FITTING MODELS OF NONSTATIONARY TIME SERIES: AN APPLICATION TO EEG DATA

by

SREENIVAS KONDA

Submitted in partial fulfillment of the requirements
For the degree of Doctor of Philosophy

Dissertation Advisor: Dr. Wojbor A. Woyczynski

Department of Statistics

CASE WESTERN RESERVE UNIVERSITY

August 2006

CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the dissertation of

SREENIVAS KONDA

candidate for the Doctor of Philosophy degree

Committee Chair: _____
Dr. Wojbor A. Woyczynski
Dissertation Advisor
Professor,
Department of Statistics

Committee: _____
Dr. Jiayang Sun
Professor,
Department of Statistics

Committee: _____
Dr. Joe Sedransk
Professor,
Department of Statistics

Committee: _____
Dr. Kenneth Loparo
Professor,
Department of Electrical Engineering & Computer Science

August 2006

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the encouragement and guidance of many individuals.

I would like to express my very deep gratitude to my adviser, Professor Wojbor Woyczynski, who guided me at each and every aspect of this dissertation and my graduate studies at Case. Wojbor has always listened to me patiently and encouraged me to ask questions - even the silliest! I value his constructive criticisms of my work and writing. He taught me the art being able to apply statistics and probability to cutting edge research. Wojbor showed me that working hard in academe can be rewarding and satisfying. It still amazes me how much work was accomplished in last few months of my research with Wojbor. I could not have asked for a better mentor and a teacher! I am looking forward to more collaboration in the future.

I would like to thank Professor Jiayang Sun for helping me all these years at Case personally and professionally and also sharing her insight into the applied statistics. She patiently and clearly answered all my questions and helped me better understand theory.

I would like to thank Professor Joe Sedransk for being a good teacher and being available for help. Despite the late notice, he still agreed to serve as a member of my dissertation committee.

I would also thank Professor Kenneth Loparo for being great help all these years at Case. He provided the crucial data to support this dissertation.

I also wish to express my gratitude to Professor Gy Terdik of University of Debrecen, Hungary. Without his help this dissertation never would have been started. He helped me to jump start this research by guiding me through theory and programming.

I also would like to thank the faculty, staff and fellow students. In particular, I would like to thank Professor James Alexander, Sharon Dingess, Professor Avi Dor, Ariadni Papana, and Dr. Alexandra Piryatinska.

My deepest gratitude goes to my parents and my sisters for their total and unwavering support during my graduate studies. They taught me by their love and living example that all things are possible with hard work and discipline. Most of all, they taught me to live humble and sincere.

Fitting Models of Nonstationary Time Series:
An Application to EEG Data

Abstract

by

Sreenivas Konda

A computationally efficient algorithm is presented for fitting models to a nonstationary time series with an evolutionary (time-varying) spectral representation. We formally define time-varying memory process and prove that this process satisfies local stationarity definition. Our procedure segments the EEG nonstationary time series into stationary or approximately stationary blocks, with and without overlapping, and then estimates the time varying parameters using the local stationarity concept. Our estimation procedure does not make any assumptions about the distribution of innovations (data generating process). We also present a systematic procedure to separate the short memory part from the nonstationary long memory part of the test example time series using a simple frequency domain procedure. Our method is simple and efficient compared to the currently available procedures to analyze the EEG data. Using our procedure, we present a thorough analysis of the sleep EEG data of fullterm and preterm neonates. Several extensions of our method to multivariate time series are also proposed.

Chapter 1

Motivation

The assumption of covariance stationarity or simply stationarity (Chapter 2) plays an important role in modeling the time series whose statistical properties do not change with time. This assumption is very useful in order to develop asymptotic properties of the estimators derived from the second order statistical properties of a time series. However, in many cases, time invariance is difficult to justify over a long period of time but provides a welcome simplification.

Nevertheless, in reality, many time series are not covariance stationary and exhibit a *time-varying* or *evolutionary* second order structure (Preistley, 1965). For example, the neonates' sleep ElectroEncephaloGram (EEG) signals in Section 1 of this chapter oscillate with greater amplitude with higher frequency during a particular state of sleep cycle compared to the other states and, also, in adults, these signals oscillate with greater amplitude and at higher frequency during an epileptic seizure than before or after seizure (Ombao, 1999). Thus, the distribution of power over frequencies changes as time changes in these time series.

For the last few decades, more and more attention has been paid to the modeling of such time series with evolutionary autocovariance structure. Priestley (1965) introduced a class of nonstationary processes based on the

concept of evolutionary spectra. The evolutionary spectral density functions are time dependent and generalize the usual definition of spectra for stationary processes.

In addition to the pioneering research on time-varying spectra of nonstationary processes (cf. Preistley, 1981), a decisive idea was recently introduced by Dahlhaus (1996, 1997) with the concept of local stationarity (cf. Chapter 3). This concept allows the modeling of a time-varying autocovariance structure which can be estimated rigorously. By this, we mean that an asymptotic theory can be developed and the usual statistical properties of estimators may be derived for certain type of time series called time-varying AutoRegressive and Moving Average (ARMA) processes (Chapter 3).

Dahlhaus (1997) applied his locally stationary concept to fit a time-varying autoregressive (AR) time series by using polynomials in time to derive *smooth* estimates of the time-varying parameters. Ombao (1999) used Dahlhaus's definition of local stationarity concept to develop the so-called Smooth Localized complex Exponential (SLEX) basis functions which are orthogonal and localized in both time and frequency domains and applied them to analyze bivariate epileptic EEG. Fourier basis functions which are used in Dahlhaus (1997) are localized only in frequency domain, unlike the SLEX basis functions. Van Bellegem (2003) also used Dahlhaus's definition of local stationarity concept to fit finance and environmental time series using Wavelets. However, Ombao and Van Bellgem applied their research to only small data sets and the results of both these methods are difficult to interpret.

This thesis is different from Ombao (1999) model fitting of epileptic EEG because of three reasons:

(i) The sleep EEG statistical properties are different from the epileptic EEG in Ombao research;

(ii) The model fitting in our case is done in two steps compared to Ombao

model fitting;

(iii) Our model fitting analysis is thorough and includes residual analysis, as you will see in Chapter 5.

This thesis contributes to the development of the local stationarity approach in the real applications of model fitting of a nonstationary time series. Herewith, by nonstationarity, we always refer to time-varying autocovariance structure, with, possibly zero mean.

The test example in this research is a nonstationary sleep EEG of neonates with possibly *infinite variance* innovations (data generating process). We use the time-varying second order properties of this signal in the model fitting procedure using Dahlhaus (1997) methodology to fit the time varying AR models with Gaussian innovations. However, our algorithm uses an approximate *Whittle* likelihood function suggested by Mikosch et.al (1995) for a stationary ARMA models with infinite variance innovations and generalizes it to time-varying model fitting framework. This approximate likelihood function suggested by Mikosch et. al (1995) is also the likelihood function of interest in Kokoszka and Taqqu (1996) to estimate the parameters of *fractional* Integrated ARMA (ARIMA) time series with infinite variance innovations. The model we consider to fit the sleep EEG is time-varying ARIMA model with nonstationary fractional or memory parameter and also with possibly infinite variance innovations (Chapter 5).

The EEG signal is of primary importance in interpreting *polysomnographic* studies associated with sleep disorders in human subjects, including neonates (cf. Scher, 1998). Parameterizing the sleep EEG of neonates helps to assess brain maturity, dysfunction, and developmental alterations during the early life. The knowledge of sleep EEG models in healthy fullterm and preterm infants facilitates a greater understanding of functional brain development that may have prognostic implications. Neurophysiologic interpretation of sleep EEG models will help a pediatric neurologist to identify

those neonates who may be at greater risk for neurologic diseases and their aftereffects.

To the best of our knowledge, this research is first of its kind in modeling the sleep EEG data of human subjects. However, there is a vast amount of literature [cf. Cohen (1989), Williams, Zaveri, and Sackellares (1995), Ombao (1999), and the references there in] on the time-varying analysis of nonstationary non-sleep EEG signals, including the modeling the epileptic seizure EEG. A search for *time-varying frequency analysis of EEG* on Google search engine resulted in more than 100 articles.

The existing methods for analyzing any nonstationary signal divide it into approximately stationary blocks of equal sizes and then compute the tapered Fourier periodograms of these stationary blocks [cf. Cohen (1989) and Williams, Zaveri, and Sackellares (1995)]. Ombao (1999) considered unequal stationary block sizes in the model fitting of seizure EEG. In our analysis and model fitting of sleep EEG of healthy neonates, we consider stationary blocs of equal sizes because the transition between sleep states of healthy neonates tends to be regular and approximately at fixed intervals. We analyze only the EEG of healthy neonates.

In the following section we provide the motivation for the analysis and model fitting of the the sleep EEG of healthy neonates and introductory analysis.

1.1 Exploratory Data Analysis

This research is motivated by the need of neurologists to study and estimate the time-varying spectrum and model fitting of nonstationary ElectroEncephaloGram (EEG) time series. The time-varying spectral analysis and model fitting of this data can provide a better understanding of the evolution of the brain via, for example, a study of the sleep cycles of the fullterm

(born after the full term) and *preterm* (prematurely born) neonates.

Our interest is the modeling of a single channel EEG data collected during the sleeping period of neonates. The ultimate goal is to find some specific statistical characteristics of sleep data in order to evaluate the brain maturation levels in neonates and to provide medical diagnostics. But the primary focus of this dissertation is the modeling of the neonates' sleep EEG data and not necessarily the discrimination of the preterm born neonates from the fullterm born neonates. Occasionally we do consider both data sets and discuss the differences between the modeling of these two data sets but not with regard to the discriminatory analysis. Before we further discuss the modeling of EEG data, we examine two examples, one for a fullterm and another for a preterm neonates by studying some of their statistical characteristics.

The analyzed fullterm and preterm neonates EEG data was provided by Dr. Mark Scher, a pediatric neurologist at the Rainbow Babies and Childrens' Hospital in Cleveland. The example data comes from a massive 25 channel EEG signal collected with a sampling frequency of 64HZ from Pittsburgh area neonates. The approximate length of the EEG signal is 6 to 8 hours.

EEG signals or brain waves are the electrical potentials recorded by electrodes placed on the scalp of subjects, such as neonates. Each EEG channel signal is the difference in electrical potential between two recording electrodes placed on the scalp. For an example, the C3-FP1 channel records the voltage difference between C3 (left central lobe) and FP1 (left frontal parietal lobe) scalp locations. It is understood that this brain region of neonates is relatively *well* developed compared to other regions. This is the channel of interest in this dissertation.

Along with EEG signals, we were also provided with manual scoring of sleep states of the first 2-3 hours EEG signal. The minute by minute scoring data was recorded manually by Dr. Scher. The following summary of sleep

states is taken from the *http://www.sleepdisorderchannel.net/*

Since the early 20th century, human sleep has been described as a succession of Rapid eye movement (REM) sleep and four non-REM sleep stages. The REM sleep is marked by extensive physiological changes, such as accelerated respiration, increased brain activity, eye movement, and muscle relaxation. People dream during REM sleep, may be as a result of excited brain activity and the paralysis of major voluntary muscles.

Sleep quality changes with transition from one sleep stage into another. Although the signals for transition between stages of sleep are mysterious, these stages are, in fact, discretely independent of one another, each marked by subtle changes in bodily function and each part of a predictable cycle whose intervals are observable. Sleep stages are monitored and examined clinically with polysomnography, which provides data regarding electrical and muscular states during sleep.

In general, the four non-REM sleep states of neonates can be recognized around or after the 36th week of post-conceptional age. These four sleep stages of neonates' sleep cycle are well defined along with some other undefined sleep states (OTH) and they are

- *Mixed frequency active* (MFA) sleep state usually begins as a sleep cycle following falling asleep. This sleep stage is characterized by stationary and continuous signals with frequencies spread across the human EEG spectrum (with most of the signal power concentrated between 0.5 and 10 Hz).

- *Low voltage irregular* (LVI) active sleep state is characterized by relatively lower amplitude signals at broadband higher frequency distributions.

- *High voltage slow* (HVS) quiet sleep state appears following the mixed

frequency active sleep, and it is a brief state, characterized by a shift in the frequency distribution to the higher power in the lower frequencies;

- *Trace alternate* (TAQ) quiet sleep is characterized by non-stationary signals of alternating broad band bursts of the activity with intermittent epochs of relative EEG quiescence, comprised over lower amplitude and lower frequencies.

The percentage of REM sleep is highest during infancy and early childhood, drops off during adolescence and young adulthood, and decreases further in older age. The stage-respective dimensions of sleep change relative to age. Stages HVS and TAQ in the sleep cycle shorten even more dramatically in older people than they do during a typical night for everyone else, so older people get less total deep sleep than younger people do. Also with age comes the lengthening of the first REM stage. Older people commonly enter REM sleep quicker and stay there longer.

The non-REM sleep comprised of 4-sleep stages lasts about 120 minutes, each stage lasting anywhere from 5 to 15 minutes. Surprisingly, however, stages HVS and LVI repeat backwards before REM sleep is attained. So, a normal sleep cycle, after falling sleep has this pattern: MFA, LVI, HVS, TAQ, HVS, LVI, REM.

Our exploratory data analysis includes only one example in each of full-term and preterm neonates groups. In the following section, for the sake of simplicity we consider only one-minute segment of the data. These one-minute segments do not characterize hours-length of EEG data sufficiently but they provide the motivation and illustration for our research. Unless stated, all the computing and plots in this thesis were provided using *matlab*.

1.1.1 A fullterm EEG

The EEG sleep data of one healthy fullterm neonate is analyzed in this section. Here we present the initial data analysis of a 1-minute segment of *C3-FP1* channel EEG. The sleep state of this data is classified as MFA. The sampling frequency of this data collection is 64HZ, hence we have 3840 observations. The original time series plot and a magnified portion of this data between 20 and 30 seconds plot are provided in Figure 1.1. We also provided in Figure 1.2 the sample AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) plots [definitions of these functions can be found in Brockwell and Davies (1996), Chapter 3]. The skewness of this data is -0.3866 and the kurtosis is 3.7654.

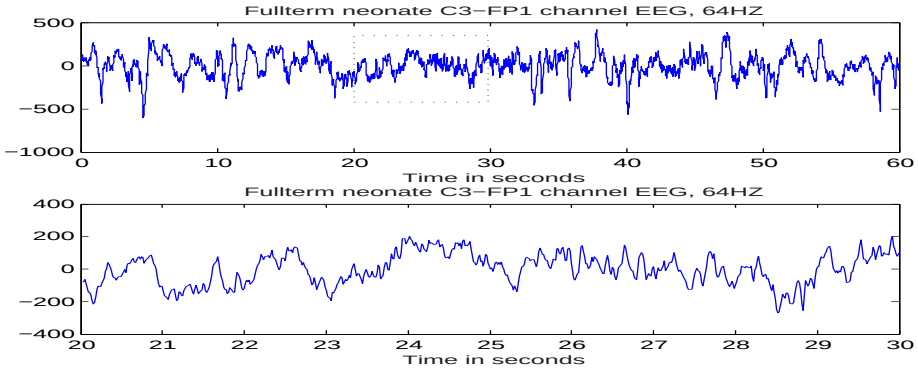

Figure 1.1: *Fullterm EEG time series.*

Observations of the time series plots indicate some local trends, most likely linear trends. The slow decay of ACF plot indicates long range dependence or *long memory* and the PACF plot (the PACF function value at lag one is close to 1) indicates nonstationarity caused, perhaps, by the stochastic trend of integration of order 1 [cf. Box and Jenkins (1970), p.174]. Such time series could be transformed to a stationary process by differencing it once [cf. Brockwell and Davies (1996), Chapter 9]. The long range dependence of the

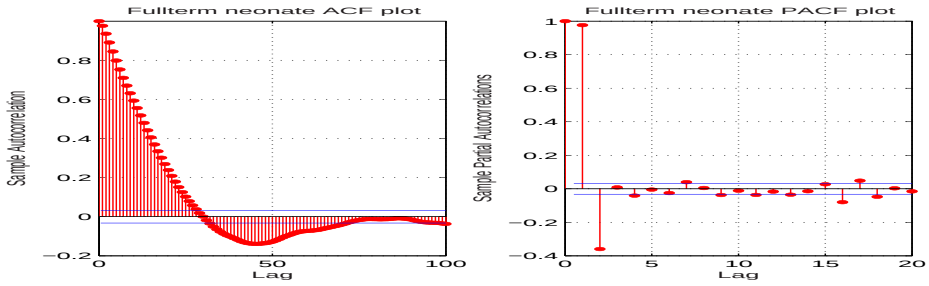

Figure 1.2: *Fullterm ACF and PACF.*

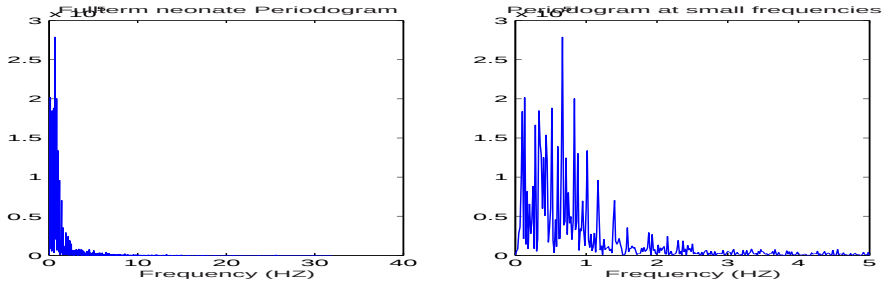

Figure 1.3: *Fullterm periodogram.*

data can also be seen in the periodogram plot and the magnified periodogram plot around zero frequency, Figure 1.3. See Chapter 2 for the definition of Periodogram. Proper definition of the long memory of a time series will be provided in Chapter 2 but for now, this property can be understood by the persistence of strong autocorrelations over a long range of observations, or by the extremely large periodogram values as frequency approaches zero.

To track the non constant second order properties over time we segmented the original time series using two second windows of length 128 observations in each, with no overlap between the windows and with no tapering and plotted the periodograms of these successive segments as a surface in 3.D. (see Figure 1.4). We call this plot the time-varying (tv) or evolutionary periodogram plot. Along with this plot we also provided a sliced version of tv-periodogram truncated at the 1000 level. This trucated plot cuts off

the original periodograms around the small frequencies and provides a better visualization of its local behavior at higher frequencies.

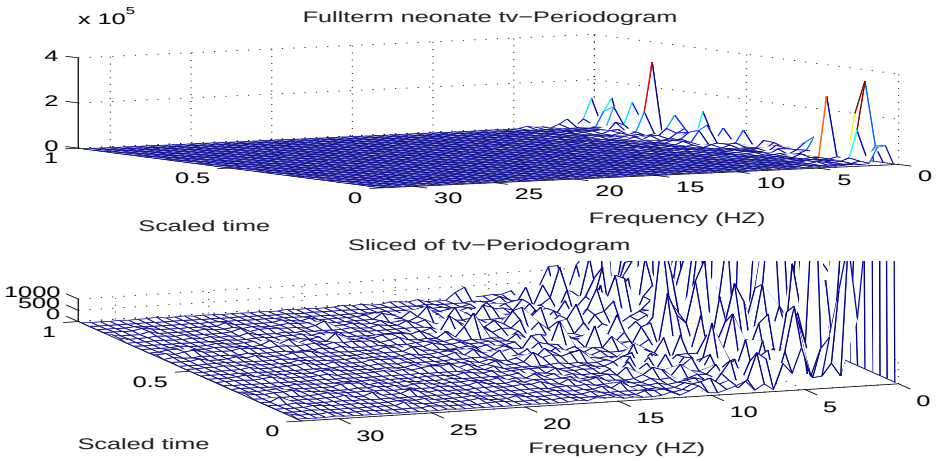

Figure 1.4: *Fullterm tv-periodogram.*

The tv-periodogram plots also show long memory properties of different segments. We can also see the variability in the long memory over different segments. The peaks around zero frequency and patterns of periodograms around small frequencies are not the same for all segments. This observation may suggest nonconstant memory parameter over the entire length of EEG data. The sliced periodogram captures changes in the time-varying second order characteristics of the original time series over the different segments and around the non-zero frequencies.

Since we suspect nonstationarity of integration of order one, we differenced the original data once and plotted all the above graphs for the differenced data again. These plots are provided in Figures 1.5 and 1.6. Clearly, we see lesser local trends in these plots. But more importantly the differenced data no longer shows the long memory property that we have seen in the original data. One more critical observation (which is not obvious from the time series plot) is that this differenced time series plot shows another type

of nonstationarity, non constant variance over the time. (This observation is more obvious if more than few minutes segments of data are analyzed).

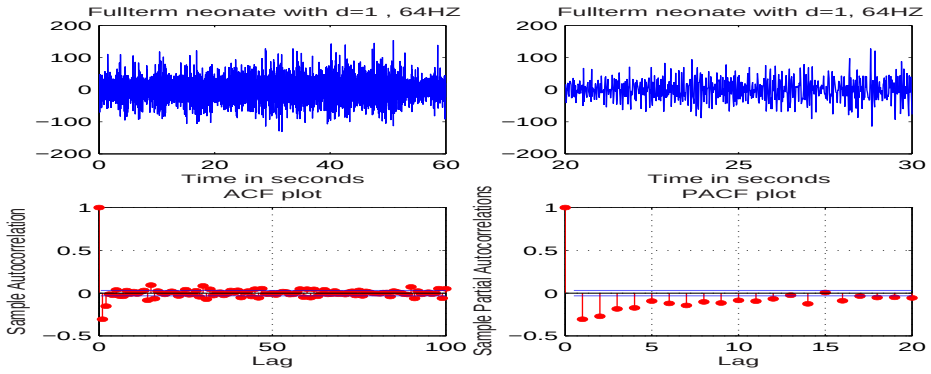

Figure 1.5: *Differenced time series with ACF and PACF plots.*

The periodogram and the tv-periodogram of the differenced data are plotted in Figure 1.6. The long memory property is no longer present in the individual segments and the tv-periodogram plot again captures changes in the second-order characteristics of the differenced time series over the different segments.

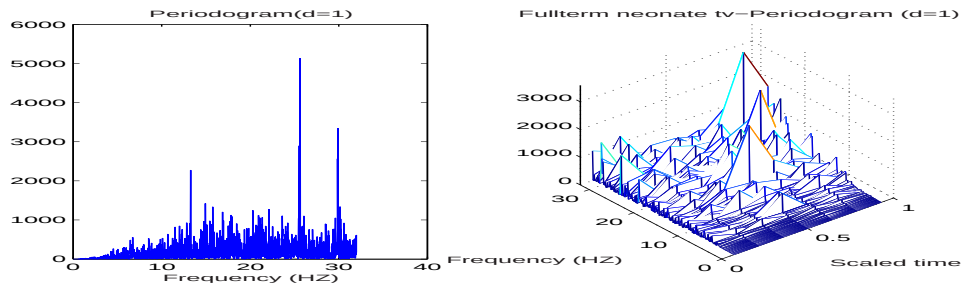

Figure 1.6: *Differenced data periodogram and tv-periodogram.*

1.1.2 A preterm EEG

The following one-minute segment shown in Figure 1.7 was taken from the *C3-FP1* channel EEG data of a prematurely born neonate (28 weeks of post conceptional age) of a teenage mother. The baby is healthy but it weighs only half the weight of the fullterm discussed in Section 1.1. The sampling frequency of data is again 64HZ, the skewness is -0.9969 and the kurtosis is 9.7923. The sleep state of this one minute length data is scored as MFA.

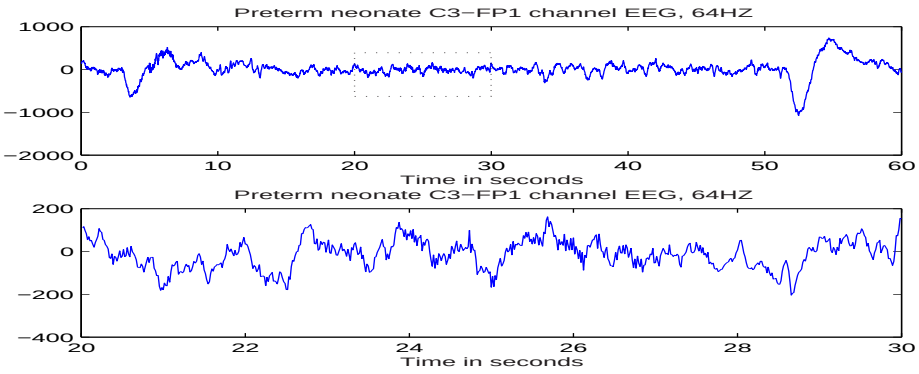

Figure 1.7: *Preterm EEG time series.*

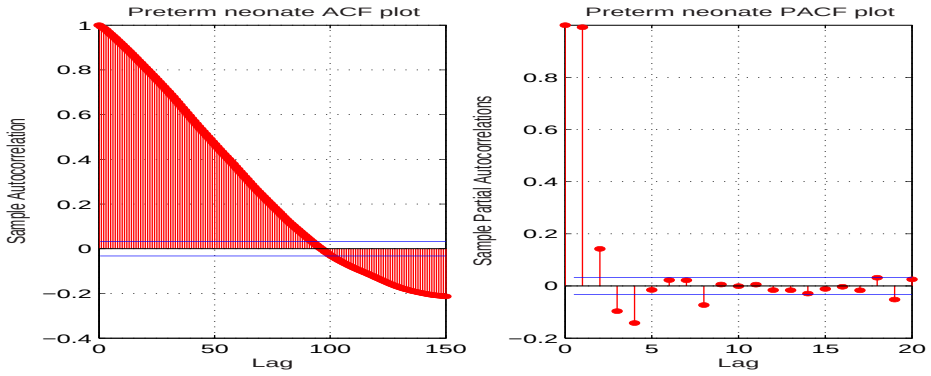

Figure 1.8: *Preterm ACF and PACF.*

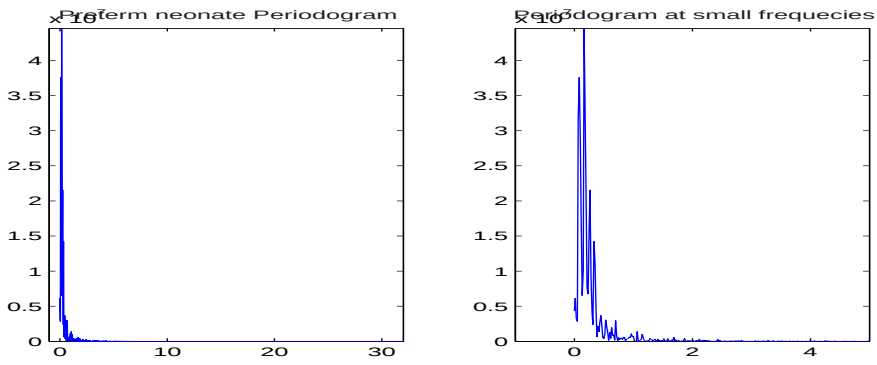

Figure 1.9: *Preterm periodogram.*

The sample ACF and PACF plots, and the periodogram plots are shown in Figures 1.8 and 1.9 respectively. Local trends are more obvious in this time series plot, which indicates *larger degree* of nonstationarity. Much stronger evidence of long memory can also be observed in the ACF plot because of the significant autocorrelations up to 100th lag. The steeper raise of periodograms around zero frequency also confirms the more pronounced long memory property. The PACF plot in this case also indicates nonstationarity of integration order 1 because the PACF value at lag one is almost one.

Since the original data shows nonstationarity, we segmented this data using 2 seconds windows, with no overlap between windows and with no tapering, and plotted the tv-periodogram plot in Figure 1.10. Along with it we also plotted a truncated version of tv-periodogram, the truncation was at 1000 level.

