AD-A104 940 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
AUTOREGRESSIVE SPECTRAL ESTIMATION, LOG SPECTRAL SMOOTHING, AND--ETC(U)
JUL 81 E PARZEN DAAG29-80-C-0070
UNCLASSIFIED TR-N-26 NL

1 OF 1
AD
A104940

END
DATE
FILMED
10-81
DTIC

(12)

TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS 77843

INSTITUTE OF STATISTICS
Phone 713 - 845-3141

LEVEL

AD A104940

AUTOREGRESSIVE SPECTRAL ESTIMATION, LOG

SPECTRAL SMOOTHING, AND ENTROPY

by Emanuel Parzen

Institute of Statistics

Texas A&M University

Technical Report No. N-26

July 1981

DTIC
ELECTE
OCT 1 1981
S D

Texas A & M Research Foundation
Project No. 4226T

"Robust Statistical Data Analysis and Modeling"

Sponsored by the Office of Naval Research

Professor Emanuel Parzen, Principal Investigator

DTIC FILE COPY

81 9 30 037

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: N-26	2. GOVT ACCESSION NO.: AD-A104940	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Autoregressive Spectral Estimation, Log Spectral Smoothing, and Entropy.		5. TYPE OF REPORT & PERIOD COVERED: Technical report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s): DAAG29-80-C-0070; ONR N0001481MP10001
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: U. S. Army Research Office, P. O. Box 12277, Research Triangle Park, NC 27709		12. REPORT DATE: Jul[redacted]981
		13. NUMBER OF PAGES: 7 (seven)
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report): Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

14 | 6 | 9 | 1 | 12 | 9

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

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)

Spectral estimation, maximum entropy, minimum information divergence, cross-entropy, maximum likelihood, quantile function, cepstral correlations, log spectral kernel estimation, iterated spectral estimation, Wolfer's sunspot data.

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

Two important methods of spectral estimation, autoregressive spectral estimation and log spectral kernel estimation, are derived from a minimum information divergence estimation principle. The fact that autoregressive spectral estimators are maximum entropy estimators is shown to be proved without the use of the calculus of variations using the properties of minimum information divergence estimation. Adaptive procedures for forming these estimators (and combining to form iterated estimators) are provided by order-determining and truncation point determining criteria, which are described. An estimated spectrum is given for Wolfer's sunspot data.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

347380

AUTOREGRESSIVE SPECTRAL ESTIMATION, LOG SPECTRAL SMOOTHING, AND ENTROPY[1]

Emanuel Parzen
Institute of Statistics
Texas A&M University
College Station, TX 77843

Abstract

Spectral estimation is motivated by information divergence distance. Two methods of spectral estimation are developed in this paper: autoregressive spectral estimation (section 3) and log spectral kernel estimation (section 4). They are motivated as parametric and non-parametric estimators which minimize "entropy" or "information divergence" distances between raw and fitted spectral densities. The role of entropy concepts in the statistical estimation of spectral densities (section 2) is explained by contrasting it with the role of entropy concepts in probability density estimation (section 1). Adaptive procedures for forming, and combining, these estimators for an observed time series are provided by order-determining and truncation (half-power) point determining criteria, which are described.

1. The role of entropy concepts in statistical estimation of probability density functions.

Let X be a continuous random variable, and $X_1,\ldots,X_n$ a random sample of X (consisting of independent random variables identically distributed as X). The distribution function $F(x), -\infty<x<\infty$, and the probability density function $f(x), -\infty<x<\infty$, are defined by

$$F(x)=\Pr(X\le x),\quad f(x)=F'(x) .$$

The entropy (or Shannon information) of X is denoted by H(f) and is defined by

$$H(f)=\int_{-\infty}^{\infty}-\log f(x)\, f(x)dx$$
$$=E_f[\log f(X)] .$$

All observed distributions are assumed to have finite entropy.

A maximum entropy density is a probability density $\hat{f}(x)$ determined by maximizing H(f) over all f satisfying certain constraints (usually involving moments of f).

Theorem 1A: Three important densities, and their characterization by a maximum entropy principle are: (1) uniform distribution over an interval a to b maximizes H(f) over the constraint that f is non-zero only on the interval a to b; (2) exponential distribution with mean μ maximizes H(f) over the constraint that f is non-zero only for x>0, and has mean μ ; (3) normal distribution with mean μ and variance σ^2 maximizes H(f) over the constraint that f has mean μ and variance σ^2.

The maximum entropy principle is a probability modeling principle in the foregoing examples. It becomes a statistical estimation principle (which fits distributions to data) when the constraints on f are expressed in terms of sample means and variances; it is then similar to the method of moments. A maximum entropy density estimator $\hat{f}$ can be expressed in symbols:

$$H(\hat{f})=\max_f H(f)$$

where f is constrained to have certain moments equal to the corresponding sample moments.

