WORKING PAPER

LINEAR SYSTEM IDENTIFICATION - A SURVEY

M. Deistler

October 1989
WP-89-078

IIASA
International Institute
for Applied Systems Analysis

LINEAR SYSTEM IDENTIFICATION - A SURVEY

M. Deistler

October 1989
WP-89-078

Institute of Econometrics, University of Vienna, Austria

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

FOREWORD

This is a contribution to the activity on the topic *From Data to Model* initiated at the Systems and Decision Sciences Program of IIASA by Professor J. C. Willems.

A. Kurzhanski
Program Leader
System and Decision Sciences Program.

LINEAR SYSTEM IDENTIFICATION - A SURVEY

M. DEISTLER

Abstract

In this paper we give an introductory survey on the theory of identification of (in general MIMO) linear systems from (discrete) time series data. The main parts are: Structure theory for linear systems, asymptotic properties of maximum likelihood type estimators, estimation of the dynamic specification by methods based on information criteria and finally, extensions and alternative approaches such as identification of unstable systems and errors–in–variables.

Keywords

Linear systems, parametrization, maximum likelihood estimation, information criteria, errors–in–variables.

1. INTRODUCTION

The problem of deducing a good model from data is a central issue in many branches of science. As such problems are often far from being trivial and on the other hand often have a lot of common structure, systematic formal approaches for their solution have been developed. A large part of statistics, parts of system theory (namely system identification) and of approximation theory are concerned with this topic.

Here a special, but important case is considered, namely identification of *linear systems* from (equally spaced discrete) *time series data*. Both with respect to the existing body of theories and with respect to applications, linear system identification is quite an extensive subject now. The most important applications are signal processing (e.g. speech processing, sonar and radar applications), control engineering, econometrics, time series analysis of geophysical and metereological data, and the analysis of medical and biological time series (e.g. EEG analysis). In different areas emphasis has been put on different problems (and there still seems to be lack of communication between scientists working in those areas). For instance in modern system and control theory, a lot of emphasis has been put on the structure theory for linear multi-input multi-output (MIMO) systems, in signal processing on on-line algorithms for real time calculation and in statistical time series analysis on asymptotic properties of (mainly off-line) estimation procedures.

Linear system identification has many different aspects and facets depending among others on the goals one wants to achieve, on the amount of a priori information available, on the nature of data and on the way that noise is modelled. Nevertheless in the last twenty years something like a "mainstream" theory has been developed.

In system identification one has to specify:

(i) The *model class* i.e. the class of all a priori feasible systems which are candidates to be fitted to the data.

(ii) The *class of observations* $(y(t))$.

(iii) The *identification procedure* which is a rule (in the automatic case a function) attaching to every finite part of the data of the form $(y(t)|t=1...T)$ a system from the model class.

The actual problem of linear system identification, however, has much additional structure. We now describe the basic assumptions and ingredients

of the mainstream approach. At the end of our contribution we indicate some deviations from this approach.

(i) The systems contained in the model class are (in general MIMO) causal, stable, finite dimensional and time-invariant linear dynamic systems. Here in addition we restrict ourselves to the discrete-time case, where the range of time points are the integers $\mathbf{Z}$. The two most important system representations in this case are the state-space and the ARMA(X) representation. For simplicity and since the differences are minor (see e.g. Hannan and Deistler, 1988, Chapter 2 for a discussion) we only discuss the second case here, i.e. the case where

$$a(z)y(t) = b(z)\epsilon(t) \tag{1.1}$$

where $y(t)$ is the s-dimensional output, $\epsilon(t)$ is the m-dimensional input, z is used for a complex variable as well as for the delay operator (i.e. $z(y(t)|t\in\mathbf{Z}) = (y(t-1|t\in\mathbf{Z}))$ and finally where

$$a(z) = \sum_{j=0}^{p} A(j)z^j, A(j)\in\mathbb{R}^{s\times s}, b(z) = \sum_{j=0}^{q} B(j)z^j, B(j)\in\mathbb{R}^{s\times m} \tag{1.2}$$

With the exception of the last section unless the contrary is stated explicitely we will assume

$$\det a(z) \neq 0 \qquad |z| \leq 1 \tag{1.3}$$

and we will only consider the steady state solution

$$y(t) = \sum_{j=0}^{\infty} K(j)\epsilon(t-j) \tag{1.4}$$

of (1.1), where

$$\sum_{j=0}^{\infty} K(j)z^j = k(z) = a^{-1}(z)b(z) \tag{1.5}$$

Thus we restrict ourselves to the stable steady state case.

(ii) Every reasonable identification procedure has to separate the "essential" part from the "noisy" part of the data. For instance, for an ARMAX system, where in general the data will not exactly fit to the deterministic part of such a system, a decision has to be made what is attributed to the deterministic part and what is attributed to noise. A

basic decision that has to be made is whether we should (explicitely) model noise or not. In statistics this is an old question and the answer to it constitutes dividing line between descriptive and inferential statistics.

Here we give a stochastic model for the noise part, and thus, from this point of view, our problem becomes part of inferential statistics. In this case, additional a priori assumptions on the stochastic noise process, such as stationarity and ergodicity have to be imposed, in order to make inference a sensible task. The advantage of such a way of noise modelling is that the quality of identification procedures can be evaluated in a formal–mathematical way, for instance by deriving asymptotic properties of estimators. On the other hand, such a priori assumptions on the noise are not innocent and in actual applications the question has to be posed whether such a priori assumptions can be justified, or at least whether such a stochastic noise process provides a meaningful "test case" for the evaluation of identification procedures. These questions in particular have be posed in applications such as in econometrics or control engineering where there is rarely any stochastic theory or even vague a priori reasoning about the nature of noise.

(iii) The next question is, how the deterministic system should be embedded in its stochastic "environment". In mainstream analysis all of the noise is added to the equations or (which is the same in most respects) to the outputs, whereas the inputs are assumed to be observed without noise. This can be modelled by distinguishing between observed inputs and unobserved noise inputs in the vector $\varepsilon(t)$. In addition in this approach, the noise process is assumed to be uncorrelated with the observed inputs. If the contrary is not stated explicitely, here, for simplicity we will assume $m = s$ and that $\varepsilon(t)$ will consist of unobserved white noise errors only, i.e.

$$E\,\varepsilon(t) = 0, \qquad E\,\varepsilon(s)\varepsilon'(t) = \delta_{st}.\Sigma \tag{1.6}$$

In this case (1.1) is called an ARMA system and its solution (1.4) is called an ARMA process. As is well known such a process is stationary with spectral density given by

$$f_y(\lambda) = (2\pi)^{-1}.k(e^{-i\lambda}).\Sigma.k(e^{-i\lambda})^*$$

(where * denotes the conjugate transpose). In addition we assume

$$k(0) = I, \qquad \Sigma > 0 \tag{1.7 a,b}$$

and the miniphase condition

$$\det b(z) \neq 0 \qquad |z| < 1 \tag{1.8}$$

As is well known, for given f_y assumptions (1.7 a) and (1.8) are costless. As is also well known, under (1.7), (1.8), k and Σ are uniquely determined from f_y. For the additional complications arising in the ARMAX case, the reader is referred to Hannan and Deistler (1988).

