FITTING TIME SERIES MODELS FOR PREDICTION[*]

by

WILLIAM S. CLEVELAND[**]
Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 661

JANUARY 1970

[*] This research was supported by the Army, Navy, Air Force and NASA under a contract, administered by the Office of Naval Research, with Yale University; by the National Science Foundation under Grant GU-2059 and the Air Force Office of Scientific Research under Contract AFOSR-68-1415 with the University of North Carolina at Chapel Hill.

[**] The major portion of the research for this paper was done while the author was at the Department of Statistics, Yale University.

Fitting Time Series Models for Prediction

by

William S. Cleveland

The mathematical theory of the best linear prediction of stationary time series presumes that the model generating the series, which can be specified by either the autocovariance function or the spectral density, is known. The true model is, of course, not known in practice, and the procedure is to fit a model and predict as if this fitted model were the truth. The question then is one of deciding whether the resulting predictions are about as good as could be gotten if the truth were known. This paper describes a method for assessing the predictions of the fitted model by an analysis of residuals. In particular, it is argued that the traditional tests of hypothesis for white noise are inappropriate if prediction is the goal, and a method is described for determining whether or not the mean square errors of the predictions arising from the fitted model can be measurably reduced.

1. Introduction

Let x_n, $-\infty < n < \infty$, be a discrete parameter, single-channel, covariance stationary time series with $E(x_n) = 0$ and spectral density f.

The linear least squares prediction $\phi_{n,p}$, for $p = 1,2,\ldots,$ of x_{n+p} from the infinite past $x_n, x_{n-1},\ldots$ is that random variable which minimizes $E(x_{n+p}-y)^2$ where y ranges over all linear combinations of x_n, $x_{n-1},\ldots$ and their limits in mean square. The calculation of $\phi_{n,p}$ can in theory be done if f is known. References describing this are (Doob, 1953, Chapter 12), (Grenander and Rosenblatt, 1957, p. 65-82), and (Whittle, 1963).

In practice, a finite number of observations $x_1,\ldots,x_N$ is available, and it is desired to calculate $\phi_{N,p}$ for various values of p. Of course, f is not known and must be estimated, and if prediction is the goal, the estimate is usually a rational spectral density

$$\hat{f}(\lambda) = \hat{c}^2 \frac{\left|1 + \sum_{k=1}^{b} \hat{\beta}_k e(k\lambda)\right|^2}{\left|1 + \sum_{k=1}^{a} \hat{\alpha}_k e(k\lambda)\right|^2}$$

where

$$(1) \qquad 1 + \sum_{k=1}^{b} \hat{\beta}_k z^k \neq 0 \quad \text{and} \quad 1 + \sum_{k=1}^{a} \hat{\alpha}_k z^k \neq 0$$

for all complex z such that $|z| \leq 1$, and

$$e(k\lambda) = \exp(2\pi ik\lambda) .$$

Such an estimate is generally gotten by one of two methods. The first is to assume that x_n is a moving-average, autoregressive process,

$$x_n + \sum_{k=1}^{a} \alpha_k x_{n-k} = \varepsilon_n + \sum_{k=1}^{b} \beta_k \varepsilon_{n-k}$$

where ε_n are independent (normal) random variables with $E(\varepsilon_n) = 0$ and $E(\varepsilon_n^2) = \sigma^2$, and (1) holds with $\hat{\alpha}_k$ replaced by α_k and $\hat{\beta}_k$ by β_k. This means that f is the rational spectral density,

$$f(\lambda) = \sigma^2 \frac{\left|1 + \sum_{k=1}^{b} \beta_k e(k\lambda)\right|^2}{\left|1 + \sum_{k=1}^{a} \alpha_k e(k\lambda)\right|^2} .$$

Then the unknown parameters α_k, β_k, and σ^2 are estimated. An excellent discussion of this method of fitting models is found in the works of Box and Jenkins cited in the bibliography.

The second method is to first estimate f by one of the standard nonparametric estimates gotten by smoothing the periodogram ((Parzen, 1968) and (Tukey, 1967)), and then approximate the estimate by a rational spectral density. The approximation is necessitated by the fact that the nonparametric estimate is not in a form from which predictions can be easily calculated.

It has been assumed that $E(x_n) = 0$. In practice, this means that a regression (a simple one if the series is assumed mean stationary) has been done and the non-zero means subtracted off, or x_n is the result of an initial series which has been differenced to the point where it is reasonable to assume a zero mean.

2. Calculation of the predictions and the residuals.

Let $\hat{\phi}_{N,p}$ be the prediction of x_{N+p} that results from calculating what would be the best prediction if $\hat{f}$ were the true density. In general, $\hat{f} \neq f$ and $\hat{\phi}_{N,p} \neq \phi_{N,p}$ so that there is an increase in the minimum mean square error $E(x_{N+p} - \phi_{N,p})^2$ to

$$E(x_{N+p}-\hat{\phi}_{N,p})^2 = E(x_{N+p}-\phi_{N,p})^2 + E(\hat{\phi}_{N,p}-\phi_{N,p})^2.$$

Let $\hat{a}_k$ and $\hat{b}_k$ be defined by

$$(2) \qquad (1 + \sum_{k=1}^{\infty} \hat{a}_k z^k)^{-1} = 1 + \sum_{k=1}^{\infty} \hat{b}_k z^k$$

$$= \frac{1 + \sum_{k=1}^{b} \hat{\beta}_k z^k}{1 + \sum_{k=1}^{a} \hat{\alpha}_k z^k} .$$

In what follows it will be convenient to let $\hat{a}_0 = \hat{b}_0 = 1$. It is shown in the Appendix (assuming the mild conditions (11) for f) that $\hat{\phi}_{N,p}$ satisfy

$$(3) \qquad \sum_{k=0}^{p-1} \hat{a}_k \hat{\phi}_{N,p-k} + \sum_{k=p}^{\infty} \hat{a}_k x_{N+p-k} = 0$$

for $p = 1,2,\ldots$. Since the $\hat{a}_k$ can be easily calculated recursively from $\hat{\alpha}_k$ and $\hat{\beta}_k$ using (2), $\hat{\phi}_{N,p}$ can be calculated recursively in p using (3).

