-j_2}^{N-j_2} e^{i(t_2-t_1+j_2-j_1)u} (\xi_c(t_1) \xi_d(t_2) - E \xi_c(t_1) \xi_c(t_2)) du \right|^{4q} \\
&\leq \left| \sum_{t_1=1-j_1}^{N-j_1} \sum_{t_2=1-j_2}^{N-j_2} (\xi_c(t_1) \xi_d(t_2) - E \xi_c(t_1) \xi_d(t_2)) \right. \\
&\quad \left. w((j_1 - j_2 + t_1 - t_2)m^{-1}) e^{i(t_2-t_1+j_2-j_1)v} \right|^{4q} \\
&\leq B_5 \left(\left| \sum_{t_1=1-j_1}^{N-j_1} \sum_{t_2=1-j_2}^{N-j_2} \xi_c(t_1) \xi_d(t_2) w((j_1 - j_2 + t_1 - t_2)m^{-1}) \cos((t_2 - t_1 + j_2 - j_1)v) \right|^{4q} \right. \\
&\quad \left. + \left| \sum_{t=\max(1-j_1, 1-j_2)}^{\min(N-j_1, N-j_2)} (\xi_c(t) \xi_d(t) - \sigma_{cd}) w(j_1 - j_2) \right|^{4q} \right. \\
&\quad \left. + \left| \sum_{t_1=1-j_1}^{N-j_1} \sum_{t_2=1-j_2}^{N-j_2} \xi_c(t_1) \xi_d(t_2) w((j_1 - j_2 + t_1 - t_2)m^{-1}) \sin((t_2 - t_1 + j_2 - j_1)v) \right|^{4q} \right)
\end{aligned}$$

For three terms above, following Part 1, it is easy to see that the expectation of them are all $O(N^{2q}m^{4q})$, that is $E |d_{j_1 j_2}|^{4q} = O(N^{2q}m^{4q})$. For $E |d_{j_1 j_2}|^{2t_1} \cdots |d_{j_{1r} j_{2r}}|^{2t_r}$, when $(j_{11}, j_{21}), \dots, (j_{1r}, j_{2r})$ are different from each other, the number of non-zero terms should be less than $E |d_{rs}(v)|^{4q}$ because some $\xi_c(t) \xi_d(t)$ in $d_{rs}(v)$ can not be found in $d_{r_k s_k}(v)$,

so when the power of those $\xi_c(j)\xi_d(j)$ is 1, $E|d_{j_1j_2}|^{2t_1} \cdots |d_{j_{1r}j_{2r}}|^{2t_r} = O(N^{2q}m^{4q})$. Since $\sum_{j=-\infty}^{+\infty} |g_{ac}(j)| < \infty, \forall 1 \leq a \leq p_x, 1 \leq c \leq p_\xi$,

$$\sum_{\substack{j_{11}j_{22} \\ \dots \\ j_{1r}j_{2r}}} |g_{ac}(j_{11})g_{bd}(j_{21})|^{2t_1} \cdots |g_{ac}(j_{1r})g_{bd}(j_{2r})|^{2t_r} < +\infty.$$

So $EU_2(v) = O(\frac{m^{2q}}{N^q})$.

$$\begin{aligned} |f_{ab}(v) - E\hat{f}_{ab}(v)|^{2q} &= \left| f_{ab}(v) - m \int W(m(v-u))f_{ab}(u)du + m \int W(m(v-u))f_{ab}(u)du \right. \\ &\quad \left. - Em \int W(m(v-u)) \frac{1}{N} \sum_{t_1=1}^N X_a(t_1)e^{-it_1u} \sum_{t_2=1}^N X_a(t_2)e^{it_2u} \right|^{2q} \\ &\leq B_6 \left(\left| \int W(u) (f_{ab}(v) - f_{ab}(v - um^{-1})) du \right|^{2q} \right. \\ &\quad \left. + \left| m \int W(m(v-u))(f_{ab}(u) - EI_{ab}(u))du \right|^{2q} \right) \\ &= R_1(v) + R_2(v) \end{aligned}$$

Since $|f_{ab}(v) - f_{ab}(u)| \leq B_{ab}|v - u|$ (uniform Lipschitz condition),

$$\begin{aligned} R_1(v) &\leq \left(\int W(u)|f_{ab}(v) - f_{ab}(v - um^{-1})|du \right)^{2q} \\ &\leq \left(\frac{B_{ab}}{m} \int W(u)|u|du \right)^{2q} = \frac{B_{ab}^{2q}}{m^{2q}} \left(\int W(u)|u|du \right)^{2q} = O\left(\frac{1}{m^{2q}}\right). \end{aligned}$$

