MARKOVIAN REPRESENTATION OF STOCHASTIC PROCESSES AND ITS APPLICATION TO THE ANALYSIS OF AUTOREGRESSIVE MOVING AVERAGE PROCESSES

HIROTUGU AKAIKE

(Received Sept. 18, 1973; revised March 28, 1974)

Summary

The problem of identifiability of a multivariate autoregressive moving average process is considered and a complete solution is obtained by using the Markovian representation of the process. The maximum likelihood procedure for the fitting of the Markovian representation is discussed. A practical procedure for finding an initial guess of the representation is introduced and its feasibility is demonstrated with numerical examples.

1. Introduction

The state space representation of a system is a fundamental concept in modern control theory. When a discrete-time system is time-invariant and linear the state space representation of the system is given in the form

$$\begin{aligned} v_{n+1} &= Av_n + Bu_{n+1} \\ y_n &= Cv_n \ , \end{aligned} \tag{1.1}$$

where n denotes the time and u_n is a $q\times 1$ vector of the input to the system, y_n is an $r\times 1$ vector of the output and v_n is a $p\times 1$ vector of the state. A, B and C are respectively $p\times p$, $p\times q$ and $r\times p$ matrices. The use of the state space representation for the design of optimal control under a quadratic cost function is well-known in the engineering literature. The state space representation of a system has been discussed in some statistical literature (for example, Whittle [19] and Akaike [1]), however, it has not yet been fully exploited by statisticians. This may partly be due to the somewhat abstract definition of the concept of state, as is described by Kalman and others ([10], Chap. 10). The state is sometimes vaguely understood as a condensed representa-

tion of information from the present and past, such that the future behaviour of the system can completely be described by the knowledge of the present state and the future input. This idea finds a precise mathematical formulation when the system is stochastic, i.e. when the input u_n and the output y_n are stochastic processes. It was shown (Akaike [6]) that by the analysis of canonical correlations between the set of the present and future output and the set of the present and past input a Markovian representation of a stochastic system can be obtained. The Markovian representation is a stochastic analogue of (1.1) and is given by

$$\begin{aligned} v_{n+1} &= Av_n + Bz_{n+1} \\ y_n &= Cv_n + w_n \,, \end{aligned} \tag{1.2}$$

where w_n is uncorrelated with v_n. In (1.2) z_{n+1} is the innovation of the input u_n at time $n+1$ and is defined by $u_{n+1}-u_{n+1|n}$, where $u_{n+1|n}$ is the projection of u_{n+1} on its past which is defined as the mean square closure of the space of finite linear combinations of the components of $u_n, u_{n-1}, \cdots$. When $u_n=y_n$, w_n vanishes from (1.2) and a Markovian representation of a stationary stochastic process y_n is given in the form

$$\begin{aligned} v_{n+1} &= Av_n + Bz_{n+1} \\ y_n &= Cv_n \,. \end{aligned} \tag{1.3}$$

This representation gives y_n as the output of a stochastic system which is time invariant and linear and driven by a white noise input z_n. v_n is called the state of the process.

The purpose of the present paper is to discuss the relation of this Markovian representation of a stationary stochastic process y_n with the familiar autoregressive moving average (AR-MA) representation

$$y_n + B_1 y_{n-1} + \cdots + B_M y_{n-M} = z_n + A_1 z_{n-1} + \cdots + A_L z_{n-L} \,. \tag{1.4}$$

First a proof of the equivalence of the Markovian and AR-MA representations, obtained by showing the existence of direct transformations from one to the other, is given. Hannan [9] discussed the problem of identifiability or the uniqueness of the AR-MA representation that is especially difficult when the process is multivariate. The simplest type of identifiability is the one which is called the block-identifiability in the present paper. The determination procedure of the AR-MA coefficient matrices A_i and B_j from the covariance matrices of y_n, under the block-identifiability condition, leads to a natural stochastic interpretation of Rissanen's [15] block triangularization procedure of block Hankel matrices. This result illustrates the inherent relationship be-

tween the problem of identification, or the determination of the Markovian or AR-MA representation, of the process y_n and the Hankel type matrices.

By a further analysis of the Markovian representation it becomes clear that the identifiability problem of the AR-MA representation can be solved completely without any restriction such as block-identifiability. This result demonstrates the merit of the Markovian representation for the purpose of analysis of stochastic systems. Generally there is not a unique structure, or a set of special forms of the matrices A, B and C in the Markovian representation (1.3), with a minimum number of undetermined parameters and that can represent every y_n with the same minimal possible dimension of v_n in the representation (1.3). Thus the identification, or the determination of the Markovian or AR-MA representation, must proceed in two steps, the first step is the selection of a special structure and the second is the determination of the parameters in the structure. Once an exhaustive set of special structures is specified the statistical identification or the determination of the structure and the parameters based on the observations of a Gaussian process can be realized through the maximum likelihood procedure with the aid of an information theoretic criterion (Akaike [2], [3]).

The statistical identification is realized by using the Markovian representation and the results may be used directly for the purpose of analysis and implementation of control of a multivariate stochastic system without recourse to the AR-MA representation. In particular in the last section of the paper it is shown that under a very mild assumption consistent estimates of the structure and the parameters within the matrix A of a special Markovian representation can be obtained by a simple procedure. This is a fundamental contribution to the subject of statistical identification of multivariate stochastic systems. The procedure is based on the canonical correlation analysis between the present and future and the present and past observations of the process and provides an initial guess of the structure and the parameters to be used for the maximum likelihood procedure. This is a significant example of the use of the canonical correlation concept in relation to the time series analysis. Numerical examples are given to show the feasibility of the procedure. Towards the end of the paper it is suggested that a special Markovian representation may be useful to give an answer to the problem of reduction of the number of measurements of a complex stochastic system. This problem was raised by Priestley and others [13].

Throughout the present paper the closure in the sense of mean square of the linear space of finite linear combinations of the components of the random vectors $x_1, x_2, \cdots$ will be denoted by $R(x_1, x_2, \cdots)$

and called the space spanned by the components of, or, simply, the space spanned by, $x_1, x_2, \cdots$. The i-step ahead predictor at time n of a stochastic process y_n is defined as the projection of y_{n+i} on $R(y_n, y_{n-1}, \cdots)$ and denoted by $y_{n+i|n}$. If $S=R(x_1, x_2, \cdots)$ holds for a linear space S, the set of the components of $x_1, x_2, \cdots$ is called a system of generators of S. In the present paper the qualities of random variables are understood in the sense of mean square.

2. Autoregressive-moving average processes and Markovian representations

The autoregressive moving average process (AR-MA process) y_n is defined by

$$y_n+B_1y_{n-1}+\cdots+B_My_{n-M}=z_n+A_1z_{n-1}+\cdots+A_Lz_{n-L}\ , \tag{2.1}$$

where B_i and A_j are the matrices of coefficients, y_n and z_n are $r\times 1$ vectors and z_n is a white noise with $\mathrm{E}\,z_n=0$, zero vector, and $\mathrm{E}\,(z_nz_n')=G$ and for $i\neq 0$ $\mathrm{E}\,(z_nz_{n-i}')=0$, zero matrix, and $\mathrm{E}\,y_nz_{n+i}'=0$ for $i=1, 2, \cdots$. It is assumed that the characteristic equations $\left|\lambda^MI+\sum_{i=1}^{M}\lambda^{M-i}B_i\right|=0$ and $\left|\lambda^LI+\sum_{j=1}^{L}\lambda^{L-j}A_j\right|=0$ have zeros outside the unit circle. This assumption assures that y_n can be expressed in a form

$$y_n=\sum_{m=0}^{\infty}W_mz_{n-m} \tag{2.2}$$

with $W_0=I$, identity matrix, and z_n is the innovation of y_n at time n, i.e. $z_n=y_n-y_{n|n-1}$ and $y_{n|n-1}$ is the one-step ahead predictor of y_n at time $n-1$. To get a Markovian representation of y_n it is only necessary to analyze the structure of the predictors $y_{n|n}, y_{n+1|n}, \cdots$. Let $x_{|n}$ denote the projection of x on $R(y_n, y_{n-1}, \cdots)$, then $y_{n+i|n}$ satisfies the relation