(iv) For many cases discussed in this paper, the decision, which system has to be chosen, given the data, is based on optimizing a function which, in general, describes a certain trade off between goodness of fit of a system to the data and the complexity of the system. Thus we have to introduce a measure for goodness of fit, a measure for the complexity of a system and we have to formulate the trade off between the contradictory goals to maximize goodness of fit and to minimize the complexity of the system used. Clearly these choices are very much related to measures for the quality of inference procedures.

In the mainstream approach the (Gaussian) likelihood or a function of the (one step ahead) prediction error variance–covariance matrix are used as measures for goodness of fit and the quality of (parameter) estimators is described in terms of consistency and relative asymptotic efficiency.

In case of "small" model classes only goodness of fit is optimized. Measures of complexity are used in addition, in particular if the original model class is so large that it has to be broken up into subclasses and the subclass has to be determined from the data too. Since in a "large" model class measures of goodness of fit alone, such as the likelihood would tend to overfit the sample, such a measure of fit has to be "penalized" by a term measuring complexity of a system usually, in terms of the dimension of the parameter space. This is explained in detail in Section 4.

Let us consider the case, where the (original) model class is T_A, the set of all ARMA systems (a,b) (satisfying our assumptions) for given s (but for arbitrary p,q) i.e. where we have no a priori assumptions besides the general ones mentioned above. By U_A we denote the set of all transfer functions $a^{-1}.b$ corresponding to T_A and by $\pi: T_A \to U_A$ we denote the mapping defined by $\pi(a,b) = a^{-1}.b$. Two ARMA systems are called observationally equivalent if they have the same transfer function k. A set $T \subset T_A$ is called *identifiable* if π restricted to T is injective; in this case the mapping $\psi: \pi(T) \to T: \psi(\pi(a,b)) = (a,b)$ is called an *(ARMA–)parametrization* of $V = \pi(T)$. For

$(a,b)\in T$, in general not all entries of the parameter matrices $A(j)$, $B(j)$ may be needed for a unique description of (a,b) due to constraints. A vector τ of entries of the $A(j)$, $B(j)$, such that $(a,b)\in T$ is uniquely determined from τ, and such that τ (whose dimension is kept constant over T) has a minimal number of entries, is called a vector of *free parameters* for T (or for $\pi(T)$). We will identify (a,b) with τ.

Every parametrization of U_A has the disadvantage that the corresponding parameter space T is infinite dimensional and clearly finite dimensional parameter spaces are more convenient for inference. What is even more cumbersome is the fact that there exists no continuous parametrization of U_A. For these reasons, U_A (and T_A) is broken into parts U_α, $\alpha\in I$, in a way that every such part can be parametrized separately, by $\psi_\alpha: U_\alpha \to T_\alpha$ say, in a convenient way.

For the sake of mathematical convenience, we may decompose an identification procedure into three steps. The first step is to determine the subclass U_α, or the index α, characterizing this subclass, from the data. Here α is a multi-index of integers, in the scalar case ($s=1$) the usual choices are $\alpha=(p,q)$ or $\alpha=(n);n=\max(p,q)$. The determination of α sometimes is called dynamic specification. Here we will almost exclusively deal with automatic procedures for dynamic specification which are in particular inference procedures based on optimization of a function describing a certain trade off between goodness of fit and complexity as has been mentioned above. However it should be emphasized that (besides the case where suitable a priori information about α is available from "physical" theories and where therefore the first step is omitted), in particular for the scalar case, dynamic specification may also be performed by non-automatic procedures (where subjective judgement based on certain patterns is involved), the most prominent of which is the Box-Jenkins procedure (Box and Jenkins 1970). Once α has been determined, for mathematical convenience, estimation of the free parameters τ and $\sigma(\Sigma)$ (where $\sigma(\Sigma)$ is the vector of on and above diagonal elements of Σ) may be decomposed into two further steps namely estimation of the transfer function k (by $\hat{k}$ say) and of $\sigma(\Sigma)$ (by $\sigma(\hat{\Sigma})$) and, finally the realization of the estimated transfer function to obtain the parameter estimator $\hat{\tau}=\psi_\alpha(\hat{k})$. Whereas the second step is concerned with statistics in the strict sense [namely with extraction of information from data], the third is concerned with (deterministic) realization and only properties of the parametrization are relevant. In order to estimate k and Σ in the second step usually a criterion for goodness of fit, such as the likelihood is optimized. The

decomposition of the problem into these two further steps is based on the observation that most of these criteria only depend on τ via k.

The structure of the problem of identification of linear systems (when the original model class is T_A[or U_A]) can be schematically represented by the following figure:

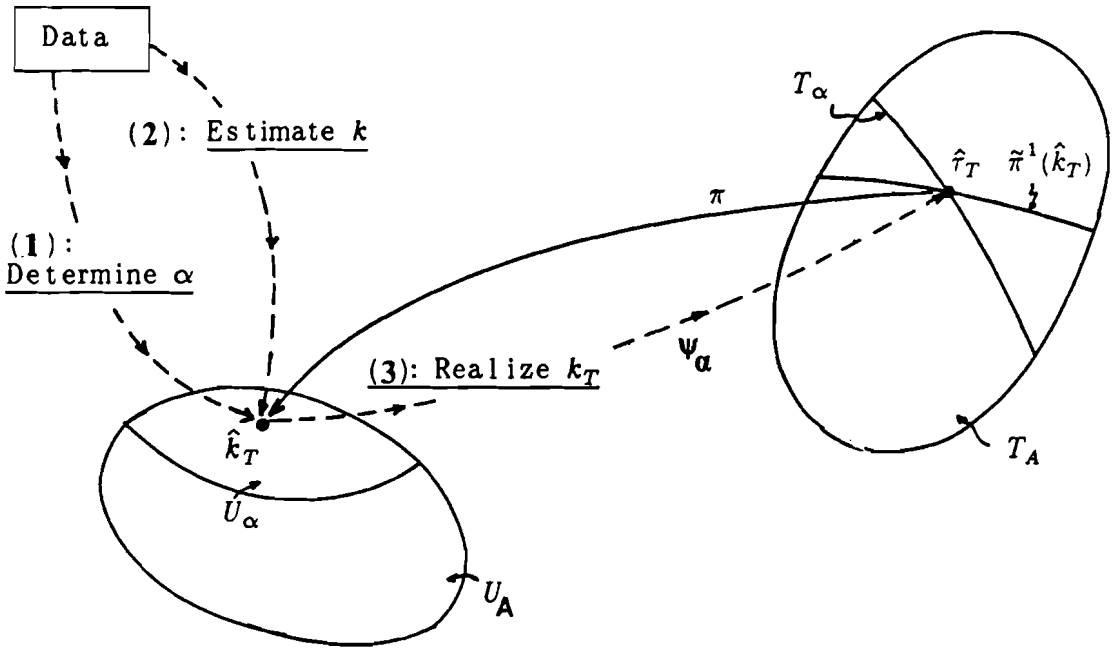

Fig. 1: The Structure of Linear System Identification

2. REALIZATION AND PARAMETRIZATION

As has been pointed out already U_A has to be broken into parts U_α, $\alpha \in I$, in order to allow for a convenient parametrization $\psi_\alpha: U_\alpha \to T_\alpha$. Clearly, there are many different ways to define such parts. From the point of view of identification some desirable properties of such parameter spaces and parametrizations are:

(i) T_α is identifiable; i.e. the mapping $\psi_\alpha: U_\alpha \to T_\alpha: \psi_\alpha(\pi(a,b)) = (a,b)$; $(a,b) \in T_\alpha$ exists.

