THE SPECTRAL ANALYSIS OF NONSTATIONARY CATEGORICAL TIME SERIES USING LOCAL SPECTRAL ENVELOPE

by

Hyewook Jeong

BS, Sookmyung Women's University, 1996

MS, Sookmyung Women's University, 1999

Submitted to the Graduate Faculty of

the Department of Statistics in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF STATISTICS

This dissertation was presented

by

Hyewook Jeong

It was defended on

August 6, 2012

and approved by

David S. Stoffer

Yu Cheng

Robert T. Krafty

Sati Mazumdar

Dissertation Director: David S. Stoffer

THE SPECTRAL ANALYSIS OF NONSTATIONARY CATEGORICAL TIME SERIES USING LOCAL SPECTRAL ENVELOPE

Hyewook Jeong, PhD

University of Pittsburgh, 2012

Most classical methods for the spectral analysis are based on the assumption that the time series is stationary. However, many time series in practical problems shows nonstationary behaviors. The data from some fields are huge and have variance and spectrum which changes over time. Sometimes,we are interested in the cyclic behavior of the categorical-valued time series such as EEG sleep state data or DNA sequence, the general method is to scale the data, that is, assign numerical values to the categories and then use the periodogram to find the cyclic behavior. But there exists numerous possible scaling. If we arbitrarily assign the numerical values to the categories and proceed with a spectral analysis, then the results will depend on the particular assignment. We would like to find the all possible scaling that bring out all of the interesting features in the data. To overcome these problems, there have been many approaches in the spectral analysis.

Our goal is to develop a statistical methodology for analyzing nonstationary categorical time series in the frequency domain. In this dissertation, the spectral envelope methodology is introduced for spectral analysis of categorical time series. This provides the general framework for the spectral analysis of the categorical time series and summarizes information from the spectrum matrix. To apply this method to nonstationary process, I used the TBAS(Tree-Based Adaptive Segmentation) and local spectral envelope based on the piecewise stationary process. In this dissertation,the TBAS(Tree-Based Adpative Segmentation) using distance function based on the Kullback-Leibler divergence was proposed to find the best segmentation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to acknowledge the advice, support, and friendship of a number of people who helped me during the writing of this thesis and my time as a graduate student. First, I would like to thank Professor David S. Stoffer, my advisor, for his suggestions and constant support during this research. His excellent guidance with endless patience always leads me to the right direction. I am honored to be his student.

I also would like to thank my other committee members, Professor Yu Chen, Professor Robert Krafty and Professor Sati Mazumdar for their helpful suggestions and encouragement. And I wish to express my appreciation to the rest of faculty, staff, especially Mary, and students of the statistics department at University of Pittsburgh for their encouragement, support and friendship throughout my graduate years.

I am grateful to my mother and parents-in-law. They have always provided me with unconditional love and ceaseless support, in joy and in sorrow.

Last but not least, I would like to thank my husband, Seohyon, for being always there for me. I would not have been able to complete it without him. In addition, I appreciate my lovely children, John and Claire, for giving me limitless happiness.

1.0 INTRODUCTION

Most classical methods for the spectral analysis are based on the assumption that the time series is stationary. The assumption of stationarity plays an important role in the theory of the estimation and inferences. However, many time series in practical problems(e.g.biomedical signal processing, speech signal processing or DNA sequences) shows nonstationary behaviors. The data from some fields are huge and have variance and spectrum which changes over time. To overcome these problems, there have been many approaches in the spectral analysis for nonstationary time series.

Priestley(1965)[10] first introduced the time-varying spectrum with Cramér representation for the analysis of nonstationary processes. Dahlhaus(1997)[4] generalized this ideas and proposed a method for estimating the parameters of the nonstationary process(locally stationary process) with an asymptotic framework. This asymptotic framework ensures a reasonable asymptotic theory in the nonstationary framework. Dahlhaus(2000) [5] also developed this method to multivariate nonstationary time series.

Adak(1998)[1] introduced a new class of nonstationary time series (piecewise locally stationary) and adopted an adaptive segmentation method to the estimation of the time-varying spectra. The segmentation algorithm is based on the optimal pruning as used in Classification and Regression Tree(CART) of Breiman et al(1984)[2] and Best Basis Algorithm(BBA) of Coifman and Wickerhause(1992)[3]. However, the estimation of the spectrum is based on windowed Fourier transform, which is not orthogonal. As a consequence, the method has some problems in establishing consistency. Moreover, the method is for univariate process and doesn't allow an extension to multivariate process.

Ombao et al(2001)[9] proposed a new method for analyzing bivariate nonstationary time series in the frequency domain. In this article, they introduced SLEX(Smoothed Localized

in time EXponential) model and used BBA(Best Basis Algorithm) as a segmentation technique. The periodogram based on this approach is localized in time and frequency, hence, is proper for nonstationary time series and also computationally efficient because they used the FFT(Fast Fourier Transform) and BBA(Best Basis Algorithm). Ombao et al(2004)[6] also applied this method to the discriminant and classification analysis of nonstationary time series.

When we are interested in the cyclic behavior of the categorical-valued time series such as EEG sleep state data or DNA sequence, the general method is to scale the data, that is, assign numerical values to the categories and then use the periodogram to find the cyclic behavior. But there exists numerous possible scaling. If we arbitrarily assign the numerical values to the categories and proceed with a spectral analysis, then the results will depend on the particular assignment. Hence, we'd like to find the all possible scaling that bring out all of the interesting features in the data.

Stoffer et al(1993a)[12] first introduced the concepts of spectral envelope in the spectral analysis for categorical time series. In this article, they established the theory of the estimation of the spectral envelope for categorical time series and discussed its applications. Stoffer et al(2002)[11] extended this idea to piecewise stationary process and developed the local spectral envelope of nonstationary time series.

Our goal is to develop a statistical methodology for analyzing nonstationary categorical time series in the frequency domain. Many models through the various approaches are introduced to estimate time-varying spectrum of nonstationary processes. There have been many attempts for the spectral analysis of nonstationary time series. In our study, we combined the spectral envelope with Tree-Based Segmentation for analyzing nonstationary process. To perform the spectral analysis for the categorical time series, we adopt the spectral envelope methodology by Stoffer et al(1993a)[12]. This provides the general framework for the spectral analysis of the categorical time series. To apply the classical method for the spectral analysis of stationary process to the nonstationary process, we use the TBAS(Tree-Based Adaptive Segmentation) in Adak(1998) and Ombao et al(2001)[8]. With this segmentation technique, we will partition the entire process into approximately stationary intervals. To find the best segmentation, we need to compute the distance measures for all possible sub-

series. It requires us too much time and efforts. However, this algorithm is expected to be computationally efficient because it uses the FFT(Fast Fourier Transform) and the data with dyadic length. For the best segmentation, we propose a decision function used in the decision criterion. This function is based on the Kullback-Leibler distance using spectral envelope. Based on the best segmentation, we will develop the local spectral envelope under the piecewise stationary process in Stoffer et al(2002)[11].

This thesis is organized in the following fashion. In section 2, we give a brief review of the previous works for the models and the spectral analysis of nonstationary processes. In section 3, some background materials are briefly described. Those are spectral envelope for stationary time series(Stoffer et al,1993), Tree-Based Adaptive segmentation(Ombao et al, 2001) and Kullback-Leibler divergence(Kullback and Leibler,1953). In chapter 4, the simulations are performed and the result of the simulations is shown. To test this algorithm, this method is applied to a real data. In the last chapter, we conclude with some discussion and future work.

2.0 LITERATURE REVIEW

The classical approach to the spectral analysis is based on the assumption that the time series is stationary. In reality, it happens that the data from various fields do not satisfy this assumption. To make an analysis of nonstationary time series within framework of classical method, some restrictions are placed on the class of nonstationary processes. In this section, we give a review of various approaches to the spectral analysis of the nonstationary processes and models of nonstationary process.

Priestley(1965)[10] first introduced the time-varying spectrum with Cramér representation for the analysis of nonstationary processes.

Definition 2.1(Priestley(1965))[10]

A stochastic process $\{X_t\}$ *is oscillatory if it has a representation of the form*

$$X_t = \int A(t,\lambda)e^{i2\pi\lambda t}dZ(\lambda) \tag{2.1}$$

where $Z(\lambda)$ *is an orthogonal increment process.*

The evolutionary power spectrum of the process is defined as

$$f(t,\lambda) = |A(t,\lambda)|^2 \tag{2.2}$$

To provide time-dependent spectra with the framework of stationary processes, the class of oscillatory processes was introduced. This definition is very similar to the Cramér representation of stationary process but the amplitude function, $A(t,\lambda)$ depends on the time t. The proposed nonstationary process is based on the assumption that the process is slowly changing, which allows the estimation of the time-dependent spectrum with some form of average spectrum in the neighborhood of any particular time-instant.

Dahlhaus(1997)[4] generalized this idea and proposed a class of nonstationary processes with an asymptotic framework. This asymptotic framework makes it possible to establish a reasonable asymptotic theory in the nonstationary time series. The Dahlhaus model has been the basis of many studies in developing models for nonstationary time series. Dahlhaus(2000) [5] also developed this model to multivariate nonstationary time series. The definition of the Dahlhaus model and corresponding time-varying spectrum are following.

Definition 2.2(Dahlhaus(1997))[4]

A sequence of zero-mean random variable $\{X_{t,T}\}$,$(t = 1, \ldots, T)$, *is called locally stationary, if there exists a representation*

$$X_{t,T} = \int_{-1/2}^{1/2} A^0_{t,T}(\omega) e^{i2\pi\omega t} dZ(\omega) \tag{2.3}$$

where $Z(\omega)$ *is an orthogonal increment and there exist constants* $\boldsymbol{C}$ *(*$\boldsymbol{C} \geq 0$*) and* D *and a function* $A : [0,1] \times [-1/2, 1/2] \longrightarrow \boldsymbol{C}$ *constant with* $A(u, -\omega)$ *such that for all* T,

$$\max_{t,\omega} \left| A^0_{t,T}(\omega) - A(t/T, \omega) \right| \leq DT^{-1} \tag{2.4}$$

The time-varying spectral density of Dahlhaus locally stationary process at time $u \in [0,1]$ *and frequency* $\omega \in [-1/2, 1/2]$ *is* $f(u,\omega) = |A(u,\omega)|^2$.

Increasing T does not provide more information about the future. It means that more data of the local structure are observed.

Adak(1998)[1] also introduced a new class of nonstationary process, *piecewise locally stationary process*, which is approximately piecewise stationary.

