Statistical Science
1987, Vol. 2, No. 2, 135–161

Rational Transfer Function Approximation

E. J. Hannan

Abstract. The problem considered is that of approximation to the structure of the stationary process generating a vector time series by a rational transfer function, i.e. ARMA, system. The structure of such systems and their coordinatization are discussed together with some deterministic approximation theory. Criteria on the basis of which to choose an approximant are considered. Theorems supporting such criteria and describing the properties of approximants are given. Finally some algorithms are described including algorithms for real time calculation of the estimates.

Key words and phrases: Time series, autoregressive-moving average, rational transfer function, Hankel matrix, Kronecker indices, Kalman filter, McMillan degree, Hankel norm approximation, canonical correlation, balanced realization, AIC, BIC, minimum description length, Levinson-Whittle recursion, lattice algorithms, forgetting, Givens transformations, real time calculation.

1. INTRODUCTION

The classical paradigm of statistics assumes that the data is generated by a stochastic process that is known save for a finite number of parameters. Around this has been built a body of theory and practice of august proportions. The paradigm has not, of course, been held sacred as the development of nonparametric methods shows. Also, in the theory of robust estimation, only part of the specification may involve a finite parametrization. In a large part of time series analysis, the classical paradigm has been seen to be inadequate. Here the amount of data increases as time passes and it is apparent that as more data accrues a more complex model will be fitted. Thus, here it is evident that the model set does not contain the truth but serves only to provide some approximation to that. Such an attitude is very obvious in the theory of spectral estimation (Brillinger, 1984, pages 1143–1153). In this paper, however, rational transfer function approximation is considered. We go on to explain this.

Statisticians will be familiar with the autoregressive-moving average (ARMA) model for a vector time series $y(t)$ of n components observed at $t = 1, 2, \cdots, T$, namely

$$\sum_0^p A(j)y(t-j) = \sum_0^q B(j)\varepsilon(t-j), \qquad A(0) = B(0) = I_n, \tag{1.1}$$

$$E\{\varepsilon(t)\varepsilon(t)'\} = \delta_{s,t}\Sigma, \quad \Sigma > 0. \tag{1.2}$$

If $n = 1$ we shall use lower case letters, $\alpha(j)$, $\beta(j)$, σ^2. When $q = 0$, (1.1) is called an autoregression, and when $p = 0$, a moving average. Each is a natural simple model although the autoregressions have proved far more important in practice both because they are simpler to estimate, computationally, and are more realistic. (For a moving average observations $q + 1$ time points apart are uncorrelated.) Computational problems are no longer so important and since the influential book, Box and Jenkins (1970), (1.1) has been widely used in statistics, at least for $n = 1$. The model (1.1) is the prototype rational transfer function. Below we explain the use of the term rational transfer function. Before doing that we emphasize that here *the stationary case alone is dealt with and the mean of* $y(t)$ *is taken as zero.* It is not difficult to extend the discussion to the ARMAX case where $z(t)$, an exogenous influence (hence the added X on ARMA), is included in a model, which is now of the form

$$\sum_0^p A(j)y(t-j) = \sum_0^r D(j)z(t-j) + \sum_0^q B(j)\varepsilon(t-j). \tag{1.1'}$$

In other parlance, $y(t)$ is said to be the output (endogenous corresponds to the exogenous usage) and $z(t)$ the (observed) input. Of course $\varepsilon(t)$ is an unobserved input. If only a constant mean is to be modeled then $r = 0$, $D(0) = \nu$, a column vector and $z(t) \equiv 1$, so that $\nu = \sum A(j)\mu$ with $E\{y(t)\} = \mu$.

Since $y(t)$ is stationary if it is generated by (1.1), we may always choose $A(j)$, $B(j)$ so that the generating functions or "transfer functions" $a(z) = \sum A(j)z^j$,

E. J. Hannan is Professor of Statistics, Institute of Advanced Studies, Australian National University, Canberra, 2601, Australia.

$b(z) = \sum B(j)z^j$ satisfy

$$\det a(z) \neq 0, \quad |z| \leq 1; \tag{1.3}$$
$$\det b(z) \neq 0, \quad |z| < 1.$$

Then also the $\varepsilon(t)$ in (1.1) are the innovations, i.e., $\varepsilon(t) = y(t) - y(t \mid t - 1)$ where $y(t \mid t - 1)$ is the best linear predictor of $y(t)$ from $y(s)$, $s \leq t - 1$. Here and below "best" is interpreted in the least squares sense. Of course, $\varepsilon(t)$, $y(t \mid t - 1)$ are theoretical quantities and cannot be observed. These statements, surrounding (1.3), are not trivially established but cannot be proved here. One reference for them, and later assertions, is Hannan (1970).

Let us use z also as the backward shift, $zy(t) = y(t - 1)$, $z\varepsilon(t) = \varepsilon(t - 1)$. Then (1.1) may be rewritten as $a(z)y(t) = b(z)\varepsilon(t)$ or, equivalently, $y(t) = k(z)\varepsilon(t)$, $k(z) = a(z)^{-1}b(z)$. The first part of (1.3), needless to say, justifies the inversion of $a(z)$ in this formula. Thus

$$y(t) = \sum_0^\infty K(j)\varepsilon(t - j), \tag{1.4}$$
$$k(z) = \sum_0^\infty K(j)z^j, \quad K(0) = I_n.$$

Now $k(z)$ is said to be the transfer function from $\varepsilon(t)$ to $y(t)$. If (1.1), (1.3) hold the $K(j)$ decrease to zero at a geometric rate but *we shall use* (1.4) *as a more general model, with* $\varepsilon(t)$ *still satisfying* (1.2), *subject to*

$$\sum_0^\infty \| K(j) \|^2 < \infty. \tag{1.5}$$

It will be convenient to use the norm $\| A \|$ to be the largest singular value of the matrix A. Then if $\varepsilon(t)$ is the innovation sequence for $y(t)$ again

$$k(z) \text{ is analytic for } |z| < 1, \tag{1.6}$$
$$\det k(z) \neq 0, \quad |z| < 1.$$

The first part of (1.3), (1.6) arises of course from the need to represent $y(t)$ in terms of $\varepsilon(t)$, $s \leq t$. Since $\varepsilon(t) = y(t) - y(t \mid t - 1)$ it must be linearly representable in terms of $y(s)$, $s \leq t$, hence the last part of (1.3), (1.6). (That the condition holds only for $|z| < 1$ in (1.3), (1.6) is subtler but is forced on us in any case by the simple model $y(t) = \varepsilon(t) - \varepsilon(t - 1)$, $b(z) = (1 - z)I_n$.) If (1.4), (1.5) and (1.6) hold and $k = a^{-1}b$ with a, b matrices of polynomials then also (1.1) holds.

Models such as (1.1) may then be thought of as approximations to (1.4), (1.5) and (1.6). The parameterization (i.e., coordinatisation) of (1.1) will be discussed in Section 2. Of course integer parameters, e.g. p, q, arise that specify the number of continuously varying parameters. Once such integer parameters arise then maximum likelihood needs modification since without that the method will uncritically choose p, q, as large as is permitted. This question is discussed in Section 4. In Section 3 deterministic (i.e., nonstatistical) approximations to $k(z)$ will be discussed. These can be important when $k(z)$ can be known (e.g., where $\varepsilon(t)$ can be experimentally formed). However, the underlying ideas also relate to some statistical technique discussed in Section 6, where algorithms are considered. Section 5 attempts to give some assessment of criteria introduced in Section 4. Statistical problems now become difficult, and the work is unfinished, because as $T \to \infty$ the number of fitted parameters will also increase.

2. LINEAR SYSTEMS

Consider a sequence, $y(t)$, generated by a stationary process. Phenomena that appear stationary abound in natural science and even in the social sciences, where there is still evolution, the stationary case is often a basis to which the evolutionary case is reduced, for example by trend removal. One natural description of the structure of such a process is through the autocovariance sequence

$$\Gamma(j) = E\{y(t)y(t + j)'\}.$$

Leaving aside some technical details any finite variance, stationary process, $y(t)$, may be represented as the sum of a purely deterministic component (perfectly predictable from its own past) and a component of the form (1.2), (1.4), (1.5) and (1.6). The purely deterministic component would be, essentially, composed of sinusoids at different frequencies (i.e. would be an almost periodic function). It could be modeled and included in the ARMAX extension (1.1) so that $z(t)$ would include components $\cos \lambda_j t$, $\sin \lambda_j t$ and again $r = 0$. We consider only the purely nondeterministic case. Then from (1.4), (1.2)

$$\Gamma(j) = \sum_{u=0}^\infty K(u)\Sigma K(u + j)',$$
$$j \geq 0, \quad \Gamma(-j) = \Gamma(j)',$$

as is easily checked. Thus if $f(\omega)$ is the matrix function with Fourier coefficient matrices $\Gamma(j)$ then

$$f(\omega) \sim \frac{1}{2\pi} \sum_{-\infty}^\infty \Gamma(j)e^{-ij\omega} = \frac{1}{2\pi} k(e^{i\omega})\Sigma k(e^{i\omega})^*, \tag{2.1}$$
$$-\pi \leq \omega \leq \pi.$$

Here the * indicates transposition combined with conjugation. Again this is easily obtained. This spectral density matrix, $f(\omega)$, has a direct physical meaning in terms of a decomposition of $y(t)$ into sinusoidal components of oscillatory frequency $\omega/2\pi$ cycles per time unit. It is much easier to interpret than the $\Gamma(j)$. However, we cannot deal with that here save to say that one reason for rational transfer function approximation procedures may be to approximate to $f(\omega)$ by means of the components in the last term in (2.1). The representation (2.1) is unique subject to (1.6), and $\Sigma > 0$ (see (1.2)). Also $f(\omega)$, and hence k, Σ, can be uniquely determined from a complete realization, given that $y(t)$ is ergodic (which is a costless assumption in our circumstances) since $f(\omega)$ is uniquely determined by the $\Gamma(j)$ and

$$C(j) = T^{-1} \sum_{1}^{T-j} y(t)y(t+j)' \rightarrow \Gamma(j), \quad \text{a.s.}$$

However, we do not have a complete realization, hence the statistical problem.

We need now to give a systematic description of all rational $k(z)$ satisfying (1.3). *Because of* (1.5) *now* $k(z)$ *is analytic for* $|z| \leq 1$ *but to avoid technicalities we shall, henceforth, also assume that* $\det\{k(z)\} \neq 0$, $|z| \leq 1$.

In case $n = 1$ one way to proceed is as follows. Put $d = \max(p, q)$, where p, q are the true degrees, i.e. $\alpha(p)$, $\beta(q) \neq 0$ and $a(z)$, $b(z)$ are prime, i.e. have no common zeros. Then list all $k(z)$ that are rational, first by the integer d and then, d being given, by the $2d$ parameters, $\alpha(j)$, $\beta(j)$, $j = 1, \cdots, d$. Of course if $p \neq q$ then some of these $2d$ are constrained to be zero. In any case, in this way, all possible rational transfer function systems are listed exactly once, for $n = 1$. For d given, and requiring $a(z)$, $b(z) \neq 0$, $|z| \leq 1$, then the $2d$ parameters prescribe an open set in $2d$-dimensional Euclidean space and thus provide a coordinatisation.

The problem becomes much less trivial when $n > 1$. It appears at first that n integers, d_j, will be needed but that is not entirely true as we shall explain. Why, it may be asked, cannot we merely follow the example for $n = 1$? First, of course, it will be necessary to exclude common factors in $a(z)$, $b(z)$ since to list $a(z)$, $b(z)$ and $u(z)a(z)$, $u(z)b(z)$ will mean that the same $k(z)$ will be listed twice. We say that $a(z)$, $b(z)$ are left coprime if $a(z) = u(z)a_1(z)$, $b(z) = u(z)b_1(z)$, all matrices being composed of polynomials, means that $u(z)$ is unimodular, i.e. $\det u(z) \equiv$ constant $\neq 0$. These unimodular matrices are precisely the matrices of polynomials for which $u(z)^{-1}$ is also a matrix of polynomials. Such factors $u(z)$ clearly cannot be "divided out" and a normalization has to be used to ensure that the only possible $u(z)$ is $u(z) = I_n$. If p, q are specified we can ensure that $u(z) = I_n$ if we require $A(0) = B(0) = I_n$ and $[A(p), B(q)]$ to be of full rank n (see Hannan, 1969). Call $M(p, q)$ the set of all such $k(z)$, i.e., that are rational matrix functions satisfying $k(z)$ analytic for $z \leq 1$ and $\det k \neq 0$, $|z| \leq 1$, having a left coprime representation $k = a^{-1}b$ with a, b of degrees p, q with $A(0) = B(0) = I_n$ and $[A(p), B(q)]$ of full rank. Then $M(p, q)$ may be mapped in a one to one manner into an open set in Euclidean space of dimension $(p + q)n^2$ by the elements of the $A(j)$, $B(j)$. However, there are still two problems. The $M(p, q)$ overlap and there are $k(z)$ belonging to no $M(p, q)$. An obvious example of the first is

$$k(z) = \begin{bmatrix} 1 & z \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -z \\ 0 & 1 \end{bmatrix}^{-1}$$

which therefore lies in $M(0, 1)$ and $M(1, 0)$. An example of the second, for $n = 2$, is $a(z) = I_2 + \frac{1}{2}Bz$, $b(z) = I_2 + Bz$, $k(z) = a(z)^{-1}b(z)$, where

$$B = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}.$$

Now $[\frac{1}{2}B, B]$ is clearly of rank unity so that $k(z)$ is not in $M(1, 1)$, and it can be shown it lies in no $M(p, q)$ for any value of p, q. We could achieve a unique representation for a given $k(z)$ by taking (say) $M(p, q)$ with q as small as possible and then p as small as possible. We shall discuss $M(p, q)$ again later.