And,

$$\begin{aligned} EI_{ab}(v) &= E \frac{1}{N} \sum_{t_1=1}^N X_a(t_1)e^{-it_1v} \sum_{t_2=1}^N X_b(t_2)e^{it_2v} \\ &= \sum_{s=-N+1}^{N-1} e^{-isv} E\hat{R}_{ab}(s) \\ &= \sum_{s=-N+1}^{N-1} e^{-isv} \left(1 - \frac{|s|}{N}\right) R_{ab}(s). \end{aligned}$$

Since $R_{ab}(s) = \int f_{ab}(u)e^{isu} du$,

$$\begin{aligned}
EI_{ab}(v) &= \frac{1}{2\pi} \sum_{s=-N+1}^{N-1} e^{-isv} \left(1 - \frac{|s|}{N}\right) \int h_{ab}(u) e^{isu} du \\
&= \frac{1}{2\pi} \sum_{s=-N+1}^{N-1} \left(1 - \frac{|s|}{N}\right) \int h_{ab}(u) e^{is(u-v)} du \\
&= \frac{1}{2\pi} \sum_{s=-N+1}^{N-1} \int h_{ab}(u) \left(1 - \frac{|s|}{N}\right) \cos(s(u-v)) du \\
&\quad + i \frac{1}{2\pi} \sum_{s=-N+1}^{N-1} \int h_{ab}(u) \left(1 - \frac{|s|}{N}\right) \sin(s(u-v)) du \\
&= \frac{1}{2\pi} \sum_{s=-N+1}^{N-1} \int h_{ab}(u) \left(1 - \frac{|s|}{N}\right) \cos(s(u-v)) du \\
&= \frac{1}{2\pi} \int h_{ab}(u) \sum_{s=-N+1}^{N-1} \left(1 - \frac{|s|}{N}\right) \cos(s(u-v)) du \\
&= \frac{1}{2\pi} \int h_{ab}(u) \frac{1}{N} \frac{\sin^2\left(\frac{N(u-v)}{2}\right)}{\sin^2\left(\frac{u-v}{2}\right)} du,
\end{aligned}$$

where the last equality follows from P417 of Priestley (1981).

$$\begin{aligned}
|f_{ab}(v) - EI_{ab}(v)| &\leq \frac{1}{2\pi N} \int \frac{\sin^2\left(\frac{N(u-v)}{2}\right)}{\sin^2\left(\frac{u-v}{2}\right)} |f_{ab}(u) - f_{ab}(v)| du \\
&\leq B_{ab} \int_{|u-v| \leq \epsilon} \frac{\sin^2\left(\frac{N(u-v)}{2}\right)}{\sin^2\left(\frac{u-v}{2}\right)} |u-v| du \\
&\quad + \int_{|u-v| > \epsilon} \frac{\sin^2\left(\frac{N(u-v)}{2}\right)}{\sin^2\left(\frac{u-v}{2}\right)} |f_{ab}(u) - f_{ab}(v)| du \\
&\leq B_6 \frac{\log N}{N} + \frac{B_6}{N} \\
&\leq \frac{B_7 \log N}{N},
\end{aligned}$$

where the first part of the third inequality is from Fejer (1910) for sufficiently small ϵ . See also Rosenblatt (1974). The second part is because $|f_{ab}(v)|$ is bounded uniformly on $[-\pi, \pi]$,

$\forall a, b$. So,

$$\left| m \int W(m(v-u))(f_{ab}(u) - EI_{ab}(u))du \right|^{2q} = O\left(\frac{m^{2q}}{N^q}\right).$$

Here we complete the proof of part 2.

Combining part 1 and part 2, we complete the proof of Lemma 1.

