MULTIPLE CHANGE-POINT DETECTION FOR PIECEWISE STATIONARY CATEGORICAL TIME SERIES

by

Cong Ye

B.S., Statistics, Wuhan University, 2008

M.A., Applied Statistics, University of Pittsburgh, 2010

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial

fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICHDIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Cong Ye

It was defended on

November 18th 2015

and approved by

David S. Stoffer, Ph.D., Department of Statistics,Professor

Yu Cheng, Ph.D., Department of Statistics, Associate Professor

Sungkyu Jung, Ph.D., Department of Statistics, Assistant Professor

Ching-Chung Li, Ph.D., Department of Electrical and Computer Engineering, Professor

Dissertation Director: David S. Stoffer, Ph.D., Department of Statistics,Professor

MULTIPLE CHANGE-POINT DETECTION FOR PIECEWISE STATIONARY CATEGORICAL TIME SERIES

Cong Ye, PhD

University of Pittsburgh, 2015

In this dissertation, we propose a fast yet consistent method for segmenting a piecewise stationary categorical-valued time series, with a finite unknown number of change-points in its autocovariance structure. To avoid loss of information, instead of arbitrarily assigning numerical numbers in analysis of the original time series, we focus on the multinomial process, which is derived by denoting each category of the original series as a unit vector. The corresponding multinomial process is then modeled by a nonparametric multivariate locally stationary wavelet process, where the piecewise constant autocovariance structure for any given variate is completely described by the wavelet periodograms for that variate at multiple scales and locations. Further, we propose a criterion that optimally selects the scalings and provides the generation of the trace statistics whose mean functions inherit the piecewise constancy. The resulting statistics will serve as input sequences for later segmentation. Change-point detection is accomplished by first examining the input sequence at each scale with a proposed binary segmentation procedure, and then combining the detected breakpoints across scales. The consistency result of our method is established under certain conditions. In addition, several simulation studies and a real-data analysis of a DNA sequence are provided to demonstrate the viability of our methodology.

Keywords: categorical-valued time series, piecewise stationarity, locally stationary wavelet process, wavelet analysis, binary segmentation, DNA sequences.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my sincere appreciation to my advisor, Professor David S. Stoffer, for his great help and valuable advice throughout my Ph.D. life. Without his continuous help and support, it would not have been possible for me to finish this dissertation. I would also want to thank my other committee members Professor Yu Cheng, Professor Sungkyu Jung and Professor Ching-Chung Li for their great and constructive advice on my research.

In addition, I wish to express my appreciation to the rest of the faculty members, staff, especially Mary Gerber, and my fellow students in the statistics department for their encouragement, support and friendship throughout my graduate years.

Last but not the least, I would like to thank my husband and my true friend, Songming Peng, for the love and happiness he has brought to me during the past ten years. I would also like to thank my parents, Zicheng Ye and Qingyu Zou, my parents-in-law, Weiguo Xu and Chunfang Peng, for their endless love and support, which helped me devote myself to the research. Moreover, I want to thank my other good friends in Pittsburgh, Shuangyan Xiong and Cong Lu, for the numerous happy times we have been through together for the past six years.

1.0 INTRODUCTION

Categorical time series are serially correlated data which are recorded in terms of categories (or states) at discrete time points. This kind of series is found in many fields of application, such as analyzing DNA sequence data (Stoffer et al. (1993a) [1]) and analyzing a person's sleep states with data obtained from electroencephalography (EEG) (Stoffer et al. (1988) [2]).

Typically, the categories in a categorical time series are assigned "scalings" (numerical values) in order to facilitate graphing and analysis. However, arbitrarily assigning numbers may mask some interesting features (such as periodic patterns) in the data. Under the assumption of stationarity, Stoffer et al. (1993a) [1] first introduced the concept of spectral envelope to the process of spectral analysis for stationary categorical time series. The spectral envelope approach, based on the Fourier analysis, could select scalings that could help emphasize any periodic feature of the series. This dissertation will focus on the dependence structure of categorical time series.

The assumption of stationarity is appealing for developing theoretical results. However, this assumption is often unrealistic since local behaviors in practical problems are widely prevalent. An example is DNA sequence (see more details in Chapter 2).

Many models and methods for dealing with non-stationarity have focused on numerical time-dependent data. Priestley (1965) [3] first proposed a time-dependent Cramér-like spectral representation for non-stationary time series, where spectral properties change slowly over time. Dahlhaus (1997) [4] refined the ideas in Priestley and introduced the class of locally stationary process, which allows for a smoothly time-varying spectrum in the asymptotic limit. These early stage models for non-stationary time series are generalized from the Cramér spectral representation by using time-varying transfer functions, while keeping the Fourier basis functions, which are not localized in time, as building blocks. Thus, those

models cannot adequately represent processes whose spectral properties evolve with time in more general ways. In order to alleviate the time localization problem, other basis functions were considered as building blocks. Nason (2000) [5] developed the locally stationary wavelet (LSW) process in exactly the same spirit as the Dahlhaus model where the Fourier basis is replaced by a wavelet basis, localized both in time and scale. Ombao et al. (2002) [6] constructed the so-called Smooth Localized complex EXponential (SLEX) basis functions and developed the SLEX model, which allows for statistical inference as well as the establishment of the estimation theory. As for (turning to) a special class of non-stationary time series, piece-wise stationary time series, many models and methods have been developed to automatically divide it into segments, where certain statistical properties are approximately the same within each segment. Adak (1998) [7] and Ombao et al. (2002) [6] proposed methods that divided the univariate time series into dyadic blocks and that selected the best segmentation according to BBA (Best Basis Algorithm, see Wickerhauser (1994) [8]). Adak's work was based on the windowed Fourier transform and Ombao's work utilized the SLEX (smooth localized complex exponentials) basis. Later, Ombao et al. extended their research to multivariate cases (Ombao et al.(2005) [9]) . Cho et al. (2012) [10] applied the binary segmentation method in the locally stationary wavelet (LSW) process, which was first introduced by Nason (2000) [5]. Their follow-up work in multivariate cases was presented in Cho et al. (2014) [11].

As discussed previously, those working with non-stationary categorical time series encounter the additional challenge of choosing appropriate scalings. There is considerably less research about the spectral domain analysis of non-stationary categorical time series. The most recent work was introduced by Stoffer et al. (2002) [12], whose approach combined dyadic tree-based adaptive segmentation (TBAS) and spectral envelope methodologies. However, one restriction of the dyadic tree-based method is that the change-points should be at dyadic locations; if not, the location of an estimated change-point would be a crude approximation. To overcome this restriction, we propose a detection method that involves extending the idea of spectral envelope with wavelet techniques and introducing the binary segmentation procedure; thus, the detection method could handle more flexible cases of change-point locations.

The rest of the dissertation is organized as follows: Chapter 2 briefly describes the common problems in DNA sequence data that will be used to validate our proposed method. Chapter 3 gives the related definitions and background techniques for our research. Chapter 4 is devoted to our proposed method and consistency result. Chapter 5 and Chapter 6 present simulation results and an application to the Epstein-Barr Virus (EBV) DNA sequence data. Finally, future directions are presented in Chapter 7.

2.0 DNA SEQUENCE DATA

This dissertation considers data from the EBV DNA sequence. This chapter briefly reviews the special problems in analyzing DNA sequence data.

DNA is commonly viewed as a genetic material in all life forms that is responsible for storing and encoding genetic information. Naturally, DNA molecules exist as double-stranded helices, consisting of two long biopolymers of nucleotides. Each nucleotide is composed of a phosphate group, a five-carbon sugar, and a nucleobase. The nucleobases are classified into four types—tadenine (A), guanine (G), cytosine (C), and thymine (T)—forming four different nucleotides. These nucleotides are linked together by a backbone made of alternating sugars and phosphate groups to form a single strand of DNA. The sequence of these four nucleobases in the strand contains genetic information specific to the organism. Two single strands of DNA, when complementary to each other, will hybridize to each other to form double-stranded DNA, following the base-paring rule (i.e. A pairs with T, and G pairs with C). This complementary base pairing rule allows the sequence in a single strand to represent the information in a DNA molecule. Thus, one strand of DNA can be recorded as a categorical sequence. For example, Table 1 shows part of the Epstein-Barr Virus DNA sequence derived from a single strand.

The information carried by DNA is held in the genes, which are pieces of a DNA sequence. Within a gene, the sequence of nucleobases along a DNA strand defines one or more protein-coding sequences (CDS) and hence can influence the phenotype of an organism. In many species, only a small fraction of the total sequence of the genome encodes protein. For instance, only about 1.5% of the human genome consists of protein-coding exons, with over 50% of human DNA carrying non-coding repetitive sequences [13]. Thus, a common problem in analyzing long DNA sequence data is identifying CDS that are dispersed throughout the

Table 1: Part of the Epstein-Barr Virus DNA sequence (read across and down)

GCCCTGGGGT	AAGTCTGGGA	GGCAGAGGGT	CGGCCTAGGC	CCGGGGAAGT	GGAGGGGGAT
CGCCCGGGTC	TCTGTTGGCA	GAGTCCGGGC	GATCCTCTGA	GACCCTCCGG	GCCCGGACGG
TCGCCCTCAG	CCCCCCAGAC	AGACCCCAGG	GTCTCCAGGC	AGGGTCCGGC	ATCTTCAGGG
GCAGCAGGCT	CACCACCACA	GGCCCCCCAG	ACCCGGGTCT	CGGCCAGCCG	AGCCGACCGG
CCCCGCGCCT	GGCGCCTCCT	CGGGGCCAGC	CGCCGGGGTT	GGTTCTGCCC	CTCTCTCTGT
CCTTCAGAGG	AACCAGGGAC	CTCGGGCACC	CCAGAGCCCC	TCGGGCCCGC	CTCCAGGCGC
CCTCCTGGTC	TCCGCTCCCC	TCTGAGCCCC	GTTAAACCCA	AAGAATGTCT	GAGGGGAGCC

sequence and separated by regions of noncoding sequences (Stoffer et al (2000) [12]).

It is well known that DNA is heterogeneous (Karlin and Macken, 1991) [14]. As Braun and Muller (1998) [15] pointed out, it is suitable to partition the DNA sequence into segments, where each segment has a certain degree of internal homogeneity. The stationary assumption might not be appropriate, but piecewise stationary assumption might be plausible here. Hence, the problem becomes how to detect the multiple change-points in piecewise stationary categorical time series.

3.0 BACKGROUND

In this chapter, we present some definitions and technical background that will help in understanding our proposed method. We first give the definition of piecewise categorical time series in Section 3.1. Next, in Section 3.2, we discuss wavelet techniques related to our research. Binary segmentation algorithm is illustrated in Section 3.3. Finally, in Section 3.4, we briefly review the local spectral envelope methodology. Our proposed method will combine aspects of all those methodologies in order to perform fast and automatic detection of multiple changepoints in piecewise categorical time series.

3.1 PIECEWISE STATIONARY CATEGORICAL TIME SERIES

Before presenting the definition of piecewise stationary categorical time series, we first introduce the stationary categorical time series. Let Y_t be a categorical-valued time series with finite state-space $\mathcal{C} = \{\mathrm{c}_1, \cdots, \mathrm{c}_p\}$. If Y_t is stationary, the probabilities $\mathrm{p}_j = \mathrm{pr}\{Y_t = \mathrm{c}_j\} > 0$ for $j = 1, 2, \cdots, p$ do not depend on time t. For $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)' \in \mathbb{R}^p$, denote by $X_t(\boldsymbol{\beta})$ the real-valued stationary time series corresponding to the scaling that assigns the category c_j the numerical value $\beta_j, j = 1, 2, \cdots, p$.

It is often very useful to represent the categories in terms of the unit vectors $\vec{\boldsymbol{e}}_1, \vec{\boldsymbol{e}}_2, \cdots, \vec{\boldsymbol{e}}_p$, where $\vec{\boldsymbol{e}}_j$ represents the $p \times 1$ vector with 1 in the j-th row and 0 elsewhere. Let $\boldsymbol{Z}_t = \vec{\boldsymbol{e}}_j$ when $Y_t = \mathrm{c}_j$. Assume that the vector process $\boldsymbol{Z}_t$ has a continuous spectral density denoted by $f_Z(\omega)$. For each ω, $f_Z(\omega)$ is a $p \times p$ complex-valued Hermitian matrix. Assume the existence of $f_X(\omega; \boldsymbol{\beta})$, the spectral density of $X_t(\boldsymbol{\beta})$. With the relationship $X_t\beta = \boldsymbol{\beta}'\boldsymbol{Z}_t$, we have $f_X(\omega; \boldsymbol{\beta}) = \boldsymbol{\beta}' f_Z(\omega)\boldsymbol{\beta} = \boldsymbol{\beta}' f_Z^{re}(\omega)\boldsymbol{\beta}$, where V is the variance-covariance matrix of $\boldsymbol{Z}_t$.

By the definition of Stoffer et al. (2002) [12], $p \times 1$ vector-valued piecewise stationary process, $\{\boldsymbol{Z}_{s,T}\}_{s=0}^{T-1}$, is defined as:

$$\boldsymbol{Z}_{s,T} = \sum_{b=1}^{B} \boldsymbol{Z}_{s,b} \mathcal{I}(s/T, U_b); \tag{3.1}$$

here, $U_b = [u_{b-1}, u_b) \subset [0, 1)$ are intervals, in which $\boldsymbol{Z}_{s,b}$ are stationary processes. $\mathcal{I}(s/T, U_b)$ is an indicator that takes the value 1 if $s/T \in U_b$, and 0 otherwise. Let M_b represent the number of observations in segment b and $\sum_{b=1}^{B} M_b = T$.

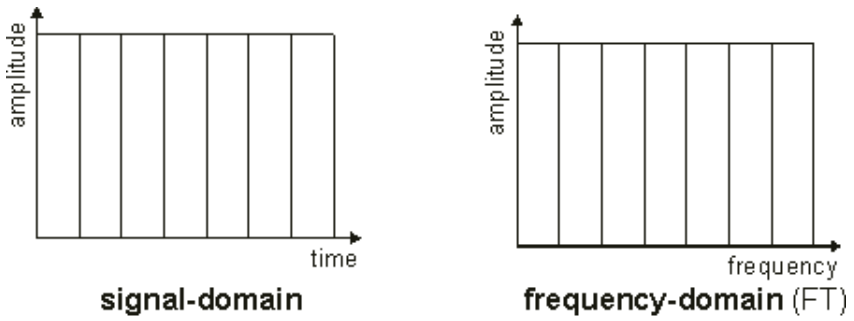
A categorical time series, $\{Y_{t,T}\}$, on a finite state-space and with nonzero marginal probabilities, is piecewise stationary if the corresponding $p \times 1$ point process, $\{\boldsymbol{Z}_{t,T}\}$, is piecewise stationary.

3.2 NON-DECIMATED DISCRETE WAVELET TRANSFORM (NDWT)

We will use the non-decimated discrete Haar wavelet transform in the first stage of our method. Some basic notation and results are presented in this section. For more details about wavelet analysis, please refer to the books by Mallat (2008) [16] and Daubechies (1992) [17]. In addition, the books by Nason (2008) [18] and Vidakovic (2009) [19] are monographs about wavelet methods in statistics. For wavelet methods specialized in time series, please refer to Percival et al. (2006) [20].

The Fourier transform is used in the classical spectral analysis of stationary time series. The big disadvantage of the Fourier transform is that it has only frequency resolution but no time resolution (Figure 1). That is, although we might be able to identify all the significant frequencies in the series, we do not know when they are present and how they evolve with time. This will be a big problem when we analyze the non-stationary time series, where the frequency domain properties may change over time. To overcome this problem, several solutions have been developed, which are intended to keep track of the information from both time and frequency domains. The wavelet transform is the most recent solution to the shortcomings of the Fourier transform.

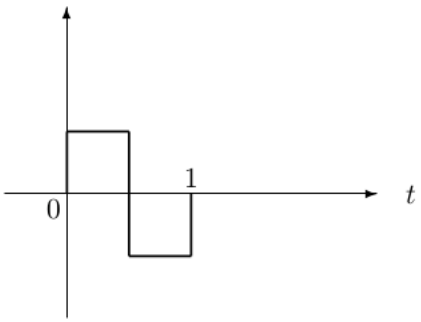
Figure 1: Time (left) and frequency (right) representations of signals



The simplest wavelet is the Haar wavelet. The Haar mother wavelet (Figure 2) is defined by

$$\psi(x) = \begin{cases} 1 & x \in [0, \frac{1}{2}), \\ -1 & x \in [\frac{1}{2}, 1), \\ 0 & otherwise. \end{cases}$$

Figure 2: Haar mother wavelet



The Haar wavelet gives us an intuition of three main characteristics of wavelets. Wavelets are oscillated, double-indexed functions with compact support (not all wavelets have compact

support, but they must decay to 0 rapidly). The double-indexing scheme enables us to obtain the information from time and frequency simultaneously. The compact support in time domain will better facilitate the capture of local behaviors in a signal.

The set of wavelets generated by dilation and translation operations, as follows,

$$\psi_{i,k}(x) = 2^{i/2}\psi(2^i x - k),$$

where i, k are integers, can form bases for various spaces of functions: e.g. $\{\psi_{i,k}(x)\}_{i,k\in\mathbb{Z}}$ is a complete orthonormal basis for $\mathcal{L}^2(\mathbb{R})$.