$$\begin{aligned} y_{n+i|n}+B_1y_{n+i-1|n}+\cdots+B_My_{n+i-M|n} \\ =z_{n+i|n}+A_1z_{n+i-1|n}+\cdots+A_{M-1}z_{n+i-L|n}\ , \end{aligned} \tag{2.3}$$

where $y_{n+h|n}=y_{n+h}$ for $h=0, -1, \cdots$, and $z_{n+h|n}=0$ for $h=0, 1, \cdots$. For $i\geqq L+1$ the right-hand side of (2.3) vanishes. Thus $y_{n+i|n}$ $(i=0, 1, \cdots)$ can be expressed as linear transforms of $y_{n|n}, y_{n+1|n}, \cdots, y_{n+K-1|n}$, where $K=\max(M, L+1)$, and the components of these vectors form a system of generators of the linear space spanned by the components of $y_{n+i|n}$ $(i=0, 1, \cdots)$. Especially it holds that

$$y_{n+K|n}=-B_1y_{n+K-1|n}-B_2y_{n+K-2|n}\cdots-B_Ky_{n|n}\ , \tag{2.4}$$

where by definition $B_m=0$ for $m=M+1, M+2, \cdots, K$. From (2.2) one

can get

$$y_{n+i+1|n+1}=y_{n+i+1|n}+W_i z_{n+1}\ . \tag{2.5}$$

From (2.4) and (2.5) it can be seen that the vector $v_n=(y'_{n|n}, y'_{n+1|n}, \cdots, y'_{n+K-1|n})'$ provides a Markovian representation

$$\begin{aligned} v_{n+1}&=\begin{bmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & I \\ -B_K & -B_{K-1} & -B_{K-2} & \cdots & -B_1 \end{bmatrix} v_n+\begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{K-2} \\ W_{K-1} \end{bmatrix} z_{n+1} \\ y_n&=[\,I\ \ 0\ \ 0\cdots 0\,]v_n\ . \end{aligned} \tag{2.6}$$

This result shows that an AR-MA process always has a Markovian representation.

It should be noted that W_m's are the impulse response matrices of the time-invariant linear system defined by (2.1) with the input z_n and the output y_n. The jth column $W_m(\cdot j)$ of W_m is obtained by the relation

$$\begin{aligned} &W_m(\cdot j)+B_1W_{m-1}(\cdot j)+\cdots+B_MW_{m-M}(\cdot j) \\ &\quad=D_n(\cdot j)+A_1D_{n-1}(\cdot j)+\cdots+A_LD_{n-L}(\cdot j)\ , \end{aligned}$$

where $W_m(\cdot j)=0$, zero vector, for $m<0$ and $D_n(\cdot j)$ denotes the jth column of a matrix D_n which is by definition equal to I, identity matrix, for $n=0$ and 0, zero matrix, for $n\neq 0$. Thus numerically it is a simple matter to derive the Markovian representation (2.6) from the AR-MA representation (2.1).

Now suppose a process y_n has a Markovian representation

$$\begin{aligned} v_{n+1}&=Av_n+Bz_{n+1} \\ y_n&=Cv_n\ , \end{aligned} \tag{2.7}$$

where it is assumed that v_n is a $p\times 1$ vector of the state and z_n is the innovation of y_n. If the characteristic polynomial of A is given by $|\lambda I-A|=\lambda^p+\sum_{m=1}^{p} a_m\lambda^{p-m}$ then by the Cayley-Hamilton theorem $A^p+\sum_{m=1}^{p} a_mA^{p-m}=0$. From (2.7), $v_{n+i}=A^iv_n+A^{i-1}Bz_{n+1}+\cdots+Bz_{n+i}$ and it follows that y_n has an AR-MA representation

$$y_{n+p}+a_1y_{n+p-1}+\cdots+a_py_n=z_{n+p}+C_1z_{n+p-1}+\cdots+C_{p-1}z_{n+1} \tag{2.8}$$

where

$$C_i=C(A^i+a_1A^{i-1}+\cdots+a_iI)B\ .$$

In (2.8) the autoregressive coefficients are scalars, or, equivalently, constant diagonal matrices. This result shows that any stationary stochastic process with the Markovian representation (2.7) also has an AR-MA representation (2.8).

Thus at least theoretically there is no distinction between the Markovian and the AR-MA representations of a stationary stochastic process. The results obtained by the analysis of the Markovian representation can be used for the analysis of the AR-MA representation. This fact is fully utilized in the following discussion of the identifiability problem of the AR-MA representation. In (2.7), predictors are given simply by

$$y_{n+i|n}=CA^{i}v_{n} \qquad i=0, 1, \cdots . \tag{2.9}$$

Thus the components of v_n form a system of generators of the space $R(y_{n|n}, y_{n+1|n}, \cdots)$ which is spanned by the predictors $y_{n|n}, y_{n+1|n}, \cdots$. This space $R(y_{n|n}, y_{n+1|n}, \cdots)$ will hereafter be called the predictor space. When the dimension of the predictor space of an arbitrary stationary process y_n is finite there is a finite K which satisfies the relation (2.4) and y_n has a Markovian representation (2.6). Thus the finiteness of the dimension of the predictor space is the fundamental characterization of a process with Markovian or AR-MA representation. And one of the two Markovian representations of a process y_n, the states of which are defined by the elements of the predictor space, can be obtained from the other by a linear transformation of the state of the latter. It is now obvious that the dimension, as a vector, of the state of a Markovian representation which is defined by using a basis of the predictor space as its state is minimal. The dimension of this basis, or the dimension of the predictor space, is a characteristic of the stochastic system which generates the process from its innovations and will be called the dimension of the system. Also the process is called a process with p-dimensional dynamics.

3. Block-identifiability of autoregressive moving average processes

Although the AR-MA representation of a stationary stochastic process has been used as one of the basic models of time series analysis there is a serious conceptual difficulty inherent in this model. This is the non-uniqueness of the representation. When (2.1) holds there are infinitely many other representations of the same process, for example those which are obtained from (2.1) by the transformations which replace n of (2.1) by some $n-k$ $(k>1)$, premultiply it with an $r\times r$ matrix D_k and add to the original (2.1). When y_n is a univariate process, the representation can be made unique by requiring the orders

L and M of (2.1) to be minimal. When y_n is a multivariate process this requirement of minimal order is not necessarily sufficient to make the representation unique. Under the assumption of non-singularity of the covariance matrix $G=(z_n z_n')$ Hannan [9] gave a necessary and sufficient condition for an AR-MA process y_n to have a unique representation of the form (2.1).

It is trivially true that the uniqueness of a representation is a prerequisite for the development of consistent estimation procedures. But this fact should not be considered as meaning the impracticability of developing a general estimation procedure of AR-MA models without identifiability conditions. If the purpose of fitting an AR-MA model is only to get an estimate of the covariance structure of the process under observation any one of the possible equivalent representations can serve for the purpose, if only it can be specified properly. The practicability of this specification procedure is the main subject of the present paper.

Although a general estimation procedure without any assumption of identifiability of the model is developed in the later sections it will be useful for the understanding of the subject to discuss the relation between the AR-MA representation (2.1) and the Markovian representation (2.6) under the assumption of a simplest type of identifiability. Under the assumption of non-singularity of $G=\mathrm{E}(z_n z_n')$, W_m is uniquely determined by (2.2) and satisfies the relation

$$A_i=W_i+B_1W_{i-1}+\cdots+B_MW_{i-M}\ , \tag{3.1}$$

where $W_m=0$ for $m<0$. Thus the representation (2.1) is uniquely determined from (2.6) if the variance matrix of v_n is non-singular. This is a simple sufficient condition for the identifiability of the AR-MA representation. Consider a general situation where a stationary Markovian process v_n is defined by

$$v_{n+1}=Av_n+Bz_{n+1}\ , \tag{3.2}$$

where z_n is a zero mean $q\times 1$ white noise with $\mathrm{E}(z_n z_n')=G$ and $\mathrm{E}(z_n z_{n-i}')=0$ $(i\neq 0)$ and v_n is $s\times 1$ and is an element of the space spanned by $z_n, z_{n-1},\cdots$. It is assumed that G is non-singular. The dimension of $R(v_n)$, the space spanned by the components of v_n, is determined by the following matrix which is often called the controllability matrix in control engineering literature:

