Resenhas IME-USP 2000, Vol. 4, No. 4, 363 - 381.

Localized Spectral Envelope

David S. Stoffer and Hernando Ombao

Abstract: The concept of the spectral envelope was introduced as a statistical basis for the frequency domain analysis and scaling of qualitative-valued time series. A major focus of this research was the analysis of DNA sequences. A common problem in analyzing long DNA sequence data is to identify coding sequences that are dispersed throughout the DNA and separated by regions of non-coding. Even within short subsequences of DNA, one encounters local behavior. To address this problem of local behavior in categorical-valued time series, we explore using the spectral envelope in conjunction with the dyadic tree-based adaptive segmentation method for analyzing locally stationary processes.

Key words: Spectral envelope, dyadic-tree based methods, adaptive segmentation, categorical-valued time series, DNA, locally stationary processes, time varying spectrum, optimal scaling, Fourier analysis, signal detection, latent roots and vectors, principal components.

1. Introduction

The concept of spectral envelope for the spectral analysis and scaling of categorical time series was first introduced in Stoffer et al (1993a). Subsequently, Stoffer et al (1993b) explored the utility of the methodology for analyzing long DNA sequences. In that article, it was noted that there may be local behavior within a single gene (coding sequence). In this article, we combine dyadic tree-based adaptive segmentation (TBAS) and spectral envelope methodologies to develop an evolutionary spectral envelope.

Before discussing the spectral envelope and adaptive segmentation methodologies, we focus on the special problems encountered when analyzing a categorical-valued time series. The spectral envelope was motivated by collaborations with researchers who collected categorical-valued time series with an interest in the cyclic behavior of the data. For example Table 1 shows the per minute sleep-state of an infant taken from a study on the effects of prenatal exposure to alcohol. Details can be found in Stoffer et al (1988), but briefly, an electro-encephalographic (EEG) sleep recording of approximately two hours is obtained on a full term infant 24 to 36 hours after birth, and the recording is scored by a pediatric neurologist for sleep state. Sleep state is categorized, per minute, into one of six possible states: `qt`: quiet sleep - trace alternant, `qh`: quiet sleep - high voltage, `tr`: transitional sleep, `al`: active sleep - low voltage, `ah`: active sleep - high voltage, and `aw`: awake. This particular infant was never awake during the study.

It is not too difficult to notice a pattern in the data if one concentrates on active versus quiet sleep. It would be difficult, however, to try to assess patterns

Table 1: Infant EEG Sleep States (per minute)
(read down and across)

ah	qt	qt	al	tr	qt	al	ah
ah	qt	qt	ah	tr	qt	al	ah
ah	qt	tr	ah	tr	qt	al	ah
ah	qt	al	ah	qh	qt	al	ah
ah	qt	al	ah	qh	qt	al	ah
ah	tr	al	ah	qt	qt	al	ah
ah	qt	al	ah	qt	qt	al	ah
ah	qt	al	ah	qt	qt	al	ah
tr	qt	tr	tr	qt	qt	al	tr
ah	qt	ah	tr	qt	tr	al	
tr	qt	al	ah	qt	al	al	
ah	qt	al	ah	qt	al	al	
ah	qt	al	ah	qt	al	al	
qh	qt	al	ah	qt	al	ah	

in a longer sequence, or if there were more categories, without some graphical aid. One simple method would be to *scale* the data, that is, *assign numerical values to the categories* according to some optimality criterion, and then draw a time plot of the scaled series.

The material on scaling time series is rather sparse and we do not know of any particular references besides those already mentioned. The basic idea, however, has been extensively used for the analysis of contingency tables and regression with qualitative variables; these come under a number of different titles such as *dual scaling*, for example, Nishisato (1980) and *correspondence analysis*, for example, Greenacre (1984). These and related topics are also discussed in Breiman and Friedman (1985) where the focus is on obtaining optimal transformations, numerically, in various situations. For a recent survey, see Michailidis and De Leeuw (1998).

Because the sleep-states have an order, one obvious scaling is

$$\texttt{qt} = 1 \quad \texttt{qh} = 2 \quad \texttt{tr} = 3 \quad \texttt{al} = 4 \quad \texttt{ah} = 5 \quad \texttt{aw} = 6, \tag{1}$$

and Figure 1 shows the time plot using this scaling. Another interesting scaling might be to combine the quiet states and the active states:

$$\texttt{qt} = 1 \quad \texttt{qh} = 1 \quad \texttt{tr} = 2 \quad \texttt{al} = 3 \quad \texttt{ah} = 3 \quad \texttt{aw} = 4. \tag{2}$$

The time plot using (2) would be similar to Figure 1 as far as the cyclic (in and out of quiet sleep) behavior of this infant's sleep pattern. Figure 2 shows the periodogram of the sleep data using the scaling in (1). Note that there is a large peak at the frequency corresponding to 1 cycle every 60 minutes. As one might

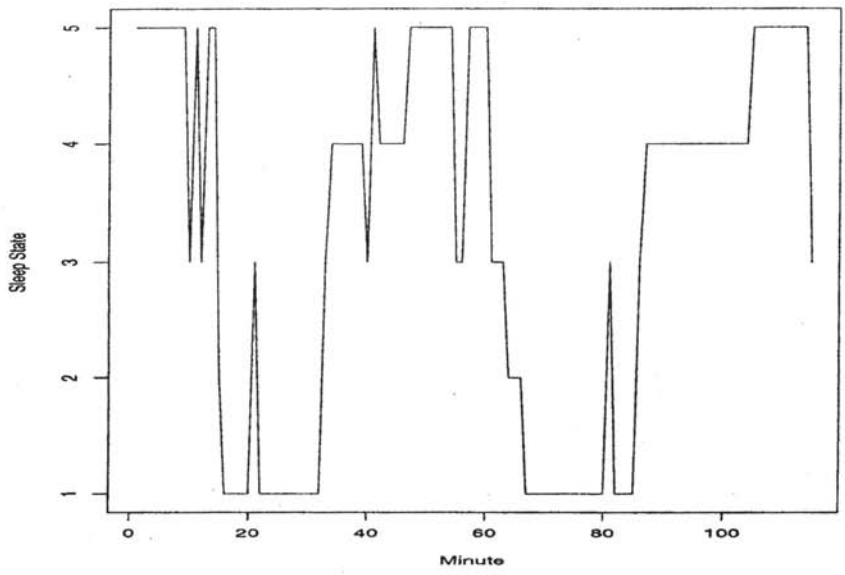

Figure 1: Time plot of the EEG sleep state data in Table 1 using the scaling in (1).

imagine, the general appearance of the periodogram using the scaling (2) (not shown) is similar to Figure 2. Most of us would feel comfortable with this analysis even though we made an arbitrary and ad hoc choice about the particular scaling. It is evident from the data (without any scaling) that if the interest is in infant sleep cycling, this particular sleep study indicates that an infant cycles between active and quiet sleep at a rate of about one cycle per hour.

