ARKIV FÖR MATEMATIK Band 2 nr 23

Communicated 10 September 1952 by H. CRAMÉR and H. FAXÉN

Estimation and information in stationary time series

By P. WHITTLE

Summary

Section (1) is devoted to a discussion of the model-fitting problem, which finds its explicit solution in equation (1.13). In section (2) the maximum likelihood, (ML), estimates of the model parameters are investigated, and for the class of series considered shown to possess the same optimum properties as in the case of independent series. Next, the covariance matrix of the parameter estimates is expressed in terms of the spectral function of the generating process (eq. 3.7). The last section is concerned with certain working approximations to the ML statistics.

(1) The ultimate objects of any time series analysis are rarely more than two in number: firstly, to estimate, for its own interest, the stochastic relation generating the terms of the series, and secondly, to obtain a forecast by the use of this relation. If the spectrum of the process is known, then both of these problems may be solved by existing methods, under the assumption that the stochastic relation is a linear one (refs. 3, 10). Thus, if we limit ourselves to the case of linearity, the analysis is reduced to the estimation of the spectrum. The word "estimation" is here used in a fairly wide sense, since we require to estimate a function, not merely a finite set of parameters. In general we must specify a definite kind of function, and it is this necessity which compels the analyst to use some sort of test or decision procedure. However, we shall consider this aspect of the problem only in passing, since it has already been treated in ref. 9.

Suppse, then, that we have a time series of N equidistant observations; $x_1, x_2, \ldots x_N$, forming an $N \times 1$ vector X. We shall assume that these observations constitute a part realisation of a real, discrete, stationary process, and our aim is then to obtain the best possible estimate of this process. As above, the generating process will for our purposes be considered as determined when we known its spectrum, $F(y)$, defined by (see ref. 2)

$$v_s = \operatorname{cov}(x_t, x_{t+s}) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} e^{isy}\, dF(y). \tag{1.1}$$

If the spectrum is differentiable, we may define the *spectral function* $A(z)$ by

$$A(e^{iy}) = \frac{\partial F(y)}{\partial y} = \sum_s v_s e^{isy}. \tag{1.2}$$

In this paper we shall limit ourselves to the case when both $A(z)$ and $[A(z)]^{-1}$ exist on the unit circle, which implies amongst other things that the spectrum is everywhere differentiable and increasing. Most of the calculations refer to the case of a Gaussian series, although this limitation may be relaxed somewhat afterwards. The degree of approximation is such that corrections for the mean do not affect the different formulae, and we may thus suppose that X has zero mean.

The author has shown (ref. 9) that the likelihood of X is under these assumptions asymptotically equal to

$$f(x) = (2\pi A)^{-\frac{N}{2}} \exp \frac{X'[A(W)]^{-1}X}{2} \tag{1.3}$$

where A is determined by the equation

$$\log A = \frac{1}{2\pi i}\int_c \log A(z)\frac{dz}{z} \tag{1.4}$$

and W is the circulant matrix

$$W = \begin{bmatrix} \cdot & 1 & \cdot & \cdots & \cdot \\ \cdot & \cdot & 1 & \cdots & \cdot \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdot & \cdot & \cdot & \cdots & 1 \\ 1 & \cdot & \cdot & \cdots & \cdot \end{bmatrix}$$

"c" in (1.4) is the unit circle in the z-plane, positively described.

A may be interpreted as normalised scale factor, or the variance of an appropriately normalised disturbance variate, ε_t. For (see ref. 11), the x process may be represented

$$x_t = \varepsilon_t + a_1\varepsilon_{t-1} + a_2\varepsilon_{t-2} + \cdots \tag{1.6}$$

where the ε_t's are uncorrelated, have zero mean and variance σ^2. Furthermore, we know that on $|z| = 1$

$$A(z) = \sigma^2 a(z)\, a(z^{-1}) = \sigma^2 e^{b(z)+b(z^{-1})} \tag{1.7}$$

where $b(z) = \sum b_j z^j = \log \sum a_j z^j$. From (1.4), (1.7) we see that $A = \sigma^2 e^{2b_0} = \sigma^2$ since $b_0 = \log(a_0) = 0$. Thus we see that A is the variance of the disturbance variate ε_t when ε_t is measured in such a scale that the leading coefficient in (1.6) is unity. This is plainly the case whatever the distribution of the sample variate x, since the representation (1.6) is valid generally.

