TIME-DEPENDENT SPECTRAL ANALYSIS OF
NONSTATIONARY TIME SERIES

BY

SUDESHNA ADAK

TECHNICAL REPORT NO. 478
MAY 1995

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS92–04864

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Time-Dependent Spectral Analysis of Nonstationary Time Series

By

Sudeshna Adak

Technical Report No. 478

May 1995

Prepared Under the Auspices

Of

National Science Foundation Grant DMS92-04864

Department of Statistics
Stanford University
Stanford, California
Time-dependent Spectral Analysis of Nonstationary Time Series

Sudeshna Adak, Department of Statistics, Stanford University, CA-94305.

Abstract — Modeling of nonstationary stochastic time series has found wide applications in speech processing, biomedical signal processing, seismology and failure detection. In this paper, the problem of defining a time-dependent spectrum for a general class of nonstationary processes is addressed. It is shown that Cohen's class of time-frequency distributions is unacceptable as time-dependent spectra for all nonstationary stochastic processes, but is appropriate for a class of nonstationary processes that are locally stationary.

A tree-based segmentation method of estimating the time-dependent spectrum is proposed for this class of locally stationary processes. A cross-validation procedure is suggested for an optimal bias-variance tradeoff in the procedure. Results of simulation studies demonstrate that the method has excellent ability to adapt to the rate at which the spectrum is changing. Applications of the method to speech signals and earthquake data are considered.

Key words: nonstationary processes, local stationarity, time-dependent spectra, tree-based methods, cross-validation.

1 Introduction

In the classical approach to spectral analysis, the time series is assumed to be weakly stationary. With this assumption, the frequency-domain behavior of a zero-mean, weakly stationary process \( \{X_t\} \) can be characterized by its Cramér representation (Priestley(1981)):

\[
X_t = \int A(\lambda) e^{i2\pi \lambda t} dZ(\lambda)
\]

(1)

where \( \text{Cov}(dZ(\lambda), dZ(\lambda')) = \begin{cases} 
0 & \text{if } \lambda \neq \lambda' \\
\frac{d\lambda}{\lambda} & \text{otherwise}
\end{cases} \)

The power spectrum of the process is defined as:

\[
f(\lambda) = |A(\lambda)|^2 = \int R_X(\tau)e^{-i2\pi \lambda \tau} d\tau
\]

(2)

where \( R_X(\tau) \) is the autocovariance function of the process \( \{X_t\} \), and is assumed to be integrable. This uniquely defines a spectrum which adequately describes the frequency domain behavior of \( \{X_t\} \).

Stationarity has been usually regarded as a valid assumption for series of short duration. However, such an assumption is rapidly losing its credibility in the enormous databases maintained by firms and organizations on a large variety of subjects such as geophysics, oceanography, meteorology, speech, and economics. A spectral description of a nonstationary stochastic process, similar to Equation 1, requires a representation in the time-frequency
plane. Unfortunately, in the time-frequency analysis of nonstationary processes, there appears to be no analogous way of defining a time-dependent spectrum uniquely. A partial list of various definitions is given in Loneyes(1968), Cohen(1989). Of these, the one that has achieved some popularity in the statistics literature is the evolutionary spectrum introduced in Priestley(1965). The topic has received far greater attention in the signal processing community starting with the introductory work of Gabor(1946) and Ville(1948). A unified formulation of the various definitions was presented in a landmark paper by Cohen(1966) and this general class of time-frequency distributions has come to be known as Cohen's class.

Let \( \{X_t\} \) be a zero-mean process with covariance function \( R_X(s, t) \). Then, Cohen's class of time-frequency distributions is defined as

\[
f(t, \lambda) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_X(u + \tau/2, u - \tau/2) g(v, \tau) e^{i2\pi \nu(u-t)} e^{-i2\pi \lambda \tau} du dv d\tau
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_X(u + \tau/2, u - \tau/2) G(t - u, \tau) e^{-i2\pi \lambda \tau} du d\tau
\]

(3)

\[
= \mathcal{F}_{\nu \rightarrow \lambda}[G(t, \tau) \star R_X(t + \tau/2, t - \tau/2)]
\]

(4)

where \( \mathcal{F} \) denotes Fourier Transform and \( \star \) denotes convolution with respect to variable \( t. \) \( G(t, \tau) \) \( (= \mathcal{F}_{\nu \rightarrow \nu}[g(v, \tau)] \) is for obvious reasons known as the time-lag kernel.

Note that for a stationary process, the spectrum is the Fourier Transform of its autocovariance function. Cohen's class generalizes this to nonstationary processes in defining a time-dependent spectrum as the Fourier Transform of a smoothed covariance. It is not based on any Cramér-like representation.

Table 1 shows the Cohen class representation of some of the definitions of time-dependent spectra proposed in the statistics and engineering literature.

<table>
<thead>
<tr>
<th>Author</th>
<th>Definition of ( f(t, \lambda) )</th>
<th>( G(t, \tau) ) : time - lag kernel</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wigner-Ville</td>
<td>( R_X(t + \tau/2, t - \tau/2)e^{-i2\pi \lambda \tau} d\tau )</td>
<td>( G(t, \tau) = \delta(t) )</td>
<td></td>
</tr>
<tr>
<td>Page(1952)</td>
<td>( R_X(t, t - \tau)e^{-i2\pi \lambda \tau} d\tau ) + ( R_X(t + \tau, t) e^{-i2\pi \lambda \tau} d\tau )</td>
<td>( G(t, \tau) = \delta(t - \tau/2) ) if ( \tau \geq 0. ) ( G(t, \tau) = \delta(t + \tau/2) ) if ( \tau \leq 0. )</td>
<td></td>
</tr>
<tr>
<td>Levin(1967)</td>
<td>( \frac{1}{2} R_X(t, t - \tau)e^{-i2\pi \lambda \tau} d\tau + ) ( \frac{1}{2} R_X(t + \tau, t) e^{-i2\pi \lambda \tau} d\tau )</td>
<td>( G(t, \tau) = \frac{1}{2} \delta(t - \tau/2) + ) ( \frac{1}{2} \delta(t + \tau/2) )</td>
<td></td>
</tr>
<tr>
<td>Spectrogram - Windowed Spectral Analysis</td>
<td>( E[</td>
<td>\int \int w(t - u)X(u)e^{-i2\pi \lambda \tau} du</td>
<td>^2] )</td>
</tr>
<tr>
<td>Priestley(1965)</td>
<td>( \frac{1}{4\pi^2} \exp \left[ -\frac{(t-u)^2}{4\pi^2r^2} \right] e^{-i2\pi \lambda \tau} d\tau )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choi-Williams(1989)</td>
<td>( R_X(u + \tau/2, u - \tau/2) )</td>
<td>( G(t, \tau) = \frac{1}{4\pi^2} e^{\frac{-t^2}{4\pi^2}} )</td>
<td></td>
</tr>
</tbody>
</table>

The assumption of stationarity is also convenient as it allows the estimation of the spec-
trum from a single record of the data. The usual procedure in estimating time-dependent spectra is to assume that the process, though nonstationary, is **slowly changing** in its spectral characteristics. This restriction to **local or quasi-stationary** time series allows the estimation of the time-dependent spectrum from a single record of the process. A **slowly changing** process is, roughly speaking, a time series for which at each time point, a **stationary interval** can be defined, within which the covariance is stationary or at least approximately so. (For a different approach, cf. Tjøstheim(1986), Mélard et al(1989).)

Amin(1987) showed that any time-averaging estimation method used to estimate the spectrum has to be adapted to the nature of the stationary interval. Priestley(1966) considered an estimate based on a windowed Fourier transform in case of processes for which the length of the stationary interval was bounded below. It was shown that the window-width has to be adapted to the bound on the length of the stationary interval. It is clear that even with an assumption of local-stationarity, an adaptive approach is called for that adjusts at each time point $t$ to use only the data from the stationary interval at $t$.

Sections 2 and 3 discuss the validity of definitions of time-dependent spectra and the nonstationary processes for which time-dependent spectra may be meaningfully defined. Sections 4 and 5 propose a method of estimation of these time-dependent spectra adaptive to the nature of nonstationarity.

