On the Consistency of Prediction Error Identification Methods

Ljung, Lennart

1975

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Ljung, L. (1975). *On the Consistency of Prediction Error Identification Methods*. (Technical Reports TFRT-7077). Department of Automatic Control, Lund Institute of Technology (LTH).

Total number of authors:
1

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

ON THE CONSISTENCY OF PREDICTION ERROR IDENTIFICATION METHODS

L. LJUNG

Report 7506 (C) February 1975
Lund Institute of Technology
Department of Automatic Control

ON THE CONSISTENCY OF 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.

[*] Division of Automatic Control, Lund Institute of Technology, Fack, S-220 07 LUND, Sweden.
At present with the Information Systems Laboratory, Stanford University, Stanford, Calif. 94305

1. INTRODUCTION

The problem of identification is to determine a model that describes input-output data obtained from a certain system.

The choice of model is as a rule made using some criterion of closeness to the data, see e.g. Åström-Eykhoff (1971), Soudack et al (1971). In output error methods the discrepancy between the model output and the measured output is minimized. The common model-reference methods are of this type, see e.g. Lüders-Narendra (1974). Equation error methods minimize the discrepancy in the input-output equation describing the model. Mendel (1973) has given a quite detailed treatment of these methods.

Output- and equation error methods are originally designed for noiseless data and detrministic systems. If they are applied to noisy data or stochastic systems they will give biased parameter estimates, unless the noise characteristics either are known or are of a very special kind.

A natural extension of these methods is to take the noise characteristics into account and compare the predicted output of the model with the output signal of the system. Minimization of criteria based on this discrepancy leads to the class of prediction error identification methods. This class contains under suitable conditions the maximum likelihood method.

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 one (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 paper 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 Section 2. There also the identification method is described.
Different identifiability concepts are introduced in Section3, where a procedure to prove consistency is outlined.

In Section 4 consistency is shown for a general system structure as well as for linear systems. The application of the results to linear time-invariant systems is discussed in Section 5.

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 inputs and previous outputs:

$$y(t+1) = f_S\left(t;y(t),y(t-1),\ldots,y(1);u(t),\ldots,u(1);Y_0\right) \tag{2.1}$$

where 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 future outputs cannot be exactly determined by previous data as in (2.1). Instead the conditional probability distribution of y(t+1) given all previous data should be considered. It turns out to be convenient to subtract out the contitional mean and consider an innovations representation of the form

$$y(t+1) = E[y(t+1)|Y_t,S] + \varepsilon(t+1,Y_t,S) \tag{2.2}$$

where $E[y(t+1)|Y_t,S)]$ is the conditional mean given all previous outputs and inputs,

$$E[y(t+1)|Y_t,S] = g_S\left(t,y(t),\ldots,y(1),u(t),\ldots,u(1);Y_0\right) \tag{2.3}$$

Here Y_t denotes the σ-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 <u>innovations</u>, see Kailath (1970).

The conditional mean $E[y(t+1)|\mathcal{Y}_t,\mathcal{S}]$ will also be called the <u>prediction</u> 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.

<u>Remark</u>. It should be remarked that the description (2.2) to some extent depends on $\mathcal{Y}_0$. Two cases of choice of $\mathcal{Y}_0$ will be discussed here. The most natural choice is of course $\mathcal{Y}_0$ = the actual a priori information about previous behaviour known to the "model-builder". A disadvantage with this choice is that in general $E(y(t+1)|\mathcal{Y}_t,\mathcal{S})$ will be time varying even if the system allows a time-invariant description. This point is further clarified below. A second choice is $\mathcal{Y}_0 = \overline{\mathcal{Y}}_0$ = the information equivalent (from the viewpoint of prediction) to knowing <u>all</u> previous $y(t)$, $u(t)$, $t < 0$. This choice gives simpler representations $E(y(t+1)|\mathcal{Y}_t,\mathcal{S})$, but has the disadvantage that $\overline{\mathcal{Y}}_0$ is often not not by the person performing the identification procedure. Both choices will be discussed in more detail for linear systems below.

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 paper will be formulated for this general system description (2.2).

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

Example 2.1 - Linear systems in state space representation.

State space representations are a common and convenient way of describing linear, time-varying systems. The input output relation for the system S is then defined by

$$\begin{aligned} x(t+1) &= A_S x(t) + B_S u(t) + e(t) \\ y(t) &= C_S x(t) + v(t) \end{aligned} \qquad (2.4)$$

where $\{e(t)\}$ and $\{v(t)\}$ are sequences of independent gaussian random vectors with zero mean values and $Ee(t)e^T(t) = R_S(t)$, $E\ e(t)v(t)^T = R_S^c(t)$ and $E\ v(t)v(t)^T = Q_S(t)$. The system matrices may very well be time-varying but the time argument is suppressed.