(ii) T_α can be embedded into an Euclidian space $\mathbb{R}^{d_\alpha}$, i.e.the parameter space is finite dimensional; in addition T_α should contain an open set of $\mathbb{R}^{d_\alpha}$.

(iii) An important property of ψ_α is its continuity in the sense that T_α is endowed with the relative Euclidean topology and U_A is endowed with the relative topology of $(\mathbb{R}^{s \times s})^{\mathbb{N}}$, where the transfer functions are identified with their power series coefficients $(K(j) | j \in \mathbb{N})$. The latter topology is called the *pointwise topology* T_{pt} and is quite natural in our context,

since the maximum likelihood estimators of the transfer functions k can be shown to be consistent in this sense. As is clear immediately, continuity of the mapping ψ_α relating the external characteristics k to the internal characteristics $\tau \leftrightarrow (a,b)$ makes the identification problem well posed and implies consistency for the estimators of τ for every estimation method (as the maximum likelihood method) which gives consistent estimators of k. As will be discussed later, also openness of U_α in $\bar{U}_\alpha$ is desirable. Note that in our analysis we do not need to show that the mapping relating *second moments* of $(y(t))$ to parameters τ is continuous, since the starting point of the analysis is consistency of *transfer functions*. For asymptotic normality of the estimators of τ, some differentiability properties are required.

(iv) A reasonable requirement is that the set of all U_α, $\alpha \in I$ is a cover for U_A, i.e. $\bigcup_{\alpha \in I} U_\alpha = U_A$.

(v) There is a certain trade off between the size of the cover U_α, $\alpha \in I$ and the dimension of the corresponding parameter spaces for the U_α. Vaguely speaking a coarser cover would tentatively make the determination of α simpler but would give a larger dimension of the parameter space T_α actually used and thus more components the parameter vector τ have to be estimated, which would cause a certain "efficiency loss". Another H in certain sense, reasonable requirement seems to be that the cover is minimal in the sense that no element of the cover can be removed without loosing the covering property.

In particular for the multi-output ($s>1$) case, there is a number of different parametrizations which are used, the most important of which are Echelon canonical forms, the overlapping parametrization of the manifold of all systems of order n and monic (in the sense that $a(0)=I$ holds) ARMA systems with prescribed column degrees. [there is a large number of references to this, see Hannan and Deistler 1988 and the references therein]. We will only describe *Echelon-forms* here. We begin with a transfer function of the form

$$\tilde{k}(z) = k(z^{-1}) = \sum_{j=0}^{\infty} K(j) z^{-j} \tag{2.1}$$

rather than with $k(z)$, for mathematical convenience. Causality of k means that $\tilde{k}$ is proper [i.e. $\lim_{|z| \to \infty} \tilde{k}(z)$ is finite]. An ARMA system $(\tilde{a},\tilde{b})$ corresponding to $\tilde{k}$ (i.e. $\tilde{a}^{-1}.\tilde{b}=\tilde{k}$) then [in an obvious notation] is of the form

$$\sum_{j=0}^{\tilde{p}} \tilde{A}(j)y(t+j) = \sum_{j=0}^{\tilde{q}} \tilde{B}(j)\varepsilon(t+j) \tag{2.2}$$

Let

$$H = \begin{pmatrix} K(1),K(2),\dots \\ K(2),K(3),\dots \\ \dots\dots\dots\dots \end{pmatrix}$$

denote the (block) Hankel matrix of k. Then from a comparison of coefficients corresponding to negative powers of z in $\tilde{a}(z).k(z^{-1})$ we obtain

$$(\tilde{A}(o),\tilde{A}(1),\dots).H = 0 \tag{2.3}$$

As is well known, since k is rational, the rank of H is finite and furthermore this rank is equal to the *order* i.e. the degree of $\det \tilde{a}$ for any (left) coprime$(\tilde{a},\tilde{b})$ [i.e. $\tilde{a},\tilde{b}$ have no nonunimodular common matrix polynomial (left) divisor; a polynomial matrix u is called unimodular if $\det u = \text{const} \neq 0$] corresponding to $\tilde{k}$. Let $M(n)$ denote the set of all $k \in U_A$ such that H has rank n. Further, let $h(i,j)$ denote the jth row in the ith block of rows of H. Due to the block Hankel structure of H, the first rows (in natural ordering) of H which form a basis for the row space of H are of the form

$$h(1,1),\dots,h(n_1,1),h(1,2),\dots,h(n_2,2),\dots,h(1,s),\dots,h(n_s,s)$$

for a suitable chosen multi-index $\alpha = (n_1,\dots,n_s)$; these $n_1 \dots n_s$ are called the *Kronecker indices*. Clearly $n = n_1 + \dots + n_s$. Expressing the respective first linear dependent rows in terms of the preceeding elements from this basis, we obtain

$$-h(n_i+1,i) = \sum_{j=1}^{s} \sum_{\mu=1}^{n_{ij}} \tilde{a}_{ij}(\mu-1)h(\mu,j) \quad , \qquad i = 1\dots s \tag{2.4}$$

where

$$n_{ij} = \begin{cases} \min\ (n_i+1,n_j) & \text{for} \quad j<i \\ \min\ (n_i,n_j) & \text{for} \quad j \geq i \end{cases}$$

Equations (2.4) define unique coefficients $\tilde{a}_{ij}(\mu)$ and they can be considered as special relations of the form (2.3) where $\tilde{a}_{ij}(\mu)$ is the (i,j)

element of $\tilde{A}(\mu)$, $\tilde{a}_{ii}(n_i)=1$, $i=1\ldots s$ and all other elements are equal to zero.

By this procedure, for every $\tilde{k} \leftrightarrow k \in U_A$ we have defined (unique) Kronecker indices $\alpha=(n_1\ldots n_s)$ and a corresponding unique ARMA realization $(\tilde{a},\tilde{b})$, with

$$\tilde{b}=\tilde{a}.\tilde{k} \tag{2.5}$$

where

$$(\tilde{a},\tilde{b}) \text{ is (left) coprime} \tag{2.6}$$

and (with $\delta(p)$ denoting the degree of polynomials)

$$\begin{aligned}
&\delta(\tilde{a}_{ij}) \le \sigma(\tilde{a}_{ii}) = n_i \quad ; \quad j \le i \\
&\delta(\tilde{a}_{ij}) < \delta(\tilde{a}_{ii}) \qquad\qquad j > i \\
&\delta(\tilde{a}_{ji}) < \delta(\tilde{a}_{ii}) \qquad\qquad j \ne i \\
&\delta(\tilde{b}_{ij}) \le \delta(\tilde{a}_{ii}) \\
&\text{the row-end matrices in } \tilde{a} \text{ and } \tilde{b} \text{ are the same.}
\end{aligned} \tag{2.7}$$

Such a unique realization is called the Echelon form. As can be shown, conversely every ARMA system satisfying (2.6) and (2.7) is in Echelon form.