The tv-periodogram plot shows the long memory property to be different for different segments because the changes in the magnitude of periodograms are different over different segments. The peaks for different segments are not of the same magnitude and the pattern of periodograms around zero frequency are different for different segments. These observations suggest that the memory parameter may not remain constant over the entire length of

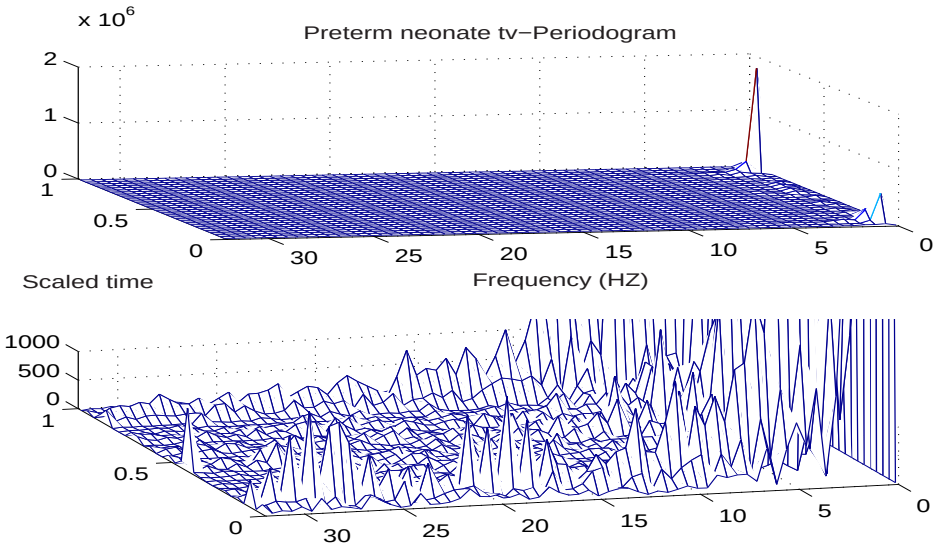

Figure 1.10: *Preterm tv-periodograms.*

this time series. The truncated tv-periodogram captures changes in the time-varying second order characteristics of the original time series over segments and around the non-zero frequency components. These plots also indicate nonconstant variance over different segments.

We differenced the preterm's original data once as we did in the case of the fullterm baby. The resulting differenced data along with its ACF and PACF plots are provided in 1.11. We still observe nonstationarity in the time series plot because the second order characteristics are not constant over the entire length of the differenced data. The ACF plot and the periodogram plots do not indicate any presence of long memory.

We again plotted the periodogram and the tv-periodogram of the differenced data in 1.12. The plots indicate no long memory property but the tv-periodogram plot again captures changes in the second order characteristics of the differenced data over the segments.

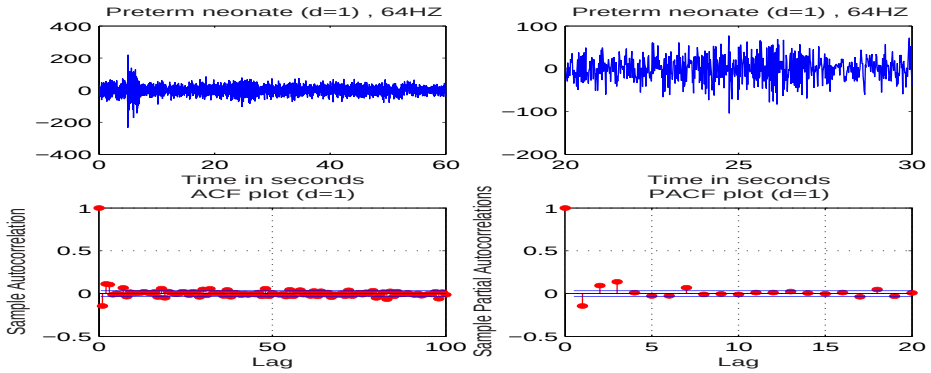

Figure 1.11: *Differenced time series with ACF and PACF plots.*

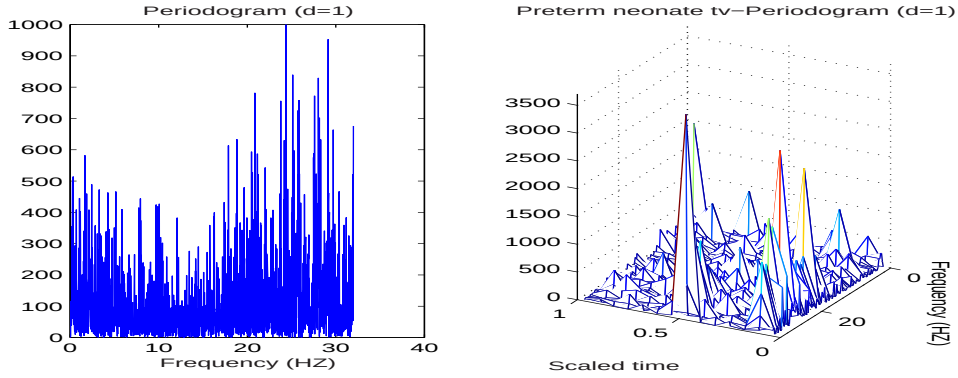

Figure 1.12: *Differenced data periodogram and tv-periodogram.*

1.2 Plan of research

To model the example sleep EEG of neonates, we aim to achieve the following objectives in our modeling technique:

1. We have to find ways to apply the stationary time series modeling techniques in the presence of nonconstant variance of a single channel EEG time series.

2. We also have to consider the possibility that the innovations of some of the sleep EEG states are non-Gaussian.

3. We have to capture time-varying behavior of the long memory property and stochastic trends of the data.

4. We also have to capture time-varying behavior of the short memory properties of the data.

We try to achieve these four objectives in reminder of the thesis. The outline of the dissertation is as follows. Chapter 2 introduces Dahlhaus (1997) definition of local stationarity concept to model a nonstationary time series. It also details a parameter estimation procedure of time-varying short memory model and also the estimation procedures of a long memory parameter. We provide a new time-varying model definition which captures the evolutionary long and short memory properties of the example time series in Chapter 3. We also provide a detailed two-step algorithm to calculate the time-varying nonstationary memory and the stationary short memory components of the sleep EEG. Chapter 4 gives a simple procedure to separate the short memory time series component from the nonstationary long memory time series component. Chapter 5 includes a thorough analysis of fullterm and a brief summary of preterm sleep EEG data, and the final Chapter 6 provides conclusions and suggests further extensions of this research.

Chapter 2

Introduction

The stationarity property of a time series was crucial in the classical theory of estimation and asymptotics of time series models. A discrete time series $X_t,\ t = 0, \pm 1, \pm 2, \cdots$ is said to be *strictly stationary* if for any $t_1, t_2, ..., t_n$ and for any k, the joint probability distribution of $\{X_{t_1}, X_{t_2}, ..., X_{t_n}\}$ is identical with the joint probability distribution of $\{X_{t_1+k}, X_{t_2+k}, ..., X_{t_n+k}\}$ (cf. Brockwell and Davis, 1996). This is a difficult requirement to satisfy or to verify for any time series. It can be relaxed by requiring stationarity only for moments up to some order. A zero-mean discrete time random process X_t is said to be *weakly stationary* or, simply, *stationary* if the autocovariance function of lag k,

$$\gamma(k) = E(X_t X_{t+k}), \tag{2.1}$$

between X_t and X_{t+k} depends only on k, but not on t.

The *spectral density function* $f(\lambda)$ of a stationary process is defined as the discrete Fourier transform of the autocovariance function,

$$f(\lambda) = 1/2\pi \sum_{j=-\infty}^{\infty} \gamma(j)\ \exp(-i\lambda j), \tag{2.2}$$

where $\lambda \in [-\pi, \pi]$. The spectral density function of a stationary process is nonnegative, integrable, and even function on $[-\pi, \pi]$.

A common nonparametric estimator of the spectral density function of X_t of length n is the periodogram,

$$I_X(\lambda) = \frac{1}{2\pi n} |J_X(\lambda)|^2 = \frac{1}{2\pi n} J_X(\lambda)\, J_X(-\lambda), \tag{2.3}$$

where $J_X(\lambda)$ is the discrete Fourier transform of X_t and it is defined as

$$J_X(\lambda) = \sum_{t=0}^{n-1} X_t \exp(-i\lambda t). \tag{2.4}$$

In general, $I_X(\lambda)$ is asymptotically unbiased but inconsistent estimator of $f(\lambda)$ (cf. Brockwell and Davies, 1996). Consistency of this estimate can be improved by applying a proper tapering to the data.

A stationary time series can also be viewed as a sum of an infinite number of randomly weighted complex exponentials, Fourier basis functions, through the use of the *Cramer* representation (cf. Brillinger, 1981),

$$X_t = \int_{-\pi}^{\pi} A(\lambda) \exp(i\lambda t)\, d\xi(\lambda), \tag{2.5}$$

where $A(\lambda)$ is the transfer function and $\xi(\lambda)$ is zero-mean random process with variance one and orthogonal increments. The spectral density function of the stationary time series X_t can then be expressed by the fomula

$$f(\lambda) = |A(\lambda)|^2 = A(\lambda)A(-\lambda). \tag{2.6}$$

The Cramer representation of a stationary time series is in terms of the Fourier functions which are perfectly localized in the frequency domain but not localized in time. Moreover, the transfer function is independent of time. Thus the distribution of power over frequencies does not change with time.

The above statement about the distribution of power over frequencies not changing in time is obviously not true for a nonstationary time series like our EEG data. If X_t is nonstationary then the autocovariance function of X_t

and X_{t+k} is a function of t and k. In practice, many time series, especially biomedical and financial time series, exhibit nonstationary behavior. In such cases, various techniques, such as specialized transformations (*differencing*) of data or considering the data over small piecewise stationary time intervals, can be employed to make the analysis of stationary techniques applicable for nonstationary time series.

One of the difficulties of nonstationary time series modeling is development of satisfactory asymptotic theory, which is needed in time series to draw statistical inferences using finite size samples. If $X_0, X_1, ..., X_{n-1}$ is an arbitrary finite segment of a nonstationary time series then letting n to infinity, may loose information about the beginning or middle of this time series. Therefore, a different type of asymptotic setup is needed for nonstationary time series.

As in the theory of nonparametric regression, it seems natural to develop the asymptotic theory for nonstationary time series by setting up inference procedures over a finite grid. In the following subsection, we adapt Dahlhaus' (1997) work to develop parameter estimation techniques and an asymptotic theory for a specific type of nonstationary time series, which, following Dahlhaus, we call *locally stationary*.

2.1 Locally stationary time series and parameter estimation

Definition. A sequence of stochastic processes $X_{t,n}$ $(t = 1, ..., n)$ is called locally stationary with transfer function A^0 and trend μ if there exists a representation

$$X_{t,n} = \mu\left(\frac{t}{n}\right) + \int_{-\pi}^{\pi} \exp(i\lambda t)\ A^0_{t,n}(\lambda)\ d\xi(\lambda) \tag{2.7}$$

where the following holds.

i. $\xi(\lambda)$ is stochastic process on $[-\pi, \pi]$ with $\overline{\xi(\lambda)} = \xi(-\lambda)$ and

$$\begin{aligned} E(d\xi(\lambda)) &= 0, \\ E(d\xi(\lambda_j), d\xi(\lambda_k)) &= \eta(\lambda_j + \lambda_k)\, d(\lambda_j) d(\lambda_k), \end{aligned}$$

where $\eta(\lambda) = \sum_{j=-\infty}^{\infty} \delta(\lambda + 2\pi j)$ is the period 2π extension of the Dirac delta function.

ii. There exists a constant K and a 2π periodic function $A : [0,1] \times \mathcal{R} \to \mathcal{C}$ with $A(u, -\lambda) = \overline{A(u, \lambda)}$, and

$$\sup_{t,\lambda} \mid A^0_{t,n}(\lambda) - A(\frac{t}{n}, \lambda) \mid \leq K n^{-1},$$

for all n; $A(u, \lambda)$ and $\mu(u)$ are assumed to be continuous in u, where t and $u = \frac{t}{n}$ denote time points in the interval $[1, n]$ and the rescaled interval $[0, 1]$ respectively.

The smoothness of A in u guarantees that the process has locally a stationary behavior. The idea behind this representation is ,essentially, that, for each fixed n, one implicitly assumes some local interval of stationarity about each time point and a smooth change from one interval to the next. We also require additional smoothness conditions on A, namely differentiability, to develop asymptotic theory.

This concept is similar to piecewise stationarity. However, there is an additional restriction on the smoothness of the transfer function.

Dahlhaus (1997) defined the time-varying (evolutionary) spectrum of a locally stationary process at time $u \in [0, 1]$ and frequency $\lambda \in [-\pi, \pi]$ by the formula

$$f(u, \lambda) = |A(u, \lambda)|^2. \tag{2.8}$$

A good example of locally stationary processes is time-varying AutoRegressive Moving Average (tv-ARMA) process, which we study in Section 2.1.3.

2.1.1 Parameter estimation via Whittle's method

The Whittle parameter estimation method [cf. Dzhaparidze (1986), and Azencott and Dacunha-Castelle (1986)] is based on the minimization of the following likelihood function,

$$L_n(\theta) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \log \ f_\theta(\lambda) + \frac{I_n(\lambda)}{f_\theta(\lambda)} \right\} d\lambda, \tag{2.9}$$

where $f_\theta(\lambda)$ is the model spectral density function (its parametric form is generally assumed) and $I_n(\lambda)$ is the periodogram. This estimate is asymptotically efficient and $L_n(\theta)$ (up to a constant) converges in probability to the Gaussian likelihood function as $n \to \infty$. In the case of stationary AR processes, the parameter estimators obtained by this method [cf. Dzhaparidze (1986), pp. 115-117] are exactly the same as the Yule-Walker estimators [cf. Brockwell and Davis (1996), Propostion 8.10.1] and their asymptotics are the same. This likelihood function is also applied in the parameter estimation of nonlinear time series. For example, Leonenko and Woyczynski (2001) used it to estimate parameters of the nonlinear diffusion equation (the Burger's turbulance problem).

The $L_n(\theta)$ may be interpreted as a distance between the parametric spectral density $f_\theta(\lambda)$ and the nonparametric periodogram estimate $I_n(\lambda)$. Hence, the Whittle estimator is called the minimum distance (or minimum contrast) estimator. In the case where the model, $f_\theta(\lambda)$, is misspecified, then the minimization of $L_n(\theta)$ leads to an estimate of the parameter θ with the best approximating parametric spectral density (Dahlhaus, 1997). If an efficient data taper is applied in the calculation of the periodogram, then the estimate also has good small sample properties (Dahlhaus, 1988). However, this requires identifiability of the model which only holds for linear processes.

The estimation theory of locally stationary processes is motivated by the generalization of the Whittle likelihood function of stationary processes

(Dahlhaus, 1997), where the usual periodogram is replaced by local periodograms over possibly overlapping stationary data segments. The generalized version of the Whittle method to a locally stationary process replaces $I_n(\lambda)$ in $L_n(\theta)$ with a local version and integrates it over time. The resulting estimator again is *efficient* (Dahlhaus, 1997).

2.1.2 Generalizing Whittle's method to locally stationary processes

Let $X_{t,n}$ $(t = 0, ..., n-1)$ be locally stationary, and let $h : R \to R$ be a data taper with $h(u) = 0$ for $u \notin [0, 1)$. Then, for even segment length $N \leq n$, and scaled time $u = \frac{t}{n}$, define

$$\begin{aligned} J_N(u, \lambda) &= \sum_{j=0}^{N-1} h\left(\frac{j}{N}\right) X_{\lfloor un \rfloor - (N/2)+j+1,n} \exp(-i\lambda j), \\ H_{k,\,N}(\lambda) &= \sum_{j=0}^{N-1} h\left(\frac{j}{N}\right)^k \exp(-i\lambda j), \\ I_N(u, \lambda) &= \frac{1}{2\pi H_{2,N}(0)} \mid J_N(u, \lambda)|^2, \end{aligned}$$

where $\lfloor un \rfloor$ is the integer part of un.

Thus, $I_N(u, \lambda)$ is the periodogram over a segment of length N with midpoint $\lfloor un \rfloor$. The shift from segment to segment is denoted by S; that is, we calculate I_N over segments with midpoints $t_j := S(j-1)+N/2$ $(j = 1, ..., M)$ where $n = S(M-1)+N$ and M is the total number of segments, or, written in rescaled time, at time points $u_j := t_j/N$. Note that, if S is the shift from segment to segment, then the overlap between the segments is $N - S$.

We now define the generalization of the Whittle likelihood function to a

locally stationary process (Dahlhaus,1997) as

$$\mathcal{L}_n(\theta) = \frac{1}{4\pi}\frac{1}{M}\sum_{j=1}^{M}\int_{-\pi}^{\pi}\left\{\log\ f_\theta(u_j,\lambda) + \frac{I_N(u_j,\lambda)}{f_\theta(u_j,\lambda)}\right\}d\lambda, \tag{2.10}$$

with resulting minimum contrast estimator

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\arg\min}\mathcal{L}_n(\theta). \tag{2.11}$$

The use of data taper which tends smoothly to zero at the boundaries of the segments, or windows, has two benefits: (i) it reduces *leakage*[1] and (ii) it reduces the bias due to nonstationarity by downweighting the observations at the boundries of the segment. It is interesting to observe that the taper, if properly selected, does not lead to an increase of the asymptotic variance for overlapping segments (see the example problem in the next section). Furthermore, some estimates are even asymptotically independent of the taper [see Dahlhaus (1997), Theorem 4.2].

Lets define,

$$\mathcal{L}(\theta) = \frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\left(\log f_\theta(u,\lambda) + \frac{f_\theta(u,\lambda)}{f(u,\lambda)}\right)d\lambda du.$$

Clearly, as $n\to\infty$ the $\mathcal{L}_n(\theta)$ should converge to $\mathcal{L}(\theta)$ [Dahlhaus (1996), Theorem 3.4].

The above motivation of the function $\mathcal{L}_n(\theta)$ is heuristic. The theoretical justification for the form of $\mathcal{L}_n(\theta)$ relies on its ability to minimize the asymptotic Kullback-Leibler (K-L) information divergence index [cf. Azencott and Dacunha-Castelle (1986), and Dahlhaus (1996)]. We summarize the details in the following theorem. Given its importance we outline the proof as well.

[1]As in stationary case, leakage is a phenomenon describing the 'leakage' of energy that is at one frequency to another in the spectral estimates. The rectangular windowing associated with the discrete Fourier transform introduces spectral leakage, the energy from low frequency coefficients leaks to the full-frequency range. The spectral leakage together with the *Gibbs* effect leads to significant phase and amplitude errors at the beginning and the end of windows [see Bloomfield (2000), Chapter 5].

Theorem 2.1 *Suppose g and f and are true probability density and spectral density of a locally stationary process $X_{(1,n)}, ..., X_{(n,n)}$. Let g_θ and f_θ are the densities of our assumed model. If g and g_θ are Gaussian with mean zero, then the asymptotic Kullback-Leibler information divergence index is*

$$\lim_{n\to\infty} \frac{1}{n}\ E_g\ \log\left(\frac{g}{g_\theta}\right) = \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \left\{\log\ f_\theta(u,\lambda) + \frac{f(u,\lambda)}{f_\theta(u,\lambda)}\right\} d\lambda\ du + \textit{ constant}$$

Proof.

The K-L index,

$$K(g, g_\theta) = \lim_{n\to\infty} \frac{1}{n} E_g \log\left(\frac{g}{g_\theta}\right) = \lim_{n\to\infty} E_g\left[L_n(\theta) - L_n(.)\right],$$

where $L_n(.)$ and $L_n(\theta)$ are (negative values of) log likelihood functions w.r.to g and g_θ respectively. (Note: Maximizing the log likelihood function is equivalent to minimizing the negative value of the log likelihood.)

The equality $K(g, g_\theta) = \lim_{n\to\infty}\ E\left[\mathcal{L}_n(\theta) - \mathcal{L}_n(.)\right]$ is true [cf. Dahlhaus(1996), Theorem 3.4] in the sense of convergence in probability, with $\mathcal{L}_n(.)$ and $\mathcal{L}_n(\theta)$ are being the generalized Whittle likelihood functions of f and f_θ respectively. We can derive the above equality in a way similar to the derivation of an analogous formula as in the special case of stationary Gaussian processes [cf. pages 51-52 in Dzhaparidze (1986), pages 51-52].

$$\begin{aligned}
K(g, g_\theta) &= \lim_{n\to\infty} E\left[\mathcal{L}_n(\theta) - \mathcal{L}_n(.)\right] \\
&= \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \left\{\left(\log f_\theta(u,\lambda) + \frac{f(u,\lambda)}{f_\theta(u,\lambda)}\right) - \left(\log\ f(u,\lambda) + \frac{f(u,\lambda)}{f(u,\lambda)}\right)\right\} d\lambda\ du \\
&= \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \left\{\log\ f_\theta(u,\lambda) + \frac{f(u,\lambda)}{f_\theta(u,\lambda)} - (\log f(u,\lambda) + 1)\right\} d\lambda\ du \\
&= \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \left\{\log\ f_\theta(u,\lambda) + \frac{f(u,\lambda)}{f_\theta(u,\lambda)}\right\} d\lambda\ du + \text{ constant},
\end{aligned}$$

where the constant is independent of the model spectral density. The final result above is due to the local prediction error [see Dahlhaus, (1996), Theorem 3.2] of locally stationary processes,

$$E\left(X_{t,n} - \hat{X}_{t,n}\right)^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log 2\pi f(t/n, \lambda) d\lambda\right\} \tag{2.12}$$

This local prediction error is a generalization of Kolmogorov's formula for stationary processes [cf. Brockwell and Davis (1996), Theorem 5.8.1]. $\diamond$

The equality $K(g, g_\theta) = \mathcal{L}(\theta)+$ constant, offers stronger justification for the particular form of $\mathcal{L}_n(\theta)$. Therefore, we may regard

$$\mathcal{L}(\theta) = \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \left\{\log\ f_\theta(u,\lambda) + \ \frac{f_\theta(u,\lambda)}{f(u,\lambda)}\right\} d\lambda du$$

as a distance between the true process with spectral density $f(u, \lambda)$ and the model with spectral density $f_\theta(u, \lambda)$. Then, the best approximating parameter value from the assumed model is

$$\hat{\theta} = \ \underset{\theta\in\Theta}{\arg\min}\mathcal{L}(\theta).$$

If the assumed model is correct, that is, $f(u, \lambda) = f_\theta(u, \lambda)$, then

$$\mathcal{L}(\theta) = \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \{\log\ f(u,\lambda)\}\, d\lambda du + \frac{1}{2}$$

Hence, $\hat{\theta} = \theta_0$, where θ_0 the true value. See Chapter 5 in Brockwell and Davis (1996) about the connection between the one-step prediction errors and the best linear unbiased estimators, and also Dahlhaus (1996).

In the following section, we apply the generalized Whittle parameter estimation procedure to estimate the parameters of an example time-varying AR problem. The final estimation equations for a time-varying AR process

are provided by Dahlhaus (1997) without any derivation for an *i.i.d* innovations with finite variance. However, we derive this algebra in the next section because we will use it to derive the estimation equations of a modified version of the generalized Whittle likelihood function. This modified likelihood function requires no prior knowledge about the existence of variance of the innovations, which may be useful in fitting the sleep EEG. We suspect that the innovations in sleep EEG has infinite variance.

2.1.3 Fitting time-varying autoregressive models

Let $X_{t,n}$ be a solution of the system of difference equations

$$\sum_{j=0}^{p} a_j\left(\frac{t}{n}\right) X_{t-j,n} = \sigma\left(\frac{t}{n}\right)\varepsilon_t \quad \text{for } t \in \mathcal{Z}, \tag{2.13}$$

where $a_0(u) = 1$ and $\{\varepsilon_t\}$ are independent random variables with mean zero and variance one. Then, $X_{t,n}$ is locally stationary (Dahlhaus, 1996) with the time-varying spectral density

$$f(u,\lambda) = \frac{\sigma^2(u)}{2\pi}\left|\sum_{j=0}^{p} a_j(u)\exp(i\lambda j)\right|^{-2}.$$

The sample autocovariance function of $X_{t,n}$

$$\begin{aligned} c_N(u,j) &= \int_{-\Pi}^{\Pi} I_N(u,\lambda)\exp(i\lambda j)d\lambda \\ &= H_{2,N}(0)^{-1}\sum_{\substack{s,t=0 \\ s-t=j}}^{N-1} h\left(\frac{s}{N}\right) h\left(\frac{t}{N}\right) X_{[un]-(N/2)+s+1,T} X_{[un]-(N/2)+t+1,T}. \end{aligned}$$

Let

$$\begin{aligned} C_N(u) &= (c_N(u,1),...,c_N(u,p))', \\ \Sigma_N(u) &= \{c_N(u,j-k)\}_{j,k=1,...,p}. \end{aligned}$$

For convenience, we split the generalized Whittle likelihood function ($\mathcal{L}_n$) into two parts $\mathcal{L}_n^1$ and $\mathcal{L}_n^2$, where

$$\mathcal{L}_n^1 = \frac{1}{4\pi}\frac{1}{M}\sum_{j=1}^{M}\int_{-\pi}^{\pi} \log\ f_\theta(u_j,\lambda)d\lambda, \tag{2.14}$$

$$\mathcal{L}_n^2 = \frac{1}{4\pi}\frac{1}{M}\sum_{j=1}^{M}\int_{-\pi}^{\pi} \left\{\frac{I_N(u_j,\lambda)}{f_\theta(u_j,\lambda)}\right\} d\lambda. \tag{2.15}$$

The $\mathcal{L}_n^1$ is a function of local prediction error (2.12) of $X_{t,n}$ as in the previous section.

Suppose that (with some abuse of notation) $a(u) = (a_1(u), ..., a_p(u))'$ and

$$\mathcal{L}_n^2(u) = \int_{-\pi}^{\pi} \left\{\frac{I_N(u,\lambda)}{f_\theta(u,\lambda)}\right\} d\lambda,$$

then

$$\begin{aligned}
\mathcal{L}_n^2(u) &= \frac{2\pi}{\sigma^2(u)}\int_{-\pi}^{\pi} I_N(u,\lambda)\left|\sum_{j=0}^{p} a_j(u)\exp(i\lambda j)\right|^2 d\lambda \\
&= \frac{2\pi}{\sigma^2(u)}\int_{-\pi}^{\pi} I_N(u,\lambda)\ \left(\sum_{j,k=0}^{p} a_j(u)a_k(u)\exp(i\lambda(j-k))\right)\ d\lambda \\
&= \frac{2\pi}{\sigma^2(u)}\left(\sum_{j=0}^{p} a_j^2(u)c_N(u,0) + \sum_{\substack{j,k=0\\ j\neq k}}^{p} a_j(u)a_k(u)c_N(u,(j-k))\right) \\
&= \frac{2\pi}{\sigma^2(u)}\left(a'(u)\Sigma_N(u)a(u) + 2a'(u)C_N(u) + c_N(u,0)\right)
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_n^2(u) &= \frac{2\pi}{\sigma^2(u)}\{a'(u)\Sigma_N(u)a(u) + a'(u)C_N(u) + C_N'(u)a(u) + c_N(u,0) \\
&\quad + C_N'(u)\Sigma_N^{-1}(u)C_N(u) - C_N'(u)\Sigma_N^{-1}(u)C_N(u)\} \\
&= \frac{2\pi}{\sigma^2(u)}\{(\Sigma_N(u)a'(u) + C_N(u))'\,\Sigma_N^{-1}\,(\Sigma_N(u)a'(u) + C_N(u)) \\
&\quad + c_N(u,0) - C_N'(u)\Sigma_N^{-1}C_N(u)\}.
\end{aligned}$$

Let $\theta = \{(a_1(u_j), \cdots, a_p(u_j), \sigma^2(u_j)\}$, where $j = 1, \cdots, M$. By combining $\mathcal{L}_n^1$ and $\mathcal{L}_n^2$, we obtain,

$$\begin{aligned}
\mathcal{L}_n(\theta) &= \frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}\{\log\sigma^2(u_j) + \frac{1}{\sigma^2(u_j)} \\
&\quad \times[(\Sigma_N(u_j)a(u_j) + C_N(u_j))'\,\Sigma_N^{-1}\,(\Sigma_N(u)a(u_j) + C_N(u_j)) \\
&\quad + c_N(u_j,0) - C_N'(u_j)\Sigma_N^{-1}C_N(u_j)]\} \qquad (2.16)
\end{aligned}$$

Using the above equation, one can derive the following partial derivatives

$$\frac{\partial\mathcal{L}_n(\theta)}{\partial a(u_j)} = \frac{2a(u_j)\Sigma_N(u_j) + 2C_N(u_j)}{\sigma^2(u_j)} \qquad (2.17)$$

$$\frac{\partial\mathcal{L}_n(\theta)}{\partial\sigma^2(u_j)} = \frac{1}{\sigma^2(u_j)} - \frac{a'(u_j)\Sigma_N(u_j)a(u_j) + 2a'(u_j)C_N(u_j) + c_N(u,0)}{\sigma^4(u_j)} \qquad (2.18)$$

Equating the above partial derivatives to zero results in the following estimators

$$\hat{a}(u_j) = -\Sigma_N^{-1}(u_j)C_N(u_j), \qquad (2.19)$$

$$\hat{\sigma}^2(u_j) = c_N(u,0) - C_N'(u_j)\Sigma_N^{-1}(u_j)C_N(u_j). \qquad (2.20)$$

In time series literature, these estimators are similar to Yule-Walker estimators of stationary processes [cf. Dzhaparidze (1986), pp. 116-117] of $a(u_j)$ and $\sigma^2(u_j)$ of segment j with midpoint u_j.

