Biometrika (1973), **60**, 2, *p.* 255
Printed in Great Britain

Maximum likelihood identification of Gaussian autoregressive moving average models

BY HIROTUGU AKAIKE
University of Hawaii and The Institute of Statistical Mathematics

SUMMARY

Closed form representations of the gradients and an approximation to the Hessian are given for an asymptotic approximation to the log likelihood function of a multidimensional autoregressive moving average Gaussian process. Their use for the numerical maximization of the likelihood function is discussed. It is shown that the procedure described by Hannan (1969) for the estimation of the parameters of one-dimensional autoregressive moving average processes is equivalent to a three-stage realization of one step of the Newton–Raphson procedure for the numerical maximization of the likelihood function, using the gradient and the approximate Hessian. This makes it straightforward to extend the procedure to the multidimensional case. The use of the block Toeplitz type characteristic of the approximate Hessian is pointed out.

Some key words: Autoregressive moving average process; Maximum likelihood; Identification; Hessian; Newton–Raphson; Block Toeplitz matrix.

1. INTRODUCTION

By an ingenious approach, Hannan (1969; 1970, pp. 377–95) developed an estimation procedure for the parameters of autoregressive moving average processes. The procedure essentially consists of modifying arbitrary consistent estimates of the parameters into estimates which are asymptotically efficient relative to the maximum likelihood estimates in the Gaussian case. This is accomplished by applying a single numerical step realized in three stages of successive computations. In spite of its asymptotic efficiency, iterative application of the procedure was suggested as a way of further improving the estimates. The extension of the procedure to the general multidimensional vector process case was considered to be too complicated for explicit presentation and only the simple moving average case was treated.

As an alternative to Hannan's intuitive approach, we formulate the problem directly as the identification of a Gaussian model by numerical maximization of the Gaussian likelihood function and we analyze its computational aspects. The evaluation of the gradient and the Hessian of the log likelihood function will then be the main subjects of this paper. It is shown that Hannan's procedure is equivalent to one step of a Newton–Raphson type iterative modification of the initial estimates for the maximization of the Gaussian likelihood function, realized in three successive stages of computations in lower dimensional spaces. This observation reveals the fact that the iterative application of Hannan's procedure can maximize the Gaussian likelihood function only when the initial estimates are close enough to the desired maximum likelihood estimates. Thus, in a practical situation a simple-minded application of Hannan's procedure does not always achieve an improvement of the

initial estimates. By our present approach, the multidimensional case can be treated explicitly and the results clearly explain the related numerical complexity. To cope with this complexity a block Toeplitz-type characterization of an approximation to the Hessian is introduced. This representation is quite useful in reducing the dimension necessary in the computation. When the orders of the autoregression and the moving average are identical, this leads to an extremely simple numerical procedure.

Dzhaparidze (1971) developed a procedure for the estimation of continuous time models. Also, there is a similarity between these procedures and the construction of the basic statistics of Le Cam (1956).

Before going into the technical details, the present status of numerical procedures for fitting the autoregressive moving average model will be reviewed briefly. Besides the procedures discussed in the introduction of Hannan (1969), there are some significant contributions in the engineering literature which are concerned with the maximization of the Gaussian likelihood function. In an unpublished report, Aström, Bohlin and Wensmark discussed the statistical and numerical aspects of the maximum likelihood procedure for a general one-dimensional scalar process model, including the autoregressive moving average model. The numerical procedure was composed for variants of the Newton–Raphson type gradient procedure and the decomposition of the one step of iteration into lower dimensional stages was not considered. Tretter & Steiglitz (1967) fitted autoregressive and moving average coefficients alternately, thus taking advantage of the linearity of the fitting equation for the autoregressive coefficients. But this was also limited to scalar cases and the interaction between the two stages was not taken into account. This deficiency does not appear in Hannan's procedure if it is used properly. A significant contribution to the multidimensional vector case was made by Kashyap (1970), who introduced a systematic procedure for the calculation of the gradient of the likelihood function. Like Tretter and Steiglitz, Kashyap left the task of maximization to available computer programs and did not go into the details of the maximization procedure. It is mentioned by Wilson (1971, p. 520) that an iterative numerical procedure which generalizes the nonlinear least square method of Marquardt (cf. Powell, 1970, p. 95) has been used for the maximization of the Gaussian likelihood function of a general model which includes the autoregressive moving average model as a special case. The examples reported by Kashyap (1970) and Wilson (1971) are restricted to two-dimensional vector cases and the orders of the models are rather low. Simple analysis shows that the effect of dimension on the computational complexity can be serious. Thus without the introduction of some new approach, practical applicability of the autoregressive moving average model is extremely dubious, at least for high-dimensional cases.

As was pointed out by Tretter & Steiglitz (1967), there is another problem, namely the decision on the order of the models to be fitted to real data. A procedure for fitting multidimensional autoregressive models (Akaike, 1971*a*) was successfully applied to real data by Otomo, Nakagawa & Akaike (1972). Although there is hope of extending the same decision procedure to the present case (Akaike, 1972*a*), the multidimensional autoregressive moving average model fitting by maximizing the Gaussian likelihood function must be considered to be still hampered by its enormous numerical complexities. In this paper we shall content ourselves by only providing some of the basic information for the development of future numerical studies.