An ARMA realization for k then is obtained from

$$(a(z),b(z)) = \operatorname{diag}\{z^{n_i}\}(\tilde{a}(z^{-1}),\tilde{b}(z^{-1})) \tag{2.8}$$

and this is called the the *reversed Echelon form*. For reversed Echelon form we have:

$$(a,b) \text{ is (left) coprime} \tag{2.9}$$

and

$$\begin{aligned}
&A(0)[=B(0)] \text{ is lower triangular and all its} \\
&\text{diagonal elements are equal to one;} \\
&\text{the degree of the } i\text{th row is } n_i; \\
&z^{n_i-n_{ij}} \text{ divides } \tilde{a}_{ij}.
\end{aligned} \tag{2.10}$$

Let U_α denote the set of all $k \in U_\alpha$ with Kronecker indices $\alpha=(n_1,\ldots,n_s)$, and T_α denote the set of all $(a,b) \in T_A$ satisfying (2.6), (2.7) and (2.8). For

$(a,b) \in T_\alpha$ a vector $\tau \in \mathbb{R}^{d_\alpha}$ of free parameters consisting of all elements of

$(\tilde{a},\tilde{b})$ which are not explicitely restricted by (2.7) is defined where

$$d_\alpha = (\sum_{i=1}^{s} n_i)(s+1) + \sum_{i,\, j:j<i} (\min(n_i,n_j)+\min(n_j,n_i+1)) \tag{2.11}$$

Then by the procedure described above in introducing (reversed) Echelon form we have defined a parametrization $\psi_\alpha: U_\alpha \to T_\alpha$. By $\bar{A}$ we denote the closure of the set A, and by $\beta = (m_1 \ldots m_s) \leq \alpha = (n_1 \ldots n_s)$ we mean $m_i \leq n_i, i = 1 \ldots s$. $\beta < \alpha$ is to indicate that $m_i < n_i$ for at least one i holds. For the next theorem we do not impose assumptions (1.3) and (1.8). We have:

Theorem 2.1:

(i) T_α is open and dense in $\mathbf{R}^{d_\alpha}$

(ii) $\psi_\alpha: U_\alpha \to T_\alpha$ is a $(T_{pt}-)$ homeomorphism

(iii) $\{U_\alpha | \sum_{i=1} n_i = n\}$ is a disjoint partition of $M(n)$
containing $\begin{pmatrix} n+s-1 \\ s-1 \end{pmatrix}$ elements

(iv) $\pi(\bar{T}_\alpha) = \bigcup_{\beta \leq \alpha} U_\beta$

(v) For every $k \in U_\beta$, $\beta \leq \alpha$, the class of all observationally equivalent ARMA systems in $\bar{T}_\alpha$ is an affine subspace of dimension

$$\sum_{i=1}^{s} \sum_{j=1}^{s} (n_{ij} - n'_{ij})$$

where

$$n'_{ij} = \begin{cases} \min(n_i+1, m_j) & \text{for} \quad j<i \\ \min(n_i, m_j) & \text{for} \quad j \geq i \end{cases}$$

(vi) U_α is $(T_{pt}-)$ open in $\bar{U}_\alpha$

(vii) $\pi(\bar{T}_\alpha) \subset \bar{U}_\alpha$ and equality holds for $s=1$

A similar result can be shown for the overlapping parametrization of $M(n)$ or for monic ARMA systems with prescribed column degrees (see e.g. Deistler 1983, Hannan and Deistler 1988, Deistler and Wang 1988). The implications of such results for estimation will be discussed in the next section.

3. ESTIMATION FOR GIVEN DYNAMIC SPECIFICATION

In most cases the estimators - at least asymptotically - only exploit information from the data $y(t), t = 1...T$, via their second moments

$$\hat{K}(s) = \begin{cases} (T)^{-1} \cdot \sum_{t=1}^{T-s} y(t+s)y'(t) & , \quad 0 \le s < T \\ \hat{K}'(-s) & , \quad 0 > s > -T \\ 0 & , \quad |s| \ge T \end{cases} \tag{3.1}$$

Clearly, these second moments can be "realized" by a moving average system of order $T-1$. [Note that typically, e.g. for the Gaussian case no data $y(t), t = 1...T$ in a deterministic sense could ever be incompatible with any system; by "realize" here we meant that we can find a system whose population second moments are given by (3.1)]. Such a system estimator however has two disadvantages. Typically it would "overfit" the data [i.e. it would use too many parameters for description] and second $\hat{K}(s) = 0$ for $|s| \ge T$, in general, is not a "good" extrapolation. So we have to "smooth" the $\hat{K}(s)$, $|s| < T$ by using (in general) less parameters for their (approximate) description and at the same time we have to extrapolate these values for $|s| \ge T$. This can also be understood as a smoothing of the periodogram

$$I(\omega) = (2\pi)^{-1} \sum_{s=-T}^{T} \hat{K}(s) e^{i\omega s} \tag{3.2}$$

by rational approximation. In addition, in general, the empirical second moments are not contained in the class of (population) second moments corresponding to the class $T_\alpha \times \underline{\Sigma}$ under consideration, so that estimation can be understood as approximating the empirical second moments of the data by an element corresponding to $T_\alpha \times \underline{\Sigma}$. Here $\underline{\Sigma} = \{\Sigma \in \mathbb{R}^{s \times s} | \Sigma > 0, \Sigma' = \Sigma\}$.

In mainstream theory the Gaussian maximum likelihood estimator (MLE) is the prototype estimator. Under Gaussian assumptions $-2T^{-1}$ times the logarithm of the likelihood of $y(1),...,y(T)$ is given up to a constant by

$$\hat{L}_T(\tau, \Sigma) = T^{-1} \log \det \Gamma_T(\tau, \Sigma) + T^{-1} y_T' \Gamma_T^{-1}(\tau, \Sigma) y_T \tag{3.3}$$

Here $y_T = (y'(1),...,y'(T))'$ denotes the stacked vector of the data and

$$\Gamma_T(\tau, \Sigma) = \left[\int e^{-i\lambda(\tau - t)} \cdot f(\lambda; \tau, \Sigma) \, d\lambda \right]_{\tau, t = 1...T} \tag{3.4}$$

denotes the matrix of second moments of a vector $(y'(1)...y'(T))'$ made from

an ARMA process with parameters τ,Σ [correspondingly $f(\lambda;\tau,\Sigma)$ denotes the spectral density of such a process]. Since no confusion can arise, $\hat{L}_T$ is also called the likelihood function. Evidently $\hat{L}_T$ depends on the parameters τ only via k and thus we can define a likelihood by.

$$L_T(\pi(\tau),\Sigma)=\hat{L}_T(\tau,\Sigma) \tag{3.5}$$

This "coordinate-free" likelihood will prove to be mathematically convenient since certain statistical properties of MLE's can be analysed in terms of transfer functions.

If $U\subset U_A$ is the set of transfer functions considered, the MLE's $\hat{k}_T,\hat{\Sigma}_T$ [over $U\times\underline{\Sigma}$] are defined as

$$(\hat{k}_T,\hat{\Sigma}_T)=\underset{(k,\Sigma)\in U\times\underline{\Sigma}}{\arg\min}\ L_T(k,\Sigma) \tag{3.6}$$

In general it is not clear whether L_T has a minimum over $U\times\underline{\Sigma}$ (see e.g. Deistler and Pötscher 1984). What is much more important and cumbersome is that in general no explicit expression for the MLE will exist. Clearly in such a situation finite sample properties of the estimators would be hard to obtain. However the asymptotic analysis of the MLE's in this case has reached a certain stage of completeness now, see e.g. Hannan 1973, Dunsmuir and Hannan 1976, Hannan and Deistler 1988.