$$C_s=[B, AB, A^2B,\cdots, A^{s-1}B]\ . \tag{3.3}$$

When the rank of C_s is less than s there is a non-zero $s\times 1$ vector g such that $g'C_s=0$. This means $\mathrm{E}(g'v_n z_{n-i}')=0$ for $i=0, 1,\cdots, s-1$. As was discussed below (2.7) there is a set of coefficients a_m $(m=1, 2,\cdots, s)$

which satisfy the relation $A^s+\sum_{m=1}^{s} a_m A^{s-m}=0$. By using this relation, $\mathrm{E}(v_n z'_{n-i})$ $(i\geqq 0)$ can always be expressed as a linear combination of $\mathrm{E}(v_n z'_n)$, $\mathrm{E}(v_n z'_{n-1}),\cdots,\mathrm{E}(v_n z'_{n-s+1})$. Since $\mathrm{E}(g'v_n z'_{n-i})=0$ holds for $i=0,1,\cdots,s-1$, this means that $\mathrm{E}(g'v_n z'_{n-i})=0$ for all non-negative values of i. As v_n is an element of the space spanned by $z_n, z_{n-1},\cdots$, this means $g'v_n=0$. Thus the distribution of v_n is degenerate. When the rank C_s is equal to s there is no non-zero g for which $\mathrm{E}(g'v_n z'_{n-i})=0$ $(i=0,1,\cdots)$ and the distribution of v_n is non-degenerate. Thus the present identifiability condition can readily be checked by analyzing the rank of the matrix C_s of (3.3) with A and B defined by (3.2) and (2.6), where W_i's are obtained by (3.1) from A_i's and B_j's. When an AR-MA process satisfies the present identifiability condition the dimension of the stochastic system or the dimension of the predictor space is, from (2.6), equal to $Kr=\max(Mr,(L+1)r)$, where r is the dimension of y_n. When y_n is multivariate $(r>1)$ the dimension of the stochastic system defined by y_n need not always be an integral multiple of r. Thus it is clear that the class of the AR-MA processes which satisfy the present condition is a rather limited one and will not be wide enough as a basis to develop a fully efficient statistical identification procedure of AR-MA processes on it.

Hereafter the identifiability described in the preceding paragraph will symbolically be called the block-identifiability and the process which satisfies the condition of block-identifiability will be called block-identifiable.

4. Determination of AR-MA coefficients from covariance sequence under the block-identifiability assumption

For an r-dimensional AR-MA process (2.1) the AR coefficient matrices $B_1, B_2,\cdots, B_M$ satisfy the Yule-Walker equation

$$C(L+i)+B_1C(L+i-1)+\cdots+B_MC(L+i-M)=0 \qquad i=1,2,\cdots,M \tag{4.1}$$

where $C(j)=\mathrm{E}\,y_n y'_{n-j}$. (4.1) can be expressed in a matrix form

$$[-C(L+1),-C(L+2),\cdots,-C(L+M)] = [B_1,B_2,\cdots,B_M] \cdot \begin{bmatrix} C(L) & C(L+1) & \cdots & C(L+M-1) \\ C(L-1) & C(L) & \cdots & C(L+M-2) \\ \vdots & \vdots & & \vdots \\ C(L-M+1) & C(L-M+2) & \cdots & C(L) \end{bmatrix}. \tag{4.2}$$

The last $Mr \times Mr$ matrix of (4.2) is a block Toeplitz matrix and a recursive numerical procedure for the solution of (4.2) was discussed by Whittle [18] and Akaike [4]. When the process satisfies the block-identifiability condition the numerical procedure produces a unique solution to (4.2). When y_n is known to be block-identifiable but the values of L and M are unknown the solution must be tried for various combinations of L and M. The block Toeplitz matrix type formulation of the Yule-Walker equation can be seen to be unsuitable for this case. Instead the block Hankel matrix type formulation, which is to be discussed shortly, is much more natural and numerically efficient.

The block-identifiability condition is equivalent to the assumption of linear independence of the components of the predictors $y_{n|n}, y_{n+1|n}, \cdots, y_{n+K-1|n}$, where $K=\max(M, L+1)$. Since $\mathrm{E}(y_{n+i|n}y'_{n-j})=\mathrm{E}(y_{n+i}y'_{n-j})$ $(i, j=0, 1, \cdots)$ the analysis of dependence of the components of the predictors $y_{n|n}, y_{n+1|n}, \cdots$ is reduced to the analysis of dependence of the elementary rows of the block Hankel matrix of the covariance between $(y_n, y_{n+1}, \cdots)$ and $(y_n, y_{n-1}, \cdots)$ defined by

$$\begin{bmatrix} C(0) & C(1) & C(2) & \cdots \\ C(1) & C(2) & C(3) & \cdots \\ C(2) & C(3) & C(4) & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}. \tag{4.3}$$

Under the block-identifiability condition the analysis can be made by a blockwise procedure. This is realized by successively finding a sequence of linear transforms $y^{(i)}_{n+i}=y_{n+i}+B_1(i)y_{n+i-1}+\cdots+B_i(i)y_n$ of $y_n, y_{n+1}, \cdots$ such that $y^{(i)}_{n+i}$ is orthogonal to $y_n, y_{n-1}, \cdots, y_{n-i+1}$. The first transformation is obtained by replacing y_{n+i+1} by $y^{(1)}_{n+i+1}=y_{n+i+1}+B_1(1)y_{n+i}$ $(i=0, 1, \cdots)$ with $B_1(1)$ defined by the relation $\mathrm{E}(y^{(1)}_{n+1}y'_n)=0$, or $B_1(1)=-C(1)\cdot C(0)^{-1}$. The covariance matrix $\mathrm{E}\{(y'_n, y^{(1)\prime}_{n+1}, y^{(1)\prime}_{n+2}, \cdots)'(y'_n, y'_{n-1}, y'_{n-2}, \cdots)\}$ has its (2, 1)th block element equal to a zero matrix and the matrix below the first block-row is a block Hankel matrix. In the second step it is desired to define $y^{(2)}_{n+i}$ in such a way that $y^{(2)}_{n+2}$ is orthogonal to y_n and y_{n-1}. To realize this, first $y^{(1)}_{n+2+i}$ is transformed into $z^{(1)}_{n+2+i}=y^{(1)}_{n+2+i}+B_{22}y_{n+i}$ $(i=0, 1, \cdots)$ with B_{22} satisfying the relation $\mathrm{E}(z^{(1)}_{n+2}y'_n)=0$. $z^{(1)}_{n+2+i}$ is further transformed into $y^{(2)}_{n+2+i}=z^{(1)}_{n+2+i}+B_{21}y^{(1)}_{n+1+i}$ with B_{21} satisfying $\mathrm{E}(y^{(2)}_{n+2}y'_{n-1})=0$. By the first of these two transformations the (3, 1)th block element of the covariance matrix is turned into a zero matrix and by the second the (3, 2)th block element is also turned into a zero matrix. By the Hankel property the (4, 1)th block element $\mathrm{E}(y^{(2)}_{n+3}y'_n)$ of $\mathrm{E}\{(y'_n, y^{(1)\prime}_{n+1}, y^{(2)\prime}_{n+2}, y^{(2)}_{n+3}, \cdots)'(y'_n, y'_{n-1}, y'_{n-2}, \cdots)\}$ is a zero matrix. The kth step of transformation is defined by $z^{(k-1)}_{n+k+i}=y^{(k-1)}_{n+k+i}+B_{k2}y^{(k-2)}_{n+k-2+i}$ and $y^{(k)}_{n+k+i}=z^{(k-1)}_{n+k+i}+B_{k1}y^{(k-1)}_{n+k-1+i}$ with B_{k2} and B_{k1} defined by the relations $\mathrm{E}(z^{(k-1)}_{n+k+i}\cdot$

$y'_{n-k+2})=0$ and $\mathrm{E}\,(y^{(k)}_{n+k+i}y'_{n-k+1})=0$, respectively. Now $\mathrm{E}\,(y^{(k)}_{n+k}y'_{n-j})=0$ for $j=0, 1, \cdots, k-1$. When k is equal to $K\,(=\max\,(M, L+1))$ then $\mathrm{E}\,(y^{(k)}_{n+k}\cdot y'_{n-j})=0$ for $j=0, 1, \cdots$, and the representation

$$y^{(K)}_{n+K}=y_{n+K}+B_1y_{n+K-1}+\cdots+B_Ky_n \tag{4.4}$$

gives the desired set of AR coefficient matrices $B_1, B_2, \cdots, B_K$. Incidentally, at the kth step of the above stated procedure the $(k+1)\times(k+1)$ block matrix $\mathrm{E}\,\{(y'_n, y^{(1)\prime}_{n+1}, \cdots, y^{(k)\prime}_{n+k})'(y'_n, y'_{n-1}, \cdots, y'_{n+k})\}$ takes the form of upper triangular block matrix. If the initial covariance matrix (4.3) is replaced by $\mathrm{E}\,\{(y'_n, y'_{n+1}, \cdots)'(u'_n, u'_{n-1}, \cdots)\}$, where u_n is a stochastic process stationarily correlated with y_n, the above procedure gives a stochastic interpretation of a block triangularization procedure of block Hankel matrices developed by Rissanen [15].