Since only a finite number of observations are available, the second sum in (3) can run only from p to $N+p-1$, but the assumption is that enough of the series has been observed so that $x_0, x_{-1},\ldots$ have little effect on the prediction of x_{N+p}.

Let $\hat{\varepsilon}_n$ be the process defined by

$$(4) \qquad x_n + \sum_{k=1}^{\infty} \hat{a}_k x_{n-k} = \hat{\varepsilon}_n .$$

$\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N$ are called the residuals. Again, since only $x_1, \ldots, x_N$ are observed the sum in (4) must run from 1 to n-1, but generally, provided no root of $1 + \Sigma_{k=1}^{b} \hat{\beta}_k z^k$ is too close to the unit circle, this will affect only the first few residuals $\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_M$ where M/N is small. In practice, the initial residuals can be plotted and M chosen to be the point where the residuals seem to have settled down to stationarity. In performing the analysis of residuals described later, these initial values should be discarded; with an abuse of notation $\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_N$ will denote the remaining residuals.

Treating $\hat{f}$ as the true spectral density of x_n, which is what is done to calculate the predictions $\hat{\phi}_{N,p}$, is equivalent to treating $\hat{\varepsilon}_n$ as white noise (a sequence of uncorrelated random variables). Thus a way of investigating the adequacy of $\hat{f}$ as an approximation of f for prediction is to examine $\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_N$ to see how much they act like white noise. How to carry out the examination will be described in the next sections. The motivation for looking at residuals to check assumptions comes largely from (Anscombe, 1961) and (Anscombe and Tukey, 1963).

3. Hypothesis tests and prediction

The question of importance is the increase in the mean square error when a prediction is calculated treating $\hat{\varepsilon}_n$ as white noise. One important point is that the traditional tests of the hypothesis of white noise, such as the cumulative periodogram test (Bartlett, 1966, p. 318), do not answer this question. The test of hypothesis judges correlation in the fitted residuals according to a criterion of no use here. For instance, suppose that x_n is the moving-average process

$$x_n = \varepsilon_n + .01\varepsilon_{n-1},$$

where $E(\varepsilon_n^2) = 1$. Suppose $\hat{f} = 1$ so that (4) is

$$x_n = \hat{\varepsilon}_n.$$

The minimum mean square error of 1-step prediction is 1, whereas the mean square error using $\hat{f}$ is 1.0001, which in almost any practical application would not be an increase at all worth worrying about. But clearly for N large enough, the probability of the cumulative periodogram test rejecting $\hat{\varepsilon}_n$ as white noise is near 1.

The inappropriateness of tests is related to the fact that an analysis for understanding the mechanism generating the series is quite different than an analysis for prediction. Models very far from the truth can give nearly best predictions. A periodic component will a small amplitude might be of importance if you are trying to understand the mechanism generating the series, but the component might contribute very little to prediction. A good example of this can be found in (Whittle, 1954). In Section 7 of this paper is an example where the residuals show a definite nonwhite noise effect, but yet the model is giving predictions that are about as good as can be gotten.

4. The mean square error of $\hat{\phi}_{N,p}$

For the purpose of analyzing mean square errors, the $\hat{a}_k$ will be taken as fixed numbers, rather than taking their sampling variability into account. The reason for doing this is that if you are about to calculate a prediction using a particualr $\hat{f}$, estimated from the sample at hand, then you are interested in knowing how that particular $\hat{f}$ performs.

(Until now, the same notation has been used for a random variable and a realization of that variable. In the following example, the two will be distinguished by writing the former in bold face.) For example, suppose $\underset{\sim}{x}_n$ is the autoregression

$$\underset{\sim}{x}_n + \tfrac{1}{2}\underset{\sim}{x}_{n-1} = \underset{\sim}{\varepsilon}_n$$

with $E(\underset{\sim}{\varepsilon}_n^{\,2}) = 1$ and $\underset{\sim}{\varepsilon}_n$ independent. Let $\hat{\underset{\sim}{a}}_1$ be a function of $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_N$ which is an estimator of $\tfrac{1}{2}$. Suppose from the sample $x_1,..,x_N$ the observed value of $\hat{\underset{\sim}{a}}_1$ is $\hat{a}_1 = \tfrac{1}{4}$. Then (4) is

$$\underset{\sim}{x}_n + \tfrac{1}{4}\underset{\sim}{x}_{n-1} = \hat{\underset{\sim}{\varepsilon}}_n .$$