An alternative (and, we believe, more general) statistical estimation principle is provided by the cross-entropy H(g;f) and information divergence I(g;f) of two probability density functions f(x) and g(x). Define

$$H(g;f) = \int_{-\infty}^{\infty} \{-\log g(x)\}f(x)dx$$
$$=E_f[-\log g(X)];$$
$$I(g;f) = \int_{-\infty}^{\infty} \{-\log \frac{g(x)}{f(x)}\}f(x)dx$$
$$=E_f[-\log \frac{g(X)}{f(X)}]$$

Note that H(f)=H(f;f). Another name for information divergence is Kullback-Liebler information number. A minimum information divergence density $\hat{g}$ is an approximator to a specified density f determined by

$$I(\hat{g};f) = \min_g I(g;f)$$

where g is constrained to belong to a specified parametric family of probability densities.

Theorem 1B: Three important examples of minimum information divergence approximators or estimators are: (1) f is assumed to be positive only over the interval a to b, and g is any uniform distribution; then $\hat{g}$ is the uniform distribution over a to b; (2) f is positive only for $x\ge a$, and has a finite mean μ, and g is the two parameter exponential distribution; then $\hat{g}$ is the exponential distribution with mean μ and domain (a,∞); (3) f has finite mean μ and variance σ^2, and g is any normal distribution; then $\hat{g}$ is $N(\mu,\sigma^2)$.

Theorem 1B may have been first explicitly formulated by Thiel (1981), although it is implicitly known through the equivalence of maximum likelihood estimation with minimum information divergence estimation. An important observation by Thiel is that Theorem 1B can be used to prove Theorem 1A, and thus avoid the use of the calculus of variations. We extend this observation to spectral density estimation in section 2.

A minimum information divergence density $\hat{g}$ can be expressed as a minimum cross-entropy density:

$$H(\hat{g};f) = \min_g H(g;f)$$

A cross-entropy can be defined for an arbitrary (including discrete) distribution function F(x) by

$$H(g;F) = \int_{-\infty}^{\infty} \{-\log g(x)\}dF(x)$$

Therefore a minimum information divergence density $\hat{g}$ can be defined for an arbitrary distribution function F by

$$H(\hat{g};F) = \min_g H(g;F)$$

where g is constrained to belong to a specified parametric family of probability densities, Theorem 1B is true for this definition.

Consider now a finite sample $X_1,\ldots,X_n$ and a parametric model $f_\theta(x)$ for the true probability density f(x), indexed by parameters θ which one would like to estimate from the sample. A maximum likelihood estimator of θ is defined as the parameter values θ maximizing

$$L_n(\theta) = \frac{1}{n} \log f_\theta(X_1,\ldots,X_n)$$
$$= \frac{1}{n} \sum_{j=1}^{n} \log f_\theta(X_j)$$

Let $\tilde{F}(x), -\infty<x<\infty$, denote the sample distribution defined by

[1] This research was supported by the Office of Naval Research (Contract N00014-81-MP-10001, ARO DAAG29-80-C0070).

$\tilde{F}(x)$ = fraction of $X_1,\ldots,X_n \le x$.

One can express

$$L_n(\theta) = \int_{-\infty}^{\infty} \log f_\theta(x)\, d\tilde{F}(x)$$
$$= -H(f_\theta;\tilde{F}) .$$

Therefore maximum likelihood parameter estimators $\hat{\theta}$ yield minimum information divergence densities $f_{\hat{\theta}}(x)$.

By introducing $\tilde{f}$ to denote the (symbolic) sample probability density of the sample, one can regard $\hat{\theta}$ as satisfying

$$I(f_{\hat{\theta}};\tilde{f}) = \min_\theta I(f_\theta;\tilde{f}) .$$

A re-interpretation of maximum likelihood is obtained by rewriting the information divergence in terms of quantile functions whose role in statistical inference is emphasized by Parzen (1979).

Introduce the sample quantile function

$$\tilde{Q}(u) = \tilde{F}^{-1}(u) .$$

Its derivative $\tilde{q}(u) = \tilde{Q}'(u)$ satisfies

$$\tilde{q}(u)\ \tilde{f}(\tilde{Q}(u)) = 1 .$$

Define

$$\tilde{d}_\theta(u) = \frac{f_\theta(\tilde{Q}(u))}{\tilde{f}(\tilde{Q}(u))} = \frac{d}{du} F_\theta(\tilde{Q}(u))$$

where $F_\theta(x)$ is the distribution function with density $f_\theta(x)$. Make the change of variable $u = \tilde{F}(x)$, $x = \tilde{Q}(u)$ to obtain

$$I(f_\theta;\tilde{f}) = \int_0^1 -\log \tilde{d}_\theta(u)\, du$$

which one can interpret as a measure of how close to a uniform density is $\tilde{d}_\theta(u)$. The full consequences of this interpretation are explored elsewhere.