The intuition used in the previous example is lost when one considers a long DNA sequence. Briefly, a DNA strand can be viewed as a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five carbon sugar, and a phosphate group. There are four different bases that can be grouped by size, the pyrimidines, thymine (T) and cytosine (C), and the purines, adenine (A) and guanine (G). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the $5'$ carbon of one sugar linked to the $3'$ carbon of the next, giving the string direction. DNA molecules occur naturally as a double helix composed of polynucleotide strands with the bases facing inwards. The two strands are complementary, so it is sufficient to represent a DNA molecule by a sequence of bases on a single strand. Thus, a strand of DNA can be represented as a sequence of letters, termed base pairs (bp), from the finite alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The order of the nucleotides contains the genetic information specific to the organism. Expression of information stored in these molecules is a

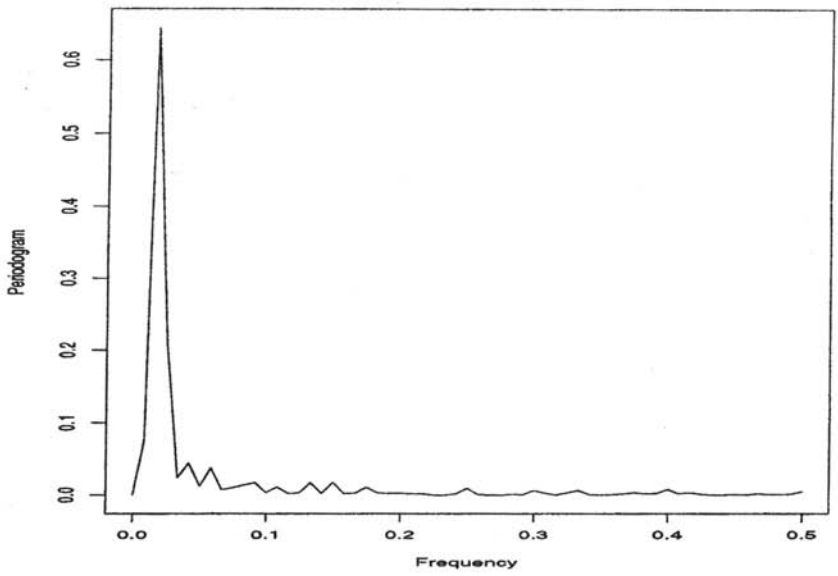

Figure 2: Periodogram of the EEG sleep state data in Table 1 based on the scaling in (1). The peak corresponds to a frequency of approximately one cycle every 60 minutes.

complex multistage process. One important task is to translate the information stored in the protein-coding sequences (CDS) of the DNA. A common problem in analyzing long DNA sequence data is in identifying CDS that are dispersed throughout the sequence and separated by regions of non-coding (which makes up most of the DNA). Table 2 shows part of the Epstein-Barr virus (EBV) DNA sequence. The entire EBV DNA sequence consists of approximately 172,000 bp.

One could try scaling according to the pyrimidine-purine alphabet, that is $\mathtt{A} = \mathtt{G} = 0$ and $\mathtt{C} = \mathtt{T} = 1$, but this is not necessarily of interest for every CDS of EBV. There are numerous possible alphabets of interest, for example, one might focus on the strong-weak hydrogen bonding alphabet $\mathtt{C} = \mathtt{G} = 0$ and $\mathtt{A} = \mathtt{T} = 1$. While model calculations as well as experimental data strongly agree that some kind of periodic signal exists in certain DNA sequences, there is a large disagreement about the exact type of periodicity. In addition, there is disagreement about which nucleotide alphabets are involved in the signals (for example, compare Ioshikhes et al, 1992 with Satchwell et al, 1986).

If we consider the naive approach of arbitrarily assigning numerical values (scales) to the categories and then proceeding with a spectral analysis, the result will depend on the particular assignment of numerical values. For example, consider the artificial sequence ACGTACGTACGT... . Then, setting $\mathtt{A} = \mathtt{G} = 0$ and

Table 2: Part of the Epstein-Barr Virus DNA Sequence (read across and down)

AGAATTCGTC	TTGCTCTATT	CACCCTTACT	TTTCTTCTTG	CCCGTTCTCT	TTCTTAGTAT
GAATCCAGTA	TGCCTGCCTG	TAATTGTTGC	GCCCTACCTC	TTTTGGCTGG	CGGCTATTGC
CGCCTCGTGT	TTCACGGCCT	CAGTTAGTAC	CGTTGTGACC	GCCACCGGCT	TGGCCCTCTC
ACTTCTACTC	TTGGCAGCAG	TGGCCAGCTC	ATATGCCGCT	GCACAAAGGA	AACTGCTGAC
ACCGGTGACA	GTGCTTACTG	CGGTTGTCAC	TTGTGAGTAC	ACACGCACCA	TTTACAATGC
ATGATGTTCG	TGAGATTGAT	CTGTCTCTAA	CAGTTCACTT	CCTCTGCTTT	TCTCCTCAGT
CTTTGCAATT	TGCCTAACAT	GGAGGATTGA	GGACCCACCT	TTTAATTCTC	TTCTGTTTGC
ATTGCTGGCC	GCAGCTGGCG	GACTACAAGG	CATTTACGGT	TAGTGTGCCT	CTGTTATGAA
ATGCAGGTTT	GACTTCATAT	GTATGCCTTG	GCATGACGTC	AACTTTACTT	TTATTTCAGT
TCTGGTGATG	CTTGTGCTCC	TGATACTAGC	GTACAGAAGG	AGATGGCGCC	GTTTGACTGT
TTGTGGCGGC	ATCATGTTTT	TGGCATGTGT	ACTTGTCCTC	ATCGTCGACG	CTGTTTTGCA
GCTGAGTCCC	CTCCTTGGAG	CTGTAACTGT	GGTTTCCATG	ACGCTGCTGC	TACTGGCTTT
CGTCCTCTGG	CTCTCTTCGC	CAGGGGGCCT	AGGTACTCTT	GGTGCAGCCC	TTTTAACATT
GGCAGCAGGT	AAGCCACACG	TGTGACATTG	CTTGCCTTTT	TGCCACATGT	TTTCTGGACA
CAGGACTAAC	CATGCCATCT	CTGATTATAG	CTCTGGCACT	GCTAGCGTCA	CTGATTTTGG
GCACACTTAA	CTTGACTACA	ATGTTCCTTC	TCATGCTCCT	ATGGACACTT	GGTAAGTTTT
CCCTTCCTTT	AACTCATTAC	TTGTTCTTTT	GTAATCGCAG	CTCTAACTTG	GCATCTCTTT
TACAGTGGTT	CTCCTGATTT	GCTCTTCGTG	CTCTTCATGT	CCACTGAGCA	AGATCCTTCT
GGCACGACTG	TTCCTATATG	CTCTCGCACT	CTTGTTGCTA	GCCTCCGCGC	TAATCGCTGG
TGGCAGTATT	TTGCAAACAA	ACTTCAAGAG	TTTAAGCAGC	ACTGAATTTA	TACCCAGTGA