2. Formulation of the problem

Assume that a set of observations, $\mathbf{y}(n)$ $(n = 1, \ldots, N)$, on a d-dimensional zero mean stationary random process is given. To these data, we fit an autoregressive moving average model

$$\sum_{m=0}^{q} \mathbf{b}(m)\,\mathbf{y}(n-m) = \sum_{m=0}^{p} \mathbf{a}(m)\,\mathbf{x}(n-m), \tag{2·1}$$

where $\mathbf{a}(m)$ and $\mathbf{b}(m)$ are $d \times d$ matrices, $\mathbf{q}(0) = \mathbf{a}(0) = \mathbf{I}_d$, the $d \times d$ identity matrix, $\mathbf{x}(n)$ is a d-dimensional white noise with $E\{\mathbf{x}(n)\} = \mathbf{0}$, a zero vector, and $E\{\mathbf{x}(n)\,\mathbf{x}(m)'\} = \mathbf{0}$, a zero matrix, for $n \neq m$, and $E\{\mathbf{x}(n)\,\mathbf{x}(n)'\} = \mathbf{G}$. We assume that the process is Gaussian and develop the maximum likelihood estimation procedure for the coefficients $\mathbf{a}(m)$, $\mathbf{b}(m)$ and $\mathbf{G}$ under this assumption. When the process is not Gaussian but stationary and ergodic, this procedure will give a Gaussian model which will asymptotically give the best fit to the finite dimensional distributions of the observed process as evaluated by a properly defined mean information of Kullback (1959, p. 5). In this sense we call the procedure maximum likelihood identification rather than maximum likelihood estimation. Since the exact evaluation of the Gaussian likelihood is rather complicated even for a one-dimensional case (Hájek, 1962, p. 432), we assume that the effect of imposing the initial conditions

$$\mathbf{y}(m) = \mathbf{x}(l) = \mathbf{0} \quad (m = 0, -1, \ldots, 1-q;\ l = 0, -1, \ldots, 1-p)$$

is negligible for the evaluation of the likelihood. Under this assumption, the values of $\mathbf{x}(n)$ $(n = 1, \ldots, N)$ can be calculated for a given set of $\mathbf{a}(m)$ and $\mathbf{b}(m)$ by using (2·1) and the set of observations $\mathbf{y}(n)$.

The corresponding likelihood is given by

$$\{(2\pi)^d\,|\mathbf{G}|\}^{-\frac{1}{2}N} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N} \mathbf{x}(n)'\,\mathbf{G}^{-1}\mathbf{x}(n)\right\}, \tag{2·2}$$

where $|\mathbf{G}|$ denotes the determinant of $\mathbf{G}$. Twice the negative of the logarithm of the likelihood is given by

$$L = Nd \log(2\pi) + N \log|\mathbf{G}| + N \operatorname{tr}(\mathbf{C}_0\,\mathbf{G}^{-1}), \tag{2·3}$$

where

$$\mathbf{C}_0 = (1/N) \sum_{n=1}^{N} \mathbf{x}(n)\,\mathbf{x}'(n). \tag{2·4}$$

Since the dependence of $\mathbf{C}_0$ on the parameters is highly nonlinear, it is not easy to develop an insight into the minimization of L. By using Fourier transforms an explicit expression of this dependence can be obtained, and for this we need another approximation. Assuming $\mathbf{y}(n) = \mathbf{0}$ for $n \leqslant 0$ and $n > N$, we define the Fourier transform of $\mathbf{y}(n)/\sqrt{N}$ by

$$\mathbf{Y}(f) = (1/\sqrt{N}) \sum_{n=-\infty}^{\infty} \exp(-i2\pi f n)\,\mathbf{y}(n).$$

The Fourier transforms of $\mathbf{a}(m)$ and $\mathbf{b}(m)$ are defined by

$$\mathbf{A}(f) = \sum_{m=0}^{p} \exp(-i2\pi f m)\,\mathbf{a}(m),$$

$$\mathbf{B}(f) = \sum_{m=0}^{q} \exp(-i2\pi f m)\,\mathbf{b}(m).$$

We assume that for the evaluation of $\mathbf{C}_0$ the Fourier transform $\mathbf{X}(f)$ of $\mathbf{x}(n)/\sqrt{N}$, analogously defined to $\mathbf{Y}(f)$, can be replaced by

$$\mathbf{X}(f) = \{\mathbf{A}(f)\}^{-1}\mathbf{B}(f)\mathbf{Y}(f). \tag{2·5}$$

This assumption is equivalent to neglecting the effect of the transient reponse of the filter specified by $\{\mathbf{A}(f)\}^{-1}\mathbf{B}(f)$, at $n > N$. Under the present assumption (2·4) is replaced by

$$\mathbf{C}_0 = \int_{-\frac{1}{2}}^{\frac{1}{2}} \mathbf{X}(f)\mathbf{X}(f)^* df, \tag{2·6}$$

where * denotes conjugate transpose. Hereafter the limits of integration are always $-\frac{1}{2}$ and $\frac{1}{2}$ and they are omitted. Also the argument f, which denotes the frequency, is sometimes omitted. Thus (2·5) may be expressed in the form $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}\mathbf{Y}$.

Our problem is now to minimize L defined by (2·3), (2·5) and (2·6) with respect to $\mathbf{a}(m)$, $\mathbf{b}(m)$ and $\mathbf{G}$. This formulation may be considered to be a special case of the least squares parameter estimation developed by Whittle (1953). Our derivation clearly shows that the transient response of $\mathbf{A}^{-1}\mathbf{B}$ plays a definite role in determining the validity of the approximations. In the following discussion it is tacitly assumed that both $\mathbf{A}$ and $\mathbf{B}$ are invertible, in the sense that the elements of $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ are the Fourier transforms of some absolutely convergent sequences which take only zeros on the negative time axis.