It should be noted that one can define other measures to minimize to form parameter estimators: examples are

$$\int_0^1 [\tilde{d}_\theta(u)-1]^2 du ,$$

whose minimization leads to "modified chi-square" estimators, and

$$\int_0^1 [F_\theta(\tilde{Q}(u))-u]^2 du ,$$

whose minimization leads to "minimum distance estimators".

Information divergence is the measure that most readily generalizes to stochastic processes.

It should be noted that only the parameter estimation problem is efficiently solved by minimizing $I(f_\theta;\tilde{f})$. The problem of goodness of fit is solved by considering the size of the difference from the uniform distribution $D(u) = u$ of $D_\theta(u) = F_\theta(\tilde{Q}(u))$ for $\theta = \hat{\theta}$. The model identification problem is to find distribution functions $\hat{F}$ such that $\hat{F}(\tilde{Q}(u))$ is parsimoniously not significantly different from the uniform distribution $D(u) = u$.

2. The role of entropy concepts in statistical estimation of spectral density functions.

Let $Y(t)$, $t = 1, \ldots, T$ be a sample of a Gaussian zero mean stationary time series with covariance function

$$R(v) = E[Y(t)Y(t+v)],\ v = 0, \pm 1, \pm 2, \ldots,$$

and correlation function

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

We assume $R(v)$ and $\rho(v)$ are absolutely summable, and define the power spectrum $S(w)$, $0 \le w \le 1$, and the spectral density $f(w)$, $0 \le w \le 1$, by

$$S(w) = \sum_{v=-\infty}^{\infty} e^{-2\pi i w v} R(v)$$

$$f(w) = \sum_{v=-\infty}^{\infty} e^{-2\pi i w v} \rho(v)$$

The spectral distribution function is defined by

$$F(w) = \int_0^w f(w')\, dw' ,\quad 0 \le w \le 1 .$$

When $\rho(v)$ is not assumed to be summable, there always exists a spectral distribution function $F(w)$, $0 \le w \le 1$, such that

$$\rho(v) = \int_0^1 e^{2\pi i w v} dF(w)$$

When $\rho(v)$ is summable, it has the spectral representation

$$\rho(v) = \int_0^1 e^{2\pi i w v} f(w) dw .$$

A stationary Gaussian time series is called White noise if

$$\rho(v) = 0,\ v > 0 ;$$
$$f(w) = 1,\ 0 \le w \le 1 ;$$
$$F(w) = w,\ 0 \le w \le 1 .$$

A stationary Gaussian time series with summable correlation function and integrable log spectral density can be represented in terms of a white noise time series $\varepsilon(t)$ representing the innovations [prediction errors $Y^\nu(t) = Y(t) - Y^\mu(t)$ of the infinite memory one-step ahead predictor $Y^\mu(t)$ of $Y(t)$]. The AR(∞), or infinite order autoregressive representation, is

$$Y(t)+a_\infty(1)Y(t-1)\ \ldots + a_\infty(n)Y(t-n)+\ldots = \varepsilon(t) .$$

The MA(∞), or infinite order moving representation, is

$$Y(t)=\varepsilon(t)+b_\infty(1)\varepsilon(t-1)+b_\infty(2)\varepsilon(t-2)+\ldots .$$

A finite parameter representation is an ARMA(p,q) of the form

$$Y(t)+a_p(1)Y(t-1)+\ldots+a_p(p)Y(t-p)$$
$$= \varepsilon(t)+b_q(1)\varepsilon(t-1)+\ldots+b_q(q)\varepsilon(t-q) .$$

The filter relating $Y(t)$ and $\varepsilon(t)$ is called a whitening filter. Parameter estimation is the theory of estimation of the parameters of the whitening filter and model identification is the theory of estimation of the structural form of the whitening filter. To develop approaches to parameter estimation for a random sample, in section 1 we defined the following concepts:

- Entropy $H(f)$,
- Maximum entropy density $\hat{f}$,
- Cross-entropy $H(g;f)$,
- Information divergence $I(g;f)$,
- Minimum information divergence density,
- Minimum cross-entropy density,
- Likelihood of a sample ,
- Maximum likelihood parameter estimator.

To develop approaches to estimation of the parameters θ of a parametric model $f_\theta(w)$ of the spectral density $f(w)$ of a stationary zero mean Gaussian time series $Y(t)$, we develop analogues of the foregoing concepts. We start with an approximate formula for the likelihood function

$$L_T(\theta) = \frac{1}{T} \log f_\theta(Y(1),\ldots,Y(T))$$

of the time series sample. We assume that $Y(t)$ has been divided by $(R(0))^{1/2}$ so that it can be considered to have variance 1, and its covariance function equals its correlation function.