C = T = 1, yields the numerical sequence 010101010101..., or one cycle every two base pairs ($\omega = 1/2$). Another interesting scaling is A = 1, C = 2, G = 3, and T = 4, which results in the sequence 123412341234..., or one cycle every four bp ($\omega = 1/4$). In this example, both scalings, {A, C, G, T} = {0, 1, 0, 1} and {A, C, G, T} = {1, 2, 3, 4}, are interesting and bring out different properties of the sequence. It should be clear that one does not want to focus on only one scaling. Instead, the focus should be on finding scalings that bring out all of the interesting features in the data. Moreover, because of heterogeneity (see e.g. Karlin and Macken, 1991), it may be the case that if one scaling works well in one region of a DNA sequence that same scaling may work poorly in another region. Rather than choose values arbitrarily, the spectral envelope approach selects scales that help emphasize any periodic feature that exists in a categorical time series of virtually any length in a quick and automated fashion.

Although it is well known that DNA is heterogeneous, in Stoffer et al (1993b) we found that heterogeneities can exist within short subsequences of a single gene. In this article, we describe a methodology that will automatically divide a DNA sequence into smaller stationary segments and then extract the pertinent information from these segments. Our methodology will be an adaptation of the TBAS method given in Adak (1998) and Ombao et al (1999). This methodology was specifically developed for real-valued non-stationary time series and has been successfully applied to a bivariate EEG data set recorded during an epileptic seizure. In the next section we will see that categorical time series (and hence DNA sequences) are real-valued multivariate time series of a special nature. In addition,

we will see that the spectral envelope is a type of principal component analysis for multivariate time series. Consequently, in the remaining sections, we adapt the dyadic TBAS methodology to the principal component analysis of multivariate time series with special attention to categorical time series.

2. The Spectral Envelope for Categorical Time Series

As a general description, the spectral envelope is a frequency based, principal components technique applied to a multivariate time series. In this section we will focus on the basic concept and its use in the analysis of categorical time series. Technical details can be found in Stoffer et al (1993a).

In establishing the spectral envelope for categorical time series, the basic question of how to efficiently discover periodic components in categorical time series was addressed. This was accomplished via nonparametric spectral analysis as follows. Let $X_t, t = 0, \pm 1, \pm 2, ...$, be a categorical-valued time series with finite state-space $\mathcal{C} = \{c_1, c_2, ..., c_k\}$. Assume that X_t is stationary and $p_j = \text{pr}\{X_t = c_j\} > 0$ for $j = 1, 2, ..., k$. For $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)' \in \mathbf{R}^k$, 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 β_j, $j = 1, 2, ..., k$. The goal is to find scalings $\boldsymbol{\beta}$ so that the spectral density is in some sense interesting, and to summarize the spectral information by what we called the spectral envelope.

We chose $\boldsymbol{\beta}$ to maximize the power (variance) at each frequency ω, across frequencies $\omega \in [-1/2, 1/2]$, relative to the total power $\sigma^2(\boldsymbol{\beta}) = \text{var}\{X_t(\boldsymbol{\beta})\}$. That is, we chose $\boldsymbol{\beta}(\omega)$, at each ω of interest, so that

$$\lambda(\omega) = \sup_{\boldsymbol{\beta}} \left\{ \frac{f(\omega; \boldsymbol{\beta})}{\sigma^2(\boldsymbol{\beta})} \right\}, \tag{1}$$

over all $\boldsymbol{\beta}$ not proportional to $\mathbf{1}_k$, the $k \times 1$ vector of ones. Note that $\lambda(\omega)$ is not defined if $\boldsymbol{\beta} = a\mathbf{1}_k$ for $a \in \mathbf{R}$ because such a scaling corresponds to assigning each category the same value a; in this case $f(\omega; \boldsymbol{\beta}) \equiv 0$ and $\sigma^2(\boldsymbol{\beta}) = 0$. The optimality criterion $\lambda(\omega)$ possesses the desirable property of being invariant under location and scale changes of $\boldsymbol{\beta}$.

As in most scaling problems for categorical data, it was useful to represent the categories in terms of the unit vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_k$, where $\boldsymbol{e}_j$ represents the $k \times 1$ vector with a one in the j-th row, and zeros elsewhere. We then defined a k-dimensional stationary time series $\boldsymbol{Y}_t$ by $\boldsymbol{Y}_t = \boldsymbol{e}_j$ when $X_t = c_j$. The time series $X_t(\boldsymbol{\beta})$ can be obtained from the $\boldsymbol{Y}_t$ time series by the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_t$. Assume that the vector process $\boldsymbol{Y}_t$ has a continuous spectral density denoted by $f_{\boldsymbol{Y}}(\omega)$. For each ω, $f_{\boldsymbol{Y}}(\omega)$ is, of course, a $k \times k$ complex-valued Hermitian matrix. Note that the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_t$ implies that $f_{\boldsymbol{Y}}(\omega; \boldsymbol{\beta}) = \boldsymbol{\beta}' f_{\boldsymbol{Y}}(\omega)\boldsymbol{\beta} = \boldsymbol{\beta}' f_{\boldsymbol{Y}}^{re}(\omega)\boldsymbol{\beta}$, where $f_{\boldsymbol{Y}}^{re}(\omega)$ denotes the real part of $f_{\boldsymbol{Y}}(\omega)$. The optimality criterion

can thus be expressed as

$$\lambda(\omega) = \sup_{\beta} \left\{ \frac{\beta' f_Y^{re}(\omega)\beta}{\beta' V \beta} \right\} \tag{2}$$

where V is the variance-covariance matrix of $\boldsymbol{Y}_t$. The resulting scaling $\beta(\omega)$ is called the optimal scaling.

The $\boldsymbol{Y}_t$ process is a multivariate point process, and any particular component of $\boldsymbol{Y}_t$ is the individual point process for the corresponding state (for example, the first component of $\boldsymbol{Y}_t$ indicates whether or not the process is in state c_1 at time t). For any fixed t, $\boldsymbol{Y}_t$ represents a single observation from a simple multinomial sampling scheme. It readily follows that $V = D - \boldsymbol{p}\boldsymbol{p}'$, where $\boldsymbol{p} = (p_1, ..., p_k)'$, and D is the $k \times k$ diagonal matrix $D = \text{diag}\{p_1, ..., p_k\}$. Since, by assumption, $p_j > 0$ for $j = 1, 2, ..., k$, it follows that $\text{rank}(V) = k - 1$ with the null space of V being spanned by $\mathbf{1}_k$. For any $k \times (k-1)$ full rank matrix Q whose columns are linearly independent of $\mathbf{1}_k$, $Q'VQ$ is a $(k-1) \times (k-1)$ positive definite symmetric matrix.

