Ch 1

OVERLAPPING PARAMETRIZATIONS FOR THE REPRESENTATION OF MULTIVARIATE STATIONARY TIME SERIES*

Michel Gevers and Vincent Wertz

ABSTRACT. When identifying a (state-space or ARMA) model for a multivariate stationary stochastic process using a black-box approach, the first problem consists in finding the order of the model and defining a uniquely indentifiable parametrization. For a given order, several "overlapping" parametrizations, all involving the same number of parameters, can usually be fitted to the same process. The question then arises as to whether some parametrizations are better than others. We present an asymptotic result showing that all overlapping parametrizations give the same value to the determinant of the Fisher information matrix and, therefore, with many identification schemes, to the determinant of the asymptotic error covariance matrix. For finite data, some structures may still be better than others, and two heuristic structure estimation methods are analyzed. Some simulation results are also presented.

1. INTRODUCTION. An important and widely studied problem in the theory of identification of multivariate stationary finite-dimensional stochastic processes is that of determining the structure of the state-space or ARMA model for that process such that the model parameters become uniquely identifiable. Two different lines of thought have been followed for this problem. The first idea is to use canonical (state-space or ARMA) forms [1]-[5]. To any finite dimensional process one can associate a canonical form in a unique way by specifying a selection procedure. Different selection procedures will lead to different canonical forms, but the parameters in any two canonical representations of a given process are related by a bijective relationship. The structure of a canonical representation of a process is determined by a set of "structural invariants" (e.g. the Kronecker invariants) which are again uniquely defined by the process and the selection procedure. The disadvantage with using canonical forms is that the estimation of those structural invariants is very critical: if they are wrongly estimated, then the parameter estimation problem becomes ill-conditioned.

*This work was performed in part while the authors were on leave at the University of Newcastle, New South Wales, Australia. The work was supported by IRSIA (Belgium) and the Australian Research Grants Committee.

In recent years an alternative approach has been proposed, namely that of using "overlapping parametrizations" [6]-[12]. It has been recognized that the set of all finite dimensional systems can be represented by a finite number of parametrizations, each parametrization being uniquely identifiable. To each parametrization there corresponds a set of integers called "structure indices." Each of these parametrizations also called "structure" is able to represent almost all finite dimensional systems, each system can normally be represented by more than one such parametrization, and any two parametrizations for a given process are related by a linear transformation which corresponds to a coordinate transformation in Euclidean space; hence the use of the word "overlapping" parametrizations. Now because a process can be represented in more than one overlapping form, the question naturally arises as to whether, for a given data set, any such form is better than the others in a statistical or numerical sense. More specifically, given an observation record $\{y_t, t = 0,1,\ldots,N\}$ generated by a finite dimensional stochastic system which can be represented by different overlapping structures, we want to know whether one particular structure will yield more efficient parameter estimates than the others, or whether using one particular structure will produce a numerically better conditioned algorithm. This question has been considered by several authors in recent years, but so far no definite answer is available. Different procedures have been proposed that select one out of several candidate parametrizations which is considered best in some ad hoc sense [7]-[9]. The "optimal" parametrization is related in a rather intuitive way to the precision with which the parameters are estimated. In [8] a different approach has been taken: the idea is not to select a best structure, but to change from one parametrization to a better one (by a coordinate transformation) when the parameter estimation algorithm runs into numerical difficulties.

In this paper we first show that, for a n-th order process with a p-dimensional observation vector y_t, the parameters of any overlapping parametrization (either in state-space or in ARMA form) are obtained from a set of $2np$ "intrinsic invariants" which are determined from the Hankel matrix of impulse responses. To any choice of p structure indices, there corresponds, for a given system, a set of $2np$ parameters which completely specify this system. The choice of the p structure indices determines in which particular local coordinate space the system is described. From these $2np$ "intrinsic invariants," a unique state-space or ARMA parametrization can then be derived; these will belong to the set of overlapping parametrizations.

Next we compare the different overlapping parametrizations in terms of asymptotic accuracy of the parameter estimates. We show that, if the determinant of the Fisher information matrix is used as a measure of asymptotic efficiency, then all overlapping parametrizations describing the same process are equivalent, in the sense that they will give the same value to this criterion.

Our result implies that if a process is modelled in state-space or ARMA form using a prediction error method, then the determinant of the covariance matrix of the parameter estimates will asymptotically be the same, whichever overlapping parametrization is used. This does not mean that, over a finite data record, one parametrization might not have a better numerical behavior than others. We proposed two heuristic methods to deal with this finite data problem. Some simulation results are also presented.

2. PARAMETRIZATION OF MULTIVARIATE SYSTEMS. We consider throughout this paper a p-dimensional stationary full rank zero-mean stochastic process $\{y_t\}$ with rational spectrum. Then it is well known that $\{y_t\}$ can be described, up to second order statistics, by the following finite-dimensional representations:

STATE-SPACE REPRESENTATION

$$x_{t+1} = Fx_t + Ke_t$$

$$y_t = Hx_t + e_t \tag{2.1}$$

where the state x_t is an n-dimensional vector; F, K and H are matrices of dimensions $n \times n$, $n \times p$, and $p \times n$, F has all its eigenvalues strictly inside the unit circle, and $\{e_t\}$ is a p-dimensional white noise sequence with covariance matrix Q.

INPUT-OUTPUT REPRESENTATION (ARMA Model)

$$y_t + A_1 y_{t-1} + \ldots + A_r y_{t-r} = e_t + B_1 e_{t-1} + \ldots + B_s e_{t-s} \tag{2.2a}$$

where $A_1,\ldots,A_r$, $B_1,\ldots,B_s$ are $p \times p$ matrices and $\{e_t\}$ is as before. This representation is equivalent with the following:

$$A(z)y_t = B(z)e_t \tag{2.2b}$$

where $A(z)$ and $B(z)$ are square polynomial matrices in the variable z (z is the advance operator: $zy_t = y_{t+1}$), with $\det A(z) \neq 0$ for $|z| \geq 1$, and $\lim_{z\to\infty} A^{-1}(z)\ (z) = I$.

Without any loss of generality, we can make the following assumptions regarding these two representations.

ASSUMPTION 1a: *The matrix triple* (H,F,K) *is of minimal order* n, *where* n *is the dimension of the state vector* x_t, *i.e.*

$$\text{rank} \begin{bmatrix} H \\ HF \\ . \\ . \\ HF^{n-1} \end{bmatrix} = n, \quad \text{rank } [K \; FK \; .... \; F^{n-1} \; K] = n \tag{2.3}$$

n *is then called the order of the process* $\{y_t\}$.

ASSUMPTION 1b: *The polynomial matrices* $A(z)$ *and* $B(z)$ *are left coprime. It can then be shown that* deg det $A(z) = n$, *the order of the process.*

DEFINITION 1a: *The set of all minimal triples* (H,F,K) *of order* n *will be denoted by* S_n.

DEFINITION 1b: *The set of all left coprime polynomial pairs* $(A(z),B(z))$ *with* deg det $A(z) = n$ *will be denoted by* S_n^*.