The first step in analyzing a time series should be to compute the sample correlation function

$$\hat{\rho}(v) = \sum_{t=1}^{T-v} Y(t)Y(t+v) \div \sum_{t=1}^{T} Y^2(t)$$

and the sample spectral density

$$\tilde{f}(w) = \left|\sum_{t=1}^{T} Y(t)e^{-2\pi iwt}\right|^2 + \sum_{t=1}^{T} Y^2(t)$$

$$= \sum_{|v|<T} e^{-2\pi iwv}\,\hat{\rho}(v) .$$

It should be emphasized that in practice one should consider using a "data window" to compute $\tilde{f}(w)$, for $w = k/Q$, $k = 0,1,\ldots,Q-1$, by

$$\tilde{f}(w) = |\tilde{\psi}(w)|^2 + \frac{1}{Q}\sum_{k=0}^{Q-1}|\tilde{\psi}(\tfrac{k}{Q})|^2 ,$$

$$\tilde{\psi}(w) = \sum_{t=1}^{T} Y(t)K(\tfrac{t}{T})\exp(-2\pi iwt)$$

for a suitable kernel $K(x)$ (properties of windows are discussed in Harris (1978)). In addition for statistical stability one should then slightly smooth $\tilde{f}(w)$: (1) compute the sample correlation function 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 $0 \le v \le Q-T$ (and therefore one may want to choose $Q \ge 2T$); (2) compute a slightly smoothed sample spectral density by

$$\bar{f}(w) = \sum_{|v|<T} \exp(-2\pi iwv)\; k(\tfrac{v}{M})\hat{\rho}(v)$$

where $M \ge T/2$ and $k(u)$ is a suitable kernel, such as the Parzen lag window:

$$\begin{aligned} k(u) &= 1 - 6u^2 + 6|u|^3 , && |u| < 0.5 , \\ &= 2(1-|u|)^3 , && 0.5 \le |u| < 1 \\ &= 0 , && |u| \ge 1 . \end{aligned}$$

Back at the likelihood ranch, one may show that approximately

$$-L_T(\theta) = \frac{1}{2}\log 2\pi + H(f_\theta;\tilde{f})$$

where

$$H(f_\theta;\tilde{f}) = \frac{1}{2}\int_0^1 \left\{\log f_\theta(w) + \frac{\tilde{f}(w)}{f_\theta(w)}\right\}dw$$

This formula for likelihood shows that the sample spectral density $\tilde{f}(w)$ is a <u>sufficient</u> <u>statistic</u> for a time series. However, it is a very wiggly function and by itself is not a consistent estimator of $f(w)$. Estimators $\hat{f}(w)$ of $f(w)$ can be regarded as "smoothings" of $\tilde{f}(w)$, but the basic problem is how much to smooth.

Another aspect of the likelihood formula is its justification as an approximation. To those misguided analysts for whom maximum likelihood provides the ultimate estimator for which no expense should be spared, there is no substitute for the <u>exact</u> likelihood (which of course is exact only if the model being assumed is exactly true). Information concepts enter estimation theory when one recognizes that maximum likelihood estimation is a technical device for carrying out minimum information divergence estimation. The information divergence for a sample $Y(t)$, $t = 1, \ldots, T$, is defined in general by

$$I_T(f_\theta;f) = \frac{1}{T}E_f\left[\log\frac{f_\theta(Y(1),\ldots,Y(T))}{f(Y(1),\ldots,Y(T))}\right]$$

where here f denotes the true probability density of the sample, and f_θ is a model for f. It should be noted that we are using the notation f and f_θ with a variety of meanings. For a Gaussian zero mean stationary time series, the probability density of the sample is specified by the spectral densities, $f(w)$ of the true distribution and $f_\theta(w)$ of the model. We continue to denote the information divergence by $I_T\,(f_\theta;f)$ but now f indicates a spectral density rather than a probability density. Pinsker (1963) proves a formula for $I_T\,(f_\theta;f)$ in the limit as $T \to \infty$:

$$\lim_{T\to\infty} I_T\,(f_\theta;f) = I(f_\theta;f)$$

where $I(f_\theta;f)$ is the information divergence defined as follows.

For two spectral densities f and g, the <u>information divergence</u> $I(g;f)$, <u>cross-entropy</u> $H(g;f)$, and <u>entropy</u> $H(f)$ are defined:

$$I(g;f) = \frac{1}{2}\int_0^1 \left\{\frac{f(w)}{g(w)} - \log\frac{f(w)}{g(w)} - 1\right\}\;dw$$

$$= H(g;f) - H(f;f) ,$$

$$H(g;f) = \frac{1}{2}\int_0^1 \left\{\log g(w) + \frac{f(w)}{g(w)}\right\}\;dw ,$$

$$H(f) = H(f;f) = \frac{1}{2}\int_0^1 \{\log f(w) + 1\}\;dw .$$

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

The information divergence can be related to the L_2 log spectral density distance

$$L_2L(f,g) = \int_0^1 \{\log f(w) - \log g(w)\}^2\,dw ,$$

using the fact that $u = \exp(\log u) = 1 + \log u + (\tfrac{1}{2})(\log u)^2$. When f and g are "neighbors" in the sense that their ratio approximates 1,

$$I(g;f) = \frac{1}{4}L_2L(f,g) ;$$

then minimizing I is equivalent to minimizing L_2L. An extensive discussion of these distances is given by Gray, Buzo, Gray, and Matsuyama (1980).

