AD-A084 142   TEXAS A AND M UNIV   COLLEGE STATION INST OF STATISTICS   F/G 12/1
MODERN EMPIRICAL STATISTICAL SPECTRAL ANALYSIS.(U)
MAY 80   E PARZEN   N00014-78-C-0599
UNCLASSIFIED   TR-N-12   NL

| OF |
AD
AD84142

END
DATE
FILMED
6-80
DTIC

TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS 77843

INSTITUTE OF STATISTICS
Phone 713 - 845-3141

(6)

Modern Empirical Statistical Spectral Analysis

by Emanuel Parzen

Texas A&M University

Technical Report No. N-12

May 1980

Texas A & M Research Foundation
Project No. 3838

"Multiple Time Series Modeling and Time
Series Theoretic Statistical Methods"

Sponsored by the Office of Naval Research

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

(14) ~~1. REPORT NUMBER~~
TR - N-12

2. GOVT ACCESSION NO.
AD-A084142

3. RECIPIENT'S CATALOG NUMBER

(6) 4. TITLE (and Subtitle)
Modern Empirical Statistical Spectral Analysis

5. TYPE OF ~~REPORT & PERIOD COVERED~~
(9) Technical rept.

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
(10) Emanuel Parzen

8. CONTRACT OR GRANT NUMBER(s)
(15) N00014-78-C-0599

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Texas A&M University
Institute of Statistics
College Station, TX 77843

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Code 436
Arlington, VA 22217

12. REPORT DATE
(11) May [redacted]80

13. NUMBER OF PAGES
26

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
(12) 28

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

Accession For
NTIS GRA&I ☑
DDC TAB □
Unannounced □
Justification________
By________
Distribution/
Availability Codes

Dist	Avail and/or special
A	

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Time series analysis, spectral analysis, minimum entropy distance spectral estimation, autoregressive spectral estimators, log spectral estimators, cepstral correlations, maximum entropy spectral estimation, empirical spectral analysis.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper has two aims: (1) to provide perspectives on the diverse paths of analysis which are available in 1980 to estimate the spectrum of an observed time series; and (2) to describe proposals for optimal statistical spectral estimation procedures which combine autoregressive spectral estimators and log spectral estimators. It is proposed that empirical statistical spectral analysis should be an adaptive procedure for forming an iterative spectral estimator (an iterative estimator is one composed of estimators obtained in different steps of the analysis). There are three parts: I. Basic concepts of time series spectral analysis; II. Entropy distances, autoregressive spectral estimators and log spectral estimators; III. An outline of empirical spectral analysis.

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)
Unclassified

347380

MODERN EMPIRICAL STATISTICAL SPECTRAL ANALYSIS

Emanuel Parzen

Institute of Statistics
Texas A & M University

INTRODUCTION*

This paper has two aims: (1) to provide perspectives on the diverse paths of analysis which are available in 1980 to estimate the spectrum of an observed time series; and (2) to describe proposals for optimal statistical spectral estimation procedures which combine autoregressive spectral estimators and log spectral estimators. It is proposed that empirical statistical spectral analysis should be an adaptive procedure for forming an iterative spectral estimator (an iterative estimator is one composed of estimators obtained in different steps of the analysis). There are three parts: I. Basic concepts of time series spectral analysis; II. Entropy distances, autoregressive spectral estimators and log spectral estimators; III. An outline of empirical spectral analysis .

I. BASIC CONCEPTS OF TIME SERIES SPECTRAL ANALYSIS

By spectral analysis of a time series Y(t) one means fitting to the time series a spectral representation of the form

$$Y(t) = \int e^{2\pi i\lambda t}\, d\Psi(\lambda) \ .$$

*This research was supported by the Office of Naval Research (Contract N00014-78-C-0599).

Spectral analysis has as its aim the determination of the properties of the function $\Psi(\lambda)$. Model identification is concerned with determining the qualitative properties of $\Psi(\lambda)$, and parameter estimation is concerned with determining the quantitative properties of $\Psi(\lambda)$. This chapter defines some basic concepts of spectral representations of time series.

1. DISCRETE PARAMETER AND CONTINUOUS PARAMETER TIME SERIES

The theory of time series analysis discusses separately discrete parameter time series $\{Y(t), \ t = 0, \pm 1, \ldots\}$ and continuous parameter time series $\{ Y(t), \ -\infty < t < \infty\}$. This paper discusses only discrete parameter time series. The range of the frequency variable λ is taken to be -0.5 to 0.5 in the discrete parameter case, and $-\infty$ to ∞ in the continuous parameter case. In many scientific fields, a discrete parameter series $Y(n)$ arises by observing a continuous parameter time series $Z(t)$ at equispaced times $t = nD$, so that $Y(n) = Z(nD)$. One calls D the sampling interval. We assume a spectral representation

$$Z(t) = \int_{-\infty}^{\infty} e^{2\pi it\omega} \psi_Z(\omega) \, d\omega \ .$$

Then

$$Y(n) = Z(nD) = \int_{-\infty}^{\infty} e^{2\pi inD\omega} \psi_Z(\omega) \, d\omega \ .$$

Let $\lambda = D\omega$. Then

$$Y(n) = \int_{-\infty}^{\infty} e^{2\pi in\lambda} \frac{1}{D} \psi_Z(\frac{\lambda}{D}) \, d\lambda \ .$$

Write the integral from $-\infty$ to ∞ as the sum of integrals over the intervals $k - 0.5$, $k + 0.5$ for $k = 0, \pm 1, \ldots$; the latter integral

$$\int_{k-0.5}^{k+0.5} e^{2\pi in\lambda} \frac{1}{D} \psi_Z(\frac{\lambda}{D}) \, d\lambda = \int_{-0.5}^{0.5} e^{2\pi in\lambda'} \frac{1}{D} \psi_Z(\frac{\lambda'+k}{D}) \, d\lambda' \ .$$

Samp!ing Theorem: $Y(n)$ has the spectral representation

$$Y(n) = \int_{-0.5}^{0.5} e^{2\pi in\lambda} \psi_Y(\lambda) \, d\lambda \ ,$$

$$\psi_Y(\lambda) = \frac{1}{D} \sum_{k=-\infty}^{\infty} \psi_Z\left(\frac{\lambda+k}{D}\right) .$$

Further, if $Z(t)$ is bandlimited, in the sense that $\psi_Z(\omega) = 0$ for $|\omega| > \frac{1}{2D}$, then

$$\psi_Y(\lambda) = \frac{1}{D} \psi_Z\left(\frac{\lambda}{D}\right) , \quad \psi_Z(\omega) = D \, \psi_Y(D\omega) .$$

Using these formulas one can rewrite the formulas one obtains for the spectrum of $Y(t)$, $t = 0, \pm 1, \ldots$ as formulas involving the spectrum of the sampled time series $Z(t)$, $-\infty < t < \infty$.

2. SOME TIME SERIES MODEL TYPES

Observed discrete parameter time series often may be regarded as sums of different types of functions.

Pure harmonics of period $p \geq 2$ are functions

$$Y(t) = A \cos \frac{2\pi}{p} t + B \sin \frac{2\pi}{p} t \quad ;$$

for which $\Psi(\lambda)$ is a function of bounded variation which changes only in jumps; $\Psi(\lambda+0) - \Psi(\lambda-0) = 0$ for all λ in $0 \leq \lambda \leq 0.5$ except $\lambda = \frac{1}{p}$.

Transients are square summable functions, $\sum_{t=-\infty}^{\infty} Y^2(t) < \infty$. Then $\Psi(\lambda)$ has a derivative $\psi(\lambda) = \Psi'(\lambda)$ satisfying

$$\psi(\lambda) = \sum_{t=-\infty}^{\infty} Y(t) \, e^{-2\pi i t\lambda} ,$$

$$Y(t) = \int_{-0.5}^{0.5} e^{2\pi i t\lambda} \, \psi(\lambda) \, d\lambda .$$

An important example of a transient time series $Y(t)$ is a spike which is non-zero at only one time t_0; then

$$\psi(\lambda) = Y(t_0) \, e^{-2\pi i t_0 \lambda}$$

and $|\psi(\lambda)|^2$ = constant for all λ .