## 2 Desirable Properties of Time-Frequency Distributions

Spectral analysis of stationary processes is a well-known concept and well used in practice. In recent years, an equal, if not more, amount of interest has been focused on time-dependent spectral analysis. In view of the large and ever-growing list of proposed definitions, it seems worth while to explore what properties such a definition should be expected to satisfy in order to justify it as a **spectrum**. It is shown that in this respect Cohen's class of time-frequency distributions is not entirely satisfactory as a time-dependent spectrum for **all** nonstationary processes. In fact, it is not apriori clear that such a definition could even exist for **all** nonstationary processes.

The spectral density of a weakly stationary, zero-mean process satisfies certain properties

(i) $f(\lambda)$ is real and even for a real-valued process.

(ii) $f(\lambda)$ is non-negative.

(iii) $E(X_t^2) = \int f(\lambda)d\lambda$, for all $t$.

The following is a list of properties, similar to the above, that one might wish a time-dependent spectrum to possess. In the case of Cohen's class of time-frequency distributions, these requirements impose certain restrictions on the time-lag window $G(t, \tau)$.

- **Property 1**: The time-frequency distribution(TFD) must be **real** and it is expected to be **non-negative** in order to give a meaningful representation of the variation in energy over frequency and time.
$f(t, \lambda)$ is real if and only if $G(t, \tau) = G^*(t, -\tau)$

Proof: Easily follows from the fact that the Fourier Transform of a function is real iff the function is even.

• **Property 2:** For a physical interpretation of $f(t, \lambda)$ as a measure of the breakdown of the energy of the process at time $t$ over so-called "frequency", the TFD must satisfy, for a zero-mean process,

$$E(X^2_t) = R_X(t, t) = \int_{-\infty}^{\infty} f(t, \lambda) d\lambda \iff G(t, 0) = \delta(t) \text{ for all } t$$

Proof: Since $\int \mathcal{F}[f(x)] = f(0)$, $\int_{-\infty}^{\infty} f(t, \lambda) d\lambda = \int_{-\infty}^{\infty} G(t - u, 0) R_X(u, u) du$, it follows that $G(t, 0)$ must be $\delta(t)$.

• **Property 3:** Another very appealing property is that if the process is stationary, then the TFD yields the usual spectral density. This can be easily shown to be equivalent to the condition:

$$\int_{-\infty}^{\infty} G(t, \tau) dt = 1 \text{ for all } \tau$$

• **Property 4:** If the processes $\{X_t, t \leq t_o\}$ and $\{X_t, t > t_o\}$ are independent, then:

For $t > t_o$, $f(t, \lambda)$ should not involve $R(u, v)$ for $u$ and $v \leq t_o$.

For $t < t_o$, $f(t, \lambda)$ should not involve $R(u, v)$ for $u$ and $v > t_o$.

A sufficient condition for this property to be valid is the condition $G(t, \tau) = 0$, if $|t| > |\tau| / 2$.

Proof: As $\{X_t, t \leq t_o\}$ and $\{X_t, t > t_o\}$ are independent, and since $G(t, \tau) = 0$, if $|t| > |\tau| / 2$,

$$f(t, \lambda) = \int_{u+|\tau|/2 \leq t_o} G(t - u, \tau) R_X(u + \tau/2, u - \tau/2) e^{-i2\pi\lambda\tau} du d\tau$$

$$+ \int_{u-|\tau|/2 > t_o} G(t - u, \tau) R_X(u + \tau/2, u - \tau/2) e^{-i2\pi\lambda\tau} du d\tau$$

$$= \int_{u+|\tau|/2 \leq t_o, |t-u| \leq |\tau|/2} G(t - u, \tau) R_X(u + \tau/2, u - \tau/2) e^{-i2\pi\lambda\tau} du d\tau$$

$$+ \int_{u-|\tau|/2 > t_o, |t-u| \leq |\tau|/2} G(t - u, \tau) R_X(u + \tau/2, u - \tau/2) e^{-i2\pi\lambda\tau} du d\tau$$

Thus,

$$f(t, \lambda) = \int_{u+|\tau|/2 \leq t_o, |t-u| \leq |\tau|/2} G(t - u, \tau) R_X(u + \tau/2, u - \tau/2) e^{-i2\pi\lambda\tau} du d\tau \text{ for } t < t_o$$

$$f(t, \lambda) = \int_{u-|\tau|/2 > t_o, |t-u| \leq |\tau|/2} G(t - u, \tau) R_X(u + \tau/2, u - \tau/2) e^{-i2\pi\lambda\tau} du d\tau \text{ for } t > t_o.$$
Accounting for all these restrictions on the magnitude and shape of the time-lag window, $G(t, \tau)$ represents a butterfly pattern in the time-lag plane. Such time-lag kernels are called butterfly functions (Amin (1987)).

2.1 A negative result about Cohen's class

One of the desired properties of a time-frequency distribution $f(t, \lambda)$ is that

$$R_X(t, t) = \int_{-\infty}^{\infty} f(t, \lambda) d\lambda$$  \hspace{1cm} (5)

This suggests that $f(t, \lambda)$ may be interpreted as a decomposition of the energy of the process at time $t$ over "frequencies".

The following theorem shows that Cohen's class of time-frequency distributions cannot be non-negative and describe the breakdown of the energy over frequencies at each time point for all nonstationary processes.

**Theorem 2.1** For any choice of the time-lag kernel $G(t, \tau)$, there is some non-stationary process $\{X_t\}$ for which $f(t, \lambda)$ as defined in Equation 4 is either negative for some $t$ or it violates the requirement of Equation 5 that $R_X(t, t) = \int_{-\infty}^{\infty} f(t, \lambda) d\lambda$.

**Proof:** Suppose Equation 5 is valid and $f(t, \lambda)$ is non-negative for each $t$, for all nonstationary processes, for some choice of the time-lag kernel $G(t, \tau)$.

It was shown earlier that Equation 5 is valid iff $G(t, 0) = \delta(t)$ for all $t$.

From Equation 4, $f(t, \lambda) = \mathcal{F}[\bar{R}_t(\tau)]$, where $\bar{R}_t(\tau) = G(t, \tau)^{(t)} \ast R_X(t + \tau/2, t - \tau/2)$

It follows from Bochner's Theorem that $f(t, \lambda)$ is non-negative for all $t$ if and only if $\bar{R}_t(\tau)$ is a non-negative definite function in $\tau$ for all $t$. Now,

$$\bar{R}_t(\tau) = \int_{-\infty}^{\infty} G(t - u, \tau) R_X(u + \tau/2, u - \tau/2) du$$

$$= \int_{-\infty}^{\infty} K_t(v + \tau, v) R_X(v + \tau, v) dv$$

where $K_t(s, s') = G(t - s' - (s - s')/2, s - s')$, $G(t, \tau) = K_t(\tau/2, -\tau/2)$

Loynes (1968) showed that $\bar{R}_t(\tau)$ is non-negative definite for each $t$, for any nonstationary process, iff $K_t(s, s')$ is a covariance function or a covariance function almost everywhere i.e. $K_t(s, s') = \text{Cov}(Y_s^{(t)}, Y_s^{(t)})$ for some random process $Y_s^{(t)}$. But,

$$\text{Var}(Y_s^{(t)}) = K_t(s, s) = G(t - s, 0) = \delta(t - s) = 0, \text{if } s \neq t$$

$$\implies Y_s^{(t)} = \text{constant if } s \neq t$$

$$\implies G(t, \tau) = K_t(\tau/2, -\tau/2) = 0, \text{unless } t = \tau = 0$$

$$\implies G(t, \tau) = \delta(t)\delta(\tau)$$

$$\implies f(t, \lambda) = R_X(t, t)$$
which is not an acceptable definition of a time-frequency distribution.

Thus, there is no universal choice of the time-lag kernel which is appropriate for all nonstationary processes. The choice of the time-lag kernel has to be adapted to the nature of the nonstationarity.

3 Locally Stationary Processes

From the previous section, it is clear that a definition of a time-dependent spectrum for all nonstationary processes that is physically meaningful and meets all the desired criteria does not yet exist. However, the general approach is to restrict the extremely large class of all nonstationary processes to a smaller class, for which a time-dependent spectrum can be meaningfully defined — this usually being a process that is slowly changing.