Definition 2.3(Adak,1998)[1]

A sequence of zero-mean stochastic processes is said to be piecewise locally stationary if it is locally stationary(By the definition 2.2) at all time point $u \in [0,1]$, *except possibly at finitely many jump points.*

Theorem 2.1(Adak,1998)[1]

Consider the class of sequences of the piecewise stationary process

$$M = \left\{ \begin{array}{l} \text{sequence} \left\{ \tilde{X}_{t,N}(t = 1, \ldots, N) \right\}_{N \geq 1} : \\ \tilde{X}_{t,N} = \sum_{j=0}^{J-1} \tilde{X}_t^{(j)} I \left(u_j \leq \frac{t}{N} \leq u_{j+1} \right) \\ \text{with } \frac{j}{N} \to 0 \text{as } N \to \infty \end{array} \right. \tag{2.5}$$

where $0 = u_1 \leq u_2 \leq \ldots \leq u_J = 1$ *form a partition of* [0,1] *that depends on* N *and* $\tilde{X}_t^{(j)}$ *are stationary processes with spectra* $f^{(j)}(\lambda)$*. Then, for any sequence of a piecewise locally stationary process* $X_{t,N}(t = 1, \ldots, N)$ *there exists a sequence* $\tilde{X}_{t,N} \in M$ *such that for all* N*,*

$$\frac{1}{N} \sum_{t=1}^{N} E(X_{t,N} - \tilde{X}_{t,N})^2 = O(N^{-2\alpha}) \tag{2.6}$$

where $0 < \alpha < 1$.

The above theorem shows that the piecewise locally stationary process can be approximated by the piecewise stationary process as the number of observations in block increases.

3.0 BACKGROUND

In this chapter, I will give some brief reviews of the previous works which are basis of my research study. First, I briefly mention the concept of the DNA sequence and give a review of Spectral envelope, Tree-Based Adaptive Segmentation and Kullback-Leibler divergence.

3.1 DNA SEQUENCE

In this thesis, I will apply my proposed method to the EBV DNA sequence. Before we discuss the methods for the spectral analysis of the DNA sequence, we need to understand the concept of the DNA sequence. The DNA sequence is a sequence of the letters which represents information of the DNA strand. The DNA strand is made up of a long string of chemical building blocks called "nucleotides". Each nucleotide is made up of nitrogenous base, a five carbon sugar, and a phosphate group. There are four different nitrogenous bases, which are labeled A(Adenine), T(Thymine), G(Guanine) and C(Cytosine). Nucleotides are arranged in two long strands that form a spiral called a double helix. The strands are complementary; Adenine with Thymine and Cytosine with Guanine. So, it is sufficient to represent a DNA molecule by the sequence of nitrogen bases on one single strand. The sequence of these bases determines the necessary information for living things to survive and reproduce. Determining the sequence is therefore useful in fundamental research into why and how organisms live, as well as in applied subjects. Because of the key nature of DNA to living things, knowledge of DNA sequence may come in useful in practically any biological research. For example, in medicine it can be used to identify, diagnose and potentially develop treatments for genetic diseases. My task is to extract the protein-coding

sequence hidden and discover patterns in the sequences. In the long DNA sequence, the coding sequences and noncoding sequences are scattered over the sequence. Coding sequence contains the instruction involved in making protein while the noncoding sequence does not. To analyze the DNA sequences, we need to extract the coding sequence in the DNA sequence and then find the genetic information stored in the coding sequence.

3.2 THE SPECTRAL ENVELOPE

Stoffer et al(1993a)[12] first introduced an approach for the spectral analysis of categorical time series. Spectral envelope is a useful tool to fulfill the spectral analysis for the categorical time series. To do spectral analysis for categorical time series, we need to assign the numerical values to categories. How to scale categorical time series is a crucial problem because the different scaling brings out different result of the sequence. One particular scaling emphasizes only one of harmonic component hidden in the sequence. Therefore, the first thing to do for the spectral analysis of the categorical time series is to find the optimal scaling. Our goal is not only to find an appropriate scaling but also to discover the periodicity in the sequence. Spectral envelope approach gives a solution of these problems, that is, it selects a proper scale and identifies the various periodic behaviors in a categorical time series. In this section, I will introduce the concept and theory of spectral envelope and then illustrate how to estimate the spectral envelope using DNA sequence.

3.2.1 The spectral envelope for stationary process

In this section, I will introduce the concepts of spectral envelope and application to categorical time series. The details are in Stoffer et al(1993a)[12].

Let X_t, $t = 0, \pm1, \pm2, \ldots$, be stationary time series with categorical values in $\{c_1, c_2, \ldots, c_k\}$. If the real value α_j is assigned to the category, c_j, and denote $h(X_t)$ be stationary time series with real values in $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)' \in \boldsymbol{R}^k$, then $h(X_t) = \alpha_j$. That is, $h(X_t)$ is the real-valued stationary time series which assigns the numerical value to each category c_j of X_t, α_j, $j = 1, 2, \ldots, k$. Then, a k-dimensional stationary time series, $\boldsymbol{Y}_t$ is defined by $\boldsymbol{z}_j$ and $\mathbf{0}$ as follows.

$$\boldsymbol{Y}_t = \begin{cases} \boldsymbol{z}_j & \text{if } X_t = c_j \text{ for } j = 1, \ldots, k-1, \\ \mathbf{0} & \text{if } X_t = c_k \end{cases} \tag{3.1}$$

where $\boldsymbol{z}_j$ is a $k \times 1$ vector with a one in the jth row and zeros elsewhere and $\mathbf{0}$ is a $k \times 1$ vector of zeros. Therefore, the time series $h(X_t)$ can be obtained from the time series $\boldsymbol{Y}_t$ by

the relationship, $h(X_t) = \boldsymbol{\alpha}'\boldsymbol{Y}_t$.

Given $x_1, x_2, \ldots, x_n$ on a categorical time series, $\boldsymbol{Y}_t$ can be formed by the equation (3.1). For any fixed t, $\boldsymbol{Y}_t$ is a single observation from a multinomial sampling scheme. The jth component of $\boldsymbol{Y}_t$ indicates whether X_t is in state c_j or not. Therefore, the variance-covariance matrix of $\boldsymbol{Y}_t$, $V = D - \boldsymbol{p}\boldsymbol{p}'$. Here, $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)'$, where $p_j = P(X_t = c_j)$, for $j = 1, 2, \ldots, k$ and D is the $k \times k$ diagonal matrix $D = \text{diag}\{p_1, p_2, \ldots, p_k\}$, therefore, the rank$(V)= k - 1$. For any $k \times (k-1)$ full rank matrix A whose columns are linearly independent of $\mathbf{1}_k$, $A'VA$ is a $(k-1) \times (k-1)$ positive definite symmetric matrix.

To find optimal scalings $\boldsymbol{\alpha}$, we maximize the variance at each frequency, ω and choose $\boldsymbol{\alpha}$ such that

$$\lambda(\omega) = \max_{\boldsymbol{\alpha}} \left\{ \frac{f_X(\omega; \boldsymbol{\alpha})}{\sigma^2(\boldsymbol{\alpha})} \right\} \tag{3.2}$$

for $\boldsymbol{\alpha} \not\propto \mathbf{1}_k$, where $f_X(\omega; \boldsymbol{\alpha})$ is the spectral density of $h(X_t)$ and $\sigma^2(\boldsymbol{\alpha})$ is var $\{h(X_t)\}$. For $\boldsymbol{\alpha} \propto \mathbf{1}_k$, $\lambda(\omega)$ is not defined because the scaling assigns each category the same value and $f_X(\omega; \boldsymbol{\alpha}) = 0$ and $\sigma^2(\boldsymbol{\alpha}) = 0$. Here, $\lambda(\omega)$ has the desirable property of being invariant under location and scale changes.

From the equation (3.1), $h(X_t)$ and spectral density of $h(X_t)$, $f_X(\omega; \boldsymbol{\alpha})$, can be represented by $\boldsymbol{Y}_t$. That is, $h(X_t) = \boldsymbol{\alpha}'\boldsymbol{Y}_t$ and $f_X(\omega; \boldsymbol{\alpha}) = \boldsymbol{\alpha}' f_Y(\omega)\boldsymbol{\alpha} = \boldsymbol{\alpha}' f_Y^{re}(\omega)\boldsymbol{\alpha}$, where $f_Y(\omega)$ is a $k \times k$ complex-valued Hermitian matrix[1] and $f_Y^{re}(\omega)$ denotes the real part of $f_Y(\omega)$. Since the entries on the main diagonal of any Hermitian matrix are necessarily real and the imaginary part of a Hermitian matrix is skew symmetric[2], that is, $\boldsymbol{\alpha}' f_Y(\omega)\boldsymbol{\alpha} = \boldsymbol{\alpha}' f_Y^{re}(\omega)\boldsymbol{\alpha}$.

[1] A Hermitian matrix is a square matrix with complex entries which is equal to its own conjugate transpose, that is, the element in the ith row and jth column is equal to the complex conjugate of the element in the jth row and ith column, for all ind ices i and j

[2] A square matrix A whose transpose is also its negative; that is, $A' = -A$

Let V be the variance-covariance matrix of $\boldsymbol{Y}_t$ then the equation (3.2) is represented as follows.

$$\lambda(\omega) = \max_{\boldsymbol{\alpha}} \left\{ \frac{\boldsymbol{\alpha}' f_Y^{re}(\omega) \boldsymbol{\alpha}}{\boldsymbol{\alpha}' V \boldsymbol{\alpha}} \right\} \tag{3.3}$$

From the above equation(3.3), we notice that the spectral envelope is the largest eigenvalue from the spectral density for any particular scaled process at any given frequency. The value of spectral envelope at any given frequency represents the largest proportion of the total power(variance) which is featured by the frequency.

The theory on estimation of the spectral density for a multivariate time series is applied to the estimation of $f_Y(\omega)$. If $f_Y(\omega)$ is estimated, then we can obtain the estimates $\hat{\lambda}(\omega)$ and $\hat{\alpha}(\omega)$. Stoffer et al(1993a)[12] established the asymptotic distribution of the spectral envelope. The details related to estimations and inferences are in Stoffer et al(1993a)[12]. The main results are as follows:

Theorem 3.1(Stoffer et al(1993a)) *If $\hat{f}_Y(\omega)$ is a consistent spectral estimator and if for each $j = 1, \ldots, J$, the largest root of $f_Y^{re}(\omega)$ is distinct, then*

$$\left\{ \eta_n[\hat{\lambda}(\omega_j) - \lambda(\omega_j)]/\lambda(\omega_j), \;\; \eta_n[\hat{\boldsymbol{\alpha}}(\omega_j) - \boldsymbol{\alpha}(\omega_j)]; j = 1, \ldots, J \right\} \tag{3.4}$$

converges jointly in distribution to independent zero-mean, normal distribution as $n \to \infty$.

The value of η_n in the equation (3.4) depends on the type of the estimator. In our study, the smoothed periodogram matrix, $I_n(\omega_j) = \hat{f}_X = \sum_{l=-m}^{m} h_l I_n(\omega_j + l/n)$. If we use the smoothed periodogram matrix with weight h_l, then $\eta_n^{-2} = \sum_{l=-m}^{m} {h_l}^2$.