With the matrix Q as previously defined, and for $\omega \in [-1/2, 1/2]$, define $\lambda(\omega)$ to be the largest eigenvalue of the determinantal equation

$$|Q' f_Y^{re}(\omega) Q - \lambda Q' V Q| = 0,$$

and let $\boldsymbol{b}(\omega) \in \mathbf{R}^{k-1}$ be any corresponding eigenvector, that is,

$$Q' f_Y^{re}(\omega) Q \boldsymbol{b}(\omega) = \lambda(\omega) Q' V Q \boldsymbol{b}(\omega).$$

The eigenvalue $\lambda(\omega) \geq 0$ does not depend on the choice of Q. Although the eigenvector $\boldsymbol{b}(\omega)$ depends on the particular choice of Q, the equivalence class of scalings associated with $\boldsymbol{\beta}(\omega) = Q\boldsymbol{b}(\omega)$ does not depend on Q. A convenient choice of Q is $Q = [I_{k-1} \mid \mathbf{0}\,]'$, where I_{k-1} is the $(k-1) \times (k-1)$ identity matrix and $\mathbf{0}$ is the $(k-1) \times 1$ vector of zeros . For this choice, $Q' f_Y^{re}(\omega) Q$ and $Q'VQ$ are the upper $(k-1) \times (k-1)$ blocks of $f_Y^{re}(\omega)$ and V, respectively. This choice corresponds to setting the last component of $\boldsymbol{\beta}(\omega)$ to zero.

The value $\lambda(\omega)$ itself has a useful interpretation; specifically, $\lambda(\omega)d\omega$ represents the largest proportion of the total power that can be attributed to the frequencies within a $d\omega$ neighborhood of ω for any particular scaled process $X_t(\boldsymbol{\beta})$, with the maximum being achieved by the scaling $\boldsymbol{\beta}(\omega)$. Because of its central role, $\lambda(\omega)$ was defined to be the *spectral envelope* of a stationary categorical time series.

The name spectral envelope is appropriate since $\lambda(\omega)$ envelopes the standardized spectrum of any scaled process. That is, given any $\boldsymbol{\beta}$ normalized so that $X_t(\boldsymbol{\beta})$ has total power one, $f(\omega; \boldsymbol{\beta}) \leq \lambda(\omega)$ with equality if and only if $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\beta}(\omega)$.

Although the law of the process $X_t(\boldsymbol{\beta})$ for any one-to-one scaling $\boldsymbol{\beta}$ completely determines the law of the categorical process X_t, information is lost when one restricts attention to the spectrum of $X_t(\boldsymbol{\beta})$. Less information is lost when one considers the spectrum of $\boldsymbol{Y}_t$. Dealing directly with the spectral density $f_Y(\omega)$

itself is somewhat cumbersome since it is a function into the set of complex Hermitian matrices. Alternatively, one can view the spectral envelope as an easily understood, parsimonious tool for exploring the periodic nature of a categorical time series with a minimal loss of information.

If we observe a finite realization of the stationary categorical time series X_t, or equivalently, the multinomial point process $\boldsymbol{Y}_t$, $t = 1, ..., T$, the theory for estimating the spectral density of a multivariate, real-valued time series is well established (e.g. Brillinger, 1975 or Hannan, 1970) and can be applied to estimating $f_Y(\omega)$, the spectral density of $\boldsymbol{Y}_t$. Given an estimate $\widehat{f}_Y(\omega)$ of $f_Y(\omega)$, estimates $\widehat{\lambda}(\omega)$ and $\widehat{\boldsymbol{\beta}}(\omega)$ of the spectral envelope, $\lambda(\omega)$, and the corresponding scalings, $\boldsymbol{\beta}(\omega)$, can then be obtained. Details on estimation and inference for the sample spectral envelope and the optimal scalings can be found in Stoffer et al (1993a), but the main result of that paper is as follows. If $\widehat{f}_Y(\omega)$ is a consistent spectral estimator and if for each $j = 1, ..., J$, the largest root of $f_Y^{re}(\omega_j)$ is distinct, then

$$\left\{ \eta_T \left[\widehat{\lambda}(\omega_j) - \lambda(\omega_j) \right] \Big/ \lambda(\omega_j), \;\; \eta_T \left[\widehat{\boldsymbol{\beta}}(\omega_j) - \boldsymbol{\beta}(\omega_j) \right]; \;\; j = 1, ..., J \right\} \tag{3}$$

converges ($T \to \infty$) jointly in distribution to independent zero-mean normal distributions, the first of which is standard normal. The term η_T in (3) depends on the type of estimator being used. For example, if the spectral estimate is obtained by a simple average of $2M_T + 1$ periodograms around a central value, then $\eta_T^2 = (2M_T + 1)$.[1] Based on these results, asymptotic normal confidence intervals and tests for $\lambda(\omega)$ can be readily constructed. Similarly, for $\boldsymbol{\beta}(\omega)$, asymptotic confidence ellipsoids and chi-square tests can be constructed; details can be found in Stoffer et al (1993a, Theorems 3.1 – 3.3).

Searching for peaks in the spectral envelope estimate can be aided using the following approximations. Using a first order Taylor expansion we have

$$\log \widehat{\lambda}(\omega) \approx \log \lambda(\omega) + \frac{\widehat{\lambda}(\omega) - \lambda(\omega)}{\lambda(\omega)}, \tag{4}$$

so that $\eta_T[\log \widehat{\lambda}(\omega) - \log \lambda(\omega)]$ is approximately standard normal under the conditions for which (3) is true. It also follows that $E[\log \widehat{\lambda}(\omega)] \approx \log \lambda(\omega)$ and $\text{var}[\log \widehat{\lambda}(\omega)] \approx \eta_T^{-2}$. If there is no signal present in a sequence of length T, we expect $\lambda(j/T) \approx 2/T$ for $1 < j < T/2$, and hence approximately $(1 - \alpha) \times 100\%$ of the time, $\log \widehat{\lambda}(\omega)$ will be less than $\log(2/T) + (z_\alpha/\eta_T)$ where z_α is the $(1 - \alpha)$ upper tail cutoff of the standard normal distribution. Exponentiating, the α critical value for $\widehat{\lambda}(\omega)$ becomes $(2/T)\exp(z_\alpha/\eta_T)$. From our experience, thresholding at very small values of α relative to the sample size works well.

A step-by-step approach to calculate the sample spectral envelope and optimal scalings for DNA sequences, using the nucleotide alphabet, is as follows. For numerical examples, see Stoffer et al (1993a, 1993b, 2000).

[1] We take $M_T \to \infty$ as $T \to \infty$ but with $M_T/T \to 0$.