Once the AR coefficient matrices $B_1, B_2, \cdots, B_K$ are obtained $y^{(K)}_{n+K}$ defined by (4.4) satisfies the representation

$$y^{(K)}_{n+K}=z_{n+K}+A_1z_{n+K-1}+\cdots+A_{K-1}z_{n+1}\ .$$

Thus the problem of determination of the MA coefficient matrices A_1, $A_2, \cdots, A_{K-1}$ of the AR-MA process y_n reduces to the problem of the determination of the MA coefficient matrices of a simple MA process $y^{(K)}_{n+K}$. A numerical solution to this problem is given also by Rissanen [15] using a blockwise recursive procedure.

5. Special Markovian representations and their use for identification

It is tempting to think that once the dimension p of the system is given the minimum number of necessary parameters to define the Markovian representation of a stationary stochastic process y_n is determined. In fact this is not true for the multivariate case. To see this a special Markovian representation is considered here. The representation is obtained by defining its state v_n as the vector of the first p linearly independent components of the pr-dimensional vector $(y'_{n|n}, y'_{n+1|n}, \cdots, y'_{n+p-1|n})'$. Now define $Y_n(k)$ $(k=1, 2, \cdots)$ by the relation

$$Y_n(jr+i)=y_{n+j}(i)_{|n}\ ,$$

where $y_n(k)$ denotes the kth component of y_n. Denote by H the set of the integers $k_1, k_2, \cdots, k_p$ such that $v_n=(Y_n(k_1), Y_n(k_2), \cdots, Y_n(k_p))'$. The set H has a special characteristic:

$$k+r\notin H\,,\qquad \text{when } k\notin H\,. \tag{5.1}$$

From the definition of $Y_n(k)$ it holds that

$$Y_{n+1}(k)_{|n}=Y_n(k+r)\ .$$

Thus the transition matrix A which satisfies the relation $v_{n+1|n}=Av_n$ is determined by the relations

$$Y_{n+1}(k_i)_{|n}=Y_n(k_j) \qquad \text{for } k_i+r=k_j \tag{5.2a}$$

$$=\sum_{j\,:\,k_j<k_i+r} A_{ij}Y_n(k_j) \qquad \text{otherwise}, \tag{5.2b}$$

where the last summation extends over the j's such that $k_j<k_i+r$. The observation matrix C which satisfies the relation $y_n=Cv_n$ is determined by the relations ($i=1, 2, \cdots, r$)

$$y_n(i)=Y_n(k_j) \qquad \text{for } i=k_j \tag{5.3a}$$

$$=\sum_{j\,:\,k_j<i} C_{ij}Y_n(k_j) \qquad \text{otherwise}, \tag{5.3b}$$

where the last summation extends over j's such that $k_j<i$. Denote the vector of the innovations $\{Y_{n+1}(k_j)-Y_{n+1}(k_j)_{|n};\ j=1, 2, \cdots, t\}$ by z_{n+1}, where t is the maximum of j such that $k_j\leqq r$. The Markovian representation is given by $v_{n+1}=Av_n+Bz_{n+1}$ and $y_n=Cv_n$, where A and C are determined by (5.2) and (5.3) and B is the matrix of regression coefficients of v_{n+1} on z_{n+1}. The matrix B can be obtained from the elements of the impulse response matrices of the system to the input z_n. From the definitions of v_n and z_n it is clear that the present Markovian representation is uniquely determined from the covariance structure of y_n. Thus the set of the r-dimensional stationary processes y_n with p-dimensional dynamics is decomposed into mutually exclusive subsets, each of which is characterized by a set H of the p integers $k_1, k_2, \cdots, k_p$ ($\leqq pr$) with the characteristic (5.1) and admits a unique Markovian representation of its elements with the transition and the observation matrices, A and C, of the forms respectively described by (5.2) and (5.3). Now for each subset specified by H consider an $r\times(p+1)$ matrix S of which (i, j)th element $S(i, j)$ is 1 when for each element y_n of the subset $y_{n+j-1}(i)_{|n}$ is retained in v_n, i.e. when $i+(j-1)r\in H$, and 0 otherwise. The $p+1$st column of S is always a zero vector. The matrix S takes a special form where each row has first several, or no, elements equal to 1 and others equal to 0. The total number of 1's within S is equal to p. Shift the columns of S one step to the right and fill in 1's in the empty fisrt column to define another $r\times(p+1)$ matrix S_+. Define $T=S+S_+$. The 1's in the first column of T correspond to the i's of (5.3b) and the 1's in other columns correspond to k_i's of (5.2b). Starting at the (1, 1)th element of T, calculate the number of 2's columnwise until the ith 1 ($i=1, 2, \cdots, r$). The sum of these numbers is equal to the number of parameters within A_{ij} and C_{ij} of (5.2) and (5.3). The number of parameters within B is $p\times t$, where t is the dimension of z_n. From the definition of the subset it is obvious that no further reduc-

tion of the number of parameters is possible. Now it is easy to see by some examples that for a given p there may be different patterns of the distribution of 1's within S which require different number of parameters within A and C.

The special representation discussed above can be applied to produce an ultimate answer to the identifiability problem of the AR-MA representation of a stationary process with a finite dimensional dynamics, under the assumption of non-singularity of the variance matrix of its innovations. For this case C takes the form $C=[I\ 0]$. Denote by p_i the number of 1's in the ith row of the above defined matrix S. By rewriting (5.2b) the following representation is obtained:

$$y_{n+p_i}(i)_{|n}=\sum_{m=0}^{p_i} C_m(i\cdot)y_{n+p_i-m|n}\ , \tag{5.4}$$

where $C_m(i\cdot)$ is a $1\times r$ vector determined by (5.2b) and $C_0(i, j)$, the $(1, j)$th element of $C_0(i\cdot)$, is always equal to zero for $j\geqq i$. Define $q=\max(p_1, p_2, \cdots, p_r)$. From (5.4) one can get a relation

$$y_{n+q|n}=\sum_{m=0}^{q} C_m y_{n+q-m|n}\ ,$$

which gives an AR-MA type representation

$$\begin{aligned}(I-C_0)y_{n+q}-C_1y_{n+q-1}-\cdots-C_qy_n \\ =D_0z_{n+q}+D_1z_{n+q-1}+\cdots+D_{q-1}z_{n+1}\ ,\end{aligned} \tag{5.5}$$

where $C_m(i, j)$, the (i, j)th element of C_m, is put equal to the $(1, j)$th element of $C_m(i\cdot)$ defined by (5.4) or equal to zero, if undefined by (5.4). From the definition of C_0, $I-C_0$ is non-singular and an AR-MA representation of y_n is given by

$$y_n+B_1y_{n-1}+\cdots+B_qy_{n-q}=z_n+A_1z_{n-1}+\cdots+A_{q-1}z_{n-q+1}\ , \tag{5.6}$$

where $B_i=-(I-C_0)^{-1}C_i$ and $A_i=(I-C_0)^{-1}D_i$. Thus it has become clear that without any assumptions such as the block-identifiability, a stationary process with a finite dimensional dynamics and a non-singular innovation variance matrix always has a uniquely identifiable AR-MA representation (5.6). This result has been obtained with the aid of the special Markovian representation introduced in this section and very clearly shows the advantage of the Markovian representation over the AR-MA representation for the purpose of multivariate stochastic system analysis.