Lemma 3.1(Stoffer et al(1993a)) *Using a first-order Taylor expansion, we have*

$$\log\hat{\lambda}(\omega) \approx \log\lambda(\omega) + \frac{\hat{\lambda}(\omega) - \lambda(\omega)}{\lambda(\omega)}, \tag{3.5}$$

thus, $\eta_n[\log\hat{\lambda}(\omega_j) - \log\lambda(\omega_j)]$ *is approximately standard normal.*

These results will be used in establishing the asymptotic property of the distance measure in our study.

3.2.2 Estimation on the spectral envelope

I briefly explained the theoretical backgrounds on the estimation of the spectral envelope for categorical time series in the previous section. To illustrate how to estimate the spectral envelope, I applied the spectral envelope to the DNA sequence.

- Let X_t for $t = 1, \ldots, n$ be the DNA sequence with the categorical value in (A,C,G,T). Each alphabet represents the type of nucleotide which contains the genetic information. The scaled sequence $h(X_t) = \boldsymbol{\alpha}'\boldsymbol{Y}_t$ and the sequence $\boldsymbol{Y}_t$ is formed by equation(3.1) as follows.

$$\begin{aligned}
\boldsymbol{Y}_t &= (1,0,0)', \text{if } \; X_t = A \\
\boldsymbol{Y}_t &= (0,1,0)', \text{if } \; X_t = C \\
\boldsymbol{Y}_t &= (0,0,1)', \text{if } \; X_t = G \\
\boldsymbol{Y}_t &= (0,0,0)', \text{if } \; X_t = T
\end{aligned}$$

- Calculate the periodogram matrix.

$$I_n(\omega_j) = \boldsymbol{d}(\omega_j)\boldsymbol{d}^T(\omega_j) \tag{3.6}$$

where $\boldsymbol{d}(\omega_j) = n^{-1/2}\sum_{t=1}^{n}\boldsymbol{Y}_t e^{-2\pi i\omega_j t}$ at $\omega_j = j/n$ for $j = 1, \ldots, [n/2]$.

- Smooth the periodogram with the weight h_l.

$$\hat{f_Y^{re}}(\omega_j) = \sum_{l=-m}^{m} h_l I_n^{re}(\omega_j + l/n) \tag{3.7}$$

where I_n^{re} denotes the real part of $I_n(\omega)$ and the weights are chosen such that $h_l = h_{-l} > 0$ and $\sum_{l=-m}^{m} h_l = 1$

- Calculate the sample variance-covariance matrix, $\boldsymbol{S}$ and determine the sample spectral envelope $\hat{\lambda}(\omega)$. The sample spectral envelope is the largest eigenvalue of the matrix in the equation (3.8)

$$\left\{2n^{-1}\boldsymbol{S}^{-1/2}\hat{f_y^{re}}(\omega_j)\boldsymbol{S}^{-1/2}\right\} \tag{3.8}$$

where $\bar{\boldsymbol{Y}} = n^{-1}\sum_{t=1}^{n}\boldsymbol{Y}_t$ and $\boldsymbol{S} = n^{-1/2}\sum_{t=1}^{n}(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})$.

- Find the optimal sample scaling. If the eigenvector corresponding to the largest eigenvalue is $\boldsymbol{a}$, then the optimal sample scaling $\hat{\boldsymbol{\alpha}} = \boldsymbol{S}^{-1/2}\boldsymbol{a}$

3.3 TREE BASED ADAPTIVE SEGMENTATION

To apply the classical methods for the stationary process to the nonstationary process, we need to place some restrictions on the nonstationary process. In our study, we assume that the time series is piecewise stationary. To estimate the local spectral envelope based on the piecewise stationary process, the segmentation should be known. For the best segmentation, I applied the Tree-Based Adaptive Segmentation algorithm to the segmentation procedure. Each segment of those obtained from segmentation should be stationary. The used segmentation techniques are delivered by Adak(1998) [1]. The main idea of the segmentation algorithm is to divide the entire series into small blocks and merge two adjacent blocks if they have similar spectral behavior; otherwise, they are left as a distinct block. The algorithm is as follows[3].

1. **Set the maximum level** K. The value of K represents the maximum depth of tree and determines the length of the smallest blocks. If K=5, the length of the smallest block is $T/2^5$, where T is the length of the entire series.
2. **Set the blocks**. For $k = 0, \ldots, K$, divide the entire sequence into 2^k blocks. Denote $B(k, l)$ to be the l-th block on level k, where $l = 1, \ldots, 2^k$. The block $B(k, l)$ contains $T/2^k$ observations, $\left\{X_{N_k(l-1)+1}, \ldots, X_{N_k(l)}\right\}$, where N_k is the length of the blocks on the level k.
3. **Estimate the distance between two blocks**. Let $D(k, l)$ be the distance measure between two adjacent blocks, $B(k + 1, 2l)$ and $B(k + 1, 2l - 1)$. Compute the estimates of the distances $D(k, l)$ for $k = K - 1, \ldots, 0$ and $l = 1, \ldots, 2^k$.
4. **Mark the blocks for final segmentation**. For $k = K - 1, \ldots, 0$ and $l = 1, \ldots, 2^k$, if $D(k, l) \leq D(k + 1, 2l - 1) + D(k + 1, 2l)$ then mark the block $B(k, l)$. Otherwise, leave the block $B(k, l)$ unmarked and set $D(k, l) = D(k + 1, 2l - 1) + D(k + 1, 2l)$.
5. **Final Segmentation** Determine the final segmentation. The final segmentation will be set of the highest marked blocks which is marked and its ancestor blocks are not marked.

[3] I follow the notation in Adak(1998)[1]

Table 1: Tree Based Adaptive Segmentation

Level	$B(j,l)$															
j=0	B(0,0)															
1	B(1,1)								B(1,2)							
2	B(2,1)				B(2,2)				B(2,3)				B(2,4)			
3	B(3,1)		B(3,2)		$\cdots$								$\cdots$		B(3,8)	
4	B(4,1)	B(4,2)	$\cdots$												$\cdots$	B(4,16)

This algorithm compares a parent block against its children blocks, such as, $D(k,l)$ versus $D(k+1,2l-1) \cup D(k+1,2l)$. If the value of distance at the parent block is smaller than the sum of the values for its children blocks, then the children blocks are chosen. If the value of distance at the parent block is greater than or equal to the sum of the values for its children blocks, then the parent block would be selected.

In searching for the best segmentation, we need to compute the distance measures for all possible sub series. It requires us too much time and efforts. However, this algorithm is computationally efficient and can handle massive datasets because it uses the FFT(Fast Fourier Transform) and the data with dyadic length.

3.4 KULLBACK-LEIBLER DIVERGENCE

The Kullback-Leibler divergence is usually used to measure the difference between two probability distributions in probability theory and information theory. The Kullback-Leibler divergence is defined as following.

Definition 3.1(Kullback and Leibler,1951[7])
Let $\boldsymbol{p}(x)$ *and* $\boldsymbol{q}(x)$ *denote the probability density functions of random variable X. Then Kullback-Leibler divergence between* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *is defined by*

$$I(\boldsymbol{p}(x), \boldsymbol{q}(x)) = \sum \left\{ \log \frac{\boldsymbol{p}(x)}{\boldsymbol{q}(x)} \right\} \boldsymbol{p}(x) \tag{3.9}$$

This divergence has some properties as follows.
(i) $I(\boldsymbol{p}(x), \boldsymbol{q}(x)) \geq 0$ *with equality if and only if* $\boldsymbol{p} = \boldsymbol{q}$ *;*
(ii) $I(\boldsymbol{p}(x), \boldsymbol{q}(x))$ *is not symmetric.*

Given random sample, $x_1, x_2, \ldots, x_n$, then the Kullback-Leibler divergence is

$$I(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)} \tag{3.10}$$

In my thesis, the Kullback-Leibler divergence is used as a discrepancy measure between the spectral envelope estimates of two subseries for the best segmentation. The value of the divergence determines the degree of dissimilarity between two adjacent blocks. If the divergence between any two adjacent blocks is small, we combine two blocks because the small value of divergence means that two adjacent blocks contain similar spectral information. If not, that means that two adjacent blocks are distinct and left them as separate blocks.

4.0 SPECTRAL ANALYSIS OF NONSTATIONARY TIME SERIES

Our goal is to explore the cyclic behavior in nonstationary categorical time series. To find the optimal scalings and extract the spectral information in the categorical time series, we propose to use the spectral envelope methodology. When we scale the categorical time series X_t in terms of unit vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$ defined in the section 3.2, the categorical time series, X_t is represented by the multiple real-valued time series $\boldsymbol{Y}_t$. To perform the spectral analysis of the multiple time series $\boldsymbol{Y}_t$, we need to use the spectrum matrix of $\boldsymbol{Y}_t$, $f_{\boldsymbol{Y}}(\omega)$. But it is difficult to deal with $f_{\boldsymbol{Y}}(\omega)$ because it is a function into the set of complex Hermitian matrix. The spectral envelope methodology summarizes the spectrum matrix into the useful information understood easily and minimizes the loss of information. As aforementioned, many time series in varied fields is very long and shows nonstationary behaviors. For the spectral analysis of nonstationary process, we propose to use the model of piecewise stationary process and Tree-Based Adaptive Segmentation algorithm in Adak(1998)[1]. We will develop the local spectral envelope based on the model of piecewise stationary process to estimate the local spectral envelope. However, the local spectral envelope on the model of piecewise stationary process is based on the assumption that the segmentation is known. If the segmentation is unknown, we need to find the best segmentation. To find the best segmentation, we adopt the the Tree-Based Adaptive Segmentation algorithm in Adak(1998)[1]. For developing a decision function, we need to establish a distance measure which can show the discrepancy well. In this dissertation, I propose the Kullback-Leibler divergence using the spectral envelope as a decision function.

4.1 LOCAL SPECTRAL ENVELOPE OF NONSTATIONARY PROCESS

The spectral envelope described in the section 3.2 is based on the assumption that the sequence is stationary. As aforementioned, the processes in the real problems do not satisfy this assumption. Thus, we need a suitable model for the spectral analysis of nonstationary process. In our study, we consider the model of the piecewise stationary process to estimate the local spectral envelope. The theory of the estimation of the local spectral envelope follows from the results in the Stoffer et al(2002))[11].

Definition 4.1(Stoffer et al(2002))[11]
A $k \times 1$ vector-valued piecewise stationary process, $\{\boldsymbol{Y}_{s,T}\}_{s=0}^{T-1}$, for $T \geq 1$, is defined to be

$$\boldsymbol{Y}_{s,T} = \sum_{b=1}^{B} \boldsymbol{Y}_{s,b} I(s/T, U_b) \tag{4.1}$$

where $\boldsymbol{Y}_{s,b}$ are stationary process with $k \times k$ spectral density matrix $f_{Y,b}(\omega)$, $U_b = [u_{b-1}, u_b) \subset [0,1)$ is an interval, and $I(s/T, U_b)$ is an indicator function which is 1 if $s/T \in U_b$.