A novel approach introduced in Dahlhaus(1993) to define a time-frequency distribution for locally stationary processes ensures a reasonable asymptotic theory in the nonstationary framework. This definition also generalizes the Cramér representation of a stationary process, which is not provided by Cohen’s class.

Definition 3.1 [Adapted from Dahlhaus(1993), Sec. 2] A sequence of zero-mean stochastic processes $X_{t,N}(t = 1, \ldots, N)$ is called locally stationary at time $u$, $0 \leq u \leq 1$ if there exists a representation

$$X_{t,N} = \int_{-1/2}^{1/2} A_{t,N}(\lambda) e^{i2\pi \lambda t} dZ(\lambda)$$

where

(i) $Z(\lambda)$ is an orthogonal increment, zero-mean process

with $\text{Cov}(dZ(\lambda), dZ(\lambda')) = \begin{cases} 0 & \text{if } \lambda \neq \lambda' \\ d\lambda & \text{otherwise} \end{cases}$

(ii) There exists a constant $K$ and a smooth function $A : [0,1] \times [-1/2, 1/2] \rightarrow \mathbb{C}$ with $A(u, \lambda) = A^*(u, -\lambda)$ such that for all $N$,

$$\sup_{\lambda \in \epsilon_N(u)} \sup_{N} | A_{t,N}(\lambda) - A(u, \lambda) | \leq KN^{-1}$$

where $\epsilon_N(u)$ is a small neighborhood of $u$, with length($\epsilon_N(u)$) = $O(N^{-\alpha})$, $0 \leq \alpha < 1$.

(iii) For all $\lambda$, $A(u, \lambda)$ is continuous in $u$.

Definition 3.2 For a sequence of processes $X_{t,N}$, locally stationary at time $u$, as given in Definition 3.1 above, the time-dependent spectrum at time $u$ is $f(u, \lambda) = |A(u, \lambda)|^2$

Definition 3.3 A sequence of zero-mean stochastic processes is said to be almost everywhere locally stationary if it is locally stationary at almost all time points $u$, $0 \leq u \leq 1$. 
Remarks —

(a) Equation 7 ensures that the process is slowly changing at time $u$. It is clear from Equation 7 that for $t/N$ in a small neighborhood $e_N(u)$ of $u$, $A^0_{t,N}(\lambda) \approx A(u, \lambda)$. So, the process is approximately stationary in a neighborhood of $u$, with the stationary interval being $e_N(u)$.

(b) (Asymptotics:) In Definition 3.1, allowing $N \to \infty$ means that more and more data of the local structure is observed. For example, consider a piecewise stationary process for which

$$A^0\left(\frac{t}{N}, \lambda\right) = I_{[0,1/2]}\left(\frac{t}{N}\right) A_1(\lambda) + I_{[1/2,1]}\left(\frac{t}{N}\right) A_2(\lambda),$$

where $A_1(\lambda)$ and $A_2(\lambda)$ correspond to the spectra of two stationary processes, $\{X_1(t)\}$ and $\{X_2(t)\}$ respectively. Then, letting $N \to \infty$ means that more and more observations are obtained on the stationary process $\{X_1(t)\}$ for the first half of the data and on $\{X_2(t)\}$ for the second half.

(c) Priestley(1965)'s definition of a time-dependent spectrum is

$$f\left(\frac{t}{N}, \lambda\right) = |A^t_{t,N}(\lambda)|^2$$

Definition 3.2 is clearly the asymptotic version of Equation 8, (letting $N \to \infty$ and $t/N \to u$). So, even though asymptotically there is no difference and the method of estimation is not affected by our choice, Definition 3.2 has been proposed as the more suitable choice. The reason for this choice is explained by the next remark.

(d) Stationary ARMA(p,q) models are characterized by rational spectra. For time-varying ARMA(p,q) models,

$$\sum_{j=0}^{p} a_j(t/N) X_{t-j,N} = \sum_{k=0}^{q} \sigma(t/N) b_k(t/N) \varepsilon_{t-k}$$

the time-dependent spectrum as given in Definition 3.2 is

$$f(u, \lambda) = \left| \sigma(u) \sum_{k=0}^{q} b_k(u) e^{-i2\pi\lambda k} \right|^2$$

(see Dahlhaus(1993), Theorem 4.1). This is analogous to the rational spectra for stationary ARMA models. However, using Equation 8 as the definition does not yield the above as the time-dependent spectra for time-varying ARMA(p,q) models. For example, it can be shown by direct verification that for a time-varying AR(1) model,

$$A^0_{t,N}(\lambda) = \sum_{l=0}^{\infty} (-1)^l \left( \prod_{j=0}^{l-1} a_1(t-j/N) \right) \sigma(t-j/N) e^{-i2\pi\lambda l}$$

(e) One of the criticisms aimed at Priestley's time-dependent spectrum is that it is not unique, though always non-negative. On the other hand, the Wigner-Ville spectrum (see Table 1) is known to be unique but can possibly take negative values. The following result
from Dahlhaus(1993) showed that asymptotically, the two agree.

**Result:**[Dahlhaus(1993), Theorem 2.2] If \( A(u, \lambda) \) is differentiable in \( u \) and \( \lambda \) with uniformly bounded derivatives and length(\( \epsilon_N(u) \)) = \( O(N^{-\alpha}) \), where \( 1/2 < \alpha < 1 \), then

\[
\begin{align*}
  f(u, \lambda) &= |A(u, \lambda)|^2 \\
  &= \lim_{N \to \infty} \sum_{s=-\infty}^{\infty} \text{Cov}(X_{[uN+s/2]}, X_{[uN-s/2]} e^{-i2\pi \lambda s} \text{ a.e. in } \lambda}
\end{align*}
\]

which is the Wigner-Ville spectrum.

( \( X_{t,N} \) is defined by Equation 6 with \( A_{t,N}^0(\lambda) = A_{t,N}^0(\lambda) \) for \( t < 1 \), and \( A_{t,N}^0(\lambda) = A_{N,N}(\lambda) \) for \( t > N \).)

### 3.1 Estimation of time-dependent spectra for locally stationary processes

Having defined a time-dependent spectrum, at least for locally stationary processes, the next step is to provide a method of estimating it. In the estimation of time-dependent spectra, the class of almost everywhere locally stationary processes (See Definition 3.3) is precisely the target class of nonstationary processes which can be universally approximated by **piecewise stationary processes** as described by the following theorem.

**Theorem 3.1** Consider the class of piecewise stationary processes

\[
\mathcal{M} = \{ \tilde{X}_{t,N} : \tilde{X}_{t,N} = \sum_{j=0}^{J-1} \tilde{X}_{t}^{(j)} I(u_j \leq \frac{t}{N} < u_{j+1}) \}
\]

where \( 0 = u_0 < u_1 < ... < u_J = 1 \) form a partition of \([0,1]\), which depends on \( N \) and \( \tilde{X}_{t}^{(j)} \) are stationary processes with spectra \( f^{(j)}(\lambda) \). Then, for any almost everywhere locally stationary process \( \tilde{X}_{t,N}(t = 1, ..., N) \), there exists a process \( \tilde{X}_{t,N} \in \mathcal{M} \) such that for all \( N \),

\[
\sum_{t=1}^{N} E(X_{t,N} - \tilde{X}_{t,N})^2 \leq KN^{-1}
\]

**Proof:** \( X_{t,N}(t = 1, ..., N) \) is an almost everywhere locally stationary process

\[\implies\] There exist constants \( K \) and \( \alpha \), \( 0 \leq \alpha < 1 \), and a smooth function \( A : [0,1] \times [-1/2, 1/2] \to \mathcal{M} \) such that for all \( N \),

\[
\sup_{\lambda} |A_{t,N}^0 - A(u_j, \lambda)| \leq KN^{-1} \text{ for } u_j \leq \frac{t}{N} < u_{j+1}, j = 0, 1, ..., J - 1
\]

where \( 0 = u_0 < u_1 < ... < u_J \) is a partition of \([0,1]\), possibly depending on \( N \), and \( J = O(N^\alpha) \)