Using the definition of *Fisher information* (F) matrix of a stationary process (cf. Dzhaparidze, 1986), one can define the Fisher information matrix

of a locally stationary process with $\theta(u_j) = (a(u_j), \sigma^2(u_j))'$ as

$$F_{\theta(u_j)} = \left\{ \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_k(u_j)} \log f(u_j, \lambda) \frac{\partial}{\partial \theta_l(u_j)} \log f(u_j, \lambda)\, d\lambda \right\}_{k,l=1,\cdots,p+1}. \tag{2.21}$$

Then, the Fisher information matrix for the generalized Whittle estimators of a time-varying AR process can be derived to be

$$F_{\theta(u_j)} = \begin{bmatrix} \Sigma_N(u_j)/\sigma^2(u_j) & 0 \\ 0 & 1/2\sigma^4(u_j) \end{bmatrix}.$$

The algebra for $F_{\theta(u_j)}$ derivation is same as the derivation of F_θ for stationary AR processes [cf. Dzhaparidze (1986), page 117].

If one assumes ε_t i.i.d. Gaussian, then the generalized Whittle likelihood estimators are asymptotically equivalent to the exact Maximum Likelihood Estimators (Dahlhaus, 1997, Section 4) and are approximately Gaussian. Hence the following asymptotic result holds in probability for $\hat{\theta}(u_j)$

$$\sqrt{n}\ \left(\hat{\theta}(u_j) - \theta(u_j)\right) \to \mathcal{N}\left(0, F^{-1}_{\theta(u_j)}\right). \tag{2.22}$$

The computational elegance of $\mathcal{L}_n(\theta)$ in (2.16) can be explained by the Yule-Walker estimation equations (2.19) and (2.20).

The $\log \sigma^2(u_j)$ term in $\mathcal{L}_n(\theta)$ can be rewritten using $\log(x) = (x-1) - \frac{1}{2}(x-1)^2 + o((x-1)^2)$

$$\begin{aligned} \log \sigma^2(u_j) &= \log \frac{\sigma^2(u_j)}{\hat{\sigma}^2(u_j)} + \log \hat{\sigma}^2(u_j) \\ &= -\log \frac{\hat{\sigma}^2(u_j)}{\sigma^2(u_j)} + \log \hat{\sigma}^2(u_j) \\ &\approx -\left[\left(\frac{\hat{\sigma}^2(u_j)}{\sigma^2(u_j)} - 1 \right) - \frac{1}{2} \left(\frac{\hat{\sigma}^2(u_j)}{\sigma^2(u_j)} - 1 \right)^2 \right] + \log \hat{\sigma}^2(u_j). \end{aligned}$$

The term $(\frac{1}{\sigma^2(u_j)}\left[(\Sigma_N(u_j)a(u_j)+C_N(u_j))'\ \Sigma_N^{-1}\ (\Sigma_N(u)a(u_j)+C_N(u_j))\right]+\frac{1}{\sigma^2}(u_j)\left[c_N(u_j,0)-C_N'(u_j)\Sigma_N^{-1}C_N(u_j)\right]$ in $\mathcal{L}_n(\theta))$ can be rewritten as

$$\frac{1}{\sigma^2(u_j)}\left[\left(a(u_j)+\Sigma_N^{-1}C_N(u_j)\right)'\Sigma_N(u)\left(a(u_j)+\Sigma_N^{-1}C_N(u_j)\right)+\hat{\sigma}^2(u_j)\right]=$$

$$\frac{(a(u_j)-\hat{a}(u_j))'\,\Sigma_N(u_j)\,(a(u_j)-\hat{a}(u_j))}{\sigma^2(u_j)}+\frac{\hat{\sigma}^2(u_j)}{\sigma^2(u_j)}.$$

If the assumed model is reasonably close to the true process, then we can expect $\sigma^2(u_j)\approx\hat{\sigma}^2(u_j)$. We therefore obtain the following expression for $\mathcal{L}_n(\theta)$ in the neighborhood of the minimum by adding the above two equations:

$$\begin{aligned}\mathcal{L}_n(\theta)\approx{}&\frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}\{2\hat{\sigma}^4(u_j)\}^{-1}\left(\sigma^2(u_j)-\hat{\sigma}^2(u_j)\right)\\&+\frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}(a(u_j)-\hat{a}(u_j))'\ \hat{\sigma}^2(u_j)^{-1}\hat{\Sigma}_N(u_j)\\&\qquad\times(a(u_j)-\hat{a}(u_j))\\&+\frac{1}{2}\frac{1}{M}\sum_{j=1}^{M}\log\hat{\sigma}^2(u_j)+\frac{1}{2}.\end{aligned}\tag{2.23}$$

Therefore, $\hat{\theta}_n$ is approximately obtained by a weighted least squares fit of $a(u_j)$ and $\sigma^2(u_j)$ to the Yule-Walker estimates on the segments. Note that, the Yule-Walker estimators of stationary AR time series with data taper have good small sample properties (Dahlhaus, 1988).

The $\mathcal{L}_n(\theta)$ expressions in (2.16) and (2.23) are useful in estimating the parameters of any time-varying AR process, in particular when $\sigma^2(u)$ is not constant over the time.

We now give an explicit formula for $\hat{\theta}$ in the following scenario. Suppose that $a(u)$ is polynomial in time, i.e., $a_j(u)=\sum_{k=0}^{K_j}b_{jk}u^k$ and $\sigma^2(u)$ is constant, σ^2. Let $b=(b_{11},...,b_{1K_1},...,b_{p1},...,b_{pK_p})'$. Then, the finite dimensional

parameter of interest is $\theta = (b', \sigma^2)$. In this case we can derive estimates for b and σ^2 explicitly via the above generalized Whittle maximum likelihood procedure. The resulting formulae are

$$\widehat{b}_n = -\left(\frac{1}{M}\sum_{j=1}^{M} F(u_j) \otimes \Sigma_N(u_j)\right)^{-1} \left(\frac{1}{M}\sum_{j=1}^{M} f(u_j) \otimes C_N(u_j)\right) \quad (2.24)$$

$$\widehat{\sigma}_n^2 = \left(\frac{1}{M}\sum_{j=1}^{M} c_N(u_j, 0)\right) + \widehat{b}_n' \left(\frac{1}{M}\sum_{j=1}^{M} f(u_j) \otimes C_N(u_j)\right), \quad (2.25)$$

where $\otimes$ is *the left Kronecker product*[2] and

$$\begin{aligned} c_N(u,j) &= \int_{-\Pi}^{\Pi} I_N(u,\lambda)\exp(i\lambda j)d\lambda \\ &= H_{2,N}(0)^{-1} \sum_{\substack{s,t=0 \\ s-t=j}}^{N-1} h\left(\frac{s}{N}\right) h\left(\frac{t}{N}\right) X_{[un]-(N/2)+s+1,n} X_{[un]-(N/2)+t+1,n}, \\ C_N(u,j) &= (c_N(u,1), ..., c_N(u,p))', \\ \Sigma_N(u) &= \{c_N(u, i-j)\}_{i,j=1,...,p}, \\ f(u) &= (u^0, u^1, ..., u^K)', \\ F(u) &= \{f(u_i)f(u_j)\}_{i,j=1,...,K}. \end{aligned}$$

The model parameters $p, K_1, ..., K_p$ may be determined by minimizing the Akaike Information Criterion (see Brockwell and Davis, 1996)

$$AIC(p, K_1, ..., K_p) = \log \widehat{\sigma}^2(p, K_1, ..., K_p) + 2\frac{\left(p + 1 + \sum_{j=1}^{p} K_j\right)}{n}. \quad (2.26)$$

[2]Let A be a $n \times n$ matrix (with entries a_{ij}) and let B be a $m \times m$ matrix. Then the Kronecker product of A and B is the $mn \times mn$

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}$$

The Kronecker product is also known as the direct product or the tensor product.

Note that:

(i) the polynomial order K_j of jth AR parameter could be different from the other AR parameters' polynomial orders and,

(ii) the above parameter estimates are valid only when $\sigma(u) = \sigma$ is independent of u.

If $\sigma(u)$ is not constant over segments then a direct minimization of $\mathcal{L}_n(\theta)$ expressions in (2.16) or (2.23) may be necessary. Such minimization generally requires good initial values.

Instead of polynomial fitting, one may consider *Wavelets* or some other orthogonal functions to fit and understand the time-varying nature of parameters. However, if the time-varying parameters change smoothly from one segment to another, then we find the polynomial fitting of time-varying parameters is *efficient* and easy to interpret.

We now briefly present a simulation example for the estimate $\hat{\theta}_n$ in a misspecified situation. If we have a locally stationary process with smoothly varying parameters, then it is likely that $\hat{\theta}_n$ leads to reasonable results. We simulated the same time-varying AR(2) example in Dahlhaus (1997) with parameters $a_1(u) = -1.8\cos(1.5 - \cos 4\pi u)$, $a_2(u) = .81$, and $\sigma(u) = 1$. We assumed ε_t to be an independent Gaussian noise and simulated this non-stationary time series. The roots of the autoregressive characteristic equation $(1 - 1.8\cos(1.5 - \cos 4\pi u)z + 0.81z^2 = 0)$ of the given example are $(1/.9)\exp\left[\mp i(1.5 - \cos 4\pi u)\right]$.

The following plot is a realization with $n = 128$ length over scaled time $u_j = (t_j/T) \in [0, 1]$ of the above example. Using cosine bell taper (Tukey-Hanning taper) $h(x) = \frac{1}{2}\left[1 - \cos(2\pi x)\right]$ (for its good leakage properties, see Bloomfield, 2000), $N = 16$, $S = 2, p = 2, K1 = 6, K2 = 0$ and using (2.24) and (2.25) we derived the estimate of the first AR parameter as shown in Figure 2.2, the estimate of the second AR parameter as .71, and the estimate of the $\sigma(u)$ as 1.71. All these estimated values are the same as the values

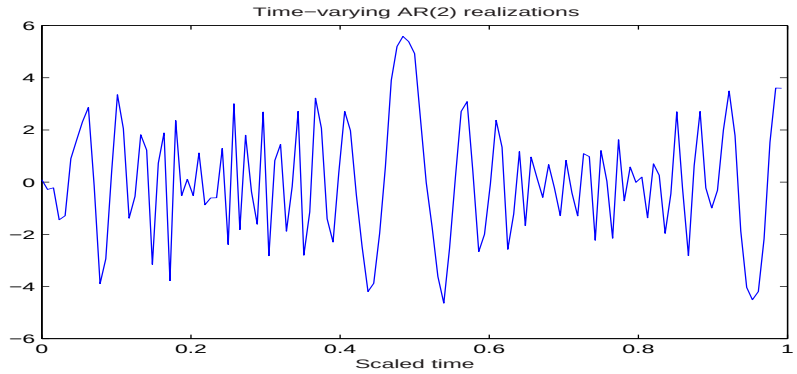

Figure 2.1: *time-varying AR(2) time series.*

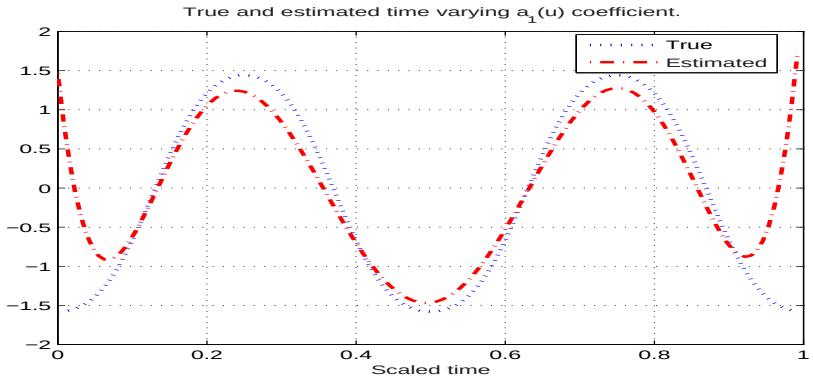

Figure 2.2: *First AR(2) parameter estimate.*

derived by Dahlhaus (1997). The model parameters $N, S, p, K1$, and $K2$ are determined by the AIC criterion. One can select smaller N and S but small values are computationally not efficient.

Notice that:

(i) divergence error of the first AR parameter estimate at the initial and final values of u and because of this $\sigma(u)$ is almost twice as large as the true value, and

(ii) the estimated spectrum peaks are underestimated as seen in the tv-periodogram plots, Figure 2.3.

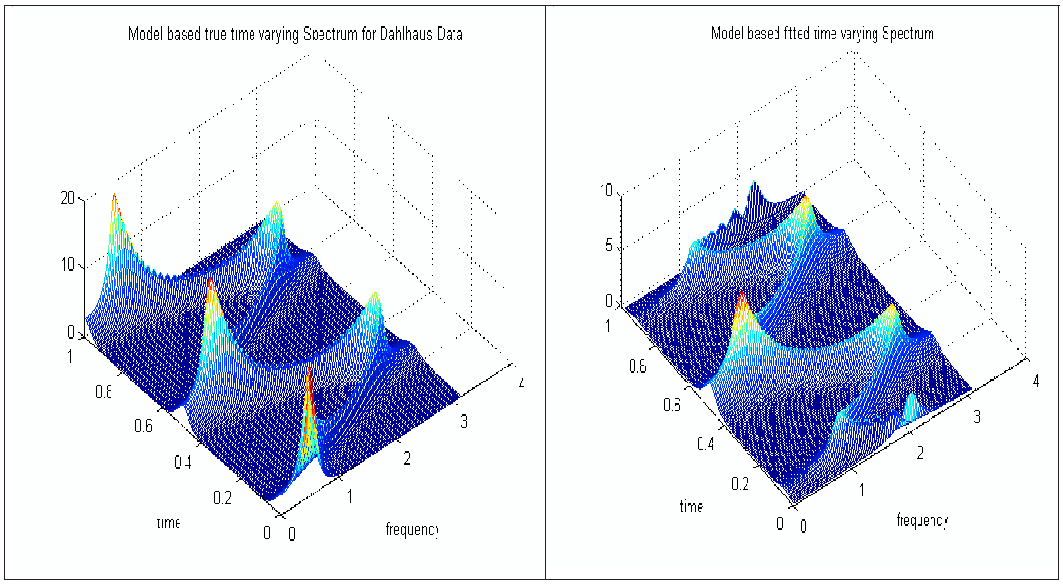

Figure 2.3: *Time-varying periodogram.*

The divergence error of the first AR parameter estimate can be attributed to the $\mathcal{L}_n(\theta)$ which penalizes bad fits inside the $[u_1, u_M]$ but not outside this interval. Dahlhaus (1997) suggested to select larger $K1$, and $K2$ values to correct it. The underestimation of spectrum can be attributed to the nonstationarity of data over the intervals $(u_j - N/2n, u_j + N/2n)$, where $I_N(u_j, \lambda)$ and $c_N(u_j, j)$ are calculated. It is obvious that smaller segment length N should decrease this negative bias but small N would increase variance of the estimates.

2.2 Long memory or fractional behavior of a time series

In literature, autocovariances and autocorrelations are often referred to as memory indicators. A simple way to classify the memory type of a stationary time series is by quantifying the rate of decay of autocovariances or autocorre-

lations. In mathematical terms long memory process' autocorrelations have a power type decay to zero as the lag increases. The autocorrelations decay to zero in a short memory process, such as ARMA processes [cf. Beran (1994), Brockwell and Davis (1996)], occurs at a much more rapid, exponential, rate. In frequency domain the spectral density of a long memory process has no singularity at zero. On the other hand, the spectral density of short memory process' poles, if any, are away from zero. Poles are the roots of the inverse of the spectral density function.

An ARMA process X_t is a short memory process since the autocovariance between X_t and X_{t+k} decreases exponentially as $k \to \infty$. In fact the autocorrelation function, $\rho(k)$, is exponentially bounded, i.e. $|\rho(k)| \leq Cr^k$, for $k = 1, 2, ...$, where $C > 0$ and $0 < r < 1$. A long memory process is a stationary process for which $\rho(k) \sim C\ r^{2d-1}$ as $k \to \infty$, where $C \neq 0$ and $d < .5$ (cf. Beran, 1994 for details).

While long memory time series can always be approximated by an ARMA(p,q) process, the orders p and q required to achieve a reasonably good approximation may be so large that parameter estimation gets extremely tedious.

For $d > -1$, we define the difference operator $\nabla^d = (1 - B)^d$, where B is the backward shift operator, by means of the binomial expansion

$$\nabla^d = (1 - B)^d = \sum_{j=0}^{\infty} \pi_j B^j,$$

where $\pi_j = \Gamma(j-d)/\Gamma(j+d)\ \Gamma(-d) = \prod_{0<k\leq j}(k-1-d)/k$, $j = 0, 1, 2...$, and Γ is the gamma function. Detailed explanation of long memory processes is provided in Chapter 4.

Let X_t be an ARMA(p,q) process with long memory parameter $d \in [-.5, .5)$ then X_t is called *Fractionally integrated* ARMA(p,q) or ARFIMA(p,d,q) process. The process $\{X_t, t = 0, 1, \cdots\}$ is said to be an ARFIMA(p,q) process with $d \in [-.5, .5)$ if X_t is a stationary solutions with zero mean of the

difference equations,

$$\varphi(B)\nabla^d X_t = \theta(B)\varepsilon_t$$

where ε_t is i.i.d noise with mean 0 and variance σ^2 and φ and θ are AR and MA polynomials of degrees p and q respectively. Again, detailed explanation is provided in Chapter 4.

One observation is that if X_t is ARMA(0,d,0), that is, $\nabla^d X_t = \varepsilon_t$ then we call X_t fractionally integrated noise. In the case if ε_t is Gaussian then we call X_t fractionally integrated Gaussian noise. In Cramer representation,

$$\nabla^d X_t = \int_{-\pi}^{\pi} \exp(i\lambda t)(1 - \exp(-i\lambda))^d d\varepsilon_X(\lambda).$$

Because ARFIMA(p,d,q) time series is a stationary process for $d \in [-5, 5)$, we can write (Theorem 13.2.2 in Brockwell and Davis, 1996) the spectral density of this time series as

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{|a(\exp(-i\lambda))|^2}{|b(\exp(-i\lambda))|^2} |1 - \exp(-i\lambda)|^{-2d} \sim \frac{\sigma^2}{2\pi}\lambda^{-2d} \text{ as } \lambda \to 0$$

where $a(\exp(-i\lambda)) = \sum_{j=0}^{p} a_j \exp(-i\lambda j))$ and $b(\exp(-i\lambda)) == \sum_{k=0}^{q} b_k \exp(-i\lambda k))$ are the AR and MA polynomials of degrees p and q respectively.

In this context a time series X_t is said to exhibit long memory property if the spectrum behaves as

$$\frac{f(\lambda)}{c_f|\lambda|^{-2d}} = 1 \text{ as } \lambda \to 0. \tag{2.27}$$

In the following section we discuss three popular methods to estimate the long memory parameter d. For other methods, refer Beran (1994).

2.2.1 Log-Periodogram method

The idea for this method comes from (2.8), the behavior of spectrum of a stationary ARFIMA process close to zero frequency. By applying the loga-

rithm on both sides of (2.7) and using the nonparametric estimate $I_n(\lambda) = \frac{1}{2\pi n}\left|\sum_{j=1}^{n-1} X_j \exp(-i\lambda j)\right|^2$ in the place of $f(\lambda)$, we can write

$$\log I_n(\lambda_j) = \log c_f - 2d\ \log \lambda_j + \log \epsilon_j. \tag{2.28}$$

The least square estimate $\hat{d}$ of the above equation for $j = 1, \cdots, m \leq \lfloor n/2 \rfloor$ is defined as the log-periodogram estimate. In general $m = n^{1/2}$ is optimal but recently Hurvich, Deo, and Brodsky (1998) showed that $m = n^{4/5}$ is *MSE-optimal*; the value $m = n^{4/5}$ is the optimal value for which $E(d - \hat{d})^2$ attains the minima.

Besides the simplicity of this method the main advantage of this estimator is that it does not require the knowledge about the short memory parameters. Consistency of the estimator can be obtained without the knowledge of the innovations and this knowledge is needed only to prove the asymptotic normality, $m^{1/2}(\hat{d} - d) \to N(0, \pi^2/24)$ in distribution.

2.2.2 Whittle's method

The Whittle estimator is the value of the vector θ which minimizes the function

$$L(\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\log\ f_\theta(\lambda) + \frac{I(\lambda)}{f_\theta(\lambda)}\right\} d\lambda. \tag{2.29}$$

The term $\int_{-\pi}^{\pi} \log\ f_\theta(\lambda) d\lambda$ can be set equal to 0 by normalization. The normalization only depends on a scale parameter, and not on the rest of the components of θ. Thus we replace f by f^*, such that $f^* = \beta f$ and $\int_{-\pi}^{\pi} \log\ f_\theta^*(\lambda) d\lambda = 0$. The first term in the above equation stays unchanged except for division by β. The parameter β is regarded as an extra parameter. This parameter does not depend on the other parameters and it will not effect the minimization of L over these parameters. In real data applications, instead of an integral, the corresponding sum over the Fourier frequencies,

$\lambda = 2\pi j/n$ is computed, where $j = 0, 1, 2, \cdots, \lfloor (n-1)/2 \rfloor$ and n is the length of the time series. Thus, the actual function which the algorithm minimizes is

$$L_n(\theta) = \frac{1}{4\pi} \sum_{j=0}^{[(n-1)/2]} \frac{I_n(\lambda_j)}{f_\theta^*(\lambda_j)}. \tag{2.30}$$

When dealing with fractional Gaussian noise of ARFIMA(0,d,0), θ is the parameter d. If the series is assumed to be ARFIMA(p,d,q) then θ includes the unknown autoregressive and moving average parameters. For such series the fractional parameter d converges to its true value at the rate $n^{1/2}$ and the asymptotic distribution is Gaussian (Fox and Taqqu, 1986).

2.2.3 Semiparametric Whittle's method

This Whittle estimator (Robinson, 1995) is semi-parametric, in that it only assumes the parametric form of the spectral density when λ is close to zero, namely, $f(\lambda) \sim G(d)\lambda^{-2d}$ as $\lambda \to 0$. The local Whittle estimator is also based on the periodogram. Its computation involves the introduction of an additional parameter m, which is an integer less than $n/2$, and satisfying, as $N \to \infty$, $1/m + m/n \to 0$. For a spectral density satisfying the above equation, the discrete analogue of Whittle maximum likelihood function is

$$L_n(G, d) = \frac{1}{m} \sum_{j=1}^{m} \left(\log G(d)\lambda_j^{-2d} + \frac{I_n(\lambda_j)}{G(d)\lambda_j^{-2d}} \right).$$

By replacing G by its estimate, $\frac{1}{m}\sum_{j=1}^{m} \frac{I_n(\lambda_j)}{\lambda_j^{-2d}}$ one obtains

$$\widehat{L}_n(G, d) = \log \left(\frac{1}{m} \sum_{j=1}^{m} \frac{I_n(\lambda_j)}{\lambda_j^{-2d}} \right) - 2d \frac{1}{m} \sum_{j=1}^{m} \log \lambda_j.$$

The value $\hat{d}$ which minimizes $\widehat{L}_T$ converges in probability to the actual value d_0 as $n \to \infty$. Robinson (1995) shows that $\sqrt{m}(\widehat{d} - d_0) \to N(0, 1/4)$ as

$n \to \infty$. This asymptotical result is significant in the sense that the variance is independent of unknown parameters.

The choice of m can be quite important. The larger the value of m the faster $\widehat{d}$ converges to d_0. On the other hand, if the series is not ideal, i.e. the time series also has short memory component, then we would want m to be small (Robinson, 1995). In such a case we only want to use the frequencies which are close to zero, since at larger frequencies the short-memory behavior of the series will affect the form of the spectral density. Still this method has several advantages over any other method to estimate the long memory parameter.

The variance $(1/4m)$ associated with this estimator happens to be smaller, almost by 40%, than the variance $(\pi^2/24m)$ of the log-periodogram estimator. In the case of Gaussian data generating process, Robinson (1995) calls the value $(1/4)$ as the efficient bound on the m times the variance of the semiparametric estimation of d.

In the final section of this chapter, we briefly discuss the effect of tapering on the polynomial trend of order r of an ARFIMA process. Such a time series has nonstationary memory parameter, d greater than .5.

2.3 Tapering and ARFIMA processes with stochastic trends

Our interest in the stochastic trends of long memory time series is limited to the integration of order 1. However, we study the general case in this section. Let the observed sequence X_t has memory parameter $d > -1/2$ where $t = 1, 2, \cdots, n$. If $r = \lfloor d + 1/2 \rfloor$, then $\nabla^r X_t = Y_t^{(r)}$ is stationary with mean zero, and the spectral density $f_{Y^{(r)}}(\lambda)$ behaving as $\lambda^{-2(d-r)}$ around the origin, $d - r \in [-1/2, 1/2)$. The function $f_X(\lambda)$, which is not integrable, can be defined by *the unit root transfer function*(cf. Chapter 4) and the spectral

density $f_{Y^{(r)}}(\lambda)$ of the stationary sequences $Y_t^{(r)}$

$$f_X(\lambda) = |1 - \exp(-i\lambda)|^{-2d} f_{Y^{(r)}}(\lambda).$$

By extending the discussion of Hurvich and Ray (1995), Velasco (1999a), and Velasco (1999b) we can write X_t in terms of the random variables X_0 and $Y_0^{(s)}$, $s = 1, 2, ..., r-1$, which do not depend on time,

$$\begin{aligned}
X_t &= X_0 + \nabla X_1 + \nabla X_2 + \cdots + \nabla X_t \\
X_t &= X_0 + \nabla X_0 + \sum_{j=1}^{1} \nabla^2 X_j + \nabla X_0 + \sum_{j=1}^{2} \nabla^2 X_j + \cdots + \nabla X_0 + \sum_{j=1}^{t} \nabla^2 X_j \\
&= X_0 + \sum_{j_1=1}^{t} \left(\nabla X_0 + \sum_{j_2=1}^{j_1} \nabla^2 X_{j_2} \right) \\
&= X_0 + \sum_{s=1}^{r-1} \nabla^s X_0 \, p^{(s)}(t) + \sum_{j_1=1}^{t} \sum_{j_2=1}^{j_1} \cdots \sum_{j_r=1}^{j_{r-1}} \nabla^r X_{j_r} \\
&= X_0 + \sum_{s=1}^{r-1} Y_0^{(s)} \, p^{(s)}(t) + \sum_{j_1=1}^{t} \sum_{j_2=1}^{j_1} \cdots \sum_{j_r=1}^{j_{r-1}} Y_{j_r}^{(r)}
\end{aligned}$$

where $p^{(s)}(t)$ are polynomials in t of order s:

$$\begin{aligned}
p^{(1)}(t) &= t, \\
p^{(2)}(t) &= \sum_{j=1}^{t} p^{(1)}(j) = \frac{t(t+1)}{2}, \\
p^{(3)}(t) &= \sum_{j=1}^{t} p^{(2)}(j) = \sum_{j=1}^{t} \frac{j(j+1)}{2}
\end{aligned}$$

and so forth.