3. MAXIMIZING THE LIKELIHOOD

When a matrix $\mathbf{F}$ is nonsingular and its elements are functions of a set of parameters θ_i, it is not hard to verify the following relations (e.g. Dwyer, 1967):

$$\frac{\partial}{\partial\theta_i}\log|\mathbf{F}| = \mathrm{tr}\left(\frac{\partial\mathbf{F}}{\partial\theta_i}\mathbf{F}^{-1}\right), \tag{3·1}$$

$$\frac{\partial}{\partial\theta_i}\mathbf{F}^{-1} = -\mathbf{F}^{-1}\frac{\partial\mathbf{F}}{\partial\theta_i}\mathbf{F}^{-1}, \tag{3·2}$$

where the derivative of a matrix denotes the matrix of the derivatives of the elements. Using these relations, we get from (2·3), for $i,j = 1,\ldots,d$,

$$\frac{\partial L}{\partial G_{ij}} = N\,\mathrm{tr}\,(\mathbf{E}_{ij}\mathbf{G}^{-1}) - N\,\mathrm{tr}\,(\mathbf{C}_0\mathbf{G}^{-1}\mathbf{E}_{ij}\mathbf{G}^{-1}),$$

where G_{ij} denotes the (i,j)th element of $\mathbf{G}$ and $\mathbf{E}_{ij}$ denotes a matrix with (i,j)th element equal to 1 and others equal to zero and with the dimension identical to that of its adjacent matrices. By equating the above derivatives to zero, we get

$$G_{ji}^{-1} = (\mathbf{G}^{-1}\mathbf{C}_0\mathbf{G}^{-1})_{ji},$$

which implies that, other parameters being fixed, $\mathbf{G} = \mathbf{C}_0$ gives the minimum of L. Inserting this into (2·3), to maximize the likelihood function, we have only to minimize M defined by

$$M = \log|\mathbf{C}_0|. \tag{3·3}$$

As was mentioned in the preceding section, except for the case of a pure autoregression, the dependence of $\mathbf{C}_0$ on the parameters is highly nonlinear and the only way to minimize M is by numerical procedures. Numerical optimization, either minimization or maximiza-

tion, is itself a subject of intensive current study, but as can be seen from a recent survey by Powell (1970) the gradient and the Hessian play the dominant roles in this field. With the aid of Fourier transforms, compact representations for these quantities are available in the present case. From (2·6) and (2·5), we have

$$\mathbf{C}_0 = \int \mathbf{A}^{-1}\mathbf{B}\mathbf{Y}\mathbf{Y}^*\mathbf{B}^*\mathbf{A}^{*-1}\,df. \tag{3·4}$$

Let $a_{ij}(m)$ and $b_{ij}(m)$ denote the (i,j)th elements of $\mathbf{a}(m)$ and $\mathbf{b}(m)$, respectively. By (3·1), we have $\partial M/\partial a_{uv}(k) = \operatorname{tr}\{\mathbf{C}_0^{-1}\,\partial\mathbf{C}_0/\partial a_{uv}(k)\}$. From (3·4) and the relation

$$\partial\mathbf{A}/\partial a_{uv}(k) = \exp(-i2\pi kf)\,\mathbf{E}_{uv},$$

we get

$$\partial\mathbf{C}_0/\partial a_{uv}(k) = \mathbf{C}_1\{a_{uv}(k)\} + \mathbf{C}_1\{a_{uv}(k)\}',$$

where

$$\mathbf{C}_1\{a_{uv}(k)\} = -\int \mathbf{A}^{-1}\mathbf{B}\mathbf{Y}\mathbf{Y}^*\mathbf{B}^*\mathbf{A}^{*-1}\mathbf{E}_{vu}\mathbf{A}^{*-1}\exp(i2\pi kf)\,df. \tag{3·5}$$

Taking into account the symmetry of $\mathbf{C}_0$ and the fact that the factors can be rotated under the trace sign, we get, from (3·1) and (3·3),

$$\partial M/\partial a_{uv}(k) = 2\operatorname{tr}[\mathbf{C}_1\{a_{uv}(k)\}\,\mathbf{C}_0^{-1}]. \tag{3·6}$$

Analogously, we get

$$\partial M/\partial b_{rs}(m) = 2\operatorname{tr}[\mathbf{C}_1\{b_{rs}(m)\}\,\mathbf{C}_0^{-1}], \tag{3·7}$$

where

$$\mathbf{C}_1\{b_{rs}(m)\} = \int \mathbf{A}^{-1}\mathbf{B}\mathbf{Y}\mathbf{Y}^*\mathbf{E}_{sr}\mathbf{A}^{*-1}\exp(i2\pi mf)\,df. \tag{3·8}$$

By rotating the factors of (3·6) and (3·7) under the trace sign and using the relation $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}\mathbf{Y}$, one representation for the gradient of M is given as follows, for $u,v,r,s = 1,\ldots,d;\ k = 1,\ldots,p;\ m = 1,\ldots,q$,

$$\frac{\partial M}{\partial a_{uv}(k)} = -2\operatorname{tr}\int \mathbf{H}\mathbf{A}\mathbf{X}\mathbf{X}^*\mathbf{E}_{vu}\exp(i2\pi kf)\,df \tag{3·9}$$

$$\frac{\partial M}{\partial b_{rs}(m)} = 2\operatorname{tr}\int \mathbf{H}\mathbf{B}\mathbf{Y}\mathbf{Y}^*\mathbf{E}_{sr}\exp(i2\pi mf)\,df, \tag{3·10}$$

where

$$\mathbf{H} = \mathbf{A}^{*-1}\mathbf{C}_0^{-1}\mathbf{A}^{-1}.$$

It can be seen that $\partial M/\partial a_{uv}(k)$ can be obtained as minus twice the (u,v)th element of the k-lag cross-covariance matrix between the time series which are the inverse Fourier transforms of $\mathbf{HAX}$ and $\mathbf{X}$. Analogously, $\partial M/\partial b_{rs}(m)$ is twice the (r,s)th element of the m-lag cross-covariance matrix between the series corresponding to $\mathbf{HBY} = \mathbf{HAX}$ and $\mathbf{Y}$. This observation reveals the identity between the present results and those obtained by Kashyap (1970, p. 29) by a time domain analysis with the help of Lagrange multipliers. By using (3·2), we have, for $u,v,r,s = 1,\ldots,d;\ k = 1,\ldots,p;\ m = 1,\ldots,q$,

$$\frac{\partial^2 M}{\partial b_{rs}(m)\,\partial a_{uv}(k)} = \operatorname{tr}\left\{\frac{\partial^2\mathbf{C}_0}{\partial b_{rs}(m)\,\partial a_{uv}(k)}\mathbf{C}_0^{-1} - \frac{\partial\mathbf{C}_0}{\partial a_{uv}(k)}\mathbf{C}_0^{-1}\frac{\partial\mathbf{C}_0}{\partial b_{rs}(m)}\mathbf{C}_0^{-1}\right\}.$$