The non-deterministic component of a time series is often assumed to be a *covariance stationary* time series Y(t) with zero mean and covariance *function* (in the notation of Parzen (1962))

$$R(v) = E[Y(t)\ Y(t+v)]\ ,\quad v = 0, \pm 1, \ldots\ .$$

The *correlation function* is

$$\rho(v) = \frac{R(v)}{R(0)} = \text{Correlation } [Y(t)\ ,\ Y(t+v)]\ .$$

We divide stationary time series into three types, (1) *white noise*, (2) *short memory*, or (3) *long memory*, whose definitions are given in the next section.

A pure harmonic of period p obeys the difference equation

$$Y(t) - \phi Y(t-1) + Y(t-2) = 0$$

where $\phi = 2 \cos \frac{2\pi}{p}$. Consequently if a time series Y(t) is the sum of harmonics and a stationary time series a useful way to identify a model for the time series Y(t) is to introduce a transformed time series

$$\tilde{Y}(t) = Y(t) - \phi Y(t-1) + Y(t-2)$$

and to model $\tilde{Y}(t)$ as a stationary time series. The final model fitted to the time series Y(t) is called an *iterated* model when it has the form

$$Y(t) \to \boxed{} \to \tilde{Y}(t) \to \boxed{} \to \varepsilon(t) \text{ white noise .}$$

To estimate the spectrum of a time series, one must identify the qualitative model types of which the time series is composed before one can estimate quantitatively their properties. It may be wisest to carry out *in parallel* several of the approaches to time series computations described in Chapter III. I express this point of view in a motto: "If one can think of two or more ways of solving the problem, one should solve it in two or more ways."

3. STATIONARY TIME SERIES MODEL TYPES

A stationary time series Y(t) has a spectral representation in terms of a stochastic integrand Ψ(λ) satisfying

$$E|d\Psi(\lambda)|^2 = R(0)\ f(\lambda)\ d\lambda = R(0)\ dF(\lambda)\ .$$

where $f(\lambda)$ and $F(\lambda)$ are spectral density and spectral distribution functions, respectively, whose definitions are given in this section.

White noise, or a no memory time series, is a time series of independent random variables; it satisfies $\sum_{v>0} |R(v)| = 0$.

To introduce the notion of a time series of short memory type, we consider a stationary time series Y(t) and assume that $\sum_{v=-\infty}^{\infty} |R(v)| < \infty$. We define the power spectrum of the time series to be

$$S(\lambda) = \sum_{v=-\infty}^{\infty} e^{-2\pi i\lambda v} R(v)\ ,\ -0.5 \leq \lambda \leq 0.5\ ;$$

it satisfies

$$R(v) = \int_{-0.5}^{0.5} e^{2\pi i v\lambda} S(\lambda)\ d\lambda\ ,\ v = 0, \pm 1, \ldots\ .$$

We define the spectral density of the time series by

$$f(\lambda) = \sum_{v=-\infty}^{\infty} e^{-2\pi i v\lambda} \rho(v)\ ,\ -0.5 \leq \lambda \leq 0.5\ ;$$

It provides a spectral representation of the correlation function,

$$\rho(v) = \int_{-0.5}^{0.5} e^{2\pi i v\lambda} f(\lambda)\ d\lambda\ ,\ v = 0, \pm 1, \ldots\ .$$

To define the spectrum of a stationary time series whose correlation function $\rho(v)$ is not summable define, for any T > 0,

$$f_T(\lambda) = \frac{1}{T} \sum_{j,k=1}^{T} e^{-2\pi i\lambda j} e^{2\pi i\lambda k} \rho(j-k)\ ;$$

it is a non-negative function by the non-negative definite property of $\rho(v)$. One can write

$$f_T(\lambda) = \sum_{|v|<T} e^{-2\pi i\lambda v} (1-\frac{|v|}{T})\rho(v) ,$$

$$(1-\frac{|v|}{T})\rho(v) = \int_{-0.5}^{0.5} e^{2\pi i\lambda v} f_T(\lambda)\, d\lambda .$$

When $\rho(v)$ is summable, $f_T(\lambda) \longrightarrow f(\lambda) \geq 0$. Otherwise, $\rho(v)$ is the limit (as $T \longrightarrow \infty$) of Fourier transforms of non-negative functions, and therefore there exists a *spectral distribution* function $F(\lambda)$, $-0.5 \leq \lambda \leq 0.5$ such that

$$\rho(v) = \int_{-0.5}^{0.5} e^{2\pi i\lambda v}\, dF(\lambda) ,$$

and $$F_T(\lambda) = \int_{-0.5}^{\lambda} f_T(u)\, du \longrightarrow F(\lambda) .$$

An important diagnostic tool of the type of a stationary time series is its *spectral log range*, defined by

$$\mathrm{SPLR} = \lim_{T\to\infty} \log \max_{\lambda} f_T(\lambda) - \log \min_{\lambda} f_T(\lambda) .$$

The *memory type* of a stationary time series is classified according to the behavior of its spectral log range:

NO MEMORY	SHORT MEMORY	LONG MEMORY
SPLR = 0	$0 < \mathrm{SPLR} < \infty$	$\mathrm{SPLR} = \infty$

A stationary time series has short memory if $\sum_{v=-\infty}^{\infty} |\rho(v)| < \infty$ and the spectral density $f(\lambda) \neq 0$ for any λ ; then there exist positive constants C_1 and C_2 such that $0 < C_1 \leq f(\lambda) \leq C_2 < \infty$ for all λ. For a short memory series, $f(\lambda)$, $f^{-1}(\lambda)$, and $\log f(\lambda)$ are all integrable over the interval $-0.5 \leq \lambda \leq 0.5$.

4. STATIONARY FILTER THEOREM

The interpretation of the power spectrum comes from the following important theorem.

Filter Theorem. If $Y(\cdot)$ is stationary with spectral density

$f_Y(\lambda)$, and

$$Z(t) = \sum_{s=-\infty}^{\infty} b(t-s)\ Y(s) = \sum_{s=-\infty}^{\infty} b(s)\ Y(t-s)$$

where $\sum_{s=-\infty}^{\infty} b^2(s) < \infty$, $B(\lambda) = \sum_{s=-\infty}^{\infty} b(s)\ e^{-2\pi i\lambda s}$

then $Z(\cdot)$ is stationary with spectral density and covariance function given by

$$f_Z(\lambda) = f_Y(\lambda)\ |B(\lambda)|^2 \frac{R_Y(0)}{R_Z(0)} ,$$

$$R_Z(v) = \sum_{s=-\infty}^{\infty} R_b(s)\ R_Y(v+s)$$

defining $R_b(v) = \sum_{k=-\infty}^{\infty} b(k)\ b(k+v)$.

5. WHITENING FILTERS

Another major aim of time series analysis is to obtain whitening filter representations of $Y(t)$, $t = 0, \pm 1, \ldots$ of the form

$$\sum_{j=0}^{p} a_j Y(t-j) = \sum_{k=0}^{q} b_k(t-k)$$

where $\{\eta(t) , t = 0, \pm 1, \ldots\}$ is a time series of "simple" structure; in particular $\eta(t)$ might be white noise or a series of impulses. Whitening filter analysis has its aim the determination of the parameters p,q, $a_0, a_1, \ldots, a_p, b_0, b_1, \ldots, b_q$, and series $\eta(t)$, especially its spectral representation

$$\eta(t) = \int_{-0.5}^{0.5} e^{2\pi it\lambda}\ d\Psi_\eta(\lambda) .$$

The whitening filter is called: an autoregressive, or AR, filter if $q = 0$; a moving average, or MA, filter if $p = 0$; and an autoregressive moving average, or ARMA, filter if p and q are both non-zero. The most frequently used filters are AR filters.