Eliminating x_t in (2.1) or premultiplying (2.2) by $A^{-1}(z)$ leads to a third representation for the process $\{y_t\}$:

$$y_t = \sum_{i=1}^{\infty} H_i \, e_{t-1} = \underline{H} \, E^t \tag{2.4}$$

where the $p \times p$ matrices H_i are called impulse response matrices (or Markov parameters). The infinite matrix $\underline{H}$ is defined as $\underline{H} = [H_0 H_1 H_2 \; ...]$ with $H_0 = I_p$ and $H(z) = \sum_{i=1}^{\infty} H_i z^i$ is analytic in $|z| \leq 1$. The infinite column vector E^t is defined as

$$E^t = \begin{bmatrix} e_t \\ e_{t-1} \\ e_{t-2} \\ \vdots \end{bmatrix}$$

The impulse response matrices are related to the representations (2.1) and (2.2) as follows:

$$H_0 = I, \quad H_i = H \, F^{i-1} \, K, \quad i = 1,2,... \tag{2.5a}$$

$$\sum_{i=0}^{\infty} H_i z^{-i} = A^{-1}(z)B(z) \tag{2.5b}$$

The impulse response representation (2.4) completely specifies the second-order statistics of the process $\{y_t\}$, namely the covariance function $R_y(k) = E\{y_t y_{t-k}\}$, $k = 0,1,...$ However there are several pairs $\{\underline{H},E^t\}$

which generate the same process $\{y_t\}$ (more specifically, a process $\{y_t\}$ with the same covariance kernel $R_y(k)$). To overcome this difficulty we require $\underline{H}$ to be not only causal, but also causally invertible, such that $\{e_t\}$ can be reconstructed from the past $\{y_t\}$ through a linear functional:

$$e_t = \sum_{i=0}^{\infty} G_i y_{t-i} = \underline{G} \, Y^t \quad , \tag{2.6}$$

with $G_0 = I_p$ and $G(z) \triangleq \sum_{i=0}^{\infty} G_i z^i$ is analytic in $|z| < 1$. $\underline{G}$ and Y^t are defined as $\underline{H}$ and E^t above.

Then $\{e_t\}$ is called the innovation of $\{y_t\}$ and

$$e_t \triangleq y_t - \hat{y}_{t/t-1} \tag{2.7}$$

where $\hat{y}_{t/t-1}$ is the linear least squares predictor of y_t given the past history Y^{t-1} of $\{y_t\}$.

We can now define identifiability up to second order statistics. Let θ be the vector of parameters in either the triple (H,F,K) or the pair $(A(z),B(z))$ and let Q be the covariance matrix of the white noise sequence $\{e_t\}$ in either (2.1) or (2.2).

DEFINITION 2: 2 *parameter pairs* (θ_1,Q_1) *and* (θ_2,Q_2) *are undistinguishable if and only if*

$$R_y(k;\theta_1,Q_1) = R_y(k;\theta_2,Q_2) \qquad \forall k \geq 0 \tag{2.8}$$

where $R_y(k;\theta_i,Q_i)$ *is the covariance function of the process* $\{y_t\}$ *generated by model* i.

Note that if $\{y_t\}$ is Gaussian (or if a second order identification method is used such as a prediction error method), then the probability law (or the loss function) is completely determined by the second order moments, and Definition 2 can be replaced by the following.

DEFINITION 2': *For a Gaussian process 2 parameter pairs* (θ_1,Q_1) *and* (θ_2,Q_2) *are undistinguishable iff*

$$p(Y_0^T;\theta_1,Q_1) = p(Y_0^T;\theta_2,Q_2) \quad \forall Y_0^T \text{ and } \forall T > 0 \quad . \tag{2.9}$$

Now it is easy to show that (θ_1,Q_1) and (θ_2,Q_2) are indistinguishable iff

$$Q_1 = Q_2 \text{ and } H_i(\theta_1) = H_i(\theta_2), \quad i = 0,1,2,... \tag{2.10}$$

Because $Q_1 = Q_2$, we shall in the sequel drop the explicit dependence of $R_y(k)$ or $p(Y_0^T)$ on Q. The indistinguishability concept induces an equivalence relation on the sets S_n and S_n^*, which we shall denote by the symbol $\sim$. It follows from (2.5) and (2.10) that

$$\theta_1 \sim \theta_2 \leftrightarrow H_i(\theta_1) = H_i(\theta_2) \quad \forall i \tag{2.11}$$

$$\leftrightarrow H_2 = H_1 T, \quad F_2 = T^{-1}F_1 T, \quad K_2 = T^{-1}K_1$$

for some nonsingular matrix T. (2.12)

$$\leftrightarrow A_2(z) = M(z)A_1(z), \quad B_2(z) = M(z)B_1(z) \tag{2.13}$$

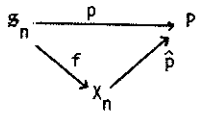
for some unimodular matrix $M(z)$.* (2.13)

Two matrix triples (H_1,F_1,K_1) and (H_2,F_2,K_2) (resp. two polynomial matrix pairs $(A_1(z),B_1(z))$ and $(A_2(z),B_2(z))$ are called *equivalent* if the relations (2.12) (resp. (2.13)) hold.

The covariance function (or the probability law in the Gaussian case) of the process $\{y_t\}$ is completely determined by specifying (H,F,K,Q) or $(A(z),B(z),Q)$. But because of the nonuniqueness induced by (2.11)-(2.12), in order to achieve identifiability, we have to find a *reparametrization* of the family $R_y(k;\theta)$ or $P(Y_0^T;\theta)$ in such a way that two different sets of parameters (in the reparametrized set) correspond to two different sequences of Markov parameters.

From now on we shall, for simplicity, assume that the process $\{y_t\}$ is Gaussian; identifiability is then defined by Definition 2'. All statements hold, up to second order statistics, for non-Gaussian processes if $p(Y_0^T;\theta)$ is replaced by $R_y(k;\theta)$.

What is needed to achieve identifiability is a factorization of the map $p : \theta \to (\cdot;\theta)$ in the following way:



(2.14)

Here $\mathcal{S}_n$ is either S_n or S_n^* (see Definitions 1 above); p is the map defined by the probability law; P is the image of p. The set X_n and the functions $f : \mathcal{S}_n \to X_n$ and $\hat{p} : X_n \to P$ must satisfy the following conditions:

*A polynomial matrix $M(z)$ is unimodular if it is square and if $\det M(z) =$ non-zero constant.

a) for each $\theta \in \mathcal{S}_n$, $\xi = f(\theta)$ is finite-dimensional (2.15a)

b) $p(\cdot;\theta) = \hat{p}(\cdot;f(\theta))$ for all $\theta \in \mathcal{S}_n$ (2.15b)

c) $\hat{p}(\cdot;\xi_1) = p(\cdot;\xi_2) \Rightarrow \xi_1 = \xi_2$. (2.15c)

The function f consists of a finite number of scalar components, say $f_1,\ldots,f_k$, which form a complete system of invariants (see [5]) for the equivalence relation (2.11), since by b) and c):

$$\theta_1 \sim \theta_2 \Leftrightarrow \hat{p}(\cdot;f(\theta_1)) = \hat{p}(\cdot;f(\theta_2)) \Leftrightarrow f(\theta_1) = f(\theta_2) \quad . \tag{2.16}$$

The set X_n can be identified with the quotient sets $S_n/\sim$ or $S_n^*/\sim$, or, equivalently, with the class of all impulse response sequences $\underline{H}$ admitting a minimal realization of order n. Now it can be shown that, when $p > 1$, no single parametrization is able to describe all n-th order systems. Rather X_n can be described by a cover of local coordinates, or equivalently by a family of $\binom{n-1}{p-1}$ overlapping parametrizations of dimension $2np$. Each set of $2np$ invariants (i.e. each local parametrization) is defined by specifying p integer valued numbers $n_1,\ldots,n_p$ called "structure indices":

$$f_{n_1,\ldots,n_p} : \mathcal{S}_n^{(n_1,\ldots,n_p)} \to X_n, \quad \text{with} \quad S_n \subset \mathbb{R}^{2np} \tag{2.17}$$

where $\mathcal{S}_n(n_1,\ldots,n_p)$ is a subset of $\mathcal{S}_n$.

Each map (2.17) is locally a complete system of surjective invariants of dimension $2np$; the subsets $\mathcal{S}_n(n_1,\ldots,n_p)$ for all possible choices of $n_1,\ldots,n_p$ overlap, and cover $\mathcal{S}_n$.

In the next section we shall define the structure indices and show how a choice of structure indices $n_1,\ldots,n_p$ defines a set of $2np$ invariants $f_{n_1,\ldots,n_p}$. These $2np$ invariants are computable functions of the impulse response matrices $H_0,H_1,H_2,\ldots$ and will be called intrinsic invariants. We shall then show how to construct overlapping state-space or ARMA parametrizations as a function of the $2np$ intrinsic invariants.