Thus, given a function $f(x) \in \mathcal{L}^2(\mathbb{R})$, the discrete wavelet transform (DWT) is to decompose f,

$$f(x) = \sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} d_{i,k}\psi_{i,k}(x),$$

where, due to the orthogonality of the wavelets, we have

$$d_{i,k} = \int_{-\infty}^{\infty} f(x)\psi_{i,k}(x)\,\mathrm{d}x$$

for integers i, k. The numbers, $\{d_{i,k}\}_{i,k\in\mathbb{Z}}$, are called the wavelet coefficients of f (Nason (2008) [18]).

However, the standard DWT is not shift-invariant. To be more specific, in the DWT, wavelet coefficients of a shift version (of the input data) may change drastically compared to those of the original data. Hence, the non-decimated wavelet transform (NDWT) with the desired property of shift-invariance is more favorable in time series analysis.

In the rest of this dissertation, the Meyer-Mallat scale numbering scheme will be adopted: scale 0 is the data itself, while -1 is the finest scale and $-J$ where $J = \log_2(T)$ is the coarsest scale for decomposition.

In practice, Nason et al. (2000) [5] constructed the compactly supported non-decimated discrete wavelets for discrete-time processes. Let $\{h_k\}$ and $\{g_k\}$ be the low- and high-pass quadrature mirror filters that are used in the construction of Daubechies' compactly supported wavelets. They first obtained the compactly supported discrete wavelets, $\psi_i = (\psi_{i,0}, \cdots, \psi_{i,(N_i-1)})$, of length N_i for scale $i < 0$ using the following formulae,

$$\psi_{-1,n} = \sum_k g_{n-2k}\delta_{0,k} = g_n, \quad \text{for} \quad n = 0, \cdots, N_{-1} - 1,$$

$$\psi_{i-1,n} = \sum_k g_{n-2k}\psi_{i,k}, \quad \text{for} \quad n = 0, \cdots, N_{i-1} - 1,$$

$$N_i = (2^{-i} - 1)(N_h - 1) + 1,$$

where $\delta_{0,k}$ is the Kronecker delta,

$$\delta_{0,k} = \begin{cases} 1 & k = 0, \\ 0 & k \neq 0, \end{cases}$$

and N_h is the number of non-zero elements of $\{h_k\}$.

Then, non-decimated discrete wavelets permit a wavelet to be shifted to any location but not only to "dyadic" locations as in the standard DWT. We note that the set of wavelet coefficients of DWT at each scale i is of length $2^i T$. In contrast, the wavelet coefficients of NDWT at each scale will have the same length T as the original series.

For example, the discrete Haar wavelets at scales -1 and -2 respectively are

$$\boldsymbol{\psi}_{-1} = (g_0, g_1) = (1, -1)/\sqrt{2},$$

$$\boldsymbol{\psi}_{-2} = (h_0 g_0, h_1 g_0, h_0 g_1, h_1 g_1) = (1, 1, -1, -1)/\sqrt{2},$$

and so on.

The non-decimated discrete Haar wavelet vectors at scale -1 and different locations for discrete-time series of length T are

$$\begin{aligned} \boldsymbol{\psi}_{-1,0,T} &= (1/\sqrt{2}, -1/\sqrt{2}, 0, 0, 0, \cdots, 0)', \\ \boldsymbol{\psi}_{-1,1,T} &= (0, 1/\sqrt{2}, -1/\sqrt{2}, 0, 0, \cdots, 0)', \\ \boldsymbol{\psi}_{-1,2,T} &= (0, 0, 1/\sqrt{2}, -1/\sqrt{2}, 0, \cdots, 0)', \end{aligned}$$

and so on. We can see from the above example that the non-decimated wavelets at the same scale are just shift versions of each other. The non-decimated discrete Haar wavelets at scale -2 and different locations are

$$\begin{aligned}
\boldsymbol{\psi}_{-2,0,T} &= (1/2, 1/2, -1/2, -1/2, 0, 0, 0, \cdots, 0)', \\
\boldsymbol{\psi}_{-2,1,T} &= (0, 1/2, 1/2, -1/2, -1/2, 0, 0, \cdots, 0)', \\
\boldsymbol{\psi}_{-2,2,T} &= (0, 0, 1/2, 1/2, -1/2, -1/2, 0, \cdots, 0)',
\end{aligned}$$

and so on.

3.3 BINARY SEGMENTATION

We will use the binary segmentation procedure in the second step of our proposed method. The binary segmentation procedure is widely used in dealing with multiple change-point detection problems. Venkatraman (1992) [21] employed the procedure to a sequence of independent normal variables with piecewise constant mean function, and proved that the detected change-points were consistent in terms of number and locations. Cho et al. (2012) [10] applied the procedure to a sequence of correlated scaled χ^2 variables with multiple change-points in its mean and establish the consistency results. The implementation of binary segmentation is conceptually easy. First, a single change-point is located, and further change-points are searched for to the left and right of the detected change-point. The procedure then recursively moves forward until no further changes are found.

To be more specific, we use $\{I_{t,T}\}$ to denote the input series and use $\mathcal{C}_{s,e}^{v}$ to denote the CUSUM-type operator on the interval with starting point s and ending point e:

$$\mathcal{C}_{s,e}^{v}(I_{t,T}) = \sqrt{\frac{e-v}{(e-s+1)\cdot(v-s+1)}} \sum_{t=s}^{v} I_{t,T} - \sqrt{\frac{v-s+1}{(e-s+1)\cdot(e-v)}} \sum_{t=v+1}^{e} I_{t,T}. \tag{3.2}$$

The first step of the binary segmentation procedure is to find the likely location of a change-point in the interval $(0, T-1)$ by searching for the point that maximizes the absolute value of

$$\mathcal{C}_{s,e}^{v}(I_{t,T}) = \sqrt{\frac{T-v-1}{T\cdot(v+1)}} \sum_{t=0}^{v} I_{t,T} - \sqrt{\frac{v+1}{T\cdot(T-v-1)}} \sum_{t=v+1}^{T-1} I_{t,T}. \tag{3.3}$$

Such a point, $b = \text{argmax}_v \mathcal{C}^v_{s,e}(I_{t,T})$, will be considered as the likely position of a change-point. Next, $d = \mathcal{C}^b_{s,e}(I_{t,T})$ is compared with a certain critical value to test the null hypothesis that no change-point is present. In Section 4.5 we will describe how to establish the critical value. If the null hypothesis is rejected, the algorithm will simultaneously locate and test to find further change-points in the left and right segments of b. The algorithm is repeated recursively until no further change-point is found. We use the binary segmentation algorithm in Cho et al. (2012) [10] in the second stage of our proposed method. Their algorithm is listed as follows:

1. Begin with $(j, l) = (1, 1)$. Let $s_{j,l} = 0$ and $e_{i,j} = T - 1$.

2. Iteratively compute $\mathcal{C}^v_{s_{j,l},e_{j,l}}(I_{t,T})$ as in (3.2) for $v \in (s_{j,l}, e_{j,l})$. Then, find $b_{j,l}$, which maximizes its absolute value while satisfying

$$\max\{\sqrt{(b_{j,l} - s_{j,l} + 1)/(e_{j,l} - b_{j,l})}, \sqrt{(e_{j,l} - b_{j,l})/(b_{j,l} - s_{j,l} + 1)}\} \leq c$$

for a fixed constant $c \in (0, \infty)$, where $n_{j,l} = e_{j,l} - s_{j,l} + 1$. Let $d_{j,l} = \mathcal{C}^{b_{j,l}}_{s_{j,l},e_{j,l}}(I_{t,T})$ and $m_{j,l} = \sum_{t=s_{j,l}}^{e_{j,l}} I_{t,T}/\sqrt{n_{j,l}}$.

3. Perform hard thresholding on $|d_{j,l}|/m_{j,l}$ with the threshold $t_{j,l} = \tau T^\theta \sqrt{\log T/n_{j,l}}$ so that $\hat{d}_{j,l} = d_{j,l}$ if $|d_{j,l}| > m_{j,l} \cdot t_{j,l}$, and $\hat{d}_{j,l} = 0$ otherwise.

4. If either $\hat{d}_{j,l} = 0$ or $\max\{b_{j,l} - s_{j,l}, e_{j,l} - b_{j,l} + 1\} < \Delta_T$ for l, stop the algorithm on the interval $[s_{j,l}, e_{j,l}]$; if neither, let $(s_{j+1,2l-1}, e_{j+1,2l-1}) = (s_{j,l}, b_{j,l})$ and $(s_{j+1,2l}, e_{j+1,2l}) = (b_{j,l} + 1, e_{j,l})$, and update the level j as the level $j \to j + 1$.

5. Repeat Steps 2-4.

3.4 LOCAL SPECTRAL ENVELOPE

The local spectral envelope analysis will be employed in the post-processing step of our proposed method. Notation, general idea, as well as basic theoretic results of the local spectral envelope methodology are given in this section. For more details, please refer to Stoffer et al. (1993) [1] and Stoffer et al. (2002) [12].

If $\{Y_{t,T}\}$ is a piecewise stationary categorical time series with known stationary segmentation, the local spectral envelope in Stoffer et al. (2002) [12] is defined as

$$\lambda_b(\omega) = \sup_{\boldsymbol{\beta} \not\propto \mathbf{1}} \{ \frac{\boldsymbol{\beta}' f^{re}_{Z,b}(\omega) \boldsymbol{\beta}}{\boldsymbol{\beta}' V_b \boldsymbol{\beta}} \}, \tag{3.4}$$

where V_b is the variance-covariance matrix of $\boldsymbol{Z}_{t,b}$ and $b = 1, \ldots, B$. The corresponding eigenvector $\boldsymbol{\beta}_b(\omega)$ is called the local optimal scaling of block b and frequency ω.

$\lambda_b(\omega)\,\mathrm{d}\omega$ can be considered as the largest proportion of the total power in block b that can be attributed to the frequencies within a $\mathrm{d}\omega$ neighborhood of ω for any particular scaled process, $X_{t,b}(\boldsymbol{\beta}) = \boldsymbol{\beta}\boldsymbol{Z}_{t,b}$. Thus, the value $\lambda_b(\omega)$ has a meaningful interpretation: it envelopes the standardized spectrum of any scaled process. That is to say, given any $\boldsymbol{\beta}$ normalized so that $X_{t,b}(\boldsymbol{\beta})$ has total power of 1, $f_b(\omega; \boldsymbol{\beta}) \leq \lambda_b(\omega)$ with equality if and only if $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\beta}(\omega)$, where $f_b(\omega; \boldsymbol{\beta})$ is the spectrum of $X_{t,b}(\boldsymbol{\beta})$. The name "spectral envelope" comes from this appealing interpretation. Furthermore, information is lost when one restricts attention to the spectrum of $X_{t,b}(\boldsymbol{\beta})$; less information is lost when one considers the spectrum of $\boldsymbol{Z}_{t,b}$. Directly dealing with the spectral density $f_{\boldsymbol{Z},b}(\omega)$ is cumbersome since it is a complex Hermitian matrix with each element as a function. From this perspective, spectral envelope can be thought of as a parsimonious tool for exploring the periodic nature of a categorical time series with minimal loss of information.

Given an estimate, $\hat{f}_{Z,b}(\omega)$ of $f_{Z,b}(\omega)$, the estimate of the local spectral envelope $\hat{\lambda}_b(\omega)$ is defined as the largest eigenvalue of $\hat{g}^{re}_b(\omega)$, where

$$\hat{g}^{re}_b(\omega) = \hat{V}_b^{-1/2} \hat{f}_{Z,b}(\omega) \hat{V}_b^{-1/2}. \tag{3.5}$$

The local sample optimal scaling, $\hat{\boldsymbol{\beta}}_b(\omega)$, is then defined by $\hat{\boldsymbol{\beta}}_b(\omega) = \hat{V_b}^{-1/2} \hat{\boldsymbol{u}}_b(\omega)$, where $\hat{\boldsymbol{u}}_b(\omega)$ is the normalized eigenvector associated with $\hat{\lambda}_b(\omega)$.

In the rest of this section, we will present asymptotic results for estimators of the local spectral envelope and the corresponding local scaling vectors established by Stoffer et al. (2002) [12]. Theorem 3.4.1 displays the results of local spectral envelope estimators based on the local periodogram $I_b(\omega)$ (see Equation(3.6)). The consistent window spectral estimate $\hat{f}_{Z,b}(\omega)$ (see Equation (3.8)) is chosen as the estimate of $f_{Z,b}(\omega)$ in Theorem 3.4.2.

The local periodogram $I_b(\omega)$ is given by

$$I_b(\omega) = \boldsymbol{d}_b(\omega)\boldsymbol{d}_b^*(\omega), \tag{3.6}$$

where

$$\boldsymbol{d}_b(\omega) = M_b^{-1/2} \sum_{t=1}^{Mb-1} \boldsymbol{Z}_{t,b} \exp\{-2\pi i t\omega\} \tag{3.7}$$

is the finite Fourier transform of the data $\{\boldsymbol{Z}_{s,T} : s/T \in U_b\}$.

Let $W[p, \nu, \Sigma]$ denote the Wishart distribution of dimension p on ν degrees of freedom and with $p \times p$ covariance Σ; $W_c[p, \nu, \Sigma]$ denotes the complex Wishart distribution. Details about Wishart distribution can be found in Brillinger (2001) [22].

Theorem 3.4.1. *Under the established notation and conditions, and for* $\hat{f}_{Z,b}(\omega) = I_b(\omega)$, *the collection,* $\{\hat{\lambda}_b(\omega_j), \hat{\boldsymbol{\beta}}_b(\omega_j) : j = 1, \cdots, J\}$, *converges in distribution to* $\lambda_{b,j}, \boldsymbol{\beta}_{b,j} : j = 1, \cdots, J\}$, *where* $\{\boldsymbol{\beta}_{b,j} = V_b^{-1/2}\boldsymbol{u}_{b,j}$ *and* $\{\lambda_{b,j}, \boldsymbol{u}_{b,j} : j = 1, \cdots, J\}$ *are the largest eigenvalue and eigenvector of independent* $W_c^{re}[p-1, 1, g_b(\omega_j)]$ *matrices, with* $\boldsymbol{u}_{b,j}$ *normalized so that* $\boldsymbol{u}'_{b,j}\boldsymbol{u}_{b,j} = 1$ *and the first nonzero entry of* $\boldsymbol{u}_{b,j}$ *is positive.*

Finally, we consider local consistent window spectral estimates. Consider a window function, $W_b(\alpha), -\infty < \alpha < \infty$, that is real-valued, even and of bounded variation, where $\int_{-\infty}^{\infty} W_b(\alpha)\,\mathrm{d}\alpha = 1$, and $\int_{-\infty}^{\infty} |W_b(\alpha)|\,\mathrm{d}\alpha < \infty$. $\hat{f}_{Z,b}(\omega)$ is defined as

$$\hat{f}_{Z,b}(\omega) = M_b^{-1} \sum_{l=0}^{M_b-1} W_{M_b}(\omega - l/M_b) I_b(l/M_b), \tag{3.8}$$

where $W_{M_b}(\alpha) = B_{M_b}^{-1} \sum_{j=-\infty}^{\infty} W_b(B_{M_b}^{-1}[\alpha + j])$ and B_{M_b} is a bounded sequence of non-negative scale parameters such that $B_{M_b} \to 0$ and $B_{M_b}M_b \to \infty$ as $T \to \infty$. Let $\nu_{M_b} = (B_{M_b}M_b)^{1/2}(\int_{-\infty}^{\infty} W_b(\alpha)^2\,\mathrm{d}\alpha)^{-1/2}$.

Theorem 3.4.2. *Under the stated conditions and assumptions, and for* $\hat{f}_{Z,b}(\omega)$ *defined by* (3.8), *if for each* $j = 1, \cdots, J$, *the largest root of* $g_b^{re}(\omega_j)$ *is distinct, then* $\{\nu_{M_b}[\hat{\lambda}_b(\omega_j) - \lambda_b(\omega_j)]/\lambda_b(\omega_j); \nu_{M_b}[\hat{\boldsymbol{\beta}}_b(\omega_j) - \boldsymbol{\beta}_b(\omega_j)] : j = 1, \cdots, J\}$ *converges jointly in distribution to* $\{z_j; \boldsymbol{y}_j : j = 1, \cdots, J\}$ *with* z_j *and* $\boldsymbol{y}_j$ *being independent for* $j = 1, \cdots, J$. *Furthermore,*

for each $j = 1, \cdots, J$, z_j *has a standard normal distribution and is independent of* $\boldsymbol{y}_j$ *which is multivariate normal with mean* 0. *The covariance matrix of* $V_b^{1/2}\boldsymbol{y}_j$ *is given by*

$$\{\lambda_b(\omega_j)H_b(\omega_j)^+ g_b^{re}(\omega_j)H_b(\omega_j)^+ - \boldsymbol{a}_b(\omega_j)\boldsymbol{a}_b(\omega_j)'\}/2, \tag{3.9}$$

where $H_b(\omega_j) = g_b^{re}(\omega_j) - \lambda_b(\omega_j)\mathbf{I}_{p-1}$, $\boldsymbol{a}_b(\omega_j) = H_b(\omega_j)^+ g_b^{im}(\omega_j)V_b^{1/2}\boldsymbol{u}_b(\omega_j)$, $\mathbf{I}_{p-1}$ *denotes the* $(p-1)\times(p-1)$ *identity matrix, and* $H_b(\omega_j)^+$ *refers to the Moore-Penrose inverse of* $H_b(\omega_j)$.