For ease of notation, we rescaled the time in $\boldsymbol{Y}_{s,b}$ using time shift from the beginning of each segment.

$$\{\boldsymbol{Y}_{s,b} : s/T \in U_b\} \longmapsto \{\boldsymbol{Y}_{t,b} : t = 0, \ldots, M_b - 1\} \tag{4.2}$$

where the number of observations in segment b is M_b and $\sum_{b=1}^{B} M_b = T$.

If the $k \times 1$ process, $\boldsymbol{Y}_{s,T}$ is piecewise stationary, then the categorical time series, $\{X_{s,T}\}$ is piecewise stationary. If $\{X_{s,T}\}$ is piecewise stationary, we can use the criterion in the equation(3.3) to define the local spectral envelope. That is,

$$\lambda_b(\omega) = \max_{\boldsymbol{\alpha} \not\propto \mathbf{1}_k} \left\{ \frac{\boldsymbol{\alpha}' f_{Y,b}^{re}(\omega) \boldsymbol{\alpha}}{\boldsymbol{\alpha}' V_b \boldsymbol{\alpha}} \right\} \tag{4.3}$$

for $b = 1, \ldots, B$, where V_b is the variance-covariance matrix of $\boldsymbol{Y}_{t,b}$.

Suppose that the categorical piecewise stationary processes, $x_{t,T}$, for $t = 0, \ldots, T-1$ is observed and the segmentation is known then we can follow the procedure in section 3.2 to estimate the sample local spectral envelope. If the $\boldsymbol{Y}_{t,b}$ is formed by the equation (3.3), then we can also define the local periodogram $I_b(\omega)$.

$$I_b(\omega) = \boldsymbol{d}_b(\omega)\boldsymbol{d}_b{}^T(\omega) \tag{4.4}$$

where $\boldsymbol{d}_b(\omega) = {M_b}^{-1/2} \sum_{t=0}^{M_{b-1}} \boldsymbol{Y}_{t,b} \exp\{-2\pi i t\omega\}$.

The local sample spectral envelope $\hat{\lambda_b}(\omega)$ is defined to be the largest value of the matrix as follows.

$$\hat{g}_b(\omega) = \hat{V}_b^{-1/2} \hat{f}_{Y,b}(\omega) \hat{V}_b^{-1/2} \tag{4.5}$$

where $\hat{f}_{Y,b} = (2m+1)^{-1} \sum_{l=-m}^{m} I_b(\omega + l/M_b)$.

When the categorical time series $X_{t,T}$ is piecewise stationary and the segmentation is known, we can find asymptotic distribution of the sample local spectral envelope using Theorem 3.1 and Lemma 3.1 in the section 3.2 under some assumptions. That is, $\nu_{M_b}\left[\log\hat{\lambda}_b(\omega) - \log\lambda_b(\omega)\right]$ is approximately standard normal under some conditions. Therefore, $E\left[\log\hat{\lambda}_b((\omega)\right] \approx \log\lambda_b(\omega)$ and $\text{var}\left[\log\hat{\lambda}_b(\omega)\right] \approx \nu_{M_b}^{-2}$.

4.2 PROPOSED DISTANCE MEASURE

In our study, the segmentation is very important step because the estimation of the local spectral envelope is based on the assumption that the time series is piecewise stationary process and the segmentation is known. The estimation of local spectral envelope depends on the segmentation. In reality, the segmentation is unknown. Thus, we need to find the best segmentation which approximate the real segmentation well. For the best segmentation, the selection of the distance measure is very important. The choice of the distance measure greatly affects the result of the segmentation. In this thesis, we propose the Kullback-Leibler distance based on the spectral envelope as a distance measure for the best segmentation. Kullback-Leibler divergence has been used to measure the difference between two probability distributions in varied circumstances. In our study, we establish a decision function based on the Kullback-Leibler distance using spectral envelope.

Let $\hat{\lambda}_{k+1,2l}(\omega_j)$, $\hat{\lambda}_{k+1,2l-1}(\omega_j)$ be the local sample spectral envelope at the frequency ω_j for each block $B(k+1,2l)$ and $B(k+1,2l-1)$. The Kullback-Leibler divergence between two blocks, $B(k+1,2l)$ and $B(k+1,2l-1)$, is defined to be

$$D(k,l) = \frac{1}{M_k/2+1} \sum_{j=0}^{M_k/2+1} \hat{\lambda}_{k+1,2l}(\omega_j) \log \frac{\hat{\lambda}_{k+1,2l}(\omega_j)}{\hat{\lambda}_{k+1,2l-1}(\omega_j)} \tag{4.6}$$

However, as we mentioned in the section 3.5, this measure is not symmetric. Thus, we use the symmetrised divergence in our study. The symmetrised divergence is defined by the following:

$$\begin{aligned} I(\mathbf{p},\mathbf{q}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{ p_i \log\frac{p_i}{q_i} + q_i \log\frac{q_i}{p_i} \right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{ [(p_i - q_i)] \log\frac{p_i}{q_i} \right\} \end{aligned} \tag{4.7}$$

Thus, the distance measure between two block ,$B(k+1,2l)$ and $B(k+1,2l-1)$, is

$$D(k,l) = \frac{1}{M_k/2+1} \sum_{j=0}^{M_k/2} \left[\hat{\lambda}_{k+1,2l}(\omega_j) - \hat{\lambda}_{k+1,2l-1}(\omega_j) \right] \log \frac{\hat{\lambda}_{k+1,2l}(\omega_j)}{\hat{\lambda}_{k+1,2l-1}(\omega_j)} \tag{4.8}$$

4.2.1 Asymptotic property of distance measure

In our study, the distance measure between two blocks $B(k+1,2l)$ and $B(k+1,2l-1)$ is defined by

$$D(k,l) = \frac{1}{M_k/2+1} \sum_{j=0}^{M_k/2} \left\{ \left[\hat{\lambda}_{k+1,2l}(w_j) - \hat{\lambda}_{k+1,2l-1}(w_j) \right] \log \frac{\hat{\lambda}_{j+1,2l}(w_j)}{\hat{\lambda}_{j+1,2l-1}(w_j)} \right\} \tag{4.9}$$

where $\hat{\lambda}_{k+1,2l}(w_j)$ is the local sample spectral envelope at the frequency w_j for the block $B(k+1,2l)$ and $\hat{\lambda}_{k+1,2l-1}(w_j)$ is for the block $B(k+1,2l-1)$.

According to Stoffer et.al(2002)[11], $\log\hat{\lambda}(w_j)$ approximately follows normal distribution with mean, $\log\lambda(w_j)$, and variance, $\nu_{M_b}^{-2}$. Thus, $\log\frac{\hat{\lambda}_1(w_j)}{\hat{\lambda}_2(w_j)}$ also approximately follows normal distribution with mean, $\log\frac{\lambda_1(w_j)}{\lambda_2(w_j)}$ and variance, $2\nu_{M_b}^{-2}$, that is,

$$\left[\log \frac{\hat{\lambda}_1(w_j)}{\hat{\lambda}_2(w_j)} \right] \approx N\left(\log \frac{\lambda_1(w_j)}{\lambda_2(w_j)}, 2\nu_{M_b}^{-2} \right) \tag{4.10}$$

In the process of the final segmentation, if $D(k,l) < D(k+1,2l) + D(k+1,2l-1)$, then mark the block $B(k,l)$. Otherwise, leave the block $B(k,l)$ unmarked and set $D(k,l) = D(k+1,2l-1) + D(k+1,2l)$. That is, if $B(k+1,2l)$ and $B(k+1,2l-1)$ have similar spectral information, then value of the distance measure $D(k,l)$ is close to 0. Thus, if the sub blocks have similar spectral information, then $Pr(D(k,l) < D(k+1,2l) + D(k+1,2l-1))$ is close to 1.

To show that our distance measure works well in searching for the best segmentation, we need to prove that

$$\lim_{M_k \to \infty} Pr\left(D(k,l) < D(k+1,2l) + D(k+1,2l-1) \right) = 1 \tag{4.11}$$

when the $B(k+1,2l)$ and $B(k+1,2l-1)$ have same spectral information.

As $M_k \to \infty$, more data of local structure are observed. That means that more observations provide more information on the subblocks.

From the equation (4.10),

$$\lim_{M_k \to \infty} Pr\left(\left|\log\frac{\hat{\lambda}_{k+1,2l}(w_j)}{\hat{\lambda}_{k+1,2l-1}(w_j)} - \log\frac{\lambda_{k+1,2l}(w_j)}{\lambda_{k+1,2l-1}(w_j)}\right| \geq \epsilon\right) \tag{4.12}$$

$$\leq \frac{Var\left(\log\frac{\hat{\lambda}_{k+1,2l}(w_j)}{\hat{\lambda}_{k+1,2l-1}(w_j)}\right)}{\epsilon^2} = \frac{Var\left(\log\hat{\lambda}_{k+1,2l}(w_j)\right) + Var\left(\log\hat{\lambda}_{k+1,2l-1}(w_j)\right)}{\epsilon^2} \tag{4.13}$$

where $\nu_{M_k}^{-2} = \sum_{-m}^{m} {h_l}^2$, $h_l = \frac{1}{2m+1}$.

As $M_k \to \infty$, $Var\left(\log\hat{\lambda}_{k+1,2l}(w_j)\right) + Var\left(\log\hat{\lambda}_{k+1,2l-1}(w_j)\right) \to 0$.
Thus,

$$\lim_{M_k \to \infty} Pr\left(\left|\log\frac{\hat{\lambda}_{k+1,2l}(w_j)}{\hat{\lambda}_{k+1,2l-1}(w_j)} - \log\frac{\lambda_{k+1,2l}(w_j)}{\lambda_{k+1,2l-1}(w_j)}\right| \geq \epsilon\right) = 0 \tag{4.14}$$

Therefore, $\log\frac{\hat{\lambda}_{k+1,2l}(w_j)}{\hat{\lambda}_{k+1,2l-1}(w_j)}$ converges in probability to $\log\frac{\lambda_{k+1,2l}(w_j)}{\lambda_{k+1,2l-1}(w_j)}$.