Here we are going to develop an approximation to the Hessian, which will practically always satisfy the numerically basic requirement of positive definiteness. Also, it will provide a good approximation to the inverse of Fisher's information matrix when the model

is correct and a large number of observations are available and therefore the estimates of the parameters are close to the true values. We already noticed the relation

$$\partial \mathbf{C}_0/\partial a_{uv}(k) = \mathbf{C}_1\{a_{uv}(k)\} + \mathbf{C}_1\{a_{uv}(k)\}',$$

where $\mathbf{C}_1\{a_{uv}(k)\}$ was given by (3·5).

When N is sufficiently large and the estimates are close to the true values, $\mathbf{A}^{-1}\mathbf{B}\mathbf{Y}\mathbf{Y}^*\mathbf{B}^*\mathbf{A}^{*-1}$ will be close to $\mathbf{G}$, the covariance matrix of $\mathbf{x}(n)$. Since $\mathbf{A}$ is physically realizable, the Fourier transform of $\mathbf{A}^{*-1}$ vanishes on the positive side of the time axis. These observations suggest an approximation to the Hessian obtained by putting $\partial \mathbf{C}_0/\partial a_{uv}(k)$ equal to zero. By using (3·8), analogous discussion can be applied to $\partial \mathbf{C}_0/\partial b_{rs}(m)$ and we come to the conclusion that a reasonable approximation to the Hessian will be obtained by assuming these first order derivatives within the above formula equal to zero. This amounts to ignoring the derivatives of $\mathbf{C}_0^{-1}$ and thus from (3·6) an approximation to $\partial^2 M/\partial b_{rs}(m)\,\partial a_{uv}(k)$ is given by $2\,\mathrm{tr}\,[\mathbf{C}_0^{-1}\,\partial \mathbf{C}_1\{a_{uv}(k)\}/\partial b_{rs}(m)]$. This approximation was discussed in the unpublished report of Aström, Bohlin and Wensmark for the one-dimensional case. Now from (3·5), we get

$$\frac{\partial \mathbf{C}_1\{a_{uv}(k)\}}{\partial b_{rs}(m)} = -\int \mathbf{A}^{-1}\mathbf{E}_{rs}\mathbf{Y}\mathbf{Y}^*\mathbf{B}^*\mathbf{A}^{*-1}\mathbf{E}_{vu}\mathbf{A}^{*-1}\exp\{i2\pi(k-m)f\}\,df$$
$$-\int \mathbf{A}^{-1}\mathbf{B}\mathbf{Y}\mathbf{Y}^*\mathbf{E}_{sr}\mathbf{A}^{*-1}\mathbf{E}_{vu}\mathbf{A}^{*-1}\exp\{i2\pi(k+m)f\}\,df.$$

On taking into account the relation $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}\mathbf{Y}$, the last term can be expressed in the form

$$\int \mathbf{X}\mathbf{X}^*\mathbf{A}^*\mathbf{B}^{*-1}\mathbf{E}_{sr}\mathbf{A}^{*-1}\,\mathbf{E}_{vu}\mathbf{A}^{*-1}\exp\{i2\pi(k+m)f\}\,df.$$

In the time domain, this represents a matrix whose elements can be expressed as various sums of $(k+m)$-lag cross-products between the elements of autocovariance matrix sequence of $\mathbf{x}(n)$ and the elements of a matrix time series; the latter have only zeros on the negative side of the time axis. Thus, when the data length N is large and the fitted model produces nearly white $\mathbf{x}(n)$, this term will nearly vanish. Based on this, we neglect the last term and adopt the approximation

$$\frac{\partial \mathbf{C}_1\{a_{uv}(k)\}}{\partial b_{rs}(m)} = -\int \mathbf{A}^{-1}\mathbf{E}_{rs}\mathbf{Y}\mathbf{X}^*\mathbf{E}_{vu}\mathbf{A}^{*-1}\exp\{i2\pi(k-m)f\}\,df.$$

By analogous reasoning we have approximately

$$\frac{\partial \mathbf{C}_1\{a_{uv}(k)\}}{\partial a_{rs}(m)} = \int \mathbf{A}^{-1}\mathbf{E}_{rs}\mathbf{X}\mathbf{X}^*\mathbf{E}_{vu}\mathbf{A}^{*-1}\exp\{i2\pi(k-m)f\}\,df.$$

We have exactly

$$\frac{\partial \mathbf{C}_1\{b_{uv}(k)\}}{\partial b_{rs}(m)} = \int \mathbf{A}^{-1}\mathbf{E}_{rs}\mathbf{Y}\mathbf{Y}^*\mathbf{E}_{vu}\mathbf{A}^{*-1}\exp\{i2\pi(k-m)f\}\,df.$$