Since the Markovian representation of a stationary stochastic process y_n is a state space representation of a time-invariant linear system which generates y_n from the input z_n, the innovation of y_n, any special Markovian representation can be defined by using a special state space re-

presentation of the corresponding system. The subject of the special state space representations of time-invariant discrete-time linear systems has been discussed extensively and several alternatives of the special Markovian representation introduced in this section can easily be introduced (Akaike [7]). Especially by replacing the definition of $Y_n(k)$ by

$$Y_n((i-1)p+j+1)=y_{n+j}(i)_{|n} \qquad i=1, 2, \cdots, r;\ j=0, 1, \cdots, p-1\,,$$

and searching for the first p linearly independent components of $Y_{n,p}=(Y_n(1), Y_n(2), \cdots, Y_n(rp))$ one can get another basis v_n of the predictor space. By this choice of the basis, the definition of z_n in the Markovian representation can be replaced by the vector of innovations of those components of y_n which are retained in v_n to make the matrix B unique. The assumption of non-singularity of the innovation matrix is now unnecessary and the unique Markovian representation thus obtained gives a corresponding unique AR-MA representation.

Once a special representation is specified, at least conceptually there is no difficulty in developing a statistical identification procedure based on the maximum likelihood method to be described in the next section. The only difficulty which prevents the practical application of the procedure is caused by the existence of the vast number of possible choices of the basis of the predictor space. It is almost prohibitive to perform the maximum likelihood computation for every possible choice of the basis. Thus the feasibility of the procedure is almost entirely depenent on how to get good initial guesses of the dimension of the system and the structure of the desired basis. A solution to this problem is given in the last section.

It should be mentioned here that in an unpublished paper by Rissanen [14] the special representation discussed at the beginning of this section was used implicitly to develop a consistent estimation procedure of the parameters of a multivariate autoregressive process.

6. Maximum likelihood procedure and information theoretic criterion

When a stationary Gaussian process y_n has a Markovian representation

$$v_{n+1}=Av_n+Bz_{n+1} \tag{6.1a}$$

$$y_n=Cv_n\,, \tag{6.1b}$$

the representation can conveniently be used to define an approximation to the likelihood function. Under the assumption of non-singularity of the innovation variance matrix, $\mathrm{E}\,(z_nz_n')$, y_n can be expressed in the form

$$y_n=CAv_{n-1}+z_n\,. \tag{6.2}$$

When a record of observations $(y_n;\ n=1, 2, \cdots, N)$ is given, by assuming $v_0=0$ a set of realization of z_n $(n=1, 2, \cdots, N)$ can be obtained by (6.2) and (6.1a). The logarithm of the approximate likelihood function is given by using the realization of z_n in the form

$$-\frac{N}{2}\{r \log 2\pi+\log |G|+\mathrm{tr}\,(G^{-1}C_0)\}\ , \tag{6.3}$$

where $C_0=\frac{1}{N}\sum_{n=1}^{N} z_n z_n'$ and G is the assumed covariance matrix of the innovations. This result corresponds to the asymptotic evaluation of the Gaussian likelihood given by Whittle ([16], (5.1)). If the Fourier transform $Y(f)$ of $(y_n;\ n=1, 2, \cdots, N)$ is defined by

$$Y(f)=\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \exp(-i2\pi f n)y_n$$

and $V(f)$ and $Z(f)$ by

$$V(f)=\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \exp(-i2\pi f n)v_n$$

$$Z(f)=\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \exp(-i2\pi f n)z_n\ ,$$

then by neglecting the effect of end conditions it holds that

$$Z(f)=[C\{I-\exp(-i2\pi f)A\}^{-1}B]^{-1}Y(f)\ .$$

By using the relation

$$C_0=\int_{-1/2}^{1/2} Z(f)Z^*(f)df\ ,$$

where * denotes the conjugate transpose, one can get an explicit representation of (6.3) in terms of the matrices A, B, C and G and the Fourier transform of y_n. Thus if only a non-redundant parametrization of A, B, C is available the maximum likelihood procedure can be realized numerically by following the line of approach developed for the AR-MA representation (Akaike [5]). The special Markovian representation introduced in the preceding section can directly be used for this purpose. The importance of this non-redundant parametrization of a Markovian representation in statistical identification of a stochastic system can be understood more clearly by persuing its analogy to the factor analysis model. If v_n in (6.1) is replaced by $u_n=Tv_n$, with T non-singular, another Markovian representation of y_n is given by $u_{n+1}=TAT^{-1}u_n+TBz_{n+1}$ and $y_n=CT^{-1}u_n$. Thus it is obvious that there is no meaning in trying to find a unique Markovian representation without any further restric-

tions. The situation is quite similar to the case of the indeterminacy of the factor analysis model (see, for example, Lawley and Maxwell [12]). The dimension of the system plays a role similar to that of the number of factors in factor analysis. As was already seen in the preceding section, the dimension is not sufficient to determine the minimum number of parameters necessary to define the Markovian representation and the parametrization is much more complicated than that of the factor analysis. As will be made clear in the following discussion of the use of an information theoretic criterion for the decision of the models, the ultimate use of these models, the Markovian or the AR-MA representation of a stationary stochastic process and the factor analysis model, is to provide a reliable estimate of the related covariance structure by controlling the number of parameters within the model when there is not sufficient prior information to limit to a unique model. The decision on the dimension and the structure of the basis or on the number of factors is obviously the crucial point in applying these models to real data.

By using (6.3) or directly from the result given by Whittle [16], the logarithm of the approximate likelihood function of the observations $(y_n;\ n=1, 2, \cdots, N)$ can be given by

$$-\frac{N}{2}\left[r \log 2\pi+\log |G|+\int_{-1/2}^{1/2} \operatorname{tr}\{Y(f)Y^*(f)P^{-1}(f)\}df\right], \tag{6.4}$$

where $P(f)=[C\{I-\exp(-i2\pi f)A\}^{-1}B]G[C\{I-\exp(-2\pi f)A\}^{-1}B]^*$ is the spectral density matrix of the assumed model. By taking the expectation of (6.4) $Y(f)Y^*(f)$ is replaced by $P_N(f)=\mathrm{E}\{Y(f)Y^*(f)\}$ which converges to the true spectral density matrix $P_\infty(f)$ as N tends to infinity. The negative of the expectation of a log-likelihood is, ignoring an additive constant, identical to the mean amount of information for discrimination between the assumed model and the true distribution per observation from true distribution as defined by Kullback and Leibler ([11], (2.4)). Thus if the negative of the expectation of the likelihood (6.4) is divided by N this gives the average of the mean information for discrimination. This quantity will be called the average information of the assumed model from the true model, or simply the average information, obtained from $(y_1, y_2, \cdots, y_N)$. When N is made infinite the average information obtained from $(y_1, y_2, \cdots, y_N)$ tends to a quantity given by

$$\frac{1}{2}\left[r \log 2\pi+\log |G|+\int_{-1/2}^{1/2} \operatorname{tr}\{P_\infty(f)P^{-1}(f)\}df\right]. \tag{6.5}$$

From the definition it would be reasonable to call this quantity the average information for discrimination of the assumed model from the

true model. The maximum likelihood estimates of the parameters are obtained by maximizing (6.4), which is equivalent to minimizing (6.5) with $P_\infty(f)$ replaced by $Y(f)Y^*(f)$. Thus the maximum likelihood estimation procedure is equivalent to first constructing a stationary Gaussian process with the spectral density matrix $Y(f)Y^*(f)$, or the covariance matrix defined by $C(j)=\frac{1}{N}\sum_{n=1}^{N-j} y_{n+j}y_n'$ $(j=0,1,2,\cdots,N-1)$, $C(-j)=C(j)'$ and $C(j)=0$ for $|j|\geqq N$, and then minimizing the average information for discrimination of the model being fitted from this constructed Gaussian process model. Since it is well known that $Y(f)Y^*(f)$ can not even be a consistent estimate of $P_\infty(f)$ the above interpretation of the maximum likelihood estimates gives a clear indication of the difficulty inherent in the model fitting by the maximum likelihood procedure. When the assumed model is too flexible with too many number of parameters, then the estimated covariance structure will come very close to the one given by $Y(f)Y^*(f)$. In that case the estimate would be unreliable. When the assumed model is too inflexible with too small number of parameters the estimated covariance structure may not be able to sufficiently approximate the true structure. In the case of the Markovian model fitting the flexibility is controlled by the dimension of the system and the decision on the dimension becomes crucial for the success of the fitting procedure.