In practice, the following approximations work well when M_b is large, according to Stoffer et al. (2002) [12]. By using a first-order Taylor expansion

$$\log\hat{\lambda}_b(\omega) \approx \log\lambda_b(\omega) + \frac{\hat{\lambda}_b(\omega) - \lambda_b(\omega)}{\lambda_b(\omega)}, \tag{3.10}$$

thus, $\nu_{M_b}[\log\hat{\lambda}_b(\omega) - \log\lambda_b(\omega)]$ is approximately standard normal under appropriate conditions. It also follows that $E[\log\hat{\lambda}_b(\omega)] \approx \log\lambda_b(\omega)$ and $\text{var}[\log\hat{\lambda}_b(\omega)] \approx \nu_{M_b}^{-2}$. Stoffer et al.'s (2002) [12] simulations showed that the average value of $\hat{\lambda}_b(j/M_b)$ is closer to $2.5/M_b$ when there is no signal present. According to this recommended value, the α critical value for $\hat{\lambda}_b(\omega)$ will be $(2.5/M_b)\exp(z_\alpha/\nu_{M_b})$. When analyzing DNA sequence data, M_b is suggested by Stoffer et al. (2002) [12] to be at least 2^8.

4.0 PROPOSED METHOD

In this chapter, we present our proposed method for dealing with multiple change-point problems in piecewise stationary categorical time series. We first put forth the general idea behind our proposed method in Section 4.1. The detailed steps of our proposed algorithm are given in Section 4.2. Consistency result of our proposed method is presented in Section 4.3.

When the stationary segmentation is unknown, previous notation needs to be slightly modified to enable the time-varying property. Let $\{Y_{t,T}\}_{t=0}^{T-1}$ be a categorical-valued time series with finite state-space $\mathcal{C} = \{c_1, \cdots, c_p\}$. For $\boldsymbol{\beta}_t = (\beta_{t,1}, \beta_{t,2}, \cdots, \beta_{t,p})' \in \mathbb{R}^p$, let $X_{t,T}(\boldsymbol{\beta}_t)$ denote the real-valued time series corresponding to the scaling that assigns the category c_j the numerical value $\beta_{t,j}$ at time t. Then $X_{t,T}(\boldsymbol{\beta}_t) = \boldsymbol{\beta}_t' \boldsymbol{Z}_{t,T}$, where $\boldsymbol{Z}_{t,T}$ is the corresponding $p \times 1$ point process. Let $\mathbf{Z}_T = (\boldsymbol{Z}_{0,T}, \boldsymbol{Z}_{1,T}, \cdots, \boldsymbol{Z}_{T-1,T}) = (\boldsymbol{Z}_T^{(1)}, \cdots, \boldsymbol{Z}_T^{(p)})'$, where $\boldsymbol{Z}_T^{(j)} = (Z_{0,T}^{(j)}, Z_{1,T}^{(j)}, \cdots, Z_{T-1,T}^{(j)})'$. The $p \times T$ matrix $\mathbf{Z}_T$ represents the data matrix. To simplify, Y always represents categorical time series, $\boldsymbol{Z}$ is a point process, and X denotes numerical series in this dissertation.

4.1 KEY IDEA

Assigning different values to the categories will bring out different features of the categorical data. Instead of arbitrarily assigning numerical values, we focus on scalings that could best assist us in locating the change-points. If there is a change in the structure of $Y_{t,T}$, the change occurs at the same location in $\boldsymbol{Z}_{t,T}$. We believe that for certain $\boldsymbol{\beta}_{i,t}$, the changes in the original categorical series could be detected by examining $X_{t,T}(\boldsymbol{\beta}_{i,t}) = \boldsymbol{\beta}_{i,t}' \boldsymbol{Z}_{t,T}$. It should be noted that our goal is change-point detection. Thus, we do not necessarily need

to estimate the exact scalings $\boldsymbol{\beta}_{i,t}$. For certain $\boldsymbol{\beta}_{i,t}$, the wavelet coefficient of $X_{t,T}(\boldsymbol{\beta}_{i,t})$ is

$$w_{i,t,T}(\boldsymbol{\beta}_{i,t}) = \sum_k X_{k,T}(\boldsymbol{\beta}_{i,t})\psi_{i,k}(t) = \boldsymbol{\beta}'_{i,t}\mathbf{Z}_T\boldsymbol{\psi}_{i,t,T},$$

where $\boldsymbol{\psi}_{i,t,T}$'s are the non-decimated discrete Haar wavelets discussed in Section 3.2, and $\psi_{i,k}(t)$ is the kth element of vector $\boldsymbol{\psi}_{i,t,T}$. The raw wavelet periodogram (RWP) is:

$$I_{i,t,T}(\boldsymbol{\beta}_{i,t}) = |w_{i,t,T}(\boldsymbol{\beta}_{i,t})|^2 = \boldsymbol{\beta}'_{i,t}\mathbf{Z}_T\boldsymbol{\psi}_{i,t,T}\boldsymbol{\psi}'_{i,t,T}\mathbf{Z}'_T\boldsymbol{\beta}_{i,t}.$$

Here, rather than using the same scaling for all scales, we allow that the scalings are different in i and t, and denote it by $\boldsymbol{\beta}_{i,t}$. We are only interested in those $\boldsymbol{\beta}_{i,t}$'s in which $\boldsymbol{\beta}_{i,t} \not\propto \mathbf{1}_p$ and $\boldsymbol{\beta}'_{i,t}\boldsymbol{\beta}_{i,t} = 1$. Thus, the RWP can be represented as:

$$I_{i,t,T}(\boldsymbol{\beta}_{i,t}) = \frac{\boldsymbol{\beta}'_{i,t}\mathbf{Z}_T\boldsymbol{\psi}_{i,t,T}\boldsymbol{\psi}'_{i,t,T}\mathbf{Z}'_T\boldsymbol{\beta}_{i,t}}{\boldsymbol{\beta}'_{i,t}\boldsymbol{\beta}_{i,t}}.$$

The idea is to focus on scalings such that at each scale i and time t, the raw wavelet sequences $I_{i,t,T}(\boldsymbol{\beta}_{i,t})$'s achieve their maximum values among arbitrary scalings where $\boldsymbol{\beta}_{i,t}$'s are subject to $\boldsymbol{\beta}'_{i,t}\boldsymbol{\beta}_{i,t} = 1$. Let $\Sigma_{I_{i,t,T}} = \mathbf{Z}_T\boldsymbol{\psi}_{i,t,T}\boldsymbol{\psi}'_{i,t,T}\mathbf{Z}'_T$, i.e.,

$$\Sigma_{I_{i,t,T}} = \begin{pmatrix} I^{(1)}_{i,t,T} & I^{(1,2)}_{i,t,T} & \cdots & I^{(1,p)}_{i,t,T} \\ I^{(1,2)}_{i,t,T} & I^{(2)}_{i,t,T} & \cdots & I^{(2,p)}_{i,t,T} \\ \vdots & \vdots & \ddots & \vdots \\ I^{(1,p)}_{i,t,T} & I^{(2,p)}_{i,t,T} & \cdots & I^{(p)}_{i,t,T} \end{pmatrix},$$

where $I^{(j)}_{i,t,T} = \boldsymbol{Z}^{(j)\prime}_T\boldsymbol{\psi}_{i,t,T}\boldsymbol{\psi}'_{i,t,T}\boldsymbol{Z}^{(j)}_T$ and $I^{(j,l)}_{i,t,T} = \boldsymbol{Z}^{(j)\prime}_T\boldsymbol{\psi}_{i,t,T}\boldsymbol{\psi}'_{i,t,T}\boldsymbol{Z}^{(l)}_T$.

Let $\lambda_{i,t,T}$ denote the largest eigenvalue of $\Sigma_{I_{i,t,T}}$. It follows immediately that

$$\lambda_{i,t,T} = \sup_{\boldsymbol{\beta}'_{i,t}\boldsymbol{\beta}_{i,t}=1} I_{i,t,T}(\boldsymbol{\beta}_{i,t}).$$

Note that $\text{rank}(\boldsymbol{\Sigma}_{I_{i,t,T}}) = 1$. We have $\lambda_{i,t,T} = \text{tr}(\boldsymbol{\Sigma}_{I_{i,t,T}}) = \sum_{j=1}^{p} I^{(j)}_{i,t,T}$.

To be consistent with the convention that random variables are usually written in upper case letters, let $I_{i,t,T} \equiv \lambda_{i,t,T} = \sum_{j=1}^{p} I^{(j)}_{i,t,T}$. $I_{i,t,T}$'s will be the input sequences for further binary segmentation.

Next, we discuss the logic explaining that breakpoints in the autocovariance structure can be detected by examining $I_{i,t,T}$'s. This dissertation focuses on what is arguably one of the simplest forms of departure from non-stationarity, piecewise stationarity, where the dependence structure is piecewise constant. The nonparametric model we use for this purpose is the multivariate locally stationary wavelet (LSW) process. The LSW process was first developed by Nason et al. (2000) [5] and was later extended to the bivariate case by Sanderson et al. (2010) [23]. Recently, Cho et al. (2014) [11] established the LSW process for the multivariate case. The p-variate LSW process $\{\boldsymbol{Z}_{t,T} = (Z_{t,T}^{(1)}, \ldots, Z_{t,T}^{(p)})'\}_{t=0}^{T-1}$ has the following representation:

$$Z_{t,T}^{(j)} = \sum_{i=-\infty}^{-1} \sum_{k=-\infty}^{\infty} W_i^{(j)}(k/T)\psi_{i,k}(t)\xi_{i,k}^{(j)} \quad \text{for each } j = 1, \ldots, p,$$

where $\boldsymbol{\xi}_{i,k} = (\xi_{i,k}^{(1)}, \xi_{i,k}^{(2)}, \cdots, \xi_{i,k}^{(p)})'$ are orthonormal, identically distributed random variables.

Cho et al. (2014) [11] defined the local autocovariance function as

$$c^{(j)}(z, h) \coloneqq \sum_i S_i^{(j)}(z)\Psi_i(h),$$

and the evolutionary wavelet spectrum as

$$S_i^{(j)}(z) \coloneqq |W_i^{(j)}(z)|^2,$$

where $\Psi_i(h) = \sum_k \psi_{i,k}\psi_{i,k}(h)$. They proved that

$$c^{(j)}(z, h) = \lim_{T\to\infty} \operatorname{cov}(Z_{[zT],T}^{(j)}, Z_{[zT]+h,T}^{(j)}),$$

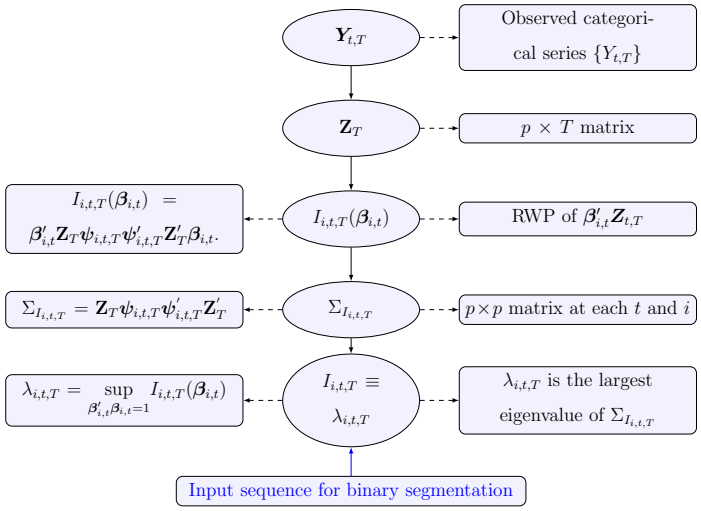
and

$$S_i^{(j)}(z) = \sum_{i'} A_{i,i'}^{-1} \sum_h c^{(j)}(z, h)\Psi_{i'}(h),$$

where $A_{i,i'} = \sum_h \Psi_i(h)\Psi_{i'}(h)$.

Although the collection $\{\Psi_i(h)\}$ is not orthogonal, it is indeed a linearly independent system. Therefore, $\{\Psi_i(h)\}$ can still serve as a wavelet basis for an non-orthogonal expansion. The redundancy in the non-orthogonal autocorrelation wavelet family $\{\Psi_i(h)\}$ is measured by $\mathbf{A} = (A_{i,i'})_{i,i'<0}$, which Nason et al. (2000) [5] first introduced. They also showed that for all Daubechies compactly supported wavelets, $\mathbf{A}$ is an invertible operator. And for each

Figure 3: Input sequence for binary segmentation at scale i



J, the norm of $\mathbf{A}_J^{-1}$ is bounded from above by some constant, where J-dimensional matrix $\mathbf{A}_J \coloneqq (A_{i,i'})_{i,i'=-1,\ldots,-J}$.

Thus, $S_i^{(j)}(z)$ can be viewed as the linear combination of $c^{(j)}(z,h)$'s. That is to say, every breakpoint in $c^{(j)}(z,h)$ will result in a breakpoint in $S_i^{(j)}(z)$ at one or multiple scales. Further, $\iota_i^{(j)}(z) = \sum_{i'} S_{i'}^{(j)}(z)A_{i,i'}$, and $T^{-1}\sum_{t=0}^{T-1} \mid \mathbb{E}I_{i,t,T}^{(j)} - \iota_i^{(j)}(\frac{t}{T}) \mid^2 \to 0$ (Proposition 2 in Cho et al. (2014) [11]). Therefore, $\mathbb{E}I_{i,t,T}^{(j)}$ is close enough to the piecewise constant function $\iota_i^{(j)}(\frac{t}{T})$ in the sense that the integrated squared bias between them converges to 0. Actually, $\mathbb{E}I_{i,t,T}^{(j)}$ is piecewise constant other than on the intervals around jumps of $\iota_i^{(j)}(\frac{t}{T})$.

4.2 PROPOSED ALGORITHM

Detailed steps of our algorithm are listed as follows:

Step 1: *Prepare the input sequences (Flow chart in Figure 3).*

Get the input sequences $\{I_{i,t,T}, t = 0, 1, \cdots, T-1\}$ at scales $i = -I^*, \ldots, -1$ for further binary segmentation, where $I^* = -\lfloor \frac{\log_2 T}{2} \rfloor$.

Step 2: *Perform binary segmentation.*

For each i, use the binary segmentation procedure described in Cho et al. (2012) [10] (Section 3.3) to examine the input sequence $\{I_{i,t,T}, t = 0, 1, \cdots, T-1\}$ at scales $i = -I^*, \ldots, -1$. Denote the detected change-points at scale i as $\hat{\mathcal{B}}_i = \{\hat{\eta}_q^{(i)}, q = 1, \ldots, \hat{N}_i\}$.

(Optional): *Monitor eigenvectors $\boldsymbol{\beta}_{i,t}$ associated with the largest eigenvalue.*

Augment: At each scale i, calculate the distance of $\boldsymbol{\beta}_{i,t}$ and $\boldsymbol{\beta}_{i,t-1}$, i.e: $dist(\boldsymbol{\beta}_{i,t}) = \|\boldsymbol{\beta}_{i,t} - \boldsymbol{\beta}_{i,t-1}\|$. Then, perform the binary segmentation on $dist(\boldsymbol{\beta}_{i,t})$ sequence at each scale i. Denote the detected change-points by $\hat{\mathcal{B}}_i^{\text{aug}} = \{\hat{\eta}_q^{\text{aug}(i)}, q = 1, \ldots, \hat{N}_i^{\text{aug}}\}$. Arrange all detected change-points in $\hat{\mathcal{B}}_i \cup \hat{\mathcal{B}}_i^{\text{aug}}$ into groups so that those within the distance of Λ_T from each other are classified in the same group. If there exists $\hat{\eta}_{q_0}^{\text{aug}(i)} \in \hat{\mathcal{B}}_i^{\text{aug}}$, such that $\hat{\eta}_{q_0}^{\text{aug}(i)}$ itself forms a group, add $\hat{\eta}_{q_0}^{\text{aug}(i)}$ to the set $\hat{\mathcal{B}}_i$. For simplicity, retain the notation of the set of the detected change-points at scale i after this step as $\hat{\mathcal{B}}_i = \{\hat{\eta}_q^{(i)}, q = 1, \ldots, \hat{N}_i\}$.

Step 3: *Combine the detected change-points across scales (here we use the across-scales post-processing step in Cho et al. (2012) [10]).*

3.1: Group the detected change-points in $\bigcup_{i=-I^*,\cdots,-1} \hat{\mathcal{B}}_i$ such that those close enough to each other (within distance $\Lambda_T = \lfloor \frac{\epsilon_T}{2} \rfloor$) are classified into the same group; denote the groups by $\mathcal{G}_1, \cdots, \mathcal{G}_{\hat{B}}$.

3.2: Find the scale index i_0 with the maximum number of detected change-points. If there is more than one scale having the same maximum number of detected change-points, choose whichever is the finest scale, i.e. $i_0 = \max\{\text{argmax}_{-I* \le i \le -1} \hat{N}_i\}$.

3.3: If $\hat{N}_{i_0} = \hat{B}$, and for any $b = 1, \cdots, \hat{B}$, there exists $\hat{\eta}_{q_0}^{(i_0)} \in \mathcal{G}_b$, set $\hat{\mathcal{B}} = \hat{\mathcal{B}}_{i_0}$. Otherwise, proceed to **3.4**.