From a whitening filter representation of Y(t) one may infer properties of its spectral representation; define

$$g_p(e^{2\pi i\lambda}) = \sum_{j=0}^{p} a_j e^{-2\pi ij\lambda} , \quad h_q(e^{2\pi i\lambda}) = \sum_{k=0}^{q} b_k e^{-2\pi ik\lambda} ,$$

called respectively the AR and MA transfer function. Then

$$\int_{-0.5}^{0.5} e^{2\pi it\lambda} g_p(e^{2\pi i\lambda}) d\Psi_Y(\lambda) = \int_{-0.5}^{0.5} e^{2\pi it\lambda} h_q(e^{2\pi i\lambda}) d\Psi_\eta(\lambda) .$$

Consequently (for all λ_0)

$$\int_{-0.5}^{\lambda_0} g_p(e^{2\pi i\lambda}) d\Psi_Y(\lambda) = \int_{-0.5}^{\lambda_0} h_q(e^{2\pi i\lambda}) d\Psi_\eta(\lambda) .$$

Knowing g_p, h_q, and Ψ_η, one can solve for Ψ_Y .

When $\eta(\cdot)$ is a stationary time series we define the spectral density of $Y(\cdot)$ by the filter theorem:

$$f_Y(\lambda) = f_\eta(\lambda) \frac{|h_q(e^{2\pi i\lambda})|^2}{|g_p(e^{2\pi i\lambda})|^2} \bar{\sigma}_\eta^2$$

where $\bar{\sigma}_\eta^2$ is a "measure" of $R_\eta(0)/R_Y(0)$, such as

$$\bar{\sigma}_\eta^2 = \frac{\Sigma\eta^2(t)}{\Sigma Y^2(t)}$$

The whitening filter is written symbolically in terms of the lag operator L defined by LY(t) = Y(t-1) . Then

$$g_p(L)Y(t) = h_q(L)\,\eta(t) , \quad \eta(t) = \frac{g_p(L)}{h_q(L)} Y(t) .$$

6. BASIC SAMPLE STATISTICS

To form estimators of parameters, such as R(v) and f(λ), we can either seek estimators which are optimal according to an estimation criterion such as maximum likelihood or we can form estimators which seem "natural" and determine their asymptotic optimality properties. A natural estimator of $R(v) = E[Y(t)Y(t+v)]$ from a sample $\{Y(t), t = 1, \ldots, T\}$ is

$$\hat{R}(v) = \frac{1}{T} \sum_{t=1}^{T-v} Y(t)\, Y(t+v)$$

called the sample covariance function. Note that we divide by T rather than by T-v in order to obtain a function $\hat{R}(v)$ which is positive-definite

$$\sum_{j,k=1}^{n} c_j c_k \hat{R}(j-k) \geq 0 \text{ for all } n, c_1, \ldots, c_n .$$

Then $\rho(v)$ is estimated by

$$\hat{\rho}(v) = \frac{\hat{R}(v)}{\hat{R}(0)} = \frac{\sum_{t=1}^{T-v} Y(t)\, Y(t+v)}{\sum_{t=1}^{T} Y^2(t)} ,$$

called the sample correlation function. These functions possess spectral representations

$$\hat{R}(v) = \int_{-0.5}^{0.5} e^{2\pi i\lambda v}\, \tilde{S}(\lambda)\, d\lambda ,$$

$$\hat{\rho}(v) = \int_{-0.5}^{0.5} e^{2\pi i\lambda v}\, \tilde{f}(\lambda)\, d\lambda ,$$

in terms of

$$\tilde{S}(\lambda) = \frac{1}{T} \left| \sum_{t=1}^{T} Y(t)\, e^{-2\pi i t\lambda} \right|^2 ,$$

$$\tilde{f}(\lambda) = \frac{\left| \sum_{t=1}^{T} Y(t)\, e^{-2\pi i t\lambda} \right|^2}{\sum_{t=1}^{T} Y^2(t)}$$

It should be noted that these functions provide a generalized harmonic analysis of $Y(\cdot)$ in the sense of Wiener (1930).

We call $\tilde{S}(\lambda)$ the sample power spectrum and $\tilde{f}(\cdot)$ the sample spectral density. They are natural estimators of $S(\lambda)$ and $f(\lambda)$ respectively, but they are very wiggly functions and lack most of the properties of optimal estimators. Thus arises the need for a sophisticated theory of statistical spectral analysis.

One reason for using $\hat{\rho}(v)$ and $\hat{f}(\lambda)$ as basic diagnostic statistics for observed time series is that they possess fast computation algorithms, using the Fast Fourier transform. Given a sample $\{Y(t), t = 1, \ldots, T\}$ one proceeds as follows.

A. Pre-processing. To analyze a time series sample $Y(t)$, $t = 1, \ldots, T$, one will proceed in stages which often involve the subtraction of or elimination of strong effects in order to see more clearly weaker patterns in the time series structure.

The aim of pre-processing is to transform $Y(\cdot)$ to a new time series $\tilde{Y}(\cdot)$ which is short memory (a zero mean stationary time series whose spectral density has finite log range). The basic pre-processing operations are memory less transformation (such as square root and logarithm), detrending, "high pass" filtering, and differencing. One usually subtracts out the sample mean $\bar{Y} = \frac{1}{T} \sum_{t=1}^{T} Y(t)$; then the time series actually processed is $Y(t) - \bar{Y}$. If the mean $\bar{Y}$ is a large number, it should be subtracted; the variations in $Y(t)$ are then the variations of $Y(t)$ about its mean. The sample mean $\bar{Y}$ and sample variance $R(0)$ should always be recorded.

B. Sample Fourier Transform by Data Windowing, Extending with Zeroes, and Fast Fourier Transform. The first step in a comprehensive analysis of a pre-processed time series sample should always be the computation of the sample Fourier transform

$$\tilde{\psi}(\lambda) = \sum_{t=1}^{T} Y(t) \exp(-2\pi i\lambda t)$$

at an equi-spaced grid of frequencies in $0 \leq \lambda \leq 1$, of the form $\lambda = \frac{k}{Q}$, $k = 0, \ldots, Q-1$. We call Q the spectral computation number. One should always choose $Q \geq T$, and we recommend $Q \geq 2T$.

Prior to computing $\tilde{\psi}(\lambda)$, one should extend the length of the time series by adding zeroes to it. Then $\tilde{\psi}(\lambda)$, $\lambda = \frac{k}{Q}$, can be computed using the Fast Fourier transform.

In addition, one should compute a sample "data windowed" Fourier transform

$$\tilde{\psi}_W(\lambda) = \sum_{t=1}^{T} Y(t)W(\tfrac{t}{T}) \exp(-2\pi i\lambda t) \ .$$

To understand the effect of the window, one replaces Y(t) by its spectral representation

$$Y(t) = \int_{-0.5}^{0.5} \exp\ (2\pi i\lambda' t)\ d\Psi(\lambda') \ ;$$

then $$\tilde{\psi}_W(\lambda) = \int_{-0.5}^{0.5} w_T(\lambda-\lambda')\ d\Psi(\lambda')$$

where $$w_T(\lambda) = \sum_{t=1}^{T} W(\tfrac{t}{T}) \exp(-2\pi i\lambda t) \ .$$

Considerations involved in the choice of data windows are discussed in Harris (1978) .

C. Sample Spectral Density. The sample spectral density $\tilde{f}(\lambda)$ is obtained essentially by squaring and normalizing the sample Fourier transform;

$$\tilde{f}(\lambda) = \frac{|\tilde{\psi}(\lambda)|^2}{\frac{1}{Q}\sum_{k=0}^{Q-1} |\tilde{\psi}(\frac{k}{Q})|^2} \ , \quad \lambda = \frac{k}{Q} \ , \ k = 0,\ 1,\ \ldots,\ Q-1 \ .$$

It is a function with period 1, whose domain is taken to be $-0.5 \leq \lambda \leq 0.5$ (or $0 \leq \lambda \leq 1$), which integrates to 1 and provides a spectral representation of $\tilde{\rho}(v)$.

D. Sample Correlation Function. The sample correlation function $\hat{\rho}(v)$ is computed (using the Fast Fourier Transform) by

$$\hat{\rho}(v) = \frac{1}{Q}\sum_{k=0}^{Q-1} \exp(2\pi i\tfrac{k}{Q}v)\ \tilde{f}(\tfrac{k}{Q}) \ ,$$

which holds for $0 \leq v < Q-T$.

