# On Consistency for Prediction Error Identification Methods

Ljung, Lennart

REPORT 7405
MARCH 1974

# On Consistency for Prediction Error Identification Methods

LENNART LJUNG

Division of Automatic Control · Lund Institute of Technology

# On Consistency For

# Prediction Error Identification Methods

Lennart Ljung

ABSTRACT.

The consistency properties for a class of identification methods, that includes the maximum likelihood method are investigated. A general way of proving consistency is suggested and sets into which the parameters converge w.p.1 are determined. Vector difference equations and state space models are used as specific examples, but the results are valid for general systems. No assumptions about ergodicity of the input and output processes are introduced and the systems may be governed by general feedback regulators.

TABLE OF CONTENTS.

## 1. INTRODUCTION.

To apply the results of modern control theory on a given system it is usually required that the system dynamics are known. In many cases it is feasible to obtain the necessary knowledge about the system dynamics from identification experiments. By this is meant that input and output data are collected from the system and used to estimate certain unknown parameters with statistical methods. These parameters describe the process dynamics. The question what happens with the parameter estimates as the number of data used increases to infinity is the problem of consistency for the identification method.

A large number of different identification methods exist. A survey of the most important ones is given in e.g. Åström-Eykhoff (1971).

In this report a certain class of identification methods, prediction error methods, is considered. This class contains, under suitable conditions, the maximum likelihood method, and also the frequently used method of least squares.

The maximum likelihood method (ML method) was first introduced by Fisher (1912) as a general method for statistical parameter estimation. The problem of consistency for this method has been investigated by e.g. Wald (1949) and Cramér (1946) under the assumption that the obtained observations are independent. The first application of the ML method to system identification is due to Åström-Bohlin (1965), who considered single input single output systems of difference equation form. In this case the mentioned consistency results are not applicable. Åström-Bohlin (1965) showed one possibility to relax the assumption on independent observations.

ML identification using state space models have been considered by e.g. Caines (1970), Woo (1970), Aoki-Yue (1970), and Spain (1971). Caines-Rissanen (1974) have discussed vector difference equations. All these authors consider consistency with probability cne (strong consistency). Tse-Anton (1972) have proved consistency in probability for more general models. Balakrishnan has treated ML identification in a number of papers, see e.g. Balakrishnan (1968).

In the papers dealing with strong consistency, one main tool usually is an ergodic theorem. To be able to apply such a result, significant idealization of the identification experiment conditions must be introduced. The possibilities to treat input signals that are partly determined as feedback are limited, and an indispensable condition is that the likelihood function must converge w.p.1. To achieve this usually strict stationarity of the output is assumed. These conditions exclude many practical identification situations. For example, to identify unstable systems some kind of stabilizing feedback must be used. Other examples are processes that inherently are under time-varying feedback, like many economic systems.

In this report strong consistency for general prediction error methods, including the ML method is considered. The results are valid for general process models, linear as well as non linear. Also quite general feedback is allowed.

A general model for stochastic dynamic systems is discussed in Chapter 2. There also the identification method is described.
Different identifiability concepts are introduced in Chapter 3, where a procedure to prove consistency is outlined.

In Chapter 4 consistency is shown in case a suitable model is to be chosen from a finite family of models.
The general case is treated in Chapter 5.
Chapter 6 is devoted to the application of the results to vector difference equations and to state space models.

# 2. SYSTEMS, MODELS AND PREDICTION ERROR IDENTIFICATION METHODS.

## 2.1. System Description.

A causal discrete time, deterministic system, denoted by $S$, can be described by a rule to compute future outputs of the systems from input and previous outputs:

$$y(t+1) = f_S\big(t;y(t),y(t-1),\ldots,y(1);u(t),\ldots,u(1);\mathcal{Y}_0\big) \quad (2.1)$$

where $\mathcal{Y}_0$, "the initial conditions", represents the necessary information to compute y(1).

Often y(t+1) is not expressed as an explicit function of old variables, but some recursive way to calculate y(t+1) is preferred. Linear difference equations and state space models are well-known examples. The advantage with such a description is that only a finite and fixed number of old values are involved in each step.

For a stochastic system it cannot be required that the future outputs be exactly determined by previous inputs and outputs as in (2.1). It is then natural to consider the probability distribution of y(t+1) given all previous data. This can be expressed as follows

$$y(t+1) = E[y(t+1)|\mathcal{Y}_t,S] + \varepsilon(t+1,\mathcal{Y}_t,S) \quad (2.2)$$

where $E[y(t+1)|\mathcal{Y}_t,S)]$ is the conditional mean given all previous outputs and inputs,

$$E[y(t+1)|\mathcal{Y}_t,S] = g_S\big(t,y(t),\ldots,y(1),u(t),\ldots,u(1);\mathcal{Y}_0\big) \quad (2.3)$$

Here $\mathcal{Y}_t$ denotes the $\sigma$-algebra generated by

$\{y(t),\ldots,y(1);u(t),\ldots,u(1);\mathcal{Y}_0\}$, and $\mathcal{Y}_0$, "the initial condition", represents the information available at time $t = 0$ about the previous behaviour of the system.

The sequence $\{\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\}$ is a sequence of random variables for which holds

$$E[\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})|\mathcal{Y}_t,\mathcal{S}] = 0$$

It consists of the innovations, see Kailath (1970).

The conditional mean $E[y(t+1)|\mathcal{Y}_t,\mathcal{S}]$ will also be called the prediction of $y(t+1)$ based on $\mathcal{Y}_t$. Since it will frequently occur in this report a simpler notation

$$\hat{y}(t+1|\mathcal{S}) = E[y(t+1)|\mathcal{Y}_t,\mathcal{S}]$$

will be used.

Remark. With abuse of notation, random variables, like $y(t)$, will often be denoted by the same symbol as the outcomes. When there is a risk of ambiguity, the argument $\omega$ (the realization) is used for the outcomes, i.e. $y(t,\omega)$.

General stochastic systems can be described by (2.2), just as (2.1) is a general description of deterministic systems. The main results of this report will be formulated for this general system description (2.2).

For practical reasons, in the usual system descriptions the output is not given explicitely as in (2.2). Various recursive ways to calculate $y(t+1)$ are used instead. Examples are given below.

Example 2.1 - Vector difference equations.

Vector difference equations are frequently used as models of multivariable systems. They constitute a general class of linear models. Consider a system *S*. Assume that the input output relationships and the noise characteristics can be described by

$$A_S(q^{-1})y(t) = B_S(q^{-1})u(t) + C_S(q^{-1})e(t) \tag{2.4}$$

where $A_S(z)$ etc. are matrix polynomials in z:

$$A_S(z) = I + z\,A_{S,1} + \dots + z^n\,A_{S,n}$$

$$B_S(z) = z\,B_{S,1} + \dots + z^n\,B_{S,n}$$

$$C_S(z) = I + z\,C_{S,1} + \dots + z^n\,C_{S,n}$$

The operator $q^{-1}$ is the backward shift operator:

$$q^{-1}y(t) = y(t-1)$$

The variables $\{e(t),\ t = 0,1,\dots\}$ form a sequence of independent random variables with zero mean values and covariance $Ee(t)e^T(t) = \Lambda$. This matrix is assumed to be nonsingular. The output y(t) is a vector of dimension $n_y$ and the input u(t) has dimension $n_u$. It has been assumed that e(t) has the same dimension as y(t), which can be shown to be no loss of generality in this case.

A square matrix polynomial D(z) is said to be stable if all zeroes of det[D(z)] are strictly outside the unit circle. If $\det[C_S(z)]$ has no zeroes on the unit circle,

$C_S(z)$ can always be chosen as a stable polynomial without changing the second order noise characteristics, cf the spectral factorization theorem. In the sequel it is assumed that $C_S(z)$ is stable.

From (2.4) we have

$$C_S^{-1}(q^{-1})A_S(q^{-1})y(t) = C_S^{-1}(q^{-1})B_S(q^{-1})u(t) + e(t)$$

and

$$y(t+1) = [I - C_S^{-1}(q^{-1})A_S(q^{-1})]y(t+1) +$$

$$+ C_S^{-1}(q^{-1})B_S(q^{-1})u(t+1) + e(t+1) \qquad (2.5)$$

($C_S(z)$ is a constant square matrix for each given z. The inverse $C_S^{-1}(z)$ is therefore straightforwardly defined.) The right hand side of (2.5) contains only y(s) and u(s) up to time t. The term e(t+1) is independent of these variables, also in case u is determined from output feedback. Hence

$$E(y(t+1)|\mathcal{Y}_t,S) = [I - C_S^{-1}(q^{-1})A_S(q^{-1})]y(t+1) +$$

$$+ C_S^{-1}(q^{-1})D_S(q^{-1})u(t+1) \qquad (2.6)$$

Denote

$$E(y(t+1)|\mathcal{Y}_t,S) = \hat{y}(t+1|S)$$

Eq. (2.6) means that $\hat{y}(t+1|S)$ is found as the solution of

$$C_S(q^{-1})\hat{y}(s+1|S) = [C_S(q^{-1}) - A_S(q^{-1})]y(s+1) +$$

$$+ B_S(q^{-1})u(s+1) \qquad (2.7)$$

Solving (2.7) requires knowledge of $y(0),\ldots,y(-n),u(0),\ldots,u(-n),\hat{y}(0),\ldots,\hat{y}(-n)$. This information is supposed to be contained in information $\mathcal{Y}_0$.

Notice that there is parameter redundancy in the representation (2.4). All matrix polynomials $A'(z)$, $B'(z)$ and $C'(z)$ such that

$$[C'(z)]^{-1}A'(z) = [C_S(z)]^{-1}A_S(z)$$

$$[C'(z)]^{-1}B'(z) = [C_S(z)]^{-1}B_S(z) \quad \text{a.e.z.}$$

give, as seen from (2.6) the same function

$E\{y(t+1)|\mathcal{Y}_t;S\}$

□

Example 2.2 - State space equations.

The input output relation for the system $S$ is defined by

$$\begin{aligned} x(t+1) &= A_S x(t) + B_S u(t) + e(t) \\ y(t) &= C_S x(t) + v(t) \end{aligned} \qquad (2.8)$$

where $\{e(t)\}$ and $\{v(t)\}$ are sequences of independent gaussian random vectors with zero mean values and $Ee(t)e^T(t) = R_1(t)$, $Ee(t)v^T(t) = 0$ and $Ev(t)v^T(t) = R_2(t)$.

The function

$$E\left(y(t+1)\,|\,\mathcal{Y}_t,\mathcal{S}\right) = \hat{y}(t+1\,|\,\mathcal{S})$$

where $\mathcal{Y}_t$ is the $\sigma$-algebra generated by $\{y(t),\ldots,y(1),$ $u(t),\ldots,u(1),\mathcal{Y}_0\}$ is obtained as follows:

$$\hat{y}(t+1\,|\,\mathcal{S}) = C_{\mathcal{S}}\hat{x}(t+1\,|\,\mathcal{S}) \qquad (2.9)$$

where the state estimate $\hat{x}$ is obtained from standard Kalman filtering:

$$\hat{x}(t+1\,|\,\mathcal{S}) = A_{\mathcal{S}}\hat{x}(t\,|\,\mathcal{S}) + B_{\mathcal{S}}u(t) + K_{\mathcal{S}}(t)\{y(t) - C_{\mathcal{S}}\hat{x}(t\,|\,\mathcal{S})\} \qquad (2.10a)$$

$K_{\mathcal{S}}(t)$ is the Kalman gain matrix, determined from $A_{\mathcal{S}}$, $B_{\mathcal{S}}$, $C_{\mathcal{S}}$, $R_1$, and $R_2$ as

$$K_{\mathcal{S}}(t) = A_{\mathcal{S}}P_{\mathcal{S}}(t)C_{\mathcal{S}}^T[C_{\mathcal{S}}P_{\mathcal{S}}(t)C_{\mathcal{S}}^T + R_2]^{-1}$$

$$P_{\mathcal{S}}(t+1) = [A_{\mathcal{S}} - K_{\mathcal{S}}(t)C_{\mathcal{S}}]P_{\mathcal{S}}(t)[A_{\mathcal{S}} - K_{\mathcal{S}}(t)C_{\mathcal{S}}]^T +$$
$$+ R_1 + K_{\mathcal{S}}(t)R_2K_{\mathcal{S}}(t)^T \qquad (2.10b)$$

In many cases it is suitable to choose a representation of (2.8) that is adapted to prediction:

$$\hat{x}(t+1) = A_{\mathcal{S}}\hat{x}(t) + B_{\mathcal{S}}u(t) + K_{\mathcal{S}}\varepsilon(t)$$

$$y(t) = C_{\mathcal{S}}\hat{x}(t) + \varepsilon(t) \qquad (2.11)$$

where $E\varepsilon(t)\varepsilon^T(t) = \Lambda(t)$

Clearly, (2.11) is obtained from (2.9) and (2.10a) directly.

In case all matrices are time invariant, $K_S$ can be taken as the steady state gain. This has no influence on the asymptotic properties of the state estimates.

To solve (2.9) and (2.10) recursively from t = 0 requires knowledge of $\hat{x}(0)$ and $P_S(0)$. (The latter one is not needed in case (2.11) is used.) This information is supposed to be contained in $\mathcal{Y}_0$.

Notice as in Example 2.1 that there is parameter redundancy in the representation (2.8) and (2.11). All matrices A', B', C', K' such that

$$C'(zI-A')^{-1}B' = C_S(zI-A_S)^{-1}B_S$$

and

$$C'(zI-A')^{-1}K' = C_S(zI-A_S)^{-1}K_S$$

where the equalities shall hold for almost every z, give the same input output relationships and the same function

$$E[y(t+1)|\mathcal{Y}_t,\mathcal{S}]$$

A continuous time state representation can be chosen instead of (2.8). In e.g. Åström-Källström (1973) and Mehra-Tyler (1973) it is shown how $E[y(t+1)|\mathcal{Y}_t,\mathcal{S}]$, where $\mathcal{Y}_t$ is as before, can be calculated. The procedure is analogous to the one described above for sampled models.

□

These examples cover linear, possibly time varying systems. Clearly, also non-linear systems can be represented by (2.3). A simple example is

$$y(t+1) = f\big(y(t),u(t)\big) + \sigma\big(y(t)\big)e(t+1)$$

It should, however, be remarked that it is in general no easy problem to transform a non linear system to the form (2.2). This is, in fact, equivalent to solving the non-linear filter problem. It is therefore advantageous to directly model the non-linear system on the form (2.2), if possible.

## 2.2. Models.

In many cases the system characteristics, i.e. the function $g_S$ and the properties of $\{\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\}$ are not known a priori. One possibility to obtain a model of the system is to use input output data to determine the characteristics. In this report we will concentrate on the problem how the function $g_S$ can be found.

Naturally, it is impossible to find a general function $g_S\big(t;y(t),\ldots,y(1);u(t),\ldots,u(1);\mathcal{Y}_0\big)$. Therefore the class of functions among which g is sought must be restricted. We will call this set of functions the model set or the model structure. Let it be denoted by $\mathcal{M}$ and let the elements of the model set be indexed by a parameter vector $\theta$. The set over which $\theta$ varies will be denoted by $D_{\mathcal{M}}$. A certain element of $\mathcal{M}$ will be called a model and be denoted by $\mathcal{M}(\theta)$ or written as

$$E[y(t+1)\,|\,\mathcal{Y}_t,\mathcal{M}(\theta)] =$$

$$= g_{\mathcal{M}(\theta)}\big(y(t),\ldots,y(1);\ u(t),\ldots,u(1);\ \mathcal{Y}_0\big) \qquad (2.12)$$

Hence

$$\mathcal{M} = \{\mathcal{M}(\theta) \mid \theta \in D_{\mathcal{M}}\}$$

A complete model of the system also models the sequence $\{\varepsilon(t+1, \mathcal{Y}_t, \mathcal{S})\}$ so that it is described by

$$y(t+1) = E[y(t+1) \mid \mathcal{Y}_t, \mathcal{M}(\theta)] + \varepsilon(t+1, \mathcal{Y}_t, \mathcal{M}(\theta)) \qquad (2.13)$$

where $\{\varepsilon(t+1, \mathcal{Y}_t, \mathcal{M}(\theta))\}$ is a sequence of random variables with properties that depend on $\mathcal{M}(\theta)$.

For brevity, the notation

$$\hat{y}(t+1 \mid \theta) = E[y(t+1) \mid \mathcal{Y}_t, \mathcal{M}(\theta)]$$

is also used for the prediction.

The case when $D_{\mathcal{M}}$ is a finite set is treated separately in this report. In such a case $\mathcal{M}$ is called a finite model set, and then it will sometimes be denoted by $\mathcal{M}_F$.

The model structures can be chosen in a completely arbitrary way. For example, g can be expanded into orthogonal function systems:

$$g_{\mathcal{M}(\theta)} = \sum_{i=1}^{n} \theta_i f_i$$

Such choices are discussed by e.g. Lampard (1955). If there is no natural parametrization of the model, such an expansion may be advantageous. Tsypkin (1973) has discussed models of this type in connection with identification of non-linear systems. However, the usual choice is to take one of the models in Example 2.1 or 2.2 and introduce unknown elements $\theta_i$ into the system matrices.

A vector difference equation model, e.g., is then described by

$$A_{M(\theta)}(q^{-1})y(t) = B_{M(\theta)}(q^{-1})u(t) + C_{M(\theta)}(q^{-1})\varepsilon\big(t;M(\theta)\big) \quad (2.14)$$

where

$$A_{M(\theta)}(z) = I + A_{1,M(\theta)}z + \ldots + A_{n(\theta),M(\theta)}z^{n(\theta)}$$

etc.

$\{\varepsilon\big(t;M(\theta)\big)\}$ is a sequence of independent random variables with zero mean values and $E\varepsilon\big(t,M(\theta)\big)\varepsilon\big(t,M(\theta)\big)^T = \Lambda_{M(\theta)}$. The unknown elements may enter quite arbitrarily in the matrices $A_{i,M(\theta)}$. Some elements may be known from basic physical laws, or a priori fixed. Other elements may be related to each other etc. Generally speaking, $M$ can be described by the way the parameter vector $\theta$ enters in the matrices: the model parameterization.

## 2.3. Identification Criteria.

The purpose of the identification is to find a model $M(\theta)$ that in some sense suitably describes the measured input and output data.