We shall also define a normalised spectral function $M(z) = A(z)/A$. If besides A the model contains parameters $\theta_1, \theta_2 \ldots \theta_p$, then $M(z)$ is a function of these parameters alone, and does not depend upon A. It is further obvious from equation (1.4) that

$$\frac{1}{2\pi i}\int_c \log M(z)\frac{dz}{z} = 0 \tag{1.8}$$

identically in $\theta_1, \theta_2 \ldots \theta_p$.

We shall now use the criterion of maximum likelihood to obtain an estimate of $A(z)$. To motivate this choice, we recall that the ML criterion leads in the case of an independent series to estimates having certain optimum properties, which, it will be shown, are in most cases shared by the estimates arrived at in the present case. Furthermore, the maximum likelihood provides an asymptotically most powerful test function for discriminating between certain hypotheses (see refs. 7, 8 and 9).

Maximising the likelihood, then, with respect to A and $M(z)$, we obtain the equation

$$\frac{X'[M(W)]^{-1}X}{N} = \min. = \hat{A} \tag{1.9}$$

where $\hat{A}$ is the ML estimate of A. The maximum likelihood, $\hat{f}(x)$, is further given by

$$\hat{f}(x) = (2\pi e \hat{A})^{-\frac{N}{2}} \tag{1.10}$$

which is a function only of $\hat{A}$, so far as the observations are concerned.

We shall write the fitting equation (1.9) somewhat differently. Let

$$\lambda(z) = \sum_s C_s z^s \tag{1.11}$$

where C_s is the circular autocovariance of lag s, $(1/N)\, X' W^s X$, although to this degree of approximation it could as well be the ordinary covariance, $(1/N) \sum_1^{N-s} x_t x_{t+s}$. Then (1.9) may be written

$$\frac{1}{2\pi i} \int_c \frac{\lambda(z)}{M(z)} \frac{dz}{z} = \min. = \hat{A} \tag{1.12}$$

where $M(z)$ is conditioned by (1.8). Combining (1.8) and (1.12), we see that the essential fitting equation is

$$\frac{1}{2\pi i} \int_c \left[\frac{\lambda(z)}{M(z)} + \frac{1}{\alpha} \log M(z) \right] \frac{dz}{z} = \min. \tag{1.13}$$

where α^{-1} is a Lagrangian multiplier.

This equation was deduced for the Gaussian process on the criterion of maximum likelihood, but we would obviously have obtained the same relation, whatever the distribution of the individual sample variate, if the criterion had instead been that of least squares, i.e., that the normalised variance of the disturbing variates be a minimum. We should still require, however, that $A(z)$ and $[A(z)]^{-1}$ exist on $|z| = 1$.

If we now minimise the expression in (1.13) freely with respect to the function $M(z)$, then the calculus of variations gives the solution

$$\frac{\partial}{\partial M} \left[\frac{\lambda}{M} + \frac{1}{\alpha} \log M \right] = 0$$

or $M(z)=\alpha\lambda(z)$, scarcely a surprising result. However, our minimisation has in general only a few degrees of freedom: that is, $M(z)$ is a specific function in which only the parameters $\theta_1, \theta_2, \ldots \theta_p$ are free to vary. (1.13) implies then that

$$\frac{\partial}{\partial\theta_j}\int\left[\frac{\lambda}{M}+\frac{1}{\alpha}\log M\right]\frac{dz}{z}=0 \tag{1.14}$$

which equations will lead to the ML estimates of $\theta_1, \theta_2, \ldots \theta_p$, if only the differential coefficients of $M(z)$ with respect to these parameters exist on $|z|=1$.

One of the special cases of greatest interest is that in which $\Lambda(z)$ is rational in z, so that we may write

$$\Lambda(z)=\gamma\frac{\prod_j(z-\alpha_j)(z^{-1}-\alpha_j)}{\prod_j(z-\beta_j)(z^{-1}-\beta_j)} \qquad |\alpha_j|\leqq 1,\ |\beta_j|<1 \tag{1.15}$$

(see ref. 4). In this case (1.4) shows that $A=\gamma$, so that equation (1.12) takes the form

$$\frac{1}{2\pi i}\int\lambda(z)\frac{\prod(z-\beta_j)(z^{-1}-\beta_j)}{\prod(z-\alpha_j)(z^{-1}-\alpha_j)}=\min. \tag{1.16}$$

where minimisation refers to the parameters α_j, β_j. However, if $[\Lambda(z)]^{-1}$ is to be developable in powers of z on $|z|=1$, we must require that $|\alpha_j|<1$ for all j, i.e., that $\Lambda(z)$ have no zeros on the unit circle.