The function

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

where Y_t is the σ-algebra generated by $\{y(t),\ldots,y(1), u(t),\ldots,u(1),Y_0\}$ is obtained as follows:

$$\hat{y}(t+1|S) = C_S\hat{x}(t+1|S) \qquad (2.5)$$

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

$$\hat{x}(t+1|S) = A_S\hat{x}(t|S) + B_S u(t) + K_S(t)\{y(t) - C_S\hat{x}(t|S)\} \quad (2.6\,a)$$

$K_S(t)$ is the Kalman gain matrix, determined from A_S, B_S, C_S, R_S, R_S^c, and $Q_S$as

$$K_S(t) = (A_S P_S(t) C_S^T + R_S^c)(C_S P_S(t) C_S^T + Q_S)^{-1}$$

$$P_S(t+1) = (A_S P_S(t) A_S^T - K_S(t)\ (C_S P_S(t) C_S^T + Q_S)^{-1}\ K_S^T(t) + R_S \quad (2.6b)$$

The information $\mathcal{Y}_0$ is translated into an estimate of the initial value $\hat{x}(0|S)$ with corresponding covariance $P_S(0)$. Then (2.6) can be solved recursively from t=0. The representation (2.6) then holds for any $\mathcal{Y}_0$, and in this case it is convenient to let $\mathcal{Y}_0$ be the actual a priori knowledge about the previous behaviour of the system. Notice that if the system matrices and covariance matrices all are time ivariant and $\mathcal{Y}_0 = \overline{\mathcal{Y}}_0$, then also K_S will be time invariant.

A continuous time state representation can be chosen instead of (2.4). In e.g. Åström-Källström (1973) and Mehra-Tyler (1973) it is shown how $E[y(t+1)|\mathcal{Y}_t, S]$, where $\mathcal{Y}_t$ is as before, can be calculated. The procedure is analogous to the one described above for sampled models. □

Example 2.2 - General linear, time-invariant systems.

A linear time-invariant system can be described as

$$y(t+1) = G_S(q^{-1})\ u(t) + H_S(q^{-1})\ e(t+1) \tag{2.7}$$

where q^{-1} is the backward shift operator: $q^{-1}\ u(t) = u(t-1)$ and $G_S(z)$ and $H_S(z)$ are matrix functions of z (z replaces q^{-1}). The variables $\{e(t)\}$ form a sequence of independent random variables with zero mean values and covariance matrices $E\ e(t)e(t)^T = \Lambda_S$ (which actually may be time-varying). It will be assumed that $G_S(z)$ and $H_S(z)$ are matrices with rational functions of z as entries and that $H_S(0) = I$. The latter assumption implies that e(t) has the same dimension as y(t), but this is no loss of generality. Furthermore, assume that det $H_S(z)$ has no zeroes on or inside the unit circle. This is no serious restriction, cf the spectral factorization theorem. Then $H_S^{-1}(q^{-1})$ is a well defined exponentially stable linear filter, which is straightforwardly obtained by inverting $H_S(z)$.

To rewrite (2.7) on the form (2.2) requires some caution regarding the initial values. If $\mathcal{Y}_0$ does not contain enough information the representation (2.2) will be time varying, even though (2.7) is time-invariant. In such a case a state space representation can be used. A simpler approach is to assume that $\mathcal{Y}_0 = \overline{\mathcal{Y}}_0$ = the information equivalent to knowing all previous y(t), u(t), t <0. It will follow from the analysis in the following sections that this assumption is quite relevant for identification problems.

From (2.7) we have

$$H_S^{-1}(q^{-1})\ y(t+1) = H_S^{-1}(q^{-1})\ G_S(q^{-1})\ u(t) + e(t+1)$$

and

$$y(t+1) = \{I - H_S^{-1}(q^{-1})\}\ y(t+1) + H_S^{-1}(q^{-1})G_S(q^{-1})\ u(t) + e(t+1) \tag{2.8}$$

Since $H_S^{-1}(0) = I$, the right hand side of (2.8) contains y(s) and u(s) only up to time t. The term e(t+1) is independent of these variables, also in the case u is determined from output feedback. Hence, if $Y_0 = \overline{Y}_0$,

$$E\left(y(t+1)|Y_t, S\right) = \{I - H_S^{-1}(q^{-1})\}\ y(t+1) + H_S^{-1}(q^{-1})\ G_S(q^{-1})\ u(t) \quad (2.9)$$