Thus as our final approximation to the Hessian we adopt, for $u, v, r, s = 1, \ldots, d$,

$$\frac{\partial^2 M}{\partial a_{rs}(m)\,\partial a_{uv}(k)} = 2\,\mathrm{tr}\int \mathbf{E}_{rs}\mathbf{X}\mathbf{X}^*\mathbf{E}_{vu}\mathbf{H}\exp\{i2\pi(k-m)f\}\,df \quad (m, k = 1, \ldots, p), \tag{3·11}$$

$$\frac{\partial^2 M}{\partial b_{rs}(m)\,\partial a_{uv}(k)} = -2\,\mathrm{tr}\int \mathbf{E}_{rs}\mathbf{Y}\mathbf{X}^*\mathbf{E}_{vu}\mathbf{H}\exp\{i2\pi(k-m)f\}\,df \quad (m = 1, \ldots, q;\ k = 1, \ldots, p), \tag{3·12}$$

$$\frac{\partial^2 M}{\partial b_{rs}(m)\,\partial b_{uv}(k)} = 2\,\mathrm{tr}\int \mathbf{E}_{rs}\mathbf{Y}\mathbf{Y}^*\mathbf{E}_{vu}\mathbf{H}\exp\{i2\pi(k-m)f\}df \quad (m,k=1,\ldots,q), \tag{3·13}$$

where $\mathbf{H} = \mathbf{A}^{*-1}\mathbf{C}_0^{-1}\mathbf{A}^{-1}$. To show the nonnegative definiteness of this approximation, we proceed as follows. Let $\mathbf{D}$ represent the positive definite symmetric square root of $\mathbf{C}_0^{-1}$. Define the stationary stochastic processes $\mathbf{X}_{uv}(n)$ and $\mathbf{Y}_{rs}(n)$ by

$$\mathbf{X}_{uv}(n) = \int \exp(i2\pi nf)\,\mathbf{D}\mathbf{A}^{-1}\mathbf{E}_{uv}d\mathbf{X}_0(f),$$

$$\mathbf{Y}_{rs}(n) = \int \exp(i2\pi nf)\,\mathbf{D}\mathbf{A}^{-1}\mathbf{E}_{rs}d\mathbf{Y}_0(f),$$

where $d\mathbf{X}_0(f)$ and $d\mathbf{Y}_0(f)$ represent orthogonal increment processes with

$$E\{d\mathbf{X}_0(f)\,d\mathbf{X}_0(f)^*\} = \mathbf{X}\mathbf{X}^*, \quad E\{d\mathbf{Y}_0(f)\,d\mathbf{X}_0(f)^*\} = \mathbf{Y}\mathbf{X}^*, \quad E\{d\mathbf{Y}_0(f)d\mathbf{Y}_0(f)^*\} = \mathbf{Y}\mathbf{Y}^*.$$

By introducing a d-dimensional random vector $\mathbf{e}$, which is independent of $\mathbf{X}_{uv}(n)$ and $\mathbf{Y}_{rs}(n)$ and with mutually independent components $e(i)$ $(i=1,\ldots,d)$ with $E\{e(i)\}=0$ and $E\{e(i)\}^2=1$, we can define stationary scalar stochastic processes $x_{uv}(n) = \mathbf{e}'\mathbf{X}_{uv}(n)$ and $y_{rs}(n) = \mathbf{e}'\mathbf{Y}_{rs}(n)$. The above approximation to the Hessian is identical to the variance covariance matrix of $\{-x_{uv}(n+k), y_{rs}(n+m)\}$ $(u,v,r,s=1,\ldots,d;\ k=1,\ldots,p;\ m=1,\ldots,q)$. This proves that the present approximation to the Hessian is nonnegative definite.

4. Newton–Raphson procedure

The Newton–Raphson procedure for successive minimization of M is realized by correcting the vector of the present estimates of the parameters by the amount equal to the negative of the inverse of the Hessian times the gradient. This gives the exact minimum in one step if the objective function is quadratic (Powell, 1970, p. 83). Thus the procedure will be useful when the initial estimates come sufficiently close to the minimizing values of the parameters. This is generally not the case in practical applications and we have to incorporate various modifications of the correcting procedure to make it useful. Thus only a reasonable approximation to the inverse of the Hessian will be sufficient for numerical procedures. It will be shown that Hannan's procedure is equivalent to just one step of the standard Newton–Raphson procedure with the gradient and the approximate Hessian obtained in the preceding section.

The scalar case where $d=1$ is especially simple to treat, since in this case the trace signs and $\mathbf{E}_{uv}$'s can be ignored and we have only $a(k) = a_{11}(k)$ and $b(m) = b_{11}(m)$ to consider. In this case if we denote by $\mathbf{U}$ the $p\times p$ matrix given by (3·11) and by $\mathbf{a}$ the vector of

$$a(k)\ (k=1,\ldots,p)$$

the part of the gradient given by (3·9) can be represented in the form $-\mathbf{c}-\mathbf{U}\mathbf{a}$, where $\mathbf{c}$ is a p-dimensional vector obtained by putting $m=0$ in (3·11) for $k=1,\ldots,p$. Analogously, the part of the gradient given by (3·10) can be represented in the form $\mathbf{d}+\mathbf{V}\mathbf{b}$, where $\mathbf{V}$ is a $q\times q$ matrix given by (3·13), $\mathbf{b}$ is the vector of $b(m)$ $(m=1,\ldots,q)$ and $\mathbf{d}$ is a q-dimensional vector obtained by putting $m=0$ in (3·13) for $k=1,\ldots,q$. Each step of the Newton–Raphson procedure modifies $\mathbf{a}$ and $\mathbf{b}$ into $\mathbf{a}_1$ and $\mathbf{b}_1$ defined by

$$\begin{bmatrix}\mathbf{a}_1\\ \mathbf{b}_1\end{bmatrix} = \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix} - \begin{bmatrix}\mathbf{U} & -\mathbf{W}'\\ -\mathbf{W} & \mathbf{V}\end{bmatrix}^{-1}\begin{bmatrix}-\mathbf{c}-\mathbf{U}\mathbf{a}\\ \mathbf{d}+\mathbf{t}\mathbf{V}\mathbf{b}\end{bmatrix}, \tag{4·1}$$

where $-\mathbf{W}$ denotes a $q \times p$ matrix obtained by (3·12) ($m = 1, \ldots, q$; $k = 1, \ldots, p$). It can be seen that if we put

$$\begin{bmatrix} \mathbf{U} & -\mathbf{W}' \\ -\mathbf{W} & \mathbf{V} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P} & \mathbf{R}' \\ \mathbf{R} & \mathbf{Q} \end{bmatrix}$$