The predictor of $\underset{\sim}{x}_{N+1}$ gotten by treating $\hat{\underset{\sim}{\varepsilon}}_n$ as white noise is $\hat{\underset{\sim}{\phi}}_{N,1} = -\tfrac{1}{4}\underset{\sim}{x}_N$. The mean square error of $\underset{\sim}{\phi}_{N,1}$ is

$$\begin{aligned} E(\underset{\sim}{x}_{N+1} - \hat{\underset{\sim}{\phi}}_{N,1})^2 &= E(\underset{\sim}{\varepsilon}_{N+1} + (\tfrac{1}{4}-\tfrac{1}{2})\underset{\sim}{x}_N)^2 \\ &= 1 + (\tfrac{1}{4}-\tfrac{1}{2})^2\,\frac{4}{3} \\ &= 1\frac{1}{12} . \end{aligned}$$

(The minimum mean square error is 1.) If the variability of $\hat{\underset{\sim}{a}}_1$ were taken into account, the mean square error of $\hat{\underset{\sim}{\phi}}_{N,1}$ would be

$$E(\underset{\sim}{\varepsilon}_{N+1} + (\hat{\underset{\sim}{a}}_1-\tfrac{1}{2})\underset{\sim}{x}_N)^2 = 1 + E((\hat{\underset{\sim}{a}}_1-\tfrac{1}{2})^2\underset{\sim}{x}_N^{\,2}).$$

This expression is the mean square error of a class of predictors, $-\hat{\underset{\sim}{a}}_1\underset{\sim}{x}_N$. But if after observing the sample $x_1,\ldots,x_N$, $\hat{\underset{\sim}{a}}_1$ takes the value $\tfrac{1}{4}$, you are no longer interested in the entire class; you are considering whether or not to use the particular predictor $-\tfrac{1}{4}\underset{\sim}{x}_N$ and want to know its mean square error.

Let h be the spectral density of $\hat{\varepsilon}_n$ and γ_k the autocovariances. From (3) and (4),

$$\sum_{k=0}^{p-1} \hat{a}_k(x_{N+p-k} - \hat{\phi}_{N,p-k}) = \hat{\varepsilon}_{N+p}$$

for $p = 1,2\ldots$. Solving recursively,

$$x_{N+p} - \hat{\phi}_{N,p} = \sum_{k=0}^{p-1} \hat{b}_k \hat{\varepsilon}_{N+p-k} \,. \tag{5}$$

Thus the mean square error E_p of $\hat{\phi}_{N,k}$ is

$$E(x_{N+p} - \hat{\phi}_{N,p})^2 = \sum_{j,k=0}^{p-1} \hat{b}_j \hat{b}_k \gamma_{k-j} \,. \tag{6}$$

For $p=1$, $E(x_{N+1} - \hat{\phi}_{N,1})^2 = E(\hat{\varepsilon}_{N+1})^2$, which will be denoted by v.

It is often the case in discussions of prediction that the mean square error E_p of $\hat{\phi}_{N,p}$ is estimated by that value which would result if $\hat{f}$ were the true density,

$$\hat{c}^2 \left(\sum_{k=0}^{p-1} \hat{b}_k^{\,2} \right).$$

But in view of (6), a more reasonable and natural estimate of the mean square error is

$$\hat{E}_p = \sum_{k,j=0}^{p-1} \hat{b}_j \hat{b}_k \hat{\gamma}_{|k-j|} \tag{7}$$

where

$$\hat{\gamma}_k = \frac{1}{N} \sum_{n=1}^{N-k} \hat{\varepsilon}_n \hat{\varepsilon}_{n+k}$$

for $k = 0,1,\ldots, N-1$ are the sample autocovariances of the residuals. In the special case $p = 1$, $\hat{E}_1 = \frac{1}{N} \Sigma_{n=1}^{N} \hat{\varepsilon}_n^2$ will be denoted by $\hat{v}$.

Assuming x_n is a normal process, from (Bartlett, 1966, p.285) the covariance of $\hat{\gamma}_j$ and $\hat{\gamma}_k$ is asymptotically

$$\frac{1}{N} \int_0^1 h^2(\lambda)[e(k\lambda-j\lambda) + e(k\lambda+j\lambda)]d\lambda .$$

Thus the variance of $\hat{E}_p$ is asymptotically

$$\frac{2}{N} \int_0^1 h^2(\lambda) \Big| \sum_{k=0}^{p-1} \hat{b}_k e(k\lambda) \Big|^4 d\lambda .$$

A short derivation shows an alternative method of computing $\hat{E}_p$ and an additional property of the estimate. Let

$$I(\lambda) = \sum_{k=-(N-1)}^{N-1} \hat{\gamma}_{|k|} e(k\lambda) = \frac{1}{N} \Big| \sum_{n=1}^{N} \hat{\varepsilon}_n e(n\lambda) \Big|^2$$

be the periodogram of the residuals. Then from (7),

$$\hat{E}_p = \int_0^1 I(\lambda) \Big| \sum_{k=0}^{p-1} \hat{b}_k e(k\lambda) \Big|^2 d\lambda$$

$$= \frac{1}{N} \sum_{n=p}^{N} (\hat{\varepsilon}_n + \hat{b}_1 \hat{\varepsilon}_{n-1} + \ldots + \hat{b}_{p-1} \hat{\varepsilon}_{n-(p-1)})^2 + 0_p(\tfrac{1}{N}) .$$

From (5)

$$\hat{\varepsilon}_n + \hat{b}_1 \hat{\varepsilon}_{n-1} + \ldots + \hat{b}_{p-1} \hat{\varepsilon}_{n-(p-1)} = x_n - \hat{\phi}_{n-p,p} .$$

Thus

$$\hat{E}_p = \frac{1}{N} \sum_{n=p}^{N} (x_n - \hat{\phi}_{n-p,p})^2 + 0_p(\tfrac{1}{N}) .$$

That is, $\hat{E}_p$ is approximately the sample mean square error of the p-step $\hat{f}$ predictor.