Now, linear systems are often not modelled directly in terms of the impulse response functions G_S and H_S. A frequently used representation is the vector difference equation (VDE):

$$\overline{A}_S(q^{-1})\ y(t) = \overline{B}_S(q^{-1})\ u(t) + \overline{C}_S(q^{-1})\ e(t) \qquad (2.10)$$

Another common representation is the state space form (in the time-invariant innovations form):

$$\begin{aligned} x(t+1) &= A_S\ x(t) + B_S\ u(t) + K_S\ e(t) \\ y(t) &= C_S\ x(t) + e(t) \end{aligned} \qquad (2.11)$$

It is easily seen that these two representations correspond to

$$G_S(z) = \overline{A}_S^{-1}(z)\ \overline{B}_S(z) \qquad H_S(z) = \overline{A}_S^{-1}(z)\ \overline{C}_S(z) \qquad (2.12)$$

and

$$G_S(z) = C_S\ (I - zA_S)^{-1}\ B_S \qquad H_S(z) = zC_S(I - zA_S)^{-1}K_S + I \ (2.13$$

respectively.In these two cases $G_S(z)$ and $H_S(z)$ will be matrices with rational functions as entries.

Inserting (2.12) into (2.9) it is seen that $E\left(y(t+1)|\ Y_t, S\right) = \hat{y}(t+1|S)$ is found as the solution of

$$C_S(q^{-1})\ \hat{y}(t+1|S) = \{C_S(q^{-1}) - A_S(q^{-1})\}\ y(t+1) + B_S(q^{-1})\ u(t) \qquad (2.14)$$

for the case of a VDE model. Solving (2.14) requires knowledge of $y(0),\dots,y(-n),u(0),\dots,u(-n),\hat{y}(0|S),\dots,\hat{y}(-n|S)$. This information is contained in the information $\overline{Y}_0$.

For the state space model (2.11) $\hat{y}(t+1|S)$ is found from

$$\begin{aligned}\hat{x}(t+1) &= A_S\hat{x}(t) + B_Su(t) + K_S\{y(t) - C_S\hat{x}(t)\} \\ \hat{y}(t+1|S) &= C_S\hat{x}(t+1)\end{aligned} \qquad (2.14\text{ b})$$

where the initial condition $\hat{x}(0)$ is obtained from $\overline{Y}_0$.

Notice that there is a parameter redundancy in the representations (2.10) and (2.11). All matrix polynomials $\overline{A}_S$, $\overline{B}_S$, and $\overline{C}_S$ and all matrices A_S, B_S, C_S, K_S that satisfy (2.12) and (2.13) respectively, correspond to the same system (2.7).

∎

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 in (2.2) and the properties of $\{\varepsilon(t+1,Y_t,S)\}$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 $g_S(t;y(t),\dots,y(1);u(t),\dots,u(1);\mathcal{Y}_0)$. 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 θ. The set over which θ 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

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

$$= g_{\mathcal{M}(\theta)}(t;\, y(t),\dots,y(1);\ u(t),\dots,u(1);\ \mathcal{Y}_0)$$

Hence

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

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

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

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

For brevity, the notation

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

is also used for the prediction.

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 θ_i into the system matrices.

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

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

where

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

etc.

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

Thus, for time-invariant linear systems the choice of model type, (vector difference equation, state space representation etc) and parameters can be understood as a way of parametrizing G and H in (2.8): $G_{M(\theta)}$ and $H_{M(\theta)}$ via (2.12) or (2.13).

Remark. Like for the system description, also the model description depends on the initial conditions $\mathcal{Y}_0$. It would be most sensible to choose $\mathcal{Y}_0$ as the actual a priori knowledge, but as remarked previously, this gives more complex algorithms for computing the prediction. For time-invariant systems it will therefore be assumed in the sequel that $\mathcal{Y}_0 = \bar{\mathcal{Y}}_0$ = knowledge of all previous history. Since $\bar{\mathcal{Y}}_0$ is in general not known it has to be included in the model: $\bar{\mathcal{Y}}_0 = \bar{\mathcal{Y}}_0(\theta)$. Often it is sufficient to take $\bar{\mathcal{Y}}_0(\theta) = 0$ all θ, i.e. $u(t)=y(t)=0, t < 0$, corresponding to zero initial conditions in (2.14) and (2.14b)

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 $E[y(t+1)|\mathcal{Y}_t,S]$ has desired behaviour. This is the separation theorem, see e.g. Åström (1970).