To understand the nature of the problem we consider the, infinite, block "Hankel matrix," $\mathscr{H}$,

$$\mathscr{H} = [K(i + j - 1)]_{i,j=1,2,\cdots},$$

where the $n \times n$ block in the ith row of blocks and the jth column of blocks is shown. To see why $\mathscr{H}$ is important consider the following, which will also be used later. Put $y^{(t)} = (y(t)', y(t+1)', y(t+2)', \cdots)'$, $y_t = (y(t)', y(t-1)', y(t-2)', \cdots)'$ with analogous definitions for $\varepsilon^{(t)}$, ε_t. Then $y^{(t+1)}$ represents the future and ε_t or y_t, equivalently, represent the present and past. (Equivalently because of (1.4) and the definition of $\varepsilon(t)$.) Of course ε_t is uncorrelated with $\varepsilon^{(t+1)}$ in the sense that any pair of elements, one from each vector, is uncorrelated. Let $\mathscr{K} = [K(i - j)]$ where $i, j = 1, 2, \cdots$ and $K(j) = 0$, $j < 0$. Then from (1.4) it is easily seen that

$$(2.2) \qquad y^{(t+1)} = \mathscr{H}\varepsilon_t + \mathscr{K}\varepsilon^{(t+1)}.$$

The relation (2.2) is a regression relation (with infinite dimensional vectors) and exhibits $\mathscr{H}$ as the (infinite) matrix that shows the dependence of the future on the past.

$\mathscr{H}$ may also be used to give a general, basic representation for the process. Because of (1.5) each row of $\mathscr{H}$ is square summable, i.e., may be considered as an element of the vector space, l_2, of all square summable sequences x_j, $j = 1, 2, \cdots$, $\sum |x_j|^2 < \infty$. Let H_0 be a subset of the rows of $\mathscr{H}$ (as elements of l_2) that span all of the rows of $\mathscr{H}$. It may be that H_0 is finite dimensional but it would not have to be. If $r(u, j)$ is the jth row of the uth block of rows, $j = 1, \cdots, n$; $u = 1, 2, \cdots$, then H_0 might consist of $r(u, j)$, $j = 1, \cdots, n - 1$; $u = 1, 2, \cdots$. (This would be a case where $y_n(t \mid t-1)$ is a linear combination of the $y_j(t \mid t-1)$, $j = 1, \cdots, n - 1$. It is not interesting.) Partition H_0 as $H_0 = [B, H_2]$ where B is the first block of n columns. The special form of H_0 means that H_2

is composed of rows of $\mathscr{H}$ so that $H_2 = AH_0$. Indeed each row of H_2 is the same row of the next block of rows in $\mathscr{H}$. Since H_0 is a basis for the rows of $\mathscr{H}$ then $H_2 = AH_0$ follows. Call H_1 the first n rows of $\mathscr{H}$. Then certainly $H_1 = CH_0$. Put $x(t) = H_0\varepsilon_{t-1}$. Then $x(t+1) = B\varepsilon(t) + H_2\varepsilon_{t-1}$, $y(t) = H_1\varepsilon_{t-1} + \varepsilon(t)$ (see (2.2)). Thus

$$x(t+1) = Ax(t) + B\varepsilon(t), \quad y(t) = Cx(t) + \varepsilon(t). \tag{2.3}$$

This (prediction error form) state space representation, which derives directly from (1.4) is simple and basic. It exhibits $x(t)$ as Markovian (for $\varepsilon(t)$ Gaussian, for example) and $y(t)$ as a linear function of $x(t)$ plus the innovation. The vector $x(t)$ is called the state vector and the first equation in (2.3) is called the state equation.

Now we can consider how to classify the rational $k(z)$ in terms of (2.3). Consider the case where d, dimension of H_0 and hence of A and $x(t)$ and the rank of $\mathscr{H}$, is finite. *Then and only then the (matrix) function $k(z)$ corresponding to $\mathscr{H}$ is rational.* (For proofs of such propositions, here and below, the reader may consult Kailath, 1980.) Then also $(I_d - Az)x(t) = Bz\varepsilon(t)$, so that $y(t) = \{C(I_d - Az)^{-1}Bz + I_n\}\varepsilon(t)$ and

$$k(z) = I_n + Cz(I_d - Az)^{-1}B, \quad K(j) = CA^{j-1}B, \quad j \geq 1. \tag{2.4}$$

Thus it can be seen that rational transfer function approximation to stationary (purely nondeterministic) processes is a process of approximating $\mathscr{H}$ by a matrix of finite rank. *We shall always take d as minimal* to avoid redundancies.

To make this more concrete we exhibit (2.3) for the case (1.1). Put

$$y(t+j \mid t) = \sum_{u=j}^{\infty} K(u)\varepsilon(t+j-u).$$

This is the (unobservable) best linear predictor of $y(t+j)$ from $y(s)$, $s \leq t$. For $j = 1$ this is obvious since, using this definition of $y(t+1 \mid t)$, we have $y(t+1) - y(t+1 \mid t) = \varepsilon(t+1)$ which agrees with the definition of $y(t+1 \mid t)$ previously given. The general case is covered, for example, in Hannan (1970). Define

$$x(t+1)' = (y(t+1 \mid t)', y(t+2 \mid t)', \cdots, y(t+m \mid t)'), \quad m = \max(p, q)$$

and

$$A = \begin{bmatrix} 0 & I_n & 0 & \cdots & 0 \\ 0 & 0 & I_n & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I_n \\ -A(m) & -A(m-1) & -A(m-2) & \cdots & -A(1) \end{bmatrix}, \tag{2.5}$$

$$B' = [K(1)', K(2)', \cdots, K(m)'], \quad C = [I_n, 0, 0, \cdots, 0].$$

Then $y(t)$, $x(t)$ satisfy (2.3) with these A, B, C. Thus $x(t+1)$ is composed of predictions of $y(t+j)$ from $y(s)$, $s \leq t$. The special feature of the finite rank case is that all $y(t+j \mid t)$ are linearly expressible in terms of only d of the low lag predictions. The proof that (2.3) can be put in this form is not difficult but is omitted. It should be emphasized that for (1.1) this form of (2.3) is not the only possible form.

The representation (2.3) for $d < \infty$ connects directly with the Kalman filter, whose development in Kalman (1960) and Kalman and Bucy (1961) began an era when most of the work discussed herein was done. Here we show only the simple case corresponding to (2.3). For further details see Anderson and Moore (1979). The Kalman filter constructs a best estimate, $\hat{x}(t+1 \mid t)$ let us say, of $x(t+1)$ from $y(s)$, $1 \leq s \leq t$. The recursive formula is as follows:

$$\begin{aligned} \hat{x}(t+1 \mid t) &= A\hat{x}(t \mid t-1) + B(t)e(t), \\ y(t) &= C\hat{x}(t \mid t-1) + e(t), \\ \hat{x}(t+j \mid t) &= A\hat{x}(t+j-1 \mid t), \quad j > 1; \\ \hat{x}(1 \mid 0) &= 0, \\ B(t) &= \{AP(t)C' + B\Sigma\}\Sigma(t)^{-1}, \\ \Sigma(t) &= \{CP(t)C' + B\Sigma B'\}, \\ P(t+1) &= AP(t)A' + \Sigma - B(t)\Sigma(t)B(t)', \\ P(1) &= AP(1)A' + B\Sigma B'. \end{aligned} \tag{2.6}$$

The first line of (2.6) mimics (2.3) but now everything can be computed, given A, B, C, Σ. As $t \to \infty$ then $B(t) \to B$, $\Sigma(t) \to \Sigma$, $e(t) \to \varepsilon(t)$, (in mean square) and $P(t) = E[\{\hat{x}(t \mid t-1) - x(t)\}\{\hat{x}(t \mid t-1) - x(t)\}'] \to 0$.

Returning to (2.3) we observe that the only arbitrary element was the choice of H_0. This arbitrariness can be avoided by instituting a rule as to the choice of a basis for the rows of $\mathscr{H}$. A simple rule is to choose the first linearly independent set of rows found as you examine the rows of $\mathscr{H}$ from top to bottom. Any such basis is always of the form, for $d < \infty$,

$$r(u, j), \quad u = 1, \cdots, d_j, \quad j = 1, \cdots, n; \quad \sum_{1}^{n} d_j = d. \tag{2.7}$$

(This is because $r(d_j + 2, j)$ is a part of $r(d_j + 1, j)$ and so on.) Thus the matrix H_0, and a canonical form for (2.3), is specified by the so-called Kronecker indices, d_j. Their sum is the rank or "order" of $\mathscr{H}$, which is also called the "McMillan degree." Let us list the possible sets $\{d_j, j = 1, \cdots, n\}$ in dictionary order and let α stand for a typical index number of that ordering. Then call V_α the set of all $\mathscr{H}$, i.e. of all $k(z)$, for which the corresponding $\{d_j\}$ are the Kronecker indices. The V_α are disjoint since the sets of d_j are distinct. Also V_α

can be mapped in a one to one manner onto an open set in Euclidean space, in a manner we shortly describe. Thus, a complete listing has been prescribed for all rational $k(z)$. This depends on n integers, d_j, and given these there are continuously varying coordinates. The latter may be found as follows. (The following is, of necessity, brief.) We may represent $r(d_j + 1, j)$ in terms of earlier rows, by the definition of d_j. Thus there are $\tilde{\alpha}_{jk}(u)$ so that

$$\sum_{k=1}^{n} \sum_{u=0}^{d_{jk}} \tilde{\alpha}_{jk}(u) r(u+1, k) = 0, \qquad d_{jj} = d_j, \quad \hat{\alpha}_{jj}(d_j) = 1. \tag{2.8}$$

Because of the selection rule for H_0 we have

$$\begin{aligned} d_{jk} &< d_k, \quad j \neq k, \\ d_{jk} &\leq d_j, \quad k \leq j, \\ d_{jk} &< d_j, \quad k > j. \end{aligned} \tag{2.9}$$

Then applying (2.8) to (2.2) we obtain, after some manipulation,

$$\begin{aligned} \sum_{k=1}^{n} \sum_{u=0}^{d_j} \alpha_{jk}(u) y_k(t-u) &= \sum_{k=1}^{n} \sum_{u=0}^{d_j} \beta_{jk}(u) \varepsilon_k(t-u), \\ \alpha_{jk}(u) = \hat{\alpha}_{jk}(d_j - u), \quad \alpha_{jk}(0) &= \beta_{jk}(0), \\ & j, k = 1, \cdots, n. \end{aligned} \tag{2.10}$$

Of course (2.9) imposes additional constraints on (2.10), so that the set of all constraints is

$$\begin{aligned} &\alpha_{jk}(u) = 0, \quad u < d_j - d_{jk}, \\ &\alpha_{jk}(u) = 0, \quad u > d_j, \\ &\alpha_{jj}(0) = 1, \\ &\alpha_{jk}(0) = \beta_{jk}(0), \quad j, k = 1, \cdots, n. \end{aligned} \tag{2.11}$$

The coordinates may be taken as the $\alpha_{jk}(u)$, $\beta_{jk}(u)$ subject to these constraints and counting shows the dimension of the space into which V_α is mapped is

$$(n+1) \sum d_j + \sum_{j<k} \{\min(d_j, d_k) + \min(d_j, d_k + 1)\}. \tag{2.12}$$