The concepts have now been defined to state some of the basic facts of parameter estimation theory.

Maximum likelihood estimators $\hat{\theta}$ are equivalent to sample minimum cross-entropy estimators $\hat{\theta}$ defined by

$$H(f_{\hat{\theta}};\tilde{f}) = \min_\theta H(f_\theta;\tilde{f})$$

They can be regarded as estimators of the population minimum cross-entropy "parameters" θ^* defined by

$$H(f_{\theta^*};f) = \min_\theta H(f_\theta;f)$$

where f is the true spectral density.

A maximum entropy spectral density $\hat{f}$ is defined by

$$H(\hat{f}) = \max_f H(f)$$

where f is constrained to satisfy a set of constraints of the form

$$\int_0^1 \psi_j(w)\,f(w)\,dw = C_j,\; j = 1, \ldots, M,$$

for M specified functions $\psi_j(w)$ and constants C_j.

When the constraints are of the form

$$\int_0^1 e^{2\pi iwj}f(w)\,dw = \rho(j),\; j = 0, \pm 1, \ldots, \pm m$$

It can be shown that $f(w)$ is the autoregressive spectral density $f_m(w)$ defined as follows:

$$f_m(w) = \sigma_m^2 |g_m(e^{2\pi i w})|^{-2} ,$$

where

$$g_m(z) = 1 + a_m(1)\, z + \ldots + a_m(m) z^m ;$$

the autoregressive coefficients $a_m(1),\ldots,a_m(m)$ satisfy normal equations (called "Yule-Walker equations)

$$\sum_{k=0}^{m} a_m(k)\rho(k-j) = 0,\ j = 1,\ \ldots,\ m ,$$

where $a_m(0) = 1$; and

$$\sigma_m^2 = \sum_{k=0}^{m} a_m(k)\ \rho(k)$$

It should be noted that from a sequence $\rho(v)$, $v = 0, \pm 1, \ldots$ one can quickly compute $f_m(w)$ for all successive values of $m = 1,2,\ldots$ using a variety of fast algorithms [see Kailath (1974)]. In practice the problem is to determine "optimal" values of m.

Some important properties of $f_m(w)$ are:

(1) $\int_0^1 e^{2\pi i w j} f_m(w) dw = \rho(j),\ j = 0, \pm 1, \ldots, \pm m$;

(2) $\int_0^1 \frac{f(w)}{f_m(w)}\ dw = 1$;

(3) $\int_0^1 \log f_m(w)\ dw = \log \sigma_m^2$

(4) $H(f_m;f) = H(f_m) = \frac{1}{2}(\log \sigma_m^2 + 1)$;

(5) $\sigma_m^2 = \int_0^1 |g_m(z)|^2 f(w)\ dw$

$$= \min_{c_1,\ldots,c_m} \int_0^1 |1+c_1 e^{2\pi i w} + \ldots + c_m e^{2\pi i w m}|^2 f(w)\ dw ;$$

(6) $g_m(z)$ has all its roots in the complex plane outside the unit circle;

(7) $\lim_{m\to\infty} \log \sigma_m^2 = \log \sigma_\infty^2$;

(8) $\log \sigma_\infty^2 = \int_0^1 \log f(w)\ dw$;

(9) $\lim_{m\to\infty} f_m(w) = f(w)$ *if f is differentiable (the* rate of convergence of $f_m(w)$ depends on the rate of convergence of σ_m^2 to σ_∞^2);

(10) $2I(f_m;f) = \log \sigma_m^2 - \log \sigma_\infty^2 \to 0$ as $m \to \infty$.

The foregoing facts explain why autoregressive spectral approximations, introduced in Parzen (1968), (1969), provide powerful, and natural, estimators of an unknown spectral density. They are generated by the "maximum entropy approach" introduced by Burg (1967). However the Burg algorithm does not compute the autoregressive coefficients $a_m(j)$ and the innovation variances σ_m^2 by the Yule-Walker equations. Indeed it does not compute either $\hat\rho(v)$ or $\hat f(w)$. It does not provide insight into how to identify "optimal" autoregressive orders m.