The tapered Fourier transform of X_t for $t = 1, \cdots, n$ with the taper

function $h(t)$, can then be written in the following form

$$\begin{aligned} J_X^T(\lambda) &= \frac{1}{\sqrt{\sum h_t^2}} \sum_{t=1}^{n} h_t X_t \exp(-i\lambda t) \\ &= \frac{1}{\sqrt{\sum h_t^2}} \sum_{t=1}^{n} h_t \left(X_0 + \sum_{s=1}^{r-1} Y_0^{(s)} p^{(s)}(t) \right) \exp(-i\lambda t) + \\ &\quad \frac{1}{\sqrt{\sum h_t^2}} \sum_{t=1}^{n} h_t \sum_{j_1=1}^{t} \sum_{j_2=1}^{j_1} \ldots \sum_{j_r=1}^{j_{r-1}} Y_{j_r}^{(r)} \exp(-i\lambda t). \end{aligned}$$

The first term in the above equation is treated as a nuisance term because it includes the information of $\{X_t\}$ from the past prior to $t = 1$. To make inferences about d we need to eliminate this dependence on the past. One solution could be to select a taper with weights such that the following expression

$$\sum_{t=1}^{n} h_t \left(1 + t^1 + t^2 + \cdots + t^{r-1}\right) \exp(-i\lambda t)$$

is zero or should approach zero with a suitable rate as $n \to \infty$ for any given λ, which may be sufficient to cancel the contribution of the polynomials in t with unknown coefficients to the Fourier transform $J_X^T(\lambda)$.

In the case of stochastic trend of integration order 1, $r = 1$, we only require

$$\sum_{t=1}^{n-1} h_t \exp(-i\lambda t) = 0.$$

The cosine bell tapered Fourier transform satisfy this condition because $\sum_{t=1}^{n} h_t \exp(-i\lambda t) = O(1/n^2)$ as derived below. In fact this taper is useful for $r = 1$ and may be for $r = 2$. Special type of tapers are defined and constructed in Velasco (1999a) for the log periodogram method estimation, and Hurvich and Chen (1998), and Velasco (1999b) for the semiparametric method of nonstationary memory parameter d.

In the following we provide the behavior of the Fourier transform of cosine bell taper $h_t = (1 - \cos(\lambda_1 t))/2$, where $\lambda_1 = 2\pi/n$, as $n \to \infty$. One can also refer Chapter 6 in Bloomfield (2000) for an excellent discussion on the behavior of the Fourier transform of cosine bell taper but our approach here is straightforward.

$$\begin{aligned}
\sum_{t=0}^{n-1} h_t \exp(-i\lambda t) &= \sum_{t=0}^{n-1} \left(\frac{1-\cos(\lambda_1 t)}{2}\right) \exp(-i\lambda t) \qquad (2.31)\\
&= \frac{1}{2}\sum_{t=0}^{n-1} \left(1 - \frac{\exp(i\lambda_1 t) + \exp(-i\lambda_1 t)}{2}\right) \exp(-i\lambda t)\\
&= \frac{1}{4}\sum_{t=0}^{n-1} (2 - \exp(i\lambda_1 t) - \exp(-i\lambda_1 t)) \exp(-i\lambda t)\\
&= \sum_{t=0}^{n-1} \frac{1}{2}\exp(-i\lambda t) - \sum_{t=0}^{n-1} \frac{1}{4}\exp(-i(\lambda-\lambda_1)t)\\
&\quad - \sum_{t=0}^{n-1} \frac{1}{4}\exp(-i(\lambda+\lambda_1)t) \qquad (2.32)
\end{aligned}$$

Each of the three parts in the RHS of the above equation can be simplified

$$\begin{aligned}
\sum_{t=0}^{n-1} \frac{1}{2}\exp(-i\lambda t) &= \frac{1}{2}\exp\left(-i\lambda\frac{(n-1)}{2}\right)\frac{\sin\left(\frac{\lambda n}{2}\right)}{\sin\left(\frac{\lambda}{2}\right)},\\
\sum_{t=0}^{n-1} \frac{1}{4}\exp(-i(\lambda-\lambda_1)t) &= \frac{1}{2}\exp\left(-i(\lambda-\lambda_1)\frac{(n-1)}{2}\right)\frac{\sin\left(\frac{(\lambda-\lambda_1)n}{2}\right)}{\sin\left(\frac{(\lambda-\lambda_1)}{2}\right)},\\
\sum_{t=0}^{n-1} \frac{1}{4}\exp(-i(\lambda+\lambda_1)t) &= \frac{1}{2}\exp\left(-i(\lambda+\lambda_1)\frac{(n-1)}{2}\right)\frac{\sin\left(\frac{(\lambda+\lambda_1)n}{2}\right)}{\sin\left(\frac{(\lambda+\lambda_1)}{2}\right)}.
\end{aligned}$$

One can write, using the trigonometric additional formulae,

$$\begin{aligned}\sin\left(\frac{(\lambda-\lambda_1)}{2}\right) &= \sin(\lambda/2)\cos(\lambda_1/2) - \sin(\lambda_1/2)\cos(\lambda/2),\\ \sin\left(\frac{(\lambda+\lambda_1)}{2}\right) &= \sin(\lambda/2)\cos(\lambda_1/2) + \sin(\lambda_1/2)\cos(\lambda/2).\end{aligned}$$

For large n, noting that $\lambda_1 = 2\pi/n$, the above two equations can be simplified as

$$\sin\left(\frac{(\lambda-\lambda_1)}{2}\right) \approx \sin(\lambda/2) - \frac{\pi}{n}\cos(\lambda/2), \tag{2.33}$$

$$\sin\left(\frac{(\lambda+\lambda_1)n}{2}\right) \approx \sin(\lambda/2) + \frac{\pi}{n}\cos(\lambda/2). \tag{2.34}$$

Then, for large n and $\lambda \neq \lambda_1$

$$\exp\left(-i(\lambda-\lambda_1)\frac{(n-1)}{2}\right) \approx \exp\left(-i\lambda\frac{(n-1)}{2}\right), \tag{2.35}$$

$$\exp\left(-i(\lambda+\lambda_1)\frac{(n-1)}{2}\right) \approx \exp\left(-i\lambda\frac{(n-1)}{2}\right). \tag{2.36}$$

The above approximations can also be derived by writing $\exp\left(-i(\lambda-\lambda_1)(n-1)/2\right)$ $= \cos\left((\lambda-\lambda_1)(n-1)/2\right) - i\sin\left((\lambda-\lambda_1)(n-1)/2\right)$ and using the trigonometric additional formulae.

Hence the RHS of (2.31) can be approximated by,

$$\frac{1}{2}\exp\left(-i\lambda(n-1)/2\right)\left(\frac{1}{\sin(\lambda/2)} - \frac{.5}{\sin(\lambda/2) - \frac{\pi}{n}\cos(\lambda/2)} - \frac{.5}{\sin(\lambda/2) + \frac{\pi}{n}\cos(\lambda/2)}\right).$$

After some algebra, this approximation can be further simplified to

$$\frac{1}{2}\exp\left(-i\lambda(n-1)/2\right)\left(\frac{-\frac{\pi^2}{n^2}\cos^2(\lambda/2)}{\sin^3(\lambda/2) - \frac{\pi^2}{n^2}\cos^2(\lambda/2)}\right).$$

Hence for large n

$$\sum_{t=0}^{n-1}\left(\frac{1-\cos(2\pi t/n)}{2}\right)\exp(-i\lambda t) \approx \frac{\pi^2}{2n^2}\exp\left(-i\lambda(n-1)/2\right) \times\left(\frac{\cos^2(\lambda/2)}{\sin^3(\lambda/2)}\right) \quad (2.37)$$

$$= O(1/n^2). \quad (2.38)$$

From (2.32), we can easily understand that the cosine bell tapered discrete Fourier transform of a time series X_t, $J_X^T(\lambda_j)$, at the discrete Fourier frequency $\lambda_j = 2\pi j/(n-1)$ is the weighted average of the ordinary discrete Fourier transforms $J_X(\lambda_{j-1}), J_X(\lambda_j), J_X(\lambda_{j+1})$ with weights $(-1/4, 1/2, -1/4)$. Then,

$$J_X^T(\lambda_j) = -\frac{1}{4}J_X(\lambda_{j-1}) + \frac{1}{2}J_X(\lambda_j) - \frac{1}{4}J_X(\lambda_{j+1}). \quad (2.39)$$

It is interesting to note that as $n \to \infty$, and at higher frequencies, $J_X^T(\lambda_j) \approx 0$.

Velasco (1999b) has showed that the cosine bell tapered semiparametric estimate $\hat{d}$ of nonstationary memory parameter $d \in [.5, 1.5)$ still has the asymptotic Gaussian distribution

$$m^{1/2}(\hat{d} - d) \to \mathcal{N}\left(0, \frac{35}{18} \times 1/4\right) \quad (2.40)$$

but with larger variance. The increase in the variance is attributed to the cosine bell taper. Similar result has also derived for the tapered log-periodogram estimate in Velasco (1999a).

Chapter 3

Time-varying ARFIMA models

Similar to the tv-ARMA(p,q) models as the generalization of ARMA(p,q) models, we define the tv-ARFIMA(p,d,q) models as the generalization of ARFIMA(p,d,q) with $d \in [-.5, .5)$. Such models are necessary to capture the time-varying properties of long memory of the EEG data over different locally stationary segments in addition to the short memory properties. In this dissertation, we assume the long memory d as a function of u, where $u \in [0, 1]$.

Consider the simplest tv-ARFIMA(p,d,q) scenario, *the time-varying fractionally integrated noise* $X_{t,n}$ is defined as

$$\nabla^{d(t/n)} X_{t,n} = \sigma\left(\frac{t}{n}\right)\varepsilon_t, \tag{3.1}$$

where ε_t is the i.i.d noise with mean zero and variance 1, and $\nabla^{(.)}$ is the fractional differencing operator as defined in Section 4.1. Then, the transfer function and the spectral density function of $X_{t,T}$ can be easily derived just

as in the case of fractionally integrated noise (cf. Sections 2.2)

$$X_{t,n} = \sigma\left(\frac{t}{n}\right)\nabla^{-d(t/n)}\varepsilon_t \tag{3.2}$$

$$= \sigma\left(\frac{t}{n}\right)\sum_{j=0}^{n}\psi_{t,n,j}\,\varepsilon_{t-j} \tag{3.3}$$

where $\nabla^{-(.)}$ is the fractional integration operator (cf. Section 4.1), with

$$\psi_{t,n,j} = \frac{\Gamma(j+d(t/n))}{\Gamma(j+1)\Gamma(d(t/n))} = \prod_{0<l\leq j}\frac{l-1-d(t/n)}{l},\ j=0,1,2,\cdots.$$

By applying the Stirling's formula, $\Gamma(x)\sim\sqrt{2\pi}\ \exp(-x+1)(x-1)^{x-1/2}$ as $x\to\infty$, to $\psi_{t,n,j}$, we obtain

$$\psi_{t,n,j}\sim\frac{j^{d(t/n)-1}}{\Gamma(d(t/n))}\quad\text{as } j\to\infty.$$

Therefore, the sequence $\{\psi_{t,n,.}\}$ is square summable , $\sum_{j=0}^{\infty}\psi_{t,n,j}^2<\infty$, for $-.5\leq d(.)<.5$.

Hence, the *time-varying unit root transfer function* is defined as the discrete Fourier transform of $\psi_{t,n,j}$ as $n\to\infty$, just as in the case of fractionally integrated noise (cf. Brockwell and Davis, 1996, Theorem 13.2.1),

$$\sum_{j=0}^{n}\psi_{t,n,j}\exp(-i\omega j)\to(1-\exp(-i\lambda))^{-d(t/n)}\quad\text{as } n\to\infty. \tag{3.4}$$

Since $X_{t,n}$ is obtained from ε_{t-j} by the application of linear filter $\{\psi_{t,n,j}\}$ [cf. Brillinger(1981), and Brockwell and Davis(1996)], the transfer function and the spectral density function of time-varying fractionally integrated noise can be derived as

$$A(u,\lambda) = \frac{\sigma(u)}{\sqrt{2\pi}}\left(1-e^{-i\lambda}\right)^{-d(u)}, \tag{3.5}$$

$$f(u,\lambda) = |A(u,\lambda)|^2 = \frac{\sigma^2(u)}{2\pi}\left|1-e^{-i\lambda}\right|^{-2d(u)}. \tag{3.6}$$

where $1/\sqrt{2\pi}$ and $1/2\pi$ are the transfer function and the spectral density function of ε respectively.

Next, we prove that a tv-ARFIMA(0,d,0) is a locally stationary process. Since ε_t is a i.i.d sequence, its Cramer representation is as follows:

$$\varepsilon_t = \int_{-\pi}^{\pi} \exp(i\lambda t)\ (2\pi)^{-1/2}\ d\xi(\lambda).$$

Then,

$$\nabla^{-d(t/n)}\varepsilon_t = \int_{-\pi}^{\pi} \exp(i\lambda t)\ \ |1 - exp(-i\lambda)|^{-d(t/n)}\ (2\pi)^{-1/2}\ d\xi(\lambda).$$

Hence $X_{t,n}$ has the transfer function

$$A^0_{t,n}(\lambda) = \frac{\sigma(\frac{t}{n})}{\sqrt{2\pi}}\,(1 - exp(-i\lambda))^{-d(t/n)}\,, \tag{3.7}$$

and

$$\sup_{t,\lambda}\left|A^0_{t,n}(\lambda) - A\left(\frac{t}{n},\lambda\right)\right| = 0 \leq Kn^{-1}.$$

Therefore, the time-varying fractionally integrated noise is locally stationary. This result can be extended to a general tv-ARFIMA(p,d,q) models by following the scheme of the proof of Theorem 2.3 in Dahlhaus (1996), which asserts that a time-varying ARMA(p,q) process is locally stationary.

Now, we extend time-varying fractionally integrated noise to more general time-varying ARFIMA(p,d,q) process.

Definition: The process $X_{t,n}$ is said to be a tv-ARFIMA(p,d,q) process, if it is the solution of the system of stochastic difference equations

$$\sum_{j=0}^{p} a_j\left(\frac{t}{n}\right)\nabla^{d(t/n)}\left(X_{t-j,n} - \mu\left(\frac{t-j}{n}\right)\right) = \sigma\left(\frac{t}{n}\right)\sum_{k=0}^{q} b_k\left(\frac{t}{n}\right)\varepsilon_{t-k}, \tag{3.8}$$

where $a_0(u) = 1$, $b_0(u) = 1$, $-.5 \leq d(.) < .5$ and the ε_t are independent random variables with mean zero and variance one.

In the following theorem, we will derive the transfer function $A(u, \lambda)$ and the spectral density of $f(u, \lambda)$ of $X_{t,n}$ with mean zero using the linear filter theory.

Theorem 3.1 *Let $Y_{t,n}$ be the the solution of the system of difference equations $\sum_{j=0}^{p} a_j\left(\frac{t}{n}\right) Y_{t-j,n} = \sigma\left(\frac{t}{n}\right) \sum_{k=0}^{q} b_k\left(\frac{t}{n}\right) \varepsilon_{t-k}$. Let $X_{t,n} = \nabla^{-d(t/n)} Y_{t,n}$, where $\nabla^{-d(t/n)}$ is as defined before with $d(.) \in [-.5, .5)$. Then, the transfer function and the spectral density function of $X_{t,n}$ are*

$$A(u, \lambda) = \frac{\sigma(u)}{\sqrt{2\pi}} \frac{\sum_{k=0}^{q} b_k(u) \exp(-i\lambda k)}{\sum_{j=0}^{p} a_j(u) \exp(-i\lambda j)} \left(1 - \exp(-i\lambda)\right)^{-d(u)}, \tag{3.9}$$

$$f(u, \lambda) = \frac{\sigma^2(u)}{2\pi} \frac{\left|\sum_{k=0}^{q} b_k(u) \exp(-i\lambda k)\right|^2}{\left|\sum_{j=0}^{p} a_j(u) \exp(-i\lambda j)\right|^2} \left|1 - \exp(-i\lambda)\right|^{-2d(u)}, \tag{3.10}$$

where $u = \frac{t}{n} \in [0, 1]$.

Proof.

$$\begin{aligned} X_{t,n} &= \nabla^{-d(t/n)} Y_t \\ &= \sum_{j=0}^{\infty} \psi_{t,n,j}\, Y_{t-j,n} \end{aligned}$$

Then, using the linear filter theory, the transfer function of $X_{t,n}$ can be written

$$A(u, \lambda) = \frac{\sigma(u)}{\sqrt{2\pi}} \frac{\sum_{k=0}^{q} b_k(u) \exp(-i\lambda k)}{\sum_{j=0}^{p} a_j(u) \exp(-i\lambda j)} \left(1 - \exp(-i\lambda)\right)^{-d(u)},$$

where $\frac{\sigma(u)}{\sqrt{2\pi}} \frac{\sum_{k=0}^{q} b_k(u) \exp(-i\lambda k)}{\sum_{j=0}^{p} a_j(u) \exp(-i\lambda j)}$ is the transfer function of $Y_{t,n}$ and $\left(1 - \exp(-i\lambda)\right)^{-d(u)}$ is the time-varying unit root transfer function.

Since, $f(u,\lambda) = |A(u,\lambda)|^2$, the second part of the theorem is also proved.

$\diamond$

3.1 Estimation of time-varying ARFIMA model parameters

The parameters of a locally stationary tv-ARFIMA model can be estimated directly using the generalized Whittle likelihood function

$$\mathcal{L}_n(\theta) = \frac{1}{4\pi}\frac{1}{M}\sum_{j=1}^{M}\int_{-\pi}^{\pi}\left\{\log\ f_\theta(u_j,\lambda) + \frac{I_N(u_j,\lambda)}{f_\theta(u_j,\lambda)}\right\}d\lambda, \tag{3.11}$$

where $f_\theta(u_j,\lambda)$ is defined in (3.10). Here, θ includes a total of $M\times(p+q+2)$ parameters: AR and MA parameters $\{a_1(u_j),\cdots,a_p(u_j)\}$ and $\{b_0(u_j),\cdots,b_q(u_j)\}$ respectively, $\sigma^2(u_j)$, and the $d(u_j)$, where $u_j \in [0,1]]$ and $j = 1,\cdots,M$.

One can try to minimize (3.11) but, according to our preliminary results, as M (the number of segments) increases, obtaining the convergence of the generalized Whittle likelihood function in (3.9) is almost impossible because of the large number of parameters and the computational effort required. Another drawback is that the minimization of this process depends on the initial values of the parameters given to the optimization routine.

As we have seen in the first chapter and will see again in Chapter 5, part of the nonstationarity of our sleep EEG signals can be characterized by the stochastic trends (the order of integration, r=1). Therefore, the memory parameter d of our EEG data is nonstationary and it is in the range of $[.5, 1.5)$. However, this trend is not the same over the entire range of the data. Some segments of the data display no trend (r=0) and other segments of the data display stochastic trend of order 1 (r=1). Hence it is important that we estimate d accurately before we proceed to estimate the AR and MA parameters. It can be seen in Chapter 5 of this dissertation that the

short memory component of the example EEG data may be characterized by locally stationary AR model with $p \leq 2$.

In the case of stochastic trend of order 1, it is a general practice to difference the data once before we apply stationary models to fit the data by assuming that the local mean is zero. But in our example case stochastic trend of order 1 can be seen only on some segments of the data. Hence the model fitting algorithm needs to identify the segments where the stochastic trend is of order 1 and difference only these segments data once. Identifying the trends and memory parameters over segments may not be an impossible task but requires large amount of computational time and effort. If overlapping between the segments is allowed then it becomes cumbersome.

The algorithm to fit the EEG data must estimate the stochastic trends of individual data segments either separately from the memory parameter or as part of the memory parameter. We considered the later case because it is easy to implement. The former case requires identifying the presence of stochastic trends over each individual segments and calculating them separately. In the later case, the memory, which includes long memory and stochastic trend of order 1, is $d(u) \in [-.5, 1.5)$ for any $u \in [0, 1]$, where the presence of stochastic trend of order 1 of the segments is represented by $d(u) > .5$. However, when $d(u) > .5$ for a given segment then the spectral density of this segment does not exist in the usual sense, because the spectral density function of this process is not integrable. However, suggested by Hurvich and Ray (1995) the $f(\lambda)$ can be interpreted as the limit of the expected value of the periodogram. The $f(\lambda)$ in this case is called *pseudo* spectral density. But we ignore the prefix, pseudo, in the rest of this thesis.

A drawback of not differencing the data initially induces bias in the estimators of d. One way to reduce the bias is to apply proper tapering to the data. But tapering increases bias in estimators of the short memory parameters and also increases the standard errors of both short and long memory

parameters' estimators (Velasco and Robinson, 2000). At the same time, for tapered data, the memory parameter $d \in [-.5, 1.5)$ estimators are asymptotically normal and are consistent. On the other hand, for the *untapered* data, the memory parameter estimators are asymptotically normal for $d < 1$ and are consistent for $d < 3/4$.

Instead of using the generalized Whittle method directly, we first estimate $d(u)$ by semiparametric method separately for each of the tapered segments. Then, we filter out this memory component of the data from each of the original (untapered) data segments to obtain the short memory component of the EEG signal. The filtering technique is provided in Chapter 4. We again segment (possibly different segment length) the short memory time series, and reapply the taper to the segments. Finally, we estimate the time-varying short memory parameters by the generalized Whittle method. We call this estimation procedure the *2-step* procedure. In the following section, we outline the sleep EEG data fitting algorithm.

3.2 An algorithm to fit EEG time series

In this section we apply the time-varying ARFIMA model with memory parameter $d \in [-.5, 1.5)$ to fit a nonstationary EEG, $X_{t,n}$.

Let $X_{t,n}$ be the solution of the system of the time-varying ARFIMA(p,d,q) model difference equations

$$\sum_{j=0}^{p} a_j\left(\frac{t}{T}\right) \nabla^{d(u)} X_{t-j,n} = \sigma\left(\frac{t}{T}\right) \sum_{k=0}^{q} b_k\left(\frac{t}{T}\right) \epsilon_{t-k} \quad t = 0, 1, \cdots,$$

with $d(u) \in [-5, 1.5)$ Then the pseudo spectral density function of $X_{t,T}$

$$f(u,\lambda) = \frac{\sigma^2(u)}{2\pi} \left| \frac{\sum_{k=0}^{q} b_k(u) \exp(i\lambda k)}{\sum_{j=0}^{p} a_j(u) \exp(i\lambda j)} \right|^2 \left|1 - e^{i\lambda}\right|^{-2d(u)}.$$

If $d(u) \in [-5, .5)$, $X_{t,n}$ is a locally stationary time-varying ARFIMA process with the usual spectral density $f(u, \lambda)$.

Algorithm to fit sleep EEG time series :

1. Segment the data into M windows with window size N, possibly without overlapping, and remove local mean of each segment. We shall call these locally mean subtracted data as *the original data* and apply the cosine bell taper to each original data segment.

   In the EEG example, the segments (one minute windows) are already defined by the sleep states. If the segments are not well defined one may consider the following simple and efficient segmenting procedure. For different window or segment sizes, calculate their skewness and kurtosis values and select the largest window size (N) for which the average of all the segments skewness values is close to zero, and the average of all the segments kurtosis values is close to 3 simultaneously. This helps segments to have the Gaussian-like properties. One may write here an optimization procedure to create Gaussian-like windows. Piryatinska (2004) provided an efficient procedure in frequency domain to identify the sleep states of the test EEG.

2. This step evaluates the memory parameter of each segment. We estimate each segment's memory parameter $d(u_j) \in [-.5, 1.5)$ by semiparametric method (cf: Section 2.2). We generate the confidence bounds for $d(u)$ by smoothing point wise confidence interval of $d(u)$ as discussed in Section 2.2.3. These confidence intervals are valid only in the case of Gaussian data generating process.

3. In this step, we filter $\widehat{d}(u_j)$ out from the j-th segment of the original data to obtain short memory component of EEG. Let's consider one segment

of length N and denote it by X_t. If X_t is ARFIMA(p,d,q), $X_t = \nabla^d Y_t$, where Y_t is ARMA(p,q), then $J_X(\lambda) = (1-\exp(-i\lambda))^{-d} J_Y(\lambda) + R_N(\lambda)$, where $J_X(.)$ and $J_Y(.)$ are the discrete Fourier transforms of $\{X_t\}$ and $\{Y_t\}$ respectively. Ignoring the error term $R_N(\lambda)$, which is of order of magnitude $o(N^{1+2d})$ (cf: Chapter 4), and replacing d by $\hat{d}$, we obtain the approximate relation $J_Y(\lambda) = (1-\exp(-i\lambda))^{\widehat{d}} J_X(\lambda)$. We now apply the inverse Fourier transform on both sides and then obtain

$$\tilde{Y}_t = \frac{1}{2\pi N} \sum_{j=1}^{N} \exp(i\lambda_j t) \left(1 - \exp(-i\lambda_j)\right)^{\widehat{d}} J_X(\lambda_j), \quad t = 0, 1, \cdots, N-1.$$

We repeat this procedure for all the original data segments for given $\widehat{d}(u_j)$. If the segment length is *large*, the obtained $\tilde{Y}_{t,T}$ is *approximately* time-varying ARMA(p,q) because of $o(N^{1+2d})$ convergence. Hence, it is important to select N large, may be in thousands.

4. We again segment the short memory $\tilde{Y}_{t,T}$ data and apply taper to each segment. The segment length N could be different from the segment we have selected in Step 1. Selecting smaller N in this step captures the local behavior of EEG better. Then, we use the following *approximate* version of the Generalized Whittle method to obtain the estimates of the parameters of tv-ARMA(p,q) model. In this step, we estimate AR and MA parameters $\{a_1(u_j), \cdots, a_p(u_j)\}$ and $\{b_0(u_j), \cdots, b_q(u_j)\}$ respectively, and $\sigma^2(u_j)$.

   Instead of minimizing

$$\mathcal{L}_T(\theta) = \frac{1}{4\pi} \frac{1}{M} \sum_{j=1}^{M} \int_{-\pi}^{\pi} \left\{ \log\ f_\theta(u_j, \lambda) + \frac{I_N(u_j, \lambda)}{f_\theta(u_j, \lambda)} \right\} d\lambda,$$

   we minimize

$$\widetilde{\mathcal{L}}_T(\theta) = \frac{1}{4\pi} \frac{1}{M} \sum_{j=1}^{M} \int_{-\pi}^{\pi} \left\{ \frac{I_N(u_j, \lambda)/c_N(u_j, 0)}{g_\theta(u_j, \lambda)} \right\} d\lambda \tag{3.12}$$

where $c_N(u_j, 0)$ is the local variance of the data in jth segment and $g(.)$ is the *power spectrum*,

$$g_\theta(u_j, \lambda) = \left| \sum_{k=0}^{q} b_k(u) \exp(i\lambda k) / \sum_{j=0}^{p} a_j(u) \exp(i\lambda j) \right|^2 . \qquad (3.13)$$

The quantity $I_N(u_j, \lambda)/c_N(u_j, 0)$ is called *normalized periodogram* (Mikosch et al., 1995) .

In the following Section, we provide a brief introduction of alpha-stable random processes. We also go through the importance of (3.13) for a locally stationary fractional AR processes generated by the independent innovations with infinite variance.

3.3 Generalized Whittle likelihood function for α-stable random processes

Any time series which exhibits sharp spikes and or occasional bursts of outlying observations suggests that the innovations of this time series has infinite variance property. The alpha-stable or α-stable distribution can be a treated as a special type of heavy-tailed distribution with infinite variance. See, Adler, Feldman, and Taqqu (1998) for a broad review on heavy-tail random processes and Taqqu and Samorodnitsky (1994) for a complete review on α-stable random processes.

A random variable X said to have stable distribution if there are parameters $0 < \alpha < 2$, $\sigma \geq 0$, $-1 \leq \beta \leq 1$, and μ real such that its characteristic function (moment generating function) has the form

$$E(\exp(itX)) = \begin{cases} \exp\left\{-\sigma^\alpha |t|^\alpha \left(1 - i\beta(\text{sign } t) \tan\frac{\pi\alpha}{2}\right) + i\mu t\right\}, & \alpha \neq 1, \\ \exp\left\{-\sigma^\alpha |t|^\alpha \left(1 - i\beta(\text{sign } t) \ln|t|\right) + i\mu t\right\}, & \alpha = 1, \end{cases}$$

where sign$(x) = 1$ if $x > 0$, 0 if $x = 0$, -1 if $x < 0$.