(2) The theory of ML estimators has been developed almost exclusively upon the assumption that the sample consists of a series of independent observations, all of which are distributed in a like fashion. With some minor restrictions upon the common distribution function, it has been shown that the ML estimator is consistent, is efficient at least in the limit, and is in the limit normally distributed (ref. 5, and see ref. 6 for a survey and bibliography of the subject). By "in the limit" is meant "for samples of infinite size". However, whatever intuition would have us believe, it is by no means obvious that these properties are conserved when the sample variates are no longer independent, e.g., in the case of a time series.

Confining ourselves as hitherto to the Gaussian series, we shall prove in this section that if $\Lambda(z)$, $[\Lambda(z)]^{-1}$, and $\frac{\partial}{\partial\theta}\Lambda(z)$ exist on $|z|=1$, and if $\frac{\partial}{\partial\theta}\Lambda(z)$ is continuous in θ at least in the neighbourhood of the true value, θ_0, then the ML estimate of θ_0, $\hat{\theta}$, is consistent and asymptotically efficient.

We shall use the author's result (ref. 9) that the cumulants of a linear function of the autocovariances $X'g(W)X$ are asymptotically given by

$$K_j=\frac{N2^{j-1}(j-1)!}{2\pi i}\int_c[g(z)\Lambda(z)]^j\frac{dz}{z} \tag{2.1}$$

if $g(z)$ exists on $|z|=1$, and has a symmetric Laurent expansion.

Now, by the above assumptions

$$N\psi(\theta) = -2A\frac{\partial}{\partial\theta}\log f(x) = X'\frac{\partial}{\partial\theta}[M(W)]^{-1}X \tag{2.2}$$

is just such a function, and $\psi(\theta)$ has mean, $\frac{A}{2\pi i}\int\left(-\frac{M_0M'}{M^2}\right)\frac{dz}{z}$, and variance $\frac{A^2}{\pi i N}\int\left(\frac{M_0M'}{M^2}\right)^2\frac{dz}{z}$. The subscript $_0$ indicates the fact that $\theta=\theta_0$. We see that $E[\psi(\theta_0)]$ is zero, since

$$-\int\frac{M_0'}{M_0}\frac{dz}{z} = -\frac{\partial}{\partial\theta_0}\int\log M_0\frac{dz}{z} = 0 \tag{2.3}$$

and also that it is increasing, as a function of θ, since

$$\left[\frac{\partial}{\partial\theta}\int\left(-\frac{M_0M'}{M^2}\right)\frac{dz}{z}\right]_{\theta=\theta_0} = \left[\frac{\partial}{\partial\theta}\int\left(-\frac{M_0M'}{M^2}\right)\frac{dz}{z}+\frac{\partial}{\partial\theta}\int\frac{M'}{M}\frac{dz}{z}\right]_{\theta=\theta_0} =$$

$$= \int\left(\frac{M_0'}{M_0}\right)^2\frac{dz}{z} > 0. \tag{2.4}$$

From this and the fact that the variance is of order $1/N$, it is clear that the probability that $\psi(\theta_0-\varepsilon)$ be negative and $\psi(\theta_0+\varepsilon)$ be positive is asymptotically unity, for arbitrarily small ε. That is, since $\psi(\theta)$ is continuous in the neighbourhood of θ_0, at least one zero of $\psi(\theta)$ corresponding to a maximum of $f(x)$ falls in the interval $\theta_0\pm\varepsilon$ with asymptotically unit probability. Thus, the ML estimate $\hat\theta$ is consistent.