As far as consistency is concerned the main complications arise due to the noncompactness of the "natural" parameter spaces. For given $U\subset U_A$ under consideration let $\bar{U}$ denote its ($T_{pt}-$) closure, $\hat{U}$ the set of all $k\in\bar{U}$ which have no pole for $|z|=1$ and U^* the set of all $k\in\hat{U}$ which have no zero for $|z|=1$. We have (see Dunsmuir and Hannan 1976 Hannan and Deistler 1988).

Theorem 3.1. Let the true system satisfy

$$k_0\in U^* \tag{3.7}$$

let

$$\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\varepsilon(t+s)\varepsilon(t)=\delta_{s,0}.\Sigma_0 \qquad \text{a.s.} \tag{3.8}$$

and let $\bar{U}\subset\bar{\mathrm{M}}(\mathrm{n})$ for a suitable n. Then the MLE's over $\hat{U}\times\underline{\Sigma}$ are strictly consistent, i.e.

$$\lim_{T\to\infty}(\hat{k}_T,\hat{\Sigma}_T)=(k_0,\Sigma_0) \qquad \text{a.s.} \tag{3.9}$$

Thus consistency of the MLE's holds under fairly general conditions. For a consistency proof in the ARMAX case see Hannan and Deistler 1988.

If the data are not generated by a system contained in the model class U^* but by a general linear regular stationary process in Wold representation

$$y(t) = k_0(z)\varepsilon(t) \tag{3.10}$$

with

$$k_0(z) = \sum_{j=0}^{\infty} K(j)z^j \quad ; \qquad \sum_{j=0}^{\infty} \|K(j)\|^2 < \infty$$

then still a generalized consistency result (see e.g. Ljung 1978, Pötscher 1987, Hannan and Deistler 1988) in the following sense holds: Let D denote the subset of $\bar{U}\times\underline{\Sigma}$ where the "asymptotic form" of the likelihood

$$L(k,\Sigma) = \log\det\Sigma + (2\pi)^{-1}\int_{-\pi}^{\pi} tr\{(k\Sigma k^*)^{-1}(k_0\Sigma_0 k_0^*)d\lambda \tag{3.11}$$

attains its minimum over $\bar{U}\times\underline{\Sigma}$. As can be shown, $L(k,\Sigma)$ is the (a.s.) limit of $L_T(k,\Sigma)$ (for $T\to\infty$) and L is a measure of goodness of fit of a system to the complete (infinite) observations. D then is the set of all (k,Σ) which are the best approximations within $\bar{U}\times\underline{\Sigma}$ to the true system (k_0,Σ_0). Now the MLE's $\hat{k}_T,\hat{\Sigma}_T$ can be shown to be (a.s.) convergent to the set D. This is an important generalization of the consistency result of Theorem 3.1 since in many cases the true system may be of higher order or even not rational and this result indicates that in such cases the MLE's still give good approximations to the true system. In a certain sense this idea is related to robustness. As has been pointed out first by Kabaila (1983), D may consist of more than one point. However (Ploberger 1982) for the usual parameter spaces (e.g. for $\bar{U}_\alpha$ corresponding to Echelon forms), there is at least a neighborhood of $\bar{U}_\alpha\times\underline{\Sigma}$ [corresponding to the weak topology of spectral measures] such that if (k_0,Σ_0) is in this neighborhood, the best approximation within $\bar{U}_\alpha\times\underline{\Sigma}$ is unique (see Fig. 2)

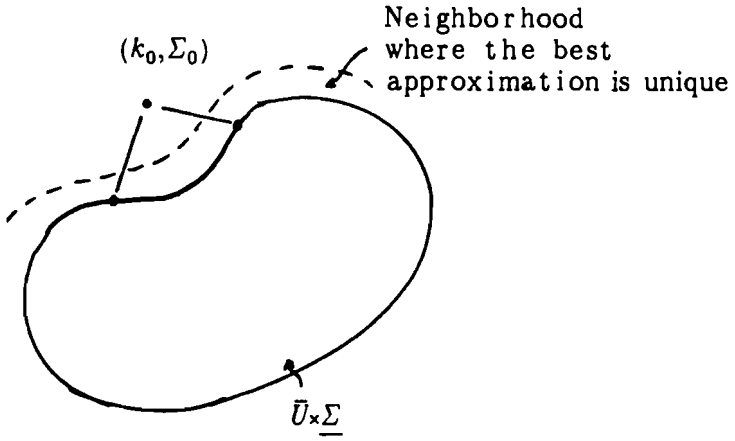

Fig. 2: Some aspects of approximation of (k_0,Σ_0) within $\bar{U}\times\underline{\Sigma}$

Let us stress again the general nature of the approach described above. In particular besides the boundedness of the degrees of the ARMA systems considered (i.e. $\bar{U}\subset M(n)$, for some n) no assumption has been imposed on the "parameter space" U (which here is a set of transfer functions). By the coordinate-free nature of the results, we had not to care about questions of existence and continuity of parametrizations. In particular, we were able to analyse the cases where k_0 is contained in the boundary U^*-U and also [since certain boundary points in the process of the optimization of the likelihood cannot be excluded a priori] the optimization of the likelihood is performed over $\hat{U}\times\underline{\Sigma}$ rather than over $U\times\underline{\Sigma}$.

However, actual calculation of the MLE's has to be performed in coordinates and in addition in many cases the parameters τ are of direct interest. Therefore we now consider estimation of the true parameter τ_0. Let $U=U_\alpha$ i.e. the set of all transfer functions $k\in U_A$ with Kronecker indices $\alpha=(n_1\ldots n_s)$ [as discussed in Section 2] and let $\psi_\alpha:U_\alpha\to T_\alpha$ be the corresponding parametrization [alternatively other standard parametrizations such as the overlapping parametrization of the manifold $M(n)$ or monic ARMA systems with prescribed column degrees may be chosen]. Then if $\hat{k}_T$ is the MLE [or any other consistent estimator] for k and if $\hat{k}_T\in\pi(\bar{T}_\alpha)$, we define a (nonnecessarely unique) MLE [or correspondingly another estimator] $\hat{\tau}_T$ of τ as any $\hat{\tau}_T\in\mathbf{R}^{d_\alpha}$ which satisfies $\pi(\hat{\tau}_T)=\hat{k}_T$. Clearly if $\hat{k}_T\in U_\alpha$, then $\hat{\tau}_T$ is uniquely given by $\psi_\alpha(\hat{k}_T)$. Investigating the behaviour of these parameter estimators we have to distinguish the following three cases:

(i) If the dynamic specification is correct in the sense that $k_0 \in U_\alpha$ holds, then $\hat{k}_T \to k_0$ and the openness of U_α in $\bar{U}_\alpha$ (Theorem 2.1) imply $\hat{k}_T \in U_\alpha$ from a certain T_0 onwards, and thus, at least for $T > T_0$, $\hat{\tau}_T = \psi_\alpha(\hat{k}_T)$ exists [Note that T_0 in general depends on the point ω in the sample space]. The continuity of ψ_α (Theorem 2.1) then implies

$$\lim_{T \to \infty} \hat{\tau}_T = \tau_0 = \psi_0(k_0) \tag{3.12}$$

and thus (under the conditions of Theorem 3.1), the MLE's $\hat{\tau}_T$ are strongly consistent for the parameter τ.