3. CONSTRUCTION OF A COMPLETE SYSTEM OF INVARIANTS. From (2.4) and (2.7) we can write the linear least squares k-step ahead predictor $\hat{y}_{t+k|t}$ as follows:

$$\hat{y}_{t+k|t} = \sum_{i=k}^{\infty} H_i e_{t-i} \qquad k = 0,1,2,\ldots \quad . \tag{3.1}$$

Therefore

$$\hat{Y}_t \triangleq \begin{bmatrix} \hat{y}_{t+1|t} \\ \hat{y}_{t+2|t} \\ \hat{y}_{t+3|t} \\ \vdots \end{bmatrix} = \begin{bmatrix} H_1 & H_2 & H_3 & \cdots \\ H_2 & H_3 & H_4 & \cdots \\ H_3 & H_4 & H_5 & \cdots \\ \vdots & & & \end{bmatrix} \begin{bmatrix} e_t \\ e_{t-1} \\ e_{t-2} \\ \vdots \end{bmatrix} = \mathbf{H} \mathcal{E}^t \qquad (3.2)$$

Since the process is of order n, the rank of the Hankel matrix $\mathbf{H}$ is n. Therefore we can choose a set of linearly independent rows of $\mathbf{H}$, which will form a basis for the whole row space of $\mathbf{H}$. Note that the corresponding components of $\hat{Y}_t$ will then form a basis for the space spanned by all components of the prediction vector $\hat{Y}_t$. Now the structure indices will define which rows of $\mathbf{H}$ will form the basis, and we shall show how to construct a set of $2np$ invariants from $\mathbf{H}$ for a given choice of structure indices.

To any choice of n linearly independent rows of $\mathbf{H}$ we shall associate a *multiindex* $\underline{i} = (i_1,\ldots,i_n)$ where the numbers $i_1,\ldots,i_n$, arranged in increasing order, are the indices of the rows of $\mathbf{H}$ that form the basis. Two restrictive conditions will be imposed on the selection of the basis rows:

CONDITION 1: *if* $j \in \underline{i}$, *then* $j-p \in \underline{i}$

CONDITION 2: $1,2,\ldots,p \in \underline{i}$

Condition 1 follows from the structure of the Hankel matrix: if the $(j-p)$-th row of $\mathbf{H}$ is in the span of the preceding rows, so is the j-th row. Condition 2 results from the full rank assumption on $\{y_t\}$: it follows that the p components of $\hat{y}_{t+1|t}$ are linearly independent.

DEFINITION 3: *If the selection of the basis vectors obeys conditions* 1 *and* 2, *then the corresponding multiindex is called "nice."*

All nice multiindices correspond to a choice of the basis inside the first $n-p+1$ row blocks of $\mathbf{H}$. For given n and p, there are only a finite number of possible nice multiindices (Example: for a 2-dimensional vector process $(p=2)$ or order 3 $(n=3)$, there are only 2 possible nice multiindices: $\underline{i}_1 = (1,2,3)$ and $\underline{i}_2 = (1,2,4)$). For most n-th order processes of dimension p, several possible choices exist for the selection of a "nice multiindex" basis. On the other hand, there are subsets of $\mathcal{S}_n$ for which only one basis exists.

DEFINITION OF THE STRUCTURE INDICES: *Let* $\underline{i} = (i_1,\ldots,i_n)$ *be a nice multiindex defining a basis for the rows of* $\mathbf{H}$. *For* $k=1,\ldots,p$, *let* n_k *be the least natural number such that* $(k+n_k p) \notin \underline{i}$. *Then* $n_1,\ldots,n_p$ *are called the "structure indices" corresponding to the basis; they specify which rows of* $\mathbf{H}$ *are taken in the basis. Note that* $\sum_{i=1}^{p} n_i = n$.

We can now define $\mathcal{S}_n(n_1,\ldots,n_p)$ as the set of all n-th order systems for which the n rows specified by $n_1,\ldots,n_p$ in the Hankel matrix are linearly independent. $\mathcal{S}_n(n_1,\ldots,n_p)$ is a proper subset of $\mathcal{S}_n$.

Consider now an element of $\mathcal{S}_n(n_1,\ldots,n_p)$ specified by its Hankel matrix $\mathbf{H}$. We shall construct a complete system of $2np$ surjective invariants for this system, i.e. a reparametrization of this system using $2np$ parameters.

Let H^i be the i-th block of p rows of the infinite Hankel matrix $\mathbf{H}$ (e.g., $H^2 = [H_2 H_3 H_4 \ldots]$ and let

$$H^i = \begin{bmatrix} h_{1i} \\ h_{2i} \\ \vdots \\ h_{pi} \end{bmatrix} \qquad (3.3)$$

where h_{ki} are rows of infinite length. Since $\underline{H}$ is an element of $\mathcal{S}_n(n_1,\ldots,n_p)$, the rows $(h_{11},\ldots,h_{1n_1}; h_{21},\ldots,h_{2n_2}; \ldots; h_{p1},\ldots,h_{pn_p})$ form a basis. Therefore the rows $h_{1(n_1+1)}, h_{2(n_2+1)},\ldots, h_{p(n_p+1)}$ can be expressed as:

$$h_{i(n_{i+1})} = \sum_{j=1}^{p} \sum_{k=1}^{n_j} \alpha_{ijk} h_{jk} \qquad i = 1,\ldots,p \qquad (3.4)$$

These relations define np scalar numbers α_{ijk}.

Now denote by $h_{ij}(k)$ the element in row i, column j of H_k. Then the $2np$ numbers

$$\{\alpha_{ijk},\ k=1,\ldots,n_j;\ h_{ij}(k),\ k=1,\ldots,n_i;\ i,\ j=1,\ldots,p\} \qquad (3.5)$$

completely coordinatize $\mathcal{S}_n(n_1,\ldots,n_p)$, i.e. they map that set in a one to one manner, on Euclidean space of dimension $2np$. The impulse response sequence $H_1, H_2, H_3,\ldots$ is completely specified by the p structure indices and these $2np$ numbers. These $2np$ numbers constitute a complete system of surjective invariants, which will be called "intrinsic invariants" of the process. In the notation of Section II:

$$f_{n_1,\ldots,n_p} : \mathcal{S}_n(n_1,\ldots,n_p) \to \mathbb{R}^{2np}$$

$$: \{H_1,H_2,H_3,\ldots\} \to f_{n_1,\ldots,n_p} = \{\alpha_{ijk}, h_{ij}(k)\} \qquad (3.6)$$

The word intrinsic is used because the invariants are defined from the (intrinsic) infinite impulse response representation, and not from a finite state-space or ARMA model.

We have thus constructed a family of cuntions $f_{n_1,\ldots,n_p}$, each of which is a complete system of surjective invariants mapping onto $\mathbb{R}^{2np}$. These functions are defined on the overlapping subsets $\mathcal{S}_n(n_1,\ldots,n_p)$ which cover $\mathcal{S}_n$. Next we show that one can construct corresponding overlapping (state-space or ARMA) parametrizations whose parameters are functions of the 2np intrinsic invariants just defined.