In general, the tails of heavy tail distributions behave $P(X > x) = t(x)$, where $t(x) < O(1/x^{\alpha})$ and $0 < \alpha$. For α-stable distributions, α is in between 0 and 2, such that $t(x) \sim 1/(x^{\alpha}w(x))$, where $w(x)$ is slowly varying process.

If $\alpha = 2$, then X is a Gaussian process. For convenience, we include the Gaussian case as part of stable processes in rest of this dissertation.

If X is stable with $\beta = \mu = 0$, then we call X a symmetric stable random process with index α, denoted by $S_{\alpha}(\sigma, 0, 0)$. A more general α-stable random process is denoted by $S_{\alpha}(\sigma, \beta, \mu)$, where σ is the scale parameter, β is the *skewness type* parameter, and μ is the location parameter. If X is $S_{\alpha}(\sigma, \beta, \mu)$, then the β parameter can be understood as the the limiting value of the ratio of the difference of the tail probabilities to the sum of the tail probabilities:

$$\beta = \lim x \to \infty \ \frac{1 - P(X > x) - P(X < -x)}{1 - P(X > x) + P(X < -x)}, \ \alpha \neq 2.$$

When β is positive, the distribution is skewed right. When it is negative, it is skewed left. When it is zero, the distribution is symmetrical. As $\alpha \to 2$, β loses its effectiveness and the distribution approaches the normal distribution regardless of β.

The estimation of stable parameters is made easy by the MuCulloch(1986) method. This simple method (Appendix A), based on fitting tabulated quantiles of stable distributions, is well suited for $\alpha \in [.6, 2.0]$ and $\beta \in [-1, 1]$ (which covers most of the cases met in practice) and all values of the other parameters. The estimation of the location parameter, may be inaccurate near $\alpha = 1$. The McCulloch estimators were originally designed for i.i.d data.

An important property of α-stable distributions is that the moments of order greater than or equal to α do not exist. Therefore, variances and covariances do not exist in the usual sense for a stable process with $\alpha < 2$. Due to this reason, if a stationary (or a locally stationary) process generated by infinite variance innovations, then the Whittle (or the generalized Whittle)

estimators do not make sense. However, one prefers the type of estimators which could be used to estimate parameters of a stationary (or a locally stationary) process generated by innovations with any $\alpha \in (0, 2)$. Doing so has the advantage of not having to determine beforehand the often unknown distributions of the innovations. Mikosch et al. (1995) prefers to estimate α parameter from the residuals of the time series rather than from the time series itself. It is necessary, however, to verify that these estimators have good properties, namely consistency, in the infinite variance case as well.

Mikosch et al. (1995) derived such consistent estimators by introducing the concept of *normalized periodogram* for ARMA processes generated by independent stable innovations with $\alpha \in (0, 2)$. It is not a priori clear the point of defining a periodogram for these processes where the second moments do not exist, let alone using it as a tool for parameter estimation. Nevertheless, there is no intrinsic problem involved in defining it for the real data. Kluppelberg and Mikosch (1993 and 1994) proved that the periodograms of stable processes converge in distribution with a rate $1/n^{\alpha}$ and also that the periodograms of stable processes are correlated.

The same authors introduced the concept of normalized periodograms for the stationary ARMA processes, which are the usual periodograms scaled by variance of the data and proved that the normalized periodograms converge weakly in distribution. Furthermore, they proved that normalized periodograms are asymptotically uncorrelated and they used these periodograms to obtain consistent estimator for the power spectrum.

Kokoszka and Taqqu (1996) extended Mikosch et.al estimators to derive consistent estimators of the parameters of stable ARFIMA processes. The Whittle likelihood function in stable ARFIMA case can be written as

$$\widetilde{\mathcal{L}}_T(\theta) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \frac{I_N(\lambda)/c_N(0)}{g_\theta(\lambda)} \right\} d\lambda, \tag{3.14}$$

where $c_N(0)$ is the variance of the data,

$$g(\lambda) = \frac{\left|\sum_{k=0}^{q} b_k \exp(i\lambda k)\right|^2}{\left|\sum_{j=0}^{p} a_j \exp(i\lambda j)\right|^2}(1-\exp(i\lambda))^{-2d},$$

and $d \in (0, .5)$

Similar objective function without normalization of the periodogram and with usual spectrum is also discussed by Brockwell and Davis (1996) for stationary ARFIMA processes with innovations with finite variance. They suggest that the estimators of a stationary ARFIMA process derived by the minimization of such objective function are slightly biased but in the case of Gaussian innovations they are similar to the Whittle likelihood estimators.

The asymptotics for the estimators of parameters of stationary time series generated by infinite variance innovations do not exist in the usual sense. This is the main reason, for not emphasizing the asymptotics of parameter estimators in this dissertation. We suspect that some of the sleep states in the test EEG have infinite variance innovations (see Chapter 5).

In the remainder of this section we derive estimators for the time varying AR process with infinite variance by the generalized Whittle likelihood function with normalized periodogram.

Let $X_{t,n}$ be a locally stationary AR process with α-stable innovations, where the stable parameter is the function of segment $j = 1, \cdots, M$ with midpoint u_j. Then, $X_{t,T}$ is the solution of the following system equations

$$\sum_{j=0}^{p} a_j\left(\frac{t}{n}\right) X_{t-j,n} = \sigma\left(\frac{t}{n}\right)\varepsilon_t \quad \text{for } t = 0, 1, \cdots, n-1, \tag{3.15}$$

where $a_0(u) = 1$ and $\{\varepsilon_t\}$ are independent α-stable random variables with α is in $(0, 2)$. We are including the Gaussian innovations ($\alpha = 2$) in this equation, because the following estimators for the time-varying AR parameters are valid for both stable and Gaussian innovations.

Let $\theta = \{a(u_j)\}$ parameter of interest. Then, the estimator of $a(u_j)$ derived from the minimization of

$$\widetilde{\mathcal{L}}_T(\theta) = \frac{1}{4\pi}\frac{1}{M}\sum_{j=1}^{M}\int_{-\pi}^{\pi}\left\{\frac{I_N(u_j,\lambda)/c_N(u_j,0)}{g_\theta(u_j,\lambda)}\right\}d\lambda \tag{3.16}$$

is

$$\hat{a}(u_j) = -\Sigma_N^{-1}(u_j)C_N(u_j). \tag{3.17}$$

We do not provide explicit derivation of these estimators because the algebra here is similar to the algebra in the finite variance case (see Section 2.1.3).

These estimators are the same Yule-Walker estimators in the finite variance case. In the finite variance case, if one minimizes $\widetilde{\mathcal{L}}_T$ to estimate $a(u_j)$, then $\sigma^2(u_j)$ can be estimated by calculating (i) calculating residuals $(X_{t,n} + \hat{X}_{t,n})$, where $\hat{X}_{t,n} = -\sum_{j=0}^{p}\hat{a}_j(u)X_{t-j,n}$ and evaluating (ii) $E(X_{t,n} - \hat{X}_{t,n})^2$. This simple procedure gives

$$\hat{\sigma}^2(u_j) \;=\; c_N(u,0) - C_N'(u_j)\Sigma_N^{-1}(u_j)C_N(u_j) \tag{3.18}$$

which is the same Yule-Walker estimator of $\sigma^2(u_j)$ (see Section 2.1.3). In the Gaussian case, the asymptotics of these estimators are same as (2.22) .

Chapter 4

Filtering the memory parameter

The aim of this chapter is to find a simple and efficient procedure to filter out the long memory or simply the memory parameter from the ARFIMA(p,d,q) time series. The effect of such a filter will be to remove the slowly varying memory and retain rapidly varying memory of a time series.

Such a filter would help to separate the parameter estimation of ARFIMA(p,d,q) into two steps with lesser computational complexity. The first step of this 2-step procedure is the estimation of the memory parameter and the second step is the estimation of ARMA(p,q) parameters. The second step could be regarded as the conditional estimation of ARMA(p,q) parameters by conditioning on the memory parameter, which is estimated from the first step. In general the long memory parameter and the ARMA(p,q) parameters are computed simultaneously [cf. Haslett and Raftery (1989), and Brockwell and Davis (1996)] by the Maximum or Whittle likelihood methods.

Our focus in this chapter is limited to stationary time series. However, we apply the filtering procedure to each segment of tv-ARFIMA(p,d,q) process separately to filter out the time-varying memory parameter and obtain tv-ARMA(p,q) process.

The underlying theory of the filtering technique in this chapter is related to the asymptotic properties of the discrete Fourier transform and the periodogram of a linear filter. It is a well known fact that the real and imaginary parts of the discrete Fourier transforms of an i.i.d series are independent and asymptotically Gaussian because the sin and cos trigonometric functions at discrete Fourier frequencies form orthonormal basis. In fact, the Fourier transforms at the discrete Fourier frequencies are independent [cf. Brockwell and Davis (1996), Proposition 10.3.2].

It is a general practice to develop asymptotics for the discrete Fourier transforms of a linear process in terms of the asymptotics of the discrete Fourier transforms of an i.i.d time series. The main application of such practice is to develop asymptotic theory for the sample periodogram.

In our case, we like to represent the discrete Fourier transform of an observed ARFIMA(p,d,q) process with the discrete Fourier transform of an unobserved ARMA(p,q) process and use this representation to approximate ARMA(p,q) process. In this chapter, we study in detail the approximation errors of such representation in the *mean square* sense and we provide the specific convergence rates of these errors.

The order of magnitude of error rates are useful in the segmenting part of nonstationary EEG time series model fitting. For example, if these rates are slower then it will be efficient to apply the filter technique to only longer segments.

We organize this chapter into four sections. In Section 4.1, we start with the discussion of discrete Fourier transform with and without tapering. In Section 4.2, we define briefly the long memory, $d \in [-.5, .5)$ of a time series and some useful properties of long memory and we develop the filtering technique to separate ARMA process from a stationary ARFIMA process. In Section 4,3, we extend the filtering technique to special type of nonstationary processes, where the memory parameter, d is grater than .5. In the final

Section 4 we evaluate the order of magnitude of the mean square error of the filtering technique by simulations. We also show in this section, the tapering of data plays more important role than originally thought.

4.1 Fourier transform of a linear process

Let ε_t is an i.i.d. series with mean zero and variance σ^2. Let X_t be a linearly nondeterministic so that it has the following representation

$$X_t = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j}.$$

In the case of stationary ARMA processes, $\sum_{j=0}^{\infty} |\phi_j| < \infty$. (cf. Brillinger, 1981)

The effect of a linear filter on a Fourier transform is (asymptotically) to multiply it by by the transfer function of the filter, provided that this is continuous at the frequency in question (cf. Walker, 2000 and references with in). This result has been known for a long time for the the discrete Fourier transform of a stationary noncausal process [cf. Hannan (1970), pp. 246-248, Brockwell and Davis (1996), Theorem 10.3.1)]. We prove similar result for the Fourier transform of a stationary causal process in the following lemma.

Lemma 4.1 *Let X_t be a stationary process with the canonical moving average representation $X_t = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j}$, with $\sum_{j=0}^{\infty} |\phi_j| < \infty$, where ε_t is an i.i.d series with mean zero and variance σ^2. Let $J_X(\lambda)$ denote the ordinary discrete Fourier transform of $\{X_1, ..., X_n\}$ and similarly we can define $J_\varepsilon(\lambda)$. Let the transfer function $\Phi(\omega) = \sum_{j=0}^{\infty} \phi_j \exp(-i\omega j)$. Then, when $\Phi(\omega)$ is continuous at $\omega = \lambda$, where $\lambda \in [0, \pi]$,*

$$J_X(\lambda) = \Phi(\lambda) J_\varepsilon(\lambda) + R_{X,n}(\lambda).$$

If in addition, $\sum_{j=0}^{\infty} |\phi_j| \; j^{1/2} < \infty$ then $\max_{\lambda_k = 2\pi k/n \in [0,\pi]} E|R_{X,n}(\lambda_k)|^2 \to 0$ as $n \to \infty$.

Proof.

$$\begin{aligned}
J_X(\lambda) &= \sum_{t=0}^{n-1} X_t \exp(-i\lambda t) \\
&= \sum_{t=0}^{n-1} \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j} \exp(-i\lambda t) \\
&= \sum_{j=0}^{\infty} \phi_j \exp(-i\lambda j) \sum_{t=0}^{n-1} \varepsilon_{t-j} \exp(-i\lambda(t-j)) \\
&= \sum_{j=0}^{\infty} \phi_j \exp(-i\lambda j) \sum_{t=j}^{n-1-j} \varepsilon_t \exp(-i\lambda t) \\
&= \sum_{j=0}^{\infty} \phi_j \exp(-i\lambda j) \left(\sum_{t=0}^{n-1} \varepsilon_t \exp(-i\lambda t) + \sum_{t=j}^{n-1-j} \varepsilon_t \exp(-i\lambda t) - \sum_{t=0}^{n-1} \varepsilon_t \exp(-i\lambda t) \right) \\
&= \Phi(\lambda) J_\varepsilon(\lambda) + R_{X,n}(\lambda)
\end{aligned}$$

where

$$R_{X,n}(\lambda) = \sum_{j=0}^{\infty} \phi_j \exp(-i\lambda j) \left(\sum_{t=j}^{n-1-j} \varepsilon_t \exp(-i\lambda t) - \sum_{t=0}^{n-1} \varepsilon_t \exp(-i\lambda t) \right). \tag{4.1}$$

Let $U_{nj} = \left(\sum_{t=j}^{n-1-j} \varepsilon_t \exp(-i\lambda t) - \sum_{t=0}^{n-1} \varepsilon_t \exp(-i\lambda t) \right)$. Note that, if $j < n$ then U_{nj} has $2j$ independent random variables, whereas if $j \geq n$ then U_{nj} has n independent random variables. It follows from that

$$E|U_{nj}|^2 \leq 2 \min(j, n/2)\sigma^2.$$

Hence

$$E|R_{X,n}|^2 \leq \left(\sum_{j=0}^{\infty} |\phi_j| \left(E|U_{nj}|^2 \right)^{1/2} \right)^2.$$

Thus

$$E|R_{X,n}|^2 \leq 2\sigma^2 \left(\sum_{j=0}^{\infty} |\phi_j| \min(j, n/2)^{1/2} \right)^2 . \tag{4.2}$$

Now, if m is a fixed positive integer we have for $n > m$

$$\sum_{j=0}^{\infty} |\phi_j| \min(j, n/2)^{1/2} \leq \sum_{j=0}^{m} |\phi_j| j^{1/2} + (n/2)^{1/2} \sum_{j>m}^{\infty} |\phi_j|,$$

hence

$$\lim_{n\to\infty} n^{-1/2} \sum_{j=0}^{\infty} |\phi_j| \min(j, n/2)^{1/2} \leq \sum_{j>m}^{\infty} |\phi_j|.$$

Since m is arbitrary it follows that the bound on $E|R_{X,n}|^2$ converges to zero as $n \to \infty$. $\diamond$

This lemma does not hold for stationary long memory processes, $d \in [-.5, .5)$, because the assumption, $\sum_{j=0}^{\infty} |\phi_j| < \infty$ is not satisfied. The Lemma 4.3 in this section makes a simpler assumption, $\sum_{j=0}^{\infty} \phi_j^2 < \infty$ and is extremely useful in the model fitting of sleep EEG. However, the result of Lemma 4.3 is weaker than the Lemma 4.1 as in the general case one can not assert that $E|R_{X,n}(\lambda)|^2 \to 0$. The Lemma 4.3 is valid for the discrete Fourier transforms of linear processes with and without taper.

Let X_t be a stationary process with a spectral density function $f(\lambda)$ then the tapered Fourier transform is defined as

$$J_X^T(\lambda) = \sum_{t=0}^{n-1} h(t/n)\, X_t \exp(-i\lambda t). \tag{4.3}$$

Here $h(.)$ is a bounded variation tapering function on $[0, 1]$. In the following Lemma 4.2, we look at the bounded variation of $h^2(.)$

Lemma 4.2 *If $h(u)$ is a bounded variation function over $[0, 1]$ with finite total variation V, then $h^2(.)$ is also a bounded variation function.*

Proof. The definition of bounded variation says, for the supremum of all partitions over $[0,1]$, we have $\sup\sum_{t=1}^{n-1}|\Delta h(u_t)| = V$, where $\Delta h(u_t) = h(u_t) - h(u_{t-1})$. Since $h(u)$ is a bounded variation function over a closed interval, $|h(u)| \leq M$ where M is finite. Let $\Delta h^2(u_t) = h^2(u_t) - h^2(u_{t-1})$ then $\Delta h^2(u_t) = (h(u_t) - h(u_{t-1}))(h(u_t) + h(u_{t-1})) = \Delta h(u_t)(h(u_t) + h(u_{t-1}))$.

$$\begin{aligned}
\sup\sum_{t=1}^{n-1}|\Delta h^2(u_t)| &= \sum_{t=1}^{n-1}\Delta h(u_t)\,|(h(u_t)+h(u_{t-1}))| \\
&\leq 2M\sum_{t=1}^{n-1}|\Delta h(u_t)| \\
&\leq 2MV \\
\sup\sum_{t=1}^{n-1}|\Delta h^2(u_t)| &< \infty.
\end{aligned}$$

Hence, $h^2(u)$ is a bounded variation function. ◇

The following Lemma 4.3, for the not necessarily absolutely-summable linear filter coefficients, is crucial in our analysis and is sketched in Walker (2000). Here, we give a comprehensive proof because it will be extensively used in Theorems 4.1 and 4.2.

Lemma 4.3 *Let ε_t i.i.d series with mean zero, variance one, and the other moments finite and let $X_t = \sum_{j=0}^{\infty} c_j\,\varepsilon_{t-j}$, where $\sum_{j=0}^{\infty} c_j^2 < \infty$ so that X_t at least exists as a mean square limit. Let $J_X^T(\lambda) = \sum_{t=0}^{n-1} h(t/n)\,X_t \exp(-i\lambda t)$, with $h(.)$ being a function of bounded variation over $[0,1]$, and let $J_\varepsilon^T(\lambda)$ be the tapered Fourier transform of i.i.d series ε_t. Let $C(\omega) = \sum_{j=0}^{\infty} c_j \exp(-i\lambda j)$. Then, when $C(\omega)$ is continuous at $\omega = \lambda$,*

$$J_X^T(\lambda) = C(\lambda)\,J_\varepsilon^T(\lambda) + R_n(\lambda) \tag{4.4}$$

where $E\{|R_n(\lambda)|^2\} = o(n), \;\; n \to \infty$.

Proof. We can write,

$$\begin{aligned}
E\{|R_n(\lambda)|^2\} &= E\{|J_X^T(\lambda)|^2\} + |C(\lambda)|^2 E\{|\,J_\varepsilon^T(\lambda)|^2\} \\
&\quad -2Re[C(\lambda)\,E\{J_e^T(\lambda)\overline{J_X^T(\lambda)}\}].
\end{aligned} \tag{4.5}$$

Now,

$$E\{|J_X^T(\lambda)|^2\} = \sum_{t,s=0}^{n-1} h(t/n)\, h(s/n) E(X_t X_s) \exp\{-i\lambda(t-s)\}.$$

Also, (X_t) is easily seen to be a stationary process with zero mean and spectral density $f_X(\lambda) = |C(\lambda)|^2/2\pi$.

Hence

$$E\{|J_X^T(\lambda)|^2\} = \int_{-\pi}^{\pi} |g_n(\lambda-\omega)|^2 f_X(\omega)\, d\omega, \tag{4.6}$$

where $g_n(\omega) = \sum_{t=0}^{n-1} h(t/n)\, \exp(-i\omega t)$.

Now, we write $g_n(\omega) \leq \sum_{t=0}^{n-1} h(t/n)\, (S_{t,\omega} - S_{t-1,\omega})$, where $S_{t,\omega} = \sum_{r=0}^{t} \exp(-i\omega r), 0 \leq t \leq n-1, S_{-1,\omega} = 0$.

Then

$$\begin{aligned} |g_n(\omega)| &\leq |\sum_{t=0}^{n-2} \{h(t/n) - h(t-1/n) S_{t,\omega} + h(n-1/n) S_{n-1,\omega})\}| \\ &\leq K M_{n,\omega}, \end{aligned}$$

where $M_{n,\omega} = \max_{0 \leq t \leq n-1} |S_{t,\omega}| \leq \frac{1}{\sin \frac{\omega}{2}}$ and K is a constant.

The result then follows by writing

$$E\{|J_X^T(\lambda)|^2\} = f_X(\lambda) \int_{-\pi}^{\pi} |g_n(\omega)|^2\, d\omega + \int_{-\pi}^{\pi} \{f_X(\lambda-\omega) - f_X(\lambda)\}\, |g_n(\omega)|^2 d\omega.$$

The first integral results in

$$f_X(\lambda) \int_{-\pi}^{\pi} |g_n(\omega)|^2\, d\omega = 2\pi n H_2 f_X(\lambda) + o(n) \tag{4.7}$$

because

$$\begin{aligned}
\int_{-\pi}^{\pi} |g_n(\omega)|^2 \, d\omega &= \int_{-\pi}^{\pi} \sum_{t,u=0}^{n-1} h(t/n)h(u/n) \exp(-i\lambda(t-u)) \, d\omega \\
&= \int_{-\pi}^{\pi} \sum_{t=0}^{n-1} h^2(t/n) d\omega = 2\pi \sum_{t=0}^{n-1} h^2(t/n) \\
&= 2\pi \sum_{t=0}^{n-1} h^2(t/n) - 2\pi n \int_0^1 h^2(t)dt + 2\pi n \int_0^1 h^2(t)dt \\
&= 2\pi n H_2 + o(n),
\end{aligned}$$

where $H_2 = \int_0^1 h^2(t)dt$. The absolute difference $\left| \frac{1}{n} \sum_{t=0}^{n-1} h^2(t/n) - \int_0^1 h^2(t)dt \right|$ is asymptotically bounded by V/n [cf. Brillinger (1981), Lemma 5.1] where V is the finite total variation of $h^2(t)$ on $[0, 1]$. This result was proved for $h(t)$ function in the given reference but it can also be applied for $h^2(t)$ because if $h(t)$ is a bounded variation function, $h^2(t)$ is also bounded variation function by the previous Lemma 4.2.

The second integral is solved by expressing it as two integrals, one over $|\omega| \leq \delta$ and the other over $|\omega| > \delta$, where δ is arbitrarily small.

$$\begin{aligned}
\int_{-\pi}^{\pi} |f_X(\lambda-\omega) - f_X(\lambda)||g_n(\omega)|^2 d\omega &\leq \epsilon \int_{-\delta}^{\delta} |g_n(\omega)|^2 d\omega \\
&\quad + \int_{(-\pi,\pi)\setminus(-\delta,\delta)} |f_X(\lambda-\omega) - f_X(\lambda)| \frac{1}{\sin^2(\frac{\omega}{2})} d\omega \\
&\leq \epsilon \int_{-\pi}^{\pi} |g_n(\omega)|^2 d\omega + \frac{1}{\delta^2(\epsilon)} 2\sigma_X^2 < \infty \\
&= o(n).
\end{aligned}$$

Here, we have used the argument that when a function $f(\omega)$ is continuous at $\omega = \lambda$ then using the so called epsilon-delta definition there ex-

ists a small neighborhood around $\lambda > 0$ such that $|\lambda - \omega| < \delta$ for which $|f_X(\lambda - \omega) - f_X(\lambda)| < \epsilon$. In the above calculation, we have also used the following properties of the Fejer kernel: (i) $\frac{\sin^2(\frac{n\omega}{2})}{\sin^2(\frac{\omega}{2})}$is always non-negative, (ii) $\int_{-\pi}^{\pi} \frac{\sin^2(\frac{n\omega}{2})}{\sin^2(\frac{\omega}{2})} d\omega = 1$, and (iii) $\int_{-\delta}^{\delta} \frac{\sin^2(\frac{n\omega}{2})}{\sin^2(\frac{\omega}{2})} d\omega \to 1$ as $n \to \infty$.

Therefore

$$E\{|J_X^T(\lambda)|^2\} = 2\pi n H_2 f_X(\lambda) + o(n). \tag{4.8}$$

In the same way, noting that ε_t has the spectral density $1/2\pi$, we obtain

$$E\{|J_\varepsilon^T(\lambda)|^2\} = nH_2 + o(1). \tag{4.9}$$

Finally,

$$\begin{aligned} E\{J_\varepsilon^T(\lambda)\overline{J_X^T(\lambda)}\} &= E\{\sum_{t=0}^{n-1} h(t/n)\varepsilon_t \exp(-i\lambda t) \sum_{u=0}^{n-1} h(u/n)(\sum_{s=0}^{\infty} c_s \varepsilon_{t-s}) \exp(i\lambda u)\} \\ &= \sum_{t,u=0}^{n-1} h(t/n)h(u/n) \exp(-i\lambda(t-u))\, c_{t-u} \\ &= \sum_{t,u=0}^{n-1} h(t/n)h(u/n) \exp(-i\lambda(t-u)) \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\omega(t-u)) C(\omega) d\omega, \end{aligned}$$

where $c(s) = 0, s < 0$.

The above expression is similar to the expression of $E\{|J_X^T(\lambda)|^2\}$ except that instead of $|C(\omega)|^2/2\pi$, we have $C(\omega)$. Therefore, we use the similar argument as before to get the following result

$$E\{J_e^T(\lambda)\overline{J_X^T(\lambda)}\} = nH_2\overline{C(\lambda)} + o(1) \tag{4.10}$$

By substituting (4.8), (4.9), and (4.10) in (4.5), we obtain the result of this lemma. ◇

Remarks: The result of Lemma 4.3 can be equally applied to the discrete Fourier transform of a stationary processes with no taper. In this case,

$h(t/n) = 1$ for any $t = 0, 1, \cdots, n-1$. Also in this lemma, we require (ε_t) to be only zero mean uncorrelated process with a spectral density.

In the following section, we study some important properties of long memory stationary processes and introduce the filtering technique in Theorem 4.1 to separate ARMA(p,q) process from a ARFIMA(p,d,q) process.

4.2 Long memory processes

Definition: A long memory process is a stationary process for which the spectral density function $f(\lambda)$ satisfies the condition

$$\lim_{\lambda \to 0} \frac{f(\lambda)}{c_f \mid \lambda \mid^{-2d}} = 1 \tag{4.11}$$

for $d \in [-.5, .5)$, where c_f is a positive constant(Beran, 1994).

Let B be the backward difference operator. Then, $BX_t = X_{t-1}$, $B^2 X_t = X_{t-2}, \cdots, B^j X_t = X_{t-j}$. For any real number $d > -1$, Hosking (1981) defines the difference operator $\nabla^d = (1-B)^d$ by means of the binomial expansion,

$$\nabla^d = (1-B)^d = \sum_{j=0}^{\infty} \pi_j B^j, \tag{4.12}$$

where

$$\begin{aligned} \pi_j &= \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \\ &= \prod_{0<k\leq j} \frac{k-1-d}{k}, \quad j = 0, 1, \cdots, \end{aligned}$$

and $\Gamma(.)$ is the gamma function,

$$\Gamma(z) = \begin{cases} \int_0^\infty t^{z-1} e^{-t} dt & z > 0, \\ \infty & z = 0, \\ z^{-1}\Gamma(1+z) & z < 0. \end{cases}$$

Similarly, we can also define the integral operator ∇^{-d}, the counterpart of ∇^d as

$$\nabla^{-d} = (1-B)^{-d} = \sum_{j=0}^{\infty} \psi_j B^j,$$

where

$$\psi_k = \frac{\Gamma(k+d)}{\Gamma(k+1)\Gamma(d)} = \prod_{0<l\leq k} \frac{l-1-d}{l},\ k = 0, 1, 2, \cdots .$$

Remark 1: By applying the Stirling's formula $\Gamma(x) \sim \sqrt{2\pi}\ e^{-x+1}(x-1)^{x-1/2}$ as $x \to \infty$ to ψ_k, we obtain

$$\psi_k \sim k^{d-1}/\Gamma(d) \quad \text{as } k \to \infty.$$

The sequence ψ_k is square summable, $\sum_{k=0}^{\infty} \psi_k^2 < \infty$, for any $d < 1$.