Indeed, since $E[\psi(\theta)]$ is increasing in this neighbourhood, we can see in the same fashion that $\psi(\theta)$ has at least asymptotically only a single zero in the neighbourhood of θ_0. Thus, for large samples

$$\text{Prob}\,[\hat\theta \gtrless \theta] = \text{Prob}\,[\psi(\theta)\lessgtr 0] = \text{const.}\ -\frac{1}{2\pi i}\int_{-\infty}^{+\infty}\Phi(s)\frac{ds}{s} \tag{2.5}$$

where $\Phi(s)$ is the characteristic function of $\psi(\theta)$. Now, from (2.1), (2.2)

$$\Phi(s) = \exp -\frac{N}{4\pi i}\int_c\log\left[1+2isA\frac{M_0M'}{M^2}\right]\frac{dz}{z}. \tag{2.6}$$

Inserting this in (2.5) we have an expression for $\hat\theta$'s distribution function. We note, however, that $\Phi(s)$ is of the form $\Phi(s)=[\phi(s)]^N$, where $\phi(s)$ is a characteristic function independent of N. That is, $\hat\theta$ distributed in the same fashion as if the sample material had consisted of N independent variates with frequency function, $p(x)$, defined by

$$\int p_0 e^{-is\frac{\partial}{\partial\theta}\log p}\,dx = \exp -\frac{1}{4\pi i}\int\log\left[1+2isA\frac{M_0M'}{M^2}\right]\frac{dz}{z}. \tag{2.7}$$

The case of a Gaussian time series has thus been reduced to that of an independent series. With the aid of this equivalence, estimator properties such as efficiency etc., may be established simply by referring back to existing theorems for independent series. The generalisation of (2.7) to the case of several parameters is immediate.

It may be of interest to demonstrate in a direct manner the optimum character of the ML estimate in the present case. Suppose that the estimating relation is

$$\int_c \frac{\lambda(z)\,Q(z)}{M(z)}\,\frac{dz}{z}=0 \tag{2.8}$$

where Q is a function of θ and z which exists together with its θ derivative on $|z|=1$, and its derivative is continuous in θ in the neighbourhood of θ. To ensure consistence we require that the integral in (2.8) have expectation zero for $\theta=\theta_0$, all θ_0, so that

$$\int Q_0(z)\,\frac{dz}{z}=0 \tag{2.9}$$

$$\frac{\partial}{\partial\theta_0}\int Q_0(z)\,\frac{dz}{z}=0. \tag{2.10}$$

We shall now search for that Q which gives $\hat{\theta}$ minimum asymptotic variance. Since the expression in (2.8) is continuous in the neighbourhood of θ_0 and the estimate is consistent, we have that

$$\int\frac{\lambda Q_0}{M_0}\,\frac{dz}{z}+(\hat{\theta}-\theta_0)\,E\left[\frac{\partial}{\partial\theta}\int\frac{\lambda Q}{M}\,\frac{dz}{z}\right]_{\theta=\theta_0}=0\,(N^{-1})$$

or, with the aid of (2.10)

$$\int\frac{\lambda Q_0}{M_0}\,\frac{dz}{z}-(\hat{\theta}-\theta_0)\,\frac{A}{2\pi i}\int\frac{Q_0 M_0'}{M_0}\,\frac{dz}{z}=0\,(N^{-1}). \tag{2.11}$$

Thus, for large samples

$$E(\hat{\theta}-\theta_0)^2=\frac{E\left[\int\frac{\lambda Q_0}{M_0}\,\frac{dz}{z}\right]^2}{\left[\frac{A}{2\pi i}\int\frac{Q_0 M_0'}{M_0}\,\frac{dz}{z}\right]^2}=\frac{\frac{1}{N\pi i}\int Q_0^2\,\frac{dz}{z}}{\left[\frac{1}{2\pi i}\int\frac{Q_0 M_0'}{M_0}\,\frac{dz}{z}\right]^2}. \tag{2.12}$$

We shall now omit the zero subscript for the sake of convenience.

Differentiating (2.12) with respect to Q we see that

$$\int Q\,\frac{dz}{z}\int Q\,\frac{M'}{M}\,\frac{dz}{z}=\int Q^2\,\frac{dz}{z}\int\frac{M'}{M}\,\frac{dz}{z} \tag{2.13}$$

which the Schwarz inequality shows is only satisfied by $Q = \text{const.}\,\frac{M'}{M}$ almost everywhere. But this value of Q corresponds to the ML estimate, which must then have asymptotically minimum variance. Setting $Q = \frac{M'}{M}$ in (2.12), we obtain

$$\text{var}\,(\hat{\theta}) = \frac{1}{\frac{N}{4\pi i}\int \left(\frac{M'}{M}\right)^2 \frac{dz}{z}} = \frac{1}{E\left(\frac{\partial \log f}{\partial \theta}\right)^2} \tag{2.14}$$

anticipating the more general result of the next section. This treatment of the ML statistic will be recognised as analogous to Fisher's, for the case of independent variates (ref. 5).