Therefore, it is very natural to choose a model that gives the best possible prediction. That is, some function of the prediction error

$$y(t+1) - E\left(y(t+1) \mid Y_t, M(\theta)\right)$$

should be minimized with respect to θ.

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

$$Q_N\left(M(\theta)\right) = \sum_{t=1}^{N} \sqrt{R(t)}\left(y(t) - \hat{y}(t|\theta)\right) \{\sqrt{R(t)}\left(y(t) - \hat{y}(t|\theta)\right)\}^T \quad (2.$$

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

<u>Properties of h.</u> Let h be a continuous function with $n_y \times n_y$, symmetric matrices as domain. Assume that

$$h(\lambda A) = g(\lambda)h(A), \ \lambda, g(\lambda) \text{ scalars and } g(\lambda) > 0 \text{ for } \lambda > 0 \quad (2.$$

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

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

for $\mathrm{tr}\, C_\varepsilon C_\varepsilon^T < \varepsilon_0$, where ε_0 depends only on δ and tr B. □

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

$$\tilde{V}_N(\theta) = h[Q_N(\mathcal{M}(\theta))]$$

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

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

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).

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 λ be the largest eigenvalue of B. Then $\lambda \geq \mathrm{tr}\ B/n_y$. 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 λ.) Now, adding C_ε to B can distort the eigenvalues at most ε/δ where $\varepsilon = \|C_\varepsilon\|$, the operator norm of C_ε, and

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

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

$$\text{for } \varepsilon < \frac{\delta\ \mathrm{tr}\ B}{2n_y\left(\frac{n_y}{\delta} + \mathrm{tr}\ B\right)} = \sqrt{\varepsilon_0}$$

which concludes the proof. □

In this chapter we will consider the limiting properties of th estimate θ that minimizes (2.17) as N tends to infinity. Of particular interest is of course whether the limiting values c θ gives models $\mathcal{M}(\theta)$ that properly describe $\mathcal{S}$. This is the problem of consistency of prediction error identification methods

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

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

where θ^* is the minimizing value. In particular, if $\{\varepsilon(t+1,\mathcal{Y}_t,\mathcal{S})\} = \{e(t+1)\}$ 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 $\frac{1}{N}\, Q_N(\mathcal{M}(\theta *))$ 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

$$y(t+1) = E(y(t+1)|\mathcal{Y}_t,\mathcal{M}(\theta)) + \varepsilon(t+1;\mathcal{M}(\theta)) \qquad (2.1$$

with

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

Let the innovations $\{\varepsilon(t,M(\theta))\}$ be assumed to be independent and normally distributed. The probability density of $y(t+1)$ given Y_t and given that (2.18) is true then is

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

$$\cdot e^{-\frac{1}{2}[x_{t+1} - \hat{y}(t+1|\theta)]^T\hat{\Lambda}^{-1}(t+1)[x_{t+1} - \hat{y}(t+1|\theta)]}$$

Here $f(x|Y_t) = F'(x|Y_t)$ where $F(x|Y_t) = P(y(t+1) \leq x|Y_t)$.

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

$$f(x_{t+1},x_t|Y_{t-1}) = f(x_{t+1}|y(t) = x_t,Y_{t-1})f(x_t|Y_{t-1}) =$$

$$= f(x_{t+1}|Y_t)f(x_t|Y_{t-1}) =$$

$$= [2\pi\ \det\ \hat{\Lambda}(t+1)\det\ \hat{\Lambda}(t)]^{-1/2} \cdot$$

$$\cdot \exp\Big\{-\tfrac{1}{2}[x_{t+1} - \hat{y}(t+1|\theta)]^T\hat{\Lambda}^{-1}(t+1) \cdot$$

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

$$\cdot \exp\Big\{-\tfrac{1}{2}[x_t - \hat{y}(t|\theta)]^T\hat{\Lambda}^{-1}(t)[x_t - \hat{y}(t|\theta)]\Big\}$$

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

Iteration directly gives the joint probability density of $y(t+1), y(t), \ldots, y(1)$ given $\mathcal{Y}_0$. The logarithm of the likelihood function, given $\mathcal{Y}_0$, then is obtained as

$$\log f\left(y(t+1), \ldots, y(1) \mid \mathcal{Y}_0\right) =$$

$$= -\tfrac{1}{2} \sum_{s=0}^{t} [y(s+1) - \hat{y}(s+1 \mid \theta)]^T \hat{\Lambda}^{-1}(s+1) [y(s+1) - \hat{y}(s+1 \mid \theta)] -$$

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

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

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

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

If the matrices $\hat{\Lambda}(t)$ are known, the MLE is consequently obtained as the minimizing point of the loss function (2.17) with $h(A) = \mathrm{tr}(A)$ 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 $\det[Q_N(\mathcal{M}(\theta))]$ giving $\theta(N)$ [where $R(t) = I$ in $Q_N(\mathcal{M}(\theta))$] and then take

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

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 way to connect the two concepts is introduced in this section. The definitions given here are consistent with those of Ljung-Gustavsson-Söderström (1974).