Define the piecewise stationary process \( \tilde{X}_{t,N} = \sum_{j=0}^{J-1} \tilde{X}_{t}^{(j)} I(u_j \leq \frac{t}{N} < u_{j+1}) \), where the
stationary process $\tilde{X}_t^{(j)} = \int_{-1/2}^{1/2} A(u_j, \lambda) e^{i2\pi \lambda t} dZ(\lambda)$

$$
\sum_{t=1}^{N} E(X_{t,N} - \tilde{X}_{t,N})^2
= \sum_{t=1}^{N} E\left[ \int_{-1/2}^{1/2} A_{t,N}(\lambda) e^{i2\pi \lambda t} dZ(\lambda) - \sum_{j=0}^{J-1} \mathcal{I}(u_j \leq \frac{t}{N} < u_{j+1}) \int_{-1/2}^{1/2} A(u_j, \lambda) e^{i2\pi \lambda t} dZ(\lambda) \right]^2
= \sum_{t=1}^{N} \int_{-1/2}^{1/2} \left| A_{t,N}(\lambda) - \sum_{j=0}^{J-1} \mathcal{I}(u_j \leq \frac{t}{N} < u_{j+1}) A(u_j, \lambda) \right|^2 d\lambda
\leq KN^{-1} \text{ by Equation 9}
$$

Remarks —
(a) Theorem 3.1 suggests that
$$
f(u, \lambda) \approx \sum_{j=1}^{J} f^{(j)}(\lambda) \mathcal{I}(u_j \leq u < u_{j+1})
$$

(b) According to the theorem, any process that is locally stationary a.e. on [0,1] can be approximated by a piecewise stationary process. Furthermore, the number of segments in the piecewise stationary approximation, $J$, may increase with $N$ (the number of observations), but $\frac{J}{N} \to 0$ as $N \to \infty$. This ensures that with increasing $N$, more and more observations fall in each stationary interval and hence the time-dependent spectrum, or at least a piecewise stationary approximation of it can be easily estimated using the windowed Fourier transform.

The major drawback in using the windowed Fourier transform is that the window-width has to be adapted to the rate of change in the spectrum. Clearly, the window has to be small enough, so that the data is locally stationary within that interval. However, too small a window size results in too few data points, which causes a bias and variability in the estimated spectrum from the stationary interval. The next two sections provide a method of estimation, using a tree-based algorithm that allows us to choose a data-dependent window size, adapting to the rate at which the spectrum changes.

4 Tree-Based Adaptive Estimation

In the previous section, it was seen that any locally stationary process can be well approximated by a piecewise stationary process. This suggests that in estimating the time-dependent spectrum of a locally stationary process, a segmentation procedure that partitions the time series into approximately stationary intervals can be used. The purpose of any segmentation algorithm for a locally stationary process is to detect changes in the spectra of the process over time and to accurately estimate these change points. An adaptive tree-based segmentation algorithm is proposed here to estimate the time-dependent spectrum, which partitions the time-domain, using the following algorithm

9
• Initially, a complete tree is grown that halves each segment recursively into very short segments, (with possibly overlapping segments) as shown in Figure 1.

![Figure 1: Successive Refinement over depth in tree](image)

• The spectrum in each block is estimated.

• Then an optimal pruning algorithm is used to recombine adjacent segments for which the spectra are the same.

### 4.1 Fast Segmentation algorithm for series of length $N = 2^J$

**Algorithm: Growing the complete tree**

| Step 1: Set $D = \lfloor \log_2 N \rfloor - 2$ | [ # Maximum Depth to which tree is grown ] |
| Step 2: Set $m = N/2^{D+1}$ | [ # Maximum overlap of segments ] |
| Step 3: For $d = 0 : D$, | |
| Divide the data into $2^d$ blocks, block(b,d): b = 0,..,2^d − 1, | |
| each block overlapping the next one over 2m-1 data points. | |
| block(b,d) corresponds to a node --- node(b,d) of the tree | |
| Step 4: For $b = 0 : (2^d - 1)$, | |
| Compute an estimate, $\hat{f}_{b,d}(\lambda)$, of the spectrum in block(b,d) | |
| end for | |
| end for |

Having estimated the spectra in each block of the tree, an optimal pruning algorithm is used to find the most stationary segmentation with the smallest number of segments — i.e. most stationary in the sense that it minimizes

$$\sum_{\text{segments}} \text{distance}(\hat{f}_{\text{left}}, \hat{f}_{\text{right}})$$

(10)
where \( \hat{f}_{\text{left}}, \hat{f}_{\text{right}} \) are the estimated spectra from the left and right half of the segment respectively and distance is a measure of discrepancy between the two spectra.

**Algorithm: Optimally Pruning the tree**

<table>
<thead>
<tr>
<th>For ( d = (D - 1) : -1 : 0 ),</th>
</tr>
</thead>
<tbody>
<tr>
<td>For ( b = 0 ) to ( (2^d - 1) ),</td>
</tr>
<tr>
<td>Step 5: Compute ( R_{b,d} = \text{distance}(\hat{f}<em>{2b,d+1}, \hat{f}</em>{2b+1,d+1}) )</td>
</tr>
<tr>
<td>Set Value((b,d) = R_{b,d} )</td>
</tr>
<tr>
<td>[ # ( R_{b,d} ) is a measure of the change in the spectra</td>
</tr>
<tr>
<td>between the two halves of block((b,d) ) ]</td>
</tr>
<tr>
<td>Step 6: If ( d = D - 1 ), Mark the block((b,d) ) as terminal</td>
</tr>
<tr>
<td>If ( d &lt; D - 1 ),</td>
</tr>
<tr>
<td>If Value((b,d) \leq \text{Value}(2b,d+1) + \text{Value}(2b+1,d+1) )</td>
</tr>
<tr>
<td>Mark the block((b,d) ) as terminal</td>
</tr>
<tr>
<td>Otherwise, leave block((b,d) ) unmarked and</td>
</tr>
<tr>
<td>Set Value((b,d) = \text{Value}(2b,d+1) + \text{Value}(2b+1,d+1) )</td>
</tr>
<tr>
<td>end for</td>
</tr>
<tr>
<td>end for</td>
</tr>
</tbody>
</table>

**Final Segmentation** = Set of highest marked blocks

\[
= \{ \text{block}(b,d) : \text{block}(b,d) \text{ is marked and its ancestors are unmarked} \}
\]

**Value of the best tree**, \( R(T^*) = \text{Value}(0,0) = \text{minimized value of Equation 10} \)

**Estimated time-dependent spectrum** = \( \{ \hat{f}_{b,d}(\lambda) : \text{block}(b,d) \in \text{the final segmentation} \} \)

The optimal pruning algorithm is similar to Wickerhauser(1994)'s best basis algorithm for minimization of entropy. The following theorem shows that the above algorithm finds the most stationary segmentation i.e. minimizes Equation 10.

**Theorem 4.1** Let the tree grown to depth \( D \) be \( T_0 = \{ \text{node}(b,d) : b = 0, ..., 2^d - 1, d = 0,1, ..., D - 1 \} \).

For any tree \( T \), with set of terminal nodes \( \tilde{T} \), let \( R(T) = \sum_{\text{node}(b,d) \in \tilde{T}} R(b,d) \)

Then, there is a unique smallest optimally pruned subtree of \( T_0 \), say \( T^* \), that minimizes \( R(T) \) among all subtrees \( T \) of \( T_0 \), which is obtained by the above optimal pruning algorithm.

and the Final Segmentation = \( \{ \text{block}(b,d) : \text{node}(b,d) \text{ is a terminal node of } T^* \} \)

and \( R(T^*) = \text{Value}(b = 0, d = 0) \) where \( \text{Value}(\cdot, \cdot) \) is defined by the optimal pruning algorithm.

**Proof**: Follows easily from Breiman et al.(1984), Theorem 10.7

**4.2 Local Spectra**

In estimating the evolutionary spectrum of a locally stationary process, a local version of the periodogram — a *windowed Fourier transform*, is used. A naive method of constructing
local estimates of the spectrum is to construct windowed periodograms of the form:

\[
\hat{f}_{b,d}(\lambda) = \frac{1}{N/2^d} \left| \sum_{t \in \text{block}(b,d)} X_t e^{-i2\pi\lambda t/N} \right|^2
\]

For a stationary time series, \(E(\hat{f}_{b,d}(\lambda)) = \int f^{1/2}(\lambda') K_{N/2^d}(\lambda - \lambda') d\lambda',\) where \(K_{N'}(\lambda)\) is the Fejer Kernel for data of length \(N'.\) The wide side lobes of the Fejer Kernel (see Figure 2) can cause a large bias, especially for series of very short length — this phenomenon is well-known as leakage.