- Let X_t denote the DNA sequence of interest. Holding the scale for T fixed at zero, form 3×1 vectors $\boldsymbol{Y}_t$:

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

  The scaling vector is $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$, and the scaled process is $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}' \boldsymbol{Y}_t$.

- Calculate the discrete Fourier transform (DFT) of the data,

$$\boldsymbol{d}(\omega_j) = T^{-1/2} \sum_{t=1}^{T} \boldsymbol{Y}_t \exp(-2\pi i t \omega_j),$$

  where $\omega_j = j/T$ for $j = 1, ..., [T/2]$. Note that $\boldsymbol{d}(\omega_j)$ is a 3×1 complex-valued vector. From these values, calculate the 3×3 periodogram matrices, $I_T(\omega_j) = \boldsymbol{d}(\omega_j)\boldsymbol{d}^*(\omega_j)$, and retain only the real part, say $I_T^{re}(\omega_j)$.

- Smooth the periodogram if desired (recommended) to obtain an estimate of the spectral matrix $f^{re}(\omega_j)$, say $\widehat{f}^{re}(\omega_j)$. For example, we could set

$$\widehat{f}^{re}(\omega_j) = \sum_{q=-M}^{M} h_q I_T^{re}(\omega_j + q/T)$$

  where the weights are chosen so that $h_q = h_{-q} > 0$ and $\sum_{q=-M}^{M} h_q = 1$. A simple average corresponds to the case where $h_q = 1/(2M+1)$ for $q = -M, ..., 0, ..., M$.

- Next, calculate the 3×3 sample variance-covariance matrix given by $S = T^{-1} \sum_{t=1}^{T} (\boldsymbol{Y}_t - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_t - \overline{\boldsymbol{Y}})'$, where $\overline{\boldsymbol{Y}}$ is the vector of sample means.

- For each ω_j, determine the largest eigenvalue and the corresponding eigenvector of the matrix $2T^{-1} S^{-1/2} \widehat{f}^{re}(\omega_j) S^{-1/2}$. Note that $S^{1/2}$ is the unique square root matrix of S.

- The sample spectral envelope $\widehat{\lambda}(\omega_j)$ is the eigenvalue obtained in the previous step. If $\boldsymbol{b}(\omega_j)$ denotes the eigenvector obtained in the previous step, the optimal sample scaling is $\widehat{\boldsymbol{\beta}}(\omega_j) = S^{-1/2} \boldsymbol{b}(\omega_j)$; this will result in 3 values, the value corresponding to the 4-th category (T in this case) being held fixed at zero.

3. Tree-Based Adaptive Segmentation for the Spectral Envelope

It is well known that long DNA sequences are heterogeneous. One subsequence (block) may contain genetic information that is unique from other blocks. Other blocks may not contain any genetic coding information at all. Genetic sequences are generally very long and our goal is to develop a fast and efficient method than can search for blocks that contain similar genetic information and to separate these blocks from other blocks that either contain different genetic information or non-coding information (noise).

In this section, we will describe an algorithm for segmenting a DNA sequence. The strategy adopted is to divide the sequence into small blocks and then recombine adjacent blocks whose estimated spectral envelopes are sufficiently close. The basic idea is that adjacent blocks with close spectral envelope estimates give similar genetic information. The main features of the algorithm are: (i) it divides the sequence in a dyadic manner and (ii) it uses a well-defined measure of distance (or similarity) between the genetic coding information contained at two adjacent (common split) blocks. Our method is inspired by the algorithm in Adak (1998).

We now give the algorithm.

- Set the maximum level J. The value of J determines the smallest possible size of the segmented blocks. For a sequence of length T, the smallest blocks have length $T/2^J$. Ideally, the block sizes should be small enough so that one can separate useful genetic information unique to that block from the non-coding material (noise). One should be careful, however, about making the blocks too small. Blocks have to be large enough to give good estimates of the spectral envelope.

- Set the blocks: For $j = 0, \ldots, J$, divide the data sequence into 2^j blocks. Denote $B(\ell, j)$ to be the ℓ-th block on level j, where $\ell = 1, \ldots, 2^j$. The first block on level j is denoted as $B(1, j)$ and the last as $B(2^j, j)$. The "inner" blocks $B(\ell, j)$, (where $\ell = 2, \ldots, 2^j - 1$), consists of the elements of the DNA sequence $\{X_{[(\ell-1)2^j + 1]}, \cdots, X_{[\ell 2^j]}\}$.

- (This step is optional) Tapers: The inner blocks can be extended towards adjacent blocks and the outer blocks (first and last) at each level can be padded with zeroes in order to apply tapers. Tapering can reduce leakage in estimates of the spectrum and the spectral envelope.

- Compute an estimate of the spectral envelope $\widehat{\lambda}_{\ell,j}(\omega_k)$ for each frequency ω_k $(k = 0, \ldots, M_j = T/2^j)$ at each block $B(\ell, j)$ where $j = 0, \ldots, J$, $\ell = 1, \ldots, 2^j$.

- Define the threshold at level j to be α_j. A discussion on significance levels and threshold is given in Equation (4).

- Form the **denoised** estimate, $\widetilde{\lambda}_{\ell,j}(\omega)$ of the spectral envelope by applying a hard threshold on the estimate in Step 4, i.e., $\widehat{\lambda}_{\ell,j}(\omega_k)$. Any value of

$\widehat{\lambda}_{\ell,j}(\omega_k)$ that does not survive the threshold α_j is considered as noise. More formally,

$$\widetilde{\lambda}_{\ell,j}(\omega_k) = \begin{cases} \widehat{\lambda}_{\ell,j}(\omega_k), & \text{if } \widehat{\lambda}_{\ell,j}(\omega_k) > \alpha_j \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

- Compute the peak information function $P_{\ell,j}(\omega_k)$ at block $B(\ell, j)$ as follows. Let $\Omega_{\ell,j,k} = \{\omega_v \text{ where } v = k - s_{\ell,j}, \ldots, v = k + s_{\ell,j}\}$ be a collection of frequencies centered about ω_k with $s_{\ell,j}$ being the span for block ℓ and level j. The peak information function is

$$P_{\ell,j}(\omega_k) = \begin{cases} 1, & \text{if } \widehat{\lambda}_{\ell,j}(\omega_k) = \max_{\omega_v \in \Omega_{\ell,j,k}} \widetilde{\lambda}_{\ell,j}(\omega_v) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

- Compute the distance, D, between two children blocks. For $j = 0, \ldots, J-1$, and $\ell = 1, \ldots 2^j$:

$$D(\ell, j) = \sum_{k=0}^{M_{j+1}/2} |P_{2\ell,j+1}(\omega_k) - P_{2\ell+1,j+1}(\omega_k)| \tag{7}$$

- Marking the blocks for final segmentation. For $j = J-1, \ldots, 0$, and $\ell = 1, \ldots, 2^j$, define $V(\ell, j) = D(\ell, j)$. Mark the blocks $B(\ell, J-1)$ as terminal. If $j < J-1$ and if $V(2\ell, j+1) + V(2\ell+1, j+1) \leq V(\ell, j)$ then mark the block $B(\ell, j)$ as terminal. Otherwise, leave the block $B(\ell, j)$ as unmarked and set $V(\ell, j) = V(2\ell, j+1) + V(2\ell+1, j+1)$.