Remark 2: The unit root transfer function is defined as the discrete Fourier transform of ψ_k as $n \to \infty$

$$\sum_{j=0}^{n} \psi_j \exp(-i\omega j) \to (1-\exp(-i\lambda))^{-d} \quad \text{as } n \to \infty. \tag{4.13}$$

Remark 3: The sequence ψ_k can be easily generated using the recursion relation $\psi_{k+1} = \psi_k \Gamma(k+d)/\Gamma(K+1)$ with $\Gamma(0) = 1$ for any $k \geq 0$.

Let X_t be an ARFIMA(p,d,q) process with $d \in [-.5, .5)$. If X_t is stationary then it satisfies the difference equations,

$$a(B)\nabla^d X_t = b(B)\varepsilon_t, \tag{4.14}$$

where ε_t is an i.i.d. series, $a(B)$ and $b(B)$ are polynomials of orders p and q respectively: $a(B) = 1 + a_1 B + \cdots + a_p B^p$ and $b(B) = 1 + b_1 B + \cdots + b_q B^q$; and ∇^d is the difference operator.

We can also rewrite the above X_t in terms of an ARMA(p,q) process, $Y_t = b(B)a(B)^{-1}\varepsilon_t$, as follows:

$$\nabla^d X_t = Y_t, \tag{4.15}$$

$$X_t = \nabla^{-d} Y_t, \tag{4.16}$$

where ∇^{-d} is the integral (summation) operator.

If a stationary process Y_t is a *causal* function of a sequence $\{\varepsilon_j\}$ then it can be represented by a canonical moving average representation $Y_t = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j}$ with $\sum_{j=0}^{\infty} |\phi_j| < \infty$ (Brockwell and Davis, 1996). But this definition of causality may not work for long memory processes, because $\sum_{j=0}^{\infty} |\psi_j| = \infty$ for this process. However, the sequence $\{\psi_j\}$ is square summable.

In view of the above discussion, we say that a long memory process X_t is causal (even though the coefficients $\{\psi_j\}$ are not be absolutely summable as in the traditional sense), if X_t can be expressed, using $X_t = \nabla^{-d} Y_t$, as

$$\begin{aligned} X_t &= \sum_{j=0}^{\infty} \psi_j Y_{t-j} \\ &= \sum_{j=0}^{\infty} \psi_j \sum_{k=0}^{\infty} \phi_k \varepsilon_{t-j-k} \\ &= \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \end{aligned}$$

where $c_j = \sum_{k=0}^{j} \psi_k \phi_{j-k}$. Since the sequences $\{\psi_j\}$ and $\{\phi_j\}$ are square summable, then the sequence $\{c_j\}$ is also square summable.

Since $\{c_j\}$ is a convolution of $\{\psi_j\}$ and $\{\phi_j\}$, the Fourier transform of $\{c_j\}$, $C(\lambda)$ is the product of the Fourier transforms of $\{\psi_j\}$ and $\{\phi_j\}$, $\psi(\lambda)$ and $\Phi(\lambda)$ respectively.

$$\begin{aligned} C(\lambda) &= \sum_{j=0}^{\infty} c_j \exp(-i\lambda j) \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^{j} \psi_k \phi_{j-k} \exp(-i\lambda j) \end{aligned}$$

$$
\begin{aligned}
C(\lambda) &= \sum_{j=0}^{\infty}\sum_{k=0}^{j}\psi_k \exp(-i\lambda k)\phi_{j-k}\exp(-i\lambda(j-k)) \\
&= \sum_{j=0}^{\infty}\psi_j \exp(-i\lambda j)\sum_{k=0}^{\infty}\phi_k \exp(-i\lambda k) \\
&= \psi(\lambda)\Phi(\lambda) \qquad (4.17)
\end{aligned}
$$

where $\psi(\lambda) = (1-\exp(-i\lambda))^{-d}$ is the unit root transfer function.

Remark 4: If a long memory process X_t is causal, then $X_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}$ with $\sum_{j=0}^{\infty} c_j^2 < \infty$.

Now, we study the behavior of unit root transfer function as $\lambda \to 0$ by analyzing the following result.

The spectral density function of an ARFIMA(p=0,d,q=0) process (*the fractional integrated noise*) is

$$\psi(\lambda) = |2\sin(\frac{\lambda}{2})|^{-d}. \qquad (4.18)$$

By the definition of stationary processes, the spectral density function of the fractional integrated noise $f(\lambda) = \frac{\sigma^2}{2\pi} \mid \psi(\lambda) \mid^2$, where σ^2 is the variance

of the i.i.d. ε_t. Then,

$$\begin{aligned}
|\psi(\lambda)|^2 &= |(1-\exp(-i\lambda))|^{-2d} \\
&= |1-\cos(\lambda)+i\sin(\lambda)|^{-2d} \\
&= \left|\sin^2(\frac{\lambda}{2})+\cos^2(\frac{\lambda}{2})-\cos^2(\frac{\lambda}{2})+\sin^2(\frac{\lambda}{2})+i2\sin(\frac{\lambda}{2})\cos(\frac{\lambda}{2})\right|^{-2d} \\
&= \left|2\sin(\frac{\lambda}{2})(\sin(\frac{\lambda}{2})+i\cos(\frac{\lambda}{2}))\right|^{-2d} \\
&= \left|4\sin^2(\frac{\lambda}{2})(\sin(\frac{\lambda}{2})+i\cos(\frac{\lambda}{2}))(\sin(\frac{\lambda}{2})-i\cos(\frac{\lambda}{2}))\right|^{-d} \\
&= \left|4\sin^2(\frac{\lambda}{2})(\sin^2(\frac{\lambda}{2})+\cos^2(\frac{\lambda}{2}))\right|^{-d} \\
&= \left|4\sin^2(\frac{\lambda}{2})\right|^{-d}
\end{aligned}$$

Hence,

$$\psi(\lambda) = \mid 2\sin(\frac{\lambda}{2}) \mid^{-d}. \tag{4.19}$$

The behavior of $\psi(\lambda)$ at higher frequencies as $\lambda \to \pm(\pi)$ is bounded. However, the behavior of this function at lower frequencies is of our interest and it can be shown that

$$\psi(\lambda) \propto \mid \lambda \mid^{-d}, \quad \text{as } \lambda \to 0. \tag{4.20}$$

Hence, the order of magnitude of unit root transfer function

$$\psi(\lambda) = O(\mid \lambda \mid^{-d}) \quad \text{as } \lambda \to 0. \tag{4.21}$$

In the case of the discrete Fourier transform and for $j = 1, 2, \cdots, (n-1)$:

$$\begin{aligned}
\psi(\lambda) &= O\left(\left|\frac{2\pi j}{n}\right|^{-d}\right) \quad \text{as } n \to \infty \\
&= O(\mid n \mid^{d}) \quad \text{as } n \to \infty
\end{aligned} \tag{4.22}$$

Note: The discrete Fourier frequency, $\lambda_{j=0} = 0$, is a special case. However, this case has no theoretical or practical impact in zero mean processes.

4.2.1 Fourier transform of a stationary long memory process

Let Y_t be linearly nondeterministic process, so that it has a canonical moving average representation $Y_t = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j}$, where ε_t is an i.i.d series with mean zero and variance σ^2, and $\sum_{j=0}^{\infty} |\phi_j| < \infty$. Then, the spectral density of Y_t can be defined as $f_Y(\lambda) = |\Phi(\lambda)|^2 \sigma^2 / 2\pi$ where $\Phi(\lambda) = \sum_{j=0}^{\infty} \phi_j \exp^{-i\lambda j}$.

Let X_t be another linearly nondeterministic process for which $X_t = \nabla^{-d} Y_t$ holds for $d \in [-1/2, 1/2)$, where $\nabla^{-d} = (1 - B)^{-d} = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. Then, we can write X_t as a linear process of Y_t, $X_t = \sum_{j=0}^{\infty} c_j Y_{t-j}$ with $\sum_{j=0}^{\infty} c_j^2 < \infty$, where c_j is the convolution of ψ_j and ϕ_j as described in (4.17). Then, the spectral density of X_t is defined as $f_X(\lambda) = |C(\lambda)|^2 \sigma^2 / 2\pi$ where $C\lambda) = \sum_{j=0}^{\infty} c_j \exp^{-i\lambda j}$.

Using Lemma 4.3, we can derive the following results for the tapered discrete Fourier transforms of Y_t and X_t

$$J_Y^T(\lambda) = \Phi(\lambda) J_\varepsilon^T(\lambda) + R_{Y,n}(\lambda) \tag{4.23}$$

where $E\{\mid R_{Y,n}(\lambda) \mid^2\} = o(n)$, as $n \to \infty$ and

$$J_X^T(\lambda) = C(\lambda) J_\varepsilon^T(\lambda) + R_{X,n}(\lambda) \tag{4.24}$$

where $E\{|R_{X,n}(\lambda)|^2\} = o(n)$, as $n \to \infty$.

The following Theorem 4.1 provides the reasoning for the filtering technique in Fourier domain to separate the ARMA(p,q) process, Y_t from the ARFIMA(p,d,q) process, X_t.

Theorem 4.1 *Let ε_t be an i.i.d. series with mean zero and variance one. Let X_t and Y_t are the processes as defined above and they are related by*

$X_t = \nabla^{-d} Y_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}$ *for* $d \in [-.5, .5)$ *with* $\sum_{j=0}^{\infty} c_j^2 < \infty$. *Let* $J_X^T(\lambda) = \sum_{t=0}^{n-1} h(t/n) X_t \exp(-i\lambda t)$, *and* J_ε^T *and* J_Y^T *are defined similarly for* ε_t *and* Y_t. *Let* $\Phi(\omega) = \sum_{j=0}^{\infty} \phi_j \exp(-i\omega j)$, $\psi(\omega) = \sum_{j=0}^{\infty} \psi_j \exp(-i\omega j) = (1 - \exp(-i\omega))^{-d}$, *and* $C(\omega) = \sum_{j=0}^{\infty} c_j \exp(-i\omega j)$. *Then, when* $\Phi(\omega)$, $\psi(\omega)$, $C(\omega)$ *are all continuous at* $\omega = \lambda$,

$$J_X^T(\lambda) = \psi(\lambda) J_Y^T(\lambda) + R_{X,Y,n}(\lambda) \tag{4.25}$$

where $E\{\mid R_{X,Y,n}(\lambda) \mid^2\} = o(n^{1+2d})$, *as* $n \to \infty$.

Proof. Using the result (4.17), $C(\lambda) = \psi(\lambda)\Phi(\lambda)$ where $\psi(\lambda) = (1 - \exp(-i\lambda))^{-d}$ is the unit root transfer function, we can write

$$\begin{aligned} J_X^T(\lambda) &= \psi(\lambda)\Phi(\lambda) J_\varepsilon^T(\lambda) + R_{X,n}(\lambda) \\ &= \psi(\lambda)\ (\Phi(\lambda) J_\varepsilon(\lambda) + R_{Y,n}(\lambda) - R_{Y,n}(\lambda)) + R_{X,n}(\lambda) \\ &= \psi(\lambda) J_Y^T + R_{X,n}(\lambda) - \psi(\lambda) R_{Y,n}(\lambda) \\ &= \psi(\lambda) J_Y^T + R_{X,Y,n}(\lambda) \end{aligned} \tag{4.26}$$

where $R_{Y,n}(\lambda) = J_Y^T(\lambda) - \Phi(\lambda) J_\varepsilon(\lambda)$, $R_{X,n}(\lambda) = J_X^T(\lambda) - C(\lambda) J_\varepsilon^T(\lambda)$ and $R_{X,Y,n}(\lambda) = R_{X,n}(\lambda) - \Phi(\lambda) R_{Y,n}(\lambda)$.

From Lemma 4.3, we also have, $E\{\mid R_{Y,n}(\lambda) \mid^2\} = o(n)$ and $E\{\mid R_{X,n}(\lambda) \mid^2 \} = o(n)$ as $n \to \infty$.

$$\begin{aligned} E\{\mid R_{X,Y,n}(\lambda) \mid^2\} &= E\{\mid R_{X,n}(\lambda) \mid^2\} + \mid \psi(\lambda) \mid^2 E\{\mid R_{Y,n}(\lambda) \mid^2\} - \\ &\quad 2\,\Re[\psi(\lambda)\ E\{R_{Y,n}(\lambda)\ \overline{R_{X,n}(\lambda)}\}] \end{aligned} \tag{4.27}$$

By Cauchy-Schwarz inequality and Lemma 4.3

$$\begin{aligned} E\{R_{Y,n}(\lambda)\ \overline{R_{X,n}(\lambda)}\} &\leq \left(E|R_{Y,n}(\lambda)|^2\right)^{1/2} \left(E|\overline{R_{X,n}(\lambda)}|^2\right)^{1/2} \\ &= o(n) \quad n \to \infty. \end{aligned} \tag{4.28}$$

Using(4.22),

$$\mid \psi(\lambda) \mid^2\ E\{\mid R_{Y,n}(\lambda) \mid^2\} = O(n^{2d})\ o(n) = o(n^{1+2d}) \quad \text{as } n \to \infty \tag{4.29}$$

and

$$\psi(\lambda)\ E\{R_{n,}(\lambda)\ \overline{R_{X,n}(\lambda)}\} = O(n^d)\ o(n) = o(n^{1+d}) \quad \text{as } n \to \infty. \tag{4.30}$$

By substituting (4.28), (4.29), and (4.30) in (4.27), we obtain,

$$E\{|\ R_{X,Y,n}(\lambda)\ |^2\} = o(n^{\max(1,1+2d)}) \quad \text{as } n \to \infty, \tag{4.31}$$

the final result. ◇

Especially for $-.5 < d \leq 0$ the order of magnitude of error is $o(n)$ and for $0 < d < .5$ the order of magnitude is $o(n^{1+2d})$. This result may not be valid at $\lambda = 0$. However, this case has no effect on zero mean X_t.

We conjecture that this order of magnitude $o(n^{1+2d})$ is *efficient*, considering that the Central Limit Theorem of long memory processes has the same rate (cf. Beran, 1994).

The immediate application of this theorem is in the filtering of the unobserved ARMA(p,q) process from the observed ARFIMA(p,d,q) process. If $\hat{d}$ is a consistent estimate of d, which is estimated for example by the semiparametric Whittle method in Section 2.2.3, then one can derive the following relation for the Fourier transform of unobserved Y_t from the Fourier transform of observed X_t

$$J_Y^T(\lambda) \approx (1 - \exp(-i\lambda))^{\hat{d}} J_X^T(\lambda) - (1 - \exp(-i\lambda))^{\hat{d}} R_{X,Y,n}(\lambda).$$

However,

$$E\{|\ (1 - \exp(-i\lambda))^{\hat{d}} R_{X,Y,n}(\lambda)\ |^2\} = o(n) \quad \text{as } n \to \infty. \tag{4.32}$$

Now, by applying the inverse Fourier transform to either side of the above equation provides an approximation to the unobserved Y_t from the observed X_t

$$\widetilde{Y}_t = \frac{1}{2\pi n} \sum_j \exp(i\lambda_j t)(1 - \exp(-i\lambda_j))^{\hat{d}} J_X(\lambda_j), \quad t = 0, 1, \cdots, n-1 \tag{4.33}$$

Remark 1: The error term in the filtering, $(1-\exp(-i\lambda))^d R_{X,Y,n}(\lambda)$ converges to zero in mean square sense.

Remark 2: Another application of Theorem 4.1 is in the development of asymptotic theory of the periodograms of a stationary ARFIMA(p,d,q) process in terms of the asymptotic theory of the periodograms of a stationary ARMA(p,q) process.

4.2.2 Fourier transform of a nonstationary long memory process

In this section, we extend Theorem 4.1 to a special type of nonstationary processes, $\nabla^d X_t = Y_t$, where Y_t is a stationary ARMA process and $d \in [.5, r+.5)$, r being an positive integer ($r \geq 1$) . When r=1, which is the case of interest, the ARFIMA process has nonstationary memory parameter, $d \in [.5, 1.5)$. When $r = 0$, we have the stationary ARFIMA process with $d \in [-.5, .5)$.

Theorem 4.2 *Let X_t be a special type of nonstationary process as defined above and Y_t be a stationary process. Let X_t and Y_t are related by $\nabla^d X_t = Y_t$, where $d \in [.5, r+.5)$ with r being any positive integer. Let $J_X^T(\lambda) = \sum_{t=0}^{n-1} h(t/n) X_t \exp(-i\lambda t)$, where $h(.)$ is the function of bounded variation. Let J_Y^T is defined similarly for Y_t. If $E(X_t^2) < \infty$,*

$$J_X^T(\lambda) = (1-\exp(-i\lambda))^{-d} J_Y^T(\lambda) + R_{X,Y,n}(\lambda) \tag{4.34}$$

where $E\{\mid R_{X,Y,n}(\lambda) \mid^2\} = o(n^{max(2r+1,\, 1+2d)})$, as $n \to \infty$.

Proof. This theorem is proved for $r = 1$, which is of our interest. The general case can be easily derived from $r = 1$ case. When $r = 1$, $\nabla X_t =$

$h(t/n)X_t - h(t-1/n)X_{t-1}$ with $\nabla X_0 = h(0/n)X_0$, then

$$\begin{aligned}
J^T_{\nabla X}(\lambda) &= \sum_{t=0}^{t=n-1} \nabla X_t \exp(-i\lambda t) \\
&= \sum_{t=0}^{t=n-1} (h(t/n)X_t - h(t-1/n)X_{t-1}) \exp(-i\lambda t) \\
&= \sum_{t=0}^{t=n-1} h(t/n)X_t - \exp(-i\lambda) \sum_{t=0}^{t=n-2} h(t/n)X_t \exp(-i\lambda t) \\
&\quad - \exp(-i\lambda)h(n-1/n)X_{n-1} \exp(-i\lambda(n-1)) \\
&\quad + \exp(-i\lambda)h(n-1/n)X_{n-1} \exp(-i\lambda(n-1)) \\
&= J^T_X(\lambda) - \exp(-i\lambda)J^T_X(\lambda) + \exp(-i\lambda n)h(n-1/n)X_{n-1} \\
&= (1-\exp(-i\lambda))J^T_X(\lambda) + h(n-1/n)X_{n-1}\exp(-i\lambda n).
\end{aligned}$$

In the case of discrete Fourier transforms, where $\lambda = 2\pi j/n$ with $j = 0, 1, \ldots, n-1$, $\exp(-i\lambda n) = 1$. Then,

$$J^T_X(\lambda) = (1-\exp(-i\lambda))^{-1} J^T_{\nabla X}(\lambda) - (1-\exp(-i\lambda))^{-1} h(n-1/n)X_{n-1}. \quad (4.35)$$

Using Theorem 4.1, we can write

$$J^T_{\nabla X}(\lambda) = (1-\exp(-i\lambda))^{-(d-1)} J^T_Y(\lambda) + R_{\nabla X,Y,n}(\lambda), \quad (4.36)$$

where $E\{\mid R_{\nabla X,Y,n}(\lambda) \mid^2\} = o(n^{\max(1,1+2(d-1))})$ as $n \to \infty$.

Then,

$$J^T_X(\lambda) = (1-\exp(-i\lambda))^{-d} J^T_Y(\lambda) + R_{X,Y,n}(\lambda), \quad (4.37)$$

where $R_{X,Y,n}(\lambda) = (1-\exp(-i\lambda))^{-1}(R_{\nabla X,Y,n} - h(n-1/n)X_{n-1})$.

The function $(1-\exp(-i\lambda))^{-1}$ is bounded at higher frequencies as $\lambda \to \pm\pi$ but at lower frequencies as $\lambda \to 0$ it behaves like $O(n)$.

Then we have,

$$\begin{aligned}
E\{\mid R_{X,Y,n}(\lambda) \mid^2\} &= \mid (1-\exp(-i\lambda)) \mid^{-2} (E\{\mid R_{\nabla X,Y,n} \mid^2\} + h^2(n-1/n)EX^2_{n-1} \\
&\quad -2h(n-1/n)\Re[E\{X_{n-1}R_{\nabla X,Y,n}\}]). \quad (4.38)
\end{aligned}$$

By Cauchy-Schwarz inequality and Lemma 4.3

$$\begin{aligned} E\{\mid X_{n-1}R_{\nabla X,Y,n}(\lambda)\mid\} &\leq \left(EX_{n-1}^2\right)^{1/2}\left(E\mid (R_{\nabla X,Y,n})^2\mid\right)^{1/2} \\ &= o(n^{\max(1/2,1+(d/2))}) \quad n\to\infty \end{aligned}$$

Hence

$$\begin{aligned} E\{\mid R_{X,Y,n}(\lambda)\mid^2\} &= O(n^2)o(n^{\max(1,1+2(d-1))}) \\ &= o(n^{\max(3,1+2d)}) \end{aligned} \tag{4.39}$$

The order of magnitude is n^3 when $d\in[.5,1)$

This result can be easily extended to for any $r>1$ with

$$\begin{aligned} E\{\mid R_{X,Y,n}(\lambda)\mid^2\} &= O(n^{2r})o(n^{\max(1,1+2(d-r))}) \\ &= o(n^{\max(2r+1,1+2d)}). \end{aligned} \tag{4.40}$$

This proves the theorem. ◇

The condition in Theorem 4.2, $E(X_t^2)<\infty$, may be relaxed in the taper case by selecting a taper such that $h^2(n-1/n)E(X_{n-1}^2)<\infty$. For example, if one select a cosine bell taper then $h(n-1/n)=0$. In this case, $E\{\mid R_{X,Y,n}(\lambda)\mid^2\}=\mid(1-\exp(-i\lambda))\mid^{-2}\ E\{\mid R_{\nabla X,Y,n}\mid^2\}$ and the proof of Theorem 4.2 gets simpler.

The advantage of tapering on the residual term $R_{X,Y,n}(\lambda)$ can not be verified in this proof. But a better explicit algebraic derivation of this theorem for a given tapering function can improve the order of magnitude of $E\{\mid R_{X,Y,n}(\lambda)\mid^2\}$. We conjecture that the order of magnitude of error in Theorem 4.2 is the worst case scenario, without tapering. In the case $r=1$ and cosine bell taper, we prove via simulations that the order of magnitude of $E\{\mid R_{X,Y,n}(\lambda)\mid^2\}$ is smaller, and it is similar to the order of magnitude of $E\{\mid R_{\nabla X,Y,n}(\lambda)\mid^2\}$, which is $o(n^{1+2(d-1)})$.

Just as in the case of Theorem 4.1, the immediate application of this theorem is in the filtering of the unobserved stationary ARMA(p,q) process

from the observed nonstationary ARFIMA(p,d,q) process. If $\hat{d}$ is a consistent estimate of d, which is estimated for example by the semiparametric Whittle method in Section 2.2.3, then one can derive the following relation for the Fourier transform of unobserved Y_t from the Fourier transform of observed X_t

$$J_Y^T(\lambda) \approx (1-\exp(-i\lambda))^{\hat{d}} J_X^T(\lambda) - (1-\exp(-i\lambda))^{\hat{d}} R_{X,Y,n}(\lambda).$$

However,

$$E\{\mid (1-\exp(-i\lambda))^{\hat{d}} R_{X,Y,n}(\lambda) \mid^2\} = o(n) \quad \text{as } n \to \infty. \tag{4.41}$$

Now, by applying the inverse Fourier transform to either side of the above equation provides an approximation to the unobserved Y_t from the observed X_t

$$\widetilde{Y_t} = \frac{1}{2\pi n} \sum_j \exp(i\lambda_j t)(1-\exp(-i\lambda_j))^{\hat{d}} J_X(\lambda_j), \quad t = 0, 1, \cdots, n-1. \tag{4.42}$$

Remark 1: The error term in the filtering, $(1-\exp(-i\lambda))^d R_{X,Y,n}(\lambda)$ converges to zero in mean square sense.

Remark 2: Another application of Theorem 4.2 is in the development of asymptotic theory of the periodograms of a nonstationary ARFIMA(p,d,q) process.

4.2.3 Simulations

In this section, we evaluate the efficiency of the filtering procedure via simulations. Instead of evaluating the rate of convergence of the mean square errors in (4.32) and (4.41), we have evaluated

$$\frac{1}{n} \sum_{j=1}^{n} (Y_t - \tilde{Y}_t)^2 \tag{4.43}$$

for different n values. We considered both stationary and nonstationary ARFIMA(p,d,q) time series, and also with and without tapering cases in our simulations.

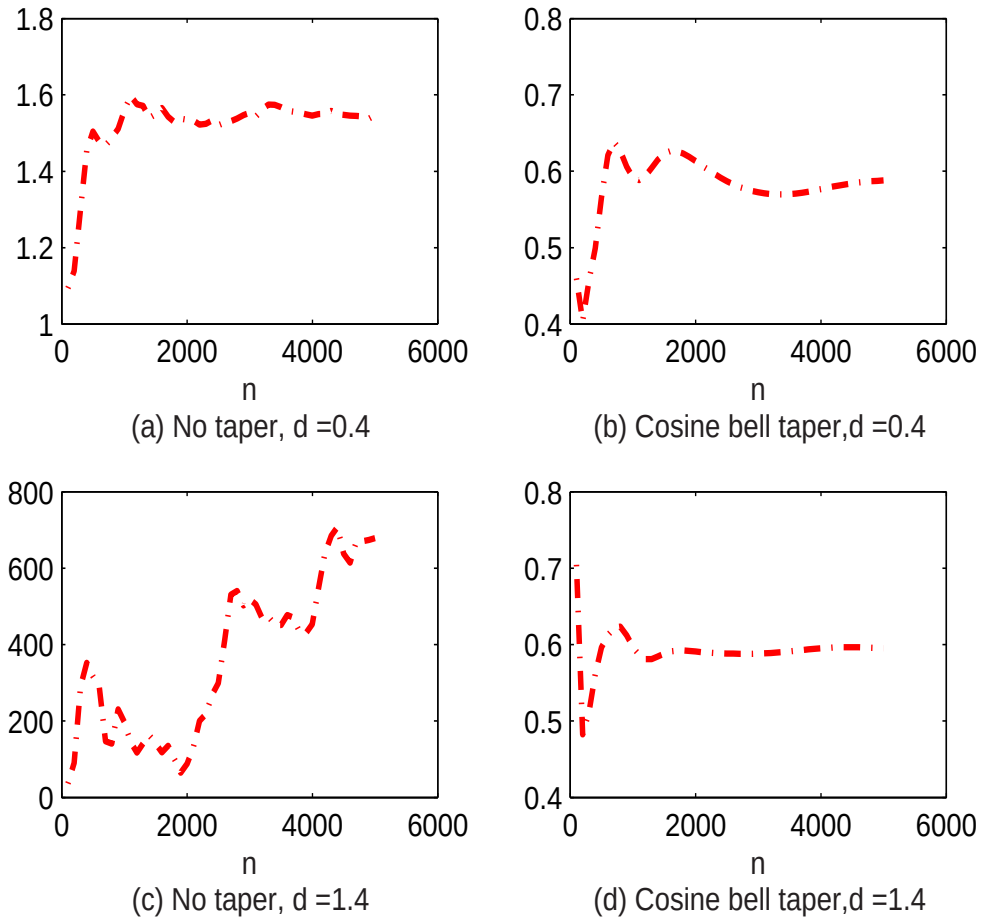

Figure 4.1: *The MSE* $\frac{1}{n}\sum_{j=1}^{n}(Y_t - \tilde{Y}_t)^2$ *of the filtering procedure in stationary ARFIMA(p=1,d=.4) case without tapering (a) and with cosine bell tapering (b). The MSE,* $\frac{1}{n}\sum_{j=1}^{n}(Y_t - \tilde{Y}_t)^2$ *of the filtering procedure in nonstationary ARFIMA(p=1,d=1.4) case without tapering (c) and with cosine bell tapering (d)*

First, we have generated an i.i.d. Gaussian noise ε_t of size 5000. Then, we have used ε_t in Davies and Harte (1987) algorithm to generate the fractional Gaussian noise $\nabla^{-d}\varepsilon_t$ with $d = .4$. Second, we have simulated an ARMA(p=1,q=0) and a stationary ARFIMA(p=1,d=.4,q=0) by the following difference equations:

$$Y_t = .2Y_{t-1} + \varepsilon_t,$$
$$X_t = .2X_{t-1} + \nabla^{-d}\varepsilon_t$$

for which $X_t = \nabla^{-d}Y_t$ holds. Note that the autoregressive coefficient is .2, and the stationary ARMA and ARFIMA time series models for this example simulations are $(1 - .2B)Y_t = \varepsilon_t$ and $(1 - .2B)\nabla^{.4}X_t = \varepsilon_t$ respectively.