One approach to defining criteria for an optimal order m is *to examine how well one has transformed to* white noise the residual series

$$\epsilon_m(t) = Y(t) + a_m(1)\ Y(t-1) + \ldots + a_m(m) Y(t-m)$$

whose spectral density is given by

$$\bar f_m(w) = \frac{1}{\sigma_m^2} |g_m(e^{2\pi i w})|^2\ f(w)$$

$$= \frac{f(w)}{f_m(w)}$$

A "model identification" determined order m has the property that the spectral distribution function

$$\bar F_m(w) = \int_0^w \bar f_m(w')\ dw' ,\ 0 \le w \le 1 ,$$

is parsimoniously not significantly different from the uniform distribution $F_0(w) = w$ representing the spectral distribution function of white noise.

In deriving autore ressive spectral estimators or approximators, we ha so far developed an analog of Theorem 1A, by stating that autoregression provides maximum entropy estimators subject to the constraint that certain correlation values are attained. We prefer an analog of Theorem 1B, which states that: the minimum information divergence parameter estimators of an autoregressive model for the true spectral density are provided by the coefficients which satisfy the Yule-Walker equations.

A very important fact (that may not be widely known) is that the maximum entropy properties of autoregressive spectral densities follow from their minimum information divergence properties, using the fact that

$$I(f_m;f) = H(f_m) - H(f) \ge 0 ;$$

consequently $H(f) \le H(f_m)$. Since f and f_m satisfy the constraint that their first m correlations equal specified values, the entropy $H(f)$ achieves its maximum value at $f = f_m$.

3. Autoregressive spectral estimators and order determining criteria.

Given a sample $Y(t)$, $t = 1, \ldots, T$, there are many approaches for forming autoregressive spectral estimators, because [as summarized in Parzen (1981)] there are four equivalent ways of parametrizing them: (A) autoregressive coefficients, (B) correlations, (C) partial correlations, and (D) innovation variances. Here we only consider starting with the sample correlations $\hat\rho(v)$. Then for $m = 1, 2, \ldots$ one forms

$$\hat f_m(w) = \hat\sigma_m^2\ |\hat g_m(w)|^{-2} ,$$

where

$$\hat g_m(z) = 1 + \hat a_m(1)\ z + \ldots + a_m(m) z^m ;$$

the sample autoregressive coefficients $\hat a_m(j)$ satisfy the sample Yule-Walker equations

$$\sum_{k=0}^{m} \hat a_m(k)\ \hat\rho(k-j) = 0,\ j = 1,\ \ldots,\ m ,$$

where $\hat a_m(0) = 1$; and the order m innovation variance

$$\hat\sigma_m^2 = \sum_{k=0}^{m} \hat a_m(k)\ \hat\rho(k) .$$

Define $\hat\sigma_\infty^2$ by

$$\log \hat\sigma_\infty^2 = \int_0^1 \log \hat f(w)\ dw .$$

Then as m tends to T,

$$2I(f_m;\hat f) = \log \hat\sigma_m^2 - \log \hat\sigma_\infty^2 \to 0 .$$

We desire $\hat{f}_m$ to be a sequence of consistent estimators of f in the sense that if one chooses m as a suitable function of T, then as $T \to \infty$

$$I(\hat{f}_m;f) \to 0 \quad \text{and} \quad \hat{f}_m(w) \to f(w) ,$$

in probability, or with probability one, or in mean square. The first rigorous proof of such results was given by Berk (1974) who also finds the asymptotic variance of $\hat{f}_m(w)$, confirming conjectures in Parzen(1969).

We now consider the problem of choosing m adaptively from a sample of size T. Conceptually one would like to choose m to minimize $I(\hat{f}_m;f)$. One approach to such a procedure is given by Akaike (1974) and leads to an order determining criterion called AIC.

The Akaike information criterion computes for $m = 1,2, \ldots$

$$AIC(m) = \log \hat{\sigma}_m^2 + \frac{2m}{T}$$

and an optimal order $\hat{m}$ satisfying

$$AIC(\hat{m}) = \min_m AIC(m)$$

The optimal order $\hat{m}$ can equal 0, indicating that the time series is white noise. The value of AIC(0) can be adjusted to the value one desires for the probability of rejecting the hypothesis of white noise, when in fact the time series is white noise. We recommend

$$AIC(0) = -\frac{1}{T} .$$

Parzen (1974), (1977) proposes autoregressive order determining criteria, called CAT, whose foundations are different from those of AIC but which usually lead to exactly equivalent orders in practice. The time series model identification problem is to estimate the infinite autoregressive transfer function

$$g_\infty(z) = 1 + a_\infty(1)\, z + \ldots + a_\infty(n)\, z^n + \ldots$$

by a sample order m autoregressive transfer function $\hat{g}_m(z)$. To evaluate the overall mean square error it is convenient to define

$$J_m = E\int_0^1 \left| \frac{1}{\hat{\sigma}_m^2}\, \hat{g}_m(e^{2\pi i w}) - \frac{1}{\sigma_\infty^2}\, g_\infty(e^{2\pi i w}) \right|^2 f(w)dw$$

which can be shown to be the sum of a variance term

$$E\int_0^1 \left| \frac{1}{\sigma_m^2}\, \hat{g}_m(e^{2\pi i w}) - \frac{1}{\sigma_m^2}\, g_m(e^{2\pi i w}) \right|^2 f(w)dw$$

and a bias term

$$\int_0^1 \left|\frac{1}{\sigma_m^2}\, g_m(e^{2\pi i w}) - \frac{1}{\sigma_\infty^2}\, g_\infty(e^{2\pi i w}) \right|^2 f(w)dw$$