then

$$\mathbf{P} = (\mathbf{U} - \mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}, \quad \mathbf{R} = \mathbf{V}^{-1}\mathbf{W}(\mathbf{U} - \mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}, \quad \mathbf{R}' = \mathbf{P}\mathbf{W}'\mathbf{V}^{-1}$$

$$\mathbf{Q} = \mathbf{V}^{-1} + \mathbf{V}^{-1}\mathbf{W}\mathbf{R}'.$$

Using these relations, we get

$$\mathbf{a}_1 = \mathbf{a} + \boldsymbol{\xi}, \quad \mathbf{b}_1 = -\mathbf{V}^{-1}(\mathbf{d} - \mathbf{W}\boldsymbol{\xi}), \tag{4·2}$$

where

$$\boldsymbol{\xi} = (\mathbf{I}_p - \mathbf{U}^{-1}\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\{\mathbf{U}^{-1}(\mathbf{c} + \mathbf{U}\mathbf{a}) + \mathbf{U}^{-1}\mathbf{W}'(-\mathbf{V}^{-1}\mathbf{d} - \mathbf{b})\}, \tag{4·3}$$

$\mathbf{I}_p$ denoting a $p \times p$ identity matrix.

If we consider the fact that in the quadratic case $-\mathbf{W}\Delta\mathbf{a}$ and $\mathbf{W}'\Delta\mathbf{b}$ give the variations of the derivatives of M with respect to $\mathbf{b}$ and $\mathbf{a}$ due to the variations $\Delta\mathbf{a}$ and $\Delta\mathbf{b}$ of $\mathbf{a}$ and $\mathbf{b}$, respectively, we can see that the above step of the Newton–Raphson procedure can be decomposed into three stages of operation. First, by equating $\mathbf{d} + \mathbf{V}\mathbf{b}$ equal to zero, $\mathbf{b}$ is replaced by $\mathbf{b}_0 = -\mathbf{V}^{-1}\mathbf{d}$. This is equivalent to minimizing M with respect to $\mathbf{b}$ and the change of $\mathbf{b}$ by the amount of $\Delta\mathbf{b} = -\mathbf{V}^{-1}\mathbf{d} - \mathbf{b}$ modifies $-(\mathbf{c} + \mathbf{U}\mathbf{a})$ of (4·1) into

$$-(\mathbf{c} + \mathbf{U}\mathbf{a} + \mathbf{W}'\Delta\mathbf{b}).$$

If we put this last quantity equal to zero and solve for $\mathbf{a}$, then we get $\mathbf{a}_0 = -\mathbf{U}^{-1}(\mathbf{c} + \mathbf{W}'\Delta\mathbf{b})$, and this introduces the modification of $\mathbf{a}$ by the amount $\Delta\mathbf{a} = \mathbf{U}^{-1}(\mathbf{c} + \mathbf{U}\mathbf{a}) + \mathbf{U}^{-1}\mathbf{W}'\Delta\mathbf{b}$. Even with this second modification the minimum of a quadratic objective function cannot be directly obtained and the process has to be iterated indefinitely to attain the minimum. The modification of $\mathbf{a}$ by $\Delta\mathbf{a}$ changes $\mathbf{d} + \mathbf{V}\mathbf{b}$ in the gradient into $\mathbf{d} + \mathbf{V}\mathbf{b} - \mathbf{W}\Delta\mathbf{a}$, and by equating this equal to zero a new value of $\mathbf{b}$ is obtained as $-\mathbf{V}^{-1}\mathbf{d} + \mathbf{V}^{-1}\mathbf{W}\Delta\mathbf{a}$. The difference between this and the original $\mathbf{b}$ is given by $\Delta\mathbf{b}_1 = -\mathbf{b} - \mathbf{V}^{-1}\mathbf{d} + \mathbf{V}^{-1}\mathbf{W}\Delta\mathbf{a}$. At this point it is easy to see that the iteration of this process will eventually lead to the solution given by (4·2) and (4·3). The Hannan procedure is a realization of this successive minimization procedure with a modification of the second stage so that the whole process will be complete at the third stage. The above $\mathbf{b}_0$ is Hannan's $\hat{\boldsymbol{\beta}}^{(1)}$, $\mathbf{a}_0$ is the second intermediate statistic of Hannan ($\hat{\boldsymbol{\alpha}}^{(2)}$ of 1969, $\hat{\boldsymbol{\alpha}}^{(1)}$ of 1970) and $\mathbf{a}_1$ and $\mathbf{b}_1$ are the final estimates $\tilde{\boldsymbol{\alpha}}^{(1)}$ and $\tilde{\boldsymbol{\beta}}^{(1)}$. Tretter & Steiglitz (1967) did not consider the possibility of this modification.

On taking into account the relations

$$\mathbf{A} = \sum_{m=0}^{p} \sum_{r=1}^{d} \sum_{s=1}^{d} a_{rs}(m)\, \mathbf{E}_{rs} \exp(-i2\pi m f),$$

$$\mathbf{B} = \sum_{k=0}^{q} \sum_{u=1}^{d} \sum_{v=1}^{d} b_{uv}(k)\, \mathbf{E}_{uv} \exp(-i2\pi k f),$$

the above discussion can be extended directly to the multidimensional vector case, if the definitions of $\mathbf{a}$ and $\mathbf{b}$ are replaced by

$$\{a_{uv}(k) : u, v = 1, \ldots, d; k = 1, \ldots, p\}, \quad \{b_{rs}(m) : r, s = 1, \ldots, d; m = 1, \ldots, q\},$$

those of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ modified correspondingly and $\mathbf{I}_p$ in (4·3) replaced by a $pd^2 \times pd^2$ identity matrix. As will be discussed in the next section, it seems that this extension to the multi-

dimensional case differs from the one suggested by Hannan for the multidimensional moving average case.