![Fejer Kernel](image)

**Figure 2: Fejer Kernel**

In the tree-growing algorithm, as shorter and shorter segments are considered with increasing depth, leakage may be a severe problem. One of the remedies for leakage is to consider **tapering** with a window whose Fourier transform has small side lobes rather than the rectangular window. A class of arbitrarily smooth tapers - tapers that decrease to 0 with an arbitrary degree of smoothness and have small side lobes are introduced here.

- We start by defining the bell over an interval \(I = [\alpha, \beta].\)
  Let \(m > 0\) be fixed such that \(\beta - \alpha \geq 2m\)
  Define the bell over \(I\) as
  \[
  w_I(t) = \begin{cases} 
  r \left( \frac{t-\alpha}{m} \right) & \text{if } \alpha - m < t < \alpha + m \\
  1 & \text{if } \alpha + m \leq t \leq \beta - m \\
  r \left( \frac{\beta-t}{m} \right) & \text{if } \beta - m < t < \beta + m \\
  0 & \text{otherwise}
  \end{cases}
  \]  
  \(r(\cdot)\) is a rising cutoff function — \(r(t)\) increases smoothly from 0 to 1 as \(t\) increases from -1 to 1.

- In the tree growing algorithm, for block\((b,d) : b = 0,1,\ldots,2^d - 1; d = 0,1,\ldots,D\)
  The interval is taken as \(I_{b,d} = [b * N/2^d + 1, (b+1) * N/2^d]\) and \(m = N/2^{D+1},\) where \(D\) is the maximum depth to which tree is grown.
  Let \(w_{b,d}(t)\) be the bell defined on the interval \(I_{b,d}\) as in Equation 11. The resulting successive refinement is as shown in Figure 1.
• We define the evolutionary periodogram on the segmentation \( \{ J_{b,d} : b = 0,1,\ldots,2^d-1\} \) by

\[
\hat{f}_{b,d}(\lambda) = \frac{1}{N/2^d} \left| \sum_t w_{b,d}(t) X_t e^{-i2\pi \lambda t/N} \right|^2 
\]

(12)

• Choice of bell functions

1. Split-Cosine Bell on [-1,1]: \( r(u) = 0.5 - 0.5 \cos(\pi.(1+u)/2) = \sin^2(\pi.(1+u)/4) \)

2. Sine and Iterated-Sine Bell on [-1,1]:

\[
\begin{align*}
  r^{(k)}(u) &= r^{(k-1)}(\sin(\pi.u/2)), \quad k \geq 1 \\
  r^{(0)}(u) &= \sin(\pi.(1+u)/4)
\end{align*}
\]

It can be shown by induction (see Wickerhauser(1994)) that the iterated sine bell of degree \( k \), \( r^{(k)} \) has \( 2^k - 1 \) vanishing derivatives at \( \pm 1 \).

3. Class of arbitrarily Smooth Tapers on [-1,1]: \( r(u) = \sin[\theta(u)] \), where \( \theta(u) = \int_{-1}^{u} \phi(s)ds \), \( \phi(\cdot) \) is a non-negative, even on [-1,1] and increasing on [0,1].

This is the bell function of the smooth, local trigonometric transforms given in Auscher et al(1992).

Remarks —

(a) Fixed Tapering: In the tree growing algorithm, at any depth, each block of data has \((2m-1)\) points tapered on each side, rather than (say) always tapering by 50% at each end. Fixed Tapering was considered desirable for two reasons —

Firstly, leakage effects decreases with increasing size of the blocks and the advantage obtained by tapering is small. So, by fixed tapering, there is 50% tapering for the smallest segments (i.e. at maximum depth) and very little tapering as the block size increases. Also, another advantage of fixed tapering is that the percentage of observations that overlap over blocks decreases with increasing block size. (See Figure 1.)

(b) Zero Padding: In estimating the local spectrum in block \((b,d)\) as defined in Equation 12, the actual estimate is obtained over a grid of frequencies which is twice as fine as that associated with the Fourier frequencies for block \((b,d)\): \( \lambda_j = 2\pi j.2^d/2N, \; j = 0,1,\ldots,N/2^d-1 \). This is achieved by zero-padding the tapered data in block \((b,d)\) and then using the Fast Fourier Transform. By using a grid twice as fine as the Fourier frequencies, it is hoped that important features of the spectrum are not missed.

(c) Fast Algorithm: One of the major breakthroughs in Fourier Analysis has been the Fast Fourier Transform algorithm which requires \( O(N \log N) \) operations rather than \( O(N^3) \). It can be easily shown that the algorithm outlined here requires \( O(N \log^2 N) \) operations to compute the spectra over dyadic segmentations in growing the tree, and \( O(N) \) comparisons in pruning the tree. So, an estimate of the time-dependent spectrum is obtained in \( O(N \log^2 N) \) time.
4.3 Distance Between Spectra

In the tree-pruning algorithm, adjacent segments are recombined based on the homogeneity of the spectra of the two segments. This requires a measure of the distance between the estimated spectra of adjacent segments. Four of the most common distance measures for spectral densities were incorporated with the algorithm.

- **Kolmogorov-Smirnov distance**

\[
R_{b,d} = \frac{N}{2^d} \max_j |\hat{F}_{2b,d+1}(\lambda_j) - \hat{F}_{2b+1,d+1}(\lambda_j)|
\]

Here, \(\hat{F}_{b,d}(\lambda) = \int_{-1/2}^{1/2} \hat{f}_{b,d}(\lambda')d\lambda'\) / \(\int_{-1/2}^{1/2} \hat{f}_{b,d}(\lambda')d\lambda'\)

- **Cramér Von Mises distance**

\[
R_{b,d} = \sqrt{\frac{N}{2^d} \|\hat{F}_{2b,d+1}(\lambda_j) - \hat{F}_{2b+1,d+1}(\lambda_j)\|_2^2}
\]

- **Transformed Cumulative Sums Distance** A transformed cumulative sums test was used in Coates and Diggle(1986) to compare between two estimated spectral densities, which has been used to construct a distance measure between spectra in a block.

\[
Z_{b,d}(\lambda_j) = 1 + \log \left( \frac{\hat{f}_{2b,d+1}(\lambda_j)}{\hat{f}_{2b+1,d+1}(\lambda_j)} \right), j = 0, 1, ..., N/2^{d+1} - 1
\]

are asymptotically independent random samples from an exponential distribution, if block\((b,d)\) is a stationary interval. Based on this, the transformed cusum distance is defined as

\[
R_{b,d} = \frac{N}{2^d} \left[ \max_j \left| \frac{\sum_{k=0}^{j} Z_{b,d}(\lambda_k)}{\sum_{k=0}^{N/2^{d+1}-1} Z_{b,d}(\lambda_k)} - \frac{j}{N/2^{d+1}} \right| \right]
\]

This is similar to the cusum test proposed in Rao(1981).

- **Anderson-Darling Distance** (See Stephens(1974)) It was also shown in Coates and Diggle(1986) that

\[
\overline{Z}_{b,d}(\lambda_j) = \frac{\sum_{k=0}^{j} Z_{b,d}(\lambda_k)}{\sum_{k=0}^{N/2^{d+1}-1} Z_{b,d}(\lambda_k)}, j = 0, 1, ..., N/2^{d+1} - 1
\]

behave approximately as the order statistics of a random sample from the \(U[0,1]\) distribution. So, based on that, the Anderson-Darling distance is

\[
R_{b,d} = \frac{N}{2^d} \left[ \frac{-N}{2^{d+1} + 1} - \frac{2^{d+1}}{N} \sum_{j=0}^{N/2^{d+1} - 1} (2j - 1) \log(\overline{Z}_{b,d}(\lambda_j)) + (N/2^d + 1 - 2j) \log(1 - \overline{Z}_{b,d}(\lambda_j)) \right]
\]
Only discrepancy measures based on overall difference in spectra have been considered here. But, it is possible to use other distance measures such as based on the difference in the mean of the processes.