(3) If $f(x)$ is the likelihood of a series of independent variates, depending upon a number of parameters $\theta_1, \theta_2 \ldots \theta_p$, then the covariance matrix of the ML estimates of these parameters, $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_p$, is $(v_{jk}) = (\alpha_{jk})^{-1}$ where (α_{jk}) is the so-called *information matrix*, whose elements are given by the relation

$$\alpha_{jk} = \int \left(\frac{\partial \log f}{\partial \theta_j}\right)\left(\frac{\partial \log f}{\partial \theta_k}\right) f\, dx. \tag{3.1}$$

However, we have seen from the previous section that the Gaussian process we have been considering is, from the point of view of ML estimation, equivalent to a certain series of independent, like distributed variates. Thus, we may immediately conclude that in this case also the covariance matrix of the estimates is asymptotically equal to the reciprocal of the information matrix, as given by (3.1). To evaluate the elements of the information matrix we shall use the fact that these are second moments of linear functions of the autocovariances, and, as such, may be expressed directly in terms of the spectral function with the help of formula (2.1).

Thus, by (2.1), (2.2)

$$E\left[\frac{\partial \log f(x)}{\partial \theta_j}\right] = \frac{N}{2\pi i}\int A\, M(z)\left[-\frac{1}{2A}\frac{\partial}{\partial \theta}\frac{1}{M(z)}\right]\frac{dz}{z} =$$

$$= \frac{\partial}{\partial \theta}\frac{N}{4\pi i}\int \log M(z)\frac{dz}{z} = 0 \quad (j = 1, 2 \ldots p) \tag{3.2}$$

by equation (1.8). In the same manner

$$E\left[\frac{\partial \log f(x)}{\partial \theta_j}\cdot\frac{\partial \log f(x)}{\partial \theta_k}\right] = \text{cov}\left[\frac{\partial \log f(x)}{\partial \theta_j}, \frac{\partial \log f(x)}{\partial \theta_k}\right] =$$

$$= \frac{N}{4\pi i}\int \frac{\partial \log M(z)}{\partial \theta_j}\cdot\frac{\partial \log M(z)}{\partial \theta_k}\frac{dz}{z}. \tag{3.3}$$

Considering now $\hat{A}$, we see that

$$\frac{\partial \log f(x)}{\partial A} = -\frac{N}{2A} + \frac{X'[M(W)]^{-1}X}{2A^2} \tag{3.4}$$

so that

$$E\left[\frac{\partial \log f(x)}{\partial A}\right] = -\frac{N}{2A} + \frac{1}{2A^2}\frac{N}{2\pi i}\int\left(\frac{A M(z)}{M(z)}\right)\frac{dz}{z} = 0$$

$$\text{var}\left[\frac{\partial \log f(x)}{\partial A}\right] = \frac{1}{4A^4}\frac{2N}{2\pi i}\int\left(\frac{A M}{M}\right)^2\frac{dz}{z} = \frac{N}{2A^2} \tag{3.5}$$

$$\text{cov}\left[\frac{\partial \log f(x)}{\partial A}, \frac{\partial \log f(x)}{\partial \theta_j}\right] = \frac{1}{2A^2}\frac{2N}{2\pi i}\int (A M)^2 M^{-1}\left[-\frac{1}{2A}\frac{\partial M^{-1}}{\partial \theta_j}\right]\frac{dz}{z} = 0 \tag{3.6}$$

$$(j = 1, 2 \ldots p)$$

(3.6) tells us already that $\hat{A}$ is uncorrelated with $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_p$. Further, from (3.1), (3.5) we see that it has variance $2A^2/N$, a result which may also be derived from (1.9), (2.1). Thus, $\hat{A}$'s sampling variance depends only upon A. Similarly, the covariance matrix of $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_p$ may be expressed independently of A or $\hat{A}$ as

$$\frac{2}{N}\left[\frac{1}{2\pi i}\int\frac{\partial \log M(z)}{\partial \theta_j}\cdot\frac{\partial \log M(z)}{\partial \theta_k}\frac{dz}{z}\right]^{-1} \quad (j = 1, 2 \ldots p). \tag{3.7}$$

This is the required expression in terms of the spectral function.