The prediction of y(t+1) plays an important role for control. In, e.g., linear quadratic control theory, the optimal input shall be chosen so that $E[y(t+1)|Y_t,S]$ has desired behaviour. This is the separation theorem, see e.g. Åström (1970).

Therefore, it is very natural to choose a model that

$$h(A+B+C_\varepsilon) \geq h(A) + p(\delta)\,\mathrm{tr}\,B \quad \text{where } p(\delta) > 0 \tag{2.16b}$$

for tr $C_\varepsilon C_\varepsilon^T < \varepsilon_0$, where $\varepsilon_0$ depends only on $\delta$ and tr B. □

If h satisfies (2.16), it defines a well posed identification criterion by

$$V_N(\theta) = h[Q_N(\mathcal{M}(\theta))]$$

or (2.17)

$$V_N(\theta) = h\left[\frac{1}{N} Q_N(\mathcal{M}(\theta))\right]$$

In particular, h(A) will be taken as tr A, which clearly satisfies (2.16). This criterion is probably the easiest one to handle, theoretically as well as computationally. Then

$$\mathrm{tr}\, Q_N(\mathcal{M}(\theta)) = \sum_1^N \left|y(t) - \hat{y}(t|\theta)\right|^2_{R(t)}$$

where $|x|^2_{R(t)} = x^T R(t) x$.

Another possible choice is h(A) = det(A), which is of interest because of its relation to the likelihood function, cf. Section 2.4.

Lemma 2.1. h(A) = det(A) satisfies (2.16).

gives the best possible prediction. That is, some function of the prediction error

$$y(t+1) - E\big(y(t+1)\,|\,\mathcal{Y}_t, \mathcal{M}(\theta)\big)$$

should be minimized with respect to $\theta$.

We will consider the following class of criteria. Introduce the matrix

$$Q_N\big(\mathcal{M}(\theta)\big) = \sum_{t=1}^{N} [y(t) - \hat{y}(t|\theta)]R(t)[y(t) - \hat{y}(t|\theta)]^T \qquad (2.15)$$

Its dimension is $n_y \times n_y$, where $n_y$ is the number of outputs. $\{R(t)\}$ is a sequence of positive definite matrices. It is assumed that $\{|R(t)|\}$ is bounded. The selection of the matrices naturally effects the relative importance given to the components of the prediction. A special choice of weighting matrices is discussed in Section 2.4.

A scalar function, $h[Q_N(\mathcal{M}(\theta))]$, of the matrix of prediction errors will be minimized with respect to $\theta$. For the minimization to make sense, some simple properties of the function h must be introduced.

Properties of h. Let h have $n_y \times n_y$, symmetric matrices as domain. Assume that

$$h(\lambda A) = g(\lambda)h(A), \; \lambda, g(\lambda) \text{ scalars and } g(\lambda) > 0 \text{ for } \lambda > 0 \qquad (2.16a)$$

Let $\delta I < A < 1/\delta I$ be a symmetric positive definite matrix, and let B be symmetric, positive semidefinite and non zero. Assume that then

Proof. Condition (2.16a) is trivially satisfied.

$$\det(A+B+C_\varepsilon) = \det A^{1/2}\det\left(I + A^{-1/2}(B+C_\varepsilon)A^{-1/2}\right)\det A^{1/2} =$$

$$= \det A \prod_{i=1}^{n_y} (1+d_i)$$

where $d_i$ are the eigenvalues of $A^{-1/2}(B+C_\varepsilon)A^{-1/2}$.

Let $\lambda$ be the largest eigenvalue of B. Then $\lambda \geq \text{tr } B/n_y$. Also, $A^{-1/2}BA^{-1/2}$ has one eigenvalue that is larger or equal to $\lambda\delta$. (Consider $A^{-1/2}BA^{-1/2}x$, where $A^{-1/2}x$ is an eigenvector to B with eigenvalue $\lambda$.) Now, adding $C_\varepsilon$ to B can distort the eigenvalues at most $\varepsilon/\delta$ and

$$\prod_{i=1}^{n_y} (1+d_i) \geq \left[\prod_{i=1}^{n_y-1} (1-\varepsilon/\delta)\right] (1+\delta\lambda-\varepsilon/\delta) \geq$$

$$\geq \left(1-\frac{n_y\varepsilon}{\delta}\right)\left(1 + \delta\frac{\text{tr } B}{n_y} - \varepsilon/\delta\right) \geq 1 + \frac{\delta}{2n_y} \text{ tr } B$$

$$\text{for } \varepsilon < \frac{\delta \text{ tr } B}{2n_y\left(\frac{n_y}{\delta} + \text{tr } B\right)}$$

which concludes the proof. □

This report deals with the question when the estimate $\theta$ that minimizes (2.17) converges as N tends to infinity to values for which the model $M(\theta)$ coincides with the system $S$. This is the problem of consistency of prediction error identification methods.

To evaluate the criterion (2.17) it is required that the initial value $\mathcal{Y}_0$ is known. This value can be regarded as known, included in the parameter vector $\theta$, or simply assigned an arbitrary value, say zero. It will turn out that often the initial value cannot be estimated consistently. Therefore, as long as consistency of the model parameters is concerned, the initial state can as well be taken as zero, cf. e.g. Åström-Bohlin-Wensmark (1965).

So far we have only discussed how the function $E[y(t+1)|$ $|\mathcal{Y}_t, S]$ can be estimated. The properties of $\{\varepsilon(t+1,\mathcal{Y}_t,S)\}$ can then be estimated from the residuals

$$y(t+1) - E[y(t+1)|\mathcal{Y}_t, \mathcal{M}(\theta^*)] = \varepsilon\big(t+1;\mathcal{Y}_t, \mathcal{M}(\theta^*)\big)$$

where $\theta^*$ is the minimizing value. In particular, if $\{\varepsilon(t+1,\mathcal{Y}_t,S)\} = \{e(t+1)\}$ is a stationary sequence of independent random variables with zero mean values and we are only interested in the second order moment properties then $\Lambda = Ee(t)e^T(t)$ can be estimated as $\frac{1}{N} Q_N\big(\mathcal{M}(\theta^*)\big)$ where $Q_N$ is defined by (2.15) with $R(t) = I$.

## 2.4. Connection with Maximum Likelihood Estimation.

It is well known that prediction error criteria are intimately connected with maximum likelihood estimates. This section contains a brief discussion of how the formal relations can be established.

Consider the model (2.13)

$$y(t+1) = E\big(y(t+1)|\mathcal{Y}_t, \mathcal{M}(\theta)\big) + \varepsilon\big(t+1;\mathcal{M}(\theta)\big)$$

with

$$E\varepsilon\big(t;\mathcal{M}(\theta)\big)\varepsilon^T\big(t;\mathcal{M}(\theta)\big) = \hat{\Lambda}(t)$$

Let the innovations $\{\varepsilon\big(t,\mathcal{M}(\theta)\big)\}$ be assumed to be independent and normally distributed. The probability density of $y(t+1)$ given $\mathcal{Y}_t$ and given that (2.13) is true then is

$$f(x_{t+1}|\mathcal{Y}_t) = \frac{1}{\sqrt{2\pi \det \hat{\Lambda}(t+1)}} \cdot$$

$$\cdot\ e^{-[x_{t+1} - \hat{y}(t+1|\theta)]^T\hat{\Lambda}^{-1}(t+1)[x_{t+1} - \hat{y}(t+1|\theta)]}$$

Here $f(x|\mathcal{Y}_t) = F'(x|\mathcal{Y}_t)$ where $F(x|\mathcal{Y}_t) = P\big(y(t+1) \leq x|\mathcal{Y}_t\big)$.

Using Bayes' rule the joint probability density of $y(t+1)$ and $y(t)$ given $\mathcal{Y}_{t-1}$ can be expressed as

$$\begin{aligned}
f(x_{t+1},x_t|\mathcal{Y}_{t-1}) &= f\big(x_{t+1}|y(t) = x_t,\mathcal{Y}_{t-1}\big)f(x_t|\mathcal{Y}_{t-1}) = \\
&= f(x_{t+1}|\mathcal{Y}_t)f(x_t|\mathcal{Y}_{t-1}) = \\
&= [2\pi \det \hat{\Lambda}(t+1)\det \hat{\Lambda}(t)]^{-1/2} \cdot \\
&\quad \cdot \exp\Big\{- [x_{t+1} - \hat{y}(t+1|\theta)]^T\hat{\Lambda}^{-1}(t+1) \cdot \\
&\quad \cdot [x_{t+1} - \hat{y}(t+1|\theta)]\Big\} \cdot \\
&\quad \cdot \exp\Big\{- [x_t - \hat{y}(t|\theta)]^T\hat{\Lambda}^{-1}(t)[x_t - \hat{y}(t|\theta)]\Big\}
\end{aligned}$$

where $y(t)$ in $\hat{y}(t+1|\theta)$ should be replaced by $x_t$. In case $E\{y\big(t+1|\mathcal{Y}_t,\mathcal{M}(\theta)\big)\}$ does not depend linearly on $y(t)$, the distribution of $\big(y(t+1),y(t)\big)$ is not jointly normal.

Iteration directly gives the joint probability density of y(t+1),y(t),...,y(1) given $Y_0$. The logarithm of the likelihood function, given $Y_0$, then is obtained as

$$\log f\left(y(t+1),\dots,y(1)\,|\,Y_0\right) =$$

$$= - \sum_{s=0}^{t} [y(s+1) - \hat{y}(s+1|\theta)]^T \hat{\Lambda}^{-1}(s+1)[y(s+1) - \hat{y}(s+1|\theta)] -$$

$$- \frac{t}{2} \log 2\pi - \frac{1}{2} \sum_{s=1}^{t} \log \det \hat{\Lambda}(s+1)$$

The maximum likelihood estimate (MLE) of θ therefore is obtained as the element that minimizes

$$\sum_{s=1}^{t} [y(s+1) - \hat{y}(s+1|\theta)]^T \hat{\Lambda}^{-1}(s+1)[y(s+1) - \hat{y}(s+1|\theta)] +$$

$$+ \frac{1}{2} \sum_{s=1}^{t} \log \det \hat{\Lambda}(s+1)$$

If the matrices $\hat{\Lambda}(t)$ are known, the MLE is consequently obtained as the minimizing point of the loss function (2.17) with h(A) = tr(A) and $R(t) = \hat{\Lambda}^{-1}(t)$.

When $\hat{\Lambda}(t)$ are unknown, the minimization should be performed also with respect to $\{\hat{\Lambda}(s)\}$. In case $\hat{\Lambda}(t)$ does not depend on t, the minimization with respect to $\hat{\Lambda}$ can be performed analytically, Eaton (1967), yielding the problem to minimize $\det[Q_N(M(\theta))]$ giving θ(N) [where R(t) = I in $Q_N(M(\theta))$] and then take

$$\hat{\Lambda} = \frac{1}{N} Q_N\left(M(\theta(N))\right)$$

Summing up, the loss function (identification criterion) (2.17) which per se has good physical interpretation, also corresponds to the log likelihood function in the case of independent and normally distributed innovations. In the analysis, however, this will not be exploited. The results are therefore valid for general distributions of the innovations.

# 3. CONSISTENCY AND IDENTIFIABILITY.

The question of identifiability concerns the possibility to determine the characteristics of a system using input output data. This question is obviously closely related to the problem of consistency of the parameter estimate θ. A suitable way to connect the two concepts is introduced in this chapter.

The consistency of the parameter estimate θ depends on a variety of conditions, such as noise structure, choice of input signal, model parametrization etc. One specific problem is that there usually is parameter redundancy in the models. It was demonstrated in Examples 2.1 and 2.2 that several sets of matrices give the same input output relationships, and hence cannot be distinguished from each other from measurements of inputs and outputs.

Introduce the set

$$D_T(\mathcal{S},\mathcal{M}) = \left\{\theta \,\middle|\, \theta \in D_{\mathcal{M}},\ E[y(t+1)\,|\,\mathcal{Y}_t,\mathcal{M}(\theta)] = E[y(t+1)\,|\,\mathcal{Y}_t,\mathcal{S}] \right.$$
$$\left. \text{all } \mathcal{Y}_t,\ t > 0\right\} \qquad (3.1)$$

For example, with the model structure $\mathcal{M}$ given by (2.14) and the system $\mathcal{S}$ described by (2.4) the set is

$$D_T(\mathcal{S},\mathcal{M}) = \left\{\theta\,|\,C^{-1}_{\mathcal{M}(\theta)}(z)A_{\mathcal{M}(\theta)}(z) = C^{-1}_{\mathcal{S}}(z)A_{\mathcal{S}}(z)\right.$$

and

$$\left.C^{-1}_{\mathcal{M}(\theta)}(z)B_{\mathcal{M}(\theta)}(z) = C^{-1}_{\mathcal{S}}(z)B_{\mathcal{S}}(z) \quad \text{a.e.z.}\right\}$$

The set $D_T(\mathcal{S},\mathcal{M})$ clearly consists of all parameters, which give models with the "true" input output relationships. The system $\mathcal{S}$ and all models $\mathcal{M}(\theta)$, $\theta \in D_T(\mathcal{S},\mathcal{M})$

cannot be distinguished from input output data only.

Clearly, it is not meaningful to consider consistency if $D_T(S,M)$ is empty. Therefore, unless otherwise stated it will be assumed that $M$ is such that $D_T(S,M)$ is non empty. Naturally, this is a very strong assumption in practice, since it implies that the actual process can be modelled exactly. However, the theory of consistency does not concern approximation of systems, but convergence to "true" values. This question is further discussed in Section 4.1.

The estimate based on N data, $\theta(N)$, naturally depends on $S$ and $M$ and on the identification method used, $I$. It also depends on the experimental conditions, like the choice of input signals, possible feedback structures etc. The experimental conditions will be denoted by $X$. When needed, these dependences will be given as arguments.

Suppose now that

$$\theta(N) \rightarrow D_T(S,M) \quad \text{w.p.1 as } N \rightarrow \infty \tag{3.2}$$

Remark. By this is meant that

$$\inf_{\theta' \in D_T} |\theta(N) - \theta'| \rightarrow 0 \text{ with probability one as } N \rightarrow \infty$$

It does not imply that the estimate converges.

Then the models that are obtained from the identification all give the same input output characteristics as the true system. If we understand a system basically as an input output relation, it is natural to say that we have identified the system if (3.2) holds:

Definition 3.1. A system $S$ is said to be System Identifiable $[SI(M,I,X)]$ under given $M$, $I$, and $X$, if $\theta(N) \rightarrow$ $\rightarrow D_T(S,M)$ w.p.1 as $N \rightarrow \infty$.

If the objective of the identification is to obtain a model that can be used to design control laws, the concept of SI is quite adequate. Since all elements in $D_T(S,M)$ give the same input output relation, they also give equivalent feedback laws.

However, if the objective is to determine some parameters that have physical significance another concept is more natural:

Definition 3.2. A system $S$ is said to be Parameter Identifiable $[PI(M,I,X)]$ under given $M$, $I$, and $X$, if it is $SI(M,I,X)$ and $D_T(S,M)$ consists of only one point.

Remark. Parameter identifiability is the normal identifiability concept, and it has been used by several authors, see e.g. Åström-Bohlin (1965), Balakrishnan (1968), Bellman-Åström (1970), Tse-Anton (1972) and Glover-Willems (1973). Usually the system matrices are assumed to correspond to a certain parameter value $\theta^0$ for the given model parametrization. In such a case the parameter $\theta^0$ is said to be identifiable w.p.1 (or in probability) if there exists a sequence of estimates that tends to $\theta^0$ w.p.1 (or in probability). Now, the sequence of estimates converges to $\theta^0$ w.p.1 if and only if it is $PI(M,$ $I,X)$ according to Def. 3.2 and $D_T(S,M) = \{\theta^0\}$. Therefore the definition just cited is a special case of the Definition 3.2 above.

Clearly, a system $S$ can be PI$(M,I,X)$ only if $D_T(S,M) = \{\theta^0\}$. This means that there exists a one to one correspondence between the transfer function and the parameter vector $\theta^0$. This one to one correspondence can hold globally or locally around a given value. The terms global and local identifiability have been used for the two cases, see e.g. Bellman and Åström (1970). Definition 3.2 clearly corresponds to global parameter identifiability.

The problem to obtain such a one to one correspondence for linear systems falls in the field of canonical representation of transfer functions. This is a field that has received much attention. The special questions related to canonical forms for identification have been treated by e.g. Åström-Eykhoff (1971), Caines (1971), Mayne (1972) and Rissanen (1973).

From the above discussion we conclude that the problem of consistency and identifiability can be treated as three different problems:

I. First determine a set $D_I(S,M,I,X)$ such that

$$\theta(N) \to D_I(S,M,I,X) \quad \text{w.p.1 as } N \to \infty$$

This is a statistical problem. To find such a set, certain conditions, mainly on the noise structure of the system, must be imposed.

II. Then demand that

$$D_T(S,M) \supset D_I(S,M,I,X)$$

i.e. that $S$ is $\mathrm{SI}(M,I,X)$. This introduces requirements on the experimental conditions, $X$, choice of input signal, feedback structures etc.

III. If so desired, require that

$$D_T(S,M) = \{\theta^0\}$$

This is a condition on the model structure only, and for linear systems it is of algebraic nature.

In this report we will mainly treat problem I. In Chapter 4 $D_I$ is determined for finite model sets and in Chapter 5 the general case is considered.

Problem II is discussed in Chapter 6 for vector difference equations and state space models. In Gustavsson-Ljung-Söderström (1974) problem II is extensively treated for vector difference equation models.

Problem III is, as mentioned, the problem of canonical representation and can be treated separately from the identification problem. It will not be discussed in this report.

Remark. In the following, the arguments $S$, $M$, $I$, $X$ in $D_I$, $D_T$, SI and PI will be suppressed when there is no risk of ambiguity.

# 4. CONSISTENCY FOR FINITE MODEL SETS.

In this chapter a set $D_I(S,M_F,I,X)$ such that $\theta(N) \to$ $\to D_I(S,M_F,I,X)$ w.p.1 as $N \to \infty$ is determined for finite model sets $M_F$. The assumption of a finite model set implies a significant simplification of the convergence problem. The relevance of such finite model sets is discussed in Section 4.1. The main result is given in Section 4.2 and in Section 4.3 some applications of the theorem are discussed.