(ii) Next, we consider the case where $k_0 \in \pi(\bar{T}_\alpha) - U_\alpha$ holds, i.e. where there is a $\beta < \alpha$ such that $k_0 \in U_\beta$ (see Theorem 2.1). In this case k_0 corresponds to an equivalence class [containing more than one element] on the boundary of T_α, and the likelihood function $\hat{L}_T$ [when defined on $\bar{T}_\alpha \times \underline{\Sigma}$] is constant along this equivalence class [for any Σ]; moreover its asymptotic form L [which again here is considered as being defined on $\bar{T}_\alpha \times \underline{\Sigma}$], attains its minimum over this equivalence class [for Σ_0]. It might be the case that for $\hat{k}_T \to k$, the corresponding $\hat{\tau}_T$ will converge to infinity, without converging to the 'true' equivalence class. However, if we impose suitable prior bounds on the norm of the elements of $\bar{T}_\alpha$, then the [not necessarily unique] $\hat{\tau}_T$ will converge to the true equivalence class, but not necessarily to a fixed point within this class. Thus an identification algorithm may search along this class.

(iii) Finally we consider the case $k_0 \in \bar{U}_\alpha - \pi(\bar{T}_\alpha)$, which can only occur for $s > 1$. In this case k_0 corresponds to the point of infinity in the one point compactification of $\bar{T}_\alpha$; even if $\hat{k}_T \in U_\alpha$, $T \in \mathbb{N}$ holds, then $\hat{k}_T \to k_0$ implies $\|\psi_\alpha(\hat{k}_T)\| \to \infty$.

In order to discriminate between different consistent estimators and in order to obtain an approximate distribution for the parameter estimators, in the asymptotic analysis central limit theorem are provided (see e.g. Dunsmuir and Hannan 1976, Hannan and Deistler 1988).

For a central limit theorem we have to consider a parameter space $T \times \underline{\Sigma}$ (and not $U \times \underline{\Sigma}$) and we have to impose additional assumptions: First the parameter space $T \subset \mathbb{R}^d$ has to be open [this is not an essential assumption; for boundary points the limiting distribution would not be Gaussian]. For

standard parameter spaces, such as T_α, we have to strengthen (1.8) to $\det b(z) \neq 0$, $|z| \leq 1$, in order to ensure openness. Also, in addition to the assumptions of Theorem 3.1, the process generating the data is assumed to satisfy the following conditions: $\varepsilon(t)$ is strictly stationary and

$$E\{\varepsilon(t)|\mathcal{F}_{t-1}\} = 0 \tag{3.13}$$

where $\mathcal{F}_t$ is the σ–algebra generated by $\varepsilon(s)$, $s \leq t$.

Condition (3.13) seems to be quite natural in our context, since it is equivalent to the condition that the best (in least squares sense) predictor $E(y(t)|\mathcal{F}_{t-1})$ of $y(t)$ from its past $y(t-1), y(t-2), \ldots$ is equal to the best linear predictor of $y(t)$ given its past, and since in cases where the difference between these two predictors is substantial, nonlinear, rather than linear systems should be used.

Theorem 3.2. Let the true system satisfy $\tau_0 \in T_\alpha$; then under the assumptions of Theorem 3.1, the assumptions above, and under the assumption

$$E\{\varepsilon(t)\varepsilon'(t)|\mathcal{F}_{t-1}\} = \Sigma_0 \tag{3.14}$$

the vector $T^{1/2}(\hat{\tau}_T - \tau_0)$ has a Gaussian limiting distribution (with mean zero and with covariance matrix given by (the inverse of the Fisher information as):

$$\left\{ \left[\frac{1}{2} \frac{\partial}{\partial \tau_i \partial \tau_j} \left((2\pi)^{-1} \int_{-\pi}^{\pi} tr\{(k\Sigma k)^{-1}(k_0 \Sigma_0 k_0)\} d\lambda \right) \right]_{i,j=1..d} \right\}^{-1} \tag{3.15}$$

Here τ_i is the i–the entry of τ. If in addition

$$E\{\varepsilon_j(t)^4\} < \infty, \qquad j = 1..s \tag{3.16}$$

[where $\varepsilon_j(t)$ is the j–th entry of εt)] and

$$E\{\varepsilon_i(t)\varepsilon_j(t)\varepsilon_k(t)|\mathcal{F}_{t-1}\} = E\varepsilon_i(t)\varepsilon_j(t)\varepsilon_k(t) \qquad 1 \leq i,j,k \leq s \tag{3.17}$$

hold then also the on – and above diagonal elements of $T^{1/2}(\hat{\Sigma}_T - \Sigma_0)$ have a Gaussian limiting distribution.

From Theorems 3.1 and 3.2 we see that asymptotic properties of MLE's

obtained from a Gaussian likelihood are also valid for a class of non-Gaussian data. For instance if the data are generated by what is sometimes called a linear process, i.e. a process of the form

$$y(t) = \sum_{j=0}^{\infty} K_0(j)\varepsilon(t-j)$$

where $(\varepsilon(t))$ is a sequence of independent (not only uncorrelated) identically distributed random variables then (3.14) is fulfilled and $T^{1/2}(\hat{\tau}_T - \tau_0)$ will have a normal limiting distribution given by (3.15) independent of the actual distribution of the $\varepsilon(t)$. Clearly, if the actual distribution of $\varepsilon(t)$ were known, for the non Gaussian case, the actual (non Gaussian) likelihood would give estimators that have a smaller limiting variance covariance matrix than (3.15). As is well known, for Gaussian processes, the Gaussian MLE's are asymptotically efficient. By the last theorem we see that the Gaussian case is the worst case among all processes satisfying (3.14) and thus Gaussian likelihood estimation can be interpreted as minimization of the worst asymptotic variance covariance matrix.

4. DYNAMIC SPECIFICATION

In most applications the dynamic specification is not known a priori and has to be determined from the data. The development and evaluation of data-based procedures for dynamic specification constitutes one of the most important contributions to the subject during the last twenty years.
Theses procedures may be classified into non-automatic and automatic ones. In the non-automatic case subjective decisions have to be made at a certain stage. A particulary successful procedure of this kind was developed by Box and Jenkins (1970) for the SO case. The advantage of automatic procedures is that they do not require a large amount of experience.

First, as the perhaps most important case we consider the problem of estimating the order n. The classical procedure for choosing a model is the maximum likelihood method. However, since $\bar{M}(n_1) \subset \bar{M}(n_2)$ for $n_1 < n_2$ holds and $M(n_1)$ has smaller dimension than $M(n_2)$, the likelihood method will usually choose the largest allowed order [the same, more generally is true for every criterion which only contains a goodness of fit term]. The common procedures for order estimation are based on minimizing a criterion of the form

$$A(n) = \log\det \hat{\Sigma}_T(n) + (2ns)\frac{c(T)}{T} \quad ; \quad 0 \le n \le N \tag{4.1}$$

where $\hat{\Sigma}_T(n)$ is the MLE of Σ_0 over $\hat{M}(n)\times\underline{\Sigma}$ with sample size T, and N is a prescribed upper bound for the order and $c(T)$ is a prescribed function. Criteria of the form (4.1) have been mentioned already in the introduction. The first term of the righthand side of (4.1), namely $\log\det \hat{\Sigma}_T(n)$ is a measure for goodness of fit of a system to the data. For given T, $\log\det \hat{\Sigma}_T(n)$ will be decreasing for increasing n. The idea is, that this increase will be not so "significant" beyond the true order n_0 (if there is any), compared with the case when we are below the true order and that this "nonsignificant" decrease can be compensated by the "penalty term"

$$(2ns)\frac{c(T)}{T}$$

which contains the dimension $2ns$ of $M(n)$ as a measure of complexity. However, criteria of the form (4.1) are also meaningful for the case where the true system is infinite dimensional. N in (4.1) may depend on sample size T too (Hannan and Deistler 1988). Another interpretation of $A(n)$ is that it provides a tradeoff between bias (due to "underfitting") and efficiency loss by using too many parameters.