3.4: Set $\hat{\mathcal{B}} = \{\hat{\eta}_b, b = 1, \ldots, \hat{B}\}$, where $\hat{\eta}_b \in \mathcal{G}_b$ with the maximum scale index i.

Step 4: *Post-process using spectral envelope technique.*

4.1: Calculate the local sample spectral envelope $\hat{\lambda}_b(\omega_k)$ at each fundamental frequency $\omega_k = \frac{k}{\hat{M}_b}, k = 0, \cdots, \hat{M}_b/2, \hat{M}_b = \hat{\nu}_b - \hat{\nu}_{b-1} + 1$ for each estimated segment $[\hat{\nu}_{b-1}, \hat{\nu}_b)$. Let $\hat{\nu}_0 = 0, \hat{\nu}_{\hat{B}+1} = T - 1$.

4.2: Define a discrepancy measure $D[\cdot, \cdot]$ between the spectral envelope estimates of two adjacent intervals $[\hat{\nu}_{b-1}, \hat{\nu}_b), [\hat{\nu}_b, \hat{\nu}_{b+1})$. If $D[\hat{\lambda}_{b+1}(\omega), \hat{\lambda}_b(\omega)] < D_r$, remove $\hat{\eta}_b$ from $\hat{\mathcal{B}}$.

4.3: Repeat **4.2** with the reduced set of change-points until the set does not change. Denote the set of estimated change-points after this step as $\hat{\mathfrak{B}} = \{\hat{\nu}_b^{\text{post}}, b = 1, \ldots, \hat{B}^{\text{post}}\}$.

Step 5: *Complete classification (for DNA sequence data).*

Use the information in the estimated local spectral envelope based on the previous segmentation to classify a segment as (i) highly likely to contain coding sequence, (ii) containing noncoding, or (iii) uncertain. The recommended classification rule established by Stoffer et al. (2001) [12] is summarized in Section 4.4 and will be used in our real data example.

Before proceeding, we note the following:

- *On threshold in binary segmentation*:

The input sequences we use are the largest eigenvalue sequences $I_{i,t,T}$'s, with different statistical properties from those in Cho et al. (2012) [10]. Thus, the critical values need to be redesigned. In Section 4.5 we will describe how to establish those critical values. We will show in the next section that the null hypothesis is rejected with probability converging to 1 under our assumptions, when a change-point exists.

- *On optional augmentation*:

The optional augmentation of monitoring $\boldsymbol{\beta}_{i,t}$ is done to ensure that, even under some extreme situations, we could still capture the structural changes of the data. Normally, the largest eigenvalue λ_1 of a matrix $\mathbf{M}$ will change as the structure of $\mathbf{M}$ changes. However, in some special cases, λ_1 will stay the same even when several elements of $\mathbf{M}$ change significantly. For example, suppose the symmetric matrix has changed from $\mathbf{M}_1$ to $\mathbf{M}_2$, where

$$\mathbf{M}_1 = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix}.$$

Both $\mathbf{M}_1$ and $\mathbf{M}_2$ have the same largest eigenvalue 6.85. The eigenvector associated with the largest eigenvalue has changed from $(-0.85, -0.53)'$ to $(0.53, 0.85)'$. Under those special situations, there will be a peak in the distance of the adjacent $\boldsymbol{\beta}_{i,t}$'s. In all types of cases, as long as the structure of a matrix changes, either the largest eigenvalue or the associated eigenvector will change. We mainly focus on examining the largest eigenvalue sequence because it is relatively less complex. One can skip this step if the above situation can be eliminated after carefully evaluating the real cases.

- *On computational efficiency*:

It is necessary to use computationally efficient methods when analyzing very long time series data sets. The computational complexity of binary segmentation is typically of order $O(T \log T)$. Both the non-decimated wavelet transform in **Step 1** and the estimates of spectral envelope in **Step 4** can be implemented by the Fast Fourier Transform (FFT), which is also of order $O(T \log T)$. With regard to obtaining the largest eigenvalue sequences in **Step 1**, as we pointed out earlier that $I_{i,t,T} = \mathrm{tr}(\mathbf{\Sigma}_{I_{i,t,T}})$, trace calculation is fast. Overall, for small p, our proposed algorithm is efficient.

- *On discrepancy measure in **Step 4***:

In this dissertation we use the distance measure recommended by Stoffer et al. (2002) [12]. Any appropriate distance measure between spectra such as Kolmogorov-Smirnov distance or Cramér Von-Mises distance described in Adak (1998) [7] may also be used.

- *Relationship to Stoffer algorithm*:

The Stoffer et al. (2002) [12] method can be viewed as "bottom-up." They first determined the smallest possible size of the segmented blocks and then combined the blocks with similar local spectral envelope information. Our algorithm is "top-down-up" in the sense that binary segmentation is "top-down," and the post-processing step goes "up" again by combining similar blocks. By going through the "top-down-up" track, we first take advantage of the good time resolution of the non-decimated wavelet technique, we then benefit from the good frequency resolution by employing the local spectral envelope methodologies. Therefore, we could gain more flexibility in estimating the locations of change-points, as well as retain good performance in capturing the signal at very narrow bands of frequency.

4.3 CONSISTENCY RESULT

Consistency results of the binary segmentation procedure have been established in different situations. Venkatraman (1992) [21] showed the consistency of the binary segmentation

procedure when it was applied to a sequence of independent normal variables with multiple breakpoints in the sequence's mean function. Cho et al. (2012) [10] proved the consistency results for the multiplicative model where dependence exists among observations. Under our assumptions, variables in the input sequence for further binary segmentation are correlated to each other and are non-normal. In this section, we would like to extend the consistency results under our assumptions.

Assumption 1. *Let* $c(z,h) = \sum_{j=1}^{p} c^{(j)}(z,h)$. *Assume that c(z,h)'s are piecewise constant functions with a finite number of change-points in set*

$$\begin{aligned}\mathbb{B} &\equiv \{u_b, b = 1, \ldots, B\} \\ &= \{u_b \in (0,1) : \exists h, \text{ such that } \lim_{z \to u_b^-} c(z,h) \neq \lim_{z \to u_b^+} c(z,h)\}.\end{aligned}$$

The one-to-one correspondence under Assumption 1 is summarized in Figure 4, where $S_i(z) = \sum_{j=1}^{p} S_i^{(j)}(z)$, and $\iota_i(z) = \sum_{j=1}^{p} \iota_i^{(j)}(z)$. Note that, $\iota_i(z) = \sum_{j=1}^{p} \sum_{i'} S_{i'}^{(j)}(z) A_{i,i'} = \sum_{i'} S_{i'}(z) A_{i,i'}$; and that

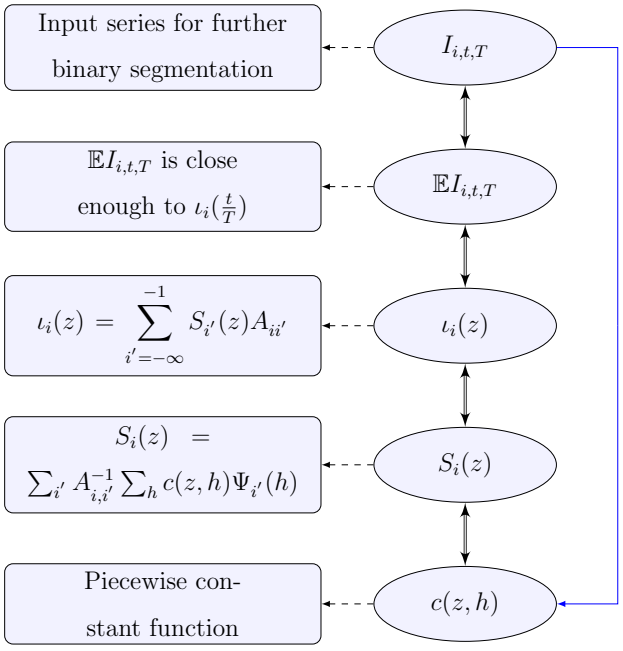
$$\begin{aligned} & T^{-1} \sum_{t=0}^{T-1} | \, \mathbb{E} I_{i,t,T} - \iota_i(\frac{t}{T}) \, |^2 \\ = \; & T^{-1} \sum_{t=0}^{T-1} | \sum_{j=1}^{p} \mathbb{E} I_{i,t,T}^{(j)} - \sum_{j=1}^{p} \iota_i^{(j)}(\frac{t}{T}) \, |^2 \\ \leq \; & T^{-1} \sum_{t=0}^{T-1} \sum_{j=1}^{p} | \, \mathbb{E} I_{i,t,T}^{(j)} - \iota_i^{(j)}(\frac{t}{T}) \, |^2 \\ \to \; & 0, \end{aligned} \tag{4.1}$$

as long as $p \ll T$ as $T \to \infty$. Denote the set of change-points in the original time domain as:

$$\mathcal{B} \equiv \{\nu_b : \nu_b = \lfloor u_b T \rfloor, b = 1, \ldots, B\}.$$

Based on the one-to-one correspondence discussed above and under Assumption 1, for each i, $\iota_i(z)$ is a piecewise constant function. Denote the number of breakpoints in $\iota_i(t/T)$ by N_i and the breakpoints themselves by $0 < \eta_1^{(i)}, \ldots, \eta_{N_i}^{(i)} < T-1$, with $\eta_0^{(i)} = 0$, $\eta_{N_i+1}^{(i)} = T-1$ for all i. Let $\mathcal{B}_i = \{\eta_q^{(i)}, q = 1, \ldots, N_i\}$. Note that $\mathcal{B} = \cup_i \mathcal{B}_i$.

Figure 4: Illustration of one-to-one correspondence



This dissertation focuses on changes in the autocovariance structure. Possible extensions of our method, based on complex wavelets, could further enable the consideration of changes in both autocovariance and cross-covariance structures.

Assumption 2. *Assume that the minimum length of each stationary segment is bounded by* δ_T*, where* $\delta_T \asymp T^{\Theta}$*,* $\Theta \in (\theta + \frac{1}{2}, 1)$ *and* $\theta \in (\frac{1}{4}, \frac{1}{2})$*. Further, there exists a positive constant* c*, such that*

$$\max_q \{ \frac{\eta_q^{(i)} - \eta_{q-1}^{(i)} + 1}{\eta_{q+1}^{(i)} - \eta_q^{(i)}}, \frac{\eta_{q+1}^{(i)} - \eta_q^{(i)}}{\eta_q^{(i)} - \eta_{q-1}^{(i)} + 1} \} \leq c \quad \textit{for each scale } i.$$

Assumption 2 ensures that each segment is long enough for the purpose of examining and that the change-points occur at "balanced" locations so that one segment is not much longer than its neighboring segment.

Assumption 3. *Assume that there exist* $M_\mu, m_\mu > 0$ *such that, for each scale* i*,*

$$\inf_q \mid \iota_i(\eta_q^{(i)}/T) - \iota_i(\eta_{q+1}^{(i)}/T) \mid > m_\mu \quad \textit{and} \quad \sup_t |\iota_i(t/T)| \leq M_\mu.$$

Assumption 4. *Let* $\rho^{(j)}(h) = \sup_{z,T} |cor(Z^{(j)}_{[zT],T}, Z^{(j)}_{[zT]+h,T})|$ *and* $\rho^{(j)}_{\infty} = \sum_h \rho^{(j)}(h)$, *where* $j = 1, \ldots, p$. *There exists a positive constant* $M^{(j)}_{\rho}$, *such that* $\rho^{(j)}_{\infty} \leq M^{(j)}_{\rho}$.

Assumption 4 requires the absence of long memory in the original data, which makes more sense when we segment the data based on the autocovariance structure.

In practice, if the data exhibit long-range dependency in addition to local features, we could first apply a high-pass filter or detrend the series to eliminate the long memory before the segmentation procedure.

In fact, in the proof of our consistency result, we only require the absence of long memory in the wavelet periodogram sequences $I^{(j)}_{i,t,T}$'s (see (A.2)). A nice feature of wavelet transform is the whitening property, also known as the decorrelating property, which is to say that the correlation between wavelet coefficients both within and between scales will generally be small even if the data are highly autocorrelated. Dijkerman et al. (1994) [24], Flandrin (1992) [25], Tewfik et al. (1992) [26], and Wornell (1993) [27] introduced the whitening property theoretically for the fractional Brownian motion process. Later, Fan (2003) [28] showed that the DWT has optimally decorrelating properties for the wider class of signals with $1/f$-like power spectral density functions. Thus, the wavelet periodogram sequences are often much less autocorrelated than the original data, and (A.2) is a much easier condition to satisfy in practice than is Assumption 4.

Theorem 4.3.1. *Suppose* $\{\boldsymbol{Z_{t,T}}\}_{t=0}^{T-1}$ *is a p-variate LSW process. Under Assumptions 1, 2, 3, and 4, and for p fixed, the number and locations of the detected change-points are consistent. That is,* $\mathbf{P}\{\hat{B} = B; |\hat{\nu}_b - \nu_b| \leq C\epsilon_T, 1 \leq b \leq B\} \to 1$ *as* $T \to \infty$, *where* $\hat{\nu}_b, b = 1, \cdots, \hat{B}$ *are detected change-points and* $\epsilon_T = T^{\frac{1}{2}} \log T$.

4.4 CLASSIFICATION RULE FOR GENE DETECTION

Based on the extensive experience with the Fourier analysis of DNA sequences (e.g. Stoffer et al. (1993) [1]; Cornette et al. (1987) [29]; Tiwari et al. (1997) [30]), Stoffer et al. (2002) [12] established the following classification rules:

- A block is designated as containing only coding if the local estimated spectral envelope exhibits a peak at frequency 1/3 (and possibly other nonzero frequencies like 1/10), but no peak exists at zero frequency.
- A block is designated as containing both coding and noncoding if the spectral envelope exhibits a peak at (or near) the zero frequency as well as a peak at frequency 1/3, and possibly other nonzero frequencies.
- A block is designated as containing noncoding (noise) if the spectral envelope is either flat, indicating white noise, or has a peak at, or near, the zero frequency and no other peaks, indicating fractional noise.
- A block is designated as containing other interesting features (e.g. repeat regions) if spectral envelope exhibits several nonzero peaks other than 1/3.
- If adjacent blocks are classified in the same way, they may be recombined.

We will use these recommended decision rules in our real data example in Chapter 6.

4.5 PRACTICAL CHOICE OF Δ_T, θ, τ

Stoffer et al. (2002) [12] suggested that asymptotic approximations of spectral envelope worked well when the number of observations in each stationary segment was at least 2^8. In the last step of our proposed algorithm, we need to use the estimate of spectral envelope and its critical value for post-processing and classification. Thus, we require each stationary segment should be at least 2^8 and set $\Delta_T = 2^8$. With regard to θ, we use $\theta = 0.251$, which is recommended by Cho et al. (2012) [10] in their binary segmentation procedure. The selection of τ is not straightforward. In theory, a certain range of thresholds may lead to consistent results. We propose a "practical" threshold selection procedure, which performs well in practice. Our final application area will be DNA sequences where at each time point there are 4 possible categories (A,C,G,T). So we consider here the case of the 4 categories, i.e., $p = 4$.

We first perform the spectral envelope analysis on the whole series $\{\boldsymbol{Z}_{t,T}\}$. Denote the smallest non-zero significant frequency by ω_τ. Simulate $X^\tau_{t,T} = A_\tau \cos(2\pi\omega_\tau t) + \epsilon_t$, where ϵ_t is Gaussian white noise process with unit variance. Categorize each simulated $X^\tau_{t,T}$ into 4 categories in the following manner

$$\boldsymbol{Z}^\tau_{t,T} = \begin{cases} (1,0,0,0)' & \text{if} \quad X^\tau_{t,T} < c^\tau_1 \\ (0,1,0,0)' & \text{if} \quad c^\tau_1 \le X^\tau_{t,T} < c^\tau_2 \\ (0,0,1,0)' & \text{if} \quad c^\tau_2 \le X^\tau_{t,T} < c^\tau_3 \\ (0,0,0,1)' & \text{if} \quad c^\tau_3 \le X^\tau_{t,T}, \end{cases}$$

where $(1,0,0,0)', (0,1,0,0)', (0,0,1,0)', (0,0,0,1)'$ represents A, C, G, T respectively.
It was stated in Stoffer et al. (2002) [12] that about 65% of the observations correspond to C and G. Here, we use $c^\tau_1 = qnorm(17.5\%)$, $c^\tau_3 = qnorm(82.5\%)$ and $c^\tau_2 = runif(c^\tau_1, c^\tau_3)$.

Next, we calculate the trace of $\mathbf{Z}^\tau_T \boldsymbol{\psi}_{i,t,T} \boldsymbol{\psi}'_{i,t,T} \mathbf{Z}^{\tau'}_T$, $\lambda^\tau_{i,t,T}$, for each scale i. After obtaining $\{\lambda^\tau_{i,t,T}\}$, we find $v \in (1,T)$ that maximizes

$$\mathcal{C}^v_{1,T}(\lambda^\tau_{i,t,T}) = |\sqrt{\frac{T-v}{T\cdot v}} \sum_{t=1}^{v} \lambda^\tau_{i,t,T} - \sqrt{\frac{v}{T\cdot(T-v)}} \sum_{t=v+1}^{T} \lambda^\tau_{i,t,T}|.$$

Let $b^\tau_i = \text{argmax}_v \mathcal{C}^v_{1,T}(\lambda^\tau_{i,t,T})$. We then compute $d^\tau_{i,T} = \frac{\mathcal{C}^{b^\tau_i}_{1,T}(\lambda^\tau_{i,t,T})}{T^{-1}\sum_{t=1}^{T} \lambda^\tau_{i,t,T} \cdot T^\theta \sqrt{\log T}}$. By repeating the above process 100 times, we obtain τ_i, which is the 95% quantile of $d^\tau_{i,T}$ for given i and T. Numerical experiments indicate that in comparison to length T, ω_τ and A_τ have relatively less impact on $d^\tau_{i,T}$.