**Example:** The algorithm was used to find the best segmentation and the corresponding time-dependent spectrum for the piecewise AR(2) model

\[
(1 - 0.9 * e^{i\pi/9} B)(1 - 0.9 * e^{-i\pi/9} B)X_{t,N} = \epsilon_t, \quad t = 1, \ldots, N/2 \tag{1}
\]
\[
(1 - 0.9 * e^{i2\pi/9} B)(1 - 0.9 * e^{-i2\pi/9} B)X_{t,N} = \epsilon_t, \quad t = N/2 + 1, \ldots, N \tag{2}
\]

where \( B \) is the back shift operator: \( BX_{t,N} = X_{t-1,N} \) and \( \epsilon_t \) is white noise. The best segmentation is at \( u = 1/2 \) and the evolutionary spectrum is:

\[
f(u, \lambda) = \begin{cases} 
\frac{1}{1-1.69e^{-i2\pi\lambda}+0.81e^{-i4\pi\lambda}} & \text{if } 0 \leq u < 1/2 \\
\frac{1}{1-1.38e^{-i2\pi\lambda}+0.81e^{-i4\pi\lambda}} & \text{if } 1/2 < u \leq 1
\end{cases}
\]

![Figure 3.1: Combination of AR(2) processes](image)

**Figure 3.1:** Combination of AR(2) processes

Poles of the AR(2) filters

*:* Pole of (1); \( \circ \): Pole of (2)

![Figure 3.2: True Spectra of the two halves](image)

**Figure 3.2:** True Spectra of the two halves

Fig 3.1 shows a realization of 1024 observations (\( N = 1024 \)) from the above piecewise AR(2) model. The true time-dependent spectrum is shown in Fig 3.2, plotted on a grey scale of 0 to 128, showing the relative power at each frequency.

Fig 3.3 - Fig 3.6 show the results of the algorithm applied to the simulated data. The height of the branches in the best segmentation trees represents the change in \( Value \) of the tree, as a result of collapsing that particular branch.
5 Penalized Minimization and Cross-Validating the Tree

In choosing the most stationary partition of the time domain, the segmentation chosen

\[ \text{minimizes} \sum_{\text{segments}} \text{distance}(\hat{f}_{\text{left}}, \hat{f}_{\text{right}}) \]

This might produce a segmentation with very small segments and result in a loss in the frequency resolution of the spectra in the segments. (See Figures 3.3-3.6). Also, it is clear
from the plots that several branches of the best segmentation tree can be collapsed without much of a change in Value, the minimized value of Equation 10.

Thus, having grown the complete tree as described in Section 4, one may opt to penalize on the number of segments, and choose instead to find the optimally pruned subtree, $T^*(\alpha)$, that minimizes

$$\sum_{\text{segments}} \text{distance}(\hat{f}_{\text{left}}, \hat{f}_{\text{right}}) + \alpha(\text{Number of segments})$$  \hspace{1cm} (13)$$

where $\alpha$ is a penalty on the number of segments.

Algorithm: Optimal Pruning with penalties

Step 0: Initialize $\alpha = -\infty$  \hspace{1cm} [# Can be initialized to any starting value]

Step 1: Set $\alpha$ at the current value.
   For $d = (D - 1) : -1 : 0$,
      For $b = 0$ to $2^d - 1$,
         Step 2: Compute $R_{b,d} = \text{distance}(\hat{f}_{2b,d+1}, \hat{f}_{2b+1,d+1})$
         Set $\text{Value}(b,d) = R_{b,d} + \alpha$
         Step 3: If $d = D - 1$, Mark the block $(b,d)$ as terminal
            If $d < D - 1$,
               If $\text{Value}(b,d) \leq \text{Value}(2b,d+1) + \text{Value}(2b+1,d+1)$
                  Mark block $(b,d)$ as terminal
                  Set $R_{\text{sum},b,d} = R_{b,d}$
                  Set $N_{b,d} = 1$
                  Set $g(b,d) = \infty$
               Otherwise, leave block $(b,d)$ unmarked and
                  Set $\text{Value}(b,d) = \text{Value}(2b,d+1) + \text{Value}(2b+1,d+1)$
                  Set $R_{\text{sum},b,d} = R_{\text{sum},2b,d+1} + R_{\text{sum},2b+1,d+1}$
                  Set $N_{b,d} = N_{2b,d+1} + N_{2b+1,d+1}$
                  Set $g(b,d) = \frac{R_{b,d} - R_{\text{sum},b,d}}{N_{b,d} - 1}$
            end if
         end if
      end for
   end for

Final Segmentation, for penalty, $\alpha = \text{Set of highest marked blocks}$

= { block$(b,d)$: block$(b,d)$ is marked and all its ancestors are unmarked } \\
[This is the optimal pruned subtree $T^*(\alpha)$]

Value of the best tree, $R(T^*(\alpha)) = \text{Value}(0,0)$ = minimized value of Equation 13

Step 4: If block$(0,0)$ is marked, STOP
   else Set $\alpha_{\text{new}} = \min \{ g(b,d): \text{block}(b,d) \in \text{Final Segmentation} \}$ and Return to Step 1.
The following result from Breiman et al (1984) shows that the above algorithm gives all the $T^* (\alpha), \alpha \in \mathcal{R}$

**Theorem 5.1** Let the complete tree grown to depth $D$ be $T_o = \{ \text{ node}(b, d); \ b = 0, \ldots, 2^d - 1, \ d = 0, 1, \ldots, D - 1 \}$. Then, there exists a positive integer $K$, real numbers $\{ \alpha_k \}_{k=1}^K$ and subtrees $\{ T_k \}_{k=1}^K$ such that $-\infty < \alpha_1 < \ldots < \alpha_K < \infty$ and $T_o < T_1 < \ldots < T_K = \text{ block}(0, 0)$ where

$$
T^* (\alpha) = \begin{cases} 
T_o & \text{if } \alpha < \alpha_1 \\
T_k & \text{if } \alpha_k \leq \alpha < \alpha_{k+1}, \ 1 \leq k \leq K \\
T_K & \text{if } \alpha \geq \alpha_K
\end{cases}
$$

and $\alpha_{k+1} = \min \{ g(t, T_k) : t \text{ is a non-terminal node of } T_k \}$, where $g(\cdot, \cdot)$ is as defined in the algorithm.

Fig 4 below shows the plots of the *Value* of the best trees, $R(T_k)$ versus the size of the trees, for the piecewise AR(2) example of Section 4.

![Plot 1](image1.png)

![Plot 2](image2.png)

![Plot 3](image3.png)

![Plot 4](image4.png)

### 5.1 Choice of the penalty parameter, $\alpha$

Let $T^* (\alpha)$ be the optimally pruned subtree with respect to the penalty parameter $\alpha$. It was shown in Theorem 10.9, Breiman et al (1984) that if $\alpha_2 \geq \alpha_1 \geq 0$, then $T^* (\alpha_2)$ is a subtree of $T^* (\alpha_1)$. So, increasing $\alpha$ leads to greater penalty for complex trees and hence in same or smaller $T^* (\alpha)$.

A small value of $\alpha$ could result in very small segments and hence to a loss in the frequency resolution, i.e. high bias and variability in the estimated spectrum of the segment. However, increasing $\alpha$ results in coarser segmentation, which implies that the series is considered to be approximately stationary in these wide segments. So, any nonstationarities occurring during the interval can cause a nonstationarity bias in the estimated time-dependent spectrum. An optimal choice of the penalty parameter is required to tradeoff between frequency resolution and nonstationarity bias. The choice of $\alpha$ is usually made by some resampling strategy and cross-validation. However, in the situation of time series, naive resampling is not advisable due to the correlation of the observed data. (Note that resampling observations from an
autoregressive series will not yield an autoregressive process.

Cross-Validation: Because of the correlations in the observed time series, it is not advisable to resample from the data but rather to resample from \( \hat{f}_{b,d}(\lambda) \)'s, which has shown to be at least asymptotically independent in Dahlhaus (1993). Based on this observation, we describe a half-sample cross-validation procedure.

Restricting to the nonnegative choices of \( \{\alpha_k\}_{k=1}^K \) for the penalty parameter as given in Theorem 5.1,

Step 1: For \( d = D - 1 : 0 \),
    
    For \( b = 0 \) to \( 2^d - 1 \),
    
    Define for block \((b,d)\)
    
    \( \hat{f}_{b,d}^{\text{even}}, \hat{f}_{b,d}^{\text{odd}} \): resample all even (respectively odd) indexed frequencies from \( \hat{f}_{b,d} \).