STATE-SPACE PARAMETRIZATION

Consider an element of $\mathcal{S}_n(n_1,\ldots,n_p)$ for a given set of structure indices, and let $\{\alpha_{ijk}, h_{ij}(k)\}$ be the instrinsic invariants of that element. Then the following is a state-space representation of that element:

$$H = \left[\begin{array}{cccc|cccc|c|cccc} 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & & 0 & 0 & \ldots & 0 \\ 0 & . & & . & 1 & 0 & \ldots & 0 & & . & . & & . \\ . & . & & . & 0 & . & & . & \cdots & . & . & & . \\ . & . & & . & . & . & & . & & . & . & & . \\ . & . & & . & . & . & & . & & 0 & 0 & & . \\ 0 & 0 & \ldots & 0 & 0 & . & \ldots & 0 & & 1 & 0 & \ldots & 0 \end{array}\right] \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \} P \qquad (3.7a)$$

$$\underbrace{\qquad\qquad}_{n_1} \qquad \underbrace{\qquad\qquad}_{n_2} \qquad\qquad \underbrace{\qquad\qquad}_{n_p}$$

$$F = \left[\begin{array}{ccc|ccc|c|ccc} 0 & & & 0 & \ldots & 0 & & 0 & & 0 \\ . & & & . & & . & & . & & . \\ . & & I_{n_1-1} & . & & . & \cdots & . & & . \\ . & & & . & & . & & . & & . \\ 0 & & & 0 & \ldots & 0 & & 0 & & 0 \\ \alpha_{111} & \cdots & \alpha_{11n_1} & \alpha_{121} & \cdots & \alpha_{12n_2} & & \alpha_{1p1} & \cdots & \alpha_{1pn_p} \\ \hline & . & & & . & & & & . & \\ & . & & & . & & & & . & \\ & . & & & . & & & & . & \\ \hline 0 & \ldots & 0 & 0 & \ldots & 0 & & 0 & & \\ . & & . & . & & . & & . & & \\ . & & . & . & & . & \cdots & . & & I_{n_p-1} \\ 0 & \ldots & 0 & 0 & \ldots & 0 & & 0 & & \\ \alpha_{p11} & \cdots & \alpha_{p1n_1} & \alpha_{p21} & \cdots & \alpha_{p2n_2} & & \alpha_{pp1} & \cdots & \alpha_{ppn_p} \end{array}\right] \qquad (3.7b)$$

$$K = \begin{bmatrix} K_1 \\ K_2 \\ . \\ . \\ . \\ K_p \end{bmatrix}, \quad \text{with} \quad K_i = \begin{bmatrix} k_{i1} \\ . \\ . \\ . \\ . \\ k_{in_i} \end{bmatrix}, \quad \text{and} \quad k_{ij} = [h_{i1}(j),\ldots,h_{ip}(j)], \quad \text{a row p-vector} \qquad (3.7c)$$

The proof is a straightforward but tedious verification that the relations (2.5a) hold with the α_{ijk} and $h_{ij}(k)$ defined from the H_i as in (3.4)-(3.5)

The process $\{y_t\}$, which is by assumption an element of $\mathcal{S}_n(n_1,\ldots,n_p)$, can be represented by (2.1) with H, F, K as in (3.7).

ARMA PARAMETRIZATION

By an argument similar to that developed in [5] and [13] we obtain the following equations for the entries A(z) and B(z) of (2.2b):

$$a_{ii}(z) = z^{n_i} - \alpha_{iin_i} z^{n_i-1} - \ldots - \alpha_{ii1} \qquad (3.12a)$$

$$a_{ij}(z) = -\alpha_{ijn_j} z^{n_j-1} - \ldots - \alpha_{ij1} \qquad (3.12b)$$

and

$$b_{ij}(z) = b_{ij\bar{n}+1} z^{n} + b_{ij\bar{n}}\, z^{\bar{n}-1} + \ldots + b_{ij1} \tag{3.12c}$$

with $\bar{n} = \max_{1\leq i\leq p} n_i$.

The α_{ijk} are defined by (3.4), while the b_{ijk} are defined as follows

$$\bar{B} = M\bar{R} \tag{3.13}$$

where

$$\underset{(\bar{n}+1)p\times p}{\bar{B}} = \begin{bmatrix} B^1 \\ \vdots \\ B^p \end{bmatrix}, \qquad \underset{(n+1)\times p}{B^i} = \begin{bmatrix} b_{i11} & & b_{ip1} \\ \vdots & & \\ b_{i1(n+1)} & \cdots & b_{ip(n+1)} \end{bmatrix} \tag{3.14}$$

$$\underset{(n+p)\times p}{\bar{K}} = \begin{bmatrix} (1 & 0 & \ldots & 0) \\ k_{11} & & & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ k_{1n_1} & & & \\ (0 & 1 & 0 \ldots & 0) \\ k_{21} & & & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ k_{pn_p} & & & \end{bmatrix} \tag{3.15}$$

with k_{ij} defined by (3.7c)

$$M = [M_{ij}] \qquad (i,j = 1,\ldots,p)$$

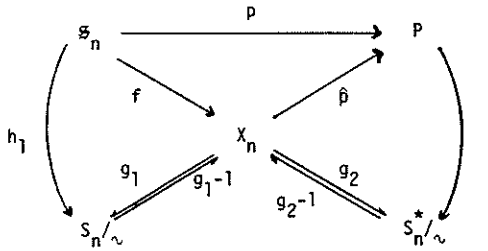
with

$$\underset{(n+1)\times(n_i+1)}{M_{ii}} = \begin{bmatrix} -\alpha_{ii1} & \cdots & -\alpha_{iin_i} & 1 \\ & & & \\ & & & \\ & & & \\ -\alpha_{iin_i} & & & \\ 1 & \cdots & & \\ 0 & \cdots & & 0 \\ 0 & \cdots & & 0 \end{bmatrix} \quad \begin{matrix} \\ \\ \\ \\ \\ \\ \bar{n}-n_i \\ \end{matrix} \tag{3.16}$$

and

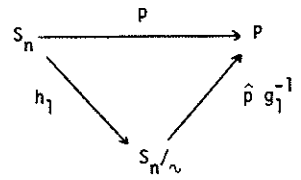
$$\underset{(n+1)\times(n_j+1)}{M_{ij}} = \begin{bmatrix} -\alpha_{ij1} & \cdots & -\alpha_{ijn_j} & 0 \\ & & & \\ & & & \\ & & & \\ -\alpha_{ijn_j} & & & \\ 0 & \cdots & & \\ 0 & \cdots & & 0 \\ 0 & \cdots & & 0 \end{bmatrix} \quad \begin{matrix} \\ \\ \\ \\ \\ \\ \bar{n}-n_j \\ \end{matrix} \tag{3.17}$$

We can now summarize the main points of this section by extending the scheme (2.14). For every element of $\mathcal{S}_n(n_1,\ldots,n_p)$, we have first defined the intrisic invariants $f_{n_1,\ldots,n_p}$. Next, by (3.7), (3.12) and (3.13), we have established two bijections $g_1(n_1,\ldots,n_p)$ and $g_2(n_1,\ldots,n_p)$ between the intrinsic invariants (defined on $\underline{H}$) and the quotient spaces $S_n/_\sim$ and $S_n^*/_\sim$

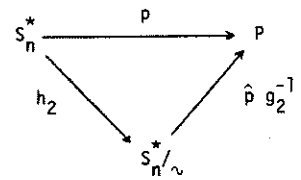


(3.18)

Hence, by a property of complete sets of surjective invariants (see, e.g., [5]), $h_i(n_1,\ldots,n_p) = g_i(n_1,\ldots,n_p) \cdot f_{n_1\ldots n_p}$ are also locally complete sets of surjective invariants. This leads to the following factorizations:



(3.19a)



(3.19b)

In figs. (3.18) - (3.19) all quantities are to be indexed by $(n_1,\ldots,n_p)$. These factorizations of the probability map make the multivariable models identifiable. Given an arbitrary element of $S_n(n_1,\ldots,n_p)$ (i.e. a matrix triple H, F, K for which the rows of $\mathcal{H}$ specified by the structure indices $n_1,\ldots,n_p$ form a basis) which is parametrized by n^2+2np parameters that are not identifiable, we replace this element by an equivalent element of the quotient space $S_n/_\sim$ via the map $h_1(n_1,\ldots,n_p)$ (see Fig. 3.19a). This last element is parametrized by the system of $2np$ invariants defined by h_1 and appearing in the form (3.7). These $2np$ invariants are uniquely identifiable. The same can be said if an ARMA form is used (see Fig. 3.19b).

4. ASYMPTOTIC EQUIVALENCE OF ALL OVERLAPPING FORMS. In most cases an n-th order system can be represented in more than one of the overlapping forms. This corresponds to the fact that in general different choices of nice multi-indices can be made, which will define different sets of linearly independent rows of the Hankel matrix.