The **final segmentation** of the DNA sequence is the set of highest marked blocks: $\{B(\ell, j)$ such that $B(\ell, j)$ is marked and its parent block and ancestor blocks are not marked$\}$.

4. Analysis of the EBV DNA Sequence

We applied our algorithm to a data set that is a subseries of the DNA sequence of the Epstein Barr virus. The subseries consists of elements of the DNA sequence with index from 46001 to 54192. This data set has length $T = 8192$. In the implementation, we report the fine tuning parameters that we have chosen.

- We chose the level $J = 5$ so that the smallest blocks have 256 elements.

- We applied a taper and extended the blocks at all levels $j = 0, \ldots, 5$ by 128 on each side. As a result, each of the blocks at level $j = 5$ had $256 + 2 \times 128 = 512$ elements. Note that the effect of the taper at larger blocks diminishes. Tapering is no longer necessary when block sizes are large.

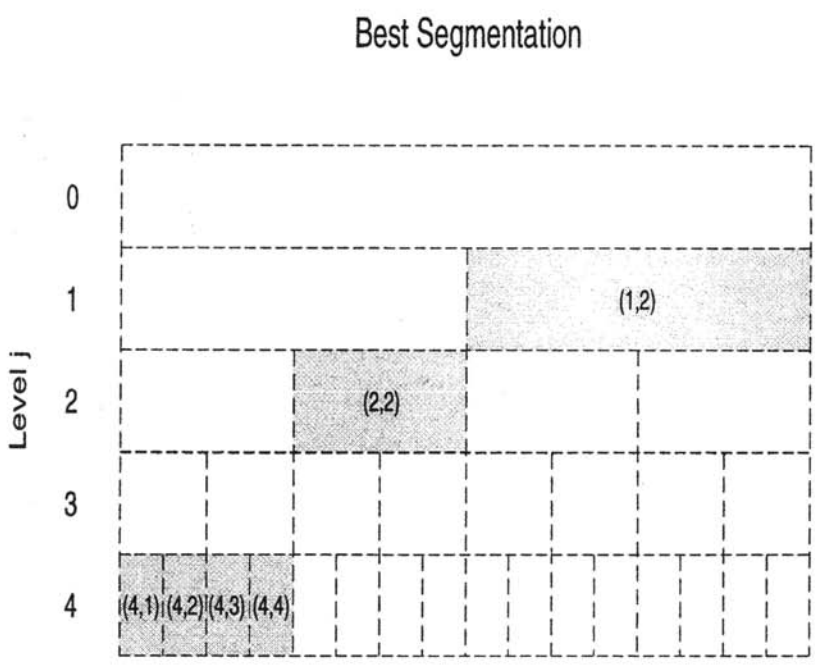

Figure 3: Best Segmentation of the EBV DNA sequence.

- We chose the threshold $\alpha_j = \exp(3/M_j)$. Note that this corresponds to the 0.01 significance level testing in Equation (4).

- We chose the span $s_{\ell,j} = 0.10M_j$ for all blocks ℓ. This ensures that peaks maintain their local behavior.

The segmentation selected by our algorithm is given in Figure 3. The segmentation consists of blocks $B(4,1)$, $B(4,2)$, $B(4,3)$, $B(4,4)$, $B(2,2)$ and $B(1,1)$. The spectral envelopes are given in Figure 4. It is very interesting to note that the DNA sequence of EB virus with index 50097 to 54192 indeed does not contain any coding information. Hence, our algorithm was able to isolate block $B(1,1)$ as containing "noise". Moreover, the DNA sequence with index 48386 to 50032 contains the coding information "EBNA-2". Our algorithm was able to isolate block $B(2,2)$ as containing this coding information. Finally, the DNA sequence with index 46333 to 47481 contains the coding information "BWRF-12". This information is captured by the three blocks $B(2,1)$, $B(2,2)$ and $B(2,3)$. It is not known, at this point, if Block $B(4,4)$ contains non-coding information or some useful genetic information that is waiting to be uncovered.

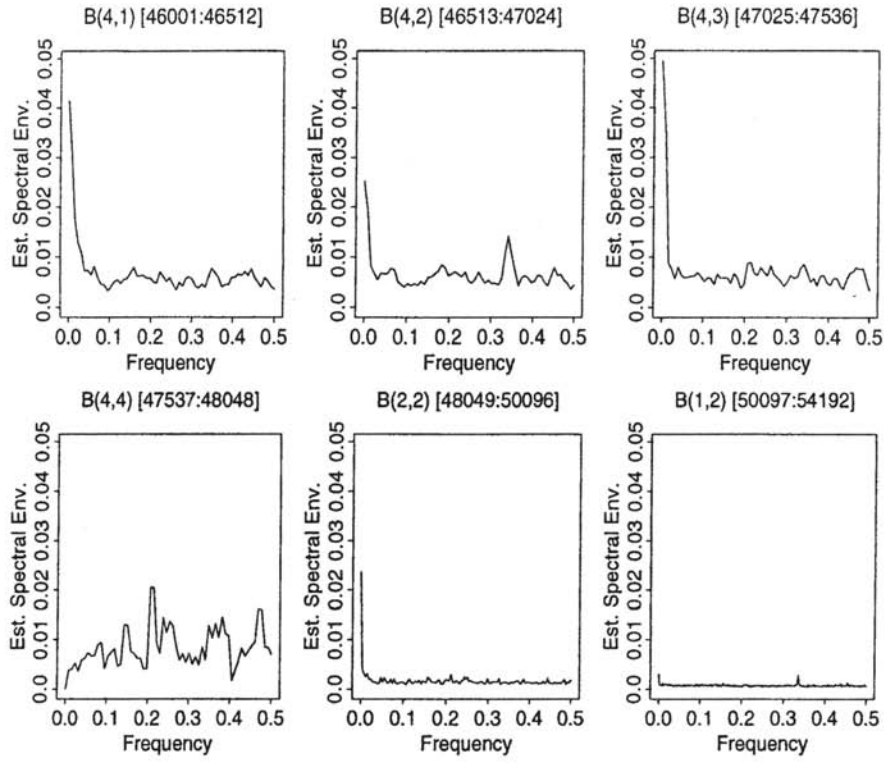

Figure 4: Estimated spectral envelope at each block of the EBV DNA sequence.

Remarks.