5. Bounds on the increase in mean square error using the fitted model

Let

$$B_p = (v-\sigma^2)\left\{\sum_{k=0}^{p-1} |\hat{b}_k|\right\}^2$$

for $p = 1,2,\ldots,$ where

(8) $$\sigma^2 = \exp \int_0^1 \mathit{log}\, h(\lambda)d\lambda.$$

In the Appendix, it is shown that

(9) $$E(\phi_{N,p} - \hat{\phi}_{N,p})^2 \leq B_p,$$

with equality holding for $p=1$. That is, the increase in mean square error due to calculating the predictions treating $\hat{f}$ as the true density is no larger than B_p. Since $v = \int_0^1 h(\lambda)d\lambda$, B_p is small if the geometric mean of h is close to the arithmetic mean of h.

B_p is derived with a view toward the practical situation and represents a middle course steered between two dangers. One danger is that a bound will involve in too complex a manner the true parameters. Indeed, the best bound for $E(\phi_{N,p}-\hat{\phi}_{N,p})^2$ is the expression itself, but to try to use it would be an extreme bootstrap method. To use it (for all p) would virtually require knowing f. The other danger is that as a bound becomes more simple, it becomes useless as a bound.

B_p involves two unknown parameters, v and σ^2. The estimation of v by $\hat{v}$ has been discussed. In view of (8) σ^2 can be estimated by ((Whittle, 1952) and (Jones and Davis, 1968))

$$(10) \qquad \hat{\sigma}^2 = \exp(\gamma + \frac{1}{N'} \sum_{k=1}^{N'} \log I(\frac{k}{N'}))$$

where $N' = [N/2] - 1$ and $\gamma = .5772157...$ is Euler's constant. Assuming x_n is normal, $\log\hat{\sigma}^2$ is asymptotically normal with mean $\log\sigma^2$ and variance $\pi^2/6N$. In practice, if N is large, the periodogram is calculated at J equally spaced points using the Fast Fourier Transform, where $J \geq N$; in this case, N' would be replaced by J' in (10), where $J' = [J/2]-1$.

B_p can therefore be estimated by

$$\hat{B}_p = (\hat{v}-\hat{\sigma}^2)\left\{\sum_{k=0}^{p-1} |\hat{b}_k|\right\}^2 .$$

6. Assessing the predictor of the fitted model

The quantity

$$\frac{E(\phi_{N,p} - \hat{\phi}_{N,p})^2}{E(x_{N+p} - \hat{\phi}_{N,p})^2}$$

represents the best possible percentage reduction in the mean square error $E_p = E(x_{N+p}-\hat{\phi}_{N,p})^2$. B_p/E_p is a bound for this quantity and $\hat{B}_p/\hat{E}_p$ is an estimate of this bound. Thus if $\hat{B}_p/\hat{E}_p$ is small, it can be concluded that the possible percentage reduction in mean square error is small, so that $\hat{f}$ is adequate for p-step prediction. The variability of $\hat{B}_p/\hat{E}_p$ can be investigated by breaking the residuals up into blocks and looking at the variability of the block estimates.

If $\hat{B}_p/\hat{E}_p$ is too large to consider the model satisfactory, there are several possibilities for courses of action. Suppose first that $p=1$. Then (apart from sampling fluctuations) the model is not giving good 1-step predictions, since for $p=1$ equality holds in (9). That is, you should be able to get roughly a $100 \times (\hat{B}_p/\hat{E}_p)\%$ reduction in the mean square error of the 1-step predictor. If, however, $p > 1$, then since B_p/E_p is a bound with equality not generally holding, a $100 \times (\hat{B}_p/\hat{E}_p)\%$ reduction in the mean square error may not be able to be realized. You should be particularly suspicious that this is the case if $\hat{B}_1/\hat{E}_1$ is small. If a different or more elaborate model is fit with the result that $\hat{E}_p$ is reduced and the new $\hat{B}_p/\hat{E}_p$ is now satisfactory then, of course, this new model will be used. If, however, $\hat{E}_p$ is not measurably reduced in the new model but the new $\hat{B}_p/\hat{E}_p$ is satisfactory then either model can be used and generally the simpler one will be chosen. The final possibility is that no simple way is seen to reduce either $\hat{E}_p$ or $\hat{B}_p/\hat{E}_p$. In this case, you must rely on your judgement of the residual autocorrelations and spectrum to decide if further fitting is really warranted.