Lemma A.6. For $X_t = \sum_{k=1}^{K+1} \sum_{j=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} \begin{pmatrix} g_{11}(j) & \cdots & g_{1p_\xi}(j) \\ \vdots & \ddots & \vdots \\ g_{p_x 1} & \cdots & g_{p_x p_\xi}(j) \end{pmatrix} \boldsymbol{\xi}(t-j)$, under Assumption 1-8,

$$\begin{aligned} |\hat{\lambda}_1(v_k, \hat{\mathbf{f}}_{xx}) - \lambda_1(v_k, \mathbf{f}_{xx})|^{2q} &= O_p\left(\frac{m^{2q}}{N^q}\right), \\ \max_k |\hat{\lambda}_1(v_k, \hat{\mathbf{f}}_{xx}) - \lambda_1(v_k, \mathbf{f}_{xx})|^{2q} &= O_p\left(\frac{m^{2q}}{N^{q-1}}\right) \end{aligned}$$

Proof:

$$\mathbf{f}_{xx} = \frac{1}{2\pi} \sum_{k=1}^{K+1} \frac{N_k}{N} \Gamma(v, k) \Sigma \Gamma^*(v, k).$$

By Wielandt-Hoffman theorem again, following the discussions in Lemma A.5,

$$\max_i |\hat{\lambda}_i(v, \hat{\mathbf{f}}_{xx}) - \lambda_i(v, \mathbf{f}_{xx})|^{2q} \leq \sum_{j,l}^{p_\xi} |\hat{f}_{xx,j,l}(v) - f_{xx,j,l}(v)|^{2q}.$$

$$\begin{aligned} \hat{f}_{xx,a,b}(v) &= m \int W(m(v-u)) \hat{I}_{xx,a,b}(u) du \\ &= m \int W(m(v-u)) \frac{1}{N} \sum_{t_1=\tau_{k-1}+1}^{\tau_k} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k-1}+1}^{\tau_k} X_b(t_2) e^{it_2 u} \\ &\quad + m \int W(m(v-u)) \frac{1}{N} \sum_{k_1 \neq k_2} \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du \end{aligned}$$

Since $\min(\tau_k - \tau_{k-1}) > m$, so we have

$$m \int W(m(v-u)) \frac{1}{N} \sum_{|k_1-k_2| \geq 2} \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du = 0.$$

So

$$\begin{aligned} \hat{f}_{xx,a,b}(v) &= m \int W(m(v-u)) \frac{1}{N} \sum_{k_1=1}^{K+1} \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_1-1}+1}^{\tau_{k_1}} X_b(t_2) e^{it_2 u} du \\ &\quad + \frac{1}{N} \sum_{k_1-k_2=1} m \int W(m(v-u)) \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du \\ &\quad + \frac{1}{N} \sum_{k_2-k_1=1} m \int W(m(v-u)) \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du. \end{aligned}$$

Then

$$\begin{aligned} &\left| \hat{f}_{xx,a,b} - f_{xx,a,b}(v) \right|^{2q} \\ &\leq B_1 \left(\left| \hat{f}_{xx,a,b}(v) - \frac{1}{N} \sum_{k=1}^{K+1} \Gamma(v, k) \Sigma \Gamma^*(v, k) \right|^{2q} + \left| \frac{1}{N} \sum_{k_1-k_2=1} m \int W(m(v-u)) \right. \right. \\ &\quad \left. \left. \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du \right|^{2q} + \left| \frac{1}{N} \sum_{k_2-k_1=1} m \int W(m(v-u)) \right. \right. \\ &\quad \left. \left. \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du \right|^{2q} \right). \\ &\leq B_2 \left(\sum_{k=1}^{K+1} \left(\frac{N_k}{N} \right)^{2q} \left| m \int W(m(v-u)) \frac{1}{N_k} \sum_{t=\tau_{k-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du \right. \right. \\ &\quad \left. \left. - \frac{1}{N_k} \Gamma(v, k) \Sigma \Gamma^*(v, k) \right|^{2q} \right. \\ &\quad + \sum_{k_2-k_1=1} \left| \frac{1}{N} m \int W(m(v-u)) \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du \right|^{2q} \\ &\quad \left. + \sum_{k_2-k_1=1} \left| \frac{1}{N} m \int W(m(v-u)) \sum_{t_1=\tau_{k_1-1}+1}^{\tau_{k_1}} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_{k_2-1}+1}^{\tau_{k_2}} X_b(t_2) e^{it_2 u} du \right|^{2q} \right) \\ &= B_2(R_1(v) + R_2(v) + R_3(v)). \end{aligned}$$

Following the proofs of Lemma A.5, $ER_1(v) = O(\frac{m^{2q}}{N^q})$. And by symmetry, we only need to prove $ER_2(v) = O(\frac{m^{2q}}{N^q})$.

Without loss of generality, set $k_1 = 1, k_2 = 2$. Next, we will separate the proofs into two parts.