- On dyadic segmentation: Dyadic tree-structured based methods are widely used and well accepted in the statistics literature. One example of a dyadic tree-based method is CART (Classification and Regression Trees) of Breiman et al (1984). In the time series literature, we now have well developed methods and theory that are based on dyadic segmentation. See for example Mallat et al (1998), Adak (1998) and Donoho et al (1998) and Ombao et al (1999). The Auto-SLEX method in Ombao et al (1999) was applied successfully to a nonstationary EEGs recorded during an epileptic seizure. The goal was to estimate the time-varying spectra of the EEGS and coherence between the two EEGs. It was clearly demonstrated in Ombao et al (1999) that the Auto-SLEX method, which is dyadic-based, does not suffer even when applied to biological signals that do not necessarily have a dyadic structure. The only condition is that these signals have to be sufficiently long. This is condition easily satisfied by all DNA sequences.

- On the distance: The distance $D(\ell, j)$ counts the number of peaks at occurring different at frequencies between the children blocks. If the peaks at the two children blocks occur at different frequencies then the two children blocks are said to contain different genetic information. The magnitude of the difference in genetic information is captured by the distance $D(\ell, j)$.

- Efficient computation: It is necessary to use computationally efficient methods when analyzing very long time series data sets. Dyadic transforms are useful tools for developing computationally efficient methods. The algorithm presented is efficient because it uses two computationally efficient methods. In computing the estimates of the spectral envelope, we used the Fast Fourier transform. Moreover, the selection of the final segmentation was delivered by Best Basis Algorithm (BBA) of Coifman and Wickerhauser (1992). Wickerhauser (1994) devotes a chapter to BBA and related cost measures.

- Relationship to the Adak algorithm: The Adak (1998) method is useful for estimating the time-varying spectrum of a univariate non-stationary process. Spectral estimates are compared between children blocks and distance measures were proposed. The algorithm presented in this article is for multivariate non-stationary categorical time series. Instead of spectral estimates, we compare the estimated **spectral envelopes** between blocks. Moreover, our distance function is unique from what was developed in Adak (1998). The distance function used in our algorithm is specific for time series data sets whose spectra have power concentrated at a very narrow band of frequencies.

5. Discussion and Conclusion

Fourier analysis of categorical time series has been applied successfully in molecular genetics for quite some time. For example, McLachlan and Stewart (1976) and Eisenberg et al (1984) studied the periodicity in proteins with Fourier analysis. They used predefined scales (or alphabets) and observed the $\omega = \frac{1}{3.6}$ frequency of amphipatic helices. Because predetermination of the scaling is arbitrary and may not be optimal, Cornette et al (1987) reversed the problem and started with a frequency of $\omega_0 = \frac{1}{3.6}$ and proposed a method to establish an 'optimal' scaling at $\omega_0 = \frac{1}{3.6}$. Viari et al (1990) generalized this approach to a systematic calculation of a type of spectral envelope (which they called λ-graphs) and of the corresponding optimal scalings over all fundamental frequencies. While the aforementioned authors dealt exclusively with amino acid sequences, various forms of harmonic analysis have been applied to DNA by, for example, Tavaré and Giddings (1989), and in connection to nucleosome positioning by Satchwell et al (1986) and Bina (1994). The basic technique of the spectral envelope for categorical time series is similar to the methods established in Tavaré and Giddings (1989) and Viari et al (1990), however, there are some differences. In particular, the techniques

differ by the optimality criterion used. Also, the spectral envelope methodology is developed in a statistical (rather than graphical) setting to allow the investigator to distinguish between significant results and those results that can be attributed to chance.

As previously indicated, the spectral envelope methodology could come under the general title of spectral domain principal component analysis of multiple time series. This topic is discussed in detail in Chapter 9 of Brillinger (1975) and there is a connection between Brillinger's work and the spectral envelope. Specifically, the spectral envelope can be viewed as a special case of Brillinger's principal components. In the language of Brillinger (1975, Section 9.3), suppose we want to approximate $\boldsymbol{Y}_t$, a $k \times 1$ stationary time series with mean $\boldsymbol{\mu}_Y$, variance-covariance matrix V, and spectral matrix $f_Y(\omega)$, by finding a scalar process, Z_t, defined by

$$Z_t = \sum_{j=-\infty}^{\infty} \boldsymbol{b}_{t-j}' \boldsymbol{Y}_j, \tag{8}$$

and absolutely summable $k \times 1$ filters $\{\boldsymbol{b}_t\}$ and $\{\boldsymbol{c}_t\}$, so that the error of approximation, $\boldsymbol{Y}_t - \widehat{\boldsymbol{Y}}_t$ is small relative to mean squared error, where $\widehat{\boldsymbol{Y}}_t = \boldsymbol{\mu}_Y + \sum_{j=-\infty}^{\infty} \boldsymbol{c}_{t-j} Z_j$. If $\boldsymbol{b}(\omega)$ is the transform of $\boldsymbol{b}_t$, and $f_Z(\omega)$ the spectral density of Z_t, then the problem becomes one of finding a *complex* vector $\boldsymbol{b}(\omega)$, subject to the constraint that $\boldsymbol{b}^*(\omega) V \boldsymbol{b}(\omega) = 1$, such that

$$f_Z(\omega) = \boldsymbol{b}^*(\omega) f_Y(\omega) \boldsymbol{b}(\omega)$$

is maximized. The solution, of course, is that $\boldsymbol{b}(\omega)$ is the eigenvector corresponding to the largest eigenvalue of $f_Y(\omega)$ in the metric of V, say $\lambda(\omega)$, and hence $f_Z(\omega) = \lambda(\omega)$ with $\boldsymbol{b}(\omega)$ so chosen.

In the language of scaling, we would state the same problem as follows. Given a vector process $\boldsymbol{Y}_t$ find a complex vector $\boldsymbol{b}$ such that at a given frequency ω the time series $Z_t(\boldsymbol{b}) = \boldsymbol{b}^* \boldsymbol{Y}_t$ has the largest possible spectrum (subject to $\boldsymbol{b}^* V \boldsymbol{b} = 1$). The solution is to choose $\boldsymbol{b} = \boldsymbol{b}(\omega)$, that is, the eigenvector corresponding to the largest eigenvalue of $f_Y(\omega)$ in the metric of V. In this case the spectrum of $Z_t(\boldsymbol{b}(\omega))$, say $f_Z(\omega, \boldsymbol{b}(\omega))$, attains the largest possible value, $\lambda(\omega)$. Hence, Brillinger's approach can be seen as a scaling problem with complex-valued scales. In our approach, we restrict $\boldsymbol{b}(\omega)$ to be real and $\boldsymbol{Y}_t$ to be the multiple indicator process associated with a categorical-valued process.

We have extended the concept of the spectral envelope for a stationary categorical time series to the situation where the time series is stationary only over short intervals. DNA sequences exhibit this kind of behavior; as seen in Section 4, the spectral envelopes differ between subsequences. The concept of an evolutionary spectral envelope is yet to be formalized. Our contribution to this new idea is the development of a computationally efficient algorithm that can segment a DNA sequence into separate blocks that give unique genetic coding information.