Let $M(d)$ be the union of all V_α for $\sum d_j = d$. Thus, $M(d)$ is just the set of all $\mathscr{H}$, i.e. of all $k(z)$, for $\mathscr{H}$ of rank d, i.e. of McMillan degree d. It may be shown that $M(d)$ is a smooth surface in high dimensional space, of the nature of an algebraic variety. However, for $n > 1$, $M(d)$ cannot be globally mapped into Euclidean space, homeomorphically. Call $U(d)$ that $V_\alpha \subset M(d)$ for which, if $d = nr + s$, $0 \leq s < n$, then $d_j = r + 1, j = 1, \cdots, s$; $d_j = r, j = s + 1, \cdots, n$. Thus H_0 is then composed of the first d rows of $\mathscr{H}$. Then, also, $U(d)$ is open and dense in $M(d)$. All other $V_\alpha \subset M(d)$ have lower dimension, as can be seen from (2.12). It is in this sense that only *one* integer is needed to list all $k(z)$. The catch is that $U(d)$ is not all of $M(d)$, for $n > 1$, and other "neighborhoods" than $U(d)$ have to be introduced to cover $M(d)$. We describe $U(d)$ in detail by describing the $A(u), B(u)$ constituted by the $\alpha_{jk}(u)$, $\beta_{jk}(u)$ in (2.10), i.e. by describing the canonical ARMA form. We use an asterisk to indicate a submatrix of freely varying elements. All partitions are after row s or column s. The following description follows immediately from (2.11):

$$\begin{aligned} A(0) = B(0) &= \begin{bmatrix} I_s & 0 \\ * & I_{n-s} \end{bmatrix}, \quad A(1) = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix}, \\ A(r+1) &= \begin{bmatrix} * \\ 0 \end{bmatrix}, \quad B(r+1) = \begin{bmatrix} * \\ 0 \end{bmatrix}. \end{aligned} \tag{2.13}$$

All other $A(j), B(j), j = 0, 1, \cdots, r + 1$, are unconstrained. Note that though $A(0) = B(0)$ so that $K(0) = I_n$ yet, unless $s = 0$, $A(0) \neq I_n$. Of course we can replace $A(j)$, $B(j)$ by $A(0)^{-1}A(j)$, $A(0)^{-1}B(j)$ so that $A(0) = B(0) = I_n$, in conformity with (1.1), but that complicates the description, i.e., coordinatization.

One might compare the set of $U(d)$ with the set of all $M(p, q)$. Each set excludes some $k(z)$ but the set excluded is slender and is constituted by limit points of the set considered. There is no double listing with the $U(d)$ (they are disjoint). As p or q increases by unity the dimension of $M(p, q)$ increases by n^2. As d increases the dimension of $U(d)$ increases by $2n$. There is only one integer d. The advantage is with $U(d)$ except when $n = 1$. It may be observed that $U(d) = M(p, p)$, $d = np$. Thus for $n = 1$ then $U(d) = M(d, d)$.

Of course one can choose to determine all d_j but this is not computationally easy. The problem will be discussed in Section 6.

3. APPROXIMATION TO A KNOWN TRANSFER FUNCTION

If the transfer function is known then there is no statistical problem so that one might ask why the problem of approximation is being considered. The answer is that, in the first place, the resulting theory may be important in relation to the statistical problem even if it does not solve that problem. There has also been some statistical use of this theory, (Jewell, Bloomfield and Bartmann, 1983) commencing from an estimate of $k(z)$. Of course the philosophy of the inference then seems rather doubtful since the problem of choosing an initial estimate, suitable for the later approximation procedure, does not seem to have been faced. There are nonstatistical contexts where $k(z)$ might be known, for example experimental situations where the $K(j)$ can be observed because $\varepsilon(t)$ can be controlled.

To determine a best rational approximation requires that a measure of the closeness of approximation be instituted. Since $k(z)$ and $\mathscr{H}$ are in one to one correspondence a natural measure might be based on the norm of $\mathscr{H}$ as an operator in l_2. To avoid complexity let us assume that

$$(3.1) \qquad \sum_1^\infty j \, \| K(j) \|^2 < \infty.$$

Then $\operatorname{tr}(\mathscr{H}\mathscr{H}') < \infty$ and $\mathscr{H}$ has all of the essential properties of a (finite) matrix and in particular it has a singular value decomposition

$$\mathscr{H} = \sum_0^\infty \mu_j \eta_j \xi_j', \quad \eta_j' \eta_k = \xi_j' \xi_k = \delta_{j,k},$$

$$\mu_j \geq 0, \quad \sum \mu_j^2 < \infty, \quad \mu_0 \geq \mu_1 \geq \cdots.$$

Here, of course ξ_j, η_j are infinite columns of real numbers. Thus $\mathscr{H}\xi_j = \mu_j \eta_j$, $\mathscr{H}'\eta_j = \mu_j \xi_j$. The norm of $\mathscr{H}$ is μ_0, the greatest singular value. If $n = 1$, since then $\mathscr{H} = \mathscr{H}'$ also $\xi_j = \pm \eta_j$.

Let us take $n = 1$ and state one result relating to the use of $\| \mathscr{H} \|$, which is spoken of as the "Hankel norm," also, for $k(z)$. Adamyan, Arov and Krein (1971) show how to construct a function $\phi(z)$, assuming $\mu_d < \mu_{d-1}$,

$$(3.2) \qquad \phi(z) = \sum_{-\infty}^\infty \phi_j z^j$$

so that the Hankel matrix $\mathscr{H}_\phi = [\phi_{j+k-1}]$, *which uses only* ϕ_j *for* $j > 0$, satisfies $\| \mathscr{H} - \mathscr{H}_\phi \| = \mu_d$, $\mathscr{H}_\phi$ is of rank d and there is no other Hankel matrix of rank d, or less, attaining this accuracy. The function ϕ is said to be the symbol of $\mathscr{H}_\phi$. We shall, for completeness, in a moment show how to obtain $\phi(z)$. However first one might ask what has been achieved. It is true that, also,

$$(3.3) \qquad \sup_\omega | \phi(e^{i\omega}) - k(e^{i\omega}) | = \mu_d$$

so that if d is large, when μ_d will be small, then ϕ is uniformly close to k on the unit circle. However, it is only the ϕ_j for positive j that occur in $\mathscr{H}_\phi$. Of course the approximation we need to $k(z)$ has to involve only positive powers of j and this is not true of $\phi(z)$. The function

$$\sum_0^\infty \phi_j z^j$$

is a candidate as an approximation to k and is certainly rational of McMillan degree d, because of our assertion below (3.2) concerning $\mathscr{H}_\phi$, but a relation (3.3) does not hold for it.

The function $\phi(z)$ is constructed as follows. Let $\xi_j(k)$, $\eta_j(k)$ be the kth components of the (infinite) vectors ξ_j, η_j. Put

$$(3.4) \qquad \xi(z) = \sum_1^\infty \xi_d(k) z^{1-k}, \quad \eta(z) = \sum_1^\infty \eta_d(k) z^k.$$

Then $\phi(z) = k(z) - \mu_d \eta(z)/\xi(z)$. The result is remarkable but that alone does not make it useful to us. Since, for $n = 1$, $\xi_j = \pm \eta_j$ then $\eta(z) = \pm z \xi(z^{-1})$.

The case $n > 1$, $d < \infty$ has been treated by Glover (1984). Of course if $d < \infty$, one might ask, why is one seeking a rational approximation, since $k(z)$ is already rational. The answer is that d might be very large (as with a system composed of a large number of, subsidiary, rational transfer function systems in series) and one may wish to find a low order approximation, for example for control purposes. Glover bases his treatment on a "balanced realization," i.e., a special form of (2.3), for d finite. This is obtained as follows. Put

$$\mathscr{C} = [B, AB, A^2B, \cdots], \quad \mathscr{O} = [C', A'C', (A')^2C', \cdots]'.$$

Then it is easily seen, from (2.4), that $\mathscr{H} = \mathscr{O}\mathscr{C}$. Put $P = \mathscr{C}\mathscr{C}'$, $Q = \mathscr{O}'\mathscr{O}$. Factor Q as $Q = R'R$ (e.g., by Choleski factorization) and put $RPR' = VS^2V'$, $VV' = I_d$, where S is diagonal with positive diagonal elements. Thus this last is the singular value decomposition of RPR'. Put $T = S^{-1/2}V'R$. It is easily checked that if we make the change, in (2.3), $x(t) \to Txt(t)$, $A \to TAT^{-1}$, $B \to TB$, $C \to CT^{-1}$ then $\mathscr{H}$, k remain unchanged and $P \to TPT'$, $Q \to (T')^{-1}QT^{-1}$ and also $TPT' = (T')^{-1}QT^{-1} = S$. *Now assume we have made this change and that we call the new* $x(t)$, A, C, B *by their old names*, to avoid introducing a new notation. Then put $\xi_i = \sigma_i^{-1/2} c_i$, where c_i' is the ith row of $\mathscr{C}$ and σ_i is the ith element of S. (We may assume $\sigma_1 \geq \sigma_2 \geq \cdots$.) Put $\eta_i = \sigma_i^{-1/2} u_i$ where u_i is the ith column of $\mathscr{O}$. Then because $\mathscr{H} = \mathscr{O}\mathscr{C}$,

$$\mathscr{H} = \sum_1^d \sigma_i \eta_i \xi_i', \quad \eta_i' \eta_j = \xi_i' \xi_j = \delta_{jk},$$

the last part following from $P = Q = S$. This is the singular value decomposition of $\mathscr{H}$. However, if we merely omit all terms σ_i, η_i, ξ_i' for $i > r$, let us say, the resulting matrix will not be a Hankel matrix, in general. One thing to do is to omit all elements, $x_j(t)$, of $x(t)$ for $j > r$ in the form of (2.3) corresponding to our A, B, C. This is the same as omitting all corresponding rows of A, B and all corresponding columns of A, C. Let $\mathscr{H}_r$ be the Hankel matrix of the system with this truncated state vector. It is *not* that Hankel matrix, $\tilde{\mathscr{H}}_r$, such that $\| \mathscr{H} - \tilde{\mathscr{H}}_r | = \sigma_{r+1}$, but Glover shows that $\| \mathscr{H} - \mathscr{H}_r \| \leq 2(\sigma_{r+1} + \sigma_{r+2} + \cdots + \sigma_d)$ and if $k_r(z)$ is the transfer function for the truncated C, A, B then

$$\sup_\omega \| k(e^{i\omega}) - k_r(e^{i\omega}) \| \leq 2(\sigma_{r+1} + \sigma_{r+2} + \cdots + \sigma_d).$$

Thus now a reasonable kind of approximation has been found. This approximation is called the truncated, balanced approximation. Glover also shows, in this case of finite d, how to obtain the optimal Hankel norm approximation.

In any case it is not clear that this norm for $\mathscr{H}$ is a suitable measure of closeness with which to work. Let us return to (2.2). We assume that the reader is familiar with the theory of canonical correlation. (See Anderson, 1958, Chapter 12.) This theory comes naturally to mind since (2.2) is a regression relation satisfying the classical requirement that the vector of independent (regressor) variables is orthogonal to the vector of residuals. The only peculiarity is the fact that all vectors and matrices are infinite. However under (3.1), for example, that is of no concern at all and the definitions can be fully carried over. In canonical correlation one finds a linear function, u_0, of $y^{(t+1)}$ and a corresponding linear function, v_0, of ε_t (called discriminant functions) so that their correlation is as high as possible. Then another pair is found, orthogonal to the first pair, so that their correlation is as high as possible, and so on. In our case this can be described as follows. We need first to replace $y^{(t+1)}$, ε_t by vectors whose components are uncorrelated. To do this for ε_t is easy, we just form $(I_\infty \otimes \Sigma^{-1/2})\varepsilon_t$ where $\Sigma^{1/2}$ is the unique positive square root of Σ and $I_\infty \otimes \Sigma^{-1/2}$ is a matrix that is block diagonal with $\Sigma^{-1/2}$ constituting all diagonal blocks. For $y^{(t+1)}$ we need to be cleverer. However, using the definition of $\Gamma(j)$ in Section 2,

$$E(y^{(t+1)}y^{(t+1)\prime}) = [\Gamma(j-i)]_{i,j=1,2,\cdots}$$

where the (i, j)th block is shown. This matrix corresponds to a stationary process with spectrum (see (2.1))

$$\frac{1}{2\pi}\sum_{-\infty}^{\infty}\Gamma(-j)e^{-ij\omega} = f(-\omega).$$

(The change from ω to $-\omega$ is due to the time reversal coming from the *increase* in the argument s in $y(s)$ as you go down the blocks of $y^{(t+1)}$.) Now, unless $n = 1$, $f(-\omega) \neq f(\omega)$ but has all of the essential properties so that as in (2.1) there is a unique factorization

$$f(-\omega) = \frac{1}{2\pi}\, l(e^{i\omega})\Omega l(e^{i\omega}), \quad \Omega > 0,$$

$$l(z) = \sum_{0}^{\infty}\Lambda(j)z^j, \quad \Lambda(0) = I_n.$$

Let $\mathscr{L}$ have $\Lambda(k-j)$ as the (j, k)th block, $\Lambda(j) = 0$, $j < 0$. Then

$$E\{\mathscr{L}^{-1}y^{(t+1)}y^{(t+1)\prime}(\mathscr{L}^{-1})'\} = I_\infty \otimes \Omega.$$