5.0 SIMULATION

5.1 SIMULATION SCHEMES

We generate $\boldsymbol{Z}_t$ by discretizing the numerical series X_t. We will apply our method to the analysis of DNA sequences, so we discretize X_t into four categories,

$$\boldsymbol{Z}_t = \begin{cases} (1,0,0,0)' & \text{if} \quad X_{t,T} < c_1 \\ (0,1,0,0)' & \text{if} \quad c_1 \leq X_t < c_2 \\ (0,0,1,0)' & \text{if} \quad c_2 \leq X_t < c_3 \\ (0,0,0,1)' & \text{if} \quad c_3 \leq X_t, \end{cases}$$

where $(1,0,0,0)', (0,1,0,0)', (0,0,1,0)', (0,0,0,1)'$ represents A, C, G, T respectively.
It was suggested by Stoffer et al. (2002) [12] that about 65% of the observations correspond to C and G. Here we use $c_1 = qnorm(17.5\%)$, $c_3 = qnorm(82.5\%)$ and $c_2 = runif(c_1, c_3)$. All the simulation models are categorized in the manner shown above.

We first test our method on series with a single change-point. A change occurs at a dyadic location in (A1) and (A2), while occurring at a non-dyadic location in (B). In (A2), a change occurs at an "unbalanced" location as the second segment is much longer than the first one.

(A) *Dyadic example with single change-point*

(A1) *Single change-point, dyadic case 1*

$$X_t = \begin{cases} 1.5\cos(\frac{2\pi t}{3}) + \epsilon_1(t) & 1 \leq t \leq 1024 \\ 1.5\cos(\frac{2\pi t}{10}) + \epsilon_2(t) & 1025 \leq t \leq 2048, \end{cases}$$

where $\epsilon_i(t)$'s are Gaussian white noise processes with unit variance as in all subsequent examples unless specified otherwise;

(A2) *Single change-point, dyadic case 2*

$$X_t = \begin{cases} 1.5\cos(\frac{2\pi t}{3}) + \epsilon_1(t) & 1 \le t \le 512 \\ 1.5\cos(\frac{2\pi t}{10}) + \epsilon_2(t) & 513 \le t \le 2048; \end{cases}$$

(B) *Non-dyadic example with single change-point*

$$X_t = \begin{cases} 1.5\cos(\frac{2\pi t}{3}) + \epsilon_1(t) & 1 \le t \le 729 \\ 1.5\cos(\frac{2\pi t}{10}) + \epsilon_2(t) & 730 \le t \le 2048. \end{cases}$$

The following two simulation models (C) and (D) are adopted from Stoffer et al. (2002) [12] to mimic the signature patterns in DNA sequences. A great many experiences with analysis of DNA sequences in the literature suggest that a CDS typically contains the frequency $\omega = 1/3$. Other frequencies, such as $\omega = 1/10$, may also present. We generate $S_1(t)$, $S_2(t)$ (representing CDS) by discretizing two sinusoidal processes:

$$X_1(t) = 2[\cos(\frac{2\pi t}{3}) + \cos(\frac{2\pi t}{10})] + \epsilon_1(t);$$

$$X_2(t) = 2\cos(\frac{2\pi t}{3}) + \epsilon_2(t).$$

The first signal contains the signature 1/3 frequency and the additional presence of a 1/10 frequency. In the second signal, only the signature 1/3 frequency is present. And we use $N_i(t)$ to denote the point process obtained from i.i.d Gaussian white noise process via the same categorized manner described at the beginning of this section (representing noncoding area). In the first simulation study, the Model (C) is divided in a dyadic manner; in Model (D), changes occur at non-dyadic locations, as follows:

(C) *Dyadic Example with multiple change-points*

$$\boldsymbol{Z}_t = \begin{cases} N_1(t) & 1 \leq t \leq 512 \\ S_1(t) & 513 \leq t \leq 1024 \\ N_2(t) & 1025 \leq t \leq 2048 \\ S_2(t) & 2049 \leq t \leq 3072 \\ N_3(t) & 3073 \leq t \leq 4096; \end{cases}$$

(D) *Non-Dyadic Example with multiple change-points*

$$\boldsymbol{Z}_t = \begin{cases} N_1(t) & 1 \leq t \leq 564 \\ S_1(t) & 565 \leq t \leq 1023 \\ N_2(t) & 1024 \leq t \leq 2199 \\ S_2(t) & 2200 \leq t \leq 3024 \\ N_3(t) & 3025 \leq t \leq 4096, \end{cases}$$

where $\boldsymbol{Z}_t$ represent the point process associated with a simulated DNA sequence of length $T = 4096$.

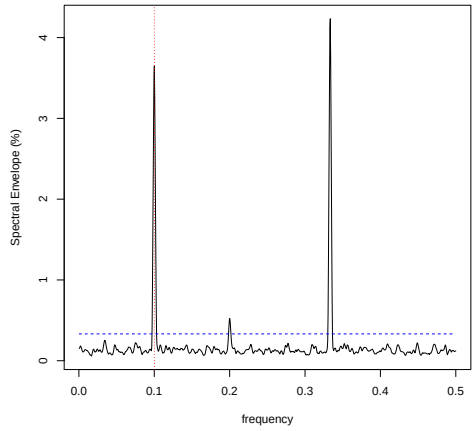
5.2 SIMULATION RESULTS

To obtain the threshold τ for the binary segmentation step (**Step 2**), we first perform the spectral envelope analysis on the entire series (Figure 5).

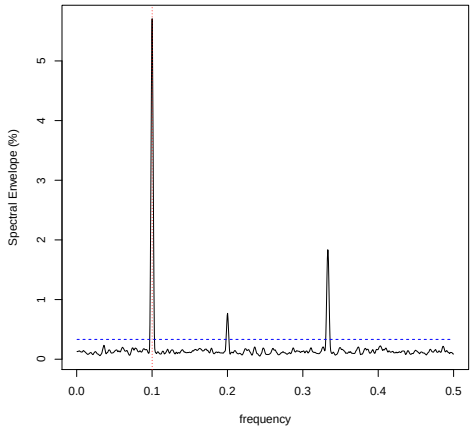
The smallest non-zero significant frequency is $1/10$, so we choose $\omega_\tau = 1/10$. Based on our extensive simulations, a range of A_τ values could generate appropriate τ. All simulations and the application in this dissertation use $A_\tau = 2$. The results of detected locations without the post-processing step (**Step 4**) are summarized in Table 2.

From Table 2, we can conclude that our method without the post-processing step works well for single change-point cases. For multiple change-point cases, our method is sensitive in the sense that every true change-point has been identified, although the number of change-points may be overestimated. To illustrate the feasibility of our proposed post-processing

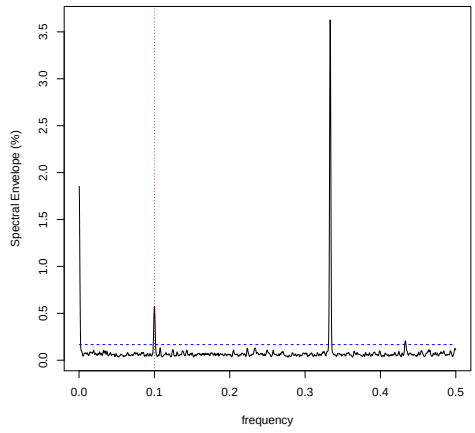
Figure 5: Estimated spectral envelope of the entire series for each simulation model



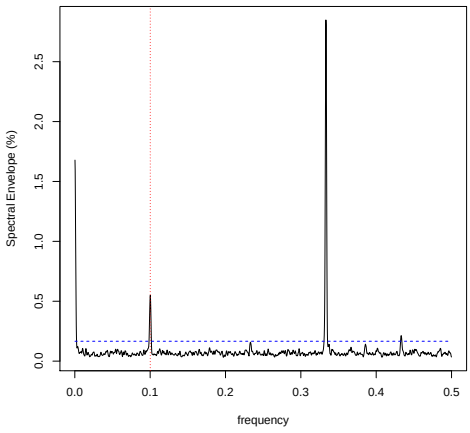
(a) model (A1)



(b) model (B)



(c) model (C)



(d) model (D)

step, we estimate the spectral envelope of each segmentation based on our detection result in Table 2 for model (C) (Figure 6) and model (D) (Figure 7).

In Figure 6, no signal is present in segments $[1, 514], [1017, 1625], [1626, 2028]$ and [3068,

Table 2: Estimated change-points before post-processing for each simulation model

τ generated by $\omega_\tau = \frac{1}{10}, A_\tau = 2$		
Simulation Model	True	Detected
(A1)	1024	1025
(A2)	513	518
(B)	729	747
(C)	513, 1025, 2049, 3073	515, 1017, 1626, 2029, 2497, 3068
(D)	565, 1024, 2200, 3025	559, 1021, 1882, 2156, 2677, 3025

4096]. In contrast, there are clear signals at $\omega = 1/10$ and $\omega = 1/3$ in the block $[515, 1016]$. Moreover, the adjacent blocks, $[2029, 2496]$ and $[2497, 3065]$, show a significant signal at $\omega = 1/3$. Based on visual inspection of Figure 6, segments $[1017, 1625], [1626, 2028]$ and $[2029, 2492], [2497, 3067]$, respectively, can be recombined, which means that $1626, 2497$ can be removed from the final set of detected change-points.

Results after the post-processing step are shown in Table 3. Figure 8 and Figure 9 display the estimated spectral envelope for each segment based on the results in Table 3 for model (C) and model (D). Both Figure 8 and Figure 9 indicate that we can not further combine any neighboring blocks.

To further examine the sensitivity of our method, we simulate each model 100 times and obtain the corresponding $\boldsymbol{Z}_t$. If the number of detected change-points $\hat{B} \geq B$, and for every true change-point ν_b, there exists a $\hat{\nu}_b$ in the set of the detected change-points $\hat{\mathcal{B}}$, $s.t.$ $\mid \nu_b - \hat{\nu}_b \mid < 5\%T$, we will count it as a successful detection. The results are summarized in Table 4.

Figure 6: Estimated spectral envelope of each segment for simulation Model (C) before post-processing

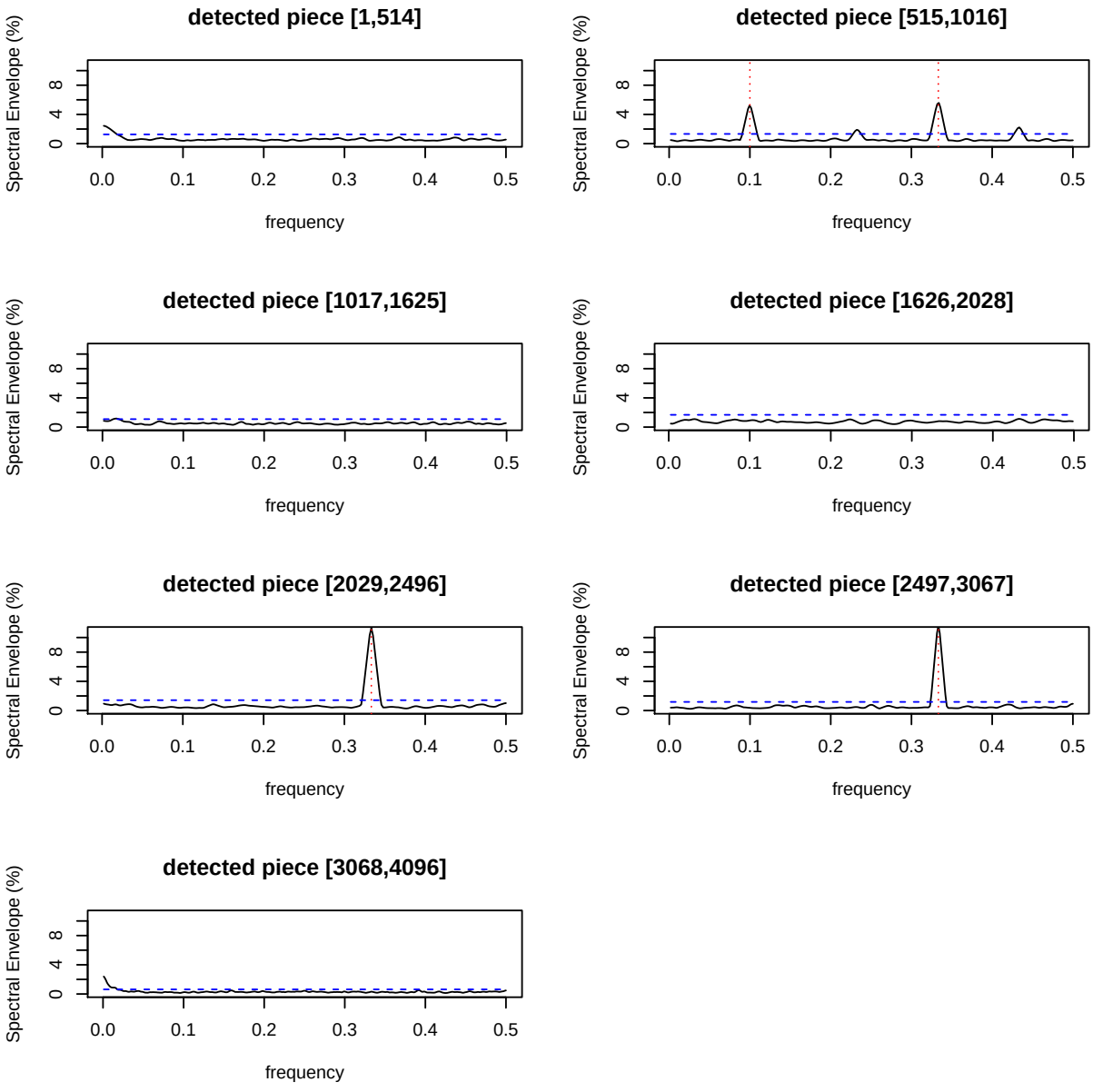


Figure 7: Estimated spectral envelope of each segment for simulation Model (D) before post-processing

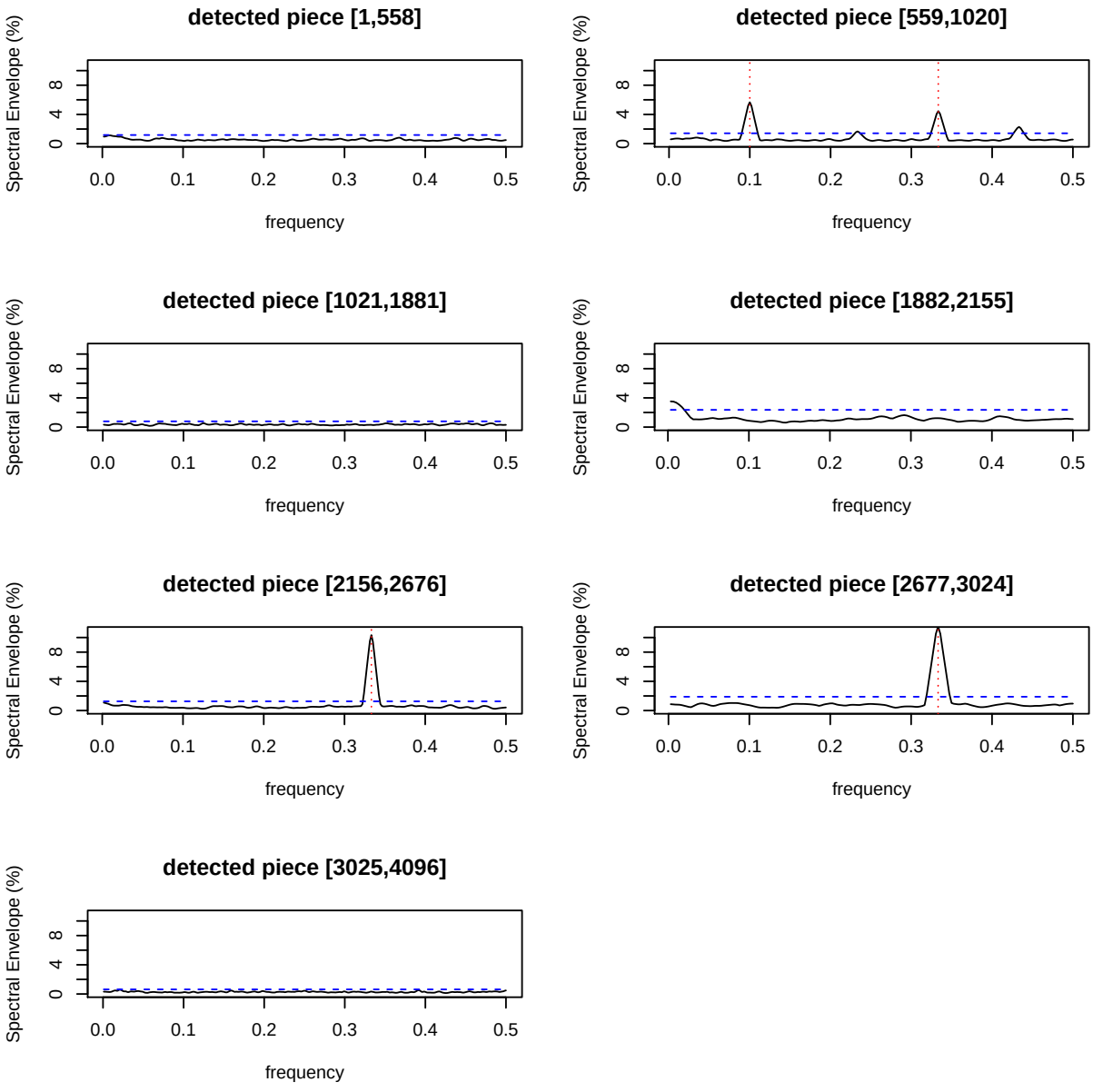


Table 3: Estimated change-points after post-processing for each simulation model

τ generated by $\omega_\tau = \frac{1}{10}, A_\tau = 2$		
Simulation Model	True	Detected
(A1)	1024	1025
(A2)	513	518
(B)	729	747
(C)	$513, 1025, 2049, 3073$	$515, 1017, 2029, 3068$
(D)	$565, 1024, 2200, 3025$	$559, 1021, 2156, 3025$