It has been found that a statistic defined by

$$(6.6)\qquad (-2)\log_e(\text{maximum likelihood}) + 2(\text{number of adjusted parameters})$$

is useful for the purpose of the above stated decision on the model flexibility (Akaike [2], [3]). In the present situation this statistic is meant to be an estimate of $2N$ times the average information for discrimination of the assumed model from the true model. The first term of (6.6) stands for the penalty of the badness of fit and the second term for the penalty of increased unreliability. (6.6) tells us that if the badness of fit is identical for two models the one with less number of parameters should be preferred. The definition of the above penalty is based on the chi-square approximation of the limiting distribution of the difference of $2\log_e$(maximum likelihood) from its expectation when the model is exact. For the justification of this chi-square approximation in the time series situation, see Whittle [17]. For the Markovian representations the number of adjusted parameters should be defined as the minimum number of parameters required to define the model and if the special representation introduced in the preceding section is used the model is specified by the dimension of the system and the set of indices H which specify the choice of a basis of the predic-

tor space. For univariate case only the dimension of the system is sufficient to specify a model. From the standpoint of statistical model fitting or identification by using (6.6) the Markovian representation of a stationary stochastic process is merely one choice of the model which will give a good approximation to an arbitrary covariance structure with rather small number of parameters.

7. Determination of dimension and basis of predictor space

Under the assumption of non-singularity of the innovation variance matrix, the search for a basis of the predictor space of y_n, which gives a special Markovian representation discussed in Section 5, can be realized by the search of the first p linearly independent elementary rows of the block Hankel matrix (4.3). If p is not known but its upper bound q is known then the search can be completed by the analysis of dependence of the qr elementary rows within the first q block rows. From the Markovian representation (2.7) the covariance matrices $C(i)=\mathrm{E}\, y_{n+i}y_n'$ $(i=0, 1, \cdots)$ can be expressed in the form

$$C(i)=CA^iPC', \tag{7.1}$$

where $P=\mathrm{E}\, v_nv_n'$ and A is a $p\times p$ matrix. As was discussed below (2.7), there is a set of constants a_m $(m=1, 2, \cdots, p)$ such that $A^p+a_1A^{p-1}+\cdots+a_pI=0$. From (7.1) it also holds that

$$C(i+p)=-\sum_{m=1}^{p} a_mC(i+p-m) \qquad i=0, 1, \cdots . \tag{7.2}$$

This shows that in the block Hankel matrix (4.3) any block column can be expressed as a linear combination of the first p block columns. Thus, when only q is known, the analysis of dependence of the elementary rows within the first q block rows of (4.3) can be realized by the analysis of dependence within the elementary rows of the $qr\times qr$ matrix

$$C_q=\begin{bmatrix} C(0) & C(1) & \cdots & C(q-1) \\ C(1) & C(2) & \cdots & C(q) \\ \vdots & \vdots & & \vdots \\ C(q-1) & C(q) & \cdots & C(2q-2) \end{bmatrix}. \tag{7.3}$$

C_q can be expressed in the form $C_q=\mathrm{E}\, y_+y_-'$, where $y_+=(y_n', y_{n+1}', \cdots, y_{n+q-1}')'$, and $y_-=(y_n', y_{n-1}', \cdots, y_{n-q+1}')'$, and the matrix of the regression coefficients of y_+ on y_- is given as $C_qD_q^{-1}$, where $D_q=\mathrm{E}\,(y_-y_-')$ and it is assumed that D_q is non-singular. The dimension p of the system, or the rank of C_q, is equal to the dimension of the linear space spanned by the components of the projection of y_+ on the linear space spanned

by the components of y_- and thus is equal to the number of non-zero canonical correlation coefficients between y_+ and y_-. When the process y_n is ergodic, $\tilde{C}(j)$ defined by

$$\tilde{C}(j)=\frac{1}{N}\sum_{n=1}^{N-j}(y_{n+j}-\bar{y})(y_n-\bar{y})'$$

where $\bar{y}=\left(\frac{1}{N}\right)\sum_{n=1}^{N}y_n$, is a consistent estimate of $C(j)$ for $j=0, 1, \cdots, 2q-2$, and the sample canonical correlation coefficients obtained by replacing $C(j)$ by $\tilde{C}(j)$ in the definition of the canonical correlation coefficients give consistent estimates of the theoretical canonical correlation coefficients, which might be useful for the decision on the dimension of the system.

The practical utility of the above stated estimates of the canonical correlation coefficients was checked with an artificially generated process. The original process y_n is defined by

$$y_n-0.9y_{n-1}+0.4y_{n-2}=z_n+0.8z_{n-1}\ ,$$

where y_n is a scalar and z_n is a Gaussian white noise with $\mathrm{E}\,z_n=0$ and $\mathrm{E}\,z_n^2=1$. For this case the dimension p of the system is equal to 2 and the Markovian representation (2.6) with $K=2$ gives a minimal representation

$$v_{n+1}=\begin{bmatrix}0 & 1\\ -0.4 & 0.9\end{bmatrix}v_n+\begin{bmatrix}1\\ 1.7\end{bmatrix}z_{n+1}$$

$$y_n=[1\quad 0]v_n\ .$$

Assuming $y_0=y_{-1}=z_0=0$ two sets of records $\{y_n;\ n=1, 2, \cdots, 550\}$ were obtained. The first fifty points of each record were discarded to suppress the effect of the initial transients and the remaining two sets of records each with the data length $N=500$ were used for the analysis. These data are designated as data #1 and 2. The results of the canonical correlation analysis are given in Table 1. The values of r_i^2, the squared sample canonical correlation coefficient, strongly suggest the choice $p=2$, the correct order in this example. The coefficients which determine the canonical variables showed consistent behaviour within the two sets of data for the canonical variables corresponding to the first two largest canonical correlation coefficients but were quite inconsistent for the rest of the variables. Analogous results were obtained for other two sets of data, data #3 and 4, with $N=100$ and are also illustrated in Table 1. In the ordinary canonical correlation analysis with independent multivariate observations the variable $\chi_{(i)}^2$ will asymptotically be distributed as a chi-square variable with the corresponding

Table 1

i	Data #1				Data #2			
	r_i^2	$\chi^2_{(i)}$	d.f. (i)	I.C. (i)	r_i^2	$\chi^2_{(i)}$	d.f. (i)	I.C. (i)
0		∞		∞		∞		∞
1	1.000	612.78	16	580.78	1.000	647.89	16	615.89
2	0.701	9.79	9	−8.10*	0.717	17.27	9	−0.73
3	0.011	4.19	4	−3.76	0.020	7.09	4	−0.91
4	0.008	0.05	1	−1.95	0.013	0.76	1	−1.23*
5	0.000	0	0	0	0.002	0	0	0

i	Data #3				Data #4			
	r_i^2	$\chi^2_{(i)}$	d.f. (i)	I.C. (i)	r_i^2	$\chi^2_{(i)}$	d.f. (i)	I.C. (i)
0		∞		∞		∞		∞
1	1.000	125.59	16	93.59	1.000	87.09	16	55.09
2	0.688	10.28	9	−7.72*	0.538	10.78	9	−7.22*
3	0.069	3.22	4	−4.78	0.090	1.50	4	−6.50
4	0.022	0.99	1	−1.01	0.014	0.14	1	−1.86
5	0.010	0	0	0	0.001	0	0	0

r_i^2=square of the ith largest sample canonical correlation coefficient between $(y_n, y_{n+1}, \cdots, y_{n+4})'$ and $(y_n, y_{n-1}, \cdots, y_{n-4})'$

$\chi^2_{(i)} = -N \log_e \prod_{j=i+1}^{5} (1-r_j^2)$ N=500 for data #1 and 2, 100 for data #3 and 4

d.f. (i)=Difference of the number of independent parameters between the full rank model and the rank i model