    \( R_{b,d}^{\text{even}} = \text{distance}(\hat{f}_{2b,d+1}^{\text{even}}, \hat{f}_{2b+1,d+1}^{\text{even}}) \)
    
    \( R_{b,d}^{\text{odd}} = \text{distance}(\hat{f}_{2b,d+1}^{\text{odd}}, \hat{f}_{2b+1,d+1}^{\text{odd}}) \)

Step 2: [# Since the optimal tree \( T^*(\alpha) \) equals \( T_k \) for \( \alpha_k \leq \alpha < \alpha_{k+1} \), a geometric midpoint of the interval \([\alpha_k, \alpha_{k+1}]\) is considered.]

Let \( \alpha_k' = \sqrt{\alpha_k \alpha_{k+1}} \), if \( \alpha_k \geq 0 \), \( k = 1, \ldots, K \)

Step 3: Find \( T_{\text{even}}^{\alpha_k'}, T_{\text{odd}}^{\alpha_k'} \), the optimal pruned subtrees, with penalty parameter \( \alpha_k' \), based on \( R_{b,d}^{\text{even}} \)'s and \( R_{b,d}^{\text{odd}} \)'s respectively.

\[
R^{CV}(k) = \frac{1}{2} \cdot \left[ R^{\text{odd}}(T_{\text{even}}^{\alpha_k'}) + R^{\text{even}}(T_{\text{odd}}^{\alpha_k'}) \right].
\]

where \( R^{(i)}(T(\alpha)) \) is the value of Equation 13 with distance defined as \( R^{(i)} \).

Step 4: Choose the Final \( \alpha = \alpha_{k_0} \), where

\[
k_0 = \min_k \{ R^{CV}(k) + \alpha_{\text{fin}} \cdot \text{Number of terminal blocks} \}
\]

and \( \alpha_{\text{fin}} \) is a small number such as 0.05*\( \min_k R^{CV}(k) \). [# In the final tree selection, Breiman et al (1984) recommend using a 1 SE rule or a rule as outlined in Step 4 as an allowance for simplicity as well as accuracy.]

Example Continued

The Cross Validation method outlined here was used to find the best segmentation for the example of a piecewise AR(2) process with a change point at 1/2. The results are given in Figures 5.1-5.4 along with the optimal value of the penalty \( \alpha \) chosen and the corresponding
value of $R(T^*(\alpha))$.

![Figure 5.1: Cross-Validated Estimate using Kolmogorov-Smirnov Distance](image)

![Figure 5.2: Cross-Validated Estimate using Cramer Von-Mises Distance](image)

![Figure 5.3: Cross-Validated Estimate using Transformed Cusum Distance](image)

![Figure 5.4: Cross-Validated Estimate using Anderson-Darling Distance](image)

6 Simulation Results, Examples and Applications

The algorithm of Section 5 was used in the following examples to determine the best segmentation (after cross-validation). The optimal choices of $\alpha$ and the "value" are shown in the figures. As the estimate of the spectrum in each "stationary interval" was taken to be the periodogram, the time-frequency pictures exhibit the characteristic roughness of the periodogram along the frequency axis.
6.1 Piecewise Stationary Processes

A random process \( \{X_{t,N}\} \) is piecewise stationary, with fixed change points \( 0 = u_0 < u_1 < \ldots < u_{m+1} = 1 \), if

\[
X_{t,N} = \sum_{k=1}^{m} I(u_k \leq t/N < u_{k+1})X^{(k)}_t
\]

where \( X^{(k)}_t, k = 0,1,\ldots, m-1 \) are independent, zero-mean stationary processes, with spectra \( f^{(k)}(\lambda) \). Then, it can easily be shown that a piecewise stationary process \( X_{t,N}(t = 1,\ldots,N) \) is locally stationary (as given in Definition 3.1) at any time \( u, 0 \leq u \leq 1, u \neq u_0,\ldots,u_{m+1} \) and the time-dependent spectrum of the process at time \( u \) is

\[
f(u, \lambda) = \sum_{k=1}^{m} I(u_k \leq t/N < u_{k+1})f^{(k)}(\lambda)
\]

Simulation of 100 runs with \( N = 1024 \)

<table>
<thead>
<tr>
<th>Definition of Process</th>
<th>Number of Terminal Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Penalty on # of blocks</td>
</tr>
<tr>
<td>white noise, ( \varepsilon_t )</td>
<td>1 node:59%, 2 nodes:26%</td>
</tr>
<tr>
<td>Piecewise AR(2) process</td>
<td>1 node:0%, 2 nodes:36%</td>
</tr>
<tr>
<td>( a_1(t/N) = -1.32, a_2(t/N) = 0.49, t \leq N/2 )</td>
<td>3-4 nodes:25%, ( \geq 5 ) nodes:39%</td>
</tr>
<tr>
<td>( a_1(t/N) = -1.07, a_2(t/N) = 0.49, t &gt; N/2 )</td>
<td>1 node:0%, 2 nodes:30%</td>
</tr>
<tr>
<td>( a_1(t/N) = -1.32, a_2(t/N) = 0.81, t \leq N/2 )</td>
<td>3-4 nodes:30%, ( \geq 5 ) nodes:40%</td>
</tr>
<tr>
<td>( a_1(t/N) = -1.32, a_2(t/N) = 0.49, t \leq N/2 )</td>
<td>1 node:0%, 2 nodes:32%</td>
</tr>
<tr>
<td>( a_1(t/N) = -1.07, a_2(t/N) = 0.81, t &gt; N/2 )</td>
<td>3-4 nodes:30%, ( \geq 5 ) nodes:38%</td>
</tr>
</tbody>
</table>

The simulation study with piecewise stationary processes suggests that the method, if anything errs on the side of being conservative i.e. it may divide even if the segment is stationary rather than use a nonstationary segment of the data to estimate a stationary spectrum.

6.2 Modulated Stationary Processes

A class of nonstationary random processes which are often encountered in seismological applications is the class of modulated stationary processes — \( X_{t,N} = A(t/N)X_\circ(t), t = 1,\ldots,N \) where \( X_\circ(t) \) is a stationary random sequence with spectrum \( f_\circ(\lambda) \), and \( A(u) \) is a continuous function on \([0,1]\).

Then, it easily follows that \( X_{t,N} \) is locally stationary at any time \( u, 0 \leq u \leq 1 \) and the evolutionary spectrum of the process at time \( u \) is

\[
f(u, \lambda) = |A(u)|^2 f_\circ(\lambda)
\]

Example: This is an example used in Section 6.6, Priestley(1988).

\[
X_{t,N} = e^{-\frac{2\pi}{N}(N-\frac{1}{2})^2}X_\circ(t), t = 1,\ldots,N
\]
where \( X_0(t) = 0.8 * X_0(t - 1) - 0.4 * X_0(t - 2) + \epsilon_t \)

in which the \( \epsilon_t \) are independent \( N(0,100^2) \) variables. The results of the simulation are shown in Figures 6.1-6.2.

![Figure 6.1: Modulated Stationary Process](image)

![Figure 6.2: Estimate of the time-dependent spectrum using Kolmogorov-Smirnov Distance](image)

The best segmentation, found with no penalty on the number of segments was found to divide the data into 33 segments. The best segmentation found by cross validating the penalty on the number of segments as described in Section 5 reduced the number of segments to 16.

### 6.3 Time-Varying ARMA(p,q) processes

For the time-varying ARMA(p,q) models,

\[
\sum_{j=0}^{p} a_j \left( \frac{t}{N} \right) X_{t-j,N} = \sum_{k=0}^{q} \sigma \left( \frac{t}{N} \right) b_k \left( \frac{t}{N} \right) \epsilon_{t-k}
\]

the time-dependent spectrum as given in Definition 3.1 is (see Dahlhaus(1993), Theorem 4.1)

\[
f(u, \lambda) = \left| \sigma(u) \sum_{k=0}^{q} b_k(u) \exp^{-i2\pi\lambda k} \sum_{j=0}^{p} a_j(u) \exp^{-i2\pi\lambda j} \right|^2
\]

The following example of a time-varying AR(2) process has been adapted from Dahlhaus(1993).

\[
(1 - 0.9 * e^{i(0.5 - \cos(\frac{\pi}{2}))} B)(1 - 0.9 * e^{-i(0.5 - \cos(\frac{\pi}{2}))} B)X_{t,N} = \epsilon_t,
\]

22
where \( \epsilon_t \) is white noise. The top panel of Figure 7 shows the simulated time-varying AR(2) process and its true time-dependent spectrum. The results with simulated data (after cross-validation) are presented in Figure 7.