One can show that the variance term is approximately

$$\frac{1}{T} \sum_{j=1}^{m} \sigma_j^{-2}$$

while the bias term is exactly

$$\sigma_\infty^{-2} - \sigma_m^{-2}$$

Therefore

$$J_m = \frac{1}{T} \sum_{j=1}^{m} \sigma_j^{-2} - \sigma_m^{-2} + \sigma_\infty^{-2}$$

Candidates for optimal orders m are obtained from an estimator of the terms in J_m which depend on m; thus one minimizes

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

where

$$\hat{\sigma}_j^{*2} = \frac{T}{T-j}\, \hat{\sigma}_j^2$$

is an "unbiased" estimator of σ_j^2. At $m = 0$, we assign $CAT(0) = -(1 + (1/T))$.

It should be noted that a multiple time series version of CAT is given in Parzen (1977).

An order determining criterion which is consistent, but whose behavior in practice is controversial, is given by Hannan and Quinn (1979).

4. Log Spectral Kernel Estimator and Cepstral Correlations

The approach we have been describing for forming "optimal" estimators $\hat{f}(w)$ of the spectral density $f(w)$ of a stationary time series is to view $\hat{f}(w)$ as a function closest to $\tilde{f}(w)$ in a distance between spectral densities given by the information divergence $I(\tilde{f};f)$. The class of functions from which $\hat{f}(w)$ is chosen has been constrained (or specified) parametrically, in the sense that $\hat{f}(w)$ is of the form $f_{\hat{\theta}}(w)$, where $\hat{\theta}$ estimates the parameters θ of a model $f_\theta(w)$ for the true $f(w)$.

A non-parametric constraint is to impose a smoothness measure on $\hat{f}$ such as the integral square of the r-th derivative of $\log \hat{f}(w)$, denoted

$$\int_0^1 |(\log \hat{f}(w))^{(r)}|^2 \, dw .$$

One then seeks to choose $\hat{f}$ to maximize smoothness, while minimizing a measure of distance of $\hat{f}$ from $\tilde{f}$. Wahba (1980) introduces the estimation distance

$$\int_0^1 |\log \tilde{f}(w) - \log \hat{f}(w)|^2 dw + K\int_0^1 |(\log \hat{f}(w))^{(r)}|^2 dw$$

where K is a penalty parameter to be determined adaptively by the data. One may show that the resulting estimators of $g(w) = \log f(w)$ are of the form, called log spectral kernel estimators,

$$\hat{g}(w) = (\log f(w))^{\hat{}} = \sum_{v=-\infty}^{\infty} \exp(-2\pi i w v)\, k(\tfrac{v}{M})\, \tilde{\gamma}(v)$$

where

$$\tilde{\gamma}(v) = \int_0^1 \exp(2\pi i w v) \log \tilde{f}(w)\, dw$$

are called cepstral correlations, and the kernel $k(x)$ is given by (compare Parzen (1958))

$$k(x) = \frac{1}{1 + x^{2r}}$$

One often considers only two values for r, 2 and 4.

The statistical properties of cepstral correlation have been extensively investigated by Bhansali (1974).

Since $k(\frac{v}{M}) = \frac{1}{2}$ for $v = M$, we call M the "half power" lag. We seek to adaptively determine M from the sample to minimize the risk function

$$R_M = J(\hat{f};f) = E\, L_2 L(\hat{f},f)$$

assuming log f(w) has a representation

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

Following Wahba (1980), to minimize R_M one minimizes an estimator of it of the form

$$\hat{R}_M = B(M) + V(M,T)$$

where B(M) and V(M,T) are measures of bias and variance given by

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

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

A closed form evaluation of the integral in V(M,T) can be obtained.

5. Iterated spectral estimation.

Observed time series do not usually obey the assumptions made in the foregoing theory that Y(t) is a zero mean Gaussian time series with summable correlation function. We call such a time series a "short memory" time series (of which white noise is a special case, called a "no memory" time series). Otherwise the time series is called "long memory" (Parzen(1982)).