A new model can be fitted by going back to the beginning and fitting a new model to $x_1,\ldots,x_N$ or by fitting a model to the residuals $\hat{\varepsilon}_1,\ldots\hat{\varepsilon}_N$. In this latter case, if the spectral density h of $\hat{\varepsilon}_n$ is estimated by

$$\dot{c}^2 \frac{\left|1 + \sum_{k=1}^{\dot{b}} \dot{\beta}_k e(k\lambda)\right|^2}{\left|1 + \sum_{k=1}^{\dot{a}} \dot{\alpha}_k e(k\lambda)\right|^2} ,$$

then the new estimate of f is

$$\dot{c}^2 \frac{\left|1 + \sum_{k=1}^{\dot{b}} \dot{\beta}_k e(k\lambda)\right|^2 \left|1 + \sum_{k=1}^{b} \hat{\beta}_k e(k\lambda)\right|^2}{\left|1 + \sum_{k=1}^{\dot{a}} \dot{\alpha}_j e(k\lambda)\right|^2 \left|1 + \sum_{k=1}^{a} \hat{\alpha}_k e(k\lambda)\right|^2} .$$

Let $\dot{h}$ be the spectral density of the new fitted residuals. That σ^2 is independent of the particular fitted model is seen by noting from the Appendix that

$$\sigma^2 = \exp \int_0^1 \mathcal{L}og f(\lambda) d\lambda .$$

That is,

$$\exp \int_0^1 \mathcal{L}og h(\lambda) d\lambda = \exp \int_0^1 \mathcal{L}og \dot{h}(\lambda) d\lambda .$$

Thus the 1-step prediction error is the same for the residuals of all fitted models, and σ^2 need only be estimated once. If the initial model is rejected and a new model is fit, the estimate of σ^2 already calculated may be used in the analysis of the residuals of the new model. Indeed σ^2 might perhaps be estimated from the periodogram of $x_1, \ldots, x_N$ (which corresponds to the fitted model $\hat{b}_k = 0$ for $k > 0$). However, a note of caution must be given. If the spectral density of the process from which σ^2 is estimated is not fairly smooth over intervals of length $\frac{1}{N}$, then the estimate $\hat{\sigma}^2$ can be biased. Thus if the residuals of an initially fit model are grossly inadequate, it is probably best to re-estimate σ^2 when a new model is fit.

7. Example

Figure 1 is a graph of the logarithms of 519 daily observations of the power flux density (10^{-22} (watt) $(\text{meter})^{-2}$ $(\text{cycle/second})^{-1}$) of 2800 MHz. solar radio noise. The observations were recorded at National Research Council, Ottawa, Canada, from April 4, 1967, to October 3, 1968.

Let x_n be the first differences multiplied by 100. It is desired to have predictions 1 and 2 steps ahead. Table 1 shows the result of analyses of residuals after fitting 0 through 4-th order autoregressions. $\hat{\sigma}^2$ was formed from the residuals of the 2-nd order model. Using the 2-nd order model, the 1-step and 2-step mean square prediction errors are 15.25 and 18.25; the 1-step error could be reduced by 3.2% and the 2-step error by no more than 5.6%. This model was judged adequate for predicting 1 and 2 steps ahead.

Figure 2 is an estimate of the spectrum of the residuals of the 2-nd order model. It was gotten by using the Fast Fourier Transform to calculate the Fourier coefficients, hanning, squaring, and then averaging in blocks of 3. The most noticeable nonwhite feature is the spike at the frequency .041 which is due to the rotation of the sun. (The phenomenon is, however, not a line in the spectrum.) The results of the analysis of residuals show that finding a new model that eliminates this feature would result in only a slight reduction in the mean square errors of the 1-step and 2-step predictions.