Table 4: Results over 100 simulations

τ generated by $\omega_\tau = \frac{1}{10}, A_\tau = 2$	
Model (A1)	97
Model (A2)	100
Model (B)	100
Model (C)	95
Model (D)	96

Figure 8: Estimated spectral envelope of each segment for simulation model (C) after post-processing

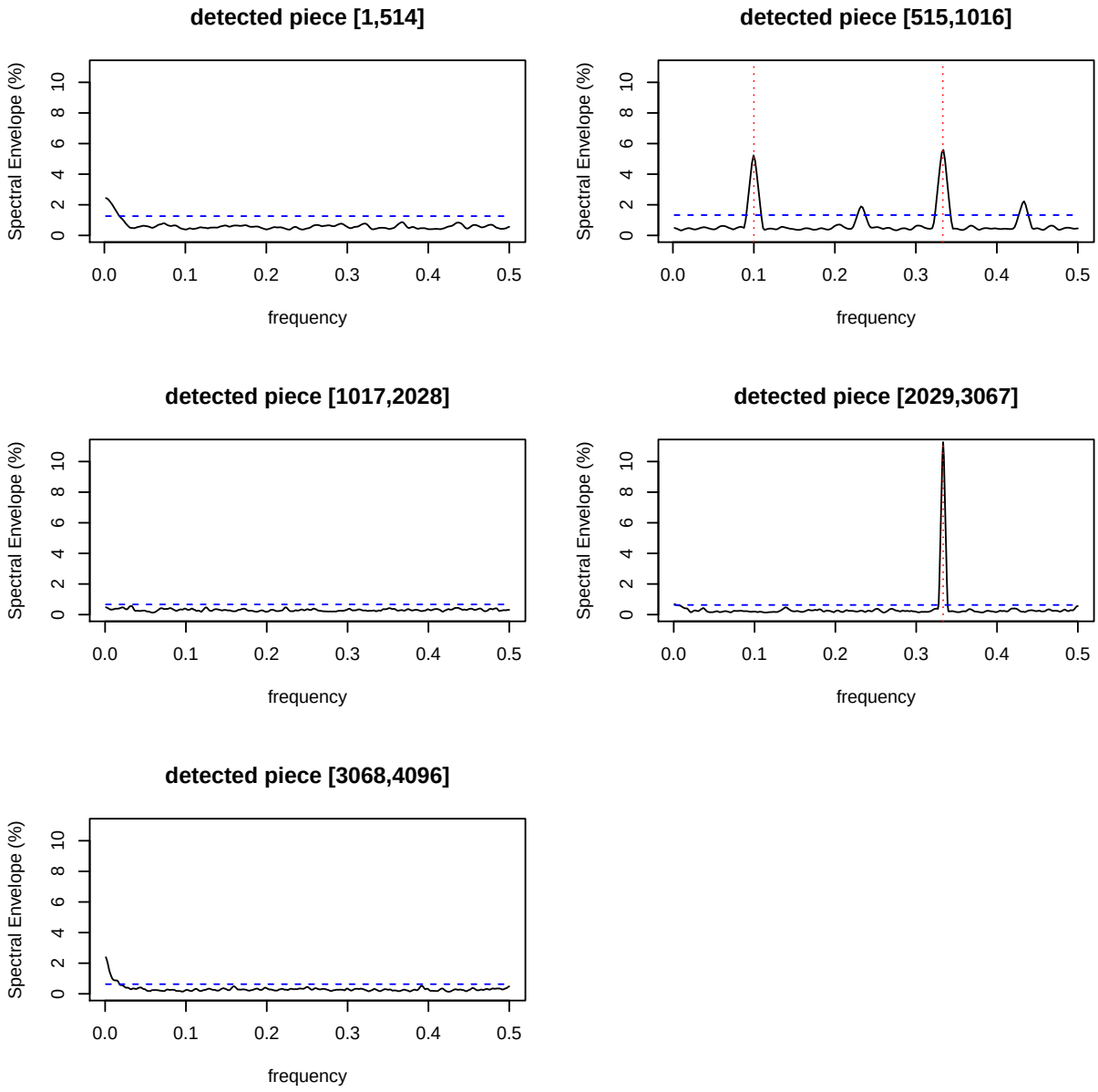
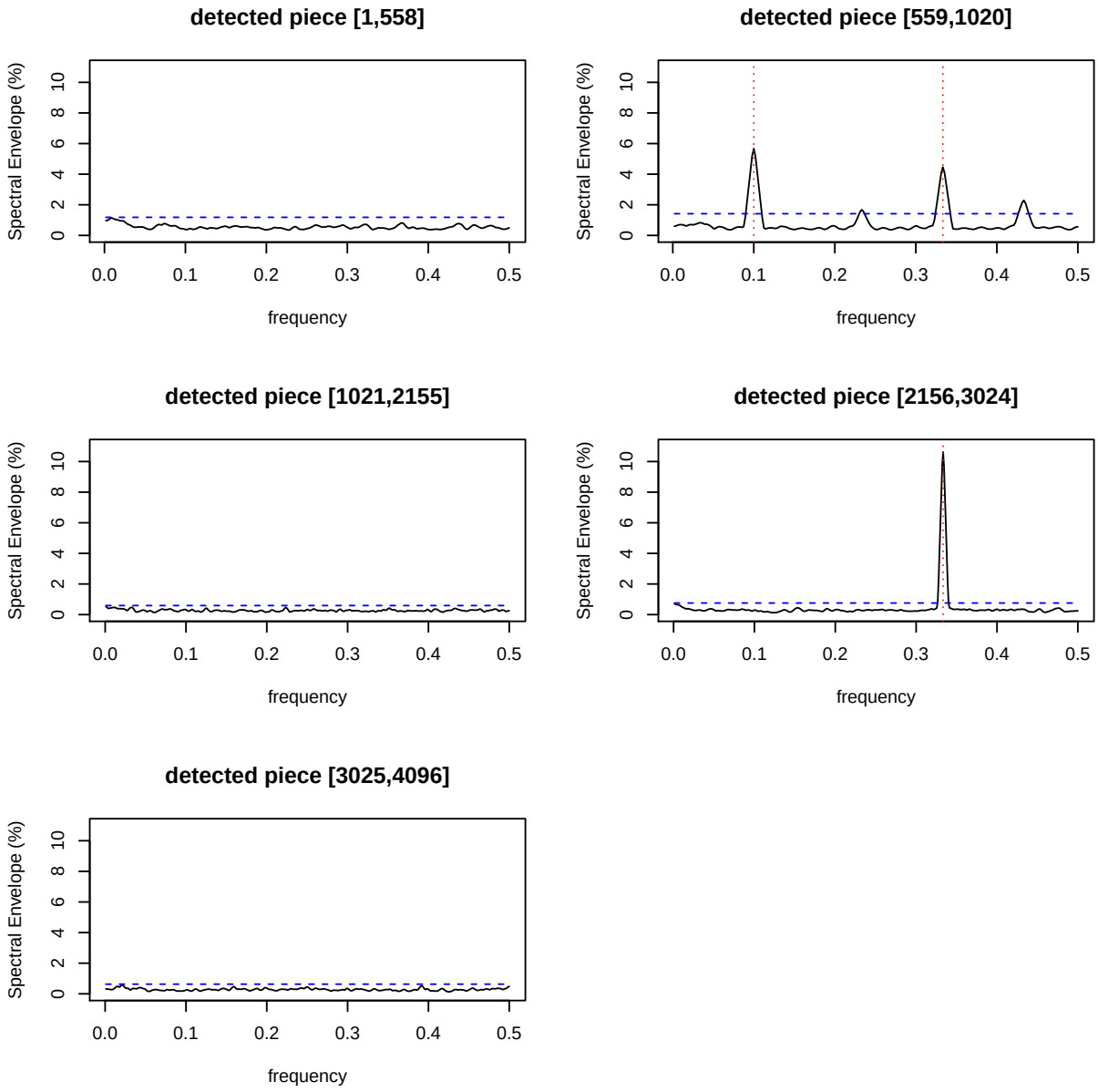


Figure 9: Estimated spectral envelope of each segment for simulation model (D) after post-processing



6.0 APPLICATION

6.1 APPLICATION: ANALYSIS OF THE EBV DNA SEQUENCE

The real data set used here is a subsequence of the EBV DNA sequence. The complete sequence information can be accessed through the National Center for Biotechnology Information (NCBI) official website: http://www.ncbi.nlm.nih.gov/nuccore/V01555.2. This subsequence consists of bp 46333 to bp 54524. Total length is: $T = 8192 = 2^{13}$. Interesting features are the following:

CDS :	$46333 \ldots 47484$
CDS :	$48385 \ldots 49967$
repeat_region :	$50578 \ldots 52115.$

The subsequence we analyzed here contains two coding sequences (CDS), one from bp 46333 to bp 47484, and the other from bp 48385 to bp 49967. In addition, there is a large repeat region from bp 50578 to 52115. By shifting the start bp 46333 to 1, the locations of the breakpoints of interest are: 1153, 2054, 3636, 4346, 5784.

We first perform the spectral envelope analysis on the entire subsequence considered here. The smallest significant non-zero frequency is $\omega = \frac{1}{10}$. We then use $\tau_\omega = \frac{1}{10}, \tau_A = 2$ to generate τ as discussed in Section 4.5.

The estimated locations of change-points before post-processing are listed in Table 5.

The estimated spectral envelope for each segment identified by our algorithm is displayed in Figure 11. Our algorithm locates three interesting segments. In particular, the first and the third estimated segments contain significant signals at $\omega = \frac{1}{10}$ and $\omega = \frac{1}{3}$ respectively.

Figure 10: Estimated spectral envelope of the entire EBV DNA subsequence , $T = 8192$

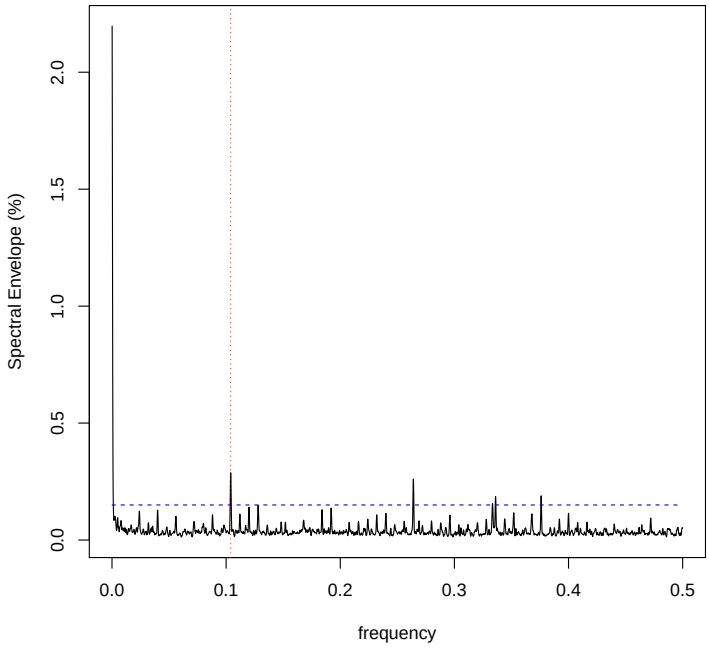
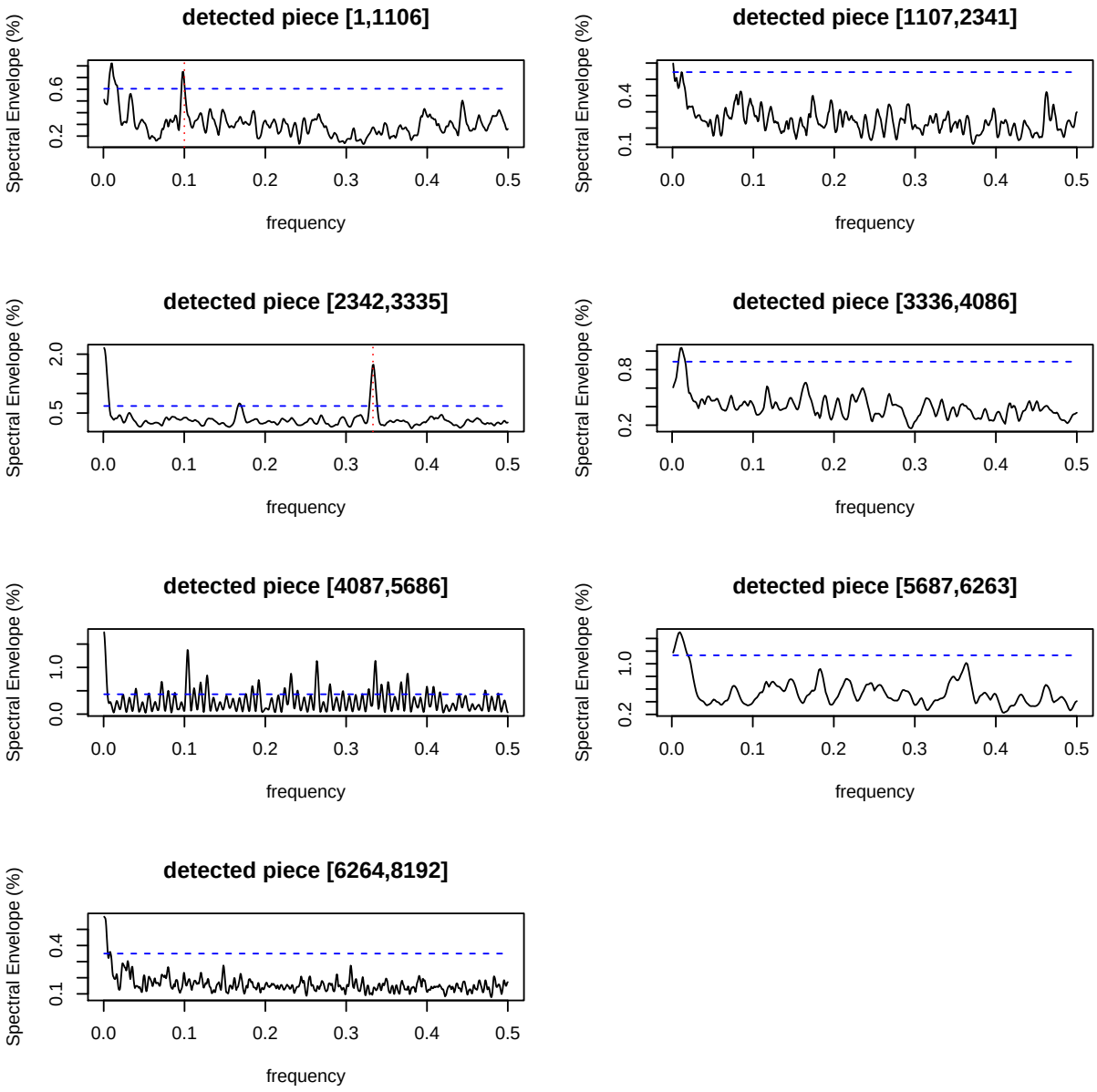


Table 5: Estimated change-points of the EBV DNA subsequence

τ generated by $\omega_\tau = \frac{1}{10}, A_\tau = 2$	
True	$1153, 2054, 3636, 4346, 5784$
Detected	$1107, 2342, 3336, 4087, 5687, 6264$