Indeed this is equivalent to saying that $y(t) = \sum \Lambda(j)\tilde{\varepsilon}(t+j)$, $E\{\tilde{\varepsilon}(s)\tilde{\varepsilon}(t)'\} = \delta_{st}\Omega$, so that $y^{(t+1)} = \mathscr{L}\tilde{\varepsilon}^{(t+1)}$. Thus $(I_\infty \otimes \Omega^{-1/2})\mathscr{L}^{-1}y^{(t+1)}$ is the replacement for $y^{(t+1)}$. This means that we have a regression relation in these two new vectors with matrix of regression coefficients

$$(3.5) \qquad \mathscr{G} = (I_\infty \otimes \Omega^{-1/2})\mathscr{L}^{-1}\mathscr{H}(I_\infty \otimes \Sigma^{1/2}).$$

It is well known that one finds the discriminant function by finding the singular value decomposition for $\mathscr{G}$, namely,

$$\mathscr{G} = \sum_{0}^{\infty}\rho_j\zeta_j\chi_j', \quad \zeta_j'\zeta_k = \chi_j'\chi_k = \delta_{jk},$$
$$\rho_0 \geq \rho_1 \geq \rho_2 \geq \cdots \geq 0.$$

The discriminant functions are

$$u_j = \chi_j'(I_\infty \otimes \Omega^{-1/2})\mathscr{L}^{-1}y^{(t+1)}, \quad v_j = \zeta_j'(I_\infty \otimes \Sigma^{-1/2})\varepsilon_t$$

and satisfy

$$E(u_ju_k) = E(v_jv_k) = \delta_{j,k}, \quad E(u_jv_k) = \delta_{jk}\rho_j,$$
$$j = 0, 1, \cdots.$$

This canonical correlation reduction must relate closely to the structure of the process since it so completely describes the relation between the future of the process and its present and past. Because of the special nature of $\mathscr{G}$ (see (3.5)) it is again a Hankel matrix. If $G(j+k-1)$ is its typical block then the $G(j)$ are generated by

$$g(z) = \Omega^{-1/2}l(z^{-1})^{-1}k(z)\Sigma^{1/2}$$
$$= \sum_{-\infty}^{\infty}G(j)z^j.$$

However, note that now we do not have $G(j) = 0$, $j < 0$. In case $n = 1$ then $l(z) = k(z)$ (since then $f(\omega) = f(-\omega)$) and $\Omega = \Sigma$ so that $g(z) = k(z)/k(z^{-1})$, which is of modulus 1 on $z = e^{i\omega}$. In general $g(e^{i\omega})$ is unitary.

It is possible for $n = 1$ to use the theory in Adamyan, Arov and Krein (1971) to construct an approximation to $k(z)/k(z^{-1})$. (The construction that will now be described is not used in the rest of this article and may be omitted.) Of course this function, unlike $k(z)$, is not analytic for $|z| \leq 1$ but that does not affect the construction since only the Hankel matrix is used. However, further steps have to be taken. Say $\tilde{\phi}(z)$ was constructed as was $\phi(z)$, below (3.4), but commencing from $zk(z)/k(z^{-1}) = \tilde{g}$, let us say. Then $\tilde{g}$ is the symbol of a Hankel matrix $\tilde{\mathscr{G}} = \mathscr{L}^{-1}\tilde{\mathscr{H}}$,

$$\tilde{\mathscr{H}} = [K(i+j-2)]_{i,j=1,2,\cdots}$$

which enters into the relation $y^{(t)} = \tilde{\mathscr{G}}\varepsilon_t + \tilde{\mathscr{K}}\varepsilon^{(t)}$, where $\tilde{\mathscr{K}}$ has $K(i-j-1)$ as the (i, j)th block, $K(j) = 0, j < 0$ (so that the first row of blocks is null). The canonical correlations are now those between present and future and present and past and thus, certainly, one of them is unity. The introduction of the factor z in $\tilde{g}(z)$ is connected with the fact that $\eta(z) = \pm z\xi(z^{-1})$ (see below (3.4)). Jonckheere and Helton (1985) now proceed as follows, making some further assumptions that we do not detail but which include assuming that all singular values of $\tilde{\mathscr{G}}$ are of unit multiplicity.

Commencing from $\tilde{\phi}(z)$ construct a further approximation, $\phi_0(z)$ to it taking $d = 0$. Now $\phi_0(z)$ can have no coefficients of z^j for $j > 0$ since $d = 0$, i.e. $\mathscr{H}_\phi$ is null. Thus $\phi_0(z) = \tilde{\phi}(z) - \tilde{\mu}_0 \tilde{\eta}_0(z)/\tilde{\xi}_0(z)$, where $\tilde{\eta}_0$, $\tilde{\xi}_0$ are constructed as in (3.4) but commencing from $\tilde{\phi}(z)$ and where $\tilde{\mu}_0$ is the largest singular value of the Hankel matrix with symbol $\tilde{\phi}(z)$. Since $\| \tilde{\mathscr{G}} \| = 1$, it is not unreasonable to assume $\tilde{\mu}_0$ is 1 or is very near to it. But then

$$\begin{aligned} \rho_d &= \sup \left| \left\{ \frac{e^{i\omega} k(e^{i\omega})}{k(e^{-i\omega})} \right\} - \tilde{\phi}(e^{i\omega}) \right| \\ &= \sup \left| \left\{ \frac{e^{i\omega} k(e^{i\omega})}{k(e^{-i\omega})} \right\} - \phi_0(e^{i\omega}) - \frac{\tilde{\mu}_0 \tilde{\eta}_0(e^{i\omega})}{\tilde{\xi}_0(e^{i\omega})} \right| \\ &= \sup \left| \left\{ \frac{e^{i\omega} k(e^{i\omega})}{k(e^{-i\omega})} \right\} - \frac{\tilde{\mu}_0 \tilde{\eta}_0(e^{i\omega})}{\tilde{\xi}_0(e^{i\omega})} \right| \end{aligned}$$

since ϕ_0 does not contribute to the Hankel matrix and, because of (3.3), that is all that determines the accuracy of the first approximation. Examining (3.4) we see that $\tilde{\eta}_0(z)/\tilde{\xi}_0(z) = z\hat{k}(z)/\hat{k}(z^{-1})$ where

$$\hat{k}(z) = c. \sum_1^\infty \tilde{\eta}_0(k) z^{k-1}.$$

The constant c may be chosen at will. The function $\hat{k}(z)$ is rational. Thus, an approximation has been constructed of the requisite form. The constant c might be chosen, for example, to provide a good approximation, via $\hat{k}(z)$, to the spectrum of $y(t)$. However, the whole procedure has an ad hoc atmosphere about it. The initial approximation by $\tilde{\phi}(z)$ is unsatisfactory so an expedient is used to convert it into $\tilde{\eta}_0/\tilde{\xi}_0$. The factorization of this again involves arbitrary elements, although no doubt c could be chosen on some objective basis. A number of assumptions are involved, e.g. $\tilde{\mu}_0 = 1$. If $k(z)$ was, in the first place, an estimate what real justification is there for the final result?

The whole procedure is susceptible to generalization to the case $n > 1$, at least when $\mathscr{H}$ has finite, albeit high, rank, d. We could, for example, use balanced truncation to find a first approximation to $zg(z)$, where $g(z)$ was defined above, and then obtain a further unitary approximation to that, which finally will have to be factored to give approximations to $k(z)$, $l(z)$. The procedure then is more complex and the ad hoc nature appears to be even more pronounced.

The ideas are clearly deep and important. As indicated earlier they relate to basic characteristics of the systems. For example it can be shown that when $k(z)$ is rational then the number of unit singular values, i.e. the number of unit canonical correlations between the future and the present and past, is the number of zeros of det $k(z)$ on $|z| = 1$ counting these with their multiplicities (Hannan and Poskitt, 1986). Of course such zeros have been excluded, by fiat, by us and this result partly validates that decision because it is not to be expected that any (linear) function of the future will be exactly predictable from the present and past.

The statistical significance of the results of this section are, however, not yet apparent and we turn therefore to a direct statistical consideration of the approximation problem. This requires also a criterion on the basis of which to choose an approximant.

4. APPROXIMATION CRITERIA

Akaike (1969) seems to have been the first to recognize the problem to be faced when the dimension of the parameter space is allowed to increase indefinitely. In the reference cited he had in mind an autoregression but the general principle is the same. Thus an integer d or a set $d = (d_1, d_2, \cdots, d_n)$, for example, of integers prescribes the parameter space and the dimension $\nu(\theta)$ of the parameter vector depends on d. In counting parameters we shall exclude the $n(n+1)/2$ elements (on and above the main diagonal) that specify Σ, since they do not involve d. Thus for $M(d)$ we have $\nu(\theta) = 2nd$, for $M(p, q)$, where $d = (p, q)$, we have $\nu(\theta) = (p+q)n^2$ and for V_α we have $\nu(\theta)$ given by (2.12). Akaike's procedure is to choose d to minimize

$$\text{(4.1)} \qquad \text{AIC}(d) = \log \det \hat{\Sigma}_d + 2\nu(\theta)/T.$$

Here $\hat{\Sigma}_d$ is the maximum likelihood estimate of Σ, on the basis of a Gaussian likelihood. It is necessary to restrict (4.1) so that very large values of d will not be examined since that can lead to bad results but these restrictions do not seem to be prescribed in practice and the chosen d seems always to be reasonably small. The criterion (4.1) can be thought of as arising as follows (Findley, 1985). Let $E_T(\theta) = E(L_T(\theta))$ where $L_T(\theta)$ is the log likelihood and expectation is with respect to the true probability law. Then $E_T(\theta)$ is maximized at θ_0, the true probability structure. Thus, given maximum likelihood estimators, $\hat{\theta}$, $\hat{\phi}$, based on different model sets one might prefer $\hat{\phi}$ to $\hat{\theta}$ if $E_T(\hat{\phi}) > E_T(\hat{\theta})$. These last are not known but, as Findley shows, $L_T(\hat{\phi}) - L_T(\hat{\theta}) - \nu(\hat{\theta}) + \nu(\hat{\phi})$ is an asymptotically unbiased estimate of $E_T(\hat{\phi}) - E_T(\hat{\theta})$ so that, under Gaussian ARMA assumptions, (4.1) gives a valid basis for comparison. The results in Findley (1985) are derived under conditions that we shall not detail here. They do not involve a Gaussian law for $y(t)$ but do involve that $\lim_{T\to\infty} E_T(\theta)$ be maximized at a point θ_∞ that is an interior point of the space over which θ varies.

Though this basis for the use of AIC has some appeal it is not entirely convincing. Nevertheless some results will be quoted in the next section showing that AIC is optimal for some situations. A different approach has been introduced by Rissanen (1978, 1983, 1986). His approach is, initially at least, marked by a resolute refusal to commit "the fallacy of misplaced concreteness," that is the fallacy of attributing reality to the model. He considers encoding the data, using a model to determine the optimal encoding. That model is then chosen which gives the smallest code length. The code must be one that can be read by any decoder armed only with the knowledge of the general coding principles and of the family of models used but not with any knowledge of the data. This is Rissanen's "minimum description length" principle. It corresponds closely to one of the three definitions of statistics given in Fisher (1944), namely as the science of "reducing" data.