## 4.1. Relevance of Finite Model Sets.

The limitation to a large (say $100^{100}$) though finite number of possible parameter values naturally is of no practical importance. The restrictive assumption is that $D_T(S,M_F)$ is non-empty, i.e. that a true description of the system is available among the finitely many models. However, already when certain models are considered, like those in Examples 2.1 and 2.2 important idealization is introduced. No process in real life can be exactly described as a linear system (2.4) even if the parameter $\theta$ can be chosen from a continuum of values.

What does now consistency mean, if the real life system cannot be described within the chosen model structure? It should mean that the identification methods are tested on artificial systems which can be exactly described by, say, (2.4). The test can be theoretical as in consistency theorems, or experimental as when simulated systems are identified. The experimental tests are most often performed on digital computers. In these only a finite number of real numbers can be represented. Therefore the model that can be simulated may take parameter values only from a finite set. Also, the minimization of the loss function is performed over the same finite set of parameter values. The theoretical tests of this chapter are of the same nature.

## 4.2. Main Result.

In Chapter 3 it was described how consistency of parameter estimates can be shown in three steps. We will in this section treat the first of these steps, to find out which values are possible limits of the estimates in case the model set is finite.

In Caines (1973) the consistency problem for the maximum likelihood method is treated for the case of finitely many possible parameter values. When the prediction error method treated here coincides with the maximum likelihood method, Caines's results and the ones of this section partly overlap. The conclusion in Caines (1973) is, however, somewhat weaker, cf Example 4.1 below.

To find a suitable set $D_I$, some conditions have to be introduced. For finite model sets they are very weak indeed. It is merely assumed that the variance of the innovations is bounded. The innovations do not have to be independent. Notice in particular that the result is valid also when the system is controlled by any kind of feedback law. The closed loop system does not even have to be stable.

Theorem 4.1. Consider the system

$$y(t+1) = E[y(t+1)|\mathcal{Y}_t, S] + \varepsilon(t+1, \mathcal{Y}_t, S) \tag{4.1}$$

where

$$E[|\varepsilon(t+1, \mathcal{Y}_t, S)|^2 | \mathcal{Y}_t] < C$$

Consider a finite set of models $\mathcal{M}_F = \{\mathcal{M}(\theta)|\theta \in D_{\mathcal{M}}\}$, such that $D_T(S, \mathcal{M}_F)$ is non empty. Let $\theta(N)$ minimize the identification criterion tr $Q_N(\mathcal{M}_F(\theta))$ where $Q_N$ is defined by (2.15). Let $\tilde{D}_I^{(1)}$ be defined by

$$\tilde{D}_I^{(1)} = \left\{\theta \mid \theta \in D_{\mathcal{M}}, \quad \sum_1^{\infty} |\hat{y}(t|S) - \hat{y}(t|\theta)|^2_{R(t)} < \infty\right\} \tag{4.2}$$

Then $\theta(N) \to \tilde{D}_I^{(1)}$ w.p.1 as $N \to \infty$.

Remark. The random variable $\hat{y}(t|S) - \hat{y}(t|\theta)$ naturally depends on the realization $\omega$. The set $\tilde{D}_I^{(1)}$ may then also depend on $\omega$ under certain circumstances. Then the conclusion should be interpreted as

$$\theta(N,\omega) \to \tilde{D}_I^{(1)}(\omega) \text{ for a.e.} \omega \text{ as } N \to \infty$$

A common situation is that

$$\tilde{D}_I^{(1)}(\omega) = D_I^{(1)} \quad \text{w.p.1}$$

where $D_I^{(1)}$ does not depend on $\omega$. It then follows that

$$\theta(N) \to D_I^{(1)} \qquad \text{w.p.1} \qquad \text{as } N \to \infty$$

□

The proof of the theorem is given in the appendix.

To obtain the same conclusion for the general criterion (2.17), some additional conditions on the system are introduced.

Corollary. Consider a general criterion $h\left(Q_N(\mathcal{M}_F(\theta))\right)$, where h satisfies (2.16). Assume, in addition to the assumptions of the theorem, that

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{1}^{N} \left|\hat{y}(t|\mathcal{S}) - \hat{y}(t|\theta)\right|^2_{R(t)} < \infty \quad \text{w.p.1} \quad \text{all } \theta \in D_{\mathcal{M}},$$

that

$$E[\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})^T|\mathcal{Y}_t] > \delta I \quad \text{all } t,$$

and that

$$E[|\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})|^4|\mathcal{Y}_t] < C$$

Then $\theta(N) \to \tilde{D}_I^{(1)}$ w.p.1 as $N \to \infty$. □

The proof of the corollary is given in the appendix.

The assumption on finiteness of $\mathcal{M}_F$ is crucial to obtain uniform convergence in $\theta$ as N tends to infinity. It does not seem to be easy to extend the result to infinite model sets using the same approach, even if the problems seem to be merely technical.

## 4.3. Applications.

The set $\tilde{D}_I^{(1)}$ is analysed for some general model structures in Chapter 6. Here, we will just point out some specific consequences of the theorem.

Consistency w.p.1 will be shown for a simple example, for which previously given criteria all seem to fail, although the criterion function converges w.p.1. Thus it is not sufficient to analyse the asymptotic loss function

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1}^{N} |y(t) - \hat{y}(t|\theta)|^2$$

to obtain full information about the convergence of the estimates.

In Tse-Anton (1972) identifiability in terms of convergence in probability is discussed. The conditional probability density is used in a similar manner as the conditional mean is used in this report. To assure identifiability the conditional probability density must be "uniformly different" for different parameter values. Also, if the densities are the same for two parameter values the system is not identifiable.

In Caines (1973) a "prediction condition" (condition A2) is postulated, that requires, loosely speaking, that the predictions corresponding to different parameter values, be uniformly different infinitely often (with a probability arbitrarily close to 1).

Example 4.1. Consider the simple system

$$y(t+1) + ay(t) = bu(t) + e(t+1) \tag{4.3}$$

with the time varying feedback law

$$u(t) = f(t)y(t) \tag{4.4}$$

where $f(t) \to f$ as $t \to \infty$ and

$$\sum_{1}^{\infty} \left(f(t) - f\right)^2$$

diverges. Let the model set $M_F$ be

$$y(t+1) + \hat{a}y(t) = \hat{b}u(t) + \varepsilon(t+1)$$

where $\theta = (\hat{a} \quad \hat{b})$ can assume a finite number of values. Then

$$\hat{y}(t+1|\theta) = -\hat{a}y(t) + \hat{b}u(t)$$

and the limit of the loss function

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1}^{N} \left[y(t+1) - \hat{y}\left(t+1|M_F(\theta)\right)\right] \tag{4.5}$$

assumes the same value for $\hat{a} = a + \mu f$, $\hat{b} = b + \mu$, $-\infty < \mu < \infty$.

For the particular feedback (4.4), condition A2 in Caines (1973) is not satisfied, so that the consistence result there cannot be applied.

Also, the system (4.3) with feedback (4.4) does not satisfy any of Tse-Anton's (1972) conditions as cited above. The question of identifiability is therefore left undecided by their criteria.

However, the set $\tilde{D}_I^{(1)}$ is given by

$$\tilde{D}_I^{(1)} = \left\{(\hat{a},\hat{b}) \,\middle|\, \sum_1^\infty \left((a-\hat{a})y(t) - (b-\hat{b})u(t)\right)^2 < \infty\right\} =$$

$$= \left\{(\hat{a},\hat{b}) \,\middle|\, \sum_1^\infty \left((a-\hat{a}) - f(t)(b-\hat{b})\right)^2 y(t)^2 < \infty\right\}.$$

and hence

$$\tilde{D}_I^{(1)} = \{(a,b)\} \quad \text{w.p.1}$$

Theorem 4.1 in this report therefore states that a and b can be estimated consistently if f(t) approaches f so slowly that

$$\Sigma\left(f(t) - f\right)^2$$

diverges. If, on the other hand this sum converges, the estimates of a and b are not consistent.

The following numerical example may illuminate this conclusion. Let $a = -0.9$, $b = 1$ in (4.3) and let $\{e(t)\}$ be a sequence of independent N(0.1) random variables. Let in (4.4)

$$f(t) = (-0.9)t^{-\alpha}$$

where $\alpha = 0.25$ or $0.75$. In Fig. 4.1 the estimates based on 500, 5000 and 50000 data are shown for 10 different realizations of $\{e(t)\}$.

It is seen that for $\alpha = 0.25$ the estimates tend to the true values. For $\alpha = 0.75$ the estimate of a tends to -0.9, while the b estimates do not approach 1 as the number of data increases. In fact, in that case $\tilde{D}_I^{(1)}$ consists of $\{(-0.9, \mu), \mu \text{ arbitrary}\}$.

It can be shown that in either case the estimate of b converges w.p.1 to a random variable as the number of data tends to infinity.

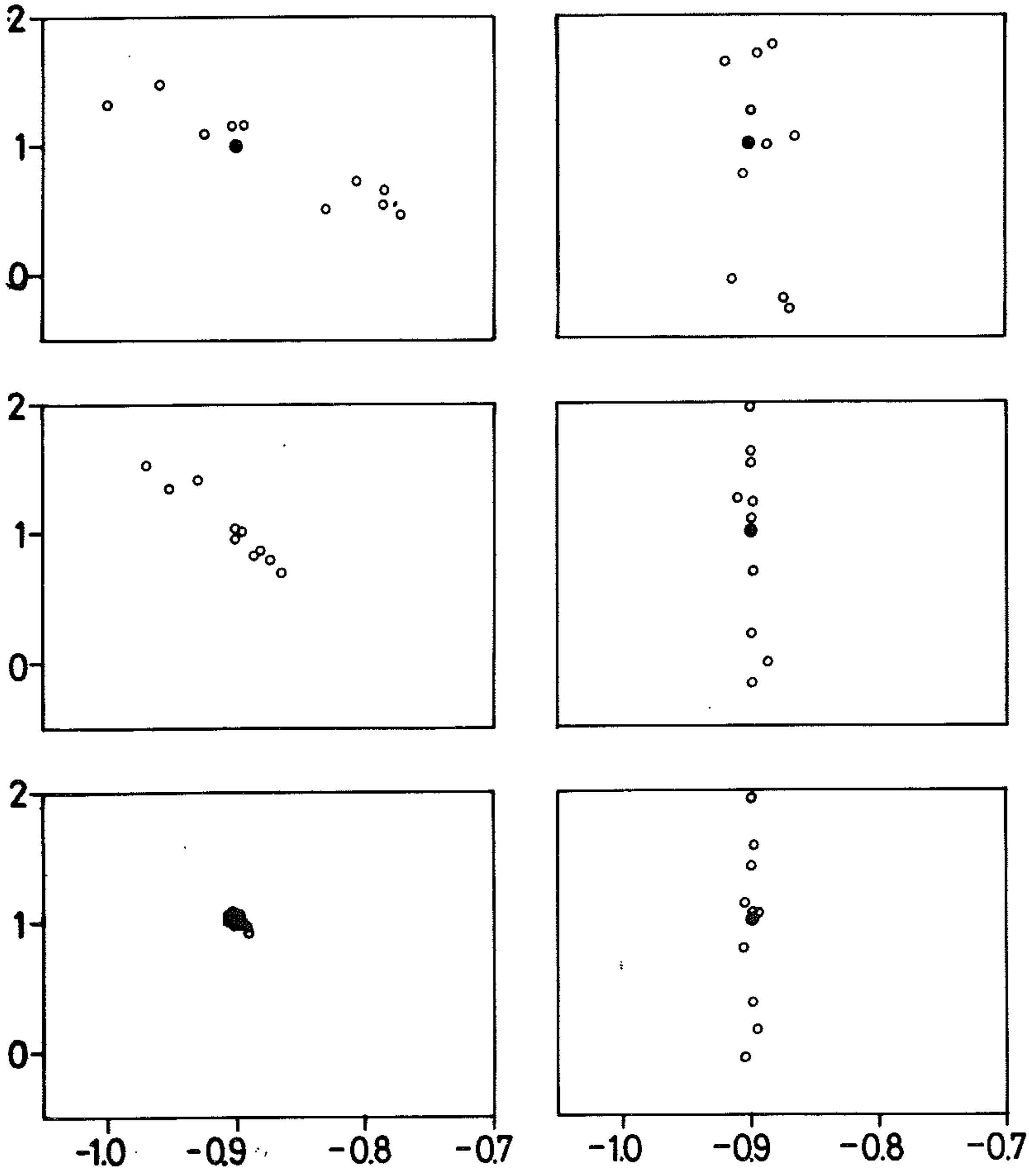

Fig. 4.1. - Identification results for the system (4.3) with regulator (4.4). Ten different realizations are shown. The left column shows the case $\alpha = 0.25$ and the right one $\alpha = 0.75$.

The number of data used are, from above, 500, 5000 and 50000 respectively.

## 5. CONSISTENCY FOR GENERAL MODEL STRUCTURES.

As was remarked in Section 4.2 the technical problems to prove consistency for infinite model sets are more difficult than for the case considered in Theorem 4.1. Mainly, this is due to the fact that uniform (in $\theta$) inequalities for the loss function must be established. In many of the previous works this problem has not been treated with sufficient care, and those proofs are, in fact, valid only for finite model sets. In this report the problem is overcome at the expense of restrictions on the noise and regularity conditions on the functions $E[y(t+1)|Y_t, M(\theta)]$.

The set into which the estimates converge will be shown to be

$$\tilde{D}_I^{(2)} = \left\{\theta \,\middle|\, \theta \in D_M \lim_{N\to\infty} \inf \frac{1}{N} \sum_1^N |\hat{y}(t+1|S) - \hat{y}(t+1)|\theta|^2_{R(t)} = 0\right\} \quad (5.1)$$

Clearly the set $\tilde{D}_I^{(1)}$, defined by (4.2), is a subset of $\tilde{D}_I^{(2)}$. Consequently the conclusion of Theorem 4.1 is sharper. However, if the behaviour of the system basically is stationary the two sets $\tilde{D}_I^{(1)}$ and $\tilde{D}_I^{(2)}$ coincide. Also, the additional conditions in this chapter are not very restrictive. They are satisfied for most model sets, and in particular for the models in Examples 2.1 and 2.2 under weak conditions on the feedback.

Like the set $\tilde{D}_I^{(1)}$, also $\tilde{D}_I^{(2)}$ may, under certain circumstances, depend on the realization $\omega$. In practice, it may be desirable to define a set into which the estimates converge that a priori is independent of the realization. It will be shown that under somewhat stronger conditions on the system and possible regulator, the estimates converge into

$$D_I^{(2)} = \left\{\theta \middle| \theta\in D_{\mathcal{M}} \lim_{N\to\infty} \inf \frac{1}{N}\sum_{1}^{N} E\left|\hat{y}(t\mid\mathcal{S}) - \hat{y}(t\mid\theta)\right|^2_{R(t)} = 0\right\} \tag{5.2}$$

These convergence results are shown in Section 5.1. The conditions that are imposed on the system are in Section 5.2 discussed for the special case of linear systems.

## 5.1. Main Results.

It will first be shown that the estimates converge into the set $\tilde{D}_I^{(2)}(\omega)$. This is achieved by a similar approach as in the proof of Theorem 4.1. An additional condition on the prediction $\hat{y}(t|\theta)$ as a function $\theta$ has to be introduced.

Theorem 5.1. Consider the system

$$y(t+1) = E[y(t+1)\mid \mathcal{Y}_t,\mathcal{S}] + \varepsilon(t+1,\mathcal{Y}_t,\mathcal{S}) \tag{5.3}$$

where $E[|\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})|^4 | \mathcal{Y}_t] < C$.

Consider a set of models $\mathcal{M}$, such that $D_T(\mathcal{S},\mathcal{M})$ is non empty. Let $\theta(N)$ minimize the identification criterion $V_N(\theta) = \mathrm{tr}[\frac{1}{N} Q_N(\mathcal{M}(\theta))]$, over a compact set $D_{\mathcal{M}}$. Let $\tilde{D}_I^{(2)}(\omega)$ be defined by (5.1). Suppose that

$$z(t) = \sup_{\theta\in D'_{\mathcal{M}}} \max_{1\le i\le n_y} \left|\frac{\partial}{\partial\theta}\hat{y}^{(i)}(t|\theta)\right| \quad \text{((i) denotes i:th row)}$$

where $D'_{\mathcal{M}}$ is a connected, open set that contains $D_{\mathcal{M}}$, satisfies the following condition

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{1}^{N} z(t)^2 < \infty \qquad \text{w.p.1} \tag{5.4}$$

Then the estimate

$$\theta(N,\omega) \to \tilde{D}_I^{(2)}(\omega) \quad \text{a.e.} \quad \text{as } N \to \infty$$

The same results hold for the general criterion (2.17) if, in addition to the above assumptions,

$$E[\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})^T|\mathcal{Y}_t] \geq \delta I \quad \text{all t.} \tag{5.5}$$

The proof of this theorem is given in the appendix.

□

Remark. The assumption on connectedness of $D_{\mathcal{M}}'$ is not essential. Another possibility is to take the supremum over $D_{\mathcal{M}}$ and to assume that

$$\limsup_{N\to\infty} V_N(\theta) < \infty \text{ w.p.1 for } \theta \in D_{\mathcal{M}}$$

□

To apply the theorem, condition (5.4) has to be satisfied. In the next section simple and weak conditions that imply (5.4) are derived for linear models.

If convergence into a set that does not depend on ω is desired, this can be achieved by showing that

$$\tilde{D}_I^{(2)}(\omega) = D_I^{(3)} \quad \text{w.p.1} \quad \text{or } \tilde{D}_I^{(2)}(\omega) \subset D_I^{(3)} \quad \text{w.p.1}$$

Then $\theta(N) \to D_I^{(3)}$ w.p.1 as $N \to \infty$.

It is probably of more practical interest to show that the estimates converge into the set $D_I^{(2)}$ defined by (5.2). To achieve this result, some additional conditions must be introduced. They essentially secure that

$$D_I^{(2)} = \tilde{D}_I^{(2)}(\omega) \quad \text{a.e.} \tag{5.6}$$

by restricting the dependence between events that occur at long time differences. In contrast to (5.4) they involve only conditions on the second moments of certain variables.