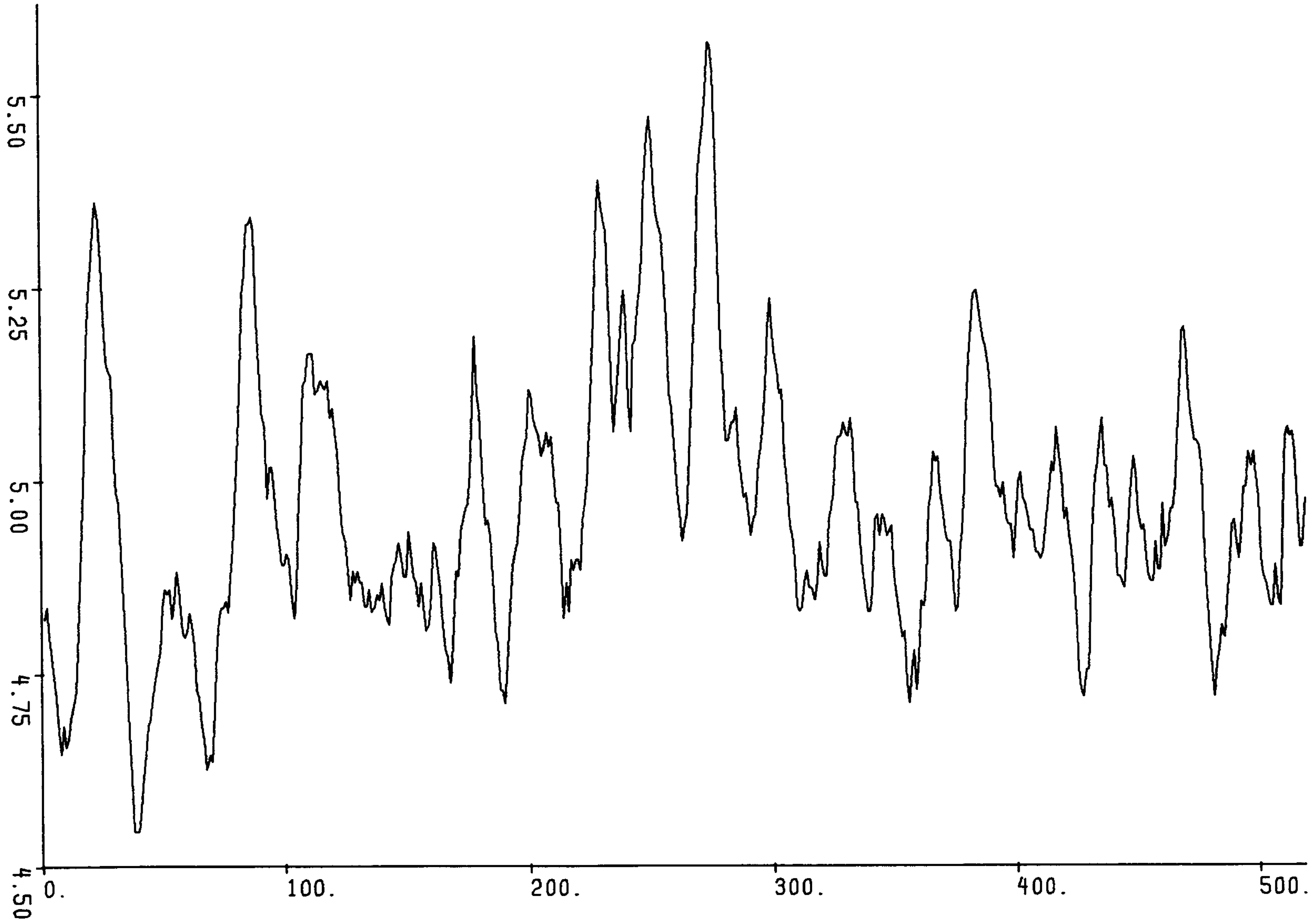

FIGURE 1

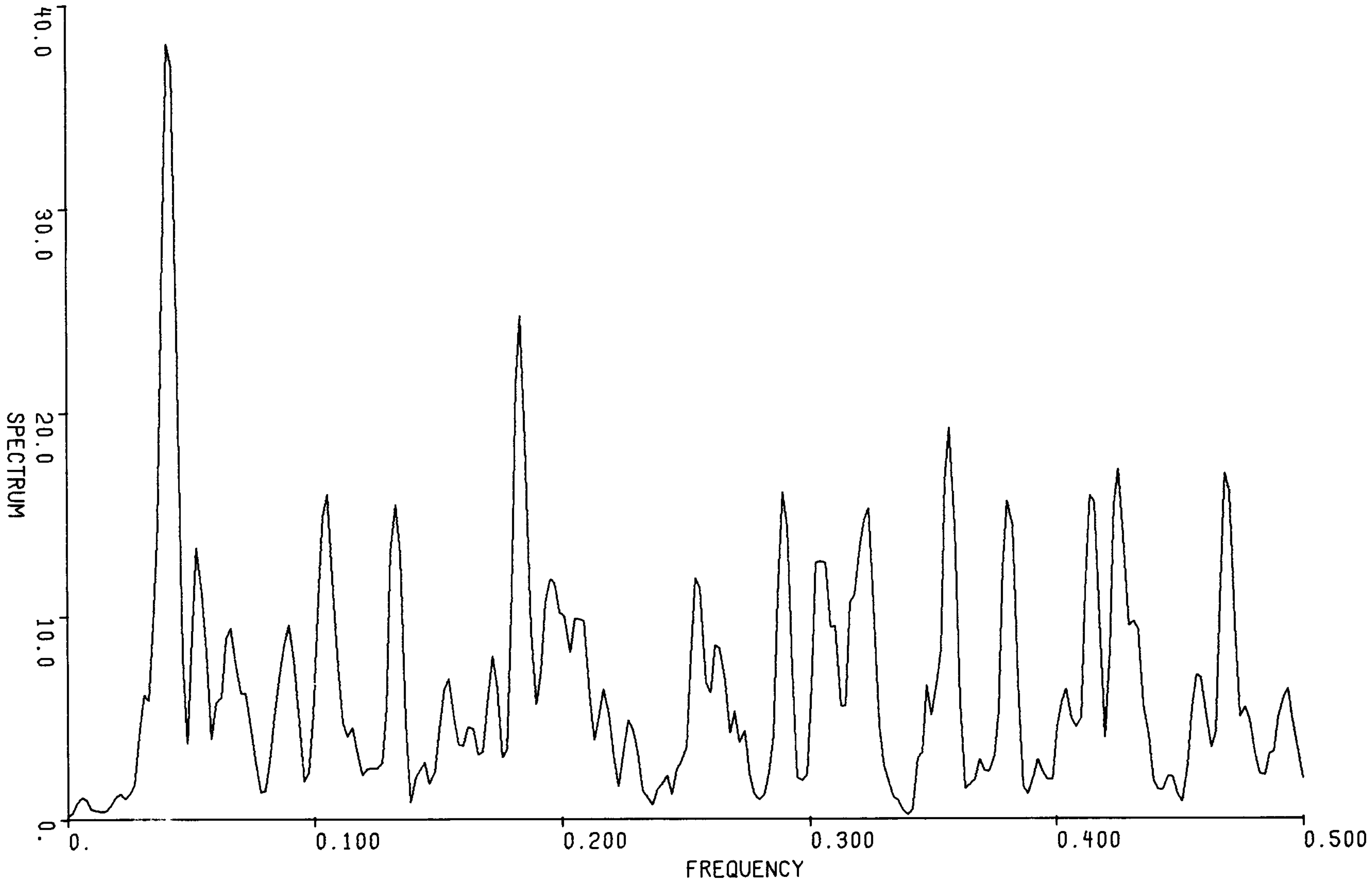
40.0
30.0
20.0
10.0
0.
SPECTRUM
0.
0.100
0.200
0.300
0.400
0.500
FREQUENCY