If the spectral envelopes of two blocks, $B(k+1, 2l)$ and $B(k+1, 2l-1)$ are same, then

$$\lambda_{k+1,2l-1}(w_j) = \lambda_{k+1,2l}(w_j) \tag{4.15}$$

at every frequency w_j, where $j = 0, 1, \ldots, M_{k+1}/2$.
Thus,

$$\left[\log\frac{\hat{\lambda}_{k+1,2l}(w_j)}{\hat{\lambda}_{k+1,2l-1}(w_j)}\right] \quad \overrightarrow{p} \quad \log\frac{\lambda_{k+1,2l}(w_j)}{\lambda_{k+1,2l-1}(w_j)} = 0 \tag{4.16}$$

By the Theorem 3.1, $\left(\hat{\lambda}_{k+1,2l}(w_j) - \hat{\lambda}_{k+1,2l-1}(w_j)\right)$ follows normal distribution with mean,

$(\lambda_{k+1,2l}(w_j) - \lambda_{k+1,2l-1}(w_j))$. Thus, $\left(\hat{\lambda}_{k+1,2l}(w_j) - \hat{\lambda}_{k+1,2l-1}(w_j)\right)$ converges in distribution to $(\lambda_{k+1,2l}(w_j) - \lambda_{k+1,2l-1}(w_j))$.

By the Slutsky's theorem,

$$\left\{ \left[\hat{\lambda}_{k+1,2l}(w_j) - \hat{\lambda}_{k+1,2l-1}(w_j)\right] \log \frac{\hat{\lambda}_{j+1,2l}(w_j)}{\hat{\lambda}_{j+1,2l-1}(w_j)} \right\} \xrightarrow{d} 0 \tag{4.17}$$

Therefore,

$$\lim_{M_k \to \infty} D(k,l) = 0 \tag{4.18}$$

$$D(k+1,2l) = \frac{1}{M_{k+1}/2+1} \sum_{j=0}^{M_{k+1}/2+1} \left[(\hat{\lambda}_{k+2,4l}(w_j) - \hat{\lambda}_{k+2,4l-1}(w_j)) \log \frac{\hat{\lambda}_{k+2,4l}(w_j)}{\hat{\lambda}_{k+2,4l-1}(w_j)} \right] \tag{4.19}$$

$$D(k+1,2l-1) = \frac{1}{M_{k+1}/2+1} \sum_{j=0}^{M_{k+1}/2+1} \left[(\hat{\lambda}_{k+2,4l-2}(w_j) - \hat{\lambda}_{k+2,4l-3}(w_j)) \log \frac{\hat{\lambda}_{k+2,4l-2}(w_j)}{\hat{\lambda}_{k+2,4l-3}(w_j)} \right] \tag{4.20}$$

That is,

$$\begin{aligned} &\lim_{M_k \to \infty} Pr(D(k,l) < D(k+1,2l) + D(k+1,2l-1)) \\ &= Pr\left(0 < (4.19) + (4.20)\right) \end{aligned} \tag{4.21}$$

But, each term ,$\left(\left[\hat{\lambda}_{k+2,4l-1}(w_j) - \hat{\lambda}_{k+2,4l-3}(w_j)\right] \log \frac{\hat{\lambda}_{k+2,4l-1}(w_j)}{\hat{\lambda}_{k+2,4l-3}(w_j)}\right)$, in (4.19) and (4.20) is non-negative for every frequency, w_j because difference of two spectral envelopes and log ratio of two spectral envelopes have same signs and the product of these is nonnegative.

Thus,

$$\lim_{M_k \to \infty} Pr(D(k,l) < D(k+1,2l) + D(k+1,2l-1)) = 1 \tag{4.22}$$

Therefore, as the sample size in the segment increases, the splitting criterion used in our segmentation become more precise in deciding the discrepancy between two subseries.

4.2.2 The segmentation algorithm

The segmentation algorithm used here is based on the Tree-Based Adaptive Segmentation for the spectral envelope as is used in Local Spectral Envelope in Stoffer et al(2002)[11]. The main idea is to partition the entire series into small blocks and recombine adjacent blocks which have similar genetic information from the estimated local spectral envelope. In this procedure, we use the Kullback-Leibler divergence using spectral envelope as a distance measure between two subblocks.

The algorithm is following :

1. **Set the maximum level** K. The value of K represents the maximum depth of tree and determines the size of the smallest blocks. For a sequence of length T, the size of the smallest blocks is $T/2^k$.
2. **Set the blocks**. For $k = 0, \ldots, K$, divide the data sequence into 2^k blocks. Denote $B(k, l)$ to be the l-th block on level k, where $l = 1, \ldots, 2^k$. The first block on level k is denoted as $B(k, 1)$ and the last as $B(k, 2^k)$. For block $B(k, l)$, it consists of the element of the sequence $\left\{X_{[(l-1)T/2^k+1]}, \ldots, X_{[lT/2^k]}\right\}$.
3. **Estimate the spectral envelope** $\widehat{\lambda}_{k,l}(\omega_j)$ **at each frequency** $\omega_j = j/M_k (j = 0, \ldots, M_k/2)$ **for block** $B(k, l)$**, where** $M_k = T/2^k, k = 0, \ldots, K$**.**
4. **Compute the Kullback-Leibler divergence** $D(k, l)$**(or relative entropy)**. For $k = 0, \ldots, K$ and $l = 1, \ldots, 2^k$,

$$D(k,l) = \frac{1}{M_k/2+1} \sum_{j=0}^{M_k/2} \left\{ \left[\hat{\lambda}_{k+1,2l}(w_j) - \hat{\lambda}_{k+1,2l-1}(w_j)\right] \log \frac{\hat{\lambda}_{j+1,2l}(w_j)}{\hat{\lambda}_{j+1,2l-1}(w_j)} \right\}$$

5. **Mark the blocks for final segmentation**. For $k = K-1, \ldots, 0$ and $l = 1, \ldots, 2^k$, if $D(k, l) \leq D(k+1, 2l-1) + D(k+1, 2l)$ then mark the block $B(k, l)$. Otherwise, leave the block $B(k, l)$ unmarked and set $D(k, l) = D(k+1, 2l-1) + D(k+1, 2l)$.
6. **Finalize the best segmentation**. The final segmentation is the set of the highest marked blocks and their ancestor blocks are unmarked.
7. **Estimated the local spectral envelope**. The local spectral envelope will be the combination of the spectral envelopes for the final segmentation.

From the estimated local spectral envelope, we can identify the coding sequence in the DNA sequence and extract the periodic component for each block.

5.0 SIMULATION RESULT AND DATA ANALYSIS

In this chapter, I apply the proposed method to the simulated data and the real data set. When the data set is simulated, the piecewise stationary time series is constructed by combining several stationary time series. To see if our algorithm works well in the real data set, a subseries of EBV DNA sequence is used.

5.1 SIMULATION RESULT

To test our algorithm, we apply our method to the simulated data. The simulated data contains the time domain of various signals and noise processes in a dyadic manner. The signal was generated as following.

$$X_1(t) = 2\cos(2\pi t/10) + \cos(2\pi t/3) + 0.3\epsilon_1(t) \quad (5.1)$$

$$X_2(t) = \cos(2\pi t/3) + 0.01\epsilon_2(t) \quad (5.2)$$

where $\epsilon_1(t)$ and $\epsilon_2(t)$ are Gaussian white noise with unit variance. The real values of two series are categorized into one of four letters (A, C, G or T). $C_1(t)$ and $C_2(t)$ represent the categorical sequence obtained by categorizing $X_1(t)$ and $X_2(t)$ and $N_1(t)$, $N_2(t)$, and $N_3(t)$ represent the simulated sequences with the categorical values from three different white noises. The first series, $X_1(t)$, contains 1/3 frequency and 1/10 frequency, whereas the second series, $X_2(t)$, contains only 1/3 frequency. In this simulation, let X_t represent a simulated DNA sequence of length of T=4096 and the decomposition is following:

$$X_t = \begin{cases} C_1(t), & 1 \leq t \leq 512 \\ N_1(t), & 513 \leq t \leq 1024 \\ N_2(t), & 1025 \leq t \leq 2048 \\ N_3(t), & 2049 \leq t \leq 3072 \\ C_2(t), & 3073 \leq t \leq 4096 \end{cases} \tag{5.3}$$

To find the best segmentation of the data set simulated by the above direction, we set the deepest level at $K = 4$ and calculated the sample spectral envelopes for all possible segments at each level, $j = 0,1,2,3$, and those are given in the Figure 1. Using those values, we performed the segmentation algorithm from the deepest level at $K = 4$. The distance between two adjacent subblocks is given in the Table 2. Table 3 is based on the recomputed distances and it shows the best segmentation. The recomputed distance is obtained by Step 5 in the algorithm and the best segmentation is finalized by Step 6 in the algorithm. The best segmentation obtained from the simulated sequence, X_t, is given in Table 4. From the best segmentation obtained from the segmentation algorithm, we notice that the best segmentation matches precisely to the segmentation of the generated data in 5.3. The local spectral envelope based on the best-segmented data is shown in the Figure 2. From the Figure 2, we notice that the block $B(3,1)$ has a narrow band peak at $\omega = 1/10$ and $\omega = 1/3$ and the block $B(2,4)$ has a significant peak at $\omega = 1/3$. We couldn't find any significant peak at any frequency from the spectral envelope for the blocks $B(3,2)$, $B(2,2)$ and $B(2,3)$. These blocks can be classified as the noises. Moreover, the block $B(3,2)$, the block $B(2,2)$ and the block $B(2,3)$ are classified as noises based on the estimated spectral envelope, the block $B(3,1)$ is classified as coding with frequencies $1/10$ and $1/3$ and the block $B(2,4)$ is classified as coding with frequency $1/3$. These results also corresponds to the way the data were generated.