Theorem 5.2. Consider the system (5.3)

$$y(t+1) = E[y(t+1)|\mathcal{Y}_t, \mathcal{S}] + \varepsilon(t+1, \mathcal{Y}_t, \mathcal{S})$$

Consider a set of models $\mathcal{M}$ such that $D_T(\mathcal{S}, \mathcal{M})$ is non empty. Let $\theta(N)$ minimize the identification criterion $V_N(\theta) = \mathrm{tr}[\frac{1}{N} Q_N(\mathcal{M}(\theta))]$ over a compact set $D_{\mathcal{M}}$. Let $D_I^{(2)}$ be defined by (5.2). Let $\theta^*$ be an arbitrary element of $D_{\mathcal{M}}$, and introduce

$$\bar{B} = \bar{B}(\theta^*, \rho) = \left\{\theta \middle| |\theta^* - \theta| \leq \rho\right\} \qquad \rho > 0$$

and

$$\eta(t, \theta^*, \rho) = \inf_{\theta \in \bar{B}} \left\{[y(t) - \hat{y}(t|\theta)]^T R(t) [y(t) - \hat{y}(t|\theta)]\right\}$$

Assume that for all $\theta^* \in D_{\mathcal{M}}$

a) $$E\left\{\sup_{\theta \in \bar{B}} \max_i \left|\frac{\partial}{\partial \theta} \hat{y}^{(i)}(t|\theta)\right|^2\right\} < C(\theta^*) \text{ for all } t$$

and some $\rho = \rho_1(\theta^*) > 0$ (5.7)

b) $\quad \mathrm{Cov}\left[\eta(t,\theta^*,\rho),\ \eta(t+s,\theta^*,\rho)\right] < K(1+|s|^{-\alpha})$

$$\alpha > 0;\ \text{all } t,s \text{ and } 0 \leq \rho \leq \rho_2^* \qquad (5.8)$$

$K$, $\alpha$ and $\rho_2^*$ may depend on $\theta^*$.

All expectations, including that in (5.2), is over the sequence of innovations $\{\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\}$.

Then $\theta(N) \to D_I^{(2)}$ w.p.1 as $N \to \infty$.

Proof. This theorem could be proved using Theorem 5.1 by showing that (5.7) and (5.8) imply (5.4) and (5.6). However, a separate proof, which is completely different from the proof of Theorem 5.1 will be given. It shows how Wald's (1949) classical proof of consistency of ML estimates for independent observations can be modified to be valid for the present case. The proof is given in the appendix.

□

The reason why the limit inferior is used in the definition of $\tilde{D}_I^{(2)}$ and $D_I^{(2)}$ is that no restrictions on time varying components should be introduced. The limit of the sum may very well fail to exist. In particular, when an adaptive controller is used, it would require a priori knowledge of the overall system behaviour to state that the limit exists.

To apply Theorem 5.2, conditions (5.7) and (5.8) must be checked. This requires some analysis of the model structures, which is the price for the general formulation of the theorem. For the common case with linear models, the conditions are discussed in the next section.

## 5.2. Linear Models.

A model $M(\theta)$ is linear if

$$E[y(t+1)\,|\,Y_t, M(\theta)]$$

is a linear function of y(s), u(s), $s \leq t$. The models discussed in Examples 2.1 and 2.2 clearly are linear. If $M$ is a set of linear models, the conditions (5.4), (5.7) and (5.8) of Theorems 5.1 and 5.2 follow from criteria, which are easily checked.

Consider first a simple example. Let the system $S$ be defined by

$$y(t+1) + ay(t) = bu(t) + e(t+1) + ce(t) \tag{5.9}$$

where $\{e(t)\}$ is white noise. Let the set of models be defined by

$$y(t+1) + \hat{a}y(t) = \hat{b}u(t) + \varepsilon(t+1) + \hat{c}\varepsilon(t) \tag{5.10}$$

where $\{\varepsilon(t)\}$ is white noise.

The parameter vector $\theta = [\hat{a}\ \hat{b}\ \hat{c}]^T$ is to be estimated. As found in Example 2.1, $\hat{y}(t+1|\theta)$ is recursively defined by

$$(1+\hat{c}q^{-1})\hat{y}(t+1|\theta) = (\hat{a}-\hat{c})y(t) + \hat{b}u(t) \tag{5.11}$$

where $q^{-1}$ is the backward shift operator.

The derivatives $\frac{\partial}{\partial\theta}\hat{y}(t+1|\theta)$ are straightforwardly found as the solutions of

$$(1+\hat{c}q^{-1})\ \frac{\partial}{\partial\hat{a}}\ \hat{y}(t+1|\theta) = y(t)$$

$$(1+\hat{c}q^{-1})\ \frac{\partial}{\partial\hat{b}}\ \hat{y}(t+1|\theta) = u(t) \tag{5.12}$$

$$(1+\hat{c}q^{-1})\ \frac{\partial}{\partial\hat{c}}\ \hat{y}(t+1|\theta) = -\ y(t)\ -\ \hat{y}(t|\theta)$$

As long as the set $M$ contains only models with stable C-polynomials, i.e. $|\hat{c}| \leq 1 - \delta$, clearly $\frac{\partial}{\partial\theta}\ \hat{y}(t+1|\theta)$ is a well defined variable.

Consider now $\frac{\partial}{\partial\hat{a}}\ \hat{y}(t|\theta)$. This variable depends only on $\hat{c}$, and let it be denoted by $\hat{y}_{\hat{a}}(t,\hat{c})$. Consider

$$\sup_{\theta\in D_M}\left|\frac{\partial}{\partial\hat{a}}\ \hat{y}\left(t|M(\theta)\right)\right| = \sup_{|\hat{c}|\leq 1-\delta}|\hat{y}_{\hat{a}}(t,\hat{c})|$$

The supremum means that for each realization $\omega$, the variable $\hat{c} = \hat{c}(\omega)$ shall be chosen such that the real number $|\hat{y}_{\hat{a}}\left(t,\hat{c}(\omega),\omega\right)|$ is maximized. From (5.12)

$$\left|\hat{y}_{\hat{a}}(t,\hat{c},\omega)\right| = \left|\sum_{k=0}^{\infty}(-\hat{c})^k y(t-k)\right| \leq \sum_{k=0}^{\infty}|\hat{c}|^k|y(t-k)| \leq$$

$$\leq \sum_{k=0}^{\infty}(1-\delta)^k|y(t-k)| \quad \text{for } \theta\in D_M \tag{5.13}$$

Consequently the variable

$$\sup_{|\hat{c}|<1-\delta}\left|\hat{y}_{\hat{a}}(t,\hat{c})\right|$$

can be obtained from $|y(s)|$ by exponentially stable filtering. It is therefore reasonable to assume that condition (5.4) in Theorem 5.1 is satisfied if $\{|y(s)|\}$ satisfies a suitable regularity condition. In fact, we have the following lemma.

Lemma 5.1. Consider the system (2.2) and the model set (2.13). Suppose that $\hat{y}(t|\theta)$ is linear in $y(s)$ and $u(s)$, $0 \leq s < t$ and in $y_0$, i.e.

$$\hat{y}(t|\theta) = \sum_{k=1}^{t} h_{k,t} y(t-k) + \sum_{k=1}^{t} f_{k,t} u(t-k) + H_t y_0$$

Suppose that the linear filter that defines $\frac{\partial}{\partial\theta} \hat{y}^{(i)}(t|\theta)$ is exponentially stable for each component i of $\hat{y}(t|\theta)$ and for $\theta \in D_M$, i.e.

$$\left|\frac{\partial}{\partial\theta} h_{k,t}^{(i)}(\theta)\right| < C\lambda^k, \quad \left|\frac{\partial}{\partial\theta} f_{k,t}^{(i)}(\theta)\right| < C\lambda^k \quad \left|\frac{\partial}{\partial\theta} H_t^{(i)}(\theta)\right| < C\lambda^t$$

for some $\lambda < 1$, all t, all $\theta \in D_M$ and $i = 1,\ldots,n_y$. $|\cdot|$ denotes the operator norm of the matrices, and $h_{k,t}^{(i)}$ denotes the i:th row of the matrix $h_{k,t}$. Let

$$z(t) = \sup_{\theta \in D_M} \max_{1 \leq i \leq n_y} \left|\frac{\partial}{\partial\theta} \hat{y}(t|\theta)\right|$$

and assume that

$$\lim_{N\to\infty} \sup \frac{1}{N} \sum_{1}^{N} [y(t)^2 + u(t)^2] < \infty \text{ w.p.1} \tag{5.14}$$

Then

$$\lim_{N\to\infty} \sup \frac{1}{N} \sum_{1}^{N} z(t)^2 < \infty \quad \text{w.p.1}$$

i.e. condition (5.4) holds.

The proof is given in the appendix.

□

Condition (5.14) is a simple and most reasonable condition on the overall behaviour of the system to be identified. The restriction of $D_M$ to parameters that give exponentially stable predictors (or rather derivatives) is also natural. Therefore, Lemma 5.1 and Theorem 5.1 yield a consistency result for linear systems that is valid under very weak conditions.

To check conditions (5.7) and (5.8) of Theorem 5.2, consider again the simple system (5.9) with model set (5.10). From (5.13) we obtain

$$E \sup_{\theta \in D_M} |\hat{y}_{\hat{a}}(t,\hat{c})|^2 \leq \sum_{s,k=0}^{\infty} (1-\delta)^{k+s} E|y(t-k)||y(t-s)| \leq K \frac{1}{\delta^2}$$

$$\text{if } Ey(s)^2 \leq K$$

Consequently, condition (5.7) of Theorem 5.2 is satisfied for this simple system.

Condition (5.8) means that the two random variables $\eta(t,\theta^*,\rho)$ and $\eta(t-s,\theta^*,\rho)$ are almost uncorrelated for large s, where

$$\eta(t,\theta^*,\rho) = \inf_{\theta \in B(\theta^*,\rho)} [y(t) - \hat{y}(t|\theta)]^2 =$$

$$= \inf_{\theta \in B} \left\{ y(t) - \sum_{k=0}^{\infty} (-\hat{c})^k [(\hat{a}-\hat{c})y(t-k-1) - \hat{b}u(t-k-1)] \right\}^2$$

If $|c^*| + \rho < 1$, the terms corresponding to large k, say $k \geq N$, in the sum, will have small influence on the variable $\eta(t,\theta^*,\rho)$. Furthermore, the variable $\eta(t-N,\theta^*,\rho)$ is defined using variables that have quite small influence on $\eta(t,\theta^*,\rho)$. Therefore, if y(t) and y(t-N) are almost uncorrelated, it should follow that condition (5.8) of Theorem 5.2 should be satisfied.

This requires some conditions on the behaviour of the closed loop system. We choose to require that the closed loop system is exponentially stable with e(t) regarded as input and y(t) as output. If part of the input is determined as linear output feedback, then the closed loop system is linear, i.e. y(t) is a linear function of e(s) and exponential stability is well defined.

If the feedback is non linear, the closed loop system is also non linear. Then the usual definition of exponential stability of ODEs , see e.g. Hahn (1967), can be used after suitable modifications to include stochastic disturbances.

Definition 5.1. Consider the linear system

$$y(t+1) = Ey(t+1|\mathcal{Y}_t,\mathcal{S}) + e(t+1)$$

where e(t) are independent random variables, and where part of the input u(t) is determined as (non linear) output feedback. Let the system and regulator be started up at time t-N, with zero initial conditions, yielding at time t the outputs and inputs, $y_N^0(t)$ and $u_N^0(t)$ respectively. Suppose that

$$|y(t) - y_N^0(t)| \leq C(\mathcal{Y}_{t-N})\lambda^N, \quad |u(t) - u_N^0(t)| \leq C(\mathcal{Y}_{t-N})\lambda^N;$$

some $\lambda < 1$, where $C(\mathcal{Y}_{t-N})$ is a scalar function of $\mathcal{Y}_{t-N}$, such that $EC(\mathcal{Y}_{t-N})^4 < C$.

Then the closed loop system is said to be *exponentially stable*.

□

*Lemma 5.2*. Consider the system (2.2) and the model set (2.13). Suppose that $E[y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta)]$ is linear in y(s) and u(s), i.e.

$$E[y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta)] = \sum_{k=0}^{\infty} h_{k,t}(\theta)y(t-k) +$$

$$+ \sum_{k=0}^{\infty} f_{k,t}(\theta)u(t-k) + H_t(\theta)\mathcal{Y}_0$$

Suppose that the linear filters that define $E[y(t+1|\mathcal{Y}_t,\mathcal{M}(\theta)]$ and $\frac{\partial}{\partial\theta} Ey\left(t+1|\mathcal{Y}_t,\mathcal{M}(\theta)\right)$ are exponentially stable for $\theta \in D_{\mathcal{M}}$, i.e.

$$\left|h_{k,t}(\theta)\right| \leq C\lambda^k, \quad \left|f_{k,t}(\theta)\right| \leq C\lambda^k \quad \left|H_t(\theta)\right| \leqslant C\lambda^t$$

$$\left|\frac{\partial}{\partial\theta} h_{k,t}^{(i)}(\theta)\right| \leq C\lambda^k \quad \left|\frac{\partial}{\partial\theta} f_{k,t}^{(i)}(\theta)\right| \leqslant C\lambda^t$$

$$\left|\frac{\partial}{\partial\theta} H_t^{(i)}(\theta)\right| \leq C\lambda^t$$

for some $\lambda < 1$ for all t and for all $\theta \in D_{\mathcal{M}}$, and all rows $i = 1,\ldots,n_y$.

Assume further that

$$Ey(t)^4 \leq C \quad \text{and} \quad Eu(t)^4 \leq C \quad .$$

and that the closed loop system is exponentially stable. Assume that the innovations $\{\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\} = \{e(t+1)\}$ are independent random variables. Then conditions (5.7) and (5.8) of Theorem 5.2, and condition (5.14) of Lemma 5.1, are satisfied.

The proof is given in the appendix.

□

In the next chapter, the lemmas will be applied to the linear models defined in Examples 2.1 and 2.2, and together with Theorems 5.1 and 5.2 this will give the desired consistency results. Naturally, the results obtained are applicable also to other more general systems.

It follows from the proof of Lemma 5.2 (Eq. (A.24)) that if the initial value is contained in the parameter vector $\theta$ to be estimated; $\mathcal{Y}_0 = \mathcal{Y}_0(\theta)$, then all $\theta$ such that

$$E[y(t+1)|\mathcal{Y}_t,\mathcal{S},\mathcal{Y}_0 = 0] = E[y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta),\mathcal{Y}_0 = 0]$$

belong to $D_I^{(2)}$.

Consequently, all $\theta$ that differ from the "correct" values only in $\mathcal{Y}_0(\theta)$ belong to $D_I^{(2)}$. It is also quite easy to show that under the assumptions of Lemma 5.1, the initial values in fact cannot be estimated consistently.

# 6. EXAMPLES.

In this chapter the results of Chapters 4 and 5 are applied to vector difference equation (VDE) models and to state space models. The set of limit points, $D_I$, is determined for these models. The particular identifiability concepts that were introduced in Chapter 3 are also applied and special attention is paid to various feedback configurations for VDEs. This is treated in Section 6.1, while state space models are considered in Section 6.2.

## 6.1. Vector Difference Equations.

Consider the VDE description of systems given in Example 2.1. Assume that the system is described by

$$A_S(q^{-1})y(t) = B_S(q^{-1})u(t) + C_S(q^{-1})e(t) \qquad (6.1)$$

Let the input to the system, u(t) be of quite general form:

$$u(t) = f_t\left(y(t),\ldots,y(0);u(t-1),\ldots,u(0)\right) + u_r(t) + w(t) \qquad (6.2)$$

where $u_r(t)$ is a measurable signal that is independent of y(s), u(s), and where w(t) is obtained from a noise source, that is independent of {e(t)}. The function $f_t$ may be unknown to the experiment designer. The set of models, $M$, is given by

$$A_{M(\theta)}(q^{-1})y(t) = B_{M(\theta)}(q^{-1})u(t) + C_{M(\theta)}(q^{-1})\varepsilon(t) \qquad (6.3)$$

where $\theta \in D_M$ which is assumed to be compact.

Finite model sets.

Suppose first that $D_{\mathcal{M}} = D_{\mathcal{M}_F}$ is a finite set, and denote the corresponding model set by $\mathcal{M}_F$. Then Theorem 4.1 can be applied, and the set of possible limit points of $\theta(N)$ is obtained as

$$\tilde{D}_I^{(1)} = \left\{\theta \,\middle|\, \theta \in D_{\mathcal{M}_F} \sum_1^{\infty} \left|\hat{y}(t|\mathcal{S}) - \hat{y}(t|\mathcal{M}(\theta))\right|_{R(t)}^2 < \infty\right\}$$

Introduce for convenience the notation

$$A = A_{\mathcal{S}}(q^{-1}) \qquad \hat{A} = A_{\mathcal{M}(\theta)}(q^{-1}) \quad \text{etc.}$$

Then according to (2.6)

$$\hat{y}(t|\mathcal{S}) = [I - C^{-1}A]y(t) + C^{-1}Bu(t)$$

$$\hat{y}(t|\theta) = [I - \hat{C}^{-1}\hat{A}]y(t) + \hat{C}^{-1}\hat{B}u(t)$$

which gives

$$\tilde{D}_I^{(1)} = \left\{\theta \,\middle|\, \theta \in D_{\mathcal{M}_F} \sum_1^{\infty} \left| (\hat{C}^{-1}\hat{A} - C^{-1}A)y(t) + \right.\right.$$

$$\left.\left. + (C^{-1}B - \hat{C}^{-1}\hat{B})u(t)\right|_{R(t)}^2 < \infty \right\}$$

With the terminology introduced in Chapter 3, the system $\mathcal{S}$ is System Identifiable under $\mathcal{M}$, $\mathcal{I}$, and $\mathcal{X}$ $\left(SI(\mathcal{M}, \mathcal{I}, \mathcal{X})\right)$ if $\tilde{D}_I^{(1)} \subset D_T$ w.p.1. Suppose that the system is not $SI(\mathcal{M},\mathcal{I},\mathcal{X})$. Then there exists a $\tilde{\theta}$, such that $\tilde{\theta} \in \tilde{D}_I^{(1)}$ and $\tilde{\theta} \notin D_T$. Introduce

$$\tilde{L} = L(\tilde{\theta}) = C^{-1}_{M(\tilde{\theta})}A_{M(\tilde{\theta})} - C_S^{-1}A_S$$

$$\tilde{M} = M(\tilde{\theta}) = C^{-1}_{M(\tilde{\theta})}B_{M(\tilde{\theta})} - C_S^{-1}B_S$$

Since $\tilde{\theta} \notin D_T$ at least one of $\tilde{L}$ and $\tilde{M}$ is non zero. Since $\tilde{\theta} \in \tilde{D}_I^{(1)}$ there exist w.p.p. a relationship between y and u such that

$$\sum_1^\infty |\tilde{L}y(t) - \tilde{M}u(t)|^2 < \infty$$

which in particular implies that

$$\tilde{L}y(t) = \tilde{M}u(t) \tag{6.5}$$

asymptotically as t tends to infinity. Furthermore, (6.5) holds for each realization such that $\tilde{\theta}$ is a possible limit point. The conclusion is that if the experimental conditions $X$ are such that no exact linear relationships (6.5) hold as t tends to infinity, then the system $S$ is SI$(M,I,X)$.