FIGURE 2

Table 1

Order of Autoregression	$\hat{E}_1$	$\hat{E}_2$	$\hat{B}_1/\hat{E}_1$	$\hat{B}_2/\hat{E}_2$
0	20.52	20.52	.281	.281
1	15.45	18.47	.045	.085
2	15.25	18.25	.032	.056
3	15.28	18.28	.034	.059
4	15.28	18.23	.034	.059

8. Topics for further study

More experience is needed with the estimate $\hat{\sigma}^2$. Other possibilities should be investigated. It is

$$\exp \int_0^1 \mathcal{L}og\, h(\lambda)\, d\lambda$$

that one is interested in estimating, and from the numerical analytic point of view, approximating an integral by a sum at a large number of equally spaced points is crude. Thus other forms of numerical integration might be used. This must be weighed with the fact that $I(\lambda)$ rather than $h(\lambda)$ is available. Another question is whether to first multiply the residuals by cosine bells (Tukey, 1967) in this situation.

There remains the problem of generalizing to multivariate time series this technique of analyzing residuals for prediction. If it is desired to have the p-step predictions of both variables in a bivariate time series, then a decision must be made how to measure errors. Two mean square errors might be considered or perhaps one generalized mean square error.

Acknowledgments

I would like to thank Professor Leonard J. Savage for many helpful suggestions, Mr. Isadore Enger for providing me with the solar flux data, and Miss Maxine Weinstein for her programming assistance.

Appendix

Theorem: If the spectral density f satisfies

$$(11)\qquad f \text{ is bounded and } \int_0^1 f^{-1}(\lambda)d\lambda < \infty,$$

then (3) and (9) hold, and

$$\exp\int_0^1 \ell og\, h(\lambda)d\lambda = \exp\int_0^1 \ell og\, f(\lambda)d\lambda.$$

Proof: Since

$$\int_0^1 |\pm\ell og\, f(\lambda)|d\lambda = \int_{f(\lambda)\geq 1} \ell og\, f(\lambda) + \int_{f^{-1}(\lambda)\geq 1} \ell og\, f^{-1}(\lambda)d\lambda$$

$$\leq \int_0^1 f(\lambda)d\lambda + \int_0^1 f^{-1}(\lambda)d\lambda$$

both f and f^{-1} have canonical factorizations (Doob, 1953, P.577)

$$f(\lambda) = \sigma^2|1 + \sum_{k=1}^{\infty} b_k e(k\lambda)|^2 = \sigma^2|b(\lambda)|^2$$

and

$$f^{-1}(\lambda) = \sigma^{-2}|1 + \sum_{k=1}^{\infty} a_k e(k\lambda)|^2 = \sigma^{-2}|a(\lambda)|^2 .$$

If f_k are the Fourier coefficients of $\frac{1}{2}\ell og\,\sigma^{-2}f(\lambda)$ then $-f_k$ are those of $\frac{1}{2}\ell og\,\sigma^2 f^{-1}(\lambda)$. Thus

$$(12)\qquad 1 + \sum_{k=1}^{\infty} a_k z^k = \exp\,(2\sum_{k=1}^{\infty} f_k z^k)$$

$$= (1 + \sum_{j=1}^{\infty} b_j z^j)^{-1}.$$

Letting L denote Lebesgue measure then

$$a(\lambda) = \lim_{r \to 1-} (1 + \sum_{k=1}^{\infty} a_k r^k e(k\lambda)) \quad \text{a.e.} \quad L .$$

(Grenander and Szego, 1958, p.25). Thus

$$\begin{aligned} a(\lambda) &= \lim_{r \to 1-} (1 + \sum_{k=0}^{\infty} b_k \gamma^k e(k\lambda))^{-1} \\ &= b^{-1}(\lambda), \end{aligned}$$

using (12). Therefore

$$e(N\lambda)\ \overline{a(\lambda)}$$

is the Stone-Kolmogorov isomorph in $L_2(f)$ of $x_N - \phi_{N,1}$. (Doob, 1953, p.575). Now the partial sums of $1+\Sigma_{k=1}^{\infty} a_k e(-k\lambda)$ converge in $L_2(L)$ to $\overline{a(\lambda)}$, therefore since f is bounded, convergence occurs also in $L_2(f)$. Thus (from the Stone-Kolmogorov isomorphism)

$$x_N + \sum_{k=1}^{\infty} a_k x_{N-k} = \varepsilon_N$$

where the sum converges in mean square. Since ε_N is the 1-step prediction error it is orthogonal to $x_{N-1}, x_{N-2}, \dots$.

$\hat{f}$ is a nonzero rational function so that the partial sums of

$$e(N\lambda)\ (1 + \sum_{k=1}^{\infty} \hat{a}_k e(-k\lambda))$$

converge uniformly and therefore in $L_2(f)$, which means the infinite sum in (4) converges in mean square. This then implies that the infinite sum in (3) also converges in mean square.