E. Sample Spectral Distribution Function.

$$\tilde{F}(\lambda) = 2 \int_0^\lambda \tilde{f}(\lambda')\, d\lambda' , \quad 0 \le \lambda \le 0.5 ;$$

the graph of $\tilde{F}(\lambda)$ provides qualitative diagnostics of the time series model type.

The foregoing basic statistics are the building blocks of the smooth spectral estimators whose theory is discussed in the rest of this paper.

II. ENTROPY DISTANCES, AUTOREGRESSIVE SPECTRAL ESTIMATORS, AND LOG SPECTRAL ESTIMATORS

The theory of statistical spectral analysis in 1980 should be based, in my opinion, on the role in statistical inference of entropy and information numbers. The credit for emphasizing this perspective should be given to the two pioneering developments of MEM (maximum entropy method) of Burg (1967) and AIC (information criterion) of Akaike (1974).

Given a sample $Y(t)$, $t = 1, 2, \ldots, T$ of a discrete parameter time series $Y(t)$, $t = 0, \pm 1, \ldots$, the general problem of statistical inference is to infer the probability distribution of the observed random variables. A probability model whose goodness of fit to the data is an ever-present hypothesis is that $Y(t)$, $t = 0, \pm 1, \ldots$ is a zero mean Gaussian stationary time series with covariance function $R(v) = E[Y(t)\ Y(t+v)]$, $v = 0, \pm , \ldots$, and correlation function $\rho(v) = R(v)/R(0)$. When discussing statistical inference, it is usual to assume that the process is <u>ergodic</u> which requires us to make an assumption such as $R(v)$ is absolutely summable: $\Sigma|R(v)| < \infty$. The power spectrum $S(\lambda)$ and spectral density function $f(\lambda)$ are defined (in Section I.3) as the Fourier transforms of $R(v)$ and $\rho(v)$ respectively.

1. APPROXIMATE LIKELIHOOD FUNCTION OF STATIONARY GAUSSIAN TIME SERIES

One approach to forming optimal estimators of statistical parameters is to obtain a formula for the likelihood or joint probability density function of $Y(1), \ldots, Y(T)$, which we denote by $f_\theta\ (Y(1), \ldots, Y(T))$; the subscript θ indicates that it is a function of the unknown parameters θ , log is natural logarithm, * is complex conjugate transpose. Then

$$-2\log f_\theta(Y(1), \ldots, Y(T)) = \log\{(2\pi)^T \det K_\theta\} + Y_T^* K_\theta^{-1} Y_T$$

where $Y_T^* = (Y(1), \ldots, Y(T))$ and $K_\theta = EY_T Y_T^*$ is a covariance matrix with (s,t) - element equal to $R_\theta(s-t)$. The subscript θ on $R_\theta(v)$, $\rho_\theta(v)$, $S_\theta(\lambda)$, and $f_\theta(\lambda)$, indicate that they are functions of unknown parameters θ (which are to be estimated).

The covariance matrix K is a Toeplitz matrix; asymptotically, as T tends to ∞, all T by T Toeplitz matrices have the same eigenvectors $\exp(-2\pi it\ j/T)$, $j = 0, 1, \ldots, T-1$. The eigenvalues of K_θ are $S_\theta(j/T)$.

We prefer to express the likelihood in terms of $f_\theta(j/T)$. Therefore, we assume that the time series Y(t) has been divided by $\{R(0)\}^{\frac{1}{2}}$ so that it can be considered to have variance 1, and its covariance function equals its correlation function. Then one can show that approximately, for large values of T,

$$-\frac{2}{T} \log f_\theta(Y(1),\ldots,Y(T)) = \log 2\pi + \int_{-0.5}^{0.5} \{\log f_\theta(\lambda) + \frac{\tilde{f}(\lambda)}{f_\theta(\lambda)}\} d\lambda$$

$$= \log 2\pi + H(\tilde{f};f_\theta)$$

where
$$\tilde{f}(\lambda) = \left|\sum_{t=1}^{T} Y(t) \exp -2\pi it\lambda\right|^2 \div \sum_{t=1}^{T} Y^2(t)$$

is the sample spectral density, and the entropy number H is defined by

$$H(f;g) = \int_{-0.5}^{0.5} \{\log g(\lambda) + \frac{f(\lambda)}{g(\lambda)}\}\ d\lambda \quad .$$

2. MINIMUM ENTROPY DISTANCE ESTIMATION

The maximum likelihood estimator $\hat{\theta}$ is equivalent to the estimator $\hat{\theta}$ minimizing over θ

$$H(\tilde{f};f_\theta) = \int_{-0.5}^{0.5} \{\log f_\theta(\lambda) + \frac{\tilde{f}(\lambda)}{f_\theta(\lambda)}\}\ d\lambda \quad .$$

In order to regard $H(\tilde{f};f_{\hat{\theta}})$ as a measure of "distance" or "fit" between the data (with representing function $\tilde{f}(\lambda)$) and the model (with representing function $f_{\hat{\theta}}(\lambda)$), we define the entropy distance

$$I(f;g) = \int_{-0.5}^{0.5} \{\frac{f(\lambda)}{g(\lambda)} - \log\frac{f(\lambda)}{g(\lambda)} - 1\}d\lambda = H(f;g) - H(f;f)$$

Since $u - \log u - 1 \geq 0$ for all u, I has two of the properties of a distance, namely $I(f;g) \geq 0$, $I(f;f) = 0$. However I does not satisfy the triangle inequality. Since

$$I(\tilde{f};f_\theta) = H(\tilde{f};f_\theta) - H(\tilde{f};\tilde{f}) ,$$

minimizing $H(\tilde{f};f_\theta)$ with respect to θ is equivalent to minimizing $I(\tilde{f};f_\theta)$. Minimum entropy distance estimators $\hat{\theta}$ are shown to be consistent (as the sample size T tends to infinity) by showing that the sequence $I(f;f_{\hat{\theta}})$ converges to zero, where f is the true spectral density function. If $f = f_{\theta_0}$ for some θ_0 , then one can infer that the sequence $\hat{\theta}$ converges to θ_0.

3. L_2 DISTANCES BETWEEN SPECTRAL DENSITIES

One can relate entropy distance to the L_2 log spectral density distance

$$L_2L(f,g) = \int_{-0.5}^{0.5} \{\log f(\lambda) - \log g(\lambda)\}^2 \, d\lambda \, .$$

Since $u = \exp(\log u) = 1 + \log u + 1/2 (\log u)^2$, for "neighboring" f and g, $I(f,g) = L(f,g)/2$ and minimizing $I(\tilde{f};f_\theta)$ could be regarded as asymptotically equivalent to minimizing $L_2L(\tilde{f};f_\theta)$. An extensive discussion of these distances is given by Gray, Buzo, Gray, and Matsuyama (1980).

The notation L_2L is chosen to emphasize the distinction between that distance and the L_2 spectral density distance

$$L_2(\tilde{f},f_\theta) = \int_{-0.5}^{0.5} \{\tilde{f}(\lambda) - f_\theta(\lambda)\}^2 \, d\lambda \, .$$

This distance has been used for spectral estimation but it seems not to be justifiable in general.

However in the case of smoothing prewhitened sample spectral densities, when $f_\theta(\lambda)$ may be expected to have a small log-range, $L_2(\tilde{f},f_\theta)$ may be a justifiable distance. It then may approximate

$$\int_{-0.5}^{0.5} \left\{\frac{f(\lambda) - f_\theta(\lambda)}{f_\theta(\lambda)}\right\}^2 d\lambda$$

which is also a useful "distance".