Table 2: Distance using Kullback-Leibler distance for simulated EBV data

0.46															
0.47								0.88							
1.01				0.07				0.11				0.07			
0.05		0.1		0.1		0.1		0.13		0.18		0.1		0.09	
0.13	0.09	0.18	0.23	0.29	0.32	0.17	0.18	0.22	0.21	0.28	0.25	0.28	0.23	0.18	0.3

Table 3: Recomputed Distance using Kullback-Leibler distance

0.46→0.40															
0.47→0.22								0.88→0.18							
1.01→0.15				**0.07**				**0.11**				**0.07**			
0.05		**0.1**		**0.1**		**0.1**		**0.13**		**0.18**		**0.1**		**0.09**	
0.13	0.09	0.18	0.23	0.29	0.32	0.17	0.18	0.22	0.21	0.28	0.25	0.28	0.23	0.18	0.3

Table 4: Best Segmentation

Level	$B(j,l)$															
j=0																
1																
2					**B(2,2)**				**B(2,3)**				**B(2,4)**			
3	**B(3,1)**		**B(3,2)**													
4																

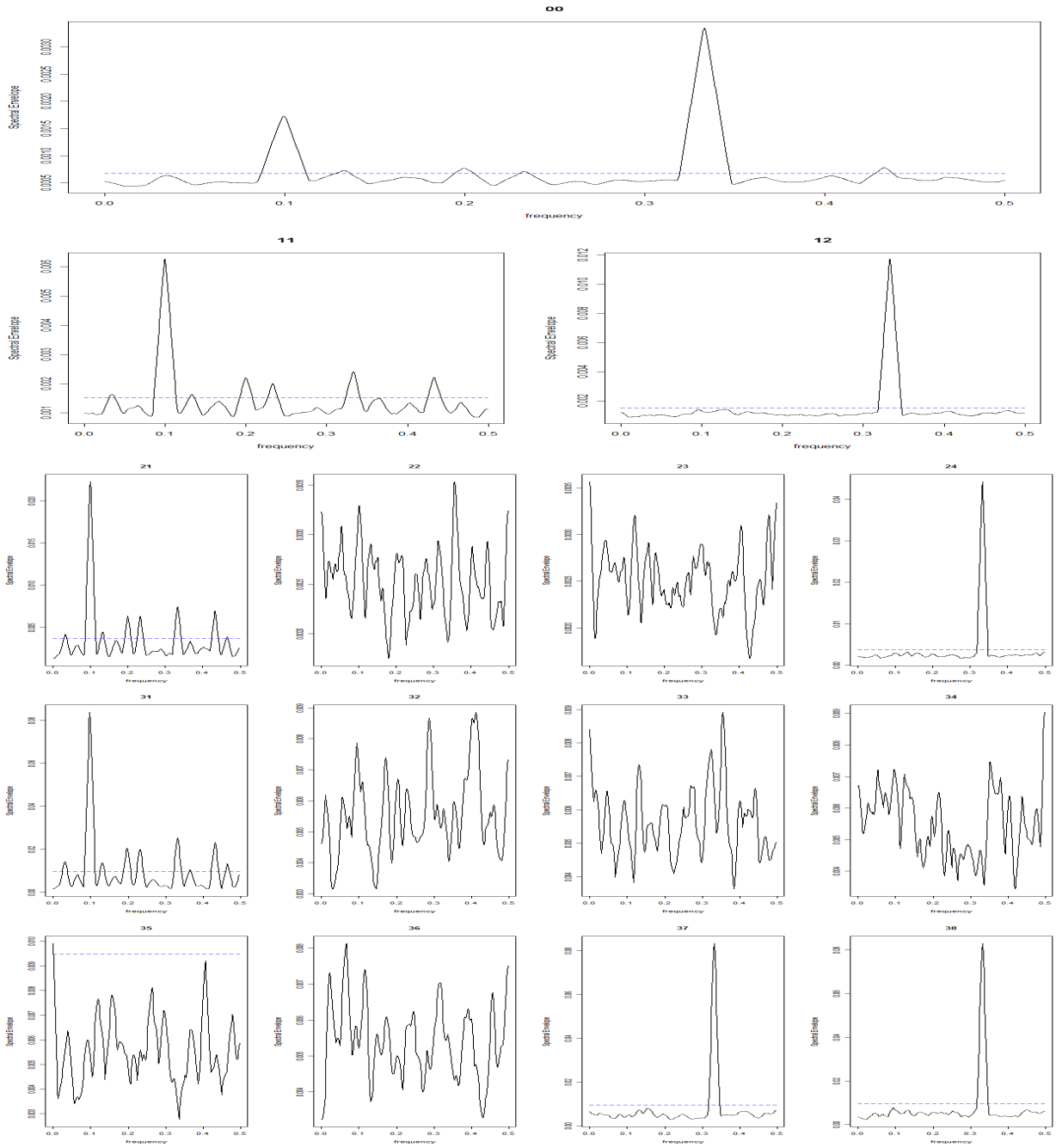

Figure 1: Estimated spectral envelopes of the Simulated data on each level

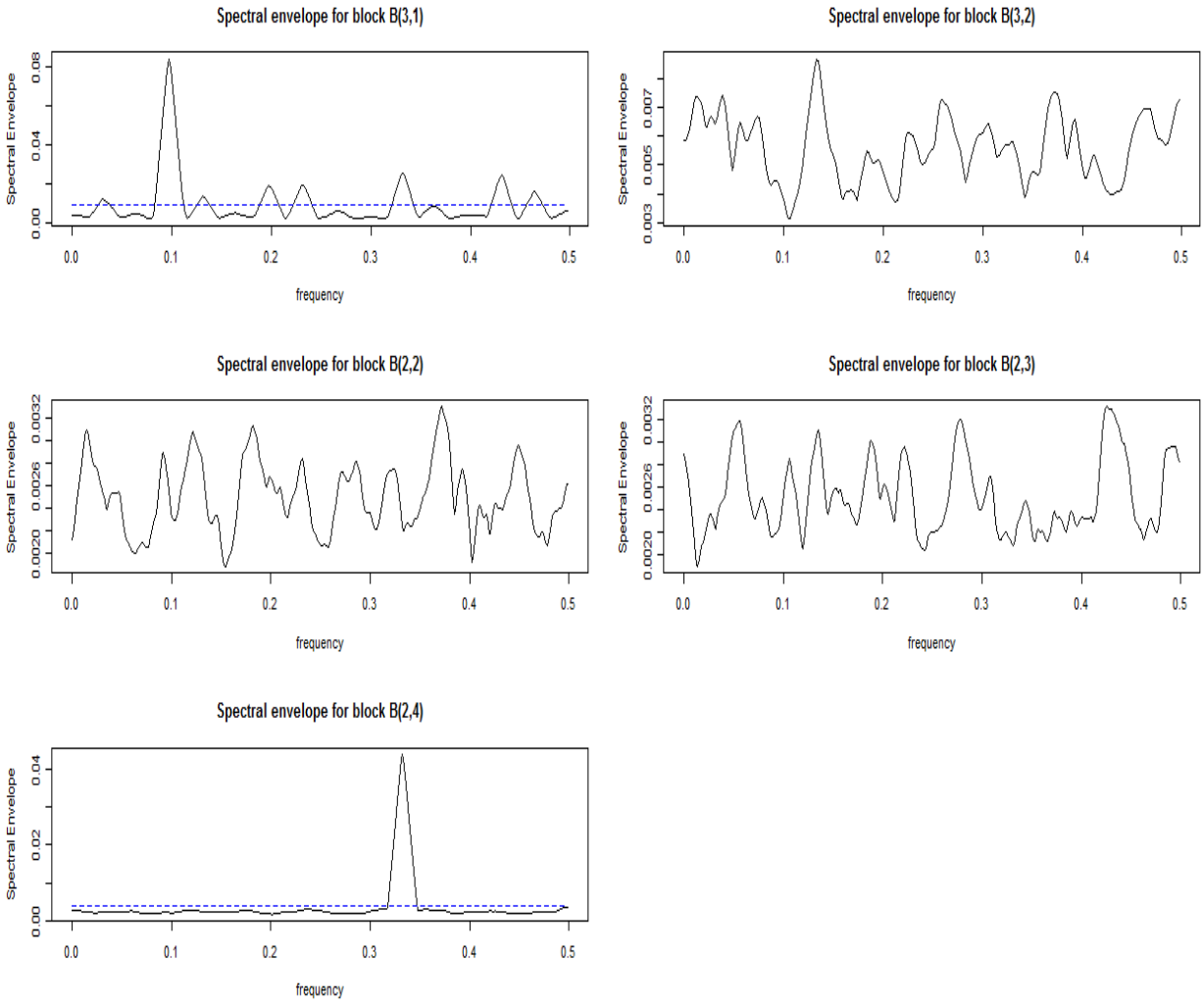

Figure 2: Estimated spectral envelopes of the Simulated data based on the best segmentation

5.2 ANALYSIS OF THE EBV DNA SEQUENCE

We applied our method to a subseries of the EBV DNA sequence which includes bp 46001 to 54192 with length of T=8192. From the EMBL(European Molecular Biology Laboratory) data file, we can find a list of the interesting part of these sequences.

```
CDS                 46333...47481
                    /Note ="BWRE1 reading fram 12"
CDS                 48386...50032
                    /Note ="Coding exon for EBNA-2"
repeat_region       50578...52115
                    /Note ="'12 x "'125bp"' repeat"
```

From the above notation, we can notice that the sequence contains two coding sequences(46333-47481 and 48386-500323) and one repeat region from bp50578 to 52115. From the experience with the Fourier analysis of DNA sequences and researches, it is known that a CDS typically contains the frequency at $w = 1/3$ and $w = 1/10$ and repeat regions may have many spectral peaks. In those repeat regions, the spectral envelope is either flat or the spectral power is concentrated around the zero frequency in Stoffer et. al(1993a)[12]. If the block contains coding only, then generally, a spectral peak will be shown at $w = 1/3$. If the block contains coding and noise, then spectral peaks will be at the frequency $w = 1/3$ and also be around at the zero frequency. If the block contains both coding and noncoding, then spectral peaks will be around the zero frequency, at frequency at $w = 1/3$ and sometimes other nonzero frequencies. If the block contains the repeat region, spectral peaks will be several nonzero frequencies other than 1/3.

In our algorithm, we set the lowest level at $K = 4$. The distance between two adjacent subblocks is given in the Table 5. Table 6 shows the recomputed distances and the best segmentation. The final segmentation by our algorithm is given in the Table.7 and the estimated spectral envelopes for the best segmented series are shown in Figure 3. From

Figure 3 and Figure 4, we notice that the estimated local spectral envelopes for the blocks $B(4,1)$, $B(3,2)$, $B(4,9)$ and $B(2,4)$ don't exhibit any significant peak at any frequency and the power is concentrated near the zero frequency. From these results, we can consider these four blocks containing noncoding regions. From the estimated local spectral envelope for $B(3,3)$ and $B(3,4)$, these two blocks have some periodic behavior at frequency $w = 1/3$ and the power is concentrated around the zero frequency. From these results, we can consider that these blocks, $B(3,3)$ and $B(3,4)$, contain both coding and noncoding regions. From the notation of EMBL data file, we can check that the two blocks $B(3,3)$ and $B(3,4)$ contain bp 48049 to 50096, contain another coding sequences(actual location is bp 48386 to 50032) and noncoding sequence (bp 50033-50096, bp 48049-48385). The estimated local spectral envelope for three block $B(4,10)$, $B(4,11)$ and $B(4,12)$ exhibit several nonzero peaks other than 1/3. From these results, we can consider that these three blocks contains repeat region. These three blocks contains bp 50609 to 52144. We can identify a large repeat region from bp 50609 to 52144 included in actual location of repeat region from bp 50578 to 51255. According to lists of coding sequences from EMBL data file, we can check that the segmentation obtained from our algorithm similarly identify the coding sequences in the DNA sequence.