The above discussion is only about a realization of one step of the Newton–Raphson procedure and it should be remembered that the simple iteration of the step with recalculation of the gradient and Hessian does not guarantee the convergence to the minimum of M in practical applications. The simple iteration will work only if the initial estimates are sufficiently close to the minimizing parameter values. There are various numerical procedures developed to overcome the difficulty when the initial estimates are far from the minimizing values (Powell, 1970), but since the choice of the best procedure is highly empirical we will leave it to future experimental studies.

5. Numerical considerations

The dimension of the Hessian is $(p+q)\,d^2 \times (p+q)\,d^2$. This shows that when d is large the computational difficulty will be quite significant. Thus even if we adopt a numerical procedure which avoids the direct calculation of the Hessian, if an estimate of the inverse of the Hessian is required the amount of storage necessary will become a serious problem. On reflection it becomes clear that we can drastically reduce this amount. As was already mentioned, the present approximation to the Hessian has the structure of a variance matrix of a set of variables of some stationary stochastic processes. Thus if in the discussion at the end of § 3 we arrange the elements of the stochastic processes in the form

$$\{-\mathbf{x}_{\cdot\cdot}(n+1), \ldots, -\mathbf{x}_{\cdot\cdot}(n+p), \mathbf{y}_{\cdot\cdot}(n+1), \ldots, \mathbf{y}_{\cdot\cdot}(n+q)\},$$

where

$$\mathbf{z}_{\cdot\cdot}(n) = \{\mathbf{z}_{1\cdot}(n), \ldots, \mathbf{z}_{d\cdot}(n)\}, \quad \mathbf{z}_{u\cdot}(n) = \{z_{u1}(n), \ldots, z_{ud}(n)\}$$

for $\mathbf{z} = \mathbf{x}$ and $\mathbf{y}$, the resulting approximate Hessian has a block Toeplitz type characteristic, i.e. we have only to store the first columns of $d^2 \times d^2$ blocks of $\mathbf{U}$ and $\mathbf{V}$ and the first rows and columns of $d^2 \times d^2$ blocks of $\mathbf{W}$. Other $d^2 \times d^2$ blocks are obtained by diagonally shifting down these blocks and using the symmetry of the Hessian. This means the reduction of the amount of necessary storage from $(p+q)\,d^2 \times (p+q)\,d^2$ to $\{2(p+q)-1\} \times d^2 \times d^2$. If we take into account the symmetry of two of the matrices, this figure further reduces to $2(p+q-1) \times d^2 \times d^2$. For the block Toeplitz matrices $\mathbf{U}$ and $\mathbf{V}$ there are efficient procedures for computing their inverses and the solutions of the simultaneous equations with coefficients equal to $\mathbf{U}$ or $\mathbf{V}$ (Kutikov, 1967; Akaike, 1973). In applying Hannan's procedure, the computation of (4·3) introduces serious trouble. Without further analysis of this stage, it looks as if the necessary dimension of the matrix jumps up again to $pd^2 \times pd^2$. The simplest case is where $p = q$. In this case we arrange the approximate Hessian so that it corresponds to the covariance matrix of

$$\{-\mathbf{x}_{\cdot\cdot}(n+1), \mathbf{y}_{\cdot\cdot}(n+1), \ldots, -\mathbf{x}_{\cdot\cdot}(n+p), \mathbf{y}_{\cdot\cdot}(n+p)\}.$$

Obviously the corresponding approximate Hessian is a block Toeplitz matrix and we have only to manipulate with p blocks each of dimension $2d^2 \times 2d^2$, one of which is symmetric. For this case the advantage of a block Toeplitz matrix can be fully exploited and without using the formulae (4·2) we can directly get the results of (4·1) with manipulations of $2d^2 \times 2d^2$ dimensional matrices. This leads to a significant computational simplicity which has not previously been mentioned in the literature. For the Newton–Raphson procedure, the explicit evaluation of the inverse of the Hessian is not necessary. But it is necessary for the

evaluation of the statistical variability of the estimates, as this provides an estimate of Fisher's information matrix. For the block Toeplitz type Hessian this can be obtained easily. A further point is that the block Toeplitz type characteristic is not disturbed by the addition of a constant $d^2 \times d^2$ diagonal matrix to the diagonal blocks of the Hessian. This permits the Marquardt type modification of the Newton–Raphson procedure (Powell, 1970, p. 95) without losing the above-stated numerical convenience.

Computations of the Hessian can be done from (3·11), (3·12) and (3·13). By replacing $\mathbf{X}$ by $\mathbf{A}^{-1}\mathbf{BY}$, these quantities can be obtained from the sequences of autocovariance matrices of $\mathbf{Y}$ with the aid of some finite duration approximation of the inverse Fourier transform of $\mathbf{A}^{-1}$. To realize this, $\mathbf{A}$ must be invertible. Since (3·9) shows that at the minimum of M every (u, v)th element of (3·9) should vanish, we have

$$\int \mathbf{HAXX}^* \exp(i2\pi kf)\, df = 0 \quad (k = 1, \ldots, p).$$

Because $\mathbf{H} = \mathbf{A}^{*-1}\mathbf{C}_0^{-1}\mathbf{A}^{-1}$, we have that

$$\int \mathbf{KAC}_0 \exp(i2\pi kf)\, df = 0 \quad (k = 1, \ldots, p),$$

where $\mathbf{K} = \mathbf{A}^{*-1}\mathbf{C}_0^{-1}\mathbf{XX}^*\mathbf{C}_0^{-1}\mathbf{A}^{-1}$. Since $\mathbf{C}_0$ is a positive definite constant matrix, we have