Clearly, $c(T)$ describes the tradeoff between goodness of fit and complexity in (4.1). The most common choices for $c(T)$ are $c(T)=2$, in which case $A(n)$ is called the *AIC criterion* $AIC(n)$and $c(T)=c.\log T$, $c\ge 1$ and then $A(n)$ is called the *BIC* criterion $BIC(n)$.

The actual choice of $c(T)$ can be motivated by a number of partially different ideas. Akaike (1969) (1977) described *AIC* from an entropy maximization principle or from ideas of optimal out of sample forecasting (see also Bhansali 1986, Findley 1985). Rissanen (1983) (1986) derived *BIC* from coding theory.

The asymptotic properties of order estimators based on (4.1) have been derived in Hannan (1980) (1981) for the case where a (finite) true order n_0 exists:

Theorem 4.1. Let $k_0 \in M(n_0)$; then under all assumptions in theorem 3.1. and under the additional assumptions (3.13), (3.14), (3.16)

$$\det k(z) \neq 0 \quad , \quad |z| < 1+\delta \quad \text{for some } \delta > 0$$

and in some coordinate system the norm of every τ is bounded a priori, the following results hold:

(i) If $c(T)/T \to 0$ (for $T \to \infty$) and $\lim_{T\to\infty} \inf[c(T)/\log T] > 0$

then

$$\hat{n}_T \to n_0$$

(ii) If $c(T)/T \to 0$ and $c(T) \uparrow \infty$, then

$$\hat{n}_T \to n_0 \text{ in probability}$$

(iii) If $\lim_{T\to\infty} \sup c(T) < \infty$ then

$$\lim_{\delta\to 0} \lim_{T\to\infty} P\{\hat{n}_T > n_0\} = 1$$

Thus in particular *AIC* gives no consistent estimator $\hat{n}_T$ for n_0. However, as has been shown by Shibata (1980), *AIC* has an optimality property if the true system is infinite dimensional.

The Kronecker indices α can also be estimated by a criterion of the form (4.1), in particular $A(n)$ gives consistent estimators of the Kronecker indices under analogous conditions as in the theorem above see Hannan and Kavalieris (1984).

Alternative inference procedures for dynamic specification are based on the investigation of the linear independence relations of an estimate of the block Hankel matrix H. Such an approach is appropriate in particular if for given n, the local coordinates in the overlapping parametrization of $M(n)$ have to be estimates, since in the case a criterion of the form (4.1) fails.

5. ALTERNATIVE APPROACHES AND EXTENSIONS

Here we give a short summary of some extensions and alternatives to the mainstream approach.

5.1. Identification of Unstable Systems

In many applications, the data show apparent non–stationarities which can be removed applying transformations such as detrending by trendregressions or (iterated) differencing before the actual identification procedure is applied. Clearly differencing removes a particular kind of instability [associated with unit roots of $\det a(z)$] however, a more general approach seems to be preferable.

For the case of unstable systems, i.e., if $\det a(z)$ has roots on or within the unit circle [and when causal solutions are considered], a complete theory is still not available.

For the *scalar* $(s=1)$ *autoregressive* case

$$y(t)=\alpha_1 y(t-1)+\ldots+\alpha_p y(t-p)+\varepsilon(t) \tag{5.1}$$

the following properties of the least squares estimator for $\tau=(\alpha_1,\ldots,\alpha_p)$, namely

$$\hat{\tau}_T=\left(\sum_{t=1}^{T} y_{t-1}y_{t-1}'\right)^{-1}\left(\sum_{t=1}^{T} y_{t-1}y_t\right) \tag{5.2}$$

where $y_t=(y(t),\ldots y(t-p+1))$, have been derived (under some additional assumptions):

(i) $\hat{\tau}_T$ is strictly consistent (Lai and Wei 1983).

(ii) For the special case $p=1$ and $\alpha_1=1$, i.e.

$$y(t)=y(t-1)+\varepsilon(t) \tag{5.2}$$

the limiting distribution of $\hat{\tau}_T(=\hat{\alpha}_{1,T})$ obeys the relation

$$T(\hat{\tau}_T-1)\xrightarrow{L}\frac{1}{2}(W^2(1)-1)/\int_0^1 W^2(t)dt \tag{5.3}$$

where $W(t)$ is a standard Brownian motion and where $\xrightarrow{L}$ indicates weak convergence of the distributions. This in particular shows that the convergence rate [for consistency] is T [rather than $T^{\frac{1}{2}}$ which is true for the stable case] and that the limiting distribution is no longer normal in general. The faster rate of convergence is quite plausible, since the regressor $y(t-1)$ becomes large in relation to the stationary error $\varepsilon(t)$. The result (5.3) is due to White (1958); this case was treated in a number of further papers, e.g. in Dickey and Fuller (1979).

(iii) The most general results seem to be those of Chan and Wei (1986). They deal with the case where all roots of $a(z)$ are on or inside the unit circle and they derive the limiting distribution of $\hat{\tau}_T$ and characterize them as a functional of stochastic integrals.

Another case of special unstable systems, namely the case of *cointegration* has attracted considerable attention in econometrics recently, see e.g. Engle and Granger (1987): Consider a nonstationary vector process $y(t)$, whose first differences $(1-z)y(t)$ are stationary [and linearly regular]. Such a process $y(t)$ is called cointegrated, if there exists a nonzero vector $a \in \mathbb{R}^s$ such that $a'y(t)$ is stationary. The interpretation is that a represents the (static) equilibrium solution of the system [where $a'y(t)$ is a stationary error which is smaller than the components of the variables]. This kind of models seems to be suited for a number of econometric applications, where in most cases the observed variables show trends in mean and variances but where there is some economic long-term "mechanism" "stabilizing" a certain linear combination of the components [such that it becomes relatively small]. An example for this would be if $y(t)$ contained consumption and income and the linear combinations correspond to a (static) comsumption function, or if $y(t)$ contained supply-side and demand-side variables for a market tending to equilibrium.
If we write

$$(1-z)y(t) = c(z)\varepsilon(t)$$

where $c(z)\varepsilon(t)$ is stationary and in Wold representation, and $c(z) = c(1) + (1-z)c^*(z)$, then we obtain

$$y(t) = (1-z)^{-1}c(1)\varepsilon(t) + c^*(z)\varepsilon(t) \tag{5.4}$$

From (5.4) we see that $y(t)$ is cointegrated iff $c(1)$ is singular. $\hat{y}(t) = (1-z)^{-1}c(1)\varepsilon(t)$ may be considered as unobserved "true" variables [since they satisfy the exact relation $a'\hat{y}(t) = 0$] and clearly they are generated by a vector autoregression, where all roots of $\det a(z)$ are equal to one; the second part on the r.h.s. of (5.4) are the stationary errors.
Estimators for a and tests for cointegration are considered e.g. in Engle and Granger (1987) and Phillips and Ouliaris (1986). Typically, here again the rate of consistency is T and the limiting distributions are obtained (via functional central limit theorems) from stochastic integrals.