![Time-Varying AR process](image1)

![True time-dependent spectrum](image2)

![Basis Tree: Time-Varying AR process](image3)

![Estimated time-dependent spectrum](image4)

**Figure 7:** Time-Varying AR(2) process – Estimate of the time-dependent spectrum using Kolmogorov-Smirnov Distance

### 6.4 Analysis of Speech Signals

Automatic recognition of speech belongs to the broader category of pattern recognition tasks, for which during the past 30 years or so, many heuristic and sophisticated methods have been tried. One of the first steps in analyzing speech fragments is to segment the signal into phonemes. Phonemes are the sets of smallest units of speech that serve to distinguish one utterance from the next. The phonemes also have the property that they have rather well-defined, stationary spectra. Rather than use an ad hoc segmentation based on the user’s ability to recognize phonemes, it is possible to use the tree algorithm to find an optimal segmentation.

The top panel of Figure 8 shows a speech recording of the word “greasy”, sampled at 16kHZ, which was analyzed using matching pursuit in Mallat & Zhang(1993). The best segmentation without any penalty produced 58 segments. The optimal choice of \( \alpha \) found using cross-validation was 0.0027, which resulted in 17 segments.
Figure 8: Analysis of Speech Signal: Greasy
The plot of the time-dependent spectrum shows the low-frequency component of the "g", the quick transition to "ea", the harmonics of "ea" and "y". The "s" component exhibits its characteristic spectrum with energy spread over a high-frequency interval.

6.5 Applications to Seismology

During an earthquake, it is a well known fact that different frequency components travel with different velocities. Various authors (see Bullen and Bolt(1985)) have determined the dispersion relation for simple structural models of the earth's surface. These dispersion relations give the relation between frequency and velocity.

A fixed window dynamic spectral analysis was used in Brillinger(1993) to obtain a non-parametric estimate of the functional relation of frequency and velocity. The adaptive tree-based algorithm can be used to first obtain a time-dependent spectrum for the non-stationary process which can then be used to estimate the relation between frequency and velocity.

The optimal tree-segmentation algorithm was applied to the seismogram data from the Pasadena station during the 7.0 magnitude Loma Prieta earthquake on Oct. 18, 1989. (Long Period data is being used here - i.e. 1 sample/sec) The earthquake trace is shown in the top panel of Figure 9 and the resultant time-dependent spectrum obtained using the algorithm is shown below.
Figure 9: Loma-Prieta Earthquake Data – Estimate of the
time-dependent spectrum using Kolmogorov-Smirnov Distance
The best segmentation without any penalty produced 26 segments. The optimal choice of
$\alpha$ found using cross-validation was 0.00296, which resulted in 11 segments.

7 Concluding Remarks

Pruning to a non-dyadic segmentation:

As was remarked earlier, the best segmentation can be achieved in $O(N \log^2 N)$ operations. Thus, the time-dependent spectrum is estimated in $O(N \log^2 N)$ time, which is comparable to the $O(N \log N)$ time required to estimate the spectrum of a stationary time series.

The price that one has to pay for the fast algorithm is that the segmentation is restricted to dyadic points. As a result, very small segments may occur if the change does not occur at a dyadic point.

Example Continued: Consider the piecewise AR(2) model of Section 4 with a change point at $1/3$. The ideal segmentation that one can hope to achieve is the one that corresponds to the dyadic expansion of $1/3$ (see top panel of Figure 10). The result with simulated data in Figure 10 shows the best segmentation obtained after cross-validation.

As most of the computation time is spent in estimating the spectrum in blocks, it appears preferable to use a fast algorithm and then recombine segments once the best dyadic segmentation has been obtained.
Figure 10: Piecewise AR(2) with a change point at 1/3

After obtaining a sparse segmentation by cross-validation, an optional recombination of segments can be made by comparing every segment with its adjacent one, rather than just segments occurring in the binary tree.

Smoothing the Evolutionary Periodogram:

The estimate of the time-dependent spectrum, the **evolutionary periodogram** obtained by the tree-segmentation method is not a consistent estimator of the time-dependent spectrum though asymptotically unbiased. Thus, it exhibits the same defects as the periodogram for a stationary time series.

Because of the erratic behavior of the periodogram, it is a standard practice to smooth the periodogram (or the log periodogram). Various methods of smoothing have been proposed to get a smoothed spectrum estimate for a stationary time series. Any such method can be applied to the **stationary segments** obtained by the tree-segmentation method.

In favor of computational simplicity, it was chosen to smooth the spectrum in each **stationary segment** by a simple moving-average filter — (See Brockwell et al (1991)). So, if block(b,d) is in the final segmentation, and $\hat{f}_{b,d}(\lambda_j)$ is the estimated spectrum in block(b,d) at frequencies $\lambda_j = 2\pi j.2^d/2n$, $j = 0, 1, ..., n/2^d - 1$, then

\[
\text{Smoothed } \hat{f}_{b,d}(\lambda_j) = \sum_{|k| \leq m_d} \frac{1}{2m_d + 1} \hat{f}_{b,d}(\lambda_{j+k})
\]

where $m_d$ is chosen so that for fixed depth $d$, $m_d \to \infty$ and $m_d/N \to 0$ as the size of the series $N \to \infty$. The resultant smoothed time-dependent spectrum, obtained after cross-validated tree-segmentation for the Example of Figure 4.1-4.6 is shown here.
Figure 11: Smoothed time-dependent spectrum for the combination of AR(2) processes

Unit Root nonstationarity: The procedure of this paper is geared towards the estimation of slowly changing processes as is characterized by Definition 3.1. Another kind of nonstationarity that is very common in the economics literature is the kind that comes from ARMA processes with unit roots or roots close to unity. This produces long range dependencies in the data. However, ARMA processes with unit roots are not included in the class of locally stationary processes. It can be easily shown that for a random walk

\[ X_{t,N} = X_{t-1,N} + \varepsilon_t, \varepsilon_t \sim N(0,1) \]
\[ A_{t,N}(\lambda) = \frac{1 - e^{i2\pi\lambda t}}{1 - e^{i2\pi\lambda}} \]

There is no function \( A: [0,1] \times [-1/2, 1/2] \rightarrow C \) such that \( A_{t,N}(\lambda) \rightarrow A(u, \lambda) \) as \( N \rightarrow \infty \) and \( t/N \rightarrow u \).

The segmentation algorithm applied to such a series would tend to segment the data to the greatest depth in the tree. The simulated random walk in Figure 12 produced 65 segments without penalization and the optimally penalized tree had 19 segments.

Figure 12: Segmentation for Random Walk

Bias-Variance trade-off: In estimating the time-dependent spectrum, one of the major issues is the trade-off between bias and variance. If the segmentation used is too fine, this might improve the bias but cause the variance to increase highly. On the other hand, a coarse segmentation may decrease the variance, but increase the nonstationarity bias that arises from considering large segments of the data to be stationary. The optimal choice of the penalty parameter \( \alpha \), determined by cross-validation is used to find the best bias-
variance trade-off.

The framework of spectral analysis of nonstationary time series here is based on a Cramér-like representation. Tree-based estimation procedures allow for an adaptive estimation of time-dependent spectra of locally stationary processes. Thus, this procedure provides for a method of analyzing nonstationary time series in the frequency domain. It provides a great deal of insight for real data applications such as in speech processing and seismology. Simulation results provide substantial empirical evidence of the validity of this method for detection of nonstationarity.

Available Software: The author has written some software (in matlab) that has a button interface that allows for an interactive use of the algorithm. Interested users may request a copy by email.

Acknowledgement

The author would like to thank Iain M. Johnstone for his helpful suggestions and incisive comments.

References