In Rissanen (1978, 1983) a "nonpredictive" principle is used. Assume the data to be suitably quantized so that a chosen number of digits is used in a binary representation. The model prescribes probabilities $P_{d,\theta}(y_T)$. If the model is true then it is known that an optimal encoding will have length approximately—$\log_2 P_{d,\theta}(y_T)$ bits. To decode, however, also the parameter vector, θ, must be available so that this also must be transmitted. A decision must be made as to how many digits should be retained for the elements of θ and a reasonable rule, if θ were scalar, would be to determine that by the standard deviation of an optimal estimate of θ (i.e. the maximum likelihood estimator) on the basis of $P_{d,\theta}(y_T)$. Since only finitely many values of θ will be involved for any d, T then these can be arranged in order according to some preassigned rule (known to the decoder) and only an integer has to be transmitted giving the index of θ in that ordering (and d). Rissanen chooses to encode the integers on the basis of a choice made on a minimax principle (see Rissanen, 1983). This choice allocates, to the integer k, $\log^* k + c^*$ binary digits where

$$\log^* k = \log_2 k + \log_2 \log_2 k + \log_2 \log_2 \log_2 k + \cdots,$$

the summation being up to the last positive term. Of course $\log^* k/\log_2 k \longrightarrow 1$. The integer $c^* \approx 2.865$ and is chosen so that

$$\sum_1^\infty 2^{-(\log^* k + \log_2 c^*)} = 1.$$

Let $L(d, \theta)$ be the length of the code for θ. Fairly evidently $L(d, \theta) \approx \frac{1}{2}\nu(\theta)\log_2 T$ because the accuracy of the best estimate of θ will be $O(T^{-1/2})$ so there will be $\log_2\{O(T^{1/2})\}$ bits retained for each element of θ and thus about $\frac{1}{2}\nu(\theta)\log_2 T$ in all. In any case

$$(4.2) \qquad -\log_2 P_{d,\theta}(y_T) + L(d, \theta)$$

is the total code length to be used. In a Gaussian ARMA context then $2T^{-1}$ by (4.2) becomes, to a first approximation, after optimization with respect to θ for given d, and after the introduction of a constant factor due to the change of base of the logarithms,

$$(4.3) \qquad \mathrm{BIC}(d) = \log \det \hat{\Sigma}_d + \nu(\theta)\log T/T,$$

which is to be compared to (4.1).

To avoid the complex argument involved in the encoding of θ Rissanen (1986) proceeds as follows, using what he calls a "predictive minimum description length principle." Let $f_{d,\theta}(y(t+1) \mid t)$ be the probability density for the (finitely many) values of $y(t+1)$ conditional on $y(s)$, $1 \le s \le t$. Then

$$(4.4) \qquad -\sum_{j=0}^{t-1} \log_2 f_{d,\theta}(y(j+1) \mid j)$$

gives the code length for the data at time t in an optimal encoding. Indeed this is just $-P_{d,\theta}(y_t)$ rewritten. Let $\theta(t)$ minimize (4.4) *for given d* and choose $d(t)$ to minimize

$$-\sum_{j=0}^{t-1} \log_2 f_{d,\theta(j)} f(y(j+1) \mid j).$$

(This is not the same as choosing $\theta(t)$, $d(t)$ to minimize (4.4).) Then there is no need to encode $d(t)$, $\theta(t)$ also since at time $t+1$ the $y(s)$, $1 \le s \le t$, will have been decoded and thus $d(t)$, $\theta(t)$ will be known and $y(t+1)$ decoded. Thus

$$(4.5) \qquad -\sum_{t=0}^{T-1} \log_2 f_{d(t),\theta(t)}(y(t+1) \mid t)$$

is a measure of a total code length for $y(1), \cdots, y(T)$ that is a valid encoding, i.e., that can be decoded. We now use $d(T-1)$, $\theta(T-1)$ as the final estimates. In some cases (4.4) could feasibly be calculated. For example consider the autoregressions (i.e. (1.1) for $q = 0$) which we write as

$$(4.6) \qquad \begin{gathered} \sum_0^h \Phi_h(j) y(t-j) = \varepsilon(t), \quad \Phi_h(0) = I_n, \\ \det\left\{\sum_0^h \Phi_h(j) z^j\right\} \neq 0, \quad |z| \le 1. \end{gathered}$$

Here θ is composed of the elements of the $\Phi_h(j)$, $j = 1, \cdots, h$ and $\nu(\theta) = hn^2$. There is an extensive literature concerning the calculation of (approximations to) the maximum likelihood estimators, on Gaussian assumptions. In particular algorithms have been developed (Friedlander, 1982) that are recursive on T and h. This is what is needed for (4.5). We shall further discuss such procedures in Section 6. However in general (4.5) would be difficult to compute and Rissanen (1986) considers the "semipredictive" principle based on

$$(4.7) \qquad -\sum_0^{T-1} \log_2 f_{d,\theta(t)}(y(t+1) \mid t) + \log^* d$$

where now d is held fixed and $\theta(t)$ is optimal for that fixed d. Now d has also to be transmitted, hence the term $\log^* d$. Then $\hat{d}$ is chosen to optimize (4.7).

Rissanen (1986) shows that (4.2), (4.5) and (4.7) do give a minimum description length, at least asymptotically and under certain conditions. The conditions include the true structure being in the model set and Gaussian requirements. No doubt both of these can be relaxed so that, for example, the true structure was only a, suitably defined, limit point of the model set.

Any theory, leading to criteria such as (4.1), (4.2) and (4.7) is of an abstract kind and should be treated with caution. It is doubtful if any principle can encompass all problems of statistical inference. What is needed is the proof of theorems relating to such criteria that show that they have good properties, plus experience with their practical use. It should be mentioned in this connection that no theorems, of the nature of those given below (5.6) in the next section, are so far available for (4.7). We go on to such questions in the next section.

5. SOME PROPERTIES OF ORDER ESTIMATES

It is necessary to restrict $\varepsilon(t)$ further if worthwhile results are to be obtained. The discussion in Section 2 shows that a very wide range of phenomena can be represented as in (1.4) with (1.5), (1.6) holding and this would include data generated by highly nonlinear mechanisms. A natural restriction, considering the linear aspects of the models, is to require the processes to be linear in the sense that the best linear predictor is the best predictor. Calling $\mathscr{F}_t$ the σ algebra of events determined by $y(s)$, $s \le t$, then the best predictor is $E\{y(t) \mid \mathscr{F}_{t-1}\}$ so that we require

$$\varepsilon(t) = y(t) - y(t \mid t-1) = y(t) - E\{y(t) \mid \mathscr{F}_{t-1}\}$$

and thus

$$(5.1) \qquad E\{\varepsilon(t) \mid \mathscr{F}_{t-1}\} = 0.$$

Conversely this implies that

$$y(t \mid t-1) = E\{y(t) \mid \mathscr{F}_{t-1}\}.$$

The $\varepsilon(t)$ are then vectors of stationary, ergodic martingale differences, with finite variances, to which a wide range of the asymptotic theory for independent identically distributed random variables carries over. For some purposes more is needed. Consider $n = 1$ and $E\{\varepsilon(s)\varepsilon(s-1)\varepsilon(t)\varepsilon(t-1)\}$. If $s > t$ this is zero using (5.1) but for $s = t$ it will involve fourth moments unless (in general, i.e. for $n > 1$)

$$(5.2) \qquad E\{\varepsilon(t)\varepsilon(t)' \mid \mathscr{F}_{t-1}\} = \Sigma.$$

Then $E\{\varepsilon(s)^2\varepsilon(s-1)^2\} = \sigma^4$, using (5.2) for $n = 1$. Thus when (5.2) holds, all quantities of the $T^{-1/2} \sum \varepsilon(t)\varepsilon(t-j)$, $j > 0$, will have the same variance and covariances as if $y(t)$ were Gaussian. As a result many estimation procedures based on Gaussian likelihoods will have the same properties as if $y(t)$ were Gaussian if (5.1), (5.2) hold. We also need

$$(5.3) \qquad E\{\varepsilon_j(t)^4\} < \infty, \quad j = 1, \cdots, n,$$

although fourth moments do not enter into most final formulae. The condition (5.2) is to be avoided if possible. For example consider the model $y(t) = \{\rho + \eta(t)\}y(t-1) + \varepsilon(t)$ where the $\eta(t)$ are stationary, serially independent and generated independently of the $\varepsilon(t)$ sequences. This is a time varying parameter model. Then $y(t) = \rho y(t-1) + \xi(t)$ where $\xi(t) = \eta(t)y(t-1) + \varepsilon(t)$. Then $\xi(t)$ satisfies a condition of the form of (5.1) but not (5.2), as is easily checked. Thus theorems using only the former will hold for the classic estimate of ρ. A condition much weaker than (5.2) that is then needed is

$$(5.4) \qquad E\{\varepsilon(t)\varepsilon(t)' \mid \mathscr{F}_{-\infty}\} = \Sigma$$

and this does hold for $\xi(t)$, above, in the sense that $E\{\xi(t)^2 \mid \mathscr{F}_{-\infty}\} = E\{\xi(t)^2\}$. We shall present some results for the scalar case for simplicity but completely analogous results hold in general. We consider a generalization of AIC, BIC of the form

$$(5.5) \quad \log\det \hat{\Sigma}_d + \nu(\theta)C_T/T, \quad C_T/T \to 0, \quad d \le D,$$

which for $n = 1$ and $M(p, q)$ becomes

$$\log \hat{\sigma}_{p,q} + (p+q)C_T/T, \quad C_T/T \to 0, \quad p \le P, \quad q \le Q.$$

For BIC, $C_T = \log T$ and for AIC it is 2. Of course $\hat{\sigma}^2_{p,q}$ is the maximum likelihood estimator. It is assumed that θ_0, the true structure is in $M(p_0, q_0)$ for $p_0 \le P$, $q_0 \le Q$. It is also necessary to assume that it is known that

$$(5.6) \qquad b_0(z) \ne 0, \quad |z| \le 1 + \delta, \quad \delta > 0,$$

i.e., that the zeros of the moving average part are bounded away from the unit circle. Call $\hat{p}$, $\hat{q}$ the estimators minimizing (5.5) subject to (5.6). *Let* (5.1), (5.3), (5.4), (5.6) *hold. Then there are* c_0, c_1, $0 < c_0 \le c_1 < \infty$, *so that the following hold.*

(a) *If* $\liminf_{T\to\infty} C_T/[2 \log\log T\} > c_1$ then $(\hat{p}, \hat{q}) \to (p_0, q_0)$ *a.s. If* (5.2) *holds we may take* $c_1 = 1$.
(b) *If* $\limsup_{T\to\infty} C_T/\{2 \log\log T\} < c_0$ *then the almost sure convergence in* (a) *does not hold and if* (5.2) *holds we may take* $c_0 = 1$.
(c) *If* $C_T \to \infty$ *then* $\operatorname{plim}_{T\to\infty}(\hat{p}, \hat{q}) = (p_0, q_0)$.
(d) *If* $\limsup C_T < \infty$ *then* (c) *fails and if also* (5.2) *holds then*

$$\lim_{\delta\to 0} \lim_{T\to\infty} P(\hat{p} > p_0, \hat{q} > q_0) = 1.$$

In (d) it must be recalled that $\hat{p}$, $\hat{q}$ are estimated subject to (5.6) so that the probability in the assertion depends on δ. This theorem (due to Hannan (1980) with generalizations in Hannan (1981) and relaxations of conditions due to An, Chen and Hannan (1982), Hannan and Kavalieris (1983)) fairly completely covers the case where the true order is finite. The last part of (d) depends on the fact that $\hat{\sigma}^2_{p,q}$ is the maximum likelihood estimator and may not hold true if some approximative procedure is used (e.g., one or two steps of a Gauss-Newton iteration). This last result is of little practical value, although of some interest, since T may have to be large before it is relevant.

This theorem grossly favors BIC over AIC but caution must be used in its interpretation. Some price must be paid for taking C_T larger, namely an increase, for fixed T, in the chance of underestimation of the order. In any case the system will not be of finite order. To discuss this further consider the case where in (1.4)

$$\sum j^{1/2} \| K(j) \| < \infty.$$

Then putting $\phi(z) = k(z)^{-1} = \sum \Phi(j)z^j$, $\Phi(0) = I_n$, we have

$$\sum_1^\infty \Phi(j)y(t-j) = \varepsilon(t), \quad \sum_1^\infty j^{1/2} \| \Phi(j) \| < \infty. \tag{5.7}$$

A standard procedure for estimating (4.6), which provides a procedure for estimating (5.7), is the solution of the equations

$$\sum_0^h \hat{\Phi}_h(j)C(j-k) = \delta_{0k}\hat{\Sigma}_h, \qquad k = 0, 1, \cdots h; \quad \hat{\Phi}_h(0) = I_h. \tag{5.8}$$

These equations can lead to bad biases if T is not large (Tjøstheim and Paulsen, 1983) and there is an extensive literature concerning that. A simple way of eliminating much of that bias is to taper the data, that is to replace $y(t)$ by $a(t/T)y(t)$, $t = 1, 2, \cdots, T$. Here $a(x)$ might be chosen to be $\frac{1}{2}\{1 + \cos(2\pi x/\rho)\}$, $x \in [0, \rho/2]$; $a(x) = 1$, $x \in [\rho/2, \frac{1}{2}]$; $a(x) = a(1-x)$, $x > \frac{1}{2}$. Thus $a(x)$ is unity over all but a proportion ρ of its support and fades to zero at the ends of its interval of support. Dahlhaus (1984) shows that this procedure materially reduces the bias. An alternative is to use a normalized, lattice algorithm. For details and references concerning these see Friedlander (1982). We shall briefly discuss these in Section 6. *It is probable that one of these methods of eliminating the bias in the Toeplitz procedure* (5.8) *should always be used.* The $\hat{\Phi}_h(j)$ are good estimates of the $\Phi(j)$ if h is large as is shown by the following result. If (5.1), (5.3) and (5.4) hold then

$$\max_{1 \le j \le h} \| \hat{\Phi}_h(j) - \Phi(j) \| = O\{(\log T/T)^{1/2}\} + \{c + o(1)\} \sum_{h+1}^\infty \| \Phi(j) \|$$

where c depends only on the true structure. This result is uniform in $h \le H_T = O\{(T/\log T)^{1/2}\}$ and the order relations hold a.s. If (5.2) holds the $o(1)$ term is also $O\{(\log T/T)^{1/2}\}$ but it will be of that order also under much weaker conditions. Thus the accuracy is ultimately determined by the size of h and the speed of decrease of the $\| \Phi(j) \|$.