Let Q be the projection operator onto the closed linear space spanned by $x_N, x_{N-1}, \ldots$. Then $\phi_{N,p} = Qx_{N+p}$. Since Q is linear and continuous applying it to both sides of

$$x_{N+p} + \sum_{k=1}^{\infty} a_k x_{N+p-k} = \varepsilon_{N+p}$$

yields

$$\phi_{N+p} + \sum_{k=1}^{p-1} a_k \phi_{N,p-k} + \sum_{k=p}^{\infty} a_k x_{N+p-k} = 0.$$

Thus if the predictions are to be calculated treating $\hat{f}$ as the true density, the result is (3).

Again since Q is linear and continuous, applying it to

$$\sum_{k=0}^{\infty} \hat{a}_k x_{N+p-k} = \varepsilon_{N+p}$$

yields

$$\sum_{k=0}^{p-1} \hat{a}_k \phi_{N,p-k} + \sum_{k=p}^{\infty} \hat{a}_k x_{N+p-k} = Q\hat{\varepsilon}_{N+p}.$$

Subtracting equation (3) from this equation gives

$$\sum_{k=0}^{p-1} \hat{a}_k (\phi_{N,p-k} - \hat{\phi}_{N,p-k}) = Q\hat{\varepsilon}_{N+p} .$$

Solving recursively,

$$\phi_{N,p} - \hat{\phi}_{N,p} = \sum_{k=0}^{p-1} \hat{b}_k Q\hat{\varepsilon}_{N+p-k} .$$

Thus

(13) $$E(\phi_{N,p}-\hat{\phi}_{N,p})^2 \le \left[\sum_{k=0}^{p-1} |\hat{b}_k| (E(Q\hat{\varepsilon}_{N+p})^2)^{\frac{1}{2}}\right]^2$$

by the Minkowski inequality. From (1) $\Sigma_{k=0}^{\infty} \hat{a}_k z^k \neq 0$ for $|z| \le 1$; thus from (Robinson, 1962, p.110) the closed linear space spanned by $\hat{\varepsilon}_N, \hat{\varepsilon}_{N-1}, \ldots$ is the same as that spanned by $x_N, x_{N-1}, \ldots$ so that $Q\hat{\varepsilon}_{N+p}$ is the prediction of $\hat{\varepsilon}_{N+p}$ from its own infinite past. Thus $E(Q\hat{\varepsilon}_{N+p})^2$ decreases with p. This fact together with (13) gives,

$$E(\phi_{N,p} - \hat{\phi}_{N,p})^2 \le E(Q\hat{\varepsilon}_{N+1})^2 \left\{\sum_{k=0}^{p-1} |\hat{b}_k|\right\}^2 .$$

Now

$$E(Q\hat{\varepsilon}_{N+1})^2 = E(\hat{\varepsilon}_{N+1}^2) - E(\hat{\varepsilon}_{N+1} - Q\hat{\varepsilon}_{N+1})^2 .$$

Since the last term is the mean square error of predicting $\hat{\varepsilon}_{N+1}$ from its infinite past it is equal to

$$\sigma^2 = \exp \int_0^1 \ell og\, h(\lambda)\, d\lambda.$$

(Doob, 1953, p.576-7). Thus

$$E(Q\hat{\varepsilon}_{N+1})^2 = v - \sigma^2$$

and (9) holds.

From (3) and (4)

$$\hat{\varepsilon}_{N+1} - Q\hat{\varepsilon}_{N+1} = x_{N+1} - Qx_{N+1}$$

so that also

$$\sigma^2 = \exp \int_0^1 \ell og\, f(\lambda)\, d\lambda.$$

References

Anscombe, F.J. (1961) Examination of residuals. *1 Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press.

Anscombe, F.J. and Tukey, J.W. (1963) The examination and analysis of residuals. *Technometrics* 5, 141-159.

Bartlett, M.S. (1966) *Stochastic Processes.* University Press, Cambridge.

Box, G.E.P. and Jenkins, G.M. (1966) Models for prediction and control. *Technical Reports 94, 95, and 99, University of Wisconsin.*

Box, G.E.P. and Jenkins, G.M. (to be published) *Models for Prediction and Control.* Holden Day, New York.

Box, G.E.P. and Jenkins, G.M. (1967) Models for forecasting seasonal and non-seasonal time series. *Spectral Analysis of Time Series,* ed. B. Harris. Wiley, New York.

Davis, H.T. and Jones, R.H. (1968) Estimation of the innovation variance of a stationary time series. *Jour. of the Amer. Stat. Assoc.* 63, 141-149.

Doob, J.L. (1953) *Stochastic Processes.* Wiley, New York.

Grenander, U. and Rosenblatt, M. (1957) *Statistical Analysis of Stationary Time Series.* Wiley, New York.

Grenander, U. and Szego, G. (1958) *Teoplitz Forms and Their Applications.* University of California Press, Los Angeles.

Parzen, E. (1968) Statistical spectral analysis (single channel case) in 1968. *Technical Report No. 11, Department of Statistics, Stanford University.*

Robinson, E.A. (1962) *Random Wavelets and Cybernetics Systems.* C. Griffin, London.

Tukey, J.W. (1967) An introduction to the calculations of numerical spectral analysis. *Spectral Analysis of Time Series,* ed. B. Harris. Wiley, New York.

Whittle, P. (1952) Tests of fit in time series. *Biometrika* 39, 309-318.

Whittle, P. (1954) The statistical analysis of a seiche record. *Sears Foundation Journal of Marine Research* 13, 76-100.

Whittle, P. (1963) *Prediction and Regulation.* Van Nostrand, Princeton.