$$\int \mathbf{KA} \exp(i2\pi kf)\, df = 0 \quad (k = 1, \ldots, p). \tag{5·1}$$

This last result shows that $\mathbf{a}(m)$ $(m = 1, \ldots, p)$ are the coefficient matrices of the pth order autoregressive model fitted to the sequence of autocovariance matrices corresponding to $\mathbf{K}$. Now $\mathbf{K}$ will never exactly correspond to a finite order autoregressive model; Whittle (1963) has shown that in this case $\mathbf{A}$ is invertible. Analogously, from (3·10), we can infer that $\mathbf{B}$ is invertible. These observations show that during the process of maximization of the Gaussian likelihood we can limit $\mathbf{A}$ and $\mathbf{B}$ to the set of invertible models. Incidentally, (5·1) is apparently in the same form as the equation proposed by Hannan [1969, (13); 1970, (5·11)] for the calculation of the intermediate statistic for the fitting of a multidimensional moving average model. It is clear from our discussion of the Newton–Raphson procedure that if in (3·9) and (3·11) we define $\mathbf{H}$ by the present values of $\mathbf{A}$ and $\mathbf{C}_0$ and equate (3·9) to zero to solve for a new $\mathbf{A}$, this will give the desired intermediate statistics. In contrast to this, if in (5·1) we define $\mathbf{K}$ by the present values of $\mathbf{A}$ and $\mathbf{C}_0$ and solve for $\mathbf{A}$ as proposed by Hannan, the relation between the present solution and the Newton–Raphson procedure is not clear. This is because (5·1) is an arbitrary linearization of the nonlinear maximum likelihood equation, obtained by equating (3·9) to zero.

The following observation clarifies the inherent difficulty in the multidimensional case. In the scalar case, once the moving average coefficients are fixed, the best fitting autoregression coefficients are directly obtained as the autoregression coefficients of the series corresponding to $\mathbf{A}^{-1}\mathbf{Y}$. Unfortunately in the multidimensional case $\mathbf{A}^{-1}$ and $\mathbf{B}$ usually do not commute and thus the gradient of the likelihood function with respect to the autoregressive coefficients becomes highly nonlinear and the maximization of the likelihood with respect to the autoregressive coefficients can only be realized through some iterative numerical procedure.

The final comments are about getting the initial estimates of the parameters and the decision on the order of the models. Since a decision procedure developed by Akaike (1971*a*) on the order of an autoregressive model to be fitted to a given set of data gives reasonable

results in practical applications (Otomo, *et al.* 1972), it seems that by using this procedure we can fit an autoregressive model to get an estimate of $\mathbf{x}(n)$. This, as was suggested by Durbin (1960), will allow us to get the initial estimates of the coefficients of an autoregressive moving average model. Based on an information theoretic consideration, the basic idea of the decision procedure on the order of autoregressive models has been extended to the general case of the maximum likelihood model identification by Akaike (1971*b*) and its practical utility has been checked with numerical examples, including one-dimensional autoregressive moving average models (Akaike, 1972*a*, *b*). For the final, and possibly also for the initial, determination of proper autoregressive moving average models, this procedure may profitably be used. Models of various orders should be tried and compared, and for the decision on the best fit the above-mentioned procedure will be useful.

The author would like to express his sincere thanks to Professors Will Gersch and Richard H. Jones, the University of Hawaii, for helpful and stimulating discussions of the present subject. This work was partially supported by National Science Foundation Grants at the University of Hawaii.

References

Akaike, H. (1971*a*). Autoregressive model fitting for control. *Ann. Inst. Statist. Math.* (*Tokyo*) **23**, 163–80.

Akaike, H. (1971*b*). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory, Tsahkadsor, Armenian* SSR. *Problems of Control and Information Theory*. To appear.

Akaike, H. (1972*a*). Use of an information theoretic quantity for statistical model identification. *Proc. 5th Hawaii International Conference on System Sciences*, pp. 249–50.

Akaike, H. (1972*b*). Automatic data structure search by the maximum likelihood. *Computers in Biomedicine, Supplement to Proc. 5th Hawaii International Conference on System Sciences*, pp. 99–101.

Akaike, H. (1973). Block Toeplitz matrix inversion. *S.I.A.M. J. Appl. Math.* **24**. To appear.

Durbin, J. (1960). The fitting of time-series models. *Rev. Inst. Int. Statist.* **28**, 233–44.

Dwyer, P. S. (1967). Some applications of matrix derivatives in multivariate analysis. *J. Am. Statist. Ass.* **62**, 607–25.

Dzhaparidze, K. O. (1971). On methods for obtaining asymptotically efficient spectral parameter estimates for a stationary Gaussian process with rational spectral density. *Theory Prob. Applic.* **16**, 550–4.

Hájek, J. (1962). On linear statistical problems in stochastic processes. *Czechoslav. Math. J.* **12**, 404–44.

Hannan, E. J. (1969). The estimation of mixed moving average autoregressive systems. *Biometrika* **56**, 579–93.

Hannan, E. J. (1970). *Multiple Time Series*. New York: Wiley.

Kashyap, R. L. (1970). Maximum likelihood identification of stochastic linear systems. *I.E.E.E. Trans. Auto. Control* AC-15, 25–34.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Kutikov, L. M. (1967). The structure of matrices which are the inverse of the correlation matrices of random vector processes. *Computational Math. and Math. Phys.* **7**, 58–71.

Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypothesis. *Proc. 3rd Berkeley Symp.* **1**, 129–56.

Otomo, T., Nakagawa, T. & Akaike, H. (1972). Statistical approach to computer control of cement rotary kilns. *Automatica* **8**, 35–48.

Powell, M. J. D. (1970). A survey of numerical methods for unconstrained optimization. *S.I.A.M. Rev.* **12**, 79–97.

Tretter, S. A. & Steiglitz, K. (1967). Power-spectrum identification in terms of rational models. *I.E.E.E. Trans. Auto. Control* AC-12, 185–8.

Whittle, P. (1953). The analysis of multiple stationary time series. *J. R. Statist. Soc.* B **15**, 125–39.

Whittle, P. (1963). On the fitting of multivariate autoregression, and the approximate canonical factorization of a spectral density matrix. *Biometrika* **50**, 129–34.

Wilson, G. T. (1971). Recent developments in statistical process control. *Bull. Inst. Int. Statist.* **44**, 1. 509–36.

[*Received August* 1972. *Revised January* 1973]