4. MINIMUM DISTANCE FORMULATION OF OPTIMAL ESTIMATION

In summary, one approach to forming "optimal" estimators $\hat{f}(\lambda)$ of the spectral density $f(\lambda)$ of a stationary time series is to view $\hat{f}(\lambda)$ as a function closest to $\tilde{f}(\lambda)$ in a "distance" between spectral density functions, such as

$$H(\tilde{f};\hat{f}) = \int_{-0.5}^{0.5} \left\{ \log \hat{f}(\lambda) + \frac{\tilde{f}(\lambda)}{\hat{f}(\lambda)} \right\} d\lambda$$

$$I(\tilde{f};\hat{f}) = \int_{-0.5}^{0.5} \left\{ \frac{\tilde{f}(\lambda)}{\hat{f}(\lambda)} - \log \frac{\tilde{f}(\lambda)}{\hat{f}(\lambda)} - 1 \right\} d\lambda = H(\tilde{f};\hat{f}) - H(\tilde{f};\tilde{f})$$

$$L_2L(\tilde{f},\hat{f}) = \int_{-0.5}^{0.5} \{\log \tilde{f}(\lambda) - \log \hat{f}(\lambda)\}^2 \, d\lambda \; .$$

The class of functions from which $\hat{f}(\lambda)$ is chosen can be specified or constrained either parametrically or non-parametrically. A parametric constraint is to choose $\hat{f}(\lambda)$ from a family of functions $f_\theta(\lambda)$ indexed by a finite number of parameters θ. A non-parametric constraint is to impose a smoothness measure on $\hat{f}$ such as the square integral of second derivatives:

$$\int_{-0.5}^{0.5} |f''(\lambda)|^2 \, d\lambda \quad \text{or} \quad \int_{-0.5}^{0.5} |(\log \hat{f}(\lambda))''|^2 d\lambda \; .$$

One then seeks to choose $\hat{f}$ to maximize smoothness while minimizing a measure of distance of $\hat{f}$ from $\tilde{f}$.

Nonparametric approaches to spectral estimation may work best for estimation of the log spectral density using an approach introduced by Wahba (1980). Motivated by the estimation distance

$$\int_{-0.5}^{0.5} \{\log \tilde{f}(\lambda) - \log \hat{f}(\lambda)\}^2 d\lambda + K \int_{-0.5}^{0.5} |(\log \hat{f}(\lambda))''|^2 \, d\lambda$$

where K is a penalty parameter to be determined adaptively by the data, she considers estimators of the form

$$\log \hat{f}(\lambda) = \sum_{v=-\infty}^{\infty} w\left(\frac{v}{M}\right) \tilde{\gamma}(v) \exp(-2\pi i v \lambda)$$

where $\tilde{\gamma}(v)$, which I call cepstral-correlations, are defined by

$$\tilde{\gamma}(v) = \int_{-0.5}^{0.5} \log \tilde{f}(\lambda) \exp (2\pi i v \lambda) \, d\lambda \; ,$$

and the weights $w(v)$ are of the form

$$w(v) = \frac{1}{1+v^{2r}} \, , \quad r = 2 \text{ or } 4 \; .$$

We call M the "half-power" lag. In Section 7 we discuss how one might choose M and r to minimize an estimator of

$$J(f,\hat{f}) = E \int_{-0.5}^{0.5} \{\log f(\lambda) - \log \hat{f}(\lambda)\}^2 \, d\lambda ,$$

assuming log $f(\lambda)$ has finite range and therefore has a representation

$$\log f(\lambda) = \sum_{v=-\infty}^{\infty} \gamma(v) \exp(-2\pi i v \lambda) .$$

5. PARAMETRIC SPECTRAL ESTIMATORS, BIAS, AND VARIANCE

A spectral density estimator is called parametric if it is based on a representation of the spectral density as a function of m parameters $\theta_1, \ldots, \theta_m$, which we denote $f_{\theta_1, \ldots, \theta_m}(\lambda)$. We call m the order, and it is also often a parameter to be estimated. The problem of model identification, or model approximation, is to estimate m, and also to estimate which parameters $\theta_1, \ldots, \theta_m$ are "significantly" different from zero.

The true spectral density is denoted by f or f_∞ . A best approximation $\bar{f} = f_{\bar{\theta}_1, \ldots, \bar{\theta}_m}$ can be determined for each order m *where* $\bar{\theta}_1, \ldots, \bar{\theta}_m$ *minimizes* $H(f;f_{\theta_1, \ldots, \theta_m})$. An estimator of f is $\hat{f} = f_{\hat{\theta}_1, \ldots, \hat{\theta}_m}$ where $\hat{\theta}_1, \ldots, \hat{\theta}_m$ minimizes $H(\tilde{f};f_{\theta_1,\ldots,\theta_m})$. The optimal estimator $\hat{f}$ minimizes $R(\hat{f}) = EI(f_\infty;\hat{f})$. When using approximating parametric densities the criterion $R(\hat{f})$ is replaced by an order determining criterion C(m) to determine the order $\hat{m}$ of the parametric density. One can write

$$C(m) = B(m) + V(m,T) ,$$

where $B(m) = I(f_\infty;\bar{f})$, $V(m,T) = EI(\bar{f};\hat{f})$. We call B(m) the model approximation error (or bias) and V(m,T) the parameter estimation error (or variance). As $m \to \infty$, $B(m) \to 0$ and $V(m,T) \to \infty$. Consequently C(m) has a minimum.

AIC introduced by Akaike (1974) may be regarded as corresponding to

$$B(m) = H(\tilde{f};f_{\hat{\theta}_1, \ldots, \hat{\theta}_m}) - H(\tilde{f};\tilde{f}) = \log \hat{\sigma}_m^2 - \log \tilde{\sigma}_\infty^2$$

$$V(m,T) = 2m/T$$

Other order determining criteria may be regarded as corresponding to different formulas for V(m,T):

$V(m,T) = (m/T) \log \log T$, Hannan and Quinn (1979);

$V(m,T) = (m/T) \log T$, Schwarz (1978) .

CAT(criterion autoregressive transfer function) is an order determining criterion for autoregressive spectral estimators introduced by Parzen (1974), (1977); one version is

$$CAT(m) = \frac{1}{T} \sum_{j=1}^{m} \hat{\hat{\sigma}}_j^{-2} - \hat{\hat{\sigma}}_m^{-2} , \quad \hat{\hat{\sigma}}_j^{-2} = (1-\frac{j}{T}) \hat{\sigma}_j^{-2} .$$

We would like to emphasize that it is also of the general form $B(m) + V(m,T)$, defining

$$B(m) = -\sigma_m^{-2} + \sigma_\infty^{-2} , \quad V(m,T) = \frac{1}{T} \sum_{j=1}^{m} \sigma_j^{-2} .$$

6. AUTOREGRESSIVE SPECTRAL ESTIMATORS

The most convenient parametric estimators are *autoregressive* spectral estimators of the form

$$f_{\theta,m}(\lambda) = \sigma^2 |1 + \alpha_1 e^{-2\pi i\lambda} + \ldots + \alpha_m e^{-2\pi i\lambda m}|^{-2} .$$

The parameters are σ^2, σ_1, ..., σ_m as well as the order m. The subscript θ,m is merely symbolic to indicate that $f(\lambda)$ is a function of m parameters (in addition to σ^2) .
Estimators of these parameters can be found by solving "normal equations"

$$\sum_{k=0}^{m} \hat{\alpha}_k \hat{K}(j,k) = 0 , \quad j = 1, \ldots, m ;$$

$$\sum_{k=0}^{m} \hat{\alpha}_k \hat{K}(0,k) = \hat{\alpha}_m .$$

where $\hat{K}(j,k)$ is an estimator of $K(j,k) = E[Y(t-j)\, Y(t-k)]$. The normal equations are called *stationary* if $\hat{K}(j,k)$ is chosen to be a function of $(j-k)$.

Stationary estimators α_1, ..., α_m may be found by minimizing

$$\tilde{J}(\alpha_1,\ldots,\alpha_m) = \int_{-0.5}^{0.5} |1 + \alpha_1 e^{-2\pi i\lambda} + \ldots + \alpha_m e^{-2\pi i m\lambda}|^2 \tilde{f}(\lambda) d\lambda ,$$

since $H(\tilde{f};f_\theta) = \log \sigma^2 + \frac{1}{\sigma^2} \tilde{J}(\alpha_1, \ldots, \alpha_m)$.

We have used the important fact that $\int_0^1 \log|1 + \alpha_1 e^{-2\pi i\lambda} + \ldots + \alpha_m e^{-2\pi i m\lambda}|^2 d\lambda = 0$ under the assumption that the characteristic polynomial $g_m(z) = 1 + \alpha_1 z + \ldots + \alpha_m z^m$ has all its roots in the complex z-plane outside the unit circle.