Example 6.1. Adaptive regulators.

Suppose that the feedback regulator is given by

$$F(\theta(t),q^{-1})u(t) = G(\theta(t),q^{-1})y(t) \tag{6.6}$$

That is, the regulator is linear and its parameters are determined from the current system parameter estimates. This is a simple adaptive controller, based on the assumption that the control can be separated from the estimation. Such a regulator is considered e.g. by Kalman (1958) and by Åström-Wittenmark (1971) in the single input output case and in case $C_S = C_{M(\theta)} = I$.

Suppose that $D_T(S,M_F)$ is non empty and that for all $\theta \in D_T(S,M_F)$

$$F(\theta,z) = F^0(z) \quad \text{a.e. } z$$

$$G(\theta,z) = G^0(z) \quad \text{a.e. } z$$

Assume that for some realization $\omega$, $\tilde{D}_I^{(1)}(\omega)$ is not contained in $D_T$. Then there exists a $\tilde{\theta}$ such that (6.5) holds as t tends to infinity. This means that the regulator (6.6) asymptotically is equivalent to a linear constant one; (6.5). For realizations such that $\tilde{D}_I^{(1)}(\omega)$ is contained in $D_T$, $F\left(\theta(t),z\right) \rightarrow F^0(z)$ and $G\left(\theta(t),z\right) \rightarrow G^0(z)$ as t tends to infinity. Consequently the adaptive regulator (6.6) always converges to a (regulator that is equivalent to) constant linear regulator. Such a piece of information is often valuable, since many results rely upon the assumption of convergence, cf. Åström-Wittenmark (1973). □

Infinite model sets.

Suppose now that $D_M$ is a general compact set, such that

$\theta \in D_M \Rightarrow C_{M(\theta)}(z)$ is stable.

The prediction $\hat{y}(t+1|\theta)$ is calculated from (2.7)

$$\hat{C}\hat{y}(t|\theta) = (\hat{C}-\hat{A})y(t) + \hat{B}u(t) \qquad (6.7)$$

The derivative $\frac{\partial}{\partial\theta}\,\hat{y}(t|\theta)$ is defined by

$$\hat{C}\,\frac{\partial}{\partial\theta}\,\hat{y}(t|\theta) = \left[\frac{\partial}{\partial\theta}\,\hat{C} - \frac{\partial}{\partial\theta}\,\hat{A}\right]y(t) + \frac{\partial}{\partial\theta}\,Bu(t) - $$

$$- \left[\frac{\partial}{\partial\theta}\,\hat{C}\right]\hat{y}(t|\theta) \tag{6.8}$$

Since $\hat{C} = C_{M(\theta)}(z)$ is stable, the linear difference equations (6.7) and (6.8) are exponentially stable, uniformly in $\theta \in D_M$.

Lemma 5.1 and Theorem 5.1 now give the following important result:

Theorem 6.1. Consider the system $S$ described by the vector difference equation (6.1). Let the model set be given by (6.3) and assume that the matrix elements are continuously differentiable functions of the parameter $\theta$. Assume that $D_M$ is compact and that $C_{M(\theta)}(z)$ is stable for all $\theta \in D_M$. Assume that

- o The fourth moments of e(t), y(t) and u(t) are uniformly bounded.
- o The closed loop system is exponentially stable (cf. Def. 5.1).

Then the estimate $\theta(N)$ that minimizes the criterion (2.17) tends w.p.1 to the set

$$D_I^{(2)} = \left\{\theta \,\middle|\, \theta \in D_M,\ \lim_{N\to\infty}\inf \frac{1}{N}\sum_{1}^{N} E\left|K_{M(\theta)}(q^{-1})y(t) - \right.\right.$$

$$\left.\left. - L_{M(\theta)}(q^{-1})u(t)\right|^2 = 0\right\}$$

as N tends to infinity.

Here

$$K_{M(\theta)}(q^{-1}) = C^{-1}_{M(\theta)}(q^{-1})A_{M(\theta)}(q^{-1}) - C^{-1}_{S}(q^{-1})A_{S}(q^{-1})$$

and

$$L_{M(\theta)}(q^{-1}) = C^{-1}_{M(\theta)}(q^{-1})B_{M(\theta)}(q^{-1}) - C^{-1}_{S}(q^{-1})B_{S}(q^{-1})$$

□

The condition that the closed loop system is exponentially stable is quite mild and very natural. Obvious examples are open loop systems with stable $A_S(q^{-1})$ or systems with linear feedback

$$F(q^{-1})u(t) = G(q^{-1})y(t)$$

such that the determinant of

$$A(z) - B(z)\left(F(z)\right)^{-1}G(z)$$

has all zeros outside the unit circle. (The determinant is a rational function of z.)

Example 6.2. Let the input be given by

$$u(t) = R(q^{-1})y(t) + v(t)$$

where $R(z) = [F(z)]^{-1}G(z)$ is such that $\det[A(z)-B(z)R(z)]$ has all zeroes outside the unit circle. The sequence $\{v(t)\}$ does not depend on $\{e(t)\}$. Then

$$y(t) = H(q^{-1})e(t) + \tilde{H}(q^{-1})v(t)$$

where

$$H(z) = [A(z) - B(z)R(z)]^{-1}C(z)$$

$$\tilde{H}(z) = [A(z) - B(z)R(z)]^{-1}B(z)$$

are stable linear filters. Assume that $E|e(t)|^4 < C$. Then the conditions of Theorem 6.1 are satisfied.

Consider with K and L as in Theorem 6.1

$$E|Ky - Lu|^2 = E|(KH-LRH)e + (K\tilde{H}-LR\tilde{H})v + Lv|^2 =$$

$$= E|(K-LR)He|^2 + |(K-LR)\tilde{H}v + Lv|^2$$

since $Ee = 0$ and $\{v(t)\}$ is independent of $\{e(t)\}$. Suppose $\theta^* \in D_I$, and denote $K^* = K_{\mathcal{M}(\theta^*)}$, $L^* = L_{\mathcal{M}(\theta^*)}$. Then

$$0 = \liminf_{N\to\infty} \frac{1}{N}\sum_1^N E|K^*y - L^*u|^2 \geq$$

$$\liminf_{N\to\infty} \frac{1}{N}\sum_1^N E|(K^*-L^*R)He|^2$$

which is possible only if $(K^*-L^*R)H = 0$ or $K^*-L^*R = 0$. Hence

$$0 = \liminf_{N\to\infty} \frac{1}{N}\sum_1^N E|K^*y - L^*u|^2 \geq \liminf_{N\to\infty} \frac{1}{N}\sum_1^N |L^*v|^2$$

Suppose now that $v(t)$ is persistently exciting of sufficiently high order, cf Mayne (1972). Then $L^* = 0$ follows, which in turn implies $K^* = 0$. Consequently, $\theta^* \in \in D_T$ and the system is $SI(\mathcal{M},\mathcal{I},\mathcal{X})$ for this choice of input signal.

In Gustavsson-Ljung-Söderström (1974) the identifiability properties for a number of different feedback configurations are discussed in detail.

## 6.2. State Space Models.

Consider the state space representation (2.11) in Example 2.2, i.e.

$$\hat{x}(t+1) = A_{\mathcal{S}}\hat{x}(t) + B_{\mathcal{S}}u(t) + K_{\mathcal{S}}\varepsilon(t)$$

$$y(t) = C_{\mathcal{S}}\hat{x}(t) + \varepsilon(t)$$

Let the input u(t) be of the general form (6.2). The model set is defined by

$$\begin{aligned} \hat{x}[t+1|\mathcal{M}(\theta)] &= A_{\mathcal{M}(\theta)}\hat{x}[t|\mathcal{M}(\theta)] + B_{\mathcal{M}(\theta)}u(t) + K_{\mathcal{M}(\theta)}\varepsilon_{\mathcal{M}(\theta)}(t) \\ y(t) &= C_{\mathcal{M}(\theta)}\hat{x}[t|\mathcal{M}(\theta)] + \varepsilon_{\mathcal{M}(\theta)}(t) \end{aligned} \tag{6.9}$$

where $\theta$ varies over a compact set $D_{\mathcal{M}}$. If $D_{\mathcal{M}}$ is finite Theorem 4.1 can be directly applied as in Section 6.11. The prediction $E[y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta)]$ is determined from u(s) and y(s), $s \leq t$, recursively as

$$\hat{x}[t+1|\mathcal{M}(\theta)] = A_{\mathcal{M}(\theta)}\hat{x}[t|\mathcal{M}(\theta)] + B_{\mathcal{M}(\theta)}u(t) + K_{\mathcal{M}(\theta)} \cdot$$

$$\cdot\ [y(t) - C_{\mathcal{M}(\theta)}\hat{x}[t|\mathcal{M}(\theta)]]$$

$$E[y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta)] = C_{\mathcal{M}(\theta)}\hat{x}[t+1|\mathcal{M}(\theta)]$$

This recursive scheme is exponentially stable if the eigenvalues of $A_{\mathcal{M}(\theta)} - K_{\mathcal{M}(\theta)}C_{\mathcal{M}(\theta)}$ are strictly inside the unit circle. Assume that this is the case for all $\theta \in D_{\mathcal{M}}$. Then $\hat{x}[t|\mathcal{M}(\theta)]$ as well as $\frac{\partial}{\partial\theta}\hat{x}[t|\mathcal{M}(\theta)]$ are determined by exponentially stable filtering of y(s) and u(s) (for $\frac{\partial}{\partial\theta}\hat{x}[t|\mathcal{M}(\theta)]$ also of $\hat{x}[t|\mathcal{M}(\theta)]$ ).

Hence, as in Section 6.1 we obtain the following theorem.

Theorem 6.2. Consider the system described by the state space equation (2.11). Let the model set be given by (6.9) and assume that the matrix elements are continuously differentiable functions of the parameter $\theta$. Assume that $D_{\mathcal{M}}$ is compact and that $A_{\mathcal{M}(\theta)} - K_{\mathcal{M}(\theta)}C_{\mathcal{M}(\theta)}$ has all eigenvalues strictly inside the unit circle for all $\theta \in$ $\in D_{\mathcal{M}}$. Assume that

- The fourth moments of e(t), y(t) and u(t) are uniformly bounded.
- The closed loop system is exponentially stable (cf. Def. 5.1).

Then the estimate $\theta(N)$ that minimizes the criterion (2.17) tends w.p.1 to the set

$$D_I^{(2)} = \left\{\theta \middle| \theta \in D_{\mathcal{M}},\ \liminf_{N\to\infty} \frac{1}{N}\sum_{1}^{N} E\left|\hat{y}(t|S) - \hat{y}(t|\theta)\right|^2 = 0\right\}$$

where

$$\hat{y}(t|S) - \hat{y}(t|\theta) = \sum_{s=0}^{t}[C(A-KC)^{t-s}B - \hat{C}(\hat{A}-\hat{K}\hat{C})^{t-s}\hat{B}]u(s) +$$

$$+ \sum_{s=0}^{t}[C(A-KC)^{t-s}K - \hat{C}(\hat{A}-\hat{K}\hat{C})^{t-s}\hat{K}]y(s) \qquad (6.10)$$

and

$A = A_S, \hat{A} = A_{M(\theta)}$ etc. □

Analysis similar to that in Section 6.1 can be applied to this system description.

Example 6.3. Consider the system (2.11) and assume that the input $\{u(t)\}$ is independent if $\{e(t)\}$, i.e. the system is in open loop. Decompose the predictions $\hat{y}(t|S)$ and $\hat{y}(t|\theta)$ into the parts that originate from $\{e(s)\}$ and the parts that originate from $\{u(s)\}$:

$$\hat{y}(y|S) = \hat{y}_e(t|S) + \hat{y}_u(t|S)$$

$$\hat{y}(t|\theta) = \hat{y}_e(t|\theta) + \hat{y}_u(t|\theta)$$

Then $\{y_e(t|S)\}$ and $\{y_u(t|S)\}$ are independent. This decomposition is possible since the system as well as the predictor are linear. If $\{e(t)\}$ is a stationary stochastic process, also $\{y_e(t|S)\}$ and $\{y_e(t|\theta)\}$ are stationary processes (neglecting initial value effects). Then

$$E|\hat{y}(t|S) - \hat{y}(t|\theta)|^2 = |\hat{y}_u(t|S) - \hat{y}_u(t|\theta)|^2 + \\ + E|\hat{y}_e(t|S) - \hat{y}_e(t|\theta)|^2$$

Since the stochastic processes in the second term of the RHS are stationary, this term does not depend on t. If $\theta \in D_I^{(2)}$, consequently the term is zero and so

$$\hat{y}_e(t|S) = \hat{y}_e(t|\theta) \quad \text{w.p.1}$$

follows. This straightforwardly implies, cf Tse-Weinert

(1973), that

$$C_{\mathcal{S}}(A_{\mathcal{S}}-zI)^{-1}K_{\mathcal{S}} = C_{\mathcal{M}(\theta)}\left(A_{\mathcal{M}(\theta)} - zI\right)^{-1}K_{\mathcal{M}(\theta)} \quad \text{a.e. } z$$

In the same manner follows, if $\{u(t)\}$ is persistently exciting of sufficiently high order, that

$$C_{\mathcal{S}}(A_{\mathcal{S}}-zI)^{-1}B_{\mathcal{S}} = C_{\mathcal{M}(\tilde{\theta})}\left(A_{\mathcal{M}(\theta)}-zI\right)^{-1}B_{\mathcal{M}(\theta)} \quad \text{a.e. } z$$

Thus, all $\theta \in D_I^{(2)}$ give models with the true transfer functions from u to y and from e to y. Consequently, for this type of experimental conditions $\mathcal{X}$ i.e. open loop experiments with a persistently exciting input signal, the system is System Identifiable, irrespectively of the orders of the models in $\mathcal{M}$, as long as they are not less than those of the system.

## 7. CONCLUSIONS.

In this report consistency and identifiability properties for prediction error identification methods have been investigated. The problem has, in various forms, been treated by several other authors, but the first entirely correct proof of consistency w.p.1 seems to have been published only recently, Rissanen-Caines (1974). That proof concerns the case with vector difference equations without an input signal. Like most other previous results, it is based on an ergodicity result. This means that when an input signal is present, restrictive conditions on it must be imposed. Feedback can usually not be allowed, and the system must be time invariant.

The main concern of the discussion of this report has been to remove such restrictions. Thus quite general feedback structures and adaptive systems can be handled.

The approach taken also allows quite general system descriptions and model structures. The results are therefore not limited to a certain class of systems, even though vector difference equations and state space models have received special attention.

## 8. ACKNOWLEDGEMENTS

I would like to thank Professor Karl Johan Åström for important discussions on the subject and Torsten Söderström and Daniel Thorburn for many valuable comments on the manuscript.

I am grateful to Gudrun Christensen, who has typed the manuscript and to Britt-Marie Carlson, who has prepared the figures.

The support of the Swedish Board for Technical Development under contract No. 733546 is also gratefully acknowledged.

## 9. REFERENCES.

Aoki, M., and Yue, P.C. (1970)
On Certain Convergence Questions in System Identification.
SIAM J. Control,Vol. 8, No. 2.

Åström, K.J., and Bohlin, T. (1965)
Numerical Identification of Linear Dynamic Systems from Normal Operating Records.
IFAC (Teddington) Symposium.

Åström, K.J., Bohlin, T., and Wensmark, S. (1965)
Automatic Construction of Linear Stochastic Dynamic Models for Stationary Industrial Processes Using Operating Records.
T.P. 18.150, Technical Paper, IBM Nordic Laboratory, Sweden.

Åström, K.J. (1970)
Introduction to Stochastic Control Theory.
Academic Press, New York.

Åström, K.J., and Eykhoff, P. (1971)
System Identification - a Survey.
Automatica, 7, 123-162.

Åström, K.J., and Wittenmark, B. (1971)
Problems of Identification and Control.
Journal of Mathematical Analysis and Applications, 34,pp. 90-113.

Åström, K.J., and Källström, C. (1973)
Application of System Identification Techniques to the Determination of Ship Dynamics.
Preprints of 3rd IFAC Symposium on Identification and System Parameter Estimation, the Hague, pp. 415-425.

Åström, K.J., and Wittenmark, B. (1973)
On Self Tuning Regulators.
Automatica, Vol. 9, pp. 185-199.

Balakrishnan, A.V. (1968)
Stochastic System Identification Techniques.
In:Stochastic Optimization and Control, M.F. Karreman, Ed., New York, Wiley, pp. 65-89.

Bellman, R., and Åström, K.J. (1970)
On Structural Identifiability.
Math. Biosc., Vol. 7, pp. 329-339.

Caines, P.E. (1970)
The Parameter Estimation of State Variable Models of Multivariable Linear Systems.
Ph.D. dissertation, Imperial College, London.

Caines, P.E. (1971)
The Parameter Estimation of State Variable Models of Multivariable Linear Systems.
Proceedings of the U.K.A.C. Conference on Multivariable Systems, Manchester, England.

Caines, P.E. (1973)
A Note on the Consistency of Maximum Likelihood Estimation for Finite Families of Stochastic Processes.
Report, Information Systems Laboratory, Stanford University.

Caines, P.E., and Rissanen, J. (1974)
Maximum Likelihood Estimation of Parameters in Multivariable Gaussian Stochastic Processes.
IEEE Trans. IT-20, No. 1

Chung, K.L. (1968)
A Course in Probability Theory.
Harcourt, Brace & World Inc.

Cramér, H. (1946)
Mathematical Methods of Statistics.
Princeton University Press, Princeton.

Cramér, H., and Leadbetter, M.R. (1967)
Stationary and Related Stochastic Processes.
Wiley, New York.

Doob, J.L. (1953)
Stochastic Processes,
Wiley, New York.

Eaton, J. (1967)
Identification for Control Purposes.
IEEE Winter meeting, New York.