Table 5: Distance using Kullback-Leibler distance for EBV data

18															
18								23							
16				8				27				17			
44		0		0		0		0		0		21		0	
0	0	0	0	0	17	21	0	0	0	0	0	0	44	0	0

Table 6: Recomputed Distance using Kullback-Leibler distance

18→17															
18→0								23→17							
16→0				8→0				27→0				**17**			
44→0		**0**		**0**		**0**		42→0		12→0		**21**		**0**	
0	**0**	0	0	0	17	21	0	**0**	**0**	**0**	**0**	0	44	0	0

Table 7: Best Segmentation

Level	$B(j,l)$															
j=0																
1																
2														**(2,4)**		
3			**(3,2)**		**(3,3)**		**(3,4)**									
4	(4,1)	(4,2)							(4,9)	(4,10)	(4,11)	(4,12)				

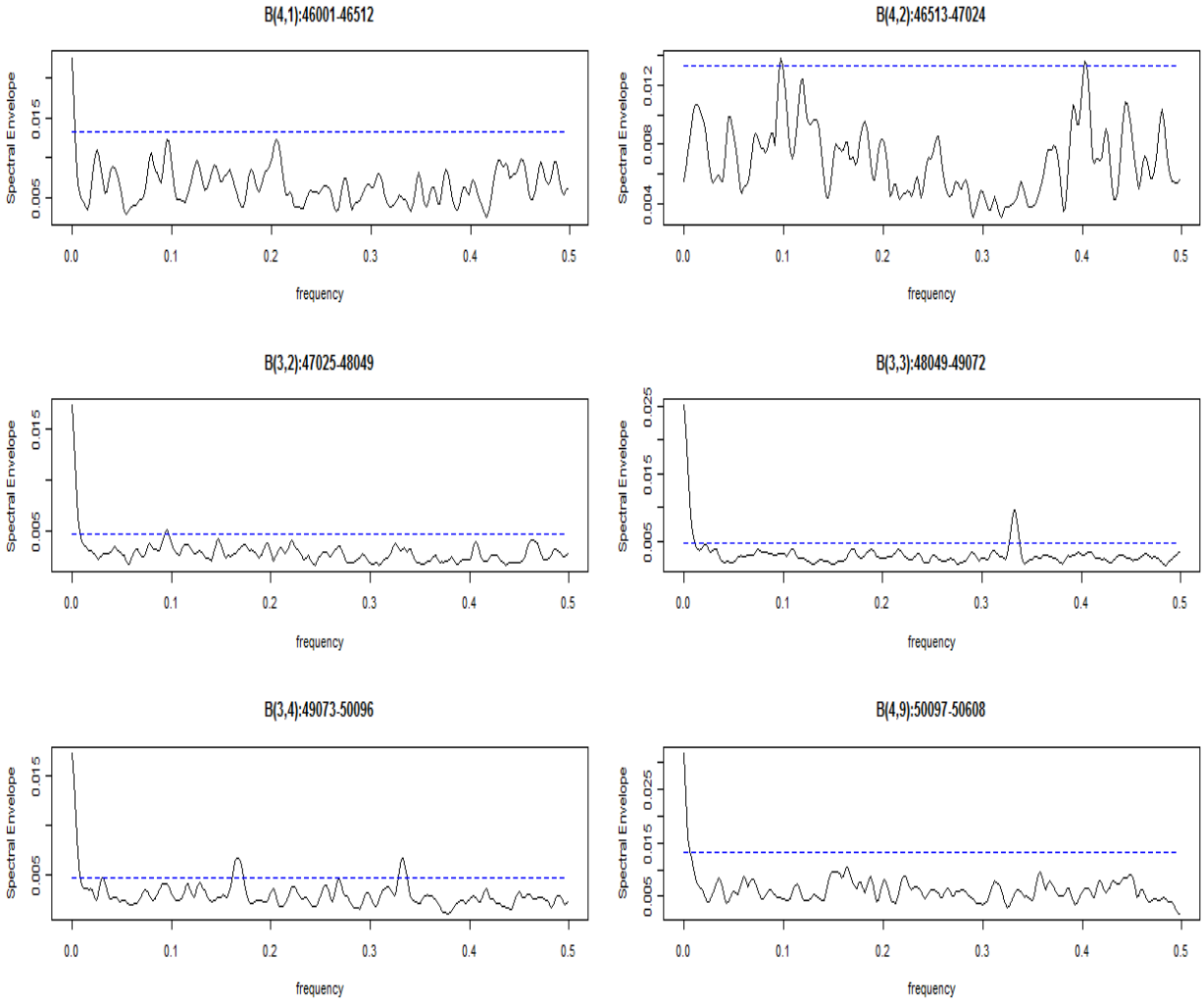

Figure 3: Estimated spectral envelopes of DNA EBV sequence

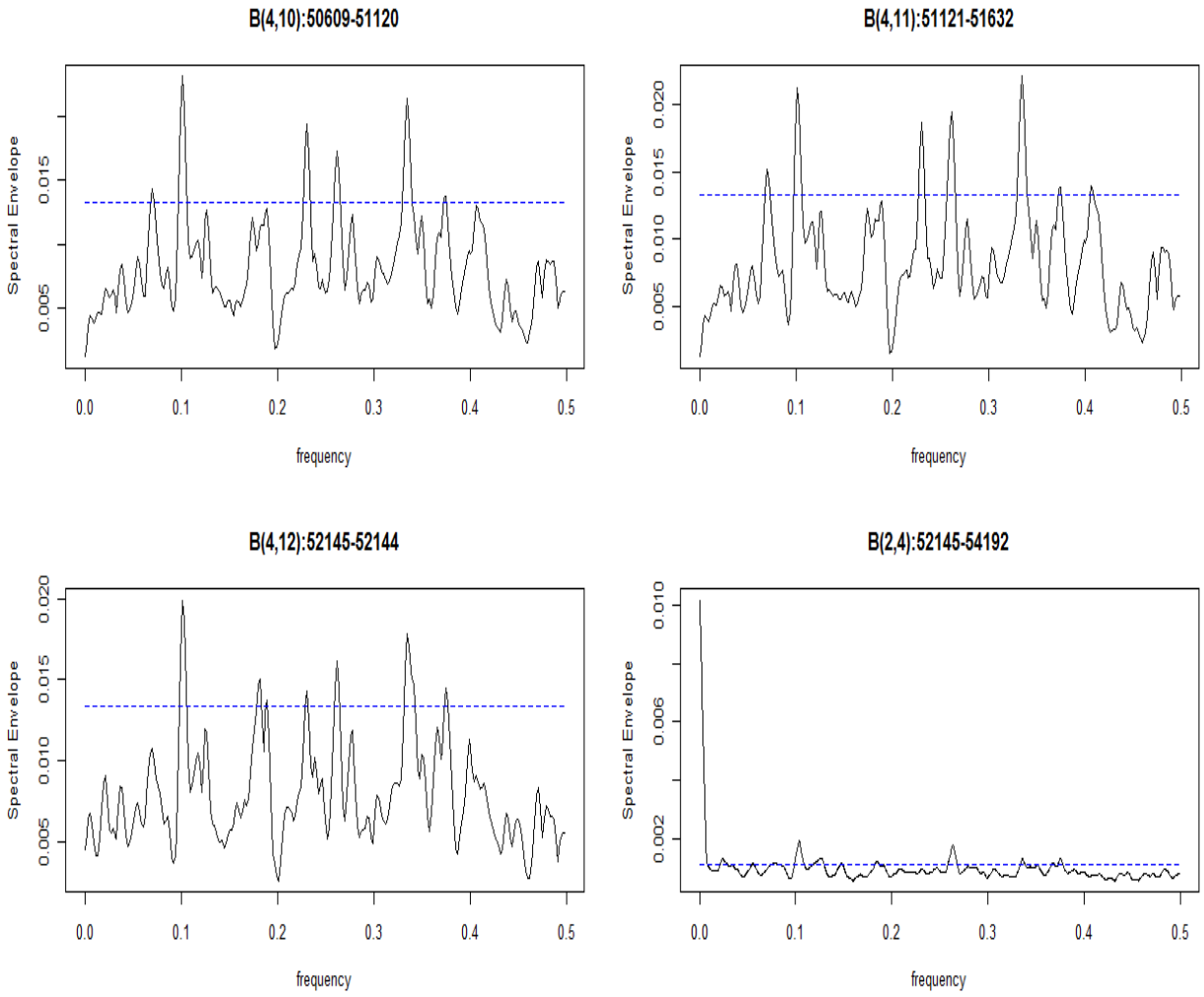

Figure 4: Estimated spectral envelopes of DNA EBV sequence

6.0 CONCLUSION AND FUTURE WORK

In this thesis, we proposed a method for the spectral analysis of nonstationary categorical time series using the spectral envelope. Spectral envelope is a useful tool to fulfill the spectral analysis for the categorical time series because it summarizes information from the spectrum matrix and presents the spectral information understood easily. Moreover, we used the local spectral envelope based on the model of piecewise stationary process. This method provides us the solution of the spectral analysis of nonstationary time series. We used the TBAS(Tree-Based Adaptive Segmentation) segmentation method to estimate the local spectral envelope based on the piecewise stationary process. When we deal with the massive data set such as DNA sequence, this method performs well in practice because it is computationally efficient and fast. In the segmentation procedure, we use the Kullback-Leibler divergence based on the spectral envelope as a distance measure. The asymptotic property of the distance measure in the section 3.1.2 verifies that the classification rule used in our algorithm is adequate for the best segmentation. By using Kullback-Leibler divergence in the segmentation procesure, I can find the best segmentation which can identify coding and noncoding in the DNA sequence. The results of the simulation study and actual data analysis support that this measure can work well in the segmentation and correctly divide the DNA sequence into coding and noncoding sequence. Even though this algorithm can't identify the exact location of a CDS but this method can find the approximate location of many CDS in a DNA sequence.

In this thesis, I apply this method to the piecewise stationary time series because the assumption of piecewise stationary process is proper for the EBV DNA sequence. I would like to extend our algorithm to the other class of nonstationary process and estimate the local spectral envelope. For example, we can consider evolutionary stationary time series. To

develop the evolutionary spectral envelope, we can adapt the model of a locally stationary process defined by Dahlhaus(1997). Also, we can find some fine tunings of methodology and it will be developed in the future; such as other distance measure or classification rule.