EXAMPLE. *Consider a bivariate process* $(p=2)$ *or order* 3 $(n=3)$. *Two nice multiindices exist, with the corresponding sets of structure indices:*

$$i_1 = (1,2,3), \quad \text{i.e.} \quad n_1 = 2, \quad n_2 = 1$$

$$i_2 = (1,2,4), \quad \text{i.e.} \quad n_1 = 1, \quad n_2 = 2 \quad .$$

The corresponding sets of intrinsic invariants are:

- for $i_1 : \alpha_{111},\alpha_{112},\alpha_{121},\alpha_{211},\alpha_{212},\alpha_{221}$

$$h_{11}(1),\ h_{12}(1),\ h_{21}(1),\ h_{22}(1),\ h_{11}(2),\ h_{12}(2).$$

- for $i_2 : \alpha_{111},\alpha_{121},\alpha_{122},\alpha_{211},\alpha_{221},\alpha_{222},$

$$h_{11}(1),\ h_{12}(1),\ h_{21}(1),\ h_{22}(1),\ h_{21}(2),\ h_{22}(2).$$

The corresponding state-space parameterizations are:

$$H_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad F_1 = \begin{bmatrix} 0 & 1 & 0 \\ \alpha_{111} & \alpha_{112} & \alpha_{121} \\ \alpha_{211} & \alpha_{212} & \alpha_{221} \end{bmatrix} \quad K_1 = \begin{bmatrix} h_{11}(1) & h_{12}(1) \\ h_{11}(2) & h_{12}(2) \\ h_{21}(1) & h_{22}(1) \end{bmatrix}$$

$$H_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad F_2 = \begin{bmatrix} \alpha_{111} & \alpha_{121} & \alpha_{122} \\ 0 & 0 & 1 \\ \alpha_{211} & \alpha_{221} & \alpha_{222} \end{bmatrix} \quad K_2 = \begin{bmatrix} h_{11}(1) & h_{12}(1) \\ h_{21}(1) & h_{22}(1) \\ h_{21}(2) & h_{22}(2) \end{bmatrix}$$

Now it can be shown (see e.g. [7]-[9]) that the state x_t of the state-space realization (2.1), with H, F, K defined by (3.7), is made up of the components $\hat{y}_{t-1}$ indexed by the elements of the selected nice multiindex. In our example, the state is defined in one of two possible ways according to which choice of structure indices is made:

for $n_1 = 2, \ n_2 = 1$: $i_1 = (1,2,3)$

$$x_t \triangleq \begin{bmatrix} \hat{y}^1_{t/t-1} \\ \hat{y}^1_{t+1/t-1} \\ \hat{y}^2_{t/t-1} \end{bmatrix}$$

for $n_1 = 1, \ n_2 = 2$: $i_2 = (1,2,4)$

$$x_t \triangleq \begin{bmatrix} \hat{y}^1_{t/t-1} \\ \hat{y}^2_{t/t-1} \\ \hat{y}^2_{t+1/t-1} \end{bmatrix}$$

The structure of the corresponding H, F, K matrices is shown above. In most cases, both choices are possible, because both sets of 3 components form a linearly independent set. The question then is whether one choice is better than the other. For example, one might think that if $\hat{y}^1_{t+1/t-1}$ is close to the linear span of $\hat{y}^1_{t/t-1}$ and $\hat{y}^2_{t/t-1}$, then the choice of i_2 would be preferable because the components of the state would be more orthogonal to one another, thereby making the ensuing parameter estimation problem numerically better behaved. More generally the question that has intrigued several researchers in recent years is whether any one of the overlapping parametrizations is optimal in the sense that, in that particular parametrization, the parameter estimates can be estimated with a higher accuracy or that the

estimation algorithm will have a better numerical behaviour. In the main result of this paper we present a partial answer to this question.

THEOREM. *Given the intrinsic invariants* $\{\alpha_{ijk},h_{ij}(k)\}$ and $\{\alpha^*_{ijk},h^*_{ij}(k)\}$ *corresponding to two different sets of structure indices* $\{n_1,\ldots,n_p\}$ *and* $\{n^*_1,\ldots,n^*_p\}$, *then the determinants of the information matrices corresponding to these two parametrizations are identical.*

The proof of the Theorem follows from the following lemma.

LEMMA. *Let* $\{\alpha_{ijk},h_{ij}(k)\}$ *and* $\{\alpha^*_{ijk},h^*_{ij}(k)\}$ *be the intrinsic invariants of a given n-th order system in* $\mathcal{S}_n$ *for two different choices of the structure indices. Then the Jacobian of the transformation between these two parameter vectors is unity.*

The proof of the lemma can be found in [9]. The theorem can then be proved as follows. Let $\theta = \{\alpha_{ijk},h_{ij}(k)\}$ and $\theta^* = \{\alpha^*_{ijk},h^*_{ij}(k)\}$. The corresponding information matrices M_θ and M^*_θ are related by†

$$M_{\theta^*} = E_{Y|\theta^*}\left\{\left(\frac{\partial \log p(Y|\theta^*)}{\partial\theta^*}\right)^T\left(\frac{\partial \log p(Y|\theta^*)}{\partial\theta^*}\right)\right\}$$

$$= E_{Y|\theta^*}\left\{\left(\frac{\partial \log p(Y|\theta)}{\partial\theta}\,\frac{\partial\theta}{\partial\theta^*}\right)^T\left(\frac{\partial \log p(y|\theta)}{\partial\theta}\,\frac{\partial\theta}{\partial\theta^*}\right)\right\}$$

$$= \left(\frac{\partial\theta}{\partial\theta^*}\right)^T E_{Y|\theta}\left\{\left(\frac{\partial \log p(Y|\theta)}{\partial\theta}\right)^T\left(\frac{\partial \log p(Y|\theta)}{\partial\theta}\right)\right\}\left(\frac{\partial\theta}{\partial\theta^*}\right)$$

$$= \left(\frac{\partial\theta}{\partial\theta^*}\right)^T M_\theta\left(\frac{\partial\theta}{\partial\theta^*}\right) \quad .$$

It follows from the bijective relationship between the intrinsic invariants $\{\alpha_{ijk},h_{ij}(k)\}$ and the corresponding overlapping parametrizations H, F, K or A(z), C(z) that the theorem also holds when two overlapping (state-space or ARMA) parametrizations are compared.

COROLLARY. *Given two overlapping parametrizations* F, K, H *and* F^*, K^*, H^* *in the form* (3.7) *(resp.* A(z), B(z) *and* $A^*(z)$, $B^*(z)$ *in the form* (3.12)) *for the same process, corresponding to two different sets of structure indices* $\{n_1,\ldots,n_p\}$ *and* $\{n^*_1,\ldots,n^*_p\}$, *then the determinants of the information matrices corresponding to these two parametrizations are identical.*

If the parameters are estimated using a maximum likelihood or a prediction error method, then the covariance matrix of the estimation errors is asymptotically equal to the inverse of the Fisher information matrix M_θ. Therefore it follows from our main result that all overlapping parametrizations are asymptotically equivalent, as far as the accuracy of the parameter estimates is concerned, when this accuracy is measured by the determinant of the covariance matrix of the estimation errors. As a consequence this criterion is unable to discriminate between two overlapping structures for the same process. Of course other criteria would be used that might be able to discriminate, even asymptotically, between different structures, see e.g. [10].

Some structures might also be better than others when only a finite data record is available. In the next two sections we describe two heuristic basis selection procedures that can be used for the finite data case. They are based on the idea of selecting the "most independent components" in the state vector.

5. A METHOD BASED ON THE CONCEPT OF COMPLEXITY. In this section, we sketch a method proposed by Ljung and Rissanen [7] and based on the concept of complexity of a random vector, defined by Van Emden [14]. We shall also propose a new iterative procedure that is closely related to that of Ljung and Rissanen.

Complexity is in fact a measure of the interaction between the components of a random vector. The more interaction there is, the larger the complexity. Van Emden shows that the complexity can be expressed using the covariance matrix and derives the following expression:

$$C = -\frac{1}{2}\sum_{i=1}^{n}\log(n\lambda_i) \qquad (5.1)$$

where λ_i are the eigenvalues of the covariance matrix of the random vector (provided this covariance matrix has been normalized so that its trace equals 1).