Part 1:

$$\begin{aligned}
& E \left| \frac{1}{N} m \int W(m(v-u)) \sum_{t_1=1}^{\tau_1} \xi_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_1+1}^{\tau_2} \xi_b(t_2) e^{it_2 u} du \right|^{2q} \\
&= E \left| \frac{1}{N} \sum_{l=1}^{N_1} w(lm^{-1}) e^{ilv} \sum_{j=N_1-l+1}^{N_1-m} \xi_a(j) \xi_b(j+l) \right|^{2q} \\
&\leq B_3 \left(E \left(\frac{1}{N^{2q}} \sum_{l=1}^m w(lm^{-1}) \cos(lv) \sum_{j=N_1-l+1}^{N_1-m} \xi_a(j) \xi_b(j+l) \right)^{2q} \right. \\
&\quad \left. + E \left(\frac{1}{N^{2q}} \sum_{l=1}^m w(lm^{-1}) \sin(lv) \sum_{j=N_1-l+1}^{N_1-m} \xi_a(j) \xi_b(j+l) \right)^{2q} \right).
\end{aligned}$$

For two terms above, following the proofs of Part 1 in Lemma A.5, we have

$$\begin{aligned}
& E \left(\frac{1}{N^{2q}} \sum_{l=1}^m w(lm^{-1}) \cos(lv) \sum_{j=N_1-l+1}^{N_1-m} \xi_a(j) \xi_b(j+l) \right)^{2q} = O\left(\frac{m^{2q}}{N^q}\right), \\
& E \left(\frac{1}{N^{2q}} \sum_{l=1}^m w(lm^{-1}) \sin(lv) \sum_{j=N_1-l+1}^{N_1-m} \xi_a(j) \xi_b(j+l) \right)^{2q} = O\left(\frac{m^{2q}}{N^q}\right).
\end{aligned}$$

Part 2:

$$\begin{aligned}
& \left| \frac{1}{N} m \int W(m(v-u)) \sum_{t_1=\tau_0+1}^{\tau_1} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_1+1}^{\tau_2} X_b(t_2) e^{it_2 u} du \right|^{2q} \\
&\leq B_4 \left(\left| \frac{1}{N} m \int W(m(v-u)) \sum_{t_1=\tau_0+1}^{\tau_1} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_1+1}^{\tau_2} X_b(t_2) e^{it_2 u} du \right. \right. \\
&\quad \left. \left. - E \frac{1}{N} m \int W(m(v-u)) \sum_{t_1=\tau_0+1}^{\tau_1} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_1+1}^{\tau_2} X_b(t_2) e^{it_2 u} du \right|^{2q} \right. \\
&\quad \left. + \left| E \frac{1}{N} m \int W(m(v-u)) \sum_{t_1=\tau_0+1}^{\tau_1} X_a(t_1) e^{-it_1 u} \sum_{t_2=\tau_1+1}^{\tau_2} X_b(t_2) e^{it_2 u} du \right|^{2q} \right) \\
&= B_4(U_1(v) + U_2(v))
\end{aligned}$$

And

$$U_2(v) \leq \sum_{p_{\xi_1}=1}^{p_\xi} \sum_{p_{\xi_2}=1}^{p_\xi} \frac{1}{N^{2q}} E \left| g_{aj_{\xi_1}}(j_1, 1) g_{bj_{\xi_2}}(j_2, 2) \xi_{j_{\xi_1}}(t_1 - j_1) \xi_{j_{\xi_2}}(t_2 - j_2) \right|^{2q}$$

Following the discussions in Part 1 of Lemma A.5 again, we have $U_2(v) = O\left(\frac{m^{2q}}{N^q}\right)$.

$$\begin{aligned} U_1(v) &\leq \sum_{p_{\xi_1}=1}^{p_\xi} \sum_{p_{\xi_2}=1}^{p_\xi} \frac{B_5}{N^{2q}} \left| m \int W(m(v-u)) \sum_{j_1=-\infty}^{+\infty} \sum_{j_2=-\infty}^{+\infty} \sum_{t_1=1}^{\tau_1} \sum_{t_2=\tau_1+1}^{\tau_2} e^{i(t_2-t_1)u} \right. \\ &\quad \left. \left(g_{aj_{\xi_1}}(j_1, 1) g_{bj_{\xi_2}}(j_2, 2) \xi_{j_{\xi_1}}(t_1 - j_1) \xi_{j_{\xi_2}}(t_2 - j_2) \right. \right. \\ &\quad \left. \left. - E \left(g_{aj_{\xi_1}}(j_1, 1) g_{bj_{\xi_2}}(j_2, 2) \xi_{j_{\xi_1}}(t_1 - j_1) \xi_{j_{\xi_2}}(t_2 - j_2) \right) \right) \right|^{2q}. \end{aligned}$$