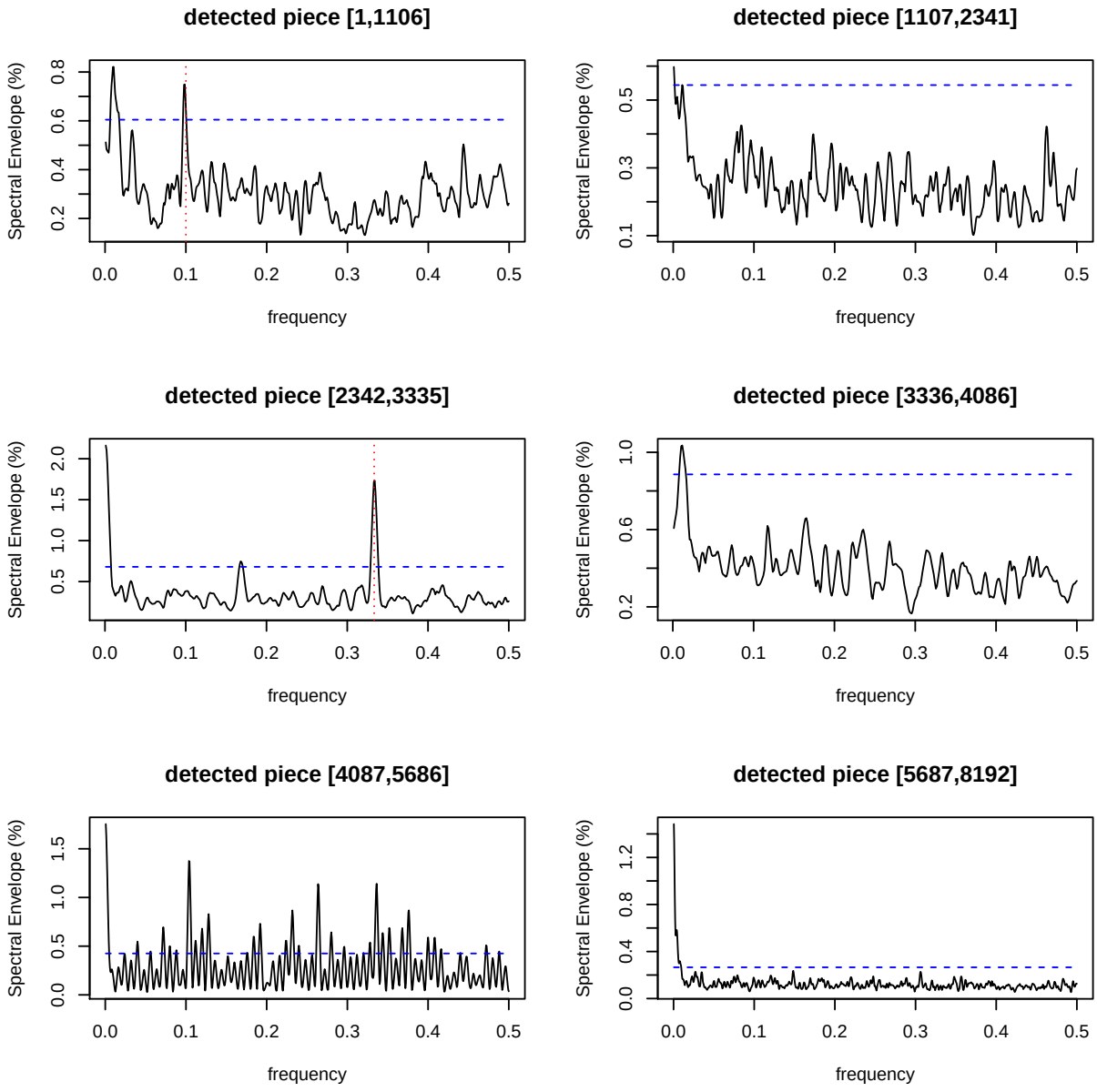
Figure 11: Estimated spectral envelope of each segment for the EBV DNA subsequence before post-processing



According to the classification rule in Section 4.4, those two segments are classified as coding sequences. In addition, the fifth segment correctly identifies the large repeat region.

Further, the last two adjacent segments are classified as noncoding sequences. Thus, they can be recombined, i.e., 6264 could be removed from the set of detected change-points. After removing 6264, the estimated spectral envelope for each segment is shown in Figure 12. No further segments can be recombined.

Figure 12: Estimated spectral envelope of each segment for the EBV DNA subsequence after post-processing



7.0 CONCLUSION AND FUTURE WORK

The challenge in analyzing long DNA sequence data is to identify coding sequences that are dispersed throughout the DNA and separated by regions of noncoding. Motivated by such kinds of local behavior problems in categorical time series, we proposed a method to deal with multiple change-point detection for piecewise categorical time series.

In particular, without loss of generality, we first represented the categorical time series as a multinomial process by denoting each category with a unit vector. We then modeled the corresponding multinomial process by the nonparametric multivariate LSW model, where the piecewise constant autocovariance structure of the process is completely described by local wavelet periodograms at multiple scales. We proposed an optimality criterion to find scalings that can help emphasize local features and summarize the information in the autocovariance structure of the multinomial process. Simultaneously as those scalings were selected, statistics inheriting the piecewise constancy in the autocovariance structure of the multinomial process were also generated, and served as input sequences for further segmentation. Multiple change-point detection was accomplished via a binary segmentation method that was applied to our input sequence separately at each scale, followed by a within-scale and across-scales post-processing procedure to obtain consistent estimators of breakpoints in the autocovariance structure. Furthermore, the consistency result of our method has been derived under certain conditions. We note that our input sequence no longer follows a multiplicative model as in Inclán et al. (1994) [31] and Cho et al. (2012) [10]. Thus, our consistency result is different from previous consistency results of binary segmentation methods in that it allows for correlated and non-normal data, such as categorical time series.

In this dissertation, we also provided several simulation studies and a real data analysis to demonstrate the viability of our methodology. In other applications, it is possible to apply

our proposed method to other real data.

Possible extensions of our work can be done by using complex-valued wavelets instead of the real-valued non-decimated Haar wavelets, as building blocks in modeling the corresponding multinomial process of categorical time series. We note that by using real-valued wavelets, the eigenvalues of the wavelet counterpart of the Fourier spectrum matrix only contain local information in the autocovariance structure of the multinomial process. However, if complex wavelets could be introduced to the LSW process, those eigenvalues could capture local information about both autocovariance and cross-covariance structure. One possible class of such wavelets could be complex-valued compactly supported wavelets generated from the complex valued wavelet transform proposed Lawton (1993) [32]. As Nason et al. (2000) [5] stated that the LSW process could theoretically admit the possibility of using this class of wavelets as building blocks. Another class of complex wavelets that might work is that generated from the dual tree complex wavelet transform ($\mathbb{C}$WT), which was first introduced by Selesnick et al. (2005) [33]. The $\mathbb{C}$WT shares the shift invariant property as the non-decimated wavelet transform, which is essential for analysis of time series. Moreover, $\mathbb{C}$WT enjoys additional properties: It is directionally selective and substantially less redundant than non-decimated wavelet transform, which are important when dealing with multivariate time series. Although complex wavelets possess additional properties when they serve as building blocks for multivariate non-stationary time series, one should note that the consistency theory might be quite involved, because the theory will be related but not limited to figuring out the asymptotic distribution of the largest root of a random matrix.

APPENDIX

PROOF

A.1 PROOF OF THE CONSISTENT RESULT

For each scale i, following from Assumption 4, there exist positive constant M_v and M_ρ^I, such that

$$\sup_{t,T} \operatorname{var}(I_{i,t,T}) \leq M_v, \tag{A.1}$$

$$\rho_\infty^{I_i} = \sum_h \rho^{I_i}(h) = \sum_h \sup_{t,T} |\operatorname{cor}(I_{i,t,T}, I_{i,t+h,T})| \leq M_\rho^I. \tag{A.2}$$

Let s and e represent the starting and ending points, respectively, of a segment. Define

$$\tilde{\mathbb{Y}}_{i,s,e}^b = \frac{\sqrt{e-b}}{\sqrt{n}\sqrt{b-s+1}} \sum_{t=s}^{b} I_{i,t,T} - \frac{\sqrt{b-s+1}}{\sqrt{n}\sqrt{e-b}} \sum_{t=b+1}^{e} I_{i,t,T}, \tag{A.3}$$

and

$$\tilde{\mathbb{S}}_{i,s,e}^b = \frac{\sqrt{e-b}}{\sqrt{n}\sqrt{b-s+1}} \sum_{t=s}^{b} \mu_{i,t,T} - \frac{\sqrt{b-s+1}}{\sqrt{n}\sqrt{e-b}} \sum_{t=b+1}^{e} \mu_{i,t,T}, \tag{A.4}$$

where $\mu_{i,t,T} = \mathbb{E}(I_{i,t,T})$ and $n = e - s + 1$. Note that, because of (4.1), the bias between $\mathbb{E}(I_{i,t,T})$ and $\iota_i(\frac{t}{T})$ does not affect the results of the following lemmas.

The lemmas of this Appendix will hold for any scale i and thus for notational convenience we drop the i index in the $\tilde{\mathbb{Y}}$, $\tilde{\mathbb{S}}$ and η (from Assumption 2). Obviously, $\eta_{q_0} \leq s < \eta_{q_0+1} < \cdots < \eta_{q_0+v} < e \leq \eta_{q_0+v+1}$ for $0 \leq q_0 \leq B - v$ at all stages of the algorithm. In Lemmas

1-2 below, we require that s and e should either be the true breakpoints or close to the true breakpoints in the sense that

$$(\eta_{q_0+1} - s) \wedge (s - \eta_{q_0}) \leq C\epsilon_T \text{ and } (\eta_{q_0+v+1} - e) \wedge (e - \eta_{q_0+v}) \leq C\epsilon_T, \tag{A.5}$$

where $\wedge$ is the minimum operator. In addition, we impose the following condition

$$\eta_{q_0+r} - s > \delta_T \text{ and } e - \eta_{q_0+r} > \delta_T \text{ for some } 1 \leq r \leq v \tag{A.6}$$

to ensure that segment $[s, e]$ contains at least one breakpoint and that the length of each subsequent segment is at least δ_T.

Note that both (A.5) and (A.6) are satisfied as long as the segment $[s, e]$ contains previously undetected breakpoints before the binary segmentation procedure stops.

Lemma 1. *Assume (A.5) and (A.6). If b maximizes $|\tilde{\mathbb{Y}}^t_{s,e}|$, then there exists a true change-point η_{q_0+r} $(1 \leq r \leq v)$, such that b is close to this change-point in the sense that $|b - \eta_{q_0+r}| \leq C\epsilon_T$ for a large T.*

Proof. We will first show

$$\mathbf{P}(|\tilde{\mathbb{Y}}^b_{s,e} - \tilde{\mathbb{S}}^b_{s,e}| > \log T) = \mathbf{P}(\frac{1}{\sqrt{n}} \mid \sum_{t=s}^{e} (I_{i,t,T} - \mu_{i,t,T}) \cdot c_t \mid > \log T) \to 0 \text{ uniformly in } \mathcal{D},$$

where $c_t = \sqrt{e-b}/\sqrt{b-s+1}$ for $t \in [s, b]$ and $c_t = -\sqrt{b-s+1}/\sqrt{e-b}$ otherwise; and $\mathcal{D} := \{1 \leq s < b < e \leq T;\, e - s + 1 \geq C\delta_T,\, \max\{\sqrt{\frac{b-s+1}{e-b}}, \sqrt{\frac{e-b}{s-b+1}}\} \leq c\}$.

For $\forall (s, b, e) \in \mathcal{D}$,

$$\begin{aligned}
&\mathbf{P}(\frac{1}{\sqrt{n}} \mid \sum_{t=s}^{e} (I_{i,t,T} - \mu_{i,t,T}) \cdot c_t \mid > \log T) \\
&\leq \frac{\mathbb{E}(\sum_{t=s}^{e} (I_{i,t,T} - \mu_{i,t,T}) \cdot c_t)^2}{n \log^2 T} \\
&= \frac{\sum_{t=s}^{e} c_t^2 \cdot \mathrm{var}(I_{i,t,T}) + \sum\limits_{\substack{l \neq k \\ s \leq l,k \leq e}} \mathrm{cov}(I_{i,l,T}, I_{i,k,T}) \cdot c_l \cdot c_k}{n \log^2 T} \\
&\leq \frac{c^2 \cdot n \cdot M_v + c^2 \cdot M_v \sum\limits_{\substack{l \neq k \\ s \leq l,k \leq e}} \mathrm{cor} \mid (I_{i,l,T}, I_{i,k,T}) \mid}{n \log^2 T}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{\substack{l\neq k\\ s\leq l,k\leq e}} \mathrm{cor}(I_{i,l,T}, I_{i,k,T}) \\
&\leq 2(n-1)\cdot\rho^{I_i}(1) + 2(n-2)\cdot\rho^{I_i}(2) + \cdots + 2\rho^{I_i}(n-1) \\
&\leq 2n\cdot(\rho^{I_i}(1) + \rho^{I_i}(2) + \cdots + \rho^{I_i}(n-1)) \\
&\leq 2n\cdot\sum_{h}\rho^{I_i}(h) \\
&\leq 2n\cdot\rho_\infty^{I_i} \\
&\leq 2n\cdot M_\rho^I.
\end{aligned}$$

We have

$$\mathbf{P}(\max_{(s,b,e)\in\mathcal{D}}\frac{1}{\sqrt{n}}\mid\sum_{t=s}^{e}(I_{i,t,T}-\mu_{i,t,T})\cdot c_t\mid>\log T)\leq O(\frac{1}{\log T}).$$

That is,

$$\lim_{T\to\infty}\mathbf{P}(|\tilde{\mathbb{Y}}^b_{s,e}-\tilde{\mathbb{S}}^b_{s,e}|\leq\log T)=1 \text{ uniformly over } \mathcal{D}. \tag{A.7}$$

Since $\mid|\tilde{\mathbb{Y}}^b_{s,e}|-|\tilde{\mathbb{S}}^b_{s,e}|\mid\leq\mid\tilde{\mathbb{Y}}^b_{s,e}-\tilde{\mathbb{S}}^b_{s,e}\mid$, then

$$\lim_{T\to\infty}\mathbf{P}(\mid|\tilde{\mathbb{Y}}^b_{s,e}|-|\tilde{\mathbb{S}}^b_{s,e}|\mid\leq\log T)=1 \text{ uniformly over } \mathcal{D}. \tag{A.8}$$

According to Lemma 4 in Cho et al. (2012) [10], if (A.5), (A.6), (A.7), and (A.8) are satisfied, for $b=\mathrm{argmax}_{s<t<e}|\tilde{\mathbb{Y}}^t_{s,e}|$, there exists $1\leq r\leq v$ such that $|b-\eta_{q_0+r}|\leq C\epsilon_T$ for a large T.

□

Note that the step after locating the point b that maximizes $|\tilde{\mathbb{Y}}^t_{s,e}|$ in the binary segmentation procedure is to perform the thresholding on $\frac{|\tilde{\mathbb{Y}}^b_{s,e}|}{T^\theta\sqrt{\log T}\cdot n^{-1}\sum_{t=s}^{e}I_{i,t,T}}$ and screen out b that $\frac{|\tilde{\mathbb{Y}}^b_{s,e}|}{T^\theta\sqrt{\log T}\cdot n^{-1}\sum_{t=s}^{e}I_{i,t,T}}<\tau$. Lemma 2 is the logic behind this step.

Lemma 2. *Under* (A.5) *and* (A.6), $\mathbf{P}(|\tilde{\mathbb{Y}}^b_{s,e}|<\tau T^\theta\sqrt{\log T}\cdot n^{-1}\sum_{t=s}^{e}I_{i,t,T})\to 0$ *for* $b=\mathrm{argmax}_{s<t<e}|\tilde{\mathbb{Y}}^t_{s,e}|$.

Proof. Denote $\tilde{d} = \tilde{\mathbb{Y}}^b_{s,e} = \tilde{d}_1 - \tilde{d}_2$ and $\tilde{m} = n^{-1/2}\sum_{t=s}^{e} I_{i,t,T} = c_1\tilde{d}_1 + c_2\tilde{d}_2$, where $\tilde{d}_1 = \frac{\sqrt{e-b}}{\sqrt{n}\sqrt{b-s+1}}\sum_{t=s}^{b} I_{i,t,T}$, $\tilde{d}_2 = \frac{\sqrt{b-s+1}}{\sqrt{n}\sqrt{e-b}}\sum_{t=b+1}^{e} I_{i,t,T}$, and $c_1 = c_2^{-1} = \sqrt{\frac{b-s+1}{e-b}}$. For simplicity, let $c_2 > c_1$. Further, let $\mu_i = \mathbb{E}\tilde{d}_i$ and $\omega_i = \mathrm{var}(\tilde{d}_i)$ for $i = 1, 2$, and define $\mu = \mathbb{E}\tilde{d}$ and $\omega = \mathrm{var}(\tilde{d})$. Finally, t_n denotes the threshold $\tau T^{\theta}\sqrt{\log T/n}$. We need to show $\mathbf{P}(|\tilde{d}| \le \tilde{m}\cdot t_n) \to 0$.

Following Lemma 5 in Cho et al. (2012) [10], $\mathbf{P}(|\tilde{d}| \le \tilde{m}\cdot t_n)$ is bounded by

$$4\mu^{-2}(1+c_1t_n)^{-2}\{(c_1t_n-1)^2\omega_1 + (c_2t_n+1)^2\omega_2 + 4c_1^2t_n^2\mu_1^2 + (c_2-c_1)^2t_n^2\mu_2^2\}.$$

Since $n > \delta_T > O(T^{\theta+\frac{1}{2}})$ and $\theta \in (\frac{1}{4}, \frac{1}{2})$, $t_n = \tau\cdot\frac{T^{\theta}\log^{\frac{1}{2}}T}{n^{\frac{1}{2}}} < O(T^{\frac{\theta}{2}-\frac{1}{4}}\log^{\frac{1}{2}}T) \to 0$ as $T\to\infty$.