(3.7) takes a particularly simple form if

$$M(z) = |z^p + \theta_1 z^{p-1} + \cdots + \theta_p|^{\pm 2} = |P(z)|^{\pm 2} \tag{3.8}$$

i.e., if the model is an autoregressive or moving average scheme, which we shall assume chosen so that $P(z)$ has real coefficients and all its roots inside the unit circle. In this case

$$\frac{N}{4\pi i}\int\frac{\partial \log M}{\partial \theta_j}\cdot\frac{\partial \log M}{\partial \theta_k}\frac{dz}{z} = \frac{N}{4\pi i}\int\left[\frac{z_j}{P} + \frac{z^{-j}}{\overline{P}}\right]\left[\frac{z^k}{P} + \frac{z^{-k}}{\overline{P}}\right]\frac{dz}{z} = \frac{N}{2\pi i}\int\frac{z^{j-k}}{|P(z)|^2}\frac{dz}{z} \tag{3.9}$$

the coefficient of z^{j-k} in the Laurent expansion of $N|P(z)|^{-2}$. It is rather remarkable that although the estimates of $\theta_1, \theta_2 \ldots \theta_p$ are obtained in such different manners, depending upon whether the scheme is a moving average or an autoregressive one (in the first case a function of all autocovariances, in the second an explicit function of the first $p + 1$), yet (3.9) indicates that the second moments of the resulting estimates are of identical form. This is another evidence of the symmetry between the two schemes.

The results of this and the preceding section have been obtained on the assumption of a Gaussian series: but we may readily see that many of them are of much wider validity. Bartlett has shown (ref. 1) that the first and second moments of the sample autocorrelation coefficients of a series generated from a linear process are for large samples independent of the distribution of the sample variate. Thus, all proofs which depend only upon a discussion of the first and second moments of linear functions of the autocorrelations will hold also in the more general case. In this way we see that estimates arrived at with the help of equation (1.13) have the following general properties:

(a) they are consistent,
(b) they have asymptotically minimum variance among the class of estimates given by 2.8,
and (c) their variances and covariances are the same as those arrived at in this section for the ML estimates of a Gaussian series.

Of course, the same restrictions apply to the spectral function as before. Also, it is not probable that the distribution of the sample variate is entirely arbitrary, but we shall not enter in on this question here.

Note that (b) states merely that the least square estimates have asymptotically least variance among the class of estimates provided by (2.8). In any particular non-Gaussian case, there is always for example the ML estimate, which will not be given by a relation of form (2.8), and which may very well have a variance asymptotically less than that of the least square estimate.

(4) One result of the previous section is that the parameters of the process fall into two well defined groups: A and the rest. The ML estimate of A is uncorrelated with those of $\theta_1, \theta_2 \ldots \theta_p$, furthermore, the variance of $\hat{A}$ depends only upon A, the covariance matrix of $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_k$ only upon $\theta_1, \theta_2 \ldots \theta_p$. We could perhaps express this by saying that the information matrix of the process may be partitioned into two non-overlapping square matrices, one of which gives information on the disturbing variate, while the other gives information on the stochastic relation generating the series.

However, while it is the values of $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_p$ which determine the model finally, it is the value of $\hat{A}$ which measures the *plausibility* of the model thus obtained. As we saw in section (1), the maximum likelihood is a function only of $\hat{A}$ (we restrict ourselves again to the Gaussian case). For different choices of model (i.e., for different kinds of function $A(z)$), one obtains different values of $\hat{A}$, and according to the theory of refs. 7, 8, 9 the best test function for discriminating between the two models is the quotient of the corresponding $\hat{A}$ statistics.