Following the discussions in Part 2 of Lemma A.5, we have $EU_1(v) = O\left(\frac{m^{2q}}{N^q}\right)$. So combing the discussions above, we complete the proof of Lemma A.6.

Lemma A.7. *Under Assumption 1-8, $\forall s = 1, \dots, K + 1$, $\max_{\tau_{s-1} \leq k < l \leq \tau_s} \vartheta_{kl} \sim O_p\left(N^{-\frac{q-3-2q\alpha}{2q}}\right)$, when $l - k = N_{kl} > ml$. Here*

$$\vartheta_{kl} = \frac{N_{kl}}{N} \int \hat{\lambda}_{kl}(u) \log \frac{st(\hat{\lambda}_{kl})}{st(\lambda_s)} du,$$

Proof: Set $A_{kl} = \left\{ \max_{v_i} \left| \frac{\hat{\lambda}_{kl}(u) - \lambda_{kl}(u)}{\lambda_{kl}(u)} \right| \leq \epsilon \frac{m}{n^{(q-3)/2q}} \right\}$. By Lemma A.5, we have

$$P(A_{kl}) \leq \frac{B_1}{n^2 \epsilon^{2q}}.$$

On set A_{kl} ,

$$\max(0, 1 - \epsilon B_2 \frac{m}{n^{(q-3)/2q}}) \hat{\Lambda}_{kl} \leq \Lambda_s \leq (1 + \epsilon B_2 \frac{m}{n^{(q-3)/2q}}) \hat{\Lambda}_{kl}.$$

Then on set $\bigcap_{k,l} A_{kl}$,

$$\begin{aligned} \max_{k,l} \vartheta_{kl} &= \max_{k,l} \frac{N_{kl}}{N} \int \hat{\lambda}_{kl}(u) \log \frac{st(\hat{\lambda}_{kl})}{st(\lambda_s)} du \\ &= \max_{k,l} \frac{N_{kl}}{N} \int \hat{\lambda}_{kl}(u) \log \left(\frac{\Lambda_s}{\hat{\Lambda}_{kl}} * \frac{\hat{\lambda}_{kl}(u)}{\lambda_s(u)} \right) du \end{aligned}$$

$$\begin{aligned}
&\leq \max_{k,l} \frac{N_{kl}}{N} \int \max_u \hat{\lambda}_{kl}(u) \log \left(\frac{\Lambda_s}{\hat{\Lambda}_{kl}} * \max_u \frac{\hat{\lambda}_{kl}(u)}{\lambda_s(u)} \right) du \\
&\leq \max_{k,l} \frac{N_{kl}}{N} \int \max_u \lambda_s(u) \left(1 + \frac{B_2 m \epsilon}{n^{(q-3)/2q}} \right) \log \left(\left(1 + \frac{B_2 m \epsilon}{n^{(q-3)/2q}} \right) \left(1 + \frac{B_1 m \epsilon}{n^{(q-3)/2q}} \right) \right) du \\
&\leq \max_{k,l} \frac{N_{k,l}}{N} \int \max_u \lambda_s(u) \left(1 + \frac{B_2 m \epsilon}{N^{(q-3)/2q}} \right) \left(1 + \frac{B_1 m \epsilon}{N^{(q-3)/2q}} \right) du \\
&\leq B_{3\epsilon} N^{-\frac{q-3-2q\alpha}{2q}}.
\end{aligned}$$