Suppose now that we know the covariance matrix $R_{\hat{Y}}$ of $\hat{Y}^N_t$ and that the order of the process, n, is also known. Then one can compute the complexity of various subvectors of order n of $\hat{Y}^N_t$, because the corresponding $n\times n$ covariance matrices are submatrices of $R_{\hat{Y}}$. The idea proposed in [7] is then to select as the state that subvector of $\hat{Y}^N_t$ that has the smallest complexith among all subvectors of dimension n that obey the conditions 1 and 2 of Section 3. The components obtained in this way are called by Ljung and Rissanen the "most independent components" of Y^N_t. The procedure they suggest is as follows:

†If x is a scalar and θ a column k-vector, then $\frac{\partial x}{\partial\theta}$ denotes the row vector $\left[\frac{\partial x}{\partial\theta_1},\ldots,\frac{\partial x}{\partial\theta_k}\right]$.

(i) compute estimates of the predictors $\hat{y}_{t+k|t}$ by first fitting a high-order autoregressive model to the data.

(ii) compute the sample covariance matrix $R_{\hat{Y}}$ from the estimate predictions.

(iii) for a given value of n, compute the complexity of various submatrices of $R_{\hat{Y}}$, subject to the constraints that the $p \times p$ upper left submatrix of $R_{\hat{Y}}$ is always included and that the j-th row of the matrix $R_{\hat{Y}}$ is chosen only if the (j-p)-th row is also chosen. (These constraints amount to meeting conditions 1 and 2 of Section 3).

(iv) select the basis for the predictor space that corresponds to the submatrix with smallest complexity.

(v) repeat the procedure for higher order models and take the order that minimizes a criterion such as Akaike's AIC criterion.

A major disadvantage of this method is that one needs to first estimate the covariance matrix of the prediction vector by fitting a high-order autoregressive model to the data and then computing sample predictions. However the procedure has the following interesting feature: if the parameters of the F matrix are estimated by least squares, one can show (see [15]) that the covariance matrix of the error of the parameter estimates is related to the inverse of the submatrix of $R_{\hat{Y}}$ selected by the procedure of Ljung and Rissanen. This seems to justify using the matrix $R_{\hat{Y}}$ as a starting point for the selection of the basis components, even though this matrix is not directly available. It also suggests minimizing some scalar measure of the inverse of the various submatrices of $R_{\hat{Y}}$ in order to disciminate between the corresponding subvectors of $\hat{Y}^N_t$. With this idea in mind, we suggest the following procedure.

The first two steps are identical to those in Ljung and Rissanen's method.

(iii) compute the inverse of the upper left $p \times p$ submatrix of $R_{\hat{Y}}$.

(iv) for an order n equal to $p+1$, select all the $(p+1) \times (p+1)$ submatrices of $R_{\hat{Y}}$, which contain the $p \times p$ upper left submatrix, and such that condition 1 of Section 3 is also satisfied. Compute the inverses of these submatrices, using the fact that the inverse of a matrix $A^1 = \begin{bmatrix} A & b \\ b^T & \alpha \end{bmatrix}$, where b is a vector and α a scalar, is given by the following inversion formulae:

$$(A^1)^{-1} = \begin{bmatrix} E & g \\ g^T & \epsilon \end{bmatrix}$$

where

$$E = A^{-1} + A^{-1} b \epsilon b^T A^{-1}$$

$$g = -A^{-1} b \epsilon$$

$$\epsilon = [\alpha - b^T A^{-1} b]^{-1} \quad .$$

(Hence, no other matrix inversion is needed once the $p \times p$ submatrix A^{-1} has been computed.)

(v) Select the submatrix for which the trace of the inverse is minimized.

(vi) Repeat the last two steps with all $(n+1) \times (n+1)$ submatrices that contain the selected $n \times n$ submatrix and whose additional row and column is chosen so as to satisfy condition 1 of Section 3.

(vii) A stopping criterion is needed. One criterion that seems to work well in practice is to stop when $J = \frac{1}{n} \operatorname{tr} [(A^1)^{-1}]$ does not decrease anymore. This is justified by the fact that when the order of the submatrices is greater than the actual order of the process, these submatrices will be ill-conditioned. Their inverses will be large and so will be J.

The method described in this Section is an off-line method: a "best basis" is chosen a priori, and the parameters are subsequently estimated in that basis. Van Overbeek and Ljung [8] have proposed an alternative on-line procedure, which is also derived from [7]. In their scheme, the parameter estimation procedure starts with any parametrization that satisfies conditions 1 and 2 of Section 3. During the estimation procedure, the covariance matrix of the state of the representation is monitored; if that matrix becomes ill-conditioned a similarity transformation is applied in such a way that the new coordinate system corresponds to a better conditioned basis.

6. A Q-R FACTORIZATION METHOD. In this section, we present a new method of structure identification, based on the assumption that a "good" structure will be one in which the predictors selected to form the basis of the prediction space are most independent. We will first assume, as has been done by Akaike [3], that we have chosen $M \in \mathbb{N}$ large enough so that $\underline{H}(Y^t_{t-M})$ (i.e. the Hilbert space spanned by the components of $y(t-k)$ for $0 \leq k \leq M$) is close enough to the sapce $\underline{H}(Y^t_{-\infty})$. This allows us to replace the analysis of dependence of the $\hat{y}^j_{t+k|t}$ by the analysis of dependence of the $\hat{y}^j_{t+k|t,t-M}$, which are defined as the projections of y^j_{t+k} onto the space $\underline{H}(Y^t_{t-M})$.

Denote by Y^t_{t-M} and Y^t_{t+N} the vectors

$$Y^t_{t-M} = \begin{bmatrix} y(t) \\ y(t-1) \\ y(t-M) \end{bmatrix} \qquad Y^t_{t+N} = \begin{bmatrix} y(t) \\ y(t+1) \\ y(t+N) \end{bmatrix} \tag{6.1}$$

and by Σ_{11}, Σ_{12} and Σ_{22} the covariance and cross covariance matrices

$$\Sigma_{11} = E\{Y^t_{t+N}(Y^t_{t+N})^T\}\ , \quad \Sigma_{12} = E\{Y^t_{t+N}(Y^t_{t-M})^T\}$$
$$\Sigma_{22} = E\{Y^t_{t-M}(Y^t_{t-M})^T\}\ . \tag{6.2}$$

By the full rank assumption Σ_{22} is positive definite so that we can have

$$\Sigma_{22}^{-T/2}\Sigma_{22}\Sigma_{22}^{-1/2} = I \tag{6.3}$$

for some nonsingular matrix $\Sigma_{22}^{1/2}$.

Let us perform the following transformation:

$$Z^t_{t-M} = \Sigma_{22}^{-T/2}\, Y^t_{t-M} \tag{6.4}$$

so that

$$E\{(Z^t_{t-M})(Z^t_{t-M})^T\} = I \tag{6.5}$$

Now $\hat{y}^j_{t+k|t,t-M}$ admits a representation:

$$\hat{y}^j_{t+k|t,t-M} = \sum_{m=0}^{M}\sum_{\ell=1}^{p} a(j,k,\ell,m) z^{\ell}_{t-m} \tag{6.6}$$

where

$$E\{\hat{y}^j_{t+k|t,t-M}\, z^{\ell}_{t-m}\} = a(j,k,\ell,m) = E\{y^j_{t+k}\, z^{\ell}_{t-m}\}\ . \tag{6.7}$$

Hence, the elements of the (kp+j)-th row of matrix $\Sigma_{12}\Sigma_{22}^{-1/2}$ are the coordinates of $\hat{y}^j_{t+k|t,t-M}$ in the basis Z^t_{t-M}. Therefore, because of (6.5), we can replace the analysis of independence of the components $\hat{y}^j_{t+k|t,t-M}$ by the analysis of independence of the rows of $\Sigma_{12}\Sigma_{22}^{-1/2}$. Now, a well known (and numerically well conditioned) method to search for the most independent rows of a given matrix is to perform a Q-R factorization by means of Householder transformations (see [16] and [17]) with row-interchange.