Fisher, R.A. (1912)
On an Absolute Criterion for Fitting Frequency Curves.
Mess. of Math., Vol. 41, p. 155.

Glover, K., and Willems, J.C. (1973)
On the Identifiability of Linear Dynamic Systems.
Preprints of the 3rd IFAC Symposium on Identification and System Parameter Estimation, the Hague, pp. 867-871.

Gustavsson, I., Ljung, L., and Söderström, T. (1974)
Identification of Linear, Multivariable Process Dynamics Using Closed Loop Experiments.
Report 7401, Division of Automatic Control, Lund Inst. of Techn.

Hahn, W. (1967)
Stability of Motion.
Springer Verlag, Berlin.

Kailath, T. (1970)
The Innovations Approach to Detection and Estimation Theory.
Proc. IEEE, Vol. 58, No. 5.

Kalman, R. (1958)
Design of a Self-Optimizing Control System.
ASME Trans, Vol. 80, No. 2, Feb. 1958, pp. 468--478; Also in: Oldenburger R (ed.) Optimal Self-Optimizing Control, MIT Press, pp. 440-449, 1966.

Lampard, D.G. (1955)
A New Method of Determining Correlation Functions of Stationary Time Series.
Proc. IEE, 102C, pp. 35-41.

Mayne, D.Q. (1972)
A Canonical Form for Identification of Multivariable Linear Systems.
IEEE Trans., AC-17, No. 5, pp. 728-729.

Mehra, R.K., and Tyler, J.S. (1973)
Case Studies in Aircraft Parameter Identification.
Preprints of 3rd IFAC Symposium on Identification and System Parameter Estimation, the Hague, pp. 117-145.

Rissanen, J. (1973)
Basis of Invariants for Linear Dynamic Systems.
Report IBM Research Laboratory, San Jose, Calif.

Rissanen, J., and Caines, P.E. (1974)
Consistency of Maximum Likelihood Estimators for Multivariable Gaussian Processes with Rational Spectrum.
Report, Department of Electrical Eng., Linköping University, Sweden.

Spain, D.S. (1971)
Identification and Modelling of Discrete, Stochastic Linear Systems.
Tech. Report No. 6302-10, Standford University.

Tse, E., and Anton, J.J. (1972)
On the Identifiability of Parameters.
IEEE Trans. AC-17, No. 5.

Tse, E., and Weinert, M. (1973)
Correction and Extension of "On the Identifiability of Parameters".
IEEE Trans. AC-18, No. 6, pp. 687-688.

Tsypkin, Ya.Z. (1973)
Foundations of the Theory of Learning Systems.
Academic Press, New York.

Wald, A. (1949)
Note on the Consistency of the Maximum Likelihood Estimate.
Ann. Math. Stat., Vol. 20, pp. 595-601.

Woo, K.T. (1970)
Maximum Likelihood Identification of Noisy Systems.
Paper 3.1, 2nd Prague IFAC Symposium on Identification and Process Parameter Estimation.

# APPENDIX.

## PROOFS OF THE THEOREMS.

For easy reference, a list of the notation used in the proofs is given at the end of this appendix.

### A.1. Proof of Theorem 4.1.

Theorem 4.1. Consider the system $S$

$$y(t+1) = E[y(t+1) | Y_t, S] + \varepsilon(t+1, Y_t, S) \tag{A.1}$$

where

$$E[|\varepsilon(t+1, Y_t, S)|^2 | Y_t] < C$$

Consider a finite set of models $M_F = \{M(\theta) | \theta \in D_M\}$, such that $D_T(S, M_F)$ is non empty. Let $\theta(N)$ minimize the identification criterion $\operatorname{tr} Q_N(M_F(\theta))$ where $Q_N$ is defined by (2.15). Let $\tilde{D}_I^{(1)}$ be defined by

$$\tilde{D}_I^{(1)} = \left\{\theta | \theta \in D_M, \sum_1^{\infty} |\hat{y}(t|S) - \hat{y}(t|\theta)|^2_{R(t)} < \infty\right\} \tag{A.2}$$

Then $\theta(N) \to \tilde{D}_I^{(1)}$ w.p.1 as $N \to \infty$.

Proof. Let the loss function be denoted by

$$V_N(\theta) = \operatorname{tr} Q_N(M_F(\theta))$$

The idea of the proof is simply to show that for a given realization $\omega$ there exists a $N_i(\omega)$ such that

$$V_N(\theta_i) > V_N(\tilde{\theta}) \qquad N > N_i(\omega) \qquad (A.3)$$

for all $\theta_i \notin \tilde{D}_I^{(1)}(\omega)$

where $\tilde{\theta}$ is an arbitrary element in $D_T$.

Then the minimizing element for

$$N > \max_i N_i(\omega)$$

must belong to $\tilde{D}_I^{(1)}(\omega)$. The inequality (A.3) will be established using the convergence theorem for martingales, Doob (1953). With no loss of generality R(t) can be taken as I. Denote for brevity $\varepsilon(t+1, Y_t, S) = e(t+1)$. Consider for an arbitrary $\theta_i \in D_{M_F}$.

$$V_N(\theta_i) = \sum_{t=1}^{N} |y(t) - \hat{y}(t|\theta_i)|^2 =$$

$$= \sum_{t=1}^{N} |e(t)|^2 + 2 \sum_{t=1}^{N} e(t)^T [\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i)] +$$

$$+ \sum_{t=1}^{N} |\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i)|^2$$

Here

$$V_N(\tilde{\theta}) = \sum_{t=1}^{N} |e(t)|^2$$

Introduce

$$\sigma(t)^2 = |\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i)|^2$$

and

$$s(t) = \sum_{r=1}^{t} \sigma(r)^2 + 1 \qquad s(0) = 1$$

Then

$$V_N(\theta_i) - V_N(\tilde{\theta}) = s(N)\left[1 - \frac{1}{s(N)} + 2\,\frac{\sum_{t=1}^{N} e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i)]}{s(N)}\right] \tag{A.4}$$

It will be shown that the term

$$\frac{1}{s(N)} \sum_{t=1}^{N} e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i)] \to 0 \quad \text{as } N \to \infty \tag{A.5}$$

for any realization such that $\theta_i \notin \tilde{D}_I^{(1)}(\omega)$. This and (A.4) clearly imply (A.3).

To do so, consider the random variable

$$z(t) = \sum_{r=1}^{t} \frac{e(r)^T[\hat{y}(r|\tilde{\theta}) - \hat{y}(r|\theta_i)]}{s(r)}; \quad z(0) = 0$$

Let $\mathcal{Y}_t$ be the $\sigma$-algebra generated by $\{y(0),\ldots,y(t); u(0),\ldots,\mathcal{Y}_0\}$. Clearly $e(t) \in \mathcal{Y}_t$. Then

$$E\left\{[e(t)^T(\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i))]^2 | \mathcal{Y}_{t-1}\right\} =$$

$$= [\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i)]^T E(e(t)e(t)^T | \mathcal{Y}_{t-1})[\hat{y}(t|\tilde{\theta}) - y(t|\theta_i)] \leq$$

$$\leq \sigma(t)^2 C$$

The sequence $\{z(t), \mathcal{Y}_t\}$ is a martingale since

$$E[z(t+1)\,|\,\mathcal{Y}_t] = z(t) + E\Big[e(t+1)^T[\hat{y}(t+1|\tilde{\theta}) - \hat{y}(t+1|\theta_i)]/s(t+1)\,|\,\mathcal{Y}_t\Big] =$$

$$= z(t) + E[e(t+1)\,|\,\mathcal{Y}_t]^T[\hat{y}(t+1|\tilde{\theta}) - \hat{y}(t+1|\theta_i)]/s(t+1) =$$

$$= z(t)$$

Consider the variance of $z(N)$, i.e.

$$Ez(N)^2 = \sum_{t=1}^{N} E\Big\{z(t)^2 - z(t-1)^2\Big\} =$$

$$= E\sum_{t=1}^{N} E\Big\{[z(t)^2 - z(t-1)^2]\,|\,\mathcal{Y}_{t-1}\Big\} =$$

$$= E\sum_{t=1}^{N} E\Big\{[z(t) - z(t-1)]^2\,|\,\mathcal{Y}_{t-1}\Big\} =$$

$$= E\sum_{1}^{N} E\left\{\left[\frac{e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i)]}{s(t)}\right]^2 \Big|\,\mathcal{Y}_{t-1}\right\} \leq$$

$$\leq CE\sum_{1}^{N}\frac{\sigma(t)^2}{s(t)^2} \leq CE\sum_{1}^{N}\frac{s(t) - s(t-1)}{s(t)s(t-1)} \leq$$

$$\leq CE\sum_{1}^{N}\left[\frac{1}{s(t-1)} - \frac{1}{s(t)}\right] = CE\left[\frac{1}{s(0)} - \frac{1}{s(N)}\right] \leq C$$

Since $z(N)$ has bounded variance, it converges w.p.1 due to the martingale convergence theorem. It now follows from Kronecker's lemma (see e.g. Chung (1968)) that

$$\frac{1}{s(N)}\sum_{1}^{N} e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta_i)] \to 0 \quad \text{as } N \to \infty$$

for those realizations for which $s(N) \to \infty$ as $N \to \infty$, i.e. for those realizations, for which $\theta_i \notin \tilde{D}_I^{(1)}(\omega)$.

Consequently (A.5) has been established, which via (A.4) implies (A.3) and the proof of the theorem is concluded. □

A.2. Proof of Corollary to Theorem 4.1.

Corollary. Consider a general criterion $h\left(Q_N(M_F(\theta))\right)$, where h satisfies (2.16). Assume, in addition to the assumptions of the theorem that

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{1}^{N}|\hat{y}(t|S) - \hat{y}(t|\theta)|^2_{R(t)} < \infty \text{ w.p.1} \quad \text{all } \theta \in D_M,$$

that

$$E[\varepsilon(t+1,Y_t,S)\varepsilon(t+1,Y_t,S)^T|Y_t] > \delta I \text{ all } t,$$

and that

$$E[|\varepsilon(t+1,Y_t,S)|^4|Y_t] < C$$

Then $\theta(N) \to \tilde{D}_I^{(1)}$ w.p.1 as $N \to \infty$. □

Proof. The matrix of prediction errors can with $\tilde{\theta} \in D_T$ be decomposed as (take $R(t) = I$)

$$Q_N\left(\mathcal{M}_F(\theta)\right) = \sum_1^N [y(t) - \hat{y}(t|\theta)][y(t) - \hat{y}(t|\theta)]^T =$$

$$= \sum_1^N e(t)e(t)^T + \sum_1^N \Big\{ e(t)[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta)]^T +$$

$$+ [\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta)]e(t)^T \Big\} +$$

$$+ \sum_1^N [\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta)][\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta)]^T =$$

$$= Q_N^{(1)} + Q_N^{(2)}(\theta) + Q_N^{(3)}(\theta)$$

Choose (for a given realization) $\theta_i \notin \tilde{D}_I^{(1)}(\omega)$. Then $\operatorname{tr} Q_N^{(3)}(\theta_i) \to \infty$ as $N \to \infty$. Consider first

$$\frac{Q_N^{(1)}}{\operatorname{tr} Q_N^{(3)}(\theta_i)}$$

Let $E[e(t)e(t)^T | \mathcal{Y}_{t-1}] = S_t$. According to the assumptions $S_t > \delta I$ for all t.

Each element of the matrix

$$Z(t) = \sum_1^t [e(k)e(k)^T - S_k]/k$$

clearly is a martingale with bounded variance, from which follows that

$$\frac{1}{N} Q_N^{(1)} - \frac{1}{N}\sum_1^N S_t \to 0$$

and

$$2/\delta' \geq \frac{1}{N} Q_N^{(1)} \geq \frac{\delta'}{2} I \quad \text{for } n > N_0$$

where $\delta' = \min(\delta, 1/C)$.

From the additional assumptions of the corollary also follows that

$$\frac{1}{\operatorname{tr} Q_N^{(3)}(\theta_i)} Q_N^{(1)} \geq \frac{\delta'}{2} I \quad \text{for } n > N_0$$

As in the proof of the theorem follows that

$$\frac{Q_N^{(2)}(\theta_i)}{\operatorname{tr} Q_N^{(3)}(\theta_i)} \to 0 \quad \text{as } N \to \infty$$

for those realizations for which $\theta_i \notin \tilde{D}_I^{(1)}(\omega)$.

According to (2.16)

$$h[Q_N(M_F(\theta_i))] = g[\operatorname{tr} Q_N^{(3)}(\theta_i)] \cdot$$

$$\cdot\, h\left[\frac{Q_N^{(1)}}{\operatorname{tr} Q_N^{(2)}(\theta_i)} + \frac{Q_N^{(2)}(\theta_i)}{\operatorname{tr} Q_N^{(3)}(\theta_i)} + \right.$$

$$\left. + \frac{Q_N^{(3)}(\theta_i)}{\operatorname{tr} Q_N^{(3)}(\theta_i)}\right] \geq g[\operatorname{tr} Q_N^{(3)}(\theta_i)] \cdot$$

$$\cdot \left\{ h\left(\frac{Q_N^{(1)}}{\operatorname{tr} Q_N^{(3)}(\theta_i)}\right) + p(\delta') \operatorname{tr}\left[\frac{Q_N^{(3)}(\theta_i)}{\operatorname{tr} Q_N^{(3)}(\theta_i)}\right]\right\} =$$

$$= h\left(Q_N^{(1)}\right) + p(\delta') g[\operatorname{tr} Q_N^{(3)}(\theta_i)]$$

where the inequality holds for sufficiently large N, yielding the matrix

$$\frac{Q_N^{(2)}(\theta_i)}{\operatorname{tr} Q_N^{(3)}(\theta_i)}$$

sufficiently small. Hence

$$h[Q_N(M_F(\theta_i))] > h[Q_N(M_F(\tilde{\theta}))] \qquad N > N_0(\omega)$$

and convergence follows as in the theorem.

□

## A.3. Proof of Theorem 5.1.

Theorem 5.1. Consider the system

$$y(t+1) = E[y(t+1|Y_t,S] + \varepsilon(t+1,Y_t,S)$$

where $E[|\varepsilon(t+1,Y_t,S)|^4|Y_t] < C$. Consider a set of models $M$, such that $D_T(S,M)$ is non empty. Let $\theta(N)$ minimize the identification criterion $V_N(\theta) = \operatorname{tr}[\frac{1}{N} Q_N(M(\theta))]$, over a compact set $D_M$. Let $\tilde{D}_I^{(2)}(\omega)$ be defined by (5.1). Suppose that

$$z(t) = \sup_{\theta \in D_M} \max_{1 \le i \le n_y} \left| \frac{\partial}{\partial \theta} \hat{y}^{(i)}(t|\theta) \right| \qquad ((i) \text{ denotes } i\text{:th row})$$

where $D'_M$ is an open, connected set that contains $D_M$, satisfies the following condition

$$\limsup_{N\to\infty} \frac{1}{N} \sum_1^N z(t)^2 < \infty \qquad \text{w.p.1}$$

Then the estimate

$$\theta(N,\omega) \to \tilde{D}_I^{(2)}(\omega) \quad \text{a.e.} \quad \text{as } N \to \infty$$

The same results hold for the general criterion (2.17) if, in addition to the above assumptions,

$$E[\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})^T \mid \mathcal{Y}_t] \geq \delta I \quad \text{all } t$$

□

Proof. The technique to achieve inequalities like (A.3), that are uniform in $\theta$, is to consider

$$\inf V_N(\theta)$$

over open spheres $\theta \in B(\theta^*,\rho) = \{\theta \mid |\theta - \theta^*| < \rho\}$.

If it can be shown that, for some $\rho^* = \rho^*(\theta^*) > 0$,

$$\inf_{\theta \in B(\theta^*,\rho^*)} V_N(\theta) > V_N(\tilde{\theta}) \quad \text{for } N > N_0(\theta^*,\rho^*(\theta^*),\omega) \qquad \text{(A.6)}$$

where $\tilde{\theta}$ is some element in $D_T$, then it follows that the minimizing point, $\theta(N)$ cannot belong to $B(\theta^*,\rho^*)$ for $N > N_0(\theta^*,\rho^*,\omega)$. This result is then extended to hold for the complement of any open region containing $\tilde{D}_I^{(2)}$, by applying the Heine Borel theorem.

To show (A.6) consider first $V_N(\theta) = \operatorname{tr} \frac{1}{N} Q_N(\mathcal{M}(\theta))$ and take, with no loss of generality, $R(t) = I$. Let $\theta^*$ be an arbitrary point in $D_{\mathcal{M}}$.

Introduce for brevity

$$e(t) = \varepsilon(t,\mathcal{Y}_{t-1},\mathcal{S})$$

Then, with inf taken over $B(\theta^*,\rho)$ and with $\tilde{\theta} \in D_T$

$$\inf V_N(\theta) = \inf \frac{1}{N} \sum_1^N |y(t) - \hat{y}(t|\theta)|^2 =$$

$$= \inf \frac{1}{N} \sum_1^N |y(t) - \hat{y}(t|\tilde{\theta}) + \hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*) +$$

$$+ \hat{y}(t|\theta^*) - \hat{y}(t|\theta)|^2 =$$

$$= \inf\Big\{\frac{1}{N} \sum_1^N |e(t)|^2 + \frac{1}{N} \sum_1^N |\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)|^2 +$$

$$+ \frac{1}{N} \sum_1^N |\hat{y}(t|\theta^*) - \hat{y}(t|\theta)|^2 +$$

$$+ \frac{2}{N} \sum_1^N e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)] +$$

$$+ \frac{2}{N} \sum_1^N e(t)^T[\hat{y}(t|\theta^*) - \hat{y}(t|\theta)] +$$

$$+ \frac{2}{N} \sum_1^N [\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)]^T[y(t|\theta^*) - \hat{y}(t|\theta)]\Big\} \quad (A.7)$$

The mean value theorem gives

$$\hat{y}^{(i)}(t|\theta) - \hat{y}^{(i)}(t|\theta^*) = (\theta-\theta^*)^T \frac{\partial}{\partial\theta} \hat{y}^{(i)}(t|\xi)$$

where $\xi$ is a point between $\theta$ and $\theta^*$. If $\theta \in B(\theta^*,\rho) \cap D_M$ and $\rho$ is sufficiently small, say $\rho < \hat{\rho}(\theta^*)$, this implies that $\xi \in D_M$. Hence

$$|\hat{y}(t|\theta) - \hat{y}(t|\theta^*)| \leq |\theta - \theta^*| \sup_{\theta' \in D_M'} \max_i \left|\frac{\partial}{\partial\theta} \hat{y}^{(i)}(t|\theta')\right| =$$

$$= |\theta - \theta^*| z(t) \qquad \text{(A.8a)}$$

Since $D_M'$ is a connected set, any two points in $D_M$ say $\theta^*$ and $\theta$, can be connected by a train of such estimates which gives

$$|\hat{y}(t|\theta) - \hat{y}(t|\theta^*)| \leq Mz(t) \qquad \text{(A.8b)}$$

where M is the maximum distance in $D_M$.