I.C. (i)=$\chi^2_{(i)}$−2(d.f. (i))

* denotes the minimum of I.C. (i)

number of degrees of freedom indicated by d.f. (i) when the theoretical values of r_j^2 $(j>i)$ are equal to zero (Anderson [8], p. 327). The values of $\chi^2_{(i)}$ and its d.f. (i) for the maximum possible value of i are set equal to zero. Also when $r_i^2=1$, $\chi^2_{(i-1)}$ and I.C. $(i-1)$ are put equal to infinity. It is not clear to what degree the asymptotic chi-square distribution can approximate the distribution of $\chi^2_{(i)}$ in the present time series situation, but the statistic I.C. $(i)=\chi^2_{(i)}-2$(d.f. (i)) will be useful as a variant of the information theoretic criterion discussed in Section 6. In this statistic the first term $\chi^2_{(i)}$ stands for the increase of badness of fit of the model by the introduction of the assumption $r_j^2=0$ $(j>i)$ and the second term -2(d.f. (i)) stands for the decrease of unreliability by the restriction of the model. If the chi-square approximation of the distribution of $\chi^2_{(i)}$ is not valid the performance of the decision procedure may depend on the structure of the process under observation. The best choice of the dimension p of the system is given as the value of i for which I.C. (i) is the minimum. In Table 1 the values of I.C. (i) show that the best choice of the dimension is given by $p=2$, the true value

of the dimension, for data #1, 3 and 4. For the data #2 $p=4$ is chosen as optimal but the difference between I.C. (i)'s $(i=2, 3, 4)$ are almost meaningless compared with the expected sampling variabilities of the related statistics under the assumption of the validity of the chi-square approximations.

If the sample canonical correlation coefficients are obtained by treating the data $(y'_{ns}, y'_{ns+1}, \cdots, y'_{(n+1)s-1})'$ and $(y'_{ns}, y'_{ns-1}, \cdots, y'_{ns-t+1})'$, with s equal to a positive integer and $n=1, 2, \cdots, N$, as if they were N independent observations, then if t is sufficiently large the asymptotic distribution of the corresponding $\chi^2_{(i)}$ will be approximated by a chi-square distribution with the number of degrees of freedom equal to $(sr-ir)\cdot(tr-ir)$, where r is the dimension of the vector y_n. This is due to the fact that the residual of y_{ns+t} after subtracting the projection on the space spanned by the components of $(y'_{ns}, y'_{ns-1}, \cdots, y'_{ns-t+1})'$ will approximately be independent for different values of n. For this case the statistic I.C. $(i)=\chi^2_{(i)}-2(\text{d.f. }(i))$ will behave as the difference between the information theoretic criterion corresponding to a restricted model for which only the $i-1$ largest canonical correlations between $(y'_{ns}, y'_{ns+1}, \cdots, y'_{(n+1)s-1})'$ and $(y'_{ns}, y'_{ns-1}, \cdots, y'_{ns-t+1})'$ are not assumed to be equal to zero and that corresponding to the unrestricted model. In spite of its structural simplicity of this modified definition of the sample canonical correlation coefficients their sampling variabilities will generally be larger compared with those of the sample canonical correlation coefficients obtained by simply replacing $C(j)$ by the sample covariance matrix $\tilde{C}(j)$ in the definition of the canonical correlation coefficients.

The decision on the value of p is sufficient for the fitting of a univariate $(r=1)$ AR-MA or Markovian model. For the multivariate $(r>1)$ case, the above stated criterion I.C. (i) can also be used to decide on the first p linearly independent elementary rows of (7.3), where the decision is also made on the value of p itself. First calculate the sample canonical correlation coefficients between $u=(y_+(1), y_+(2), \cdots, y_+(r+1))'$ and y_-, where $y_+(k)$ denotes the kth component of y_+, $y_+=(y'_n, y'_{n+1}, \cdots, y'_{n+q-1})'$, $y_-=(y'_n, y'_{n-1}, \cdots, y'_{n-q+1})'$ and it is assumed that the dimension p of the system is less than q. If the criterion attains its minimum at $i=r+1$, i.e. I.C. $(r)>$I.C. $(r+1)$ $(=0)$, then retain $y_+(r+1)$ in u, otherwise drop $y_+(r+1)$ from u. When $y_+(r+1)$ is dropped from u discard $y_+(jr+1)$ $(j=2, 3, \cdots, q-1)$ from y_+ and pick up the canonical variable $a'_{r+1}u$ which corresponds to the minimum canonical correlation coefficient where a_{r+1} is an $(r+1)$-dimensional vector $(a_{r+1}(1), a_{r+1}(2), \cdots, a_{r+1}(r+1))'$. An estimate of A_{ij} of (5.2b) with $k_i=1$ is given by $-(a_{r+1}(j))\cdot(a_{r+1}(r+1))^{-1}$. Now include $y_+(r+2)$ into u to define the new u and repeat the analysis to decide on the rejection of $y_+(r+2)$ from u. When $y_+(r+2)$ is rejected, $y_+(jr+2)$ $(j=2, 3, \cdots, q-1)$ are discarded from y_+

and an estimate of A_{ij} of (5.2b) with $k_i=2$ is obtained analogously as in the case with $y_+(r+1)$. Repeat the canonical correlation analysis and decision with the remaining components of y_+ until the last one. The u at this final stage defines a choice on the minimal basis and the corresponding estimates of A_{ij}'s of (5.2b) provide an estimate of the transition matrix A. Many other variants of the selection procedures are conceivable.

The feasibility of the above stated type procedure on the decision of the basis was checked with an artificially generated two-dimensional process y_n which is defined by

$$(7.4)\quad y_n+\begin{bmatrix}-0.9 & 0\\ 0 & -1.5\end{bmatrix}y_{n-1}+\begin{bmatrix}0.4 & 0\\ 0 & 1.2\end{bmatrix}y_{n-2}+\begin{bmatrix}0 & 0\\ 0 & -0.448\end{bmatrix}y_{n-3}$$
$$=z_n+\begin{bmatrix}0.8 & 0\\ 0 & 0\end{bmatrix}z_{n-1}\,,$$

where z_n is a Gaussian white noise with zero mean and unit variance matrix. Assuming the zero initial condition, i.e. $y_0=y_{-1}=y_{-2}=z_0=0$, two realizations of z_n ($n=1, 2, \cdots, 550$) were used to generate two realizations of y_n ($n=1, 2, \cdots, 550$) and the first fifty points were discarded from both realizations. The following numerical results are based on these two sets of data, #5 and #6 each with $N=500$. The dimension p of the system defined by (7.4) is 5 and the first 5 independent components of $(y'_{n|n}, y'_{n+1|n}, \cdots)$ are $y_n(1)_{|n}$, $y_n(2)_{|n}$, $y_{n+1}(1)_{|n}$, $y_{n+1}(2)_{|n}$, $y_{n+2}(2)_{|n}$ and the process is not block-identifiable. By following the discussion of Section 5 and using $v_n=(y_n(1)_{|n}, y_n(2)_{|n}, y_{n+1}(1)_{|n}, y_{n+1}(2)_{|n}, y_{n+2}(2)_{|n})'$ as the state of the representation one can get a minimal Markovian representation of the process in the following form:

$$v_{n+1}=\begin{bmatrix}0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ -0.4 & 0 & 0.9 & 0 & 0\\ 0 & 0 & 0 & 0 & 1\\ 0 & 0.448 & 0 & -1.2 & 1.5\end{bmatrix}v_n+\begin{bmatrix}1 & 0\\ 0 & 1\\ 1.7 & 0\\ 0 & 1.5\\ 0 & 1.05\end{bmatrix}z_{n+1}$$

$$y_n=\begin{bmatrix}1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\end{bmatrix}v_n\,.$$

When autoregressive models were fitted to the data using a criterion which is equivalent to the present information criterion (Akaike [1]), the orders chosen for the simple bivariate autoregressive representations were 6 and 5, respectively, for data #5 and data #6. Based on this observation $q=5$ was used as a tentative choice of q in the definition of $y_-=(y'_n, y'_{n-1}, \cdots, y'_{n-q+1})'$. Canonical correlation analysis between $u_m=$