We set $B_{4\epsilon} > B_{3\epsilon}$,

$$\begin{aligned}
P(\max_{k,l} \vartheta_{kl} \geq B_{4\epsilon} N^{-\frac{q-3-2q\alpha}{2q}}) &= P \left(\max_{k,l} \vartheta_{kl} \geq B_{4\epsilon} N^{-\frac{q-3-2q\alpha}{2q}} \mid \bigcap_{kl} A_{kl} \right) P \left(\bigcap_{kl} A_{kl} \right) \\
&\quad + P \left(\max_{k,l} \vartheta_{kl} \geq B_{4\epsilon} N^{-\frac{q-3-2q\alpha}{2q}} \mid \bigcup_{kl} A_{kl}^c \right) P \left(\bigcup_{kl} A_{kl}^c \right) \\
&\leq P \left(\max_{k,l} \vartheta_{k,l} \geq B_{4\epsilon} N^{-\frac{q-3-2q\alpha}{2q}} \mid \bigcap_{kl} A_{kl} \right) + P \left(\bigcup_{kl} A_{kl}^c \right) \\
&\leq 0 + \frac{B_1}{\epsilon^{2q}},
\end{aligned}$$

and the last inequality can be arbitrarily small by choosing sufficiently large ϵ . Here we complete the proof of Lemma A.7.

Lemma A.8. *Under Assumptions, we have $\max_{\tau_0 \leq k < l \leq \tau_{K+1}} \vartheta_{kl} \sim O_p(n^{-\frac{q-3-2q\alpha}{2q}})$, when $l - k = n_{kl} > ml$. Here*

$$\vartheta_{kl} = \frac{N_{kl}}{N} \int \hat{\lambda}_{kl}(u) \log \frac{st(\hat{\lambda}_{kl})}{st(\lambda_{kl})} du,$$

Proof: Following the proof of Lemma A.7, it can be easily obtained.

Proofs of Theorem 4.1

Set $A_{kl} = \left\{ \omega : \max_{v_i} \frac{\hat{\lambda}_{kl}(v_i) - \lambda_{kl}(v_i)}{\lambda_{kl}(v_i)} \leq \epsilon \frac{m}{N^{(q-3)/2q}} \right\}$, where ϵ is sufficiently small. Then $\forall \omega \in \bigcap_{k,l} A_{k,l}$, we prove that $\hat{\kappa}_j \rightarrow \kappa_j, \forall j$.

Since $\hat{\kappa}_j$ is bounded, there exists a subsequence $\{n_s\}$, such that $\hat{\kappa}_j \rightarrow \kappa_j^*$. Then by Lemma A.6 and A.7, we have

$$\frac{1}{N}R(\hat{\kappa}_1, \dots, \hat{\kappa}_K) = \sum_{k=1}^{K+1} (\kappa_k^* - \kappa_{k-1}^*) \int \lambda_{\kappa_{k-1}^*, \kappa_k^*}(u) \log \left(\frac{st(\lambda_{\kappa_{k-1}^*, \kappa_k^*})}{st(\lambda)} \right) du + O_p(N^{-\frac{q-3-2q\alpha}{2q}}).$$

When $\kappa_{i-1}^0 \leq \kappa_{j-1}^* < \kappa_i^0 < \dots < \kappa_{i+k}^0 < \kappa_j^* < \kappa_{i+k+1}$, by Lidskii's inequality (Tao 2012),

$$\begin{aligned} \lambda_{kl} &= \lambda \left(\frac{\kappa_i^0 - \kappa_{j-1}^*}{\kappa_j^* - \kappa_{j-1}^*} f_{xx,i} + \frac{\kappa_{i+1}^* - \kappa_i^*}{\kappa_j^* - \kappa_{j-1}^*} f_{xx,i+1} + \dots + \frac{\kappa_j^* - \kappa_{i+k}^0}{\lambda_j^* - \lambda_{j-1}^*} f_{xx,i+k+1} \right) \\ &\leq \frac{\kappa_i^0 - \kappa_{j-1}^*}{\kappa_j^* - \kappa_{j-1}^*} \lambda(f_{xx,i}) + \frac{\kappa_{i+1}^* - \kappa_i^*}{\kappa_j^* - \kappa_{j-1}^*} \lambda(f_{xx,i+1}) + \dots + \frac{\kappa_j^* - \kappa_{i+k}^0}{\lambda_j^* - \lambda_{j-1}^*} \lambda(f_{xx,i+k+1}) \\ &= \frac{\kappa_i^0 - \kappa_{j-1}^*}{\kappa_j^* - \kappa_{j-1}^*} \lambda_i + \frac{\kappa_{i+1}^* - \kappa_i^*}{\kappa_j^* - \kappa_{j-1}^*} \lambda_{i+1} + \dots + \frac{\kappa_j^* - \kappa_{i+k}^0}{\lambda_j^* - \lambda_{j-1}^*} \lambda_{i+k+1} \end{aligned}$$