The method we propose is thus as follows:

(i) compute an upper-triangular square root of Σ_{22}, i.e. an uppertriangular matrix $\Sigma_{22}^{1/2}$ such that $\Sigma_{22}^{T/2}\Sigma_{22}^{1/2} = \Sigma_{22}$ (Cholesky factorization).

(ii) compute the product $\Sigma_{12}\Sigma_{22}^{-1/2}$. Note that the first p rows of this product are lower triangular. This saves p steps in the subsequent QR factorization procedure.

(iii) compute a recursive Q-R factorization of $\Sigma_{12}\Sigma_{22}^{-1/2}$ using Householder transformations, where the triangularization is always performed on the row leading to the largest pivot.

(iv) stop when the pivots do not significantly decrease any more.

We briefly illustrate one step of the Q-R factorization.

After k steps of the triangularization procedure, we have the following factorization:

$$S_k^T\Sigma_{12}\Sigma_{22}^{-1/2}P_k = T_k$$

where S_k is the product of k permutation matrices, P_k is the product of k Householder transformation matrices, and T_k is a lower triangular matrix of the following form:

$$T_k = \left[\begin{array}{ccccc|ccc} t_1 & 0 & & & & 0 & \cdots & 0 \\ X & t_2 & 0 & \cdots & 0 & \vdots & & \vdots \\ \vdots & & & & \vdots & & & \\ & & & & 0 & & & \\ X & \cdots & X & & t_k & 0 & \cdots & 0 \\ \hline & & & & & & & \\ & & T_{21} & & & & T_{22} & \\ & & & & & & & \end{array}\right]$$

It is clear from the structure of T_k that the Euclidean norm of the rows in T_{22} are the distances of the last $(N+1)p-k$ rows of T_k to the space spanned by the first k rows. From the properties of Householder transformations, the pivot of the next triangularization step is the Euclidean norm of the selected row in T_{22}. So choosing the row leading to the largest pivot amounts to choosing the row with the largest projection onto the space orthogonal to the span of the first k rows. This is what we call the "most independent" row.

Upon completion of the triangularization procedure, the following factorization is obtained:

$$\Sigma_{12}\Sigma_{22}^{-1/2} = STP^T \tag{6.8}$$

where S is a permutation matrix, T is a lower triagular matrix (at least for its first n rows) and P is a product of Householder transformation matrices (which implies that $P^TP = PP^T = I$). The permutation matrix S indicates which components of $\hat{Y}^N_t$ are to be chosen in the basis. In order

to be consistent with the conditions 1 and 2 of Section 3, we should also introduce some constraints in the procedure of row selection for the triangularization: the first p rows are to be triangularized, and one can choose the j-th row only if the (j-p)-th row has already been chosen.

We now show that our triangularization method can also be related very nicely to Akaike's method [3] which is based on a canonical correlation analysis on the vectors of future and past observations (see Anderson [18]). We shall not go back in detail to this method, but we recall that it amounts to a singular value decomposition of the matrix $\Sigma_{11}^{-T/2}\Sigma_{12}\Sigma_{22}^{-1/2}$ and that the idea of the canonical correlation analysis in this case is to search for independent linear combinations of the vector Y_{t+N}^t which are most correlated with independent linear combinations of Y_{t-M}^t. Actually, in Akaike's method, this only gives the number of independent components of a sub-vector of the predictor vector, and hence it leads to the choise of the *first* independent components of the prediction vector in the basis. It is not possible with this method to discriminate between various independent components and to take the "most independent" ones, and the reason for this is that the canonical correlation analysis uses linear combinations of Y_{t+N}^t rather than single components of this vector. By searching for the *single* components of Y_{t+N}^t that are most correlated with independent linear combinations of the past, one can establish a close relationship between our procedure and the method of Akaike.

In the canonical correlation analysis, one searches in a first step for two vectors α_1 and γ_1 such that $\alpha_1^T\Sigma_{12}\gamma_1$ is maximized subject to $\alpha_1^T\Sigma_{11}\alpha_1 = \gamma_1^T\Sigma_{22}\gamma_1 = 1$. In a second step, two linear combinations $\alpha_2 Y_{t+N}^t$ and $\gamma_2 Y_{t-M}^t$ are sought, which are orthogonal to the first ones, and which have maximum correlation with one another. Now consider the following variations.

In the first step we maximize $a_1(\Sigma_{12})_{j_1}\gamma_1$, where a_1 is a scalar, $(\Sigma_{12})_{j_1}$ is the j_1-th row of Σ_{12}, γ_1 is a vector of dimension $(M+1)p$, under the constraints $a_1^2 = 1$, $\gamma_1^T\Sigma_{22}\gamma_1 = 1$. The maximum is taken over the values of a_1, γ_1 and the index j_1. For the second step, we maximize $a_2(\Sigma_{12})_{j_2}\gamma_2$, $j_2 \neq j_1$, under the constraints $a_2^2 = 1$, $\gamma_2^T\Sigma_{22}\gamma_2 = 1$, $\gamma_2^T\Sigma_{22}\gamma_1 = 0$, and so on.

With computations similar to those of the canonical correlation analysis, this amounts to finding two matrics A and Γ such that $A^TA = I$ and A is a permutation matrix (with possibly some changes of sign), $\Gamma^T\Sigma_{22}\Gamma = I$ and

$$A^T\Sigma_{12}\Gamma = T \tag{6.9}$$

T is a lower triangular matrix with decreasing pivots. If we compare (6.8) and (6.9), we can identify S with A and $\Sigma_{22}^{-1/2}P$ with Γ, and hence establish the equivalence between the two approaches.

In this section we have proposed a new structure estimation method as an alternative to the method of Rissanen and Ljung. It is based on another heuristic definition of "most independent rows" of a matrix. We do not claim that this new method is superior from a theoretical point of view. However, from a computational point of view, our method has the major advantage that it works with the covariance function of the observation process, R_Y, which can be readily estimated from the data, while the method of Rissanen and Ljung works with the covariance of the predictors, $R_{\hat{Y}}$, which requires that the data be filtered first, using an AR model that has to be identified.

7. SIMULATION RESULTS. In this section, we present some simulation results in order to compare the various methods that we have introduced. Data sequences have been generated from the Markovian model:

$$x_k = Ax_{k-1} + Be_k$$

$$y_k = Cx_k \quad .$$

A member of different models have been simulated but, for reasons of brevity, we restrict ourselves here to two models which will hopefully give a sufficient illustration for the previous sections. In both models, y_k is a two-dimensional vector process, x_k is of dimension 4, and e_k is a two-dimensional Gaussian white noise, with mean zero and unit covariance matrix.

Tables 1 and 2 gives the values of the matrices A, B and C for each of the two models

$$A_1 = \begin{bmatrix} -0.50 & 4.83 & -0.63 & 0.72 \\ 0.20 & -3.67 & 0.50 & -0.58 \\ -0.22 & -2.42 & -0.19 & -0.36 \\ -1.55 & 22.50 & -2.87 & 3.69 \end{bmatrix} \qquad B_1 = \begin{bmatrix} 3.38 & -3.13 \\ -2.5 & 2.5 \\ -0.69 & 1.56 \\ 15.13 & -14.88 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} -70 & 38.0 & 16.0 & 9.0 \\ -4.39 & 85.33 & 37.5 & 17.92 \end{bmatrix}$$

Table 1: Parameter values for model 1

$$A_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -0.1 & 0.65 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -2/3 & -5/3 & -0.25 & 1 \end{bmatrix} \qquad B_2 = \begin{bmatrix} 0 & 2 \\ 0.25 & 0.8 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Table 2: Parameter values for model 2

In the first model, the entries of the matrices have been chosen such that the third row of the Hankel matrix $H_{N,M}$ is nearly in the linear span of the first two rows. Hence the best structural vector of the process is h_1 = (1,2,4,6).