Note that

$$\begin{aligned}
\omega_1 &= \mathrm{var}\Big(\frac{\sqrt{e-b}}{\sqrt{n}\sqrt{b-s+1}}\sum_{t=s}^{b} I_{i,t,T}\Big)\\
&= \frac{e-b}{n(b-s+1)}\sum_{t=s}^{b}\mathrm{var}(I_{i,t,T}) + \frac{e-b}{n(b-s+1)}\sum_{s<l,k<b}\mathrm{cov}(I_{i,l,T}, I_{i,k,T})\\
&\le \frac{e-b}{n(b-s+1)}\cdot(b-s+1)\max\mathrm{var}(I_{i,t,T}) + \frac{e-b}{n(b-s+1)}\cdot 2\cdot(b-s+1)\cdot\rho_{\infty}^{I_i}\cdot M_v\\
&\le (2M_{\rho}^{I}+1)M_v.
\end{aligned}$$

Similarly, $\omega_2 \le (2M_{\rho}^{I}+1)M_v$.

Also, μ_i's are bounded by the following:

$$\begin{aligned}
\mu_1 &= \mathbb{E}\Big(\frac{\sqrt{e-b}}{\sqrt{n}\sqrt{b-s+1}}\sum_{t=s}^{b} I_{i,t,T}\Big)\\
&\le \frac{\sqrt{e-b}}{\sqrt{n}\sqrt{b-s+1}}\cdot(b-s+1)M_{\mu}\\
&\le \sqrt{n}M_{\mu}.
\end{aligned}$$

The proof of Lemma 4 in Cho et al. (2012) [10] indicates that if $b = \mathrm{argmax}_{s<t<e}|\tilde{\mathbb{Y}}^t_{s,e}|$, then $\tilde{\mathbb{S}}^b_{s,e} \ge \tilde{\mathbb{S}}_{s,e} - 2\log T$, where $\tilde{\mathbb{S}}_{s,e} = \max_{s<t<e}|\tilde{\mathbb{S}}^t_{s,e}|$.

According to Lemma 1 in Cho et al. (2012) [10], $\tilde{\mathbb{S}}_{s,e} \ge C\delta_T/\sqrt{T} > T^{\theta}\sqrt{\log T}$. Recall that under Assumption 2, $\delta_T = CT^{\Theta}$, where $\Theta \in (\theta+\frac{1}{2}, 1)$. Thus, we have $\mu = \tilde{\mathbb{S}}^b_{s,e} \ge \tilde{\mathbb{S}}_{s,e} - 2\log T > O(\delta_T/\sqrt{T}) > O(T^{\theta}\sqrt{\log T})$.

As a result, we have

$$\mathbf{P}(|\tilde{d}| \leq \tilde{m} \cdot t_n) \leq O(\frac{t_n^2 \cdot n}{\delta_T^2/T}) = O(\frac{T^{2\theta} \log T}{\delta_T^2/T}) \to 0.$$

□

Next, Lemma 3 deals with the situation that, except for the true breakpoint that is close enough to s and e, no other breakpoint is in the segment $[s, e]$. This situation occurs when no further breakpoint can be detected; the binary segmentation procedure will then stop.

Lemma 3. *For some positive constants C, C', let s, e satisfy either*

(i) $\exists! 1 \leq q \leq B$ *such that* $s \leq \eta_q \leq e$ *and* $[\eta_q - s + 1] \wedge [e - \eta_q] \leq C\epsilon_T$, *or*

(ii) $\exists 1 \leq q \leq B$ *such that* $s \leq \eta_q \leq \eta_{q+1} \leq e$ *and* $[\eta_q - s + 1] \vee [e - \eta_{q+1}] \leq C'\epsilon_T$.

Then, for a large T,

$$\mathbf{P}(|\tilde{\mathbb{Y}}_{s,e}^b| > \tau T^\theta \sqrt{\log T} \cdot n^{-1} \sum_{t=s}^{e} I_{i,t,T}) \to 0,$$

where $b = \text{argmax}_{s<t<e} |\tilde{\mathbb{Y}}_{s,e}^t|$.

Proof. First we assume (i). Let $\mathcal{E} = \{|\tilde{\mathbb{Y}}_{s,e}^b| > \tau T^\theta \sqrt{\log T} \cdot n^{-1} \sum_{t=s}^{e} I_{i,t,T}\}$ and

$$\mathcal{F} = \{\frac{1}{n} |\sum_{t=s}^{e} (I_{i,t,T} - \mu_{i,t,T})| < d = \frac{(\eta_q - s + 1)\mu_{i,\eta_q,T} + (e - \eta_q)\mu_{i,\eta_{q+1},T}}{2n}\}.$$

We have $\mathbf{P}(\mathcal{E}) = \mathbf{P}(\mathcal{E} \cap \mathcal{F}) + \mathbf{P}(\mathcal{E}|\mathcal{F}^c)\mathbf{P}(\mathcal{F}^c) \leq \mathbf{P}(\mathcal{E} \cap \mathcal{F}) + \mathbf{P}(\mathcal{F}^c)$.

Note that

$$\mathcal{F} = \{\sum_{t=s}^{e} \mu_{i,t,T} - n \cdot d < \sum_{t=s}^{e} I_{i,t,T} < \sum_{t=s}^{e} \mu_{i,t,T} + n \cdot d\},$$

and $\sum_{t=s}^{e} \mu_{i,t,T} = 2n \cdot d$.

Thus, by Markov's inequalities,

$$\begin{aligned} \mathbf{P}(\mathcal{E} \cap \mathcal{F}) &\leq \mathbf{P}(|\tilde{\mathbb{Y}}_{s,e}^b| > \tau T^\theta \sqrt{\log T} \cdot n^{-1} (\sum_{t=s}^{e} \mu_{i,t,T} - nd)) \\ &= \mathbf{P}(|\tilde{\mathbb{Y}}_{s,e}^b| > \tau T^\theta \sqrt{\log T} \cdot d) \\ &\leq \frac{\mathbb{E}|\tilde{\mathbb{Y}}_{s,e}^b|^2}{\tau^2 T^{2\theta} \log T d^2}. \end{aligned}$$

Based on Lemmas 2.2 and 2.3 of Venkatraman (1993) [21], $|\tilde{\mathbb{S}}^t_{s,e}|$ achieves its maximum value only at a breakpoint. In case (i), only one breakpoint η_q is in segment $[s, e]$; one of the subsegments is less than $C\epsilon_T$ in length, and the other subsegment is less than n in length. Then,

$$\max_{s<t<e} |\tilde{\mathbb{S}}^t_{s,e}| = |\tilde{\mathbb{S}}^{\eta_q}_{s,e}| = | \frac{\sqrt{\eta_q - s + 1}\sqrt{e - \eta_q}}{\sqrt{n}} (\mu_{i,\eta_q,T} - \mu_{i,\eta_{q+1},T}) | \leq O(\sqrt{\epsilon_T}).$$

From Lemma 1, $|\tilde{\mathbb{Y}}^b_{s,e} - \tilde{\mathbb{S}}^b_{s,e}| \leq \log T$. Note that $|\tilde{\mathbb{Y}}^b_{s,e}| - |\tilde{\mathbb{S}}^b_{s,e}| \leq |\tilde{\mathbb{Y}}^b_{s,e} - \tilde{\mathbb{S}}^b_{s,e}| \leq \log T$, so we have

$$\begin{aligned} |\tilde{\mathbb{Y}}^b_{s,e}| &\leq |\tilde{\mathbb{S}}^b_{s,e}| + \log T \\ &\leq \max_{s<t<e} |\tilde{\mathbb{S}}^t_{s,e}| + \log T \\ &\leq |\tilde{\mathbb{S}}^{\eta_q}_{s,e}| + \log T \\ &= O(\sqrt{\epsilon_T}). \end{aligned}$$

Thus,

$$\begin{aligned} \mathbf{P}(\mathcal{E} \cap \mathcal{F}) &\leq \frac{\mathbb{E}|\tilde{\mathbb{Y}}^b_{s,e}|^2}{\tau^2 T^{2\theta} \log T d^2} \\ &\leq O(\frac{\epsilon_T}{T^{2\theta} \log T}) \\ &= O(\frac{T^{\frac{1}{2}} \log T}{T^{2\theta} \log T}) \\ &= O(T^{\frac{1}{2}-2\theta}) \\ &\to 0. \end{aligned}$$

Note that by Assumption 2, $\theta \in (\frac{1}{4}, \frac{1}{2})$.

Now, consider $\mathbf{P}(\mathcal{F}^c) = \mathbf{P}(\frac{1}{n}|\sum_{t=s}^{e}(I_{i,t,T} - \mu_{i,t,T})| > d)$. According to Markov's inequality and the similar argument in the proof of Lemma 1, we have

$$\begin{aligned}
\mathbf{P}(\mathcal{F}^c) &\leq \frac{\mathbb{E} \mid \sum_{t=s}^{e}(I_{i,t,T} - \mu_{i,t,T}) \mid^2}{n^2 d^2} \\
&\leq \frac{\sum_{t=s}^{e} \mathrm{var}(I_{i,t,T}) + \sum\limits_{\substack{l \neq k \\ s \leq l,k \leq e}} \mathrm{cov}(I_{i,l,T}, I_{i,k,T})}{n^2 d^2} \\
&\leq \frac{n \cdot M_v + 2n \cdot \rho_{\infty}^{I_i} \cdot M_v}{n^2 d^2} \\
&\leq \frac{M_v + 2M_{\rho}^{I} \cdot M_v}{n d^2} \\
&\to 0
\end{aligned}$$

For case (ii), d in $\mathcal{F}$ is different from case (i) and should be set as $(\eta_q - s + 1)\mu_1 + (\eta_{q+1} - \eta_q)\mu_2 + (e - \eta_{q+1})\mu_3$, where $\mu_1 = \mu_{i,\eta_q,T}, \mu_2 = \mu_{i,\eta_{q+1},T}, \mu_3 = \mu_{i,\eta_{q+2},T}$. Obviously, $\sum_{t=s}^{e} \mu_{i,t,T} = 2n \cdot d$ still holds. Another difference is that $\max_{s<t<e} |\tilde{\mathbb{S}}_{s,e}^{t}| = \max(|\tilde{\mathbb{S}}^{\eta_q}|, |\tilde{\mathbb{S}}^{\eta_{q+1}}|)$. Under condition (ii),

$$\begin{aligned}
|\tilde{\mathbb{S}}^{\eta_q}| &= \mid \frac{\sqrt{e - \eta_q}}{\sqrt{n}\sqrt{\eta_q - s + 1}} \sum_{t=s}^{\eta_q} \mu_{i,t,T} - \frac{\sqrt{\eta_q - s + 1}}{\sqrt{n}\sqrt{e - \eta_q}} \sum_{t=\eta_q+1}^{e} \mu_{i,t,T}) \mid \\
&= \mid \frac{\sqrt{e - \eta_q}\sqrt{\eta_q - s + 1}}{\sqrt{n}} \mu_1 - \frac{\sqrt{\eta_q - s + 1}}{\sqrt{n}\sqrt{e - \eta_q}}((\eta_{q+1} - \eta_q)\mu_2 + (e - \eta_{q+1})\mu_3) \mid \\
&\leq \max(\mid \frac{\sqrt{e - \eta_q}\sqrt{\eta_q - s + 1}}{\sqrt{n}} \mu_1 - \frac{\sqrt{e - \eta_q}\sqrt{\eta_q - s + 1}}{\sqrt{n}} \min(\mu_2, \mu_3) \mid, \\
&\quad \mid \frac{\sqrt{e - \eta_q}\sqrt{\eta_q - s + 1}}{\sqrt{n}} \mu_1 - \frac{\sqrt{e - \eta_q}\sqrt{\eta_q - s + 1}}{\sqrt{n}} \max(\mu_2, \mu_3) \mid) \\
&\leq O(\sqrt{\epsilon_T}).
\end{aligned}$$

Similarly, $|\tilde{\mathbb{S}}^{\eta_{q+1}}| \leq O(\sqrt{\epsilon_T})$. Hence, $\max_{s<t<e} |\tilde{\mathbb{S}}_{s,e}^{t}| \leq O(\sqrt{\epsilon_T})$ still holds. A similar argument as the proof in case (i) leads to $\mathbf{P}(|\tilde{\mathbb{Y}}_{s,e}^{b}| > \tau T^{\theta}\sqrt{\log T} \cdot n^{-1} \sum_{t=s}^{e} I_{i,t,T}) \to 0$ as $T \to \infty$ for case (ii).

□

According to Theorem 2 in Cho et al. (2012) [10], post-processing across-scales preserves the consistency result. Then, the theorem of Section 4.3 follows immediately from Lemmas 1-3.

Bibliography

[1] David S Stoffer, David E Tyler, and Andrew J McDougall. Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, 80(3):611–622, 1993.

[2] David S Stoffer, Mark S Scher, Gale A Richardson, Nancy L Day, and Patricia A Coble. A walshfourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. *Journal of the American Statistical Association*, 83(404):954–963, 1988.

[3] Maurice B Priestley. Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 204–237, 1965.

[4] Rainer Dahlhaus et al. Fitting time series models to nonstationary processes. *The annals of Statistics*, 25(1):1–37, 1997.

[5] Guy P Nason, Rainer Von Sachs, and Gerald Kroisandt. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):271–292, 2000.

[6] Hernando Ombao, Jonathan Raz, Rainer von Sachs, and Wensheng Guo. The slex model of a non-stationary random process. *Annals of the Institute of Statistical Mathematics*, 54(1):171–200, 2002.

[7] Sudeshna Adak. Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association*, 93(444):1488–1501, 1998.

[8] Mladen Victor Wickerhauser. Adapted wavelet analysis from theory to software. 1994.

[9] Hernando Ombao, Rainer Von Sachs, and Wensheng Guo. Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100(470):519–531, 2005.

[10] Haeran Cho and Piotr Fryzlewicz. Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22(1):207–229, 2012.

[11] Haeran Cho and Piotr Fryzlewicz. Multiple change-point detection for high-dimensional time series via sparsified binary segmentation. *Preprint*, 2014.

[12] David S Stoffer, Hernando C Ombao, and David E Tyler. Local spectral envelope: An approach using dyadic tree-based adaptive segmentation. *Annals of the Institute of Statistical Mathematics*, 54(1):201–223, 2002.

[13] Tyra G Wolfsberg, Johanna McEntyre, and Gregory D Schuler. Guide to the draft human genome. *Nature*, 409(6822):824–826, 2001.

[14] Samuel Karlin and Catherine Macken. Some statistical problems in the assessment of inhomogeneities of dna sequence data. *Journal of the American Statistical Association*, 86(413):27–35, 1991.

[15] Jerome V Braun and Hans-Georg Muller. Statistical methods for dna sequence segmentation. *Statistical Science*, pages 142–162, 1998.

[16] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[17] Ingrid Daubechies et al. *Ten lectures on wavelets*, volume 61. SIAM, 1992.

[18] Guy P Nason. *Wavelet methods in statistics with R*. Springer, 2008.

[19] Brani Vidakovic. *Statistical modeling by wavelets*, volume 503. John Wiley & Sons, 2009.

[20] Donald B Percival and Andrew T Walden. *Wavelet methods for time series analysis*, volume 4. Cambridge University Press, 2006.

[21] Ennapadam Seshan Venkatraman. *Consistency results in multiple change-point problems*. PhD thesis, to the Department of Statistics.Stanford University, 1992.

[22] David R Brillinger. *Time series: data analysis and theory*, volume 36. Siam, 2001.

[23] Jean Sanderson, Piotr Fryzlewicz, and MW Jones. Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika*, 97(2):435–446, 2010.

[24] R.W. Dijkerman and R.R. Mazumdar. On the correlation structure of the wavelet coefficients of fractional brownian motion. 40(5):1609–1612, 1994.

[25] P. Flandrin. Wavelet analysis and synthesis of fractional brownian motion. 38(2):910–917, 1992.

[26] A.H. Tewfik and M. Kim. Correlation structure of the discrete wavelet coefficients of fractional brownian motion. 38(2):904–909, 1992.

[27] Gregory W. Wornell. Wavelet-based representations for the 1/f family of fractal processes. 81(10):1428–1450, 1993.

[28] Yanqin Fan. On the approximate decorrelation property of the discrete wavelet transform for fractionally differenced processes. 49(2):516–521, 2003.

[29] James L Cornette, Kemp B Cease, Hanah Margalit, John L Spouge, Jay A Berzofsky, and Charles DeLisi. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *Journal of molecular biology*, 195(3):659–685, 1987.

[30] Shrish Tiwari, S Ramachandran, Alok Bhattacharya, Sudha Bhattacharya, and Ramakrishna Ramaswamy. Prediction of probable genes by fourier analysis of genomic sequences. *Bioinformatics*, 13(3):263–270, 1997.

[31] Carla Inclan and George C Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994.

[32] Wayne Lawton. Applications of complex valued wavelet transforms to subband decomposition. *Signal Processing, IEEE Transactions on*, 41(12):3566–3568, 1993.

[33] Ivan W Selesnick, Richard G Baraniuk, and Nick G Kingsbury. The dual-tree complex wavelet transform. *Signal Processing Magazine, IEEE*, 22(6):123–151, 2005.