So

$$\begin{aligned} &(\kappa_j^* - \kappa_{j-1}^*) \int \lambda_{\kappa_{k-1}^*, \kappa_k^*}(u) \log \left(\frac{st(\lambda_{\kappa_{k-1}^*, \kappa_k^*})}{st(\lambda)} \right) du - (\kappa_i^0 - \kappa_{j-1}^*) \int \lambda_i \log \frac{st(\lambda_i)}{st(\lambda)} du \\ &- (\kappa_{i+1}^0 - \kappa_i^0) \int \lambda_{i+1} \log \frac{st(\lambda_{i+1})}{st(\lambda)} du - \dots - (\kappa_j^* - \kappa_{i+k}^0) \int \lambda_{i+k} \log \frac{st(\lambda_{i+k})}{st(\lambda)} du \\ &\leq (\kappa_i^0 - \kappa_{j-1}^*) \int \lambda_i \log \frac{st(\lambda_{\kappa_{k-1}^*, \kappa_k^*})}{st(\lambda_i)} du + (\kappa_{i+1}^0 - \kappa_i^0) \int \lambda_{i+1} \log \frac{st(\lambda_{\kappa_{k-1}^*, \kappa_k^*})}{st(\lambda_{i+1})} du + \dots + \\ &(\kappa_j^* - \kappa_{i+k}^0) \int \lambda_{i+k} \log \frac{st(\lambda_{\kappa_{k-1}^*, \kappa_k^*})}{st(\lambda_{i+k})} du < 0. \end{aligned}$$

When $\kappa_{j-1}^0 < \kappa_k^* < \kappa_{k+1}^* < \kappa_j^0$, then

$$\frac{1}{N}R(\hat{\kappa}_{j-1}, \hat{\kappa}_j) = (\kappa_j^* - \kappa_{j-1}^*) \int \lambda_j \log \left(\frac{st(\lambda_j)}{st(\lambda)} \right) du + O(N^{-\frac{q-3-2q\alpha}{2q}}).$$

So

$$\begin{aligned} &\frac{1}{N}R(\hat{\kappa}_1, \dots, \hat{\kappa}_K) - \frac{1}{N}R(\kappa_1^0, \dots, \kappa_K^0) \\ &= \frac{1}{N}R(\kappa_1^*, \dots, \kappa_K^*) + O(N^{-\frac{q-3-2q\alpha}{2q}}) - \frac{1}{N}R(\kappa_1^0, \dots, \kappa_K^0) < 0 \end{aligned}$$

as $N \rightarrow \infty$. This is a contradiction since $\hat{\kappa}_j$ are the maximizer of $R(\kappa_1, \dots, \kappa_K)$. And

$$P \left(\bigcap_{k,l} A_{kl} \right) = 1 - P \left(\bigcup_{k,l} A_{kl}^c \right) \geq 1 - \sum_{l-k \geq ml}^N P(A_{kl}^c) = 1 - \frac{B_1}{\epsilon^{2q}}.$$

By choosing ϵ sufficiently large, we have $\hat{\kappa}_j \rightarrow \kappa_j^0$ in probability.

Proofs of Theorem 4.2

If $L < K$, there should be a change point κ_j^0 that can not be estimated consistently. So

$$\begin{aligned} & -\frac{1}{N}R(\hat{\kappa}_1, \dots, \hat{\kappa}_L) + LC_N/N + \frac{1}{N}R(\hat{\kappa}_1, \dots, \hat{\kappa}_K) - KC_N/N \\ & < -c + O_p(N^{-\frac{q-3-2q\alpha}{2q}}) - (K-L)C_N/N < 0, \end{aligned}$$

as N tends to infinity, where c is a positive constant.

If $L > K$, then still every change point κ_j^0 should be estimated consistently, so following the proof of Theorem 1, there should be a change point κ_k^* such that $\kappa_{j-1}^0 < \kappa_k^* < \kappa_j^0$. Then by Lemma A.6, A.7 and A.8,

$$R(\kappa_{j-1}^0, \kappa_k^*, \kappa_j) - R(\kappa_{j-1}, \kappa_j) = O_p\left(N^{-\frac{q-3-2q\alpha}{2q}}\right).$$

So $BIC_L - BIC_K = O_p\left(N^{-\frac{q-3-2q\alpha}{2q}}\right) + (K-L)C_N/N < 0$, as N tends to infinity. Here we complete the proof of Theorem 2.