Moreover, $\hat{A}$ statistics corresponding to different models are directly comparable, a fact of value in the first stages of an analysis. That is, if we obtain values $\hat{A}_1$ and $\hat{A}_2$ for hypotheses H_1 and H_2, and $\hat{A}_1 < \hat{A}_2$, then we may immediately conclude that H_1 is more plausible than H_2, insomuch as that H_1 "explains away" a greater proportion of the total variance than H_2. This is obvious, if we consider the interpretation of A as the variance of the normalised disturbance variate, and the fact that the moments of $\hat{A}$ depend asymptotically only upon A.

We see from the above that the only statistics of direct interest in the early part of an analysis are the $\hat{A}$ statistics, but these cannot be calculated without first calculating $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_p$. This is regretable, since the solution of (1.13) can be very laborious. One is, of course, willing to take some trouble in calculating the parameter estimates of a model which one is convinced is best, but it is a tedious task to reckon out the ML estimates of a number of models which are only of a preliminary, hypothetical nature. What would be desirable, then, is a direct method of calculating $\hat{A}$ without the intermediate necessity of calculating $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_p$. One is tempted, for example, to set simpler estimates of $\theta_1, \theta_2 \ldots \theta_p$ in the expression

$$\frac{1}{N} X' [M(W)]^{-1} X$$

and to use this instead of $\hat{A}$. A few practical examples are enough to show, however, that such a procedure may be attended with severe error, not to mention the fact that property (1.10) is now lost.

Example. Consider a first order moving average scheme with spectral function $A(z-\theta)(z^{-1}-\theta)$. The ML estimate of θ is given by the equation

$$\frac{\partial}{\partial\theta}\left[\frac{1}{1-\theta^2}[C_0+2\theta C_1+2\theta^2 C_2+\cdots]\right]=0$$

and we can also obtain a simple estimate θ^* from

$$r_1=\frac{c_1}{c_0}=-\frac{\theta^*}{1+\theta^{*2}}.$$

Equation (3.7) shows that $\text{var}\,(\hat{\theta})=\dfrac{1-\theta^2}{N}$, while, using Bartlett's formula for the asymptotic variance of r_1 in ref. 1, we see that

$$\text{var}\,(\theta^*)=\frac{1+\theta^2+4\theta^4+\theta^6+\theta^8}{N(1-\theta^2)^2}.$$

Thus, for $\theta=\frac{1}{2}$, say

$$\frac{\text{var}\,(\theta^*)}{\text{var}\,(\hat{\theta})}=3.8$$

indicating an enormous difference in efficiency. From an artificial experiment with $\theta=0.382$ and $N=150$ the following values are obtained:

$$\hat{\theta}=0.407 \qquad \hat{A}=0.8054$$
$$\theta^*=0.634 \qquad A^*=0.8656.$$

The imprecision of θ^* is obviously reflected in the corresponding value of A, A^*.

We shall now describe a method by means of which the precision of the estimate A^* may be improved without the need to recalculate the estimates of the remaining parameters. To do this, we shall make use of the fact that if $A(z)$ may be expanded in a Taylor series in $\theta_1, \theta_2 \ldots \theta_p$ in the neighbourhood of their true values, then

$$A(\theta)=\frac{X'[M(W)]^{-1}X}{N} \tag{4.1}$$

is approximately parabolic in this neighbourhood, and thus also in that of $\theta_1, \theta_2 \ldots \theta_p$. Thus, since $\theta_1^*, \theta_2^* \ldots \theta_p^*$ lie presumably also in this neighbourhood, we may write

$$A(\theta) \approx A^* + \sum_i A_i^* (\theta_i - \theta_i^*) + \tfrac{1}{2} \sum_i \sum_j A_{ij}^* (\theta_i - \theta_i^*)(\theta_j - \theta_j^*) \tag{4.2}$$

where

$$A_i^* = \left(\frac{\partial A(\theta)}{\partial \theta_i}\right)_{\theta=\theta^*} \qquad A_{ij}^* = \left(\frac{\partial^2 A(\theta)}{\partial \theta_i \, \partial \theta_j}\right)_{\theta=\theta^*}.$$