APPENDIX

R CODES

A.1 FUNCTION FOR CALCULATION OF SPECTRAL ENVELOPE

```
# R code to calculate the SPECTRAL ENVELOPE for a categorical time series
# *** Must source mvspec.R first ***
# The data set is a column of INTEGERS representing the categories

specenv<-function(u,s)
{
x=model.matrix(~u-1)[,1:3]
# makes indicator matrix
#x=x[1:1000,]                  # select subsequence if desired
Var=var(x)                     # var-cov matrix
#
source("C:/Users/Hyewook Jeong/Desktop/code/function/mvspec.R")
n=length(u)
s=n/(2^6)
xspec=mvspec(x, spans=c(s,s))  # must source mvspec.R
fxxr= Re(xspec$fxx)            # fxxr is real(fxx)
#
ev=eigen(Var)
Q=matrix(0,3,3)                # Q is Var^-1/2
 for (i in 1:3){
 Q=(1/sqrt(ev$values[i]))*ev$vectors[,i]%*%t(ev$vectors[,i]) + Q
 }
#
num=xspec$n.used
nfreq=length(xspec$freq)
specenv=matrix(0,nfreq,1)
beta=matrix(0,nfreq,3)
 for (k in 1:nfreq){
  ev = eigen(2*Q%*%fxxr[,,k]%*%Q/num)
  specenv[k]=ev$values[1]
  b=Q%*%ev$vectors[,1]
  beta[k,]=b/sqrt(sum(b^2))
```

```
  }
frequency=(0:(nfreq-1))/num
m=xspec$kernel$m
etainv=sqrt(sum(xspec$kernel[-m:m]^2))
thresh=(2/num)*exp(qnorm(0.999)*etainv)*matrix(1,nfreq,1)
l=log2(8192/num)
#thresh=(1/(2^(13-l)))*matrix(1,nfreq,1)lines(frequency,thresh, lty="dashed", col="blue")
output = cbind(frequency, specenv, beta,thresh)
list(freq=frequency, l=l,spec=specenv, beta=beta,thresh=thresh,etainv=etainv)
}
```

A.1.1 Subfunction-mvspec.R

```
mvspec <- function(x, spans = NULL, kernel = NULL, taper = 0, pad = 0,
    fast = TRUE, demean = TRUE, detrend = FALSE, plot = FALSE,
    na.action = na.fail,...)
{
    series <- deparse(substitute(x))
    x <- na.action(as.ts(x))
    xfreq <- frequency(x)
    x <- as.matrix(x)
    N <- N0 <- nrow(x)
    nser <- ncol(x)
    if (!is.null(spans))
        kernel <- {
            if (is.tskernel(spans))
                spans
            else kernel("modified.daniell", spans%/%2)
        }
    if (!is.null(kernel) && !is.tskernel(kernel))
        stop("must specify 'spans' or a valid kernel")
    if (detrend) {
        t <- 1:N - (N + 1)/2
        sumt2 <- N * (N^2 - 1)/12
        for (i in 1:ncol(x)) x[, i] <- x[, i] - mean(x[, i]) -
            sum(x[, i] * t) * t/sumt2
    }
    else if (demean) {
        x <- sweep(x, 2, colMeans(x))
    }
    x <- spec.taper(x, taper)
    u2 <- (1 - (5/8) * taper * 2)
    u4 <- (1 - (93/128) * taper * 2)
    if (pad > 0) {
        x <- rbind(x, matrix(0, nrow = N * pad, ncol = ncol(x)))
        N <- nrow(x)
    }
    NewN <- if (fast)
        nextn(N)
    else N
    x <- rbind(x, matrix(0, nrow = (NewN - N), ncol = ncol(x)))
    N <- nrow(x)
    Nspec <- floor(N/2)
```

```
    freq <- seq(from = xfreq/N, by = xfreq/N, length = Nspec)
    xfft <- mvfft(x)
    pgram <- array(NA, dim = c(N, ncol(x), ncol(x)))
    for (i in 1:ncol(x)) {
        for (j in 1:ncol(x)) {
            pgram[, i, j] <- xfft[, i] * Conj(xfft[, j])/(N0 *
                xfreq)
            pgram[1, i, j] <- 0.5 * (pgram[2, i, j] + pgram[N,
                i, j])
        }
    }
    if (!is.null(kernel)) {
        for (i in 1:ncol(x)) for (j in 1:ncol(x)) pgram[, i,
            j] <- kernapply(pgram[, i, j], kernel, circular = TRUE)
        df <- df.kernel(kernel)
        bandwidth <- bandwidth.kernel(kernel)
    }
    else {
        df <- 2
        bandwidth <- sqrt(1/12)
    }
    df <- df/(u4/u2^2)
    df <- df * (N0/N)
    bandwidth <- bandwidth * xfreq/N
    pgram <- pgram[2:(Nspec + 1), , , drop = FALSE]
    spec <- matrix(NA, nrow = Nspec, ncol = nser)
    for (i in 1:nser) spec[, i] <- Re(pgram[1:Nspec, i, i])
    if (nser == 1) {
        coh <- phase <- NULL
    }
    else {
        coh <- phase <- matrix(NA, nrow = Nspec, ncol = nser *
            (nser - 1)/2)
        for (i in 1:(nser - 1)) {
            for (j in (i + 1):nser) {
                coh[, i + (j - 1) * (j - 2)/2] <- Mod(pgram[,
                  i, j])^2/(spec[, i] * spec[, j])
                phase[, i + (j - 1) * (j - 2)/2] <- Arg(pgram[,
                  i, j])
            }
        }
    }
    for (i in 1:nser) spec[, i] <- spec[, i]/u2
    spec <- drop(spec)
#========================
    fxx=array(NA, dim=c(nser,nser,Nspec))
    for (i in 1:nser){
 for (j in 1:nser){
   for (k in 1:Nspec){
fxx[i,j,k]=pgram[k,i,j]
    }
       }
  }
#========================
```

```
    spg.out <- list(freq = freq, spec = spec, coh = coh, phase = phase,
        kernel = kernel, df = df, bandwidth = bandwidth, n.used = N,
        fxx=fxx,
        orig.n = N0, series = series, snames = colnames(x), method = ifelse(!is.null(kernel),
            "Smoothed Periodogram", "Raw Periodogram"), taper = taper,
        pad = pad, detrend = detrend, demean = demean)
    class(spg.out) <- "spec"
    if (plot) {
        plot(spg.out, ...)
        return(invisible(spg.out))
    }
    else return(spg.out)
}
```

A.2 R-CODE FOR CALCULATING DISTANCES BETWEEN SUBSEQUENCE AND FINDING BEST SEGMENTATION

A.2.1 Best segmentation of EBV DNA sequence

```
u=factor(scan("C:/Users/Hyewook Jeong/Desktop/code/ebv2.dat"))
x<-u[46001:54192]
T=length(x)

source("C:/Users/Hyewook Jeong/Desktop/code/function/specenv_ebv.R")
D=matrix(c(rep(0,5*2^4)),nrow=5)

for (i in 1:5){
  j=6-i
  blksz=T/(2^j)
  span=blksz/128
  #par(mfrow=n2mfrow(2^j))
  for (l in (seq(1,2^j,by=2))){
       ind1=(blksz*(l-1)+1)
       ind2=(blksz*l)
       ind3=ind2+1
       ind4=ind2+blksz
       z1=x[ind1:ind2]
       z2=x[ind3:ind4]
       spec1=specenv(z1)$spec*(specenv(z1)$spec>specenv(z1)$thresh)
       spec2=specenv(z2)$spec*(specenv(z2)$spec>specenv(z2)$thresh)


       ## Kullback_leibler divergence
       rel_spec1=spec1/sum(spec1,na.rm=TRUE)
       rel_spec2=spec2/sum(spec2,na.rm=TRUE)
       I=(rel_spec1-rel_spec2)*log((rel_spec1+0.0000000001)/(rel_spec2+0.0000000001))
       m=(l+1)/2
       D[i,m]= sum(I,na.rm=TRUE)
 }
}
```

A.2.2 Best segmentation of simulated data

```
t1<-1:512
x1<-2*cos(2*pi*t1/10)+cos(2*pi*t1/3)+0.3*rnorm(512,0,1)
t2<-3073:4096
x2<-cos(2*pi*t2/3)+.01*rnorm(1024,0,1)
s1<-as.numeric(x1<=quantile(x1,.25))+2*as.numeric(x1>quantile(x1,.25) & x1<=quantile(x1,.5))
+   +3*as.numeric(x1>quantile(x1,.5)& x1<=quantile(x1,.75))+4*as.numeric(x1>quantile(x1,.75))
s2<-as.numeric(x2<=quantile(x2,.25))+2*as.numeric(x2>quantile(x2,.25) & x2<=quantile(x2,.5))
+   +3*as.numeric(x2>quantile(x2,.5)& x2<=quantile(x2,.75))+4*as.numeric(x2>quantile(x2,.75))
y1<-3*rnorm(512,0,1)
y2<-rnorm(1024,0,1)
y3<-2*rnorm(1024,0,1)
n1<-randomNumbers(512, 1, 4,1)
```

```
n2<-randomNumbers(1024, 1, 4,1)
n3<-randomNumbers(1024, 1, 4,1)
x<-append(append(append(append(s1,n1),n2),n3),s2)
t<-1:4096
T=length(t)
x<-as.factor(x)

source("C:/Users/Hyewook Jeong/Desktop/code/function/specenv.R")
D=matrix(c(rep(0,5*2^4)),nrow=5)

for (i in 1:5){
  j=6-i
  blksz=T/(2^j)
  for (l in (seq(1,2^j,by=2))){
       ind1=(blksz*(l-1)+1)
       ind2=(blksz*l)
       ind3=ind2+1
       ind4=ind2+blksz
       z1=x[ind1:ind2]
       z2=x[ind3:ind4]
       spec1=specenv(z1)$spec
       spec2=specenv(z2)$spec

     ## Kullback_leibler divergence
       rel_spec1=spec1/sum(spec1,na.rm=TRUE)
       rel_spec2=spec2/sum(spec2,na.rm=TRUE)
       I=(rel_spec1-rel_spec2)*log((rel_spec1+0.00000000001)/(rel_spec2+0.00000000001))
       m=(l+1)/2
       D[i,m]= sum(I,na.rm=TRUE)
 }
}
```

BIBLIOGRAPHY

[1] S. Adak. Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association*, 93(444):1488–1489, 1998. American Statistical Association.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. wadsworth and brooks. *Monterey, Calif*, 1984.

[3] R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *Information Theory, IEEE Transactions on*, 38(2 Part 2):713–718, 1992.

[4] R. Dahlhaus. Fitting time series models to nonstationary processes. *ANNALS OF STATISTICS*, 25:1–37, 1997.

[5] R. Dahlhaus. A likelihood approximation for locally stationary processes. *ANNALS OF STATISTICS*, 28(6):1762–1794, 2000.

[6] H. Y. Huang, H. Ombao, and D. S. Stoffer. Discrimination and classification of nonstationary time series using the slex model. *JOURNAL-AMERICAN STATISTICAL ASSOCIATION*, 99:763–774, 2004.

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[8] H. Ombao, J. Raz, R. von Sachs, and W. Guo. The slex model of a non-stationary random process. *Annals of the Institute of Statistical Mathematics*, 54(1):171–200, 2002.

[9] H. C. Ombao, J. A. Raz, R. von Sachs, and B. A. Malow. Automatic statistical analysis of bivariate nonstationary time series. *parameters*, 1(1):1, 2001.

[10] M. B. Priestley. Evolutionary spectra and non-stationary processes. *J. Roy. Statist. Soc. Ser. B*, 27(2):204?37, 1965.

[11] D. S. Stoffer, H. C. Ombao, and D. 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.

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