Autoregressive spectral estimators are especially suitable for matching the large scale oscillations of the spectral density of a long memory time series. The role of the autoregressive filter is then to transform the time series to a short memory time series [obtained as the residuals $\tilde{\epsilon}_m(t)$ described in section 2]. The spectral density of the short memory series, which can be regarded as the fine structure of the original spectral density, can be estimated by a log spectral smoothing estimator as well as by an autoregressive spectral estimator. Employing two different approaches to short memory spectral estimation is desirable since the problem of spectral estimation is not simply a problem of parameter estimation but is also one of model identification.

Iterated models for forecasting long memory time series are used by Parzen (1981) under the name of "ARARMA models" (see Appendix for an example).

REFERENCES

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

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

3. Berk, K.M. (1974). Consistent Autoregressive Spectral Estimates, Annals of Statistics, 2, 489-503.

4. Bhansali, R.J. (1974). Asymptotic Properties of the Wiener-Kolmogorov Predictor I, Journal of the Royal Statistical Society B, 36, 61-73.

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

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

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

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

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

10. Haykin, S. et al (1979). Nonlinear Methods of Spectral Analysis, Springer Verlag: Berlin.

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

12. Parzen, E. (1958). On asymptotically efficient consistent estimates of the spectral density of a stationary time series. Journal of the Royal Statistical Society B, 303-322 (Reprinted in Parzen (1967)).

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

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

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

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

17. Parzen, E. (1969). Multiple time series modeling, Multivariate Analysis II, ed. by P. Krishnaiah, Academic Press; New York, 389-410.

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

19. 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.

20. Parzen, E. (1979). Nonparametric statistical data modeling, Journal of the American Statistical Assoc., 74, 105-131.

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

22. 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.

23. Parzen, E. (1981). ARARMA Models for Time Series Analysis and Forecasting, Journal of Forecasting, Vol. 1, No. 1.

24. Parzen, E. (1981). Modern Empirical Spectral Analysis, Underwater Acoustics and Signal Processing, ed. L. Bjorno. Reidel: Dordrecht, Holland, 470-497.

25. Parzen, E. (1982). Time Series Model Identification and Prediction Variance Horizon, Applied Time Series Analysis II, ed. D. Findley, Academic Press; New York.

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

27. Thiel, H. (1981). Maximum entropy and minimum Kullback-Liebler information. Reprint, University of Chicago Business School.

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

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

APPENDIX: Wolfer Sunspot Numbers 1846-1963.

To illustrate the application of some of the foregoing ideas, we report an iterated autoregressive model fitted to the annual time series Y(t) of Wolfer's sunspot data for the years 1846-1963 (which is a sample of length T = 118). Our ARARMA model fitting algorithm automatically proposes the following model (which it hopes will have the best medium range, if not long range, forecasting capability):

$$\tilde{Y}(t) = Y(t) - .482\, Y(t-10) - .554\, Y(t-11)$$

$$\tilde{Y}(t) - 1.009\, \tilde{Y}(t-1) + .362\, \tilde{Y}(t-2) = \epsilon(t)$$

The series $\tilde{Y}(t)$ is a short memory time series to which Y(t) has been transformed by the initial autoregression on Y(t). As an estimator of the true log spectral density f(w) we take, up to a normalizing constant,

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

where

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

is the autoregressive spectral density corresponding to the transformation from Y(t) to $\tilde{Y}(t)$.

Figure 1 is a graph of the Wolfer sunspot data (the crosses represent the one-step ahead predictors of the model above). Figure 2 graphs the iterated autoregressive log spectral estimator $\{\log f_Y(\omega)\}^{\wedge}$.

Fig. 2

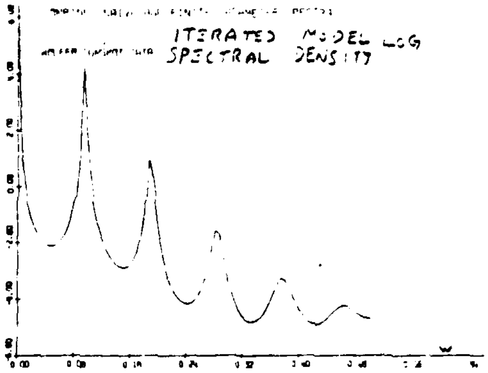
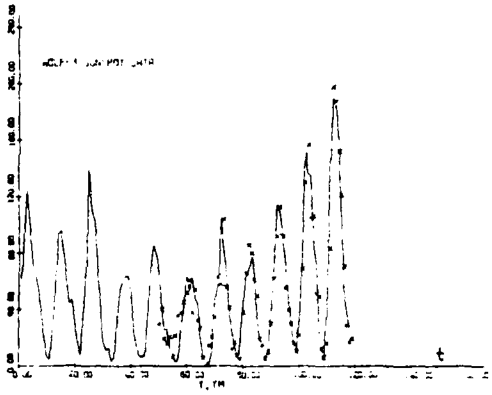


Fig. 1



END

DATE
FILMED
10-81

DTIC