5.2. Alternative Measures of Goodness of Fit

In particular in control engineering in many cases uniform approximation of transfer functions, in the sense that approximation in the norm $\sup_{\omega\in[\pi,\pi]} \|k(e^{-i\omega})\|$ is considered, is appropriate. However for such an approximation actual calculation would be difficult to perform. Balanced realizations and Hankel norm approximations are relatively easy to calculate and it is still possible to derive error bounds in the uniform norm for them (Glover 1984). However, most of the work done in this area commences from a known true transfer-function, rather than from data, and there are only a few results available on the statistical properties of procedures commencing from data, e.g. via a first estimate of the second moments.

5.3. Errors-in-Variables

Consider an ARMAX system, i.e.

$$a(z)y(t) = d(z)x(t) + b(z)\epsilon(t) \tag{5.6}$$

where $d(z) = \Sigma D(j)x(t-j)$, $D(j)\in\mathbb{R}^{s\times m}$ and $x(t)$ are observed inputs where $E\epsilon(s)x'(t) = 0$ for all s and t. ARMAX modelling, or more general errors-in-equations modelling is the "conventional" approach to embed a deterministic (input-output) system into a stochastic environment. However, there is a certain amount of unsymmetry in this way of modelling, since first we have to know a-priori the classification into inputs and outputs and second, and even more important, all of the noise is added to the equations or (for our analysis) equivalently to the outputs. *Linear errors-in-variables* (EV) modelling provides a more general way of modelling of the form:

$$w(z)\hat{z}(t) = 0 \quad ; \quad w(z) = \sum_{j=-\infty}^{\infty} W(j)z^j \; ; \; W(j)\in\mathbb{R}^{g\times(s+m)} \tag{5.7}$$

$$z(t) = \hat{z}(t) + w(t) \tag{5.8}$$

where $z(t)$ is the stacked vector of all observations at time t, i.e. $z(t) = (x(t)',y(t)')'$; $\hat{z}(t)$ is the corresponding vector of, in general unobserved, true, variables (which are related by the deterministic system (5.7) and $w(t)$ is a noise vector, where noise is added, in general, to each component. The main cases, when this more general EV setting is appropriate

are:

(i) If we are interested in the "true" system generating the data, rather than in encoding the data by system parameters, and is we cannot be sure a priori that the inputs are not corrupted by noise.

(ii) If we have no a priori classification of the observed variables into inputs and outputs or if even the number of outputs (i.e. the number of equations) is not known a priori and thus has to be determined from the data. Clearly $z(t)$ could also be modelled by a (vector) ARMA system, however in general, this leads to parameter spaces with dimension being considerably higher compared to the corresponding EV system.

(iii) Under certain additional assumptions on the noise structure EV–models are equivalent to dynamic principal component models or to dynamic factor analysis models. If we assume that the noise components are mutually uncorrelated then the model provides a decoupling of common and individual effects between the variables, where all common effects are attributed to the system.

One of the main problems in this context is identifiability of transfer functions (see e.g. Kalman 1982, Deistler and Anderson 1988, Picci and Pinzoni 1986). The statistical analysis is far from being complete.

REFERENCES

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

Akaike, H. (1977). On entropy maximisation principle. In *Applications of Statistics* (ed. P.R. Krishnaiah), 27–41. Amsterdam, North–Holland.

Bhansali, R.J. (1986). The criterion autoregressive transfer function of Parzen. *J. Time Series Anal.* **7**, 79–104.

Box, G.E.P., and Jenkins. G.M. (1970). *Time Series Analysis, Forecasting and Control*, San Fransisco, Holden Day.

Chan, N.A., and Wei, C.Z. (1986). Limiting distributions of least squares estimates of unstable autoregressive processes. To appear.

Deistler, M. (1983). The properties of the parametrization of ARMAX systems and their relevance for structural estimation. *Econometrica* **51**, 1187–1207.

Deistler, M, and Pötscher, B.M. (1984). The behavior of the likelihood function for ARMA models. *Adv. Appl. Probab.* **16**, 843–865.

Deistler, M., and Anderson, B.D. (1988). Linear dynamic errors invariables models: some structure theory. To appear in *J. Econometrics*.

Deistler, M., and Wang, Liqun (1987). The common structure of parametrizations for linear systems. To appear.

Dickey, D.A., and Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* 74,

427–431.

Dunsmuir, W., and Hannan, E.J. (1976). Vector linear time series models. *Adv. Appl. Probab.* **8**. 339–364.

Engle, R., and Granger, C.W.J. (1987). Co–integration and error–correction: Representation, estimation and testing. *Econometrica* **55**, 251–276.

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

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

Hannan, E.J. (1973). The asymptotic theory of linear time series models. *J. Appl. Probab.* **10**, 130–145.

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

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

Hannan, E.J., and Deistler, M. (1988). The Statistical Theory of Linear Systems, New York, John Wiley.

Hannan, E.J., and Kavalieris, L. (1984). Multivariate linear time series models. *Adv. Appl. Prob.* **16**, 492–561.

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

Kabaila, P. (1983). Parameter values of ARMA models minimising the one–step–ahead prediction error when the true system is not in the model set. *J. Applied Prob.* **20**, 405–408.

Kalman, R.E. (1982). System identification from noisy data, in (A. Bednarak and L. Cesari, eds) *Dynamical Systems II, a University of Florida International Symposium*, Academic Press, New York.

Lai, T.L., and Wei, C.Z. (1983). Asymptotic properties of general autoregressive models and strong consistency of least–squares estimates of their parameters. *J. Multivariate Anal.* **13**, 1–23.

Ljung, L. (1978). Convergence analysis of parametric identification methods. *IEEE Trans. Autom. Control* **AC–23**, 770–783.

Phillips, P.C.B., and Ouliaris, S. (1986). Testing for cointegration. *Cowles Foundation Discussion Paper* **809**.

Picci, G., and Pinzoni, S. (1986). Dynamic factor–analysis models for stationary processes. *IMA Math. Control and Information* **3**, 185–210.

Ploberger, W. (1982). Slight misspecifications of linear systems. In *Operations Research in Progress* (eds. G. Feichtinger and P. Kall), 413–424. Dordrecht, The Netherlands, D. Reidel.

Pötscher, B.M. (1987). Convergence results for maximum likelihood type estimators in multivariable ARMA models. *J. Multivariate Anal.* **21**. 29–52.

Rissanen, J. (1983). Universal prior for parameters and estimation by minimum description length. *Ann. Statist.* **11**, 416–431.

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

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

Solo, V. (1986). topics in advanced time series analysis. In *Lectures in Probability and Statistics* (eds. G. del Pino and R. Rebodedo). Berlin, Springer–Verlag.

White, J.S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *Ann. Math. Statist.* **23**, 1188–1237.

Willems, J.C. (1986). From time series to linear system, Part I: Finite dimensional linear time invariant systems. *Automatica*, **22**, 561–580.