In the theoretical development of an evolutionary spectral envelope, we can use the model of a locally stationary process of Dahlhaus (1997) or its special case

given in Chiann and Morettin (1999). For ease of exposition, we just state the model for the univariate case. The extension to multivariate is given in Dahlhaus (1999).

A sequence of zero-mean stochastic processes $\{X_{t,T}, t = 1, \ldots, T\}$ is called locally stationary if it admits a Cramèr-like representation

$$X_{t,T} = \int_{-1/2}^{1/2} \exp(2\pi i\omega) A(t/T, \omega) dZ(\omega), \tag{9}$$

where $Z(\omega)$ is a stochastic process whose increments are orthogonal and satisfy regularity conditions on its cumulants. The function $A(\cdot)$ is the time-varying filter. Under this model, the evolutionary spectrum is defined to be $f(u, \omega) = |A(u, \omega)|^2$. Under the Dahlhaus model, one can form consistent estimators of the evolutionary spectrum by computing the spectrograms. One problem with this approach is that it will be computationally burdensome particularly when the time series is very long.

The dyadic segmentation framework is computationally efficient and provides a remedy to the above problem. It is in the tradition of the growing body of work used in regression and signal processing. Under the dyadic segmentation framework, consistent estimators for the Dahlhaus time-varying spectrum are formed when the segmentation is known. This result is given in Ombao et al (1999). Thus, one conjecture that can be given at this point is that under known segmentation, one can also form a consistent estimator for the true spectral envelope if the evolving spectral envelope follows the same smoothness assumptions of the Dahlhaus evolving spectrum. The next step is to rigorously define that evolving spectral envelope. Moreover, the over-all consistency still has to be addressed given that the segmentation has to selected from the data.

Acknowledgment

David Stoffer was supported, in part, by a grant from the National Science Foundation. Hernando Ombao was supported, in part, by a grant from the National Institutes of Mental Health.

References

ADAK, S. (1998). Time dependent spectral analysis of non-stationary time series, *Journal of the American Statistical Association,* **93**, 1488-1501.

BINA, M. (1994). Periodicity of dinucleotides in nucleosomes derived from siraian virus 40 chromatin. *Journal of Molecular Biology*, **235**, 198-208.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. AND STONE, C. (1984). *Classification and Regression Trees.* Monterey, CA: Wadsworth and Brooks/Cole.

BREIMAN, L. AND FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, **80**, 580-619.

BRILLINGER, D. R. (1975). *Time Series: Data Analysis and Theory.* 2nd ed: 1981. San Francisco: Holden-Day.

CHIANN, C. AND MORETTIN, P. (1999). Estimation of Time Varying Linear Systems. *Statistical Inference for Sthocastic Processes*, **2**, 253-285.

COIFMAN, R. AND WICKERHAUSER, M. (1992). Entropy based algorithms for best basis selection. *IEEE Transactions on Information Theory*, **32**, 712-718.

CORNETTE, J.L., CEASE, K.B., MARGAHT, H., SPOUGE, J.L., BERZOFSKY, J.A. AND DELISI, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *Journal of Molecular Biology*, **195**, 659-685.

DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics*, **25**, 1-37.

DAHLHAUS, R. (1999). A likelihood approximation for locally stationary processes. Beiträge zur Statistik 56, Universität Heidelberg.

DONOHO, D., MALLAT, S. AND VON SACHS, R. (1998). Estimating covariances of locally stationary processes: rates of convergence of best basis methods. *Technical Report 517*, Department of Statistics, Stanford University. Submitted for publication and under revision.

EISENBERG, D., WEISS, R.M. AND TERWILLGER, T.C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proceedings of the National Academy of Science*, **81**, 140-144.

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis.* London: Academic Press.

HANNAN, E. J. (1970). *Multiple Time Series.* New York: Wiley and Sons.

IOSHIKHES, I., BOLSHOY, A. AND TRIFONOV, E.N. (1992). Preferred positions of AA and TT dinucleotides in aligned nucleosomal DNA sequences. *Journal of Biomolecular Structure and Dynamics*, **9**, 1111-1117.

KARLIN S. AND MACKEN, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *Journal of the American Statistical Association*, **86**, 27-35.

MALLAT, S., PAPANICOLAU, G. AND ZHANG, Z. (1998). Adaptive covariance estimation of locally stationary processes. *Annals of Statistics*, **26**, 1-47.

McLachlan, A.D. and Stewart. M. (1976). The 14-fold periodicity in alpha-tropomyosin and the interaction with actin. *Journal of Molecular Biology*, **103**, 271-298.

Michailidis, G. and De Leeuw, J. (1988). The Gifi system of descriptive multivariate analysis. *Statistical Science*, **18**, 307-336.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto: University of Toronto Press.

Ombao, H., Raz, J., von Sachs, R. and Malow, B. (1999). Automatic statistical analysis of bivariate nonstationary time series. *Technical Report 9908*, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Vols 1 and 2. London: Academic Press.

Satchwell, S.C., Drew, H.R. and Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*, **191**, 659-675.

Shumway, R.H. and Stoffer, D.S. (2000). *Time Series Analysis and Its Applications*. New York: Springer-Verlag.

Stoffer, D.S., Scher, M., Richardson, G., Day, N. and Coble, P. (1988) A Walsh-Fourier Analysis of the Effects of Moderate Maternal Alcohol Consumption on Neonatal Sleep-State Cycling. *Journal of the American Statistical Association*, **83**, 954-963.

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

Stoffer, D.S., Tyler, D.E., McDougall, A.J. and Schachtel, G. (1993b). Spectral analysis of DNA sequences (with discussion). *Bulletin of the International Statistical Institute*, Bk I, pp 345-361; Discussion: Bk IV, pp 63-69, 1994.

Stoffer, D.S., Tyler, D.E., Wendt, D.A. (2000). The spectral envelope and its applications. *Statistical Science*, **15**, in press.

Tavaré, S. and Giddings, B.W. (1989). Some statistical aspects of the primary structure of nucleotide sequences. In *Mathematical Methods for DNA Sequences*, M.S. Waterman ed., pp. 117-131, Boca Raton, Florida: CRC Press.

VIARI, A., SOLDANO, H. AND OLLIVIER, E. (1990). A scale-independent signal processing method for sequence analysis. *Computer Applications in the Biosciences*, **6**, 71-80.

WICKERHAUSER, M. (1994). *Adapted Wavelet Analysis from Theory to Software*. IEEE Press, Wellesley, MA.

David S. Stoffer
Department of Statistics,
University of Pittsburgh,
stoffer@stat.pitt.edu
USA

Hernando Ombao
Department of Statistics,
University of Pittsburgh,
ombao@stat.pitt.edu
USA