Inserting (A.8a) in (A.7) yields for $\rho < \hat{\rho}$

$$\inf_{B(\theta^*,\rho)} V_N(\theta) \geq V_N(\tilde{\theta}) + \frac{1}{N}\sum_1^N |\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)|^2 +$$

$$+ \frac{2}{N}\sum_1^N e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)] - \frac{2}{N}\rho\sum_1^N |e(t)|z(t) -$$

$$- \frac{2}{N}\rho\sum_1^N |\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)| z(t) \qquad \text{(A.9)}$$

Consider the different terms of the RHS of (A.9):

- $\underline{\frac{2}{N}\sum_1^N e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)]}$

Exactly as in the proof of Theorem 4.1, Eqn. (A.5), it follows that

$$\sum_1^N e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)]/s(t) \text{ converges w.p.1 as } N\to\infty \qquad \text{(A.10)}$$

where s(t) is defined on p. And so (Kronecker's lemma)

$$\frac{1}{s(N)} \sum_{1}^{N} e(t)^{T}[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)] \to 0 \text{ as } N \to \infty \tag{A.11}$$

for any realization such that

$$s(N) = \sum_{1}^{N} |\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)|^2 \to \infty \quad \text{as } N \to \infty$$

In particular, for realizations such that $\theta^* \notin \tilde{D}_I^{(2)}(\omega)$, i.e.

$$s(N) > \delta(\theta^*)N$$

(A.11) holds. Furthermore, using (A.8b) we obtain

$$s(N) \leq M^2 \sum_{1}^{N} z(t)^2$$

If

$$\frac{1}{N} \sum_{1}^{N} z(t)^2$$

is bounded it now follows from (A.11) that

$$\frac{1}{N} \sum_{1}^{N} e(t)^{T}[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)] \to 0 \quad \text{as } N \to \infty \tag{A.12}$$

for realizations for which $\theta^* \notin \tilde{D}_I^{(2)}(\omega)$.

o $\underline{\frac{1}{N}\sum_{1}^{N}|e(t)|z(t)}$

From Schwarz's inequality, this term is less than

$$\left[\frac{1}{N}\sum_{1}^{N}|e(t)|^2 \cdot \frac{1}{N}\sum_{1}^{N} z(t)^2\right]^{1/2}$$

Consider first

$$r(N) = \sum_{1}^{N}[|e(t)|^2 - E\{|e(t)|^2|Y_{t-1}\}]/t$$

Clearly $(r(N), Y_{N-1})$ is a martingale with bounded variance. Hence r(N) converges w.p.1 and from Kronecker's lemma,

$$\frac{1}{N}\sum_{1}^{N}|e(t)|^2 - \frac{1}{N}\sum_{1}^{N} E\{|e(t)|^2|Y_{t-1}\} \to 0 \quad \text{w.p.1 as } N \to \infty \qquad (A.13)$$

which implies that

$$\frac{1}{N}\sum_{1}^{N}|e(t)|^2 < 2C \text{ for } N > N_0(\omega)$$

Hence

$$\frac{1}{N}\sum_{1}^{N}|e(t)|z(t) \leq \sqrt{2C} \cdot \sqrt{\frac{1}{N}\sum_{1}^{N} z(t)^2} \quad \text{for } N > N_0(\omega) \qquad (A.14)$$

o $\underline{\frac{1}{N}\sum_{1}^{N}|\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)|z(t)}$

Since

$$|\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)| \leq Mz(t)$$

the sum is less than

$$M\frac{1}{N}\sum_{1}^{N} z(t)^2$$

Now introduce a countable subset of $D_M$, that is dense in $D_M$. Let it be denoted by $\tilde{D}_M$. Also, introduce a subset $\Omega^*$ of the sample space such that

$$\Omega^* = \{(A.13) \text{ holds}\} \cap \{(A.10) \text{ holds for all } \theta^* \in \tilde{D}_M\} \cap$$
$$\cap \left\{\limsup_{N\to\infty} \frac{1}{N}\sum_{1}^{N} z(t)^2 < \infty\right\}$$

Clearly, $P(\Omega^*) = 1$. Consider from now on a fixed realization $\omega \in \Omega^*$ and introduce the set

$$D_M^*(\varepsilon,\omega) = \left\{\theta \,|\, \theta \in D_M, \inf_{\theta' \in \tilde{D}_I^{(2)}(\omega)} |\theta - \theta'| \geq \varepsilon\right\}$$

Choose $\theta^* \in D_M^*(\varepsilon,\omega) \cap \tilde{D}_M$.[1]

Suppose that

$$\frac{1}{N}\sum_{1}^{N} z(t)^2 < H(\omega) \qquad N > N_1(\omega) \tag{A.15}$$

[1] If this set is empty for any $\varepsilon > 0$, the assertion of the theorem is trivially true.

Since $\theta^* \notin \tilde{D}_I^{(2)}(\omega)$

$$\frac{1}{N}\sum_1^N |\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)|^2 > \delta(\theta^*) \qquad N > N_2(\omega)$$

Then from (A.9), (A.14) and (A.15), it follows that for $N > N_3(\omega) = \max\left(N_0(\omega), N_1(\omega), N_2(\omega)\right)$

$$\inf_{\theta \in B(\theta^*,\rho)} V_N(\theta) \geq V_N(\tilde{\theta}) + \delta(\theta^*) + \frac{1}{N}\sum_1^N e(t)^T[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)] -$$

$$- 2\rho \sqrt{2C}\sqrt{H} - 2\rho MH \qquad \text{(A.16)}$$

Choose

$$\rho^*(\theta^*) \leq \min\left[\frac{\delta(\theta^*)}{(MH+\sqrt{2CH})8}, \hat{\rho}(\theta^*)\right]$$

and $N_4(\omega)$ such that

$$\frac{1}{N}\sum_1^N e(t)^T[y(t|\tilde{\theta}) - y(t|\theta^*)] < \frac{\delta(\theta^*)}{4} \qquad N > N_4(\omega)$$

which is possible in view of (A.12). Then for $N > N_5(\omega) = \max\left(N_3(\omega), N_4(\omega)\right)$

$$\inf_{\theta \in B(\theta^*,\rho)} V_N(\theta) \geq V_N(\tilde{\theta}) + \delta(\theta^*)/2$$

which is the desired relation (A.6).

The case with a general criterion is treated analogously to the corollary of Theorem 4.1.

The family of open sets

$$\left\{B\left(\theta^*,\rho^*(\theta^*)\right),\ \theta^* \in D_M^*(\varepsilon,\omega)\ \cap\ \tilde{D}_M\right\}$$

clearly covers the compact set $D_M^*(\varepsilon,\omega)$. According to the Heine Borel theorem there exists a finite set

$$\left\{B\left(\theta_i,\rho^*(\theta_i)\right),\ i = 1,\ldots,K\right\}$$

that covers $D_M^*(\varepsilon,\omega)$. Let

$$N_0(\omega,\varepsilon) = \max_{1\le i\le K} N_0\left(\theta_i,\rho^*(\theta_i),\omega\right)$$

It then follows from (A.6) that

$$\inf_{\theta\in D_M^*(\varepsilon,\omega)} V_N(\theta) > V_N(\theta^0) \quad \text{for } N > N_0(\omega,\varepsilon)$$

which means that the minimizing element $\theta(N)$ cannot belong to $D_M^*(\varepsilon,\omega)$ for $N > N_0(\omega,\varepsilon)$, i.e.

$$|\theta(N) - \tilde{D}_I^{(2)}| < \varepsilon \quad \text{for } N > N_0(\omega,\varepsilon)$$

which, since $\varepsilon$ is an arbitrary small number, is the conclusion of the theorem.

## A.4. Proof of Theorem 5.2.

Theorem 5.2. Consider the system (5.3)

$$y(t+1) = E[y(t+1)|\mathcal{Y}_t,\mathcal{S}] + \varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})$$

Consider a set of models $\mathcal{M}$ such that $D_T(\mathcal{S},\mathcal{M})$ is non empty. Let $\theta(N)$ minimize the identification criterion $V_N(\theta) = \mathrm{tr}[\frac{1}{N} Q_N(\mathcal{M}(\theta))]$ over a compact set $D_{\mathcal{M}}$. Let $D_I^{(2)}$ be defined by (5.2). Let $\theta^*$ be an arbitrary element of $D_{\mathcal{M}}$, and introduce

$$\bar{B} = \bar{B}(\theta^*,\rho) = \left\{\theta \,\middle|\, |\theta^* - \theta| \leqslant \rho\right\} \qquad \rho > 0$$

and

$$\eta(t,\theta^*,\rho) = \inf_{\theta\in B}\Big\{[y(t+1) - \hat{y}(t+1|\mathcal{M}(\theta))]^T R(t) \cdot$$

$$\cdot\, [y(t+1) - \hat{y}(t+1|\mathcal{M}(\theta))]\Big\}$$

Assume that for all $\theta^* \in D_{\mathcal{M}}$

a) $E\left\{\sup_{\theta\in B} \max_i \left|\frac{\partial}{\partial\theta}\, \hat{y}^{(i)}(t+1|\mathcal{M}(\theta))\right|^2\right\} < C(\theta^*)$ for all t

and some $\rho = \rho_1(\theta^*) > 0$

b) $\mathrm{Cov}(\eta(t,\theta^*,\rho),\ \eta(t+s,\theta^*,\rho)) < K(1+|s|^{-\alpha})$;

$\alpha > 0$; all t,s and $0 \leq \rho \leq \rho_2^*$

K, $\alpha$ and $\rho_2^*$ may depend on $\theta^*$.

All expectations, including that in (5.2), is over the sequence of innovations $\{\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\}$.

Then $\theta(N) \to D_I^{(2)}$ w.p.1 as $N \to \infty$.

<u>Proof</u>. The proof of this theorem is inspired by Wald's (1949) proof of consistency of maximum likelihood estimates. As in the previous proof a relation like (A.6) has to be established. Since the observed variables are not independent in this case, the strong law of large numbers cannot be applied as in Wald's case. Instead the following result by Cramér-Leadbetter (1967) is used:

Let $\{f_i\}$ be a sequence of random variables, with zero mean values such that

$$Ef_i f_j \leq K \frac{i^p + j^p}{1 + |i - j|^q} \qquad 0 \leq 2p < q < 1$$

Then

$$\frac{1}{N} \sum_1^N f_i \to 0 \quad \text{w.p.1 as } N \to \infty \tag{A.17}$$

We now turn to the proof of the counterpart of (A.6), i.e. that for any $\theta^* \in D_{\mathcal{M}}$, $\theta^* \notin D_I^{(2)}$ there exists a $\rho^* = \rho^*(\theta^*)$ and a $N_0 = N_0(\theta^*,\rho^*,\omega)$ such that

$$\inf_{B(\theta^*,\rho^*)} V_N(\theta) > V_N(\tilde{\theta}) \qquad \text{for } N > N_0 \tag{A.18}$$

where $\tilde{\theta}$ is an arbitrary element in $D_T$.

Consider, for $\theta^* \in D_{\mathcal{M}}^*(\varepsilon) = \left\{\theta \,\middle|\, \inf_{\theta' \in D_I^{(2)}} |\theta - \theta'| \geq \varepsilon, \theta \in D_{\mathcal{M}}\right\}$

the quantity

$$\inf V_N(\theta) = \inf \frac{1}{N} \sum_1^N [y(t) - \hat{y}(t|\theta)]^T R(t) [y(t) - \hat{y}(t|\theta)]$$

where inf is taken over $\theta \in B(\theta^*,\rho)$.
Introduce for brevity the notation

$$\xi(t,\theta) = [y(t) - \hat{y}(t|\theta)]^T R(t) [y(t) - \hat{y}(t|\theta)]$$

$$\text{i.e. } \eta(t,\theta^*,\rho) = \inf_{B(\theta^*,\rho)} \xi(t,\theta)$$

$$\chi(t,\theta) = [\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta)]^T R(t) [\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta)]; \ \tilde{\theta} \in D_T$$

$$\Psi(t,\theta^*,\rho) = \sup_{\theta \in B(\theta^*,\rho)} \ \max_i \left| \frac{\partial}{\partial \theta} \hat{y}^{(i)}(t|\theta) \right|$$

Then

$$\inf V_N(\theta) \geq \frac{1}{N} \sum_1^N \eta(t,\theta^*,\rho) = \frac{1}{N} \sum_1^N E\xi(t,\theta^*) +$$

$$+ \frac{1}{N} \sum_1^N E[\eta(t,\theta^*,\rho) - \xi(t,\theta^*)] +$$

$$+ \frac{1}{N} \sum_1^N [\eta(t,\theta^*,\rho) - E\eta(t,\theta^*,\rho)] \qquad \text{(A.19)}$$

The three sums in the RHS of (A.19) will be considered separately.

o The term $\frac{1}{N} \sum E\xi(t,\theta^*)$.

$$E\xi(t,\theta^*) = Ee(t)^T R(t)e(t) + 2Ee(t)^T R(t)[\hat{y}(t|\tilde{\theta}) - \hat{y}(t|\theta^*)] +$$

$$+ E\chi(t,\theta^*)$$

The middle term is zero, since

$$E[e(t)|\mathcal{Y}_{t-1}] = 0$$

Since $\theta^* \notin D_I^{(2)}$

$$\liminf_{N\to\infty} \frac{1}{N} \sum_1^N E\chi(t,\theta^*) = \delta(\theta^*) > 0$$

Then

$$\frac{1}{N} \sum_1^N E\xi(t,\theta^*) = \frac{1}{N} \sum_1^N Ee(t)^T R(t)e(t) + \frac{1}{N} \sum_1^N E\chi(t,\theta^*) >$$

$$> EV_N(\tilde{\theta}) + \frac{\delta(\theta^*)}{2} \quad \text{for } N > N_1(\theta^*)$$

Consider

$$V_N(\tilde{\theta}) - EV_N(\tilde{\theta}) = \frac{1}{N} \sum_1^N e(t)^T R(t)e(t) - Ee(t)^T R(t)e(t)$$

According to (A.17) and assumption b) with $\theta^* = \tilde{\theta}$ and $\rho = 0$ it follows that

$$V_N(\tilde{\theta}) - EV_N(\tilde{\theta}) \to 0 \quad \text{w.p.1} \quad \text{as } N \to \infty$$

Hence

$$\frac{1}{N}\sum_{1}^{N} E\xi(t,\theta) > V_N(\tilde{\theta}) + \frac{\delta(\theta^*)}{2} \quad \text{for } N > N_2(\theta^*,\omega) \tag{A.20}$$

o <u>The term $\frac{1}{N}\sum_{1}^{N} E[\eta(t,\theta^*,\rho) - \xi(t,\theta^*)]$</u>

Choose first $\rho(\theta^*) = \min\left(\rho_1(\theta^*),\rho_2(\theta^*)\right)$

where $\rho_1$ and $\rho_2$ are defined in (5.7) and (5.8). Clearly

$|\hat{y}(t|\theta) - \hat{y}(t|\theta^*)| \leq |\theta - \theta^*|\Psi(t,\theta^*,\rho)$

for $\theta \in B(\theta^*,\rho)$

where $E\Psi(t,\theta^*,\rho)^2 < C(\theta^*)$ according to assumption a).

Consider

$$\begin{aligned}
|\xi(t,\theta) - \xi(t,\theta^*)| &= [y(t) - \hat{y}(t|\theta)]^T R(t)[y(t) - \hat{y}(t|\theta)] - \\
&\quad - [y(t) - \hat{y}(t|\theta^*)]^T R(t)[y(t) - \hat{y}(t|\theta^*)] = \\
&= 2y(t)^T R(t)[\hat{y}(t|\theta^*) - \hat{y}(t|\theta)] + \\
&\quad + \hat{y}(t|\theta)^T R(t)[\hat{y}(t|\theta) - \hat{y}(t|\theta^*)] + \\
&\quad + \hat{y}(t|\theta^*)^T R(t)[\hat{y}(t|\theta) - \hat{y}(t|\theta^*)] = \\
&= |\hat{y}(t|\theta) - \hat{y}(t|\theta^*)|\,|R(t)|\,|\hat{y}(t|\theta) + \hat{y}(t|\theta^*) - 2y(t)| \leq \\
&\leq |\hat{y}(t|\theta) - \hat{y}(t|\theta^*)|\,|R(t)|\,[|\hat{y}(t|\theta) - \hat{y}(t|\theta^*)| + \\
&\quad + 2|y(t) - \hat{y}(t|\theta^*)|] \leq \\
&\leq \rho\Psi(t,\theta^*,\rho)|R(t)|\,[\rho\Psi(t,\theta^*,\rho) + 2|y(t) - \hat{y}(t|\theta^*)|]
\end{aligned}$$

Notice that the bound does not depend on $\theta$, as long as $\theta \in B(\theta^*,\rho)$. Therefore, since $\eta(t,\theta^*,\rho) = \inf \xi(t,\theta)$

$$E|\eta(t,\theta^*,\rho) - \xi(t,\theta^*)| \leq \rho^2|R(t)|E\Psi(t,\theta^*,\rho)^2 +$$

$$+ \rho E\Psi(t,\theta^*,\rho)^2 4E|y(t) - \hat{y}(t|\theta^*)|^2 \leq$$

$$\leq \rho C_1 C(\theta^*)$$

where

$$C_1 = \rho|R(t)| + 4E|y(t) - \hat{y}(t|\theta^*)|^2 \leq \rho|R(t)| + 4K(\theta^*)$$

according to assumption b).