The second model has been taken from a paper by H. El Sherief and N. K. Sinha [19]; using their own method, they find the structural vector h_2 = (1,2,3,4) for this model.

In the sequel, the method of Ljung and Rissanen will be referred to as the LR method, the variant that we have introduced in section 4 will be called WG1, while the QR factorization procedure will be called method WG2.

Order	Structural vector	LR method Complexity	WG1 method trace	WG1 method J	WG2 method selected row	WG2 method |pivot|
3	(1,2,3)	4.37	4.27			
3	(1,2,4)	3.72 ←	4.03 ←	1.33	1	0.408
					2	0.054
4	(1,2,3,4)	7.49	21.52		4	0.110
4	(1,2,4,6)	5.00 ←	4.14 ←	1.04	6	0.104
4	(1,2,3,5)	6.72			8	0.100
					10	0.105
5	(1,2,3,4,6)	8.71	21.68			
5	(1,2,4,6,8)	7.17 ←	7.01 ←	1.40		
5	(1,2,3,4,5)	9.86				
5	(1,2,3,5,7)	9.08				

Table 3: Methods LR, WG1 and WG2 applied to model 1

Table 3 shows the results of the three methods applied to the first model. Note that in the LR method, the program computes the complexity of all admissible structures for each different order; method WG1, on the other hand, is recursive: for an order n, only those structural vectors are considered which contain all the rows selected in the optimal (n-1)-th order structural vector. The stopping criterion J used with WG1 has been defined at the end of section 4.

All three methods reject the odd rows in the structural vector. (Recall that the first p rows are always chosen following condition 2 of section 3). The LR method gives no estimate of the order. In the WG1 method, the criterion $J = \frac{1}{n} \operatorname{tr}(A^1)^{-1}$ is minimum for n = 4, while in method WG2, after the triangularization of row 4 (third step of the procedure) the decrease of the pivots is not significant anymore, which suggests an order 3.

Finally we add that the use of Akaike's canonical correlation method [3] on this model leads to the structural vector h = (1,2,4).

Order	Structural vector	LR method Complexity	WG1 method trace	WG1 method J	WG2 method selected row	WG2 method |pivot|
3	(1,2,3)	0.92 ←	2.72			
3	(1,2,4)	1.14	2.28 ←	0.76	1	0.735
					2	0.964
4	(1,2,3,4)	2.06 ←	4.89 ←	1.22	4	0.230
4	(1,2,4,6)	2.68	9.22		3	0.063
4	(1,2,3,5)	4.08			5	0.062
					7	0.063
5	(1,2,3,4,5)	7.17	7266.8			
5	(1,2,3,4,6)	3.76 ←	15.11 ←	5.04		
5	(1,2,4,6,8)	7.91				
5	(1,2,3,5,7)	6.24				

Table 4: Methods LR, WG1 and WG2 applied to model 2

In table 4 we see that the three methods lead to a structural vector h = (1,2,3,4) if the supposed order is 4, but if n = 3, the LR method leads to h = (1,2,3) while the other two methods indicate h = (1,2,4). The estimation of the order in WG1 leads to n = 3, while visual inspection of the pivots in WG2 leads to n = 4. Again, the simulation of Akaike's method leads to h = (1,2,4).

The conclusion we have drawn from our simulations is that in most cases (with results of Table 4 being the sole exception) method WG1 seems to give the best estimate of the order of the model. We recall that Akaike's method

gives an order estimate but does not select a "best basis," while the LR method does not estimate the order but selects a "best basis" within a prescribed order. As for method WG2, a better criterion than the visual inspection of the decrease of the pivots could probably be found.

We believe that all three methods give fairly good results for the determination of the structure, the advantage of the methods WG1 and WG2 being that they provide some estimate of the order as well, which avoids the fitting of too many parametrizations.

8. CONCLUSIONS. We have shown that the problem of specifying identifiable parametric structures for multivariable systems can be solved by a factorization of the probability map in such a way as to define a finite set of invariants which completely characterize the process. Proceeding in this way we have constructed a family of overlapping parametrizations which completely cover the set of finite-dimensional minimal-order systems. Since a given process can in general be represented by different overlapping parametrizations, the question then arises as to whether some parametrizations might yield more accurate parameter estimates than others. Our main result is that all overlapping parametrizations yield asymptotically the same value to the determinant of the information matrix. Therefore, when a prediction error identification method is used for the estimation of the parameters, all overlapping parametrizations will give the same value to the determinant of the asymptotic error covariance matrix. This is an asymptotic result. It does not imply that some structures might not be better than others when only a finite data record is available. Two heuristic selection schemes that can be used for the finite data case have been presented.

REFERENCES

1. D. G. Luenberger, *Canonical forms for linear multivariable systems*, IEEE Trans. Aut. Cont., vol. 12, pp. 290-293, 1967.

2. M. J. Denham, *Canonical forms for the identification of multivariable linear systems*, IEEE Trans. Aut. Cont., vol. 19, pp. 646-656, 1974.

3. H. Akaike, *Canonical Correlation Analysis of Time Series and the Use of an Information Criterion*, System Identification: Advances and Case Studies, (R. Mehra and D. Lainiotis, eds.), Academic Press, 1976.

4. J. Rissanen, *Basis of invariants and canonical forms for linear dynamic systems*, Automatica, vol. 10, pp. 175-182, 1974.

5. R. P. Guidorzi, *Invariants and canonical forms for systems structural and parametric identification*, Automatica, vol. 17, pp. 117-133, 1981.

6. K. Glover, J. C. Willems, *Parametrization of linear dynamical systems, canonical forms and identifiability*, IEEE Trans. Aut. Cont., vol. 19, pp. 640-646, 1974.

7. L. Ljung, J. Rissanen, *On canonical forms, parameter identifiability and the concept of complexity*, 4th IFAC Symp. on Identification and System Parameter Estimation, Tbilisi, USSR, 1976.

8. A. J. M. van Overbeek, L. Ljung, *On line structure selection for multivariable state space models*, 5th IFAC Symp. on Identification and System Parameter Estimation, Darmstadt, FRG, 1979.

9. V. Wertz, M. Gevers, E. Hannan, *The determination of optimum structures for the state space representation of multivariate stochastic processes*, submitted for publication.

10. J. Rissanen, *Estimations of structure by minimum description length*, Proc. Intern. Workshop on Rational Approximations for Systems, Leuven, Belgium, 1981.

11. M. Deistler, E. J. Hannan, *Some properties of the parametrizations of ARMA systems with unknown order*, to appear in J. Multivariate Analysis.

12. G. Picci, *Some numerical aspects of multivariable systems identification*, Proc. of the Workshop on Numerical Methods for Systems Engineering Problems, Lexington, Kentucky, June 1980.

13. B. D. O. Anderson, M. R. Gevers, *Overlapping state-space and ARMA parametrizations for multivariable systems*, in preparation.

14. M. van Emden, *Analysis of complexity*, Math. Cent. Tracts, 35, Amsterdam, 1971.

15. V. Wertz, *Détermination de la structure de processes multivariés*, Department of Electrical Engineering, Louvain University, IRSIA Report 79-80.

16. G. H. Golub, V. Klema, G. W. Stewart, *Rank degeneracy and least squares problems*, Tech. Reprt. Stan. CS76.559, Stanford University, August 1976.

17. G. H. Golub, G. P. H. Styan, *Numerical computations for univariate linear models*, J. Stat. Comp. Simul., 1973, vol. 2, pp. 253-274.

18. T. W. Anderson, *An introduction to multivariate statistical analysis*, Wiley, 1958.

19. M. El Sherief, N. K. Sinha, *Determination of the structure of a canonical model for the identification of linear multivariable systems*, IFAC Symp. on Identification and Parameter Estimation, Darmstadt, FRG, Sept. 1979.

LABORATOIRE D'AUTOMATIQUE ET D'ANALYSE DES SYSTÈMES
LOUVAIN UNIVERSITY, BÂTIMENT MAXWELL
B-1348 LOUVAIN-LA-NEUVE
BELGIUM