Differentiating $H(\tilde{f};f_\theta)$ with respect to σ^2 one obtains $\hat{\sigma}^2 = \tilde{J}(\hat{\alpha}_1, \ldots, \hat{\alpha}_m)$.

The problem of minimizing $J(\alpha_1, \ldots, \alpha_m)$ can be viewed as a problem of projection in the Hilbert space of functions on the unit circle with the *inner product*

$$(g_1,g_2)_{\tilde{f}} = \int_{-0.5}^{0.5} g_1(e^{2\pi i\lambda}) \{g_2(e^{2\pi i\lambda})\}^* \tilde{f}(\lambda)\, d\lambda .$$

$J(\hat{\alpha}_1, \ldots, \hat{\alpha}_m)$ is the norm squared of the best approximation of 1 by a linear combination of $e^{2\pi i\lambda}$, ..., $e^{2\pi i m\lambda}$. The coefficients

$\hat{\alpha}_1, \ldots, \hat{\alpha}_m$ are determined by the condition that $\hat{g}_m(z) = 1 + \hat{\alpha}_1 z + \ldots + \hat{\alpha}_m z^m$, $z = e^{2\pi i\lambda}$, is orthogonal to z^j, $j = 1, \ldots, m$. Thus

$$0 = \int_{-0.5}^{0.5} \hat{g}_m(e^{2\pi i\lambda})\, e^{-2\pi i\lambda j}\, \tilde{f}(\lambda)\, d\lambda$$

$$= \sum_{k=0}^{m} \hat{\alpha}_k \hat{\rho}(k-j), \quad j = 1, \ldots, m, \text{ where } \hat{\alpha}_0 = 1 .$$

These are the celebrated sample Yule-Walker equations, or Toeplitz normal equations for the autoregressive coefficients. The estimator of σ^2 is $\hat{\sigma}_m^2$, called the residual variance or prediction error variance, given by

$$\hat{\sigma}_m^2 = (1,\hat{g}_m)_{\tilde{f}} = \sum_{k=0}^{m} \hat{\alpha}_k\, \hat{\rho}(k) .$$

It cannot be too strongly emphasized that there are several ways to form estimators of parameters to form an autoregressive spectral density

$$\hat{f}_m(\lambda) = \hat{\sigma}_m^2 |1 + \hat{\alpha}_1\, e^{2\pi i\lambda} + \ldots + \hat{\alpha}_m e^{2\pi i\lambda m}|^{-2}.$$

Various approaches are outlined in section 9. When the coefficients are computed by the Yule-Walker equations $\hat{f}_m$ is called the Yule-Walker autoregressive spectral estimator, and it satisfies

$$H(\tilde{f};\hat{f}_m) = \int_{-0.5}^{0.5} \{\log \hat{f}_m(\lambda) + \frac{\tilde{f}(\lambda)}{\hat{f}_m(\lambda)}\}\, d\lambda = \log \hat{\sigma}_m^2 ,$$

since $\int_{-0.5}^{0.5} \log \hat{f}_m(\lambda)\, d\lambda = \log \hat{\sigma}_m^2$,

$$\int_{-0.5}^{0.5} \frac{\tilde{f}(\lambda)}{\hat{f}_m(\lambda)}\, d\lambda = \hat{\sigma}_m^2 ||\hat{g}_m||_{\tilde{f}}^2 = \frac{1}{\hat{\sigma}_m^2} (1,\hat{g}_m)_{\tilde{f}} = 1$$

Akaike's AIC (to be minimized to determine significant orders m) is

$$AIC(m) = B(m) + V(m,T) = H(\tilde{f};\hat{f}_m) + \frac{2m}{T} = \log \hat{\sigma}_m^2 + \frac{2m}{T} .$$

An important consequence of our derivation is that one can evaluate a similar criterion for other ways of computing an autoregressive spectral estimator $\hat{f}_m(\lambda)$;

$$\int_{-0.5}^{0.5} \frac{\tilde{f}(\lambda)}{\hat{f}_m(\lambda)}\, d\lambda = \frac{\sigma_m^{*2}}{\hat{\sigma}_m^2} \quad (\geq 1 \text{ , usually})$$

defining $\sigma_m^{*2} = \int_{-0.5}^{0.5} |1 + \hat{\alpha}_1 e^{2\pi i\lambda} + \ldots + \hat{\alpha}_m e^{2\pi i\lambda m}|^2 \tilde{f}(\lambda)\, d\lambda$.

Consequently an order determining criterion could be

$$H(\tilde{f};\hat{f}_m) + \frac{2m}{T} = \frac{\sigma_m^{*2}}{\hat{\sigma}_m^2} + \log \hat{\sigma}_m^2 + \frac{2m}{T} \ ,$$

if the criterion one uses for the match of model to data is one based on the spectral matching of $\tilde{f}$ to $\hat{f}$ (although $\tilde{f}$ is estimated using non-stationary autoregressive models).

7. LOG SPECTRAL SMOOTHING AND CEPSTRAL CORRELATIONS

The problem of estimation of the spectral density of a time series $Y(\cdot)$ can be regarded *in theory* as determining a smooth function $\hat{f}_Y(\lambda)$ which optimally fits a sample spectral density $\tilde{f}_Y(\lambda)$. (Note that to compute $\tilde{f}_Y(\lambda)$ one may have used a data window). We believe that the best fit is often obtained by an *iterated spectral estimator* which uses an autoregressive estimator to match the large scale excursions of $\tilde{f}_Y(\lambda)$, and then uses log spectral smoothing to match the smaller excursions. The autoregressive filter often has the effect of reducing the log-range of the spectrum, without following fine structure which is present. The fine structure which is left in the residual process is estimated by the *log spectral smoothing estimator*.

For long memory time series, the iterated spectral estimator combines (1) an order 1 or 2 autoregression to transform to a short memory time series, (2) an autoregression to prewhiten, (3) log-spectral smoothing.

Autoregressive spectral estimation phase. Using an order determining criterion, and either stationary or non-stationary estimators of coefficients, one determines an autoregressive filter $\hat{g}_m(L)$, autoregressive residual variance $\hat{\sigma}_m^2$, and autoregressive spectral density estimator

$$\hat{f}_m(\lambda) = \hat{\sigma}_m^2 \ |\hat{g}_m(e^{2\pi i\lambda})|^{-2} \ .$$

The residual time series $\tilde{Y}(t)$ is defined by

$$\tilde{Y}(t) = \hat{g}_m(L) \ Y(t) \ .$$

Autoregressive spectral estimators are superior to other spectral estimators when the length of the observed segment of a time series is short compared to the (long) memory of the correlation function of the time series.

If Y were regarded as white noise, one would regard $\hat{f}_m(\lambda)$ as the estimated spectral density of the time series. To compensate for the fact that $\tilde{Y}$ may not be white noise, and to ease the burden of requiring $\tilde{Y}(t)$ to be white noise, we estimate its spectral density.