Perhaps more importantly under (5.1), (5.2), (5.4) and a slight strengthening of (5.2) we have

$$\begin{aligned} \log \det \hat{\Sigma}_h + hC_T/T &= \log \det\left\{\frac{1}{T}\sum_1^T \varepsilon(t)\varepsilon(t)'\right\} \\ &\quad + \{1 + o_P(1)\}hn^2(C_T - 1)/T \\ &\quad + \operatorname{tr}\{\hat{\Sigma}^{-1}(\Sigma_h - \Sigma)\}. \end{aligned} \tag{5.9}$$

This kind of relation was first stated in Shibata (1980) and is stressed further in Hannan and Kavalieris (1984). Here $\Sigma_h = E\{\varepsilon_h(t)\varepsilon_h(t)'\}$, $\varepsilon_h(t) = \sum_0^h \Phi_h((j) \cdot y(t-j)$ and the $\Phi_h(j)$ have been chosen, subject to $\Phi_h(0) = I_n$, to make $E\{\varepsilon_h(t)'\varepsilon_h(t)\}$ as small as possible. Clearly $\hat{\Sigma}_h$ is an estimate of Σ_h. The relation (5.9) is crucial as we now explain. It is clear that the a.s. accuracy of $\hat{\Sigma}_h$ as an estimator of Σ_h cannot exceed that given by the law of the iterated logarithm. But that is $O\{(\log\log T/T)^{1/2}\}$ which is much bigger than $\log T/T$. The position is saved by the fact that in (5.9) the error of order $O\{(\log\log T/T)^{1/2}\}$ is in the first term which is independent of h. The result (5.9) shows that $\hat{h}$ is essentially determined by h^*, where that minimizes

$$\mathscr{L}_n(C_T, h) = hn^2(C_T - 1)/T + \operatorname{tr}\{\Sigma^{-1}(\Sigma_h - \Sigma)\}. \tag{5.10}$$

As a result it can be shown that when (5.8) is used for $n = 1$ and $k(z)$ is rational then $\hat{h} = \{\log T/(2 \log \rho_0)\} \cdot \{1 + o(1)\}$, ρ_0 being the modulus of a zero of $k(z)$ nearest to $|z| = 1$. Moreover Shibata (1981) obtains the following interesting result. Let us put, for $n = 1$,

$$2\pi\hat{f}_h(\omega) = \hat{\sigma}_h^2 \left| \sum_0^h \hat{\phi}_h(j)e^{ij\omega} \right|^{-2},$$

again using lower case letters for $n = 1$. (See (2.1) for the motivation of $\hat{f}_h(\omega)$.) We may consider as a measure of the merit of this estimate the quantity

$$m_h(T) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \frac{\hat{f}_h(\omega) - f(\omega)}{f(\omega)} \right|^2 d\omega.$$

Shibata shows that $m_h(T)/\mathscr{L}_1(2, h)$ converges in probability to 2, uniformly in $1 \leq h \leq H_T$, provided $h \to \infty$. This choice of C_T is of course that for AIC and as a consequence it follows that $m_h(T)$ is optimized, asymptotically, when h is chosen by AIC. Since $m_h(T)$ is a reasonable measure this is an argument in favor of AIC. However even this result must be viewed with care. Franke, Gasser and Steinberg (1985) considered the classification of individuals on the basis of an examination of estimates of spectra of EEG records, the classification being compared with that effected by experts. They found that AIC did badly on this problem compared to BIC. The reason may perhaps be seen from $m_h(T)$. The classification is mainly on the basis of predominant features of the spectrum but $m_h(T)$ treats all features equally in the sense that it is relative error that matters. Thus since AIC optimizes in relation to $m_h(T)$ it may choose too high an order (as found in the cited reference) and thus introduce too much random variation from individual to individual in the predominant features, thus making classification difficult.

There is no means by which it can be established that AIC is always to be preferred to BIC, or the reverse. Rissanen's general principle has great appeal but in practice must be used via *ad hoc* choices (such as Gaussian likelihoods). These can, no doubt, be shown to be somewhat immaterial, asymptotically, as the results of this section indicate. In practice the purpose of the statistical analysis may be very special so that, if the criterion is to relate to the purpose, the nature of the data to be summarized will need also to be related to that purpose. Thus in the situation discussed in Franke, Gasser and Steinberg (1985) the data might be, for each individual, a "filtered" form of the original record in which frequencies, other than those in the band of frequencies of interest, are eliminated. (The "filtering" might be done via fast Fourier transformation.) The problem is not easy to phrase in a general form.

6. ALGORITHMS

Apart from the fast Fourier transform algorithms and the Kalman filter apparatus the most important algorithms in time series analysis are those that compute, in an effective manner, (5.8). It is evident that this must be important since an iterative optimization will consist of a series of linear calculations, i.e., autoregressions. The basic form of such an algorithm was originally discovered by Levinson but in the vector form needed here is due to Whittle (1963), where references to earlier developments (see Durbin, 1960) may also be found. Let us put this in a general setting so that it can be used via special identifications later. Thus let the data be $v(t)$, $t = 1, \cdots, T$, where $v(t)$ has s components. Then $C_v(j)$ is computed from the $v(t)$ as was $C(j)$ from the $y(t)$ in Section 2. The algorithm computes $F_h(j)$, $j = 0, 1, \cdots, h$, which correspond to the $\hat{\Phi}_h(j)$ computed in (5.8), by a recursion on h as follows:

$$
\begin{aligned}
F_h(j) &= F_{h-1}(j) + F_h(h)\tilde{F}_{h-1}(h-j),\\
\tilde{F}_h(j) &= \tilde{F}_{h-1}(j) + \tilde{F}_h(h)F_{h-1}(h-j),\\
F_h(h) &= -\Delta_{h-1}\tilde{S}_{h-1}^{-1},\\
\tilde{F}_h(h) &= -\Delta'_{h-1}S_{h-1}^{-1},\\
\Delta_h &= \sum_0^h F_h(j)C_v(j-h-1),\\
S_h &= (I_S - F_h(h)\tilde{F}_h(h))S_{h-1},\\
\tilde{S}_h &= (I_s - \tilde{F}_h(h)F_h(h))\tilde{S}_{h-1}\\
F_h(0) &= \tilde{F}_h(0) = I_s,\\
S_0 &= \tilde{S}_0 = C_v(0).
\end{aligned}
\tag{6.1}
$$

Here the $\tilde{F}_h(j)$ relate to a time reversed form of (5.8). As indicated below (5.8) the equations (5.8) have disadvantages and have been modified and correspondingly modifications of (6.1) are needed and have been provided. *We emphasize again that in practice such modifications should, probably, always be used.* We shall use (6.1) later but first discuss a statistical use of the canonical correlation analysis discussed below (3.5), due to Akaike (1976).

Akaike's technique is not related directly to the Hankel norm approximations discussed in Section 3. Instead he attempts to find estimates of the Kronecker indices d_j and the corresponding $\alpha_{jk}(u)$ defined in (2.10). Indeed these are known if the $\tilde{\alpha}_{jk}(u)$ in (2.8) are known and these provide "null functions" for (2.2), i.e., these define linear combinations of the elements of $y^{(t+1)}$ that annihilate $\mathscr{H}$. If $\mathscr{H}$ were of finite rank, as would be the case if $k(z) \in V_\alpha$, α indexing $(d_1, d_2, \cdots, d_n)$, then we can seek to find the d_j and the $\tilde{\alpha}_{jk}(u)$ by effecting the canonical correlation analysis (if that were possible) between $y^{(t+1)}$ and y_t, $y'_t = (y(t)', y(t-1)', \cdots)$, which is the same as the canonical correlation analysis of the relation between $y^{(t+1)}$ and ε_t. Akaike's procedure is to replace y'_t by $\tilde{y}'_t = (y(t)', y(t-1)', \cdots, y(t-\hat{h})')$ where $\hat{h}$ is chosen by AIC to minimize $\log\det \hat{\Sigma}_h + 2hn^2/T$, $\hat{\Sigma}_h$ being the estimate computed from (5.8). Of course (6.1) with $y(t) = v(t)$ could be used for this purpose. The vector $y^{(t+1)}$ is replaced, in a series of canonical correlation analyses, by $y_t(l, m)$, $0 \leq m < n$, $l = 1, 2, \cdots$,

$$
\begin{aligned}
y_t(l, m)' = (y(t+1)', y(t+2)', \cdots, y(t+l)',\\
y_1(t+l+1), \cdots, y_m(t+l+1)).
\end{aligned}
$$

Examining (2.8) we see that if d_m is the smallest d_j, $j = 1, \cdots, n$, and $d_m = l$ then the row $r(l + 1, m)$ is linearly dependent on earlier rows of $\mathscr{H}$ and is the first row for which this is so. Thus, assuming $\hat{h} \gg d_j$, $j = 1, \cdots, n$ (as will eventually be the case since we have seen that $\hat{h} = O(\log T)$, in Section 5), then in the "true" canonical correlation analysis of $\tilde{y}_t$ with $y_t(l, m)$ there will be $ln + m - 1$ nonzero canonical correlations, and there will be one true zero canonical correlation, this corresponding to the discriminant function given by the left side of (2.10) for $j = m$. All that is needed is a criterion on the basis of which to judge whether the smallest *observed* canonical correlation can be reasonably regarded as null. (We do not give details of the calculation of the canonical correlations $\hat{\rho}$ since this is classical; see Anderson, 1958, page 298.) Akaike uses

$$(6.2) \qquad \begin{aligned} &(T - \nu_{l,m})\log(1 - \hat{\rho}^2_{ln+m}) + \nu_{l,m}, \\ &\qquad \nu_{l,m} = n(\hat{h} - l) - m + 1, \end{aligned}$$

where $\hat{\rho}^2_{ln+m}$ is the smallest canonical correlation between $y_t(l, m)$ and $\tilde{y}'_t$, $t = \hat{h} + 1, \cdots, T - l - 1$. If this is positive for the first time at $l(1)$, $m(1)$ then $\hat{d}_{m(1)} = l(1)$. Then $y_{m(1)}(t + l(1) + \ )$, $j > 0$, is eliminated from all future $y_t(l, m)$ and the procedure is repeated to find the second smallest $\hat{d}_j$. The number $\nu_{l,m}$ is always $n\hat{h} - \dim(y_t(l, m)) + 1$. At each step the coefficients in the "null function" for which the last canonical correlation is judged to be zero estimate the $\alpha_{jk}(u)$ (using the normalization $\alpha_{jj}(0) = 1$).

Thus a first estimate of the d_j and of the $\alpha_{jk}(u)$ has been obtained. It is possible to replace the parameters in Σ and $\beta_{jk}(u)$, $u = 1, 2, \cdots, d_j$, $j, k = 1, \cdots, n$ (see (2.10)), by those in $\Gamma(0)$ and $\gamma_{jk}(u)$, $u = 1, 2, \cdots, d_j$, $j, k = 1, \cdots, n$, where $\gamma_{jk}(u)$ is the typical element of $\Gamma(u)$ (see Solo, 1984). Since the $C(j)$ provide estimates of the $\Gamma(j)$ then having determined the $\hat{d}_j$, $\hat{\alpha}_{jk}(u)$, $u = 0, \cdots, \hat{d}_j$ by Akaike's procedure a first estimate of k, Σ has been obtained from which a better estimate could be got by an iterative solution of the equations of (Gaussian) maximum likelihood. This calculation could be based on (2.6). Indeed $-\frac{1}{2}T$ by the Gaussian log likelihood, ignoring a constant, is

$$\log \det \Sigma(t) + \frac{1}{T}\sum_1^T e(t)'\Sigma(t)^{-1}e(t)$$

where $\Sigma(t)$, $e(t)$ depend on Σ, A, B, C or equivalently on the $\alpha_{jk}(u)$, $\gamma_{jk}(u)$, $u = 1, \cdots, d_j$, and $\Gamma(0)$. Initial values of these and of the d_j are provided by Akaike's method.