$(y_+(1), y_+(2), \cdots, y_+(m))$ and y_- were performed, where $y_+(i)$ is the ith component of $y_+=(y_n', y_{n+1}', \cdots, y_{n+q-1}')'$. The results are illustrated in Table 2. From the behaviour of the information criterion, which is analogously defined as in Table 1, $y_+(5)_{|n}$, or $y_{n+2}(1)_{|n}$, is judged to be linearly dependent on the preceding components of $u_{m|n}$ for the both sets of data at $m=5$. At $m=6$, $y_+(6)_{|n}$ is judged to be linearly independent of the preceding components of $u_{m|n}$. At $m=7$ and 8 the selected number of linearly independent components remains at 5 for the both sets of data. This shows that $y_+(7)_{|n}$ and $y_+(8)_{|n}$ are judged to be linearly dependent on the preceding components of $u_{r|n}$ and the

Table 2

m	i	Data #5			Data #6		
		$\chi^2_{(i)}$	d.f. (i)	I.C. (i)	$\chi^2_{(i)}$	d.f. (i)	I.C. (i)
3	2	405.57	8	389.57	375.03	8	359.03
	3	0	0	0 *	0	0	0 *
4	2	824.73	16	792.73	722.09	16	690.09
	3	395.32	7	381.32	315.49	7	301.49
	4	0	0	0 *	0	0	0 *
5	2	951.76	24	903.76	855.29	24	807.29
	3	419.12	14	391.12	341.06	14	313.06
	4	2.84	6	−9.16*	8.75	6	−3.25*
	5	0	0	0	0	0	0
6	2	1059.98	32	995.98	958.95	32	894.95
	3	517.25	21	475.25	427.50	21	385.50
	4	29.30	12	5.30	36.84	12	12.84
	5	1.65	5	−8.35*	8.47	5	−1.53*
	6	0	0	0	0	0	0
7	2	1113.74	40	1033.74	1042.71	40	962.71
	3	526.16	28	470.16	449.31	28	393.31
	4	35.61	18	−0.39	45.04	18	9.04
	5	7.57	10	−12.43*	15.62	10	−4.38*
	6	0.72	4	−7.28	6.83	4	−1.17
	7	0	0	0	0	0	0
8	2	1115.40	48	1019.40	1044.81	48	948.81
	3	527.45	35	457.45	451.79	35	381.79
	4	37.35	24	−10.66	47.68	24	−0.32
	5	9.26	15	−20.74*	18.19	15	−11.81*
	6	2.06	8	−13.94	8.93	8	−7.07
	7	0.52	3	−5.48	2.08	3	−3.92
	8	0	0	0	0	0	0

* denotes the minimum of I.C. (i)

search for the basis is terminated. The selected basis is $(y_+(1)_{|n}, y_+(2)_{|n}, y_+(3)_{|n}, y_+(4)_{|n}, y_+(6)_{|n})$ for the both sets of data, which is identical to the theoretical result. The present identity between the experimental and theoretical result is obtained only by chance, but the behaviour of the statistics in Table 2 gives a clear feeling of the utility of this type of procedure for general practical applications. In this experiment, for the simplicity of the computer program, the discarding of the components judged to be dependent was not implemented. The definition of $\chi^2_{(i)}$ was analogous to that of Table 1 but with N replaced by $N-0.5(2m+2q+1)$. This difference of the definition of $\chi^2_{(i)}$ has little effect on the present application.

It is almost certain that if the $(i+1)$st largest canonical correlation coefficient r_{i+1} is equal to zero then $N^{-a}\chi^2_{(i)}$ $(0<a<1)$ will converge to zero in probability as N tends to infinity. If this convergence is assumed and I.C. (i) is replaced by i.c. $(i)=N^{-1}\chi^2_{(i)}-2N^{a-1}$(d.f. (i)) the present decision procedure provides consistent estimates of the desired basis and the corresponding matrix A. This can easily be seen from the fact that $N^{-1}\chi^2_{(i)}$ converges in probability to a positive constant and dominates the behaviour of i.c. (i) when r_{i+1} is not zero while $-2N^{a-1}$ $\cdot$d.f. (i) dominates when r_{i+1} is zero. This result may be considered to be a fundamental one in the subject of statistical identification of multivariate stochastic systems with finite dimensional dynamics. In practical applications where the dimension of the system is infinite the replacement of I.C. (i) by i.c. (i) will stress the tendency to pick up a rather low dimensional model.

Once the estimate of the basis is determined, an estimate of the corresponding matrix B can be obtained from any estimate of W_m $(m=1, 2, \cdots)$ of (2.2). An estimate of W_m is obtained by fitting an autoregressive model and computing the response of the system to an impulsive input. This will be useful to produce an initial estimate of B to start the maximum likelihood computation.

It should be remembered here that by a modification of the search procedure for a basis of the predictor space, which was described in Section 5, the assumption of non-singularity of the innovation variance matrix can be eliminated. With this modification the procedure described in this section will give a minimal set of components of y_n which are judged to be useful for the description of the stochastic system and retained for the further analysis. The application of this type of approach to the problem raised by Priestley and others [13] concerning the reduction of the number of measurements in the control of a complex system will be a subject of further study. Also the procedures described in this section are only for the initial determination of the dimension or the basis of the predictor space. The final decision will

be made by comparing the values of the information criterion defined by the maximized likelihood for several possible choices of the dimension or the basis. Theoretical analysis of the asymptotic distribution of $\chi^2_{(i)}$ is also a subject of further study.

The procedure described in this paper provides only a starting point for the development of practical procedures of fitting AR-MA models. Much remains to be done for the development of a computationally and statistically efficient fitting procedure.

Acknowledgements

The author wishes to thank Professor M. B. Priestley, Dr. T. Subba Rao, Dr. H. Tong and Mrs. V. Haggan for many stimulating discussions and to Miss E. Arahata for conducting the numerical experiment. The author is also grateful to Professor W. Gersch for the helpful comments. Thanks are especially due to Professor E. J. Hannan for kindly pointing out a serious error in the original version of Section 3. The research was supported by a Science Research Council grant at the University of Manchester Institute of Science and Technology.

INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. (1971). Autoregressive model fitting for control, *Ann. Inst. Statist. Math.*, **23**, 163-180.

[2] Akaike, H. (1972). Use of an information theoretic quantity for statistical model identification, *Proc. 5th Hawaii International Conference on System Sciences*, 249-250.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Proc. 2nd International Symposium on Information Theory*, (B. N. Petrov and F. Csaki eds.), *Akademiai Kiado*, Budapest, 267-281.

[4] Akaike, H. (1973). Block Toeplitz matrix inversion, *SIAM J. Appl. Math.*, **24**, 234-241.

[5] Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models, *Biometrika*, **60**, 255-265.

[6] Akaike, H. (1973). Markovian representation of stochastic processes by canonical variables, to be published in *SIAM J. Control.*

[7] Akaike, H. (1973). Stochastic theory of minimal realizations, to be published in *IEEE Trans. Automat. Contrl.*

[8] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

[9] Hannan, E. J. (1969). The identification of vector mixed autoregressive-moving average systems, *Biometrika*, **56**, 223-225.

[10] Kalman, R. E., Falb, P. L. and Arbib, M. A. (1969). *Topics in Mathematical System Theory*, McGraw-Hill, New York.

[11] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statist.*, **22**, 79-86.

[12] Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, Butterworths, London.

[13] Priestley, M. B., Subba Rao, T. and Tong, H. (1972). Identification of the structure of multivariable stochastic systems, to appear in *Multivariate Analysis* III, Ed. P. R. Krishnaiah, Academic Press.

[14] Rissanen, J. (1972). Estimation of parameters in multi-variate random processes, unpublished.

[15] Rissanen, J. (1973). Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factorising positive matrix polynomials, *Mathematics of Computation*, **27**, 147-154.

[16] Whittle, P. (1953). The analysis of multiple stationary time series, *J. R. Statist. Soc.*, B, **15**, 125-139.

[17] Whittle, P. (1962). Gaussian estimation in stationary time series, *Bulletin L'Institut International de Statistique*, **39**, 2e Livraison, 105-129.

[18] Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate factorization of a spectral density matrix, *Biometrika*, **50**, 129-134.

[19] Whittle, P. (1969). A view of stochastic control theory, *J. R. Statist. Soc.*, A, **132**, 320-334.