Residual log spectral estimation phase. Between the sample spectral densities of $\tilde{Y}(t)$ and $Y(t)$ there exists a basic relation:

$$\tilde{f}_{\tilde{Y}}(\lambda) = \sigma_m^{*-2} |\hat{g}_m(e^{2\pi i\lambda})|^2 \ \tilde{f}_Y(\lambda)$$

where $$\sigma_m^{*2} = \int_{-0.5}^{0.5} |\hat{g}_m(e^{2\pi i\lambda})|^2\, \tilde{f}_Y(\lambda)\, d\lambda .$$

This relation can be written:

$$\tilde{f}_{\tilde{Y}}(\lambda) = \frac{\hat{\sigma}_m^2}{\sigma_m^{*2}} \frac{\tilde{f}_Y(\lambda)}{\hat{f}_m(\lambda)}$$

$$\log \tilde{f}_{\tilde{Y}}(\lambda) = \log \frac{\hat{\sigma}_m^2}{\sigma_m^{*2}} - \log \hat{f}_m(\lambda) + \log \tilde{f}_Y(\lambda) .$$

Assuming that $\tilde{f}_{\tilde{Y}}(\lambda)$ has been "prewhitened" in the sense of having moderate log range, we smooth $\log \tilde{f}_{\tilde{Y}}(\lambda)$ to form an estimator $\{\log f_{\tilde{Y}}(\lambda)\}^{\wedge}$. Then as a final estimator of the true log spectral density $f(\lambda)$ we take, up to a normalizing constant,

$$\{\log f_Y(\lambda)\}^{\wedge} = \{\log f_{\tilde{Y}}(\lambda)\}^{\wedge} + \log \hat{f}_m(\lambda) .$$

To smooth a log spectral density, compute cepstral correlations

$$\tilde{\gamma}(v) = \frac{1}{Q} \sum_{k=0}^{Q-1} \exp\left(-2\pi i v \frac{k}{Q}\right) \log \tilde{f}_{\tilde{Y}}\left(\frac{k}{Q}\right) ,$$

for $v = 0, 1, \ldots, T$. Define, following Wahba (1980),

$$\{\log f_{\tilde{Y}}(\lambda)\}^{\wedge} = .57721 + \sum_{|v|<T} \exp(2\pi i v\lambda)\, \tilde{\gamma}(v) \frac{1}{1 + (v/M)^{2r}}$$

where r is an integer ≥ 2 ; usually one takes $r = 4$ or $r = 2$, and M is a real number chosen in practice to be an integer satisfying $2 \leq M \leq 12$. One calls M the half power point of the estimate.

To introduce a criterion for the choice of M, define $g(\lambda) = \log f_{\tilde{Y}}(\lambda)$, $\tilde{g}(\lambda) = \log \tilde{f}_{\tilde{Y}}(\lambda)$, $\gamma(v) = \int_{-0.5}^{0.5} \exp(2\pi i\lambda v) g(\lambda) d\lambda$, $\hat{g}_M(\lambda) = \{\log f_{\tilde{Y}}(\lambda)\}^{\wedge}$ defined above. A measure of the goodness of an estimator is

$$R_M = E \int_0^1 |\hat{g}_M(\lambda) - g(\lambda)|^2 d\lambda = \frac{1}{T} \sum_{j=0}^{T-1} E\{\hat{g}_M(\tfrac{j}{T}) - g(\tfrac{j}{T})\}^2 .$$

Following Wahba (1980), to minimize R_M one minimizes $\tilde{R}_M = B(M) + V(M,T)$, defining

$$B(M) = \frac{1}{M^{4r}} \sum_{|v|<T/2} |\tilde{\gamma}(v)|^2\, v^{4r} \{1 + (\tfrac{v}{M})^{2r}\}^{-2} ,$$

$$V(M,T) = \frac{M}{T} \frac{\pi^2}{6}\, 4\int_0^\infty (1 + u^{2r})^{-1}\, du .$$

One evalues $\tilde{R}_M$ for various values of M (and r); one chooses for these parameters the values minimizing $\tilde{R}_M$. The iterated spectral estimator is data adaptive, since the parameters m and M required to compute the estimator are chosen adaptively through order-determining (or model selection) criteria.

8. MAXIMUM ENTROPY SPECTRAL ESTIMATION

To discuss the philosophical basis of the maximum entropy method of spectral estimation introduced by Burg (1967), we need to discuss further the role of information numbers in statistics. To a sample $\{Y(t)\ ,\ t = 1, \ldots, T\}$ there is a true probability density $f(Y(1), \ldots, Y(T))$; we denote by $f_\theta(Y(1),\ldots,Y(T))$ a probability density function which is a function of parameters and which represents a model for the true probability density. A measure of the discrepancy between f and f_0 is the Kullback-Liebler information number or directed divergence

$$I_T(f;f_0) = \frac{1}{T} E_f \left[\log \frac{f}{f_0}\right]$$

$$= \frac{1}{T}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(y_1,\ldots,y_T)\log\frac{f(y_1,\ldots,y_T)}{f_\theta(y_1,\ldots,y_T)}dy_1\ldots dy_T\ .$$

Pinsker (1963) shows that in the limit as T ⟶

$$2I_T(f;f_\theta) = \int_{-0.5}^{0.5}\left\{\frac{f(\lambda)}{f_\theta(\lambda)} - 1 - \log\frac{f(\lambda)}{f_\theta(\lambda)}\right\} d\lambda$$

$$= H(f;f_\theta) - H(f;f)\ .$$

We can distinguish two ways to use this formula, (1) a statistical or data analysis approach, and (2) a probability approach. A data analysis approach to parameter estimation is to use a raw estimator $\tilde{f}$ of f (which, while a wiggly estimator of f, is satisfactory when only used as an integrand) to form an estimator $I_T(\tilde{f};f_\theta)$ of $I_T(f;f_\theta)$.

In contrast to the data based approach which minimizes $H(\tilde{f};f_\theta)$ over θ, is the probability approach which maximizes $H(f;f)$ over all functions f satisfying a set of constraints

$$\int_{-0.5}^{0.5} \psi_j(\lambda)\ f(\lambda)\ d\lambda = C_j\ ,\quad j = 1, \ldots, M$$

for specified functions $\psi_j(\lambda)$. An example of a set of constraints is to require the first m correlations of $f(\lambda)$ to equal sample correlations $\hat{\rho}(j)$:

$$\int_{-0.5}^{0.5} e^{2\pi i\lambda j}\ f(\lambda)\ d\lambda = \hat{\rho}(j)\ ,\quad j = 0, \pm 1, \ldots, \pm m\ .$$

Since $\quad H(f;f) = \int_{-0.5}^{0.5}\{1 + \log f(\lambda)\}\ d\lambda$,

the optimal function $\hat{f}(\lambda)$ is called a maximum entropy estimator of $f(\lambda)$. It is well known that $\hat{f}(\lambda)$ has the form of an autoregressive spectral density:

$$\hat{f}(\lambda) = \sigma_m^2 \, |1 + \alpha_1 e^{2\pi i\lambda} + \ldots + \alpha_m e^{2\pi i\lambda m}|^{-2}$$

The maximum entropy principle provides a motivation or justification for the use of autoregressive spectral estimators. However the maximum entropy principle provides no insight into how to identify an optimal order m, or even what are the effects of different methods of estimating the parameters $\sigma_m^2, \alpha_1, \ldots, \alpha_m$. It provides no guidance for how to learn from the data whether the time series is non-stationary (long memory) or stationary (short memory), or whether the best time series model is AR, MA, or ARMA. In my view it is a principle for deriving probability models, rather than statistically fitting models to data.

It should be realized that the maximum entropy principle justifies autoregressive estimators only for short memory time series (for whom log $f(\lambda)$ is integrable). Autoregressive estimators are justified for long memory time series by the fact that a pure harmonic $Y(t) = A \cos \frac{2\pi}{p} t + B \sin \frac{2\pi}{p} t$ satisfies $Y(t) - \phi Y(t-1) + Y(t-2) = 0$ where $\phi = 2 \cos \frac{2\pi}{p}$.

A justification of autoregressive estimators for short memory time series that I prefer is the existence of the infinite autoregressive scheme representation for a stationary time series satisfying: spectral density $f(\lambda)$ is continuous and differentiable; $f(\lambda)$ is bounded above and below; $f'(\lambda)$ is square integrable. Then $f(\lambda)$ has an infinite autoregressive representation

$$f(\lambda) = \sigma_\infty^2 \, |g_\infty(e^{2\pi i\lambda})|^{-2}$$

where $g_\infty(z) = 1 + \alpha_{1,\infty} z + \ldots + \alpha_{m,\infty} z^m + \ldots$.

9. PARAMETRIZATION OF AUTOREGRESSIVE SPECTRAL ESTIMATORS

There are many approaches for forming autoregressive spectral estimators, because there are four equivalent ways of parametrizing them: (A) autoregressive coefficients, (B) correlations, (C) partial correlations, and (D) residual variances.