There are some problems with the procedure and it does not so far seem to have been widely used, even for $n = 1$. *The procedure does not lend itself easily to a generalization to the case where there are observed inputs* (see (1.1)′). The d_j are not reestimated at a stage where more efficient estimates may be obtained. There is no asymptotic theory justifying the method at present.

A more direct approach is to seek to find θ, d to minimize $\log \det \Sigma_\theta + dC_T/T$ by a Gauss-Newton iteration. Two things are needed for this to be effective. The first is an initial value from which to commence and the second is a recursion on d which will enable the calculations at each iteration to be done reasonably cheaply. Since Σ_θ depends on θ only through $k(z)$, Σ then an initial estimate of these is needed. These can be obtained through (5.8) using (6.1) to make the calculation recursive. The first Gauss-Newton iteration is then as follows, *taking for example the case where only the $U(d)$ are examined.* (We use "iterate" to mean "repeat" so that the first iteration is the first calculation after the initial one.) Having computed

$$\hat{\varepsilon}(t) = \sum_0^{\hat{h}} \hat{\Phi}_{\hat{h}}(j)y(t - j)$$

form the autoregression (5.8), choosing $\hat{h}$ by minimizing the left side of (5.9), say for $C_T = 2$ or $\log T$ and using (6.1) for $v(t) = y(t)$, then $y(t)$ is regressed upon $-y(t - j)$, $j = 1, \cdots, p$, $\hat{\varepsilon}(t - j)$, $j = 1, \cdots, p$, and $\hat{p}$ is chosen to minimize $\log \det \hat{\Sigma}_p + 2p^2C_T/T$, again with $C_T = 2$ or $C_T = \log T$. Here the coefficient matrices in the regression and $\hat{\Sigma}_p$ may be found recursively from (6.1) with $v(t)' = (-y(t)', \hat{\varepsilon}(t)')$, the first n rows of $F(j)$ providing the coefficient matrices $[\hat{A}_p(j), \hat{B}_p(j)]$ of $(-y(t - j)', \hat{\varepsilon}(t - j)')'$ and the top left $n \times n$ block of S_p providing $\hat{\Sigma}_p$. A more economical modification of this procedure is described in Franke (1985). *Again it will be wise to taper $v(t)$.* Having determined $\hat{p}$ then $\hat{q}$ in $\hat{d} = n\hat{p} + \hat{q}$, $0 \leq \hat{q} < n$, is determined by a further set of $n - 1$ regressions, the nature of which can be determined from (2.13). Thus for the qth of these multiple systems of regressions, $1 \leq q \leq n - 1$, we regress, for $j = 1, \cdots, q$, $y_j(t)$ on $-y_k(t - i)$, $k = 1, \cdots, q$, $i = 1, \cdots, \hat{p} + 1$; $-y_k(t - i)$, $k = q + 1, \cdots, n$, $i = 2, \cdots, \hat{p}$; $\hat{\varepsilon}_k(t - i)$, $k = 1, \cdots, n$, $i = 1, \cdots, \hat{p} + 1$. For $j = q + 1, \cdots, n$ we regress $y_j(t)$ on $-y_k(t - i)$, $k = 1, \cdots, n$, $i = 1, \cdots, \hat{p}$; $-\{y_k(t) - \hat{\varepsilon}_k(t)\}$, $k = q + 1, \cdots, n$; $\hat{\varepsilon}_k(t - i)$, $k = 1, \cdots, n$, $i = 1, \cdots, \hat{p}$. The residuals, $\tilde{\varepsilon}_j(t)$ from the jth of these regressions, for fixed q, are used to calculate

$$\hat{\Sigma}_{\hat{p},q} = [T^{-1} \sum \tilde{\varepsilon}_j(t)\tilde{\varepsilon}_k(t)]$$

where the (j, k)th element is shown. For $q = 0$, n the calculation has already been done since $\hat{\Sigma}_{p,0} = \hat{\Sigma}_p$, defined above. The criterion is $\log \det \hat{\Sigma}_{\hat{p},q} + 2ndC_T/T$, $d = n\hat{p} + q$, $0 \leq q < n$. The next Gauss-Newton step is more complicated to describe but again it can be done substantially recursively using (6.1). Details and theorems concerning the method appear in Hannan and Kavalieris (1984) for example. There is a lot more to be said about the method, for which theorems of justification can be found in the reference just cited. It was first proposed, for $n = 1$ and p, q known, by Durbin (1960) and in the wider present context by Hannan and Rissanen (1982). *We emphasize again what is being done.* A criterion of the form introduced in Section 4 is chosen, e.g. (4.3). We seek to optimize this over the model set of rational transfer function systems (or some subset such as $U(d)$, $d = 0, 1, 2, \cdots$). This is done by a Gauss-Newton iteration. At each Gauss-Newton step the values of d have to be scanned and to reduce this labour a recursion on d is used. To initiate the procedure values of k, Σ are estimated by a linear procedure, namely (5.8), choosing $\hat{h}$ by (4.3) once more. The procedure has an evident appeal because of its simple, logical justification.

To conclude this section and this paper we will discuss briefly on-line calculation of an ARMA approximation. By this we mean a calculation which is effected as T increases, with a new estimate being obtained as each time point passes. Such procedures, in a wide context, are of paramount importance in many connections. For example they could be used to conserve channel capacity in, say, telephone traffic by using an ARMA model to encode data minimally. The data being encoded would be a fraction of a second of speech so that the procedure would forget the past as rapidly as it took account of the present. In this situation the calculation would have to be effected in real time, but on-line calculation can be important even without a real time context since such on-line procedures are adaptive, i.e. will change with an evolving situation. For a recent survey of such procedures see Ljung and Söderström (1983). One way to proceed would be to effect each of the Gauss-Newton steps discussed in the previous paragraph recursively also in time. If different values of the order parameters have also to be determined then some fixed upper bound will have to be imposed on them if the calculation is to be in real time. This will not be a problem in practice because the forgetting of the past ensures that the amount of data being used at each time point will not increase indefinitely and with a fixed amount of data the orders to be used will be uniformly bounded. Here we outline some details only for the case $n = 1$ and we consider only $p = q$. We emphasize again that a procedure with a simple logical justification is being used. Thus a criterion, (4.3) or (4.7), let us say, is being instituted and a real time Gauss-Newton procedure introduced. The criterion (4.7) is particularly well adapted to this problem as will be emphasized below. At each Gauss-Newton step a regression is being effected. At each step after the first the regressors in the regression have to be constructed and these can be obtained from a parallel calculation of the previous step. We briefly illustrate by the first two Gauss-Newton steps. The basic formula is that for the calculation of a set of regression parameters, $\theta(t)$, from data to time t on a dependent variable $y(s)$ and a vector of independent variables $v(s)$, $s = 1, \cdots, t$. Such formulae are for $n = 1$,

$$\begin{aligned} &\theta(t) = \theta(t-1) + P(t)v(t)e(t), \\ (6.3)\qquad &e(t) = y(t) - \theta(t-1)'v(t), \\ &P(t) = \left\{\sum_1^t v(s)v(s)'\right\}^{-1}. \end{aligned}$$

It is possible to calculate $P(t)$ recursively by

$$\begin{aligned} (6.4)\qquad P(t) = P(t-1) - \{1 + v(t)'P(t-1)v(t)\}^{-1} \\ \cdot P(t-1)v(t)v(t)'P(t-1). \end{aligned}$$

To make this forget the past at an exponential rate one multiplies $y(s)$, $v(s)$ by $l_t(s)^{1/2}$, $s = 1, \cdots, t$, so that the least squares problem is to minimize $\sum l_t(s)\{y(s) - \theta'v(s)\}^2$,

$$l_t(s) = \prod_{s+1}^{t} \lambda(u), \quad l_t(t) = 1, \quad \lambda(u) > 0.$$

For example we might choose $\lambda(u) \equiv \lambda$, $0 < \lambda < 1$. It is not necessary to effect the multiplications of $y(s)$, $v(s)$ by $l_t(s)^{1/2}$ (which could not be done in real time) but it is enough merely to adjust (6.4) so that it is replaced by

$$\begin{aligned} P(t) = \frac{1}{\lambda(t)}[P(t-1) - \{\lambda(t) + v(t)P(t-1)v(t)\}^{-1} \\ \cdot P(t-1)v(t)v(t)'P(t-1)]. \end{aligned}$$

In the case of the autoregressive step then $v(t)' = (-y(t-1)', \cdots, -y(t-h)')$.

However it is necessary also cheaply to compute this for $h = 1, \cdots, H$ if the order is to be determined at each t. We now describe briefly a procedure for doing this. A maximum value, H, of h is prescribed so that now $v(t)' = (-y(t-1)', \cdots, -y(t-H)')$. Thus at time t the data matrix is

$$X(t) = \begin{bmatrix} v(1)' & y(1) \\ v(2)' & y(2) \\ \vdots & \vdots \\ v(t)' & y(t) \end{bmatrix}.$$

The procedure is to reduce $X(t)$ to upper triangular form by a sequence of plane reflections (fast Givens transformations). As another row is added (when t increases to $t + 1$) the next calculation is easily determined. During the process the quantities needed to update the residual variances, $\hat{\sigma}_h^2(t)$, for all h to the maximum value, H, to be considered are obtained, $\hat{\sigma}_h^2(t)$ being the nonrecursive mean square to time t, but also the recursive residuals $e_h(t)$ (see (6.3)). We discuss these further below. The incorporation of forgetting, of the kind described by $l_t(s)$, above, is effected by one additional multiplication at each time point. For details we refer the reader to Hannan, Kavalieris and Mackisack (1986) and references contained therein. To choose h, we consider, taking $n = 1$ for illustration,

$$\log\{s_h^2(t)/\tau(t)\} + h \log\{\tau(t)\}/\tau(t), \tag{6.5}$$

$$s_h^2(t) = \sum_1^t \hat{\varepsilon}_h(s)^2,$$

where $\hat{\varepsilon}_h(s) = y(s) - \theta(t)'v(s)\tau(t)$ is a measure of how much data is used at time t and might be defined by

$$\tau(t-1) = \lambda(t-1)\tau(t) + 1, \quad \tau(0) = 0. \tag{6.6}$$

This is because the effective sample size at time t might reasonably be considered to be

$$\sum_{s=1}^t l_t(s) = \sum_{s=1}^t \prod_{s+1}^t \lambda(u),$$

which satisfies (6.6). For the second Gauss-Newton step $v(t)' = (-y(t-1), \cdots, -y(t-P), \hat{\varepsilon}(t-1), \cdots, \hat{\varepsilon}(t-P))$ where P is the maximum p to be considered and $\hat{\varepsilon}(t)$ is $e(t)$ from the first Gauss-Newton step (see (6.3)). There may be some modifications to this (see the cited reference). The procedure may be continued to any fixed number of Gauss-Newton iterations. The procedure at each Gauss-Newton step is essentially the same, being a regression. It should be mentioned that at later steps not only will $v(t)$ involve the output from previous steps, but also $y(t)$ will be replaced by a vector involving the output of previous steps. In practice one might not wish to change h or p too frequently and simple rules can be used to effect that. Again the method has considerable appeal, having a simple, logical basis. Some theorems relating to the method are given in Hannan, Kavalieris and Mackisack (1986), but these relate to the case where p, q are true values and are known. However the methods are more generally described. These calculations based on Givens reflections required, for the first Gauss-Newton step, about $4H^2$ operations of multiplication followed by addition (and the storage of about $H^2/2$ quantities). An alternative procedure would be to use one of the lattice algorithms mentioned in Section 5, below (5.8). Let $e_h(t)$ be the "forwards residuals," i.e., the residuals from the (Toeplitz) regression of $y(t)$ on $y(t-j)$, $j = 1, \cdots, h$. (The equations (5.8) are Toeplitz in form because the matrix on the left is "block Toeplitz," having all blocks the same down a diagonal.) Let $r_h(t)$ be the "backwards" residuals, from the Toeplitz regression of $y(t-h)$ on $y(t-1), \cdots, y(t-h+1)$. These are Toeplitz orthogonal, of course, to $y(t-1), \cdots, y(t-h+1)$ and occur because (6.1) proceeds by the standard method for adding a variable into a regression of making the new variable orthogonal to those already in the regression. Using the first line of (6.1), for $v(t) = y(t)$, we obtain

$$\begin{aligned} e_h(t+1) &= e_{h-1}(t+1) + F_{h,t}(h)r_{h-1}(t), \\ &\qquad r_h(0) = e_h(0) = 0, \\ r_h(t+1) &= r_{h-1}(t+1) + \tilde{F}_{h,t}(h)e_{h-1}(t+1), \\ &\qquad r_0(t) = e_0(t) = y(t). \end{aligned} \tag{6.7}$$