The approximate $\tilde{Y}_t$ are then computed using the filtering procedure (4.33) for different n values starting from 100 to 5000 in increments of 100 observations. The MSE (4.43) is then computed and plotted in Figure 4.1(a). The estimation procedure is repeated with cosine bell taper and the MSE plot is provided in Figure 4.1(b).

To create a nonstationary ARFIMA time series, $X_t^{(1)}$, with memory parameter, $d = 1.4$, we integrated the above simulated time series X_t with order 1. It means, $X_0^{(1)} = X_0,\ X_1^{(1)} = X_0 + X_1, \cdots, X_{n-1}^{(1)} = \sum_{t=0}^{n-1} X_t$. The nonstationary ARFIMA(p=1,d=1.4,q=0) can be represented by the difference equation: $(1 - .2B)\nabla^{1.4}X_t = \varepsilon_t$.

The above procedure described for stationary ARFIMA(p,d,q) time series is repeated for the nonstationary ARFIMA(p,d,q) time series and the MSE plots without and with cosine bell tapering are provided in Figures 4.1(c) and 4.1(d) respectively .

These plots illustrate the importance of the tapering in the application of the filtering procedure, especially in the nonstationary case. When tapering is applied, the rate of convergence of the mean square error of the filtering procedure for stationary $d = .4$ and nonstationary $d = 1.4$ are almost similar.

However, these convergence rates when tapering is not applied are quite different. The MSE of the filtering procedure in the nonstationary $d = 1.4$ case never converges.

This simulation is by all means not a elaborative one. However, we repeated this simulation several times and found that the rate of convergence of the MSE in (43) is not much different from the one showed in the Figure 4.1.

We have also verified the efficiency of parameter estimates of ARFIMA(p,d,q) time series by the *direct* minimization of Whittle likelihood function for ARFIMA processes with the parameter estimates by the *2-step method*: first estimating d with the semiparametric method, then filtering the ARMA(p,q) process using the estimate of d, and finally estimating ARMA(p,q) parameters separately by the Whittle likelihood method of ARMA processes. We shall call the first method as the *direct* method. The results of the simulations are provided in different tables. We also considered nonstationary ARFIMA(p,d,q) time series with $d \in (.5, 1.5)$, and implemented parameter estimation procedures by the direct and 2-step methods. All the optimization routines are written in *matlab* code using the function *fminsearch*.

We have considered the following four different time series models in our simulations :

$$(1 - .4B)\nabla^{.2} X_t = \varepsilon_t, \tag{4.44}$$

$$(1 - .4B)\nabla^{1.2} X_t = \varepsilon_t, \tag{4.45}$$

$$(1 - .4B)\nabla^{.2} X_t = (1 + .3B)\varepsilon_t, \tag{4.46}$$

$$(1 - .4B)\nabla^{1.2} X_t = (1 + .3B)\varepsilon_t. \tag{4.47}$$

where ε_t is i.i.d Gaussian with mean 0 and variance 1 and simulated the data using the function *arima.fracdiff.sim* in S-plus. This function simulates a stationary ARFIMA time series by the algorithm developed by Haslett and Raftery (1989). In each simulation, we generated $n = 500$ observations for

each example time series and replicated these simulations $m = 200$ times.

The simulation results (mean, standard deviation, and the square root of mean squared error (MSE) of parameter estimates) for the direct and 2-step methods are summarized in the following tables. For each method, we considered three scenarios to estimate the parameters:
(i) *no-tapering* of the data,
(ii) cosine bell *tapering* of the data, and
(iii) segmenting the data with *overlap* between the segments and cosine bell *tapering* of segments.

The (iii) scenario is a special case. Here, we formed a longer time series (n=10000) by placing the 200 replications side-by-side.

The simulation results show that the combination of filtering and 2-step method provide *better* results compared to the direct method. The tapered estimates are less biased, but with slightly higher standard deviation. In the case of models (4.46 and 4.47) with moving average parameter and nonstationary memory, we did not provide any summary statistics of the parameter estimates by the direct method because this method failed to converge.

True Values	Direct method			2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = .2$	0.1001	0.1636	0.1917	0.1958	0.1555	0.1556
$\phi = .4$	0.4860	0.1614	0.1829	0.3978	0.1536	0.1536

Table 4.1: $(1 - .4B)\nabla^{.2}X_t = \varepsilon_t$ *model parameter estimates, without tapering.*

True Values	Direct method			2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = .2$	.0661	.1984	.2394	0.1779	0.2074	0.2086
$\phi = .4$	.5116	.2029	.2316	0.4163	0.2038	0.2045

Table 4.2: $(1 - .4B)\nabla^{.2}X_t = \varepsilon_t$ *model parameter estimates, with tapering.*

True Values	Direct method			2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = .2$	0.0735	0.1874	0.2261	0.1812	0.2008	0.2017
$\phi = .4$	0.5046	0.1890	0.2160	0.4116	0.1959	0.1962

Table 4.3: $(1-.4B)\nabla^{.2}X_t = \varepsilon_t$ *model parameter estimates, overlap=250 and with tapering.*

True Values	Direct method			2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = 1.2$	1.0558	0.1074	0.1797	1.0769	0.1217	0.1731
$\phi = .4$	0.0762	0.1688	0.3652	0.5430	0.1579	0.2130

Table 4.4: $(1 - .4B)\nabla^{1.2}X_t = \varepsilon_t$ *model parameter estimates, without tapering.*

True Values	Direct method			2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = 1.2$	1.1467	0.1920	0.1993	1.2570	0.2166	0.2240
$\phi = .4$	0.4422	0.1992	0.2036	0.3469	0.2103	0.2169

Table 4.5: $(1 - .4B)\nabla^{1.2}X_t = \varepsilon_t$ *model parameter estimates, with tapering.*

True Values	Direct method			2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = 1.2$	1.1479	0.1864	0.1935	1.2632	0.2048	0.2143
$\phi = .4$	0.4416	0.1896	0.1941	0.3395	0.1983	0.2073

Table 4.6: $(1 - .4B)\nabla^{1.2}X_t = \varepsilon_t$ *model parameter estimates, overlap=250 and with tapering.*

True Values	2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = .2$	0.2097	0.1387	0.1390
$\phi = .4$	0.3557	0.4595	0.4616
$\theta = .3$	0.2741	0.4574	0.4581

Table 4.7: $(1 - .4B)\nabla^{.2}X_t = (1 + .3B)\varepsilon_t$ *model parameter estimates, without tapering.*

True Values	2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = .2$	0.1750	0.1841	0.1858
$\phi = .4$	0.3100	0.5352	0.5427
$\theta = .3$	0.2042	0.5121	0.5210

Table 4.8: $(1-.4B)\nabla^{.2}X_t = (1+.3B)\varepsilon_t$ *model parameter estimates, with tapering.*

True Values	2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = .2$	0.1815	0.1862	0.1871
$\phi = .4$	0.3644	0.4919	0.4932
$\theta = .3$	0.2609	0.4804	0.4820

Table 4.9: $(1-.4B)\nabla^{.2}X_t = (1+.3B)\varepsilon_t$ *model parameter estimates, overlap=250 and with tapering.*

True Values	2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = 1.2$	1.0858	0.1300	0.1730
$\phi = .4$	0.6232	0.3744	0.4359
$\theta = .3$	0.4203	0.3497	0.3698

Table 4.10: $(1-.4B)\nabla^{1.2}X_t = (1+.3B)\varepsilon_t$ *model parameter estimates, without tapering.*

True Values	2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = 1.2$	1.2661	0.1827	0.1943
$\phi = .4$	0.3419	0.4785	0.4820
$\theta = .3$	0.3136	0.4889	0.4891

Table 4.11: $(1-.4B)\nabla^{1.2}X_t = (1+.3B)\varepsilon_t$ *model parameter estimates, with tapering.*

True Values	2-step method		
	Mean	Stdev	$\sqrt{\text{MSE}}$
$d = 1.2$	1.2713	0.1884	0.2014
$\phi = .4$	0.3458	0.4621	0.4653
$\theta = .3$	0.3207	0.4761	0.4765

Table 4.12: $(1-.4B)\nabla^{1.2}X_t = (1+.3B)\varepsilon_t$ *model parameter estimates, overlap=250 and with tapering.*

Chapter 5

EEG data analysis

In this chapter, we present a thorough analysis of an EEG C3-FP1 channel time series for a fullterm baby and also a brief summary of an analysis of for a preterm neonate. The fullterm EEG was collected for little more than 3 hours at a sampling frequency of 64HZ. An initial look at this signal in Figure 5.1 suggests some spikes in the data, which could be artifacts and could be filtered out using a suitable filter. However, we treated them as a part of the signal because we did not know the exact cause of such spikes.

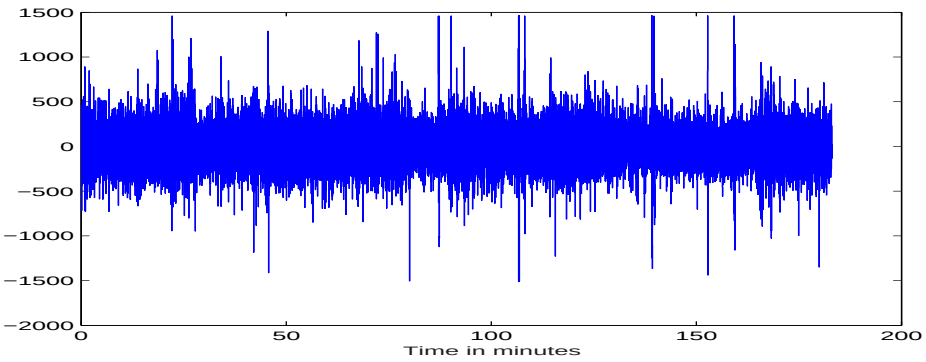

Figure 5.1: *Fullterm EEG.*

The ACF and PACF plots in Figure 5.2 provide the information about

short and long memory properties of this signal. The rate of decay of ACF plot suggests that this signal has long-range dependence and the PACF value (≈ 1) at lag one indicates the presence of stochastic trend of order one. The tv-periodogram plot (logarithmic scale), in Figure 5.3, with one minute cosine bell tapered segments illustrates the time-varying nature of short and long memory properties of the EEG signal.

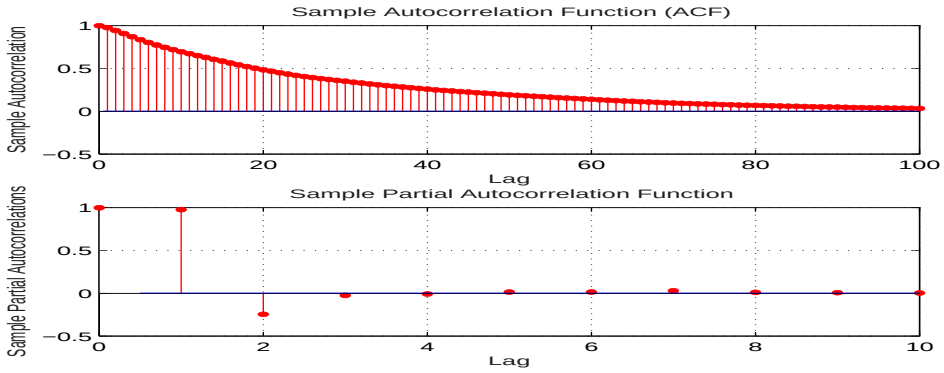

Figure 5.2: *ACF and PACF* .

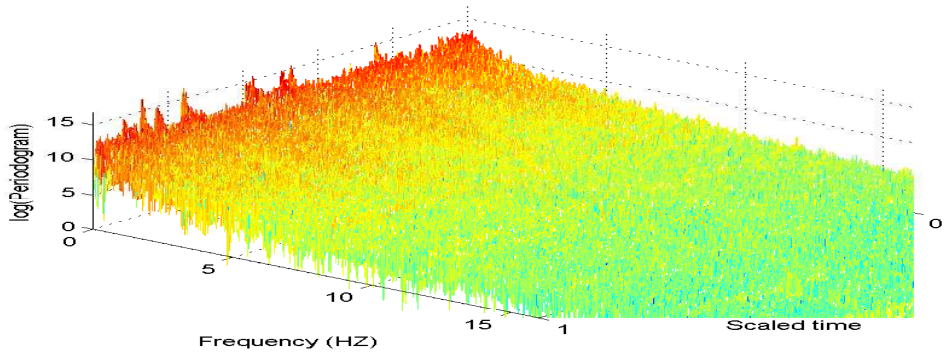

Figure 5.3: *Time-varying Periodogram.*

The presence of stochastic trend of order one and long memory makes $(0, 1.5)$ as the acceptable range of values for the memory parameter d. Since

the sleep states are defined minute by minute, we segmented this EEG signal with $N = 60 \times 64$ length windows with no overlapping ($S = N$), where S indicates the shift from one segment to the next segment. Note that the value of $(N - S)$ represents the actual overlap between the successive segments.

In our case N is predetermined. However one could select any N-value and any S-value. If segment length are unknown then one must start with small $N \geq 1$ and $S \geq 0$ values. Smaller N-values improve the variance of the estimators but increases the bias and requires more computational time. On the other hand, small S improves the the smoothness of the parameter estimators and again requires more computational time. We could have selected segments of length shorter than a minute. This would effect the efficiency of the filtering procedure to separate the short memory component from the EEG signal.

The choice of cosine bell taper is due to the presence of stochastic trend of order 1 in the signal. This taper, in addition to good bias properties with respect to leakage (cf. Section 2.3), also has the advantage of the observations at the edge of each segment being weighted down which makes the estimate heuristically less sensitive to the nonstationarity within the segments.

The results of application of the algorithm in Section 3.2 are summarized in Figures 5.4 to 5.15. Figure 5.4 shows the time-varying nature of the semiparametric estimate of the nonstationary memory. These estimates are smoothed by the polynomials of 8th degree. Figure 5.5 shows smoothed time-varying nonstationary memory estimates and their point-wise confidence intervals (Section 2.2.3) assuming that the innovations of the EEG signal is Gaussian. Figure 5.6 shows the same smoothed time-varying memory estimates along with the prediction interval of the polynomial fit (See *polyfit* command in MATLAB manual). This figure suggests that the 8th degree polynomial we selected to smooth the time-varying memory parameter provides a good fit. One could use lower or higher degree polynomials to smooth

the memory estimates. The higher the degree of polynomial, the lesser is the smoothness of the fit. One could also look at the prediction intervals of this polynomial smoothing. Smoothing of this parameter is not required but it provides a better visualization of time-varying nature of the memory over segments.

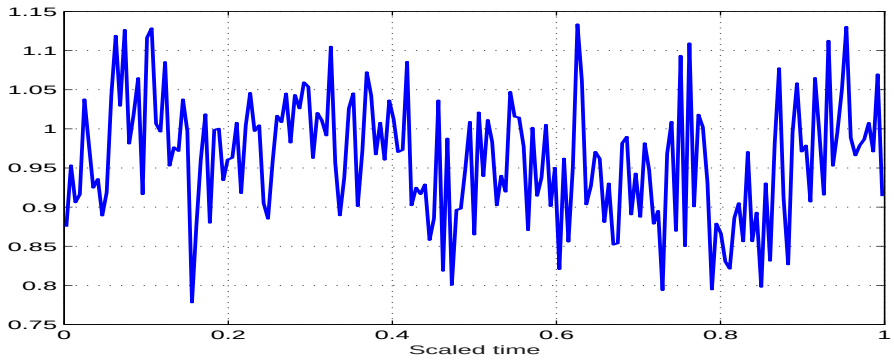

Figure 5.4: *Time-varying memory parameter estimate.*

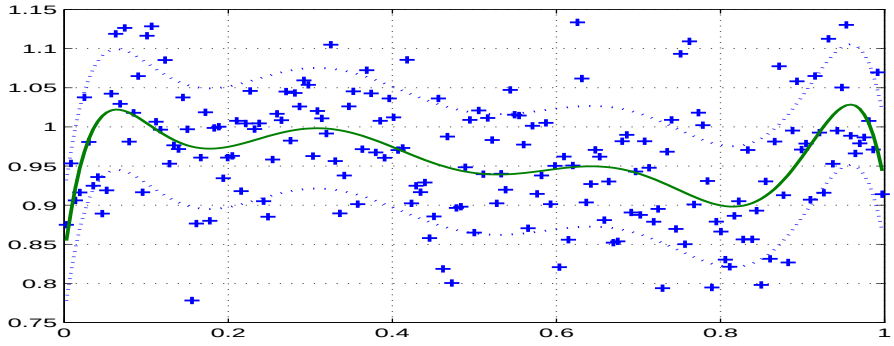

Figure 5.5: *Smoothed memory parameter estimate and confidence interval.*

We obtained the short memory component of EEG signal, by applying the filtering technique in Chapter 4, and plotted in Figure 5.7. For verification, we plotted the ACF and PACF plots of the short memory signal in Figure 5.8.

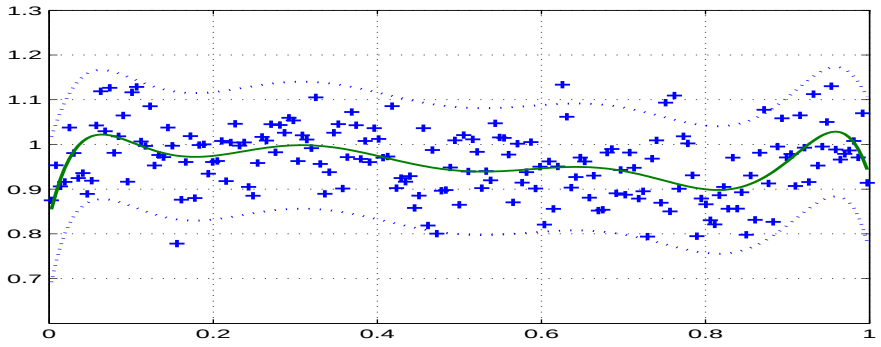

Figure 5.6: *Smoothed memory parameter estimate and prediction interval.*

These plots indicate the absence of memory and stochastic trend of order 1 in this signal. Now, it may be safe to claim it is a short memory signal. We again segmented this signal with a smaller window size, *N=32*(half a second) with no overlap between the segments (N=S). The time-varying periodogram plot in Figure 5.9 shows *no significant peaks* around zero frequency over different segments. One could apply a much smaller or larger window sizes. Smaller windows capture local short memory properties better and larger windows are computationally efficient. We also looked at the AIC criterion (cf. Section 2.1.3), Tables 1 to 3, for selecting the window size. We selected time-varying AR(2) model to fit the short memory data.

Only the time-varying AR models are considered in this dissertation. The obvious reason is that, the generalized likelihood function (3.16) for tv-AR models produces closed form estimators (Yule-Walker estimators $\hat{a}(u_j)$ and $\hat{\sigma}^2(u_j)$ in (3.17) and (3.18) respectively). The variance estimator in (3.18) is valid only in the case of Gaussian innovations. In the case of Gaussian innovations, one can use the Yule-Walker estimators $\hat{a}(u_j)$ and $\hat{\sigma}^2(u_j)$ as the initial estimators to obtain smooth final estimates which minimizes (3.16). This additional step produces smooth final estimates. The final smoothed first time-varying AR parameter estimate $\hat{a}_1(u_j)$ is shown in Figure 5.12. We

again employed polynomial smoothing to get a better visualization of this estimate. The order of the polynomial is selected from the AIC criterion. The confidence intervals for this parameter are derived from (2.22) by assuming Gaussian innovations.

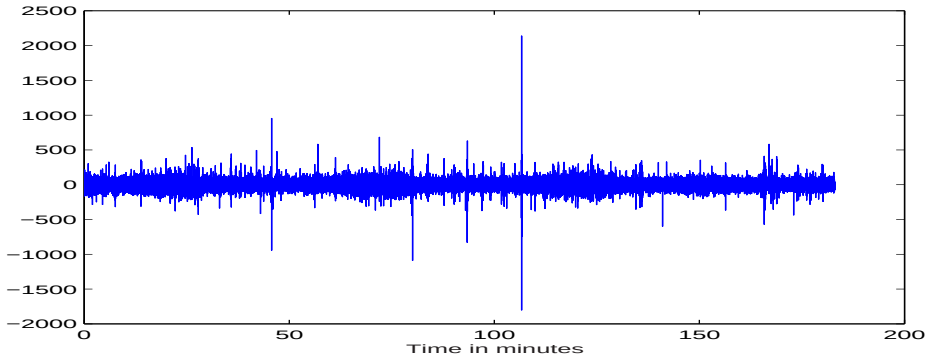

Figure 5.7: *Fullterm short memory EEG.*

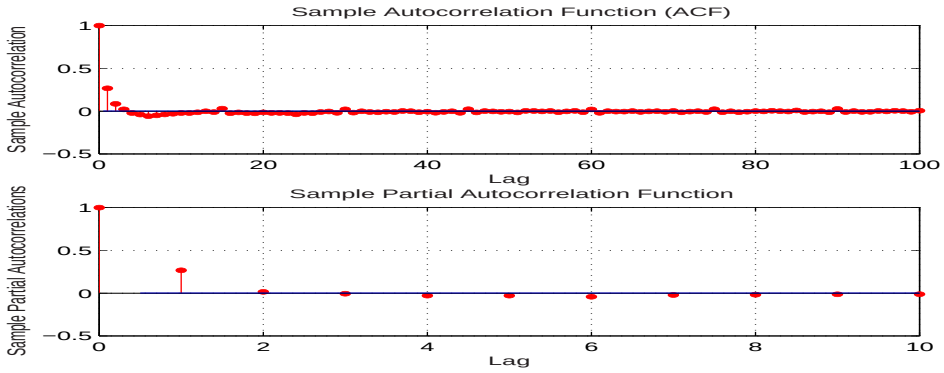

Figure 5.8: *ACF and PACF plot of short memory EEG.*

To calculate the time-varying MA and ARMA parameters one needs to optimize (3.16) directly which is computationally intensive and the minimization of the generalized Whittle likelihood function is almost impossible to achieve for large number of segments. Also, the asymptotics of time-

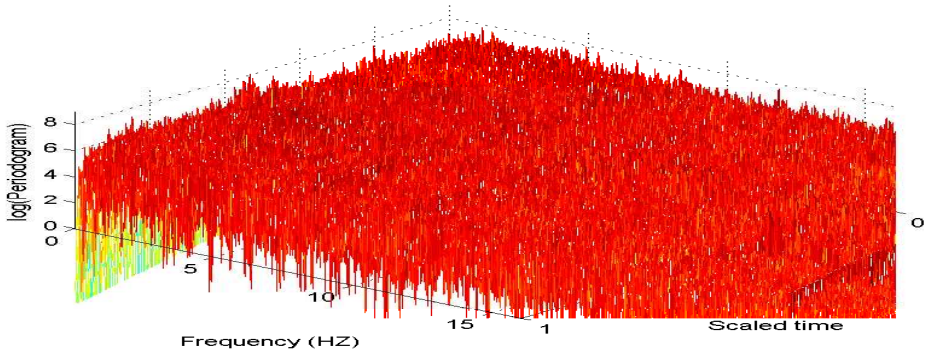

Figure 5.9: *Time-varying Periodogram of short memory EEG.*

varying MA estimators are difficult to derive, even in the case of Gaussian innovations.

The selection of the order of the AR model (p) and the degree of polynomial ($K_j, \quad j = 1, \cdots, p$) for smoothing the time varying AR parameters are selected by the minimization of the AIC criterion (2.26). Knight(1988) showed that the AIC based criterion for model selection of stable AR processes is still consistent. The variance term $\hat{\sigma}^2$ in AIC criterion (2.26) is the variance estimate obtained from the Yule-Walker estimator. There are two problems with the AIC criterion for short memory EEG data fitting: (i) variance changes over time, and (ii) the length of the time series is so large that the total number of parameters term in (2.26) is almost negligible. We still found the AIC criterion is somewhat useful in our model selection, in spite of these problems.

We modified the AIC criterion 4 different ways by changing the $\hat{\sigma}^2$ in (2.26):

(i) In *AIC1*, by the mean of the $\{\hat{\sigma}^2(u_j)\}$, where $j = 1, \cdots, M$,

(ii) In *AIC2*, by the variance estimate of the entire residual time series of the short memory signal,

(iii) In *AIC3*, by maximum of all the $\{\hat{\sigma}^2(u_j)\}$, and

(iv) In *AIC4*, by the minimum Whittle likelihood function value (3.16).

We found the AIC3 criterion to be the only one that is really useful. The following Table 5.1 is constructed by selecting $K_2 = 0$, where K_2 is the smoothing polynomial order of the second AR parameter.

K_1	4	5	6	7	8	9	10	11	12
AIC1	7.014	7.013	7.013	7.012	7.011	7.011	7.009	7.009	7.007
AIC2	6.998	6.998	6.998	6.996	6.996	6.996	6.994	6.994	6.992
AIC3	12.065	12.078	12.076	12.053	12.055	12.055	12.059	12.058	12.069
AIC4	11.059	11.059	11.059	11.058	11.058	11.058	11.058	11.058	11.057

Table 5.1: Values of AIC for $p = 2$ and different values of K_1 for N=64 and $K_2 = 0$

It is obvious from Table 5.1 that all the AIC criteria, except AIC3, are non-increasing as K_1 is increasing. The minimum value of the AIC3 criterion, in the Table 5.1, can be seen at $K_1 = 8$.

The different AIC criteria may also be applied to select the window size, N, of short memory EEG and the order of the AR model, p. Table 5.2 summarizes different AIC criteria values for N=64 and p=2, and Table 5.3 summarizes AIC criteria values N=32 and p=1.

K_1	4	5	6	7	8	9	10	11	12
AIC1	6.983	6.982	6.982	6.980	6.980	6.980	6.977	6.977	6.976
AIC2	6.953	6.953	6.953	6.951	6.951	6.951	6.9482	6.948	6.946
AIC3	12.723	12.734	12.733	12.713	12.711	12.712	12.716	12.715	12.728
$AIC4$	11.066	11.066	11.066	11.065	11.065	11.065	11.065	11.065	11.064

Table 5.2: Values of AIC for $p = 2$ and different values of K_1 for N=32 and $K_2 = 0$.

We preferred to over-fit the data by selecting $N = 32$, $p = 2$, $K_1 = 8$, and $K_2 = 0$ to get the residuals. The second time-varying AR parameter, $a_t(u_j)$ in fitting the short memory EEG signal is almost independent of segments and is zero for most of the segments.

K_1	4	5	6	7	8	9	10	11	12
AIC1	7.015	7.014	7.014	7.013	7.012	7.012	7.010	7.010	7.008
AIC2	6.984	6.984	6.984	6.982	6.982	6.982	6.979	6.979	6.978
AIC3	12.732	12.742	12.742	12.724	12.723	12.723	12.726	12.727	12.737
$AIC4$	11.067	11.067	11.067	11.067	11.067	11.067	11.066	11.066	11.065

Table 5.3: Values of AIC for $p = 1$ and different values of K_1 for N=32.

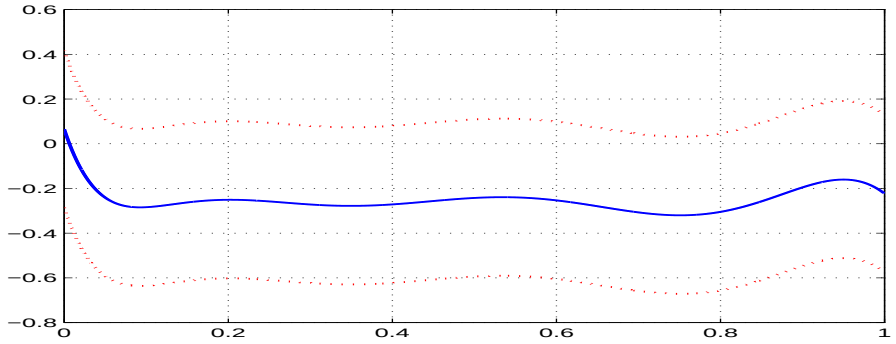

Figure 5.10: *First time-varying AR estimate from generalized Whittle method.*

Figure 5.10 shows the smoothed estimate of the first time-varying AR parameter, $a_1(u_j)$ along with it's smoothed 95% point-wise confidence interval. Figure 5.11 shows the predicted short-memory EEG signal and Figure 5.12 shows the *standardized* residuals of the model fit. Figures 5.13 and 5.14 show the *Q-Q plot* and, ACF and PACF plots. The ACF and PACF plots show no significant ACF and PACF values.