A. Consider autoregressive coefficients $0 < \sigma_m^2 \leq 1, \alpha_{1,m}, \ldots, \alpha_{m,m}$ such that $g(z) = 1 + \alpha_{1,m} z + \ldots + \alpha_{m,m} z^m$ satisfies $g(z) \neq 0$ for complex z such that $|z| \leq 1$. Thus g(z) is a minimum phase filter transfer function. These coefficients define the autoregressive spectral estimator $f_m(\lambda) = \sigma_m^2 \, |g_m \, (e^{2\pi i\lambda})|^{-2}$.

B. Consider correlation coefficients $\rho(1), \rho(2), \ldots, \rho(m)$ which are positive definite. The correlation coefficients determine autoregressive coefficients by solving the Yule Walker equation (with $\alpha_{0,m} = 1$)

$$\sum_{j=0}^{m} \alpha_{j,m}\rho(j-k) = 0, \; k = 1, \ldots, m; \; = \sigma_m^2 , \; k = 0 .$$

The autoregressive coefficients determine the correlation coefficients by

$$\rho(j) = \int_{-0.5}^{0.5} \exp\,(2\pi i\lambda j) \; f_m(\lambda) \; d\lambda \; .$$

C. Consider coefficients $\Pi(1), \ldots, \Pi(m)$ satisfying $|\Pi(1)| < 1, \ldots, |\Pi(m)| < 1$. They represent partial correlation coefficients defined theoretically by: $\Pi(j)$ = partial correlation between $Y(t)$ and $Y(t-j)$, conditioned on $Y(t-1), \ldots, Y(t-j+1)$.

D. Consider coefficients $\sigma_1^2, \ldots, \sigma_m^2$, sign $\Pi(1), \ldots,$ sign $\Pi(m)$ satisfying $1 > \sigma_1^2 > \sigma_2^2 > \ldots > \sigma_m^2 > 0$. They represent residual variances defined by: σ_j^2 = mean square prediction error of $Y(t)$ given $Y(t-1), \ldots, Y(t-j)$, expressed in units of $E[|Y(t)|^2]$.

Partial correlation coefficients determine autoregressive coefficients and residual variances by the Levinson recursion (see Makhoul (1977)):

$$\alpha_{k,k} = - \Pi(k) \; ,$$

$$\alpha_{j,k} = \alpha_{j,k-1} - \Pi(k) \; \alpha_{k-j,k-1},$$

$$\sigma_k^2 = \sigma_{k-1}^2 \, \{1 - \Pi^2(k)\} \quad .$$

Residual variances determine partial correlation coefficients by a formula due to Dickenson (1978)

$$\Pi(k) = \text{sign } \Pi(k) \left\{ 1 - \frac{\sigma_k^2}{\sigma_{k-1}^2} \right\}^{\frac{1}{2}}$$

Autoregressive coefficients determine partial correlations by the recursion (Barndorf-Nielsen and Schon (1973))

$$\alpha_{j,k-1} = \{1 - \Pi^2(k)\}^{-1} \{\alpha_{j,k} + \Pi(k) \; \alpha_{k-j,k}\} \; .$$

In summary, to form $f_m(\lambda)$ one can specify any one of the four parametrizations. Given correlations, to solve the Yule-Walker equations one has many approaches: (1) SWEEP, (2) Cholesky decomposition, (3) Levinson-Durbin recursion, which computes partial correlation coefficients by

$$-\Pi(k) = \sum_{j=0}^{k-1} \alpha_{j,k-1}\, \rho(k-j) / \sigma^2_{k-1} ,$$

and (4) Levinson-Whittle-Robinson recursion which computes $\Pi(k)$ using forward and backward prediction error coefficients (see Kailath (1974)).

III. AN OUTLINE OF EMPIRICAL SPECTRAL ANALYSIS

Successive stages of analysis whose outputs are combined to form estimators of the spectrum of a single time series are:

Data Transformation and Detrending

Data Windowing

Extend with Zeroes

Fourier Transform

Average Short-time Segment Spectral Density Estimators

Sample Spectral Density, Sample Spectral Distributions

Spectral Average Direct Spectral Density Estimators

Sample Correlations

Indirect Lag Window Spectral Density Estimators

Autoregressive Coefficients, Yule Walker Equations

Autoregressive Spectral Density Estimators

Autoregressive Order Determination AIC CAT

Memory Identification, ARMA Identification

Autoregressive Coefficients: Nonstationary Least Squares, Lattice Algorithms. Kalman Filtering

Autoregressive Transformation of Y to $\tilde{Y}$

If Y long memory, either seek $\tilde{Y}$ short memory and return to data transformation stage or go to long memory mixed or band-limited methods listed below.

If Y short memory, seek whitening filters

Log Spectral Density Estimators of $\tilde{Y}$, via cepstral correlations

Iterated Adaptive Spectral Density Estimators of Y

Subset ARMA Identification

S-Array ARMA Identification

ARMA Spectral Density Estimator of Y

Other spectral analysis procedures:

Robust Autoregressive Transformation of Y to $\tilde{Y}$

Mixed Spectral Estimation (Long Memory)

Bandlimited Noise Spectral Estimation (Long Memory)

New techniques under research:

Nonparametric Data Modeling of Sample Spectral Density

Spectral De-whitening

A good description of techniques for reliably estimating the spectrum is in Thomson (1977). We must conclude our outline here due to space limitations.

REFERENCES

Akaike, H. (1974). A new look at the Statistical model identification, IEEE Trans. Automat. Contr., AC-19, 716-723.

Akaike, H. (1977). On entropy maximization principle, Applications of Statistics, P.R. Krishnaiah, ed., North-Holland, Amsterdam, 27-41.

Barndorf-Nielsen, O. and Schon, G. (1973). On the parametrization of autoregressive models by partial autocorrelations, J. Multivariate Analysis, 3, 408-419.

Burg, John P. (1967). Maximum entropy spectral analysis, Reprinted in Childers (1978).

Childers, D.G. (1978). Modern Spectrum Analysis, New York: IEEE Press.

Dickenson, Bradley W. (1978). Autoregressive estimation using residual energy ratios, IEEE Transactions on Information Theory, IT-24, 503-505.

Gray, R.M., Buzo, A., Gray, A.H.Jr., and Matsuyama, Y. (1980). Distortion measures for speech processing, submitted for publication.

Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression, Journal of the Royal Statistical Society, 41, 190-195.

Harris, F. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. IEEE, 66, 51-83.

Kailath, T. (1974). A view of three decades of linear filtering theory, IEEE Trans. Inform. Theory, IT-20, 145-181.

Makhoul, J. (1977). Stable and efficient lattice methods for linear prediction, IEEE Trans. Acous. Speech, and Signal Processing, ASSI-25, 423-428.

Parzen, E. (1962). Stochastic Processes, Holden Day: San Francisco.

Parzen, E. (1964). An approach to empirical time series, J. Res. Nat. Bur. Standards, 68D, 937-951.

Parzen, E. (1967). Time Series Analysis Papers, Holden Day: San Francisco.

Parzen, E. (1974). Some recent advances in time series modeling, IEEE Transactions on Automatic Control, Vol. AC-19, No. 6, December 1974, 723-730.

Parzen, E. (1977). Multiple time series: determining the order of approximating autoregressive schemes, Multivariate Analysis IV, ed. by P. Krishnaiah, North Holland: Amsterdam, 283-295.

Parzen, E. (1979). Forecasting and whitening filter estimation, TIMS Studies in the Management Sciences, 12, 149-165.

Parzen, E. (1980). Time series modeling, spectral analysis, and forecasting, Directions in Time Series Analysis, ed. D.R. Brillinger and G.C. Tiao, Institute of Mathematical Statistics.

Pinsker, M. (1963). Information and Information Stability of Random Variables, Holden Day: San Francisco.

Schwarz, G. (1976). Estimating the dimension of a model, Ann. Statist. 6, 461-467.

Thomson, D.J. (1977). Spectrum estimation techniques, Bell System Technical Journal, 56, 1769-1815.

Wahba, Grace (1980). Automatic smoothing of the log periodogram, Journal of the American Statistical Assn., 75, 122-132.

Wiener, N. (1930). Generalized harmonic analysis, Act. Math., 117-258.

Parzen, E. (1968). Statistical spectral analysis (single channel case) in 1968, Proceedings of NATO Advanced Study Institute on Signal Processing, Enschede, Netherlands.