Note that $e_h(s)$ is not the same as $\hat{e}_h(s)$ in (6.5), which is computed using data to time t, while $e_h(s)$ uses only data to time s. The latter are the "recursive residuals." Thus if the $r_h(t)$ are stored, $h = 0, 1, \cdots, H-1$, then $e_h(t+1)$, $r_h(t+1)$ may be formed $h = 0, 1, \cdots, H$, once $y(t+1)$ is available, using $2h$ calculations in the case $n = 1$. However $F_{h,t+1}(h)$, $\tilde{F}_{h,t+1}(h)$ have also to be calculated for the next stage. These are essentially got from the matrices of variances and covariances of $e_h(t)$ with $r_h(t-1)$, which provides Δ_{h-1} (see (6.1)). Indeed $F_{h,t+1}(h)$ is the coefficient of (Toeplitz) regression of $y(s)$ on $r_{h-1}(s-1) = \sum \tilde{F}_{h-1,t+1}(j)y(s-h+j)$. This is the same as the regression of $e_{h-1}(s)$ on $r_{h-1}(s-1)$. The normalized lattice algorithms spoken of below (5.8) replace $e_h(t)$, $r_h(t)$ by normalized forms by transformation so that their variance, covariance matrices are unity. They attempt also to mitigate the effect of the Toeplitz calculation by appropriate definitions of the estimate of $F_{h,t}(h)$, for example by confining the autocovariances to the use of $1 \le s \le t-h$. We cannot give full details here but refer the reader to Friedlander (1982). The use of (6.7) would not enable $\hat{\sigma}_h^2(t)$ to be formed but only the analogous quantity, for $n = 1$,

$$\tilde{\sigma}_h^2(t) = \frac{1}{t}\sum_1^t e_h(s)^2. \tag{6.8}$$

If we consider (4.5) we see that, for the scalar autoregressive case, for $T = t$ and multiplying by $-2/t$ we obtain

$$\sum_{s=0}^{t-1} \left\{\log \hat{\sigma}_h^2(s) + \frac{e_h(s+1)^2}{\hat{\sigma}_h^2(s)}\right\}. \tag{6.9}$$

This seems well adapted to an algorithm of the present type, but while it can be calculated by the methods of Hannan, Kavalieris and Mackisack (1986) it cannot be calculated easily by the lattice methods which provide only (6.8). This latter could be inserted in (6.9) in place of $\hat{\sigma}_h^2(s)$. If forgetting is used (see below (6.3)), as will almost always be the case, the divisor, t, in (6.8) should be replaced by $\tau(t)$ from (6.6). In fact consideration of (6.9) suggests that, whether or not there is forgetting (6.9) may as well be replaced by $\tilde{\sigma}_h^2(t)$. A key further step is the further analysis of this, at least for the stationary case. This is proceeding and there seems no doubt that theorems analogous to those below (5.6) can be established for it.

It is dangerous indeed to make strong assertions about any of the on-line methods introduced here in the case of an evolving system, where forgetting is used and h is to be determined, since they are so difficult to analyze so that intuition may be misleading. Simulations and experience are needed. Some experience shows the methods based on (6.5) working well.

7. CONCLUSION

The problem of rational transfer function approximation is attractive, for there is a considerable, and relevant, mathematical theory underlying it and the problem is physically important. There is much still to be done. Such subjects tend to develop an impetus of their own. The main problems of time series analysis may relate to the development of nonlinear models. As the work of Priestley (1980) and Tjøstheim (1986) shows, one way of proceeding is via a development commencing from (2.3), and Tjøstheim's models in particular seem to be of the nature of the simple model discussed below (5.3). Thus the full understanding of the problem of (linear) rational transfer function approximation may be essential for the later nonlinear theories.

REFERENCES

ADAMYAN, V. M., AROV, D. F. and KREIN, M. G. (1971). Analytic properties of Schmidt pairs for a Hankel operator and the generalised Schur-Takagi problem. *Math. USSR-Sb.* **15** 31–73.

AKAIKE, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21** 243–247.

AKAIKE, H. (1976) Canonical correlation analysis of time series and the use of an information criterion. In *Systems Identification: Advances and Case Studies* (R. K. Mehra and D. G. Lainiotis, eds.) 29–91. Academic, New York.

AN, HONGH-ZHI, CHEN, ZHAO-GUO and HANNAN, E. J. (1982). Autocorrelation, autoregression and autoregressive approximation. *Ann. Statist.* **10** 926–936.

ANDERSON, B. D. O. and MOORE, J. B. (1979). *Optimal Filtering.* Prentice Hall, Englewood Cliffs, N.J.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

BOX, G. E. P. and JENKINS, G. M. (1970). *Time Series Analysis, forecasting and control.* Holden-Day, San Francisco.

BRILLINGER, D. R. (ed.). (1984). *The Collected Works of John W. Tukey: Time Series, 1965–1984* **2**. Wadsworth, Monterey, Calif.

DAHLHAUS, R. (1984). Parameter estimation of stationary processes with spectra containing strong peaks. In *Robust and Nonlinear Time Series Analysis* (J. Franke, W. Härdle and D. Martin, eds.). *Lecture Notes in Statist.* **26** 50–67. Springer, New York.

DURBIN, J. (1960). The fitting of time series models. *Rev. Internat. Inst. Statist.* **28** 233–244.

FINDLEY, D. F. (1985). On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. *J. Time Ser. Anal.* **6** 229–252.

FISHER, R. A. (1944). *Statistical Methods for Research Workers*, 9th ed. Oliver and Boyd, Edinburgh.

FRANKE, J. (1985a). A Levinson-Durbin algorithm for autoregressive-moving average processes. *Biometrika* **72** 573–581.

FRANKE, J., GASSER, TH. and STEINBERG, H. (1985). Fitting autoregressive processes to EEG time series: an empirical comparison of estimates of the order. *IEEE Trans. Acoust. Speech Signal Process.* **33** 143–150.

FRIEDLANDER, B. (1982). Lattice filters for adaptive processing. *Proc. IEEE* **70** 830–867.

GLOVER, K. (1984). All optimal Hankel-norm approximations of linear multivariate systems and their L^∞ error bounds. *Internat. J. Control* **39** 1115–1193.

HANNAN, E. J. (1969). The identification of vector mixed autoregressive-moving average systems. *Biometrika* **56** 223–225.

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

HANNAN, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8** 1071–1081.

HANNAN, E. J. (1981). Estimating the dimension of a linear system. *J. Multivariate Anal.* **11** 459–473.

HANNAN, E. J. and KAVALIERIS, L. (1983). The convergence of autocorrelations and autoregressions. *Austral. J. Statist.* **25** 287–297.

HANNAN, E. J. and KAVALIERIS, L. (1984). Multivariate linear time series models. *Adv. Appl. Probab.* **16** 492–561.

HANNAN, E. J. and KAVALIERIS, L. (1986). Regression, autoregression models. *J. Time Ser. Anal.* **7** 27–49.

HANNAN, E. J., KAVALIERIS, L. and MACKISACK, M. (1986). Recursive estimation of linear systems. *Biometrika* **73** 119–134.

HANNAN, E. J. and POSKITT, D. S. (1986). Unit canonical correlation between future and past. To appear in *Ann. Statist.*

HANNAN, E. J. and RISSANEN, J. (1982). Recursive estimation of ARMA order. *Biometrika* **69** 81–94.

JEWELL, N. P., BLOOMFIELD, P. and BARTMANN, F. C. (1983). Canonical correlations of past and future for time series: bounds and computations. *Ann. Statist.* **11** 848–855.

JONCKHEERE, E. A. and HELTON, J. W. (1985). Power spectrum reduction by optimal Hankel norm approximation of the phase of the outer spectral factor. *IEEE Trans. Automat. Control* **AC-30** 1192–1201.

KAILATH, T. (1980). *Linear Systems.* Prentice Hall, Englewood Cliffs, N.J.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Engrg.* **82** 35–45.

KALMAN, R. E. and BUCY, R. S. (1961). New results in linear filtering and prediction theory. *Trans. ASME J. Basic Engrg.* **83** 95–108.

LJUNG, L. and SÖDERSTRÖM, T. (1983). *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, Mass.

PRIESTLEY, M. B. (1980). State dependent models: a general approach to non-linear time series analysis. *J. Time Series Anal.* **1** 57–71.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431.

RISSANEN, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080–1100.

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164.

SHIBATA, R. (1981). An optimal autoregressive spectral estimate. *Ann. Statist.* **9** 300–306.

SOLO, V. (1984). The exact likelihood for a multivariate ARMA model. *J. Multivariate Anal.* **15** 164–173.

TJØSTHEIM, D. (1986). Some doubly stochastic time series models. *J. Time Ser. Anal.* **7** 51–72.

TJØSTHEIM, D. and PAULSEN, J. (1983). Bias of some commonly-used time series estimates. *Biometrika,* **70** 389–399.

WHITTLE, P. (1963a). On the fitting of multivariate autoregressions and the approximate canonical factorisation of a spectral density matrix. *Biometrika* **50** 129–134.

Comment

R. J. Bhansali

I would like to congratulate Ted Hannan on a masterly survey of the current state of the art for fitting multivariate autoregressive moving average models, ARMA(p, q). Hannan is quite correct in emphasizing that there may not be a true (p, q) and thus a fitted ARMA model is at best thought of as an approximation to the generating structure of the observed time series. The question then arises: what are the properties of the order selected by minimizing AIC when viewed in this light rather than as an estimator of an underlying "true" order? The work of Shibata (1980) would suggest that the order selected by AIC is such that the one-step mean square error of prediction is minimized within the class of all order selection procedures. However, the more-than-one-step mean square error of prediction may not be minimized (see also Whittle, 1963b, page 36). Indeed, for autoregressive model fitting, Findley (1983) has advocated that a different order should be selected for each forecast lead, and he has suggested that a criterion introduced by Shibata (1980, page 163) may be used for this purpose. However, a justification for introducing this criterion has not been given. A related but different criterion is suggested by the work of Hannan and Rissanen (1982).

As has already been noted by Franke (1985a) and Chen (1985), at the second stage of the Hannan-Rissanen procedure for ARMA model selection, "autoregressive" estimates of the coefficients $b(u)$, say, in the moving average representation of a univariate stationary nondeterministic process $\{x_t\}$ are obtained as

$$\hat{b}_h(u) = \hat{c}_h(u)/\hat{c}_h(0), \quad u = 0, 1, \cdots,$$

where

$$\hat{c}_h(u) = \sum_{j=0}^{h} \hat{a}_h(j) R^{(T)}(u + j)$$

provides the corresponding estimator of the cross-covariance $c(u)$ say, between x_{t+u} and the linear innovations ε_t; the $\hat{a}_h(j)$ are the hth order "Yule-Walker" estimates of the autoregressive coefficients;

$$R^{(T)}(u) = T^{-1} \sum_{t=1}^{T-u} x_t x_{t+u}, \quad u = 0, 1, \cdots,$$

is a "positive definite" estimator of the covariance function of $\{x_t\}$; and $x_1, \cdots, x_T$ denotes an observed realization of length T of $\{x_t\}$.

Now, if the complete past $\{x_t, t \le 0\}$, say of $\{x_t\}$ is known, the s step mean square error of prediction is given by

$$V(s) = \sigma^2 \sum_{j=0}^{s-1} b^2(j), \quad s = 1, 2, \cdots,$$

where $\sigma^2 = c(0)$ is the variance of ε_t.

Bhansali (1978) and Lewis and Reinsel (1985) consider the effect on the mean square error of prediction of estimating the prediction constants by fitting an autoregressive model of order h, when h is a function of T and tends to infinity simultaneously with it. It is clear from their work that if $o_p(h^{1/2}/T^{1/2})$ terms are ignored and certain additional regularity conditions are satisfied, the resulting mean square error of prediction may be approximated by

$$L(s) = V(s)\left(1 + \frac{h}{T}\right).$$

On adopting an argument similar to that used by Akaike (1970) for deriving his FPE criterion, which as discussed by Bhansali (1986) is closely related to the argument used for deriving AIC, one may consider

R. J. Bhansali is Senior Lecturer, Department of Statistics and Computational Mathematics, University of Liverpool, Liverpool L69 3BX, England.