The goodness-of-fit of model fitting is verified by the Kolmogorov-Smirnov test to check the Gaussianity (cf. Denker and Woyczynski, 1998), the Fisher test for hidden periodicities (cf. Bloomfield, 2000), and the Ljung-Box Q-test (LBQ test) for the lack of independence (cf. Brockwell and Davis, 1996) of the residuals.

The Fisher test statistic is easy to understand because it looks at the maximum periodogram of the data and tests whether it is significantly larger

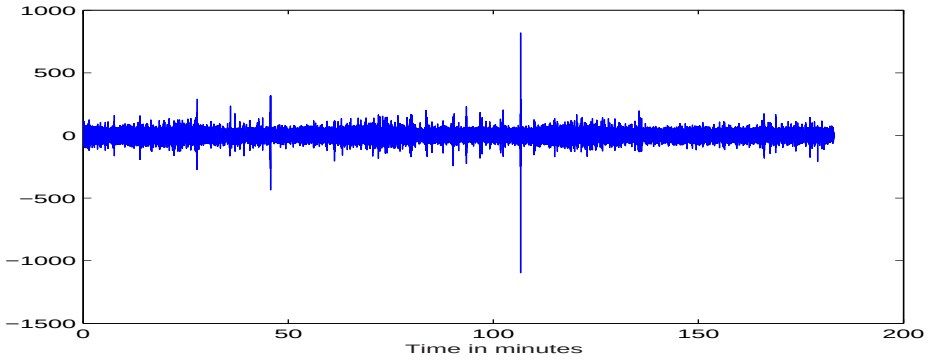

Figure 5.11: *Predicted short memory time series.*

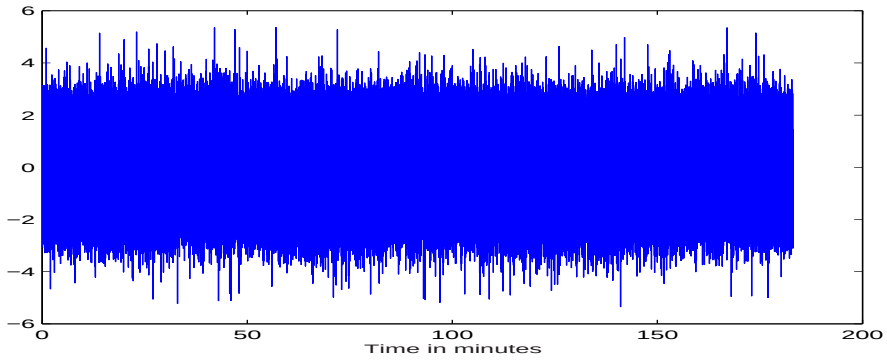

Figure 5.12: *Standardized residuals time series.*

than any other periodograms. The test statistic is easy to construct because periodograms are approximately independent exponential random variables for the Gaussian white noise. Then, the maximum order statistic of independent exponential random variables is *Gumbel* (cf. Ferguson, 1996). The test statistic and significant values are easy to compute for the Gumbel distribution.

On the other hand LBQ test is a powerful lack-of-fit test for the departure from randomness. The *Q-statistic* is based on the sum of the squared autocorrelations over a range of lags. Under the null hypothesis that the model

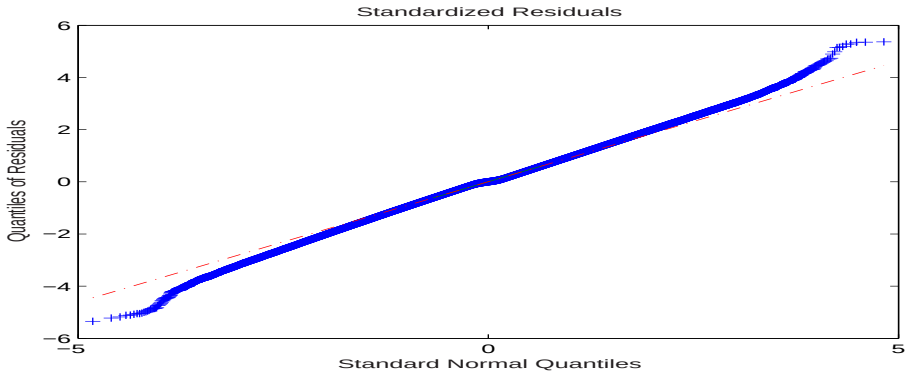

Figure 5.13: *Standardized residuals QQ plot .*

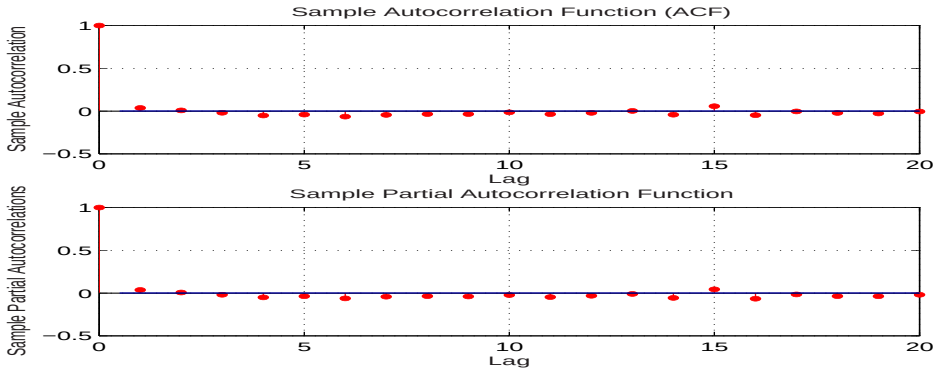

Figure 5.14: *ACF and PACF plots of residuals.*

fit is adequate, the test statistic is asymptotically Chi-square distributed.

The whole residual series of the fullterm EEG fails all these three tests. However, how individual segments residuals fare is more interesting to see. The following Table 5.4 looks at *the percentage of segments* which rejected the null hypotheses of the three goodness-of-fit tests. We have also verified these tests for different polynomial degree (K_1) fit of the first AR parameter.

Of M=21960 segments' residuals, only 78 (.35%) segments residuals have rejected Kolmogorov-Smirnov test for Gaussianity. Of M=21960 segments' residuals, only 2081 (9.48%) segments have rejected Fisher test for white

K_1	4	5	6	7	8	9	10	11	12
KS-test	0.37	0.37	0.37	0.34	0.35	0.35	0.36	0.36	0.35
Fisher	9.54	9.57	9.55	9.49	9.48	9.47	9.49	9.49	9.41
LBQ	20.88	20.81	20.81	20.76	20.69	20.71	20.69	20.68	20.49

Table 5.4: Goodness-of-fit tests for tv-AR($p = 2$) model and different values of K_1, $K_2 = 0$ and N=32.

noise. Of M=21960 segments residuals, 4543 (20.69%) segments have rejected LBQ-test for lack of randomness.

For comparison, we generated and tested Gaussian random variables of same sample size (702720) as the size of EEG signal. Just as in the case of EEG signal, we segmented this sequence by $N = 32$ and obtained $M = 21960$ segments. Of these M segments 1158 segments (5.7%) of segments have rejected KS-test for Gaussianity, 792 segments (3.6%) of segments have rejected Fisher test for no periodic components. Of M=21960 segments residuals, 1873(8.53%) segments have rejected LBQ-test for lack of randomness. The interesting observation comes from the last statement because 8.53% is larger than the usual significance value of 5%. However, this is one sample and we do not try to interpret too much out of it.

The last four plots in this section are most interesting. Figure 5.15 shows the actual residual series of model fitting. The signal to noise ratio may not have improved from the short memory model fitting. Figure 5.16 shows the estimate of the time varying stable parameter (α) for the residual series' segments of a minute size. This estimate is calculated using the McCulloch (1986) method and it is summarized in the Appendix A. We plotted side-by-side *boxplots*, Figure 5.17, of segments' sleep states and $\hat{\alpha}$.

One can see that, the Trace Alternate Quiet (TAQ) sleep state has smaller $\hat{\alpha}$) values, which indicate these sleep states have more heavy tail distribution than the other sleep states. Figure 5.18 shows side-by-side boxplots of sleep states and memory parameter (d) estimates. The quiet sleep states (TAA

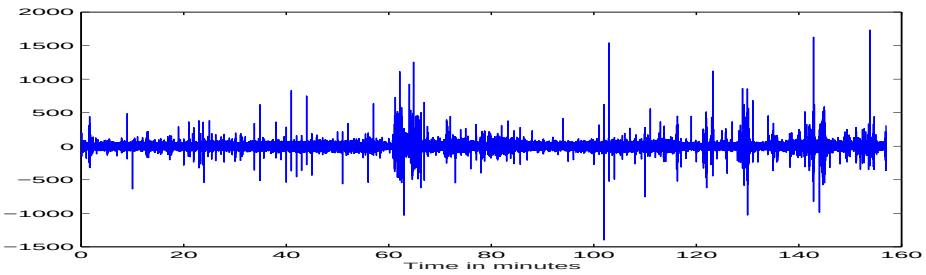

Figure 5.15: *Residuals plot* .

and HVS) have lesser degree of long-range dependence than the active sleep states. Memory estimate of a sleep state could help in discriminating it from other states (cf. Piryatinska, 2004).

The model fitting of preterm EEG has produced similar results but comparatively larger estimates for the time-varying memory and α-stable parameters. We do not provide details about the fitting of this baby data because the procedure is to fit this data is similar to the fullterm EEG. The model fitting parameters such as segment lengths, overlap between the segments, and the type of model to fit the short memory signal are similar to the full term baby model fitting parameter. We used the AIC criterion to select these model fitting parameters. Also, we verified the goodness-of-fit of this preterm baby model fitting.

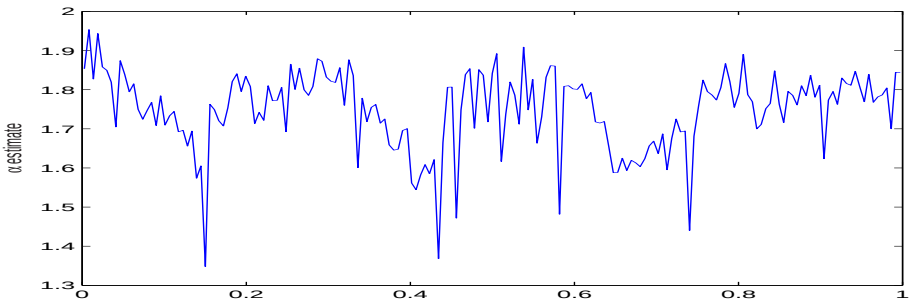

Figure 5.16: *Time-varying α estimates.*

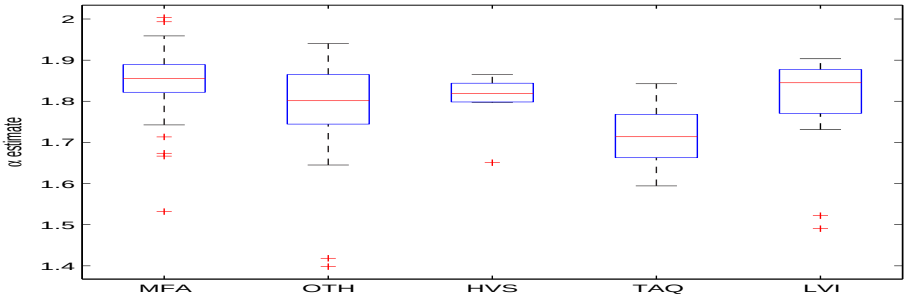

Figure 5.17: *α boxplots by sleep states.*

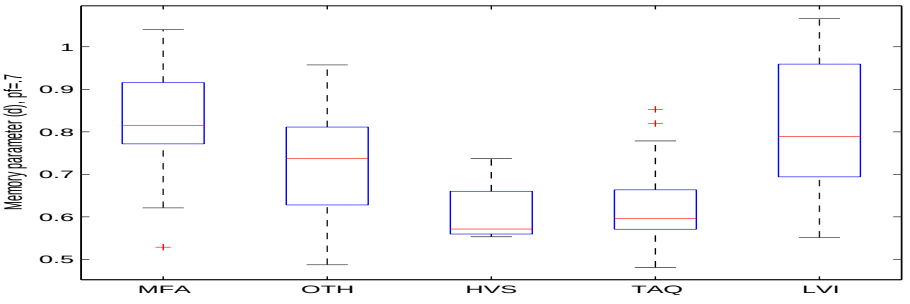

Figure 5.18: *Memory boxplots by sleep states.*

Chapter 6

Conclusions

We have defined a new locally stationary time-varying ARFIMA model with long memory parameter, d, contained in the interval $[-.5, .5)$. We have also estimated parameters of this model by developing a computationally efficient 2-step algorithm. These parameter estimators are minimum distance estimators in the spectral domain. Since our distance function is an approximate Gaussian likelihood, the parameter estimation method, in principle, can be applied to models whose parameters can be identified from this distance function, mainly time-varying linear processes.

Our algorithm estimates the time-varying long memory parameter and short memory parameters separately in two steps. In general, the long memory and short memory parameters of a stationary process are estimated simultaneously to reduce the bias of the estimators. However, we found such method practically infeasible when the number of parameters to estimate is large, which is the case with the time-varying ARFIMA models.

The first step of our algorithm uses a semiparametric method (Robinson, 1995) to estimate the long memory parameter and the second step uses a modified version of the minimum distance method (Dahlhaus, 1997) to estimate the time-varying short memory parameters. We have adapted a computationally efficient filtering theory that separates the unobserved ARMA

time series (Y_t) from an observed ARFIMA process (X_t) and derived the order of magnitude of the error term, $E\{\mid R_{X,Y,n}(\lambda \mid^2$. We have demonstrated via simulations that this filtering technique is more efficient compared to the derived order of magnitude, when an appropriate taper is applied to the original data.

The model selection of time-varying ARFIMA processes with Gaussian innovations using the AIC criterion seems to work reasonably well. It is a computationally intensive criterion. Even after all its shortcomings, it is possible to apply the AIC criterion to decide between stationary and locally stationary models, for example by testing the order of polynomials (which may be used to fit the time-varying parameters) for zero.

We have modified the minimum distance method to fit the time-varying fractional AR models with memory ($d > .5$) and infinite variance innovations, which has been the case with the neonates' sleep EEG. This modified method provides the same time-varying AR estimators as the minimum distance method. The main advantage of this method is that it does not require a priori assumptions about the distribution of innovations, especially whether the variance of innovations is finite or not. In the case of Gaussian innovations, the estimators of the modified minimum distance method are still the same. The nonstationary memory parameter is estimated by applying the cosine bell taper to the original data, which slightly increases the variance of time-varying memory estimators.

We have presented a thorough analysis of sleep EEG model fitting using the 2-step procedure and the filtering technique. At this point, it is premature to say that the time-varying ARFIMA models provide the best fit for sleep EEG. However, it is impossible to deny the fact that the results from this model fitting are *exciting*. More can be done in fitting the sleep EEG data and exploring the application of such model fitting. In the near future, more runs will be executed on several neonates' EEG.

The future work should include further theoretical and applied work on the local stationarity concept. In the following, we specify only a few of the possible extensions but potential can be seen on various fronts.

(i) Our immediate interest is to develop a *reasonable* order of magnitude of the error term ($R_{X,Y,n}(\lambda)$) in the filtering technique to separate the short memory signal from a long memory process. In particular, when the memory is nonstationary, $d > .5$. The order of magnitude of $R_{X,Y,n}(\lambda)$ term in Theorems 4.1 and 4.2 do not reflect the benefits of tapering. However, our simulations suggest that the order of magnitude of this error term improves as one applies tapering to the original data. Our immediate focus is to develop explicit order of magnitude of error term for the cosine bell tapered discrete Fourier transforms of ARFIMA processes with memory parameter greater than .5. This research can be applied in theory and applications. It will help to develop asymptotics of periodograms of nonstationary processes in terms of the well known asymptotics of stationary processes. Moreover, it will help in parameter estimation of ARMA processes with nonstationary memory in spectral domain and in the development of asymptotics of these parameter estimators. This research also has tremendous potential in economics and financial time series model fitting and forecasting.

(ii) Our second future goal is to apply *wavelets* to fit the time-varying parameters of sleep EEG using the theory developed by Neumann and Von Sachs (1997). This theory can be applied to develop a new adaptive wavelet estimator of the time-dependent spectrum of a locally stationary time series. Such method will overcome the difficulty of how to choose the right amount of smoothing, that is, how to adapt to the appropriate resolution, for reconstructing the local structure of the evolutionary spectrum in the localized time-frequency plane. Wavelets may fit unhealthy neonates' sleep EEG signals better, because of the discontinuities in these signals. Here, we have done some introductory work in fitting the healthy neonates' sleep EEG

data with the *Haar* wavelets.

(iii) Our third future goal is to improve the 2-step method to obtain more accurate estimates. This goal can be achieved because the step 2 of this method is similar to the conditional likelihood method. One can use the 2-step method estimates as initial values to minimize the exact likelihood function. Also, we can develop conditional likelihood function for time-varying short memory process by conditioning on the time-varying long memory process and viceversa and repeat the 2-step procedure until a local (or the global) minima reached. This method is similar to the *Gibbs sampling*.

(iv) Our fourth future goal is to develop asymptotic theory for the time-varying fractional AR models with Gaussian innovations. In particular, developing the asymptotics for time-varying memory parameter. One could unify the theory of asymptotics of long memory in Robinson (1995) and asymptotics of time-varying nonstationary models in Dahlhaus (1997).

(v) Our fifth future goal is to to extend this dissertation to multivariate settings and ultimately to random fields. Here, we emphasize specific applications. One such application, which is of tremendous interest, is the model fitting of cancerous tumor growth by contributing to the concept of locally stationary random fields (cf. Dahlhaus and Sahm, 2000).

(vi) We have also done some introductory research in Principle Component Analysis (PCA) of multivariate locally stationary time series similar to the PCA of multivariate stationary time series in frequency domain (cf. Brillinger, 1981). We see application of such research in dimension reduction of multivariate EEGs. Currently, EEG signals are recorded from 256 channels. Understanding the multivariate structure and model fitting (time-varying ARIMA and ARFIMA models) of these signals have significant applications in medical diagnostics. Applications can also be seen in financial and economics time series.

(vii) A possible extension can be seen in the forecasting of a time series with evolutionary structure using local stationary models. It is possible by fitting the data up to $u = u_1 < 1$ (where $u = t/n$) and predicting the data for the interval from u_1 to 1.

(viii) Another possible extension can be seen in the parameter estimation of time-varying fractional MA processes.

(ix) In this dissertation, we have not emphasized enough about the effect of time-varying mean, consistency of time-varying memory parameter estimators, local polynomial trends, and stochastic trends associated with the sleep states in the model fitting. Especially, the effect of zero frequency in the analysis of long memory time series. A thorough analysis of these properties of sleep EEG will provide better results.

Appendix A

McCulloch estimator of stable parameters

Let X be a $S_\alpha(\sigma, \beta, \mu)$ process (Section 3.3) and denote p-th quantile of this process by $X_{(p)}$. McCulloch(1986) estimator uses five quantiles of data to estimate $\alpha \in [.6, 2.0]$ and $\beta \in [-1, 1]$, and are derived as follow:

Set

$$\begin{aligned} \phi_1(\alpha, \beta) &= \frac{X_{(95)} - X_{(5)}}{X_{(75)} - X_{(25)}}, \\ \phi_2(\alpha, \beta) &= \frac{X_{(95)} + X_{(5)} - 2X_{(50)}}{X_{(95)} - X_{(5)}}. \end{aligned}$$

Since ϕ_1 is monotonic in α and ϕ_2 monotonic in β (for fixed α), we can invert these functions to obtain

$$\begin{aligned} \alpha &= \psi_1(\phi_1, \phi_2), \\ \beta &= \psi_2(\phi_1, \phi_2). \end{aligned}$$

McCulloch tabulated ψ_1 and ψ_2 for various values of ϕ_1 and ϕ_2.

To form an estimator, take a random sample from a stable distribution and define $\hat{\phi}_1$ and $\hat{\phi}_2$ by replacing the quantiles $X_{(p)}$ with the corresponding

sample quantiles $\hat{X}_{(p)}$. Since the sample quantiles are consistent estimators for the population quantiles, the $\hat{\phi}_1$ and $\hat{\phi}_2$ are consistent estimators of ϕ_1 and ϕ_2. Then, using tables,

$$\hat{\alpha} = \psi_1(\hat{\phi}_1, \hat{\phi}_2),$$
$$\hat{\beta} = \psi_2(\hat{\phi}_1, \hat{\phi}_2).$$

To obtain estimates of the scale and location parameters, McCulloch defined similar functions using these same quantiles, which were also tabulated for various scale and location values. These tables can be used similar way to obtain estimates for scale and location parameters.

Appendix B

Simulating a fractional Gaussian time series

The following method (Davies and Harte, 1987) was used to generate the fractional Gaussian noise (FGN) in Chapter 5 simulations. This method generates a series of FGN of length $n+1$ by performing $O(n\log(n))$ operations.

Let $X_0, X_1, \cdots, X_n$ denote a series of stationary FGN with fractional parameter d in the range $(0, .5)$, then the covariance c_k between X_1 and X_k is defined (cf. Beran, 1994) by

$$c(k) = \sigma^2 \left(\frac{1}{2} \mid k+1 \mid^{2d-1} + \frac{1}{2} \mid k-1 \mid^{2d-1} - \mid k \mid^{2d-1} \right), \tag{B.1}$$

where σ^2 is the variance of the Gaussian noise. For convenience, we assume $\sigma^2 = 1$.

Let g_k denote the inverse Fourier transform of the sequence $\{c_0, c_1, \cdots, c_{n-1}, c_n, c_{n-1}, \cdots, c_1\}$. That is

$$g_k = \sum_{j=0}^{n-1} c_j \exp(i\lambda_k j) + \sum_{j=n}^{2n-1} c_{2n-j} \exp(i\lambda_k j)$$

where $\lambda_k = 2\pi k/(2n)$ for $k = 0, 1, \cdots, 2n-1$.

The series g_k will be real, but for the method to work it must be nonnegative. If any of the g_k are negative then the method fails. This has not been a problem in the simulations of this dissertation.

Let $\{Z_k\}$ be a sequence of independent zero mean complex Gaussian random variables with independent real and imaginary parts, for $0 \leq k \leq n$, and let $Z_k = \bar{Z}_{2n-k}$ for $n < k < 2n$. Let Z_0 and Z_n be real with variances equal to two, otherwise, for $0 < k < n$, let the real and imaginary parts of Z_k have unit variance.

Then the FGN time series

$$X_j = \frac{1}{2} n^{-1/2} \sum_{k=0}^{2n-1} Z_k g_k^{1/2} \exp(i\lambda_k j)$$

for $0 \leq j \leq n$ has the required covariance structure (A.1).

Bibliography

[1] Adler, R., Feldman, R., and Taqqu, M. (1986). *A Practical Guide to Heavy Tails*. Birkhauser, Boston.

[2] Azencott, R. and Dacunha-Castelle, D. (1986). *Series of Inrregular Observations: Forecasting and Model Building*. Springer-Verlag, New York.

[3] Beran, J. (1994). *Statistics for Long-Memory Processes*. Chapman-Hall, New York.

[4] Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction*. Wiley, New York.

[5] Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. Holden-Day Inc., San Francisco.

[6] Brockwell, P. and Davis, R. (1996). *Time Series: Theory and Methods*. Springer, New York.

[7] Cohen, L. (1989). Time-Frequency: a review. *Proceedings of the IEEE*. **77** 941-981.

[8] Davies, R. B. and Harte, D. S. (1987). Tests for Hurst effect. *Biometrika*. **74**(1) 95-101.

[9] Dahlhaus, R. (1988). Efficient Parameter Estimation for Self-Similar Processes. *Annals of Statistics*. **17**(4) 1749-1766.

[10] Dahlhaus, R. (1996). On the Kullback-Leibler Information Divergence of Locally Stationary Processes. *Stochastic Processes and Their Applications*. **62** 139-168.

[11] Dahlhaus, R. (1997). Fitting Time Series Models to Nonstationary Processes. *Annals of Statistics*. **25** 1-37.

[12] Dahlhaus, R. and Sahm, M. (2000). Likelihood Methods for Nonstationary Time Series and Random Fields. *Resenhas*. **4** 457-477.

[13] Denker, M. and Woyczynski, W. A. (1998). *Introductory Statistics and Random Phenomena*. Birkhauser, Boston.

[14] Dzhaparidze, K.(1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*, Springer-Verlag, New York.

[15] Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman-Hall, New York.

[16] Fox, R. and Taqqu, M. S. (1986). Large Sample Properties of Paramter Estimates for Strongly Dependent Stationary Gaussian Time Series. *Annals of Statistics*. **14** 517-532.

[17] Hosking, J. R. M. (1981). Fracional Differencing. *Biometrica*. **68**(1) 165-176.

[18] Hurvich, C. M. and Ray, B. K. (1995). Estimation of the Memory Parameter for Nonstationary or Noninvertible Fractionally Integrated Processes. *J. of Time Series Analysis*. **16** 17-41.

[19] Kluppelburg, C. and Mikosch, T. (1993). Spectral Estimates and Stable Processes. *Stochastic Processes and their Applications*. **47** 323-344.

[20] Kluppelburg, C. and Mikosch, T. (1994). Some Limit Theory for the Self-normalised Periodogram of Stable Processes. *Scand. J. of Statistics*. **21** 485-491.

[21] Knight, K. (1989). Consistency of Akaike's Information Criterion for Infinite Variance Autoregressive Processes. *Annnals of Statistics*. **17** 824-840.

[22] Leonenko, N. and Woyczynski, W. (2001). Parameters Identification for Stochastic Burgers; Flows via Parabolic Scaling. *Probability and Mathematical Statistics*. **21**(1) 1-55.

[23] McCulloch, J. H. (1986). Simple Consistent Estimators of Stable Distribution Parameters. *Commun. Statistics - Simulations*. **15**(4) 1109-1136.

[24] Mikosch, T., Gadrich, T., Kluppelburg, C. and Adler, R. J. (1995). Parameter Estimation for ARMA Models with Infinite Variance Innovations. *Annals of Statistics*. **23** 305-326.

[25] Neumann, M. H. and von Sachs, R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Annnals of Statistics*. **25** 38-76.

[26] Ombao, H. C. Jr (1999). Statistical Analysis of Nonstationary Time Series. *UMI Dissertation Services*. Ann Arbor, Michigan, USA.

[27] Piryatinska, A. (2004). Inference for the Levy Models and their Applications in Medicine and Statistical Physics. *CWRU Dissertation Services*. Cleveland, OH, USA.

[28] Preistley, M. (1981). *Spectral Analysis and Time Series*, Academic Press.

[29] Raftery, A. E. and Haslett, J. (1989). Space -Time Modeling with Long-memory Dependence: Assessing Ireland's Wind Power Resource. *Applied Statistics*. **38** 1-50.

[30] Robinson, P. M. (1995). Gaussian Semiparametric Estimation of Long Range Dependence. *Annals of Statistics*. **23** 1630-61.

[31] Samorodnitsky, G. and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes*. Chapman-Hall, New York.

[32] Taqqu, M. S., Teverovsky, V. and Willinger, W. (1995). Estimators for Long-Range Dependence: an Empirical Study. *Fractals*. **3** 785-798.

[33] Velasco, C. (1999a). Non-stationary log-periodogram regression. *J. of Econometrics*. **91** 325-371.

[34] Velasco, C. (1999b). Gaussian Semiparametric Estimation of Non-stationary Time Series. *J. of Time Series Analysis*. **20** 87-127.

[35] Walker, A. M. (2000). Some Results Concerning the Asymptotic Distribution of Sample Fourier Transforms and Periodograms for a Discrete-Time Stationary Process with a Continuous Spectrum. *J. of Time Series Analysis*. **21** 95-109.

[36] William, W., Zaveri, H. and Sackellares (1995). Time-Frequency Analysis of Electrophysiology Signals in Epilepsy. *IEEE Engineering in Medicine and Biology*. **March/April** 133-143.