Regarding the right hand side of (4.2) as a quadratic function of

$$1,\ \theta_1 - \theta_1^*,\ \theta_2 - \theta_2^* \ldots \theta_p - \theta_p^*$$

we see from the well known theorem on the minimum of a positive definite Hermite form that it has a minimum

$$A_{\min} = \frac{\begin{vmatrix} 2A^* & A_1^* & A_2^* & \ldots & A_p^* \\ A_1^* & A_{11}^* & A_{12}^* & \ldots & A_{1p}^* \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ A_p^* & A_{p1}^* & A_{p2}^* & \ldots & A_{pp}^* \end{vmatrix}}{2 \begin{vmatrix} A_{11}^* & A_{12}^* & \ldots & A_{1p}^* \\ A_{21}^* & A_{22}^* & \ldots & A_{2p}^* \\ \ldots & \ldots & \ldots & \ldots \\ A_{p1}^* & A_{p2}^* & \ldots & A_{pp}^* \end{vmatrix}} = A^* - \tfrac{1}{2} P' Q^{-1} P \tag{4.3}$$

where P and Q are respectively the vector and matrix of first and second differential coefficients of $A(\theta)$ at $\theta = \theta^*$. The quantities used to calculate A^* may also be used to calculate A_i^*, A_{ij}^*, with relative ease, so that the calculation times for A^*, $A_{\min}$, are of the same order of magnitude. We see, however, that $A_{\min}$ departs from $\hat{A}$, the ML estimate, only by so much as $A(\theta)$ departs from parabolism in the neighbourhood of $\hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_p$, so that we may expect the estimate of A to have been considerably improved. More precisely, it can be readily shown that if $\theta_i^* - \hat{\theta}_i = 0(N^{\alpha})$, $(i = 1, 2 \ldots p;\ \alpha < 0)$, so that $A^* - \hat{A} = 0(N^{2\alpha})$, then $A_{\min} - \hat{A} = 0(N^{3\alpha})$.

To explicitly calculate the variance of $A_{\min}$ is no easy matter, but numerical examples would seem to indicate that its efficiency is not greatly inferior to that of the ML estimate.

Example. The same as the preceding one. A was calculated from the formula

$$A(\theta) = \frac{1}{1-\theta^2}\left[C_0 + 2 \sum_1^{12} \theta^j C_j\right].$$

With $\theta^* = 0.6343$ we find that $A^* = 0.8656$, $A_{\min} = 0.8105$. Comparing these with $\hat{A} = 0.8054$, we see that $A_{\min}$ has an error of 0.6 %, while A^*'s is 7.4 %. In the case of the autoregressive process, $\hat{A}$ and $A_{\min}$ actually coincide, for if

$$A(z) = A\,|z^p + \theta_1 z^{p-1} + \cdots + \theta_p|^{-2}$$

it may be shown that

$$A(\theta) = \hat{A} + (\theta - \hat{\theta})' \begin{bmatrix} C_0 & C_1 & \dots & C_{p-1} \\ C_1 & C_0 & \dots & C_{p-2} \\ \dots & \dots & \dots & \dots \\ C_{p-1} & C_{p-2} & \dots & C_0 \end{bmatrix} (\theta - \hat{\theta})$$

i.e., $A(\theta)$ *is* parabolic in θ. But, of course, it is in just this case that $\hat{A}$ is explicitly calculable, so that recourse to the method above is unnecessary.

University Institute of Statistics, Uppsala.

REFERENCES

1. M. S. BARTLETT: (1946 J. R. Statist. Soc., *7*, No. 1, 27.
2. H. CRAMÉR: (1942 Arkiv för matemtik, astronomi och fysik, *28 B*, No. 12.
3. C. L. DOLPH and M. A. WOODBURY: Unpublished memorandum, University of Michigan.
4. J. L. DOOB: (1949) Berkeley Symposium, p. 303.
5. R. A. FISHER: (1938) Statistical Theory of Estimation. Calcutta Readership Lectures.
6. M. G. KENDALL: (1946) The Advanced Theory of Statistics, *2*. C. Griffin.
7. E. L. LEHMANN and C. STEIN: (1948) Ann. Math. Stat., *19*, 495.
8. A. WALD: (1943) Trans. Amer. Math. Soc., *36*, 426.
9. P. WHITTLE: (1951) Hypothesis Testing in Time Series Analysis. Uppsala.
10. N. WIENER: (1949) Extrapolation, Interpolation: and Smoothing of Stationary Time Series. M.I.T.
11. H. WOLD: (1938) A Study in the Analysis of Stationary Time Series. Uppsala.

Tryckt den 14 mars 1953

Uppsala 1953. Almqvist & Wiksells Boktryckeri AB