Now choose

$$\rho^* = \rho^*(\theta^*) = \min\left(\frac{\delta(\theta^*)}{8C_1C(\theta^*)}, \rho(\theta^*)\right)$$

Then

$$E|\eta(t,\theta^*,\rho^*) - \xi(t,\theta^*)| < \frac{\delta(\theta^*)}{8}$$

and

$$\frac{1}{N}\sum_1^N E|\eta(t,\theta^*,\rho^*) - \xi(t,\theta^*)| < \frac{\delta(\theta^*)}{8} \qquad (A.21)$$

o <u>The term $\frac{1}{N}\sum_{1}^{N} \eta(t,\theta^*,\rho) - E\eta(t,\theta^*,\rho)$</u>

According to assumption b) and (A.17) this term tends to zero w.p.1. Hence

$$\left|\frac{1}{N}\sum_{1}^{N} \eta(t,\theta^*,\rho) - E\eta(t,\theta^*,\rho)\right| < \frac{\delta(\theta^*)}{8}$$

$$\text{for } N > N_3(\theta^*,\rho^*,\omega) \qquad \text{(A.22)}$$

Inserting (A.22), (A.21) and (A.20) in (A.19) gives

$$\inf_{\theta\in B(\theta^*,\rho^*)} V_N(\theta) \geq V_N(\tilde{\theta}) + \frac{\delta(\theta^*)}{4} \qquad N > N_4(\theta^*,\rho^*,\omega)$$

where

$$N_4(\theta^*,\rho^*,\omega) = \max\left(N_2(\theta^*,\omega),\ N_3(\theta^*,\rho^*,\omega)\right)$$

which is the desired relation (A.18).

The proof is now completed using the Heine Borel theorem as in the previous proof.

□

## A.5. Proof of Lemma 5.1.

Lemma 5.1. Consider the system (2.2) and the model set (2.13). Suppose that $\hat{y}(t|\theta)$ is linear in $y(s)$ and $u(s)$ $0 \leq s < t$ and in $y_0$, i.e.

$$\hat{y}(t|\theta) = \sum_{k=1}^{t} h_{k,t}y(t-k) + \sum_{k=1}^{t} f_{k,t}u(t-k) + H_t y_0$$

Suppose that the linear filter that defines $\frac{\partial}{\partial\theta}\hat{y}^{(i)}(t|\theta)$ is exponentially stable for each component i of $\hat{y}(t|\theta)$ and for $\theta \in D_M$, i.e.

$$\left|\frac{\partial}{\partial\theta} h^{(i)}_{k,t}(\theta)\right| < C\lambda^k, \qquad \left|\frac{\partial}{\partial\theta} f^{(i)}_{k,t}(\theta)\right| < C\lambda^k \quad \left|\frac{\partial}{\partial\theta} H^{(i)}_{t}(\theta)\right| < C\lambda^t$$

for some $\lambda < 1$, all t, all $\theta \in D_M$ and $i = 1,\dots,n_y$. $|\cdot|$ denotes the operator norm of the matrices, and $h^{(i)}_{k,t}$ denotes the i:th row of the matrix $h_{k,t}$. Let

$$z(t) = \sup_{\theta\in D_M} \max_{1\leq i\leq n_y} \left|\frac{\partial}{\partial\theta}\hat{y}(t|\theta)\right|$$

and assume that

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{1}^{N}[y(t)^2 + u(t)^2] < \infty \quad \text{w.p.1}$$

Then

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{1}^{N} z(t)^2 > \infty \quad \text{w.p.1}$$

i.e. condition (5.4) holds.

Proof. As in the simple example on p. 41 we have

$$
\begin{aligned}
z(t) &= \sup \max \left| \frac{\partial}{\partial \theta} \hat{y}^{(i)}(t,\theta) \right| = \\
&= \sup \max \left| \sum_{k=0}^{t} \frac{\partial}{\partial \theta} h_{k,t}^{(i)}(\theta) y(t,k) + \right. \\
&\quad \left. + \sum_{k=0}^{t} \frac{\partial}{\partial \theta} f_{k,t}^{(i)}(\theta) u(t-k) + \frac{\partial}{\partial \theta} H_t^{(i)}(\theta) y_0 \right| \le \\
&\le \sum_{k=0}^{t} \left\{ \sup \max \left| \frac{\partial}{\partial \theta} h_{k,t}^{(i)}(\theta) \right| |y(t-k)| + \right. \\
&\quad \left. + \sup \max \left| \frac{\partial}{\partial \theta} f_{k,t}^{(i)}(\theta) \right| |u(t-k)| \right\} + \\
&\quad + \sup \max \left| \frac{\partial}{\partial \theta} H_t^{(i)}(\theta) \right| |y_0| \le \\
&\le C \left\{ \sum_{k=0}^{t} \lambda^k [|y(t-k)| + |u(t-k)|] + \lambda^t |y_0| \right\}
\end{aligned}
$$

where sup is taken over $\theta \in D_M$, and max over $1 \le i \le n_y$.

Introduce for brevity the notation

$$\eta(t) = C[|y(t)| + |u(t)|]$$

and define

$$\tilde{z}(t) = \sum_{1}^{t} \lambda^k \eta(t-k) + \lambda^t |y_0|$$

Then $z(t) \le \tilde{z}(t)$ and

$$\tilde{z}(t+1) = \lambda \tilde{z}(t) + \eta(t) \qquad \tilde{z}(0) = |y_0|$$

or

$$\tilde{z}(t+1)^2 = \lambda^2\tilde{z}(t)^2 + \eta(t)^2 + 2\lambda\tilde{z}(t)\eta(t)$$

Sum over $t = 0,\ldots,N$ and divide by N:

$$\frac{1}{N}\sum_1^N \tilde{z}(t)^2 \leq \lambda^2 \frac{1}{N}\sum_1^N \tilde{z}(t)^2 + \lambda^2 \frac{1}{N}|y_0| + \frac{1}{N}\sum_0^N \eta(t)^2 +$$

$$+ 2\lambda \frac{1}{N}\sum_0^N z(t)\eta(t)$$

or

$$(1-\lambda^2)\frac{1}{N}\sum_1^N \tilde{z}(t)^2 \leq \frac{1}{N}\sum_0^N \eta(t)^2 + 2\lambda\left[\frac{1}{N}\sum_0^N \eta(t)^2 \cdot \frac{1}{N}\sum_0^N \tilde{z}(t)^2\right]^{1/2} +$$

$$+ \frac{1}{N}\lambda^2|y_0|$$

According to the assumptions of the lemma,

$$\frac{1}{N}\sum_0^N \eta(t)^2$$

is bounded w.p.1, from which directly follows that

$$\frac{1}{N}\sum_1^N \tilde{z}(t)^2$$

is bounded w.p.1. Since $z(t) \leq \tilde{z}(t)$ this concludes the proof of the lemma.

□

## A.6. Proof of Lemma 5.2.

Lemma 5.2. Consider the system (2.2) and the model set (2.13). Suppose that $E[y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta)]$ is linear in y(s) and u(s), i.e.

$$E[y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta)] = \sum_{k=0}^{\infty} h_{k,t}(\theta)y(t-k) +$$

$$+ \sum_{k=0}^{\infty} f_{k,t}(\theta)u(t-k) + H_t(\theta)\mathcal{Y}_0$$

Suppose that the linear filters that define $E[y(t+1)|\mathcal{Y}_t, \mathcal{M}(\theta)]$ and $\frac{\partial}{\partial\theta} E[y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta)]$ are exponentially stable for $\theta \in D_{\mathcal{M}}$, i.e.

$$\left|h_{k,t}(\theta)\right| \le C\lambda^k, \qquad \left|f_{k,t}(\theta)\right| \le C\lambda^k \quad \left| H_t(\theta) \right| \leqslant C\lambda^t$$

$$\left|\frac{\partial}{\partial\theta} h_{k,t}^{(i)}(\theta)\right| \le C\lambda^k \quad \left|\frac{\partial}{\partial\theta} f_{k,t}^{(i)}(\theta)\right| \leqslant C\lambda^k$$

$$\left|\frac{\partial}{\partial\theta} H_t^{(i)}(\theta)\right| \le C\lambda^t$$

for some $\lambda < 1$ for all t and for all $\theta \in D_{\mathcal{M}}$, and all rows $i = 1,\dots,n_y$.

Assume further that

$$Ey(t)^4 \le C \quad \text{and} \quad Eu(t)^4 \le C$$

and that the closed loop system is exponentially stable. Assume that the innovations $\{\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\} = \{e(t+1)\}$ are independent random variables. Then conditions (5.7) and (5.8) of Theorem 5.2, and condition (5.14) of Lemma 5.1 are satisfied.

Proof. The conclusion about condition (5.7) follows immediately as in the simple example on p. 41.

Let the variables $y_N^0(t)$ and $u_N^0(t)$ be defined as in Definition 5.1. They are the input and the output to the system at time t, if the system and the regulators were started up at time t-N with zero initial conditions. Clearly $y_N^0(t)$ and $u_N^0(t)$ are independent of $\{e(s);\ s<t-N\}$.

Let the prediction based on $y_N^0(s)$ and $u_N^0(s)$ be denoted by $\hat{y}_N^0(t|\theta)$, i.e.

$$\hat{y}_N^0(t|\theta) = \sum_{k=1}^{N} h_{k,t} y_N^0(t-k) + \sum_{k=1}^{N} f_{k,t} u_N^0(t-k)$$

Then $\hat{y}_N^0(t|\theta)$ is independent of $\{e(s);\ s < t-N\}$.

Introduce also

$$\xi(t,\theta) = [y(t) - \hat{y}(t|\theta)]^T R(t)[y(t) - \hat{y}(t|\theta)]$$

$$\xi_N^0(t,\theta) = [y(t) - \hat{y}_N^0(t|\theta)]^T R(t)[y(t) - \hat{y}_N^0(t|\theta)]$$

$$\eta(t,\theta^*,\rho) = \inf_{\theta \in B(\theta^*,\rho)} \xi(t,\theta)$$

$$\eta_N^0(t,\theta^*,\rho) = \inf_{\theta \in B(\theta^*,\rho)} \xi_N^0(t,\theta)$$

The idea of the proof is quite simple. In view of the exponential stability of the closed loop system, the variables $\eta(t,\theta^*,\rho)$ and $\eta_N^0(t,\theta^*,\rho)$ will differ little for large N. Since $\eta(t-N,\theta^*,\rho)$ belongs to the $\sigma$-algebra generated by $\{e(s);\ s < t-N\}$, this variable is independent of $\eta_N^0(t,\theta^*,\rho)$.

Then $\eta(t-N,\theta^*,\rho)$ will depend only to a small extent on the variable $\eta(t,\theta^*,\rho)$, which is the conclusion that condition b) holds.

To formally prove the first claim, introduce

$$\nu_N(t,\theta^*,\rho) = \eta(t,\theta^*,\rho) - \eta_N^0(t,\theta^*,\rho)$$

From the obvious inequality

$$|\inf_x f_1(x) - \inf_x f_2(x)| \leq \sup_x |f_1(x) - f_2(x)|$$

we have

$$|\nu_N(t,\theta^*,\rho)| \leq \sup_{\theta \in B(\theta^*,\rho)} |\xi(t,\theta) - \xi_N^0(t,\theta)|$$

As in the proof of Theorem 5.2 we obtain by straightforward calculation

$$|\xi(t,\theta) - \xi_N^0(t,\theta)| \leq |\hat{y}(t|\theta) - \hat{y}_N^0(t|\theta)|\,|R(t)| \cdot$$

$$\cdot\ [\,|\hat{y}(t|\theta) - \hat{y}_N^0(t|\theta)| +$$

$$+ 2|y(t) - \hat{y}(t|\theta)|\,] \qquad \text{(A.23)}$$

Hence

$$\nu_N(t,\theta^*,\rho)^2 \leq 2|R(t)|\sup|\hat{y}(t|\theta) - \hat{y}_N^0(t|\theta)|^4 +$$

$$+ 4 \sup|\hat{y}(t|\theta) - \hat{y}_N^0(t|\theta)|^2 \sup|y(t) -$$

$$- \hat{y}(t|\theta)|^2$$

Consider first

$$\sup|\hat{y}(t|\theta) - \hat{y}_N^0(t|\theta)| = \sup \sum_{k=N+1}^{t} [h_{k,t}(\theta)y(t-k) + f_{k,t}(\theta)u(t-k)] +$$

$$+ \sum_{k=1}^{N} \Big\{ h_{k,t}(\theta)[y(t-k) - y_N^0(t-k)] +$$

$$+ f_{k,t}(\theta)[u(t-k) - u_N^0(t-k)] \Big\} +$$

$$+ H_t(\theta)\mathit{y}_0 \Big| \leq \sum_{k=N+1}^{t} \Big\{ \sup|h_{k,t}(\theta)|\cdot|y(t-k)| +$$

$$+ \sup|f_{k,t}(\theta)|\cdot|u(t-k)| \Big\} +$$

$$+ \sum_{k=1}^{N} \Big\{ \sup|h_{k,t}(\theta)|\cdot|y(t-k) - y_N^0(t-k)| +$$

$$+ \sup|f_{k,t}(\theta)|\cdot|u(t-k) - u_N^0(t-k)| \Big\} + \sup|H_t(\theta)||\mathit{y}_0| \leq$$

$$\leq C\left[ \sum_{N+1}^{t} \lambda^k[|y(t-k)| + |u(t-k)|] + \sum_{1}^{N} \lambda^k\lambda^{N-k}C(\mathit{y}_{t-N}) + \lambda^t|\mathit{y}_0| \right]$$

Hence

$$E \sup|\hat{y}(t|\theta) - \hat{y}_N^0(t|\theta)|^4 \leq 16E\, C^4 \left[ \sum_{k_1=N+1}^{t} \cdots \sum_{k_4=N+1}^{t} \cdot \right.$$

$$\cdot \lambda^{k_1+\ldots+k_4} \prod_{i=1}^{4} \Big\{ |y(t-k_i)| + |u(t-k_i)| \Big\} +$$

$$\left. + \lambda^{4N}N^4 C(\mathit{y}_{t-N})^4 + \lambda^{4t}|\mathit{y}_0| \right]$$

But, by applying Schwarz's inequality twice,

$$E \prod_{i=1}^{4} \left\{ |y(t-k_i)| + |u(t-k_i)| \right\} \leq$$

$$\leq \prod_{i=1}^{4} \left[ E\left\{ |y(t-k_i)| + |u(t-k_i)| \right\}^4 \right]^{1/4} \leq C$$

according to the assumptions of the lemma. Furthermore,

$$\sum_{k_1=N+1}^{t} \cdots \sum_{k_4=N+1}^{t} \lambda^{k_1+\ldots+k_4} \leq \left( \frac{\lambda^{N+1}}{1 - \lambda} \right)^4$$

and $N^4 \lambda^N < C$ for sufficiently large N. Consequently

$$E \sup |\hat{y}(t|\theta) - \hat{y}_N^0(t|\theta)|^4 \leq \tilde{C}\lambda^N \tag{A.24}$$

Consider now the other term in (A.23), $E \sup |y(t) - \hat{y}(t|\theta)|^2$

$$\sup |y(t) - \hat{y}(t|\theta)|^4 \leq 8 \sup |y(t)^4 + \hat{y}(t|\theta)^4| =$$

$$= 8\, y(t)^4 + 8 \sup \hat{y}(t|\theta)^4$$

It follows directly from the exponentially stable, linear expansion of $\hat{y}(t|\theta)$ into $y(t-s)$ and $u(t-s)$ that $E \sup \hat{y}(t|\theta)^4$ is bounded. Hence

$$E \sup |y(t) - \hat{y}(t|\theta)|^4 \leq C \tag{A.25}$$

From (A.23), (A.24) and (A.25) it follows that

$$E\nu_N(t,\theta^*,\rho)^2 < C\lambda^N \text{ where } \lambda < 1 \tag{A.26}$$

Consider now

$$\begin{aligned}
&\mathrm{Cov}\left(\eta(t+N,\theta^*,\rho),\ \eta(t,\theta^*,\rho)\right) = \\
&\quad = \mathrm{Cov}\left(\eta_N^0(t+N,\theta^*,\rho) + \nu_N(t+N,\theta^*,\rho),\ \eta(t,\theta^*,\rho)\right) = \\
&\quad = \mathrm{Cov}\left(\nu_N(t+N,\theta^*,\rho),\ \eta(t,\theta^*,\rho)\right) \le \\
&\quad \le [E\nu_N(t+N,\theta^*,\rho)^2 E\eta(t,\theta^*,\rho)^2]^{1/2} \le C\lambda^N
\end{aligned}$$

where the second inequality follows since $\eta_N^0(t+N,\theta^*,\rho)$ is independent of $\{e(s),\ s \le t\}$ and hence of $\eta(t,\theta^*,\rho)$. This concludes the proof (5.8). Condition (5.14) is shown analogously.

□

## A.7. List of Symbols Used in the Proofs.

$D_T = D_T(\mathcal{S},\mathcal{M})$ : set of $\theta$ giving models with the same "transfer function" as $\mathcal{S}$.

$D_{\mathcal{M}}$ : set of $\theta$ over which the minimization is performed.

$D_{\mathcal{M}_F}$ : as above for finite model sets.

$\tilde{\theta}$ : arbitrary element in $D_T$.

$\theta_i$ : arbitrary element in $D_{\mathcal{M}_F}$.

$\theta$ : arbitrary element in $D_{\mathcal{M}}$.

$\theta^*$ : arbitrary element in $D_{\mathcal{M}}$, center of spheres.

$$B(\theta^*,\rho) = \{\theta \,|\, |\theta - \theta^*| < \rho\}$$

$M$ : maximal length of arc connecting two elements in $D_{\mathcal{M}}$.

$z(t)$ : $\sup\limits_{\theta\in D_{\mathcal{M}}} \max\limits_{1\le i\le n_y} \left|\frac{\partial}{\partial\theta}\, \hat{y}^{(i)}(t|\theta)\right|$

$$\Psi(t,\theta^*,\rho) = \sup_{\theta\in B(\theta^*,\rho)} \max_{1\le i\le n_y} \left|\frac{\partial}{\partial\theta}\, \hat{y}^{(i)}(t|\theta)\right|$$

$$\xi(t,\theta) = [y(t) - \hat{y}(t|\theta)]^T R(t)[y(t) - \hat{y}(t|\theta)]$$

$$\eta(t,\theta^*,\rho) = \inf_{\theta\in B(\theta^*,\rho)} \xi(t,\theta)$$

$$\chi(t,\theta) = [y(t|\tilde{\theta}) - y(t|\theta)]^T R(t)[y(t|\tilde{\theta}) - y(t|\theta)] \qquad \tilde{\theta} \in D_T$$

$$e(t) = \varepsilon(t,\mathcal{Y}_{t-1},\mathcal{S})$$