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

$$D_T(S,M) = \{\theta \,|\, \theta \varepsilon D_M,\ \lim_{N\to\infty} \frac{1}{N} \sum_1^N \ \{E[y(t+1)|\mathcal{Y}_t,\mathcal{Y}_0,S] - E[y(t+1|\mathcal{Y}_t,\mathcal{Y}_0(\theta),M(\theta)\}^T$$

$$\cdot\{E[y(t+1)|\mathcal{Y}_t,\mathcal{Y}_0,S] - E[y(t+1)|\mathcal{Y}_t,\mathcal{Y}_0(\theta),M(\theta))]\} = 0 \text{ all } \mathcal{Y}_t\ \} \tag{3.1}$$

The set $D_T(S,M)$ consists of all parameters in D_M which give models that describe the system without error in the mean square sense. There might be differences between S and $M(\theta)$, $\theta \varepsilon D_T(S,M)$ due to initial conditions and discrepances at certain time instances, but on the whole they are indistinguishable from input output data only.

For the case of linear, time-invariant systems it is easy to see that $D_T(S,M)$ can be described as

$$D_T(S,M) = \{\theta \,|\, \theta \varepsilon D_M,\ G_{M(\theta)}(z) = G_S(z)\ ;\ H_{M(\theta)}(z) = H_S(z) \text{ a.e.z}\} \tag{3.2}$$

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.

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 $\theta(N) \to$ $\to D_T(S,M)$ w.p.1 as $N \to \infty$.

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.

When minimizing the criterion function, it may however lead to numerical difficulties if there is a non-unique minimum. 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 θ^0 for the given model parametrization. In such a case the parameter θ^0 is said to be identifiable w.p.1 (or in probability) if there exists a sequence of estimates that tends to θ^0 w.p.1 (or in probability). Now, the sequence of estimates converges to θ^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 θ^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 is related to 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 (1974).

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 $SI(M,I,X)$. 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 Lemma 4.1 and in Theorems 4.1 and 4.2 of the following section the set D_I is determined for general model structures (2.18), and linear systems respectively.

Problem II is discussed in Section 5 for linear time-invariant systems. In Gustavsson-Ljung-Söderström (1974) problem II is extensively treated for vector difference equations.

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 paper.

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 GENERAL MODEL STRUCTURES.

The problem to determine a set D_I is, as mentioned above, a statistical problem. The approach used in most works is to apply an ergodicity result to the criterion function (2.17) and then show that D_I is the set of global minima of the limit of the criterion function. However, to assume ergodicity of the involved processes introduces rather limiting conditions on the system, possible feedback structure and on the input signal. Furthermore, uniform (in θ) inequalities for the loss functions must be established. This is a fairly difficult problem, which in fact has been overlooked in many of the previous works.

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

$$D_I = \{\theta | \theta \varepsilon D_M \, , \quad \lim_{N\to\infty} \inf \frac{1}{N} \sum_{1}^{N} \quad |\hat{y}(t+1|S) - \hat{y}(t+1|\theta)|^2_{R(t)} = 0\} \quad (4.1)$$

The reason for using limit inferior is that, under some cirkumstances, the limit may fail to exist.

It should also be noted that D_I may depend on the realization ω, $D_I(\omega)$, although in most applications it does not (a.e), see Section 5. For adaptive regulators it is, however, sometimes useful to consider D_I as a function of ω.

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

$$D_I (\omega) = \bar{D}_I \quad \text{w.p.1} \quad \text{or } D_I (\omega) \subset \bar{D}_I \quad \text{w.p.1} \qquad (4.2)$$

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

4.1 Main result.

Lemma 4.1. Consider the system

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

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

Consider a set of models, $\mathcal{M}$, such that $D_T(S,\mathcal{M})$ is non-empty. Let $\theta(N)$ minimize the identification criterion (2.17), $V_N(\theta) = h\left\{\frac{1}{N} Q_N(\mathcal{M}(\theta))\right\}$, where h satisies (2.16), over a compact set $D_\mathcal{M}$. Let $D_I(\omega)$ be defined by (4.1). Suppose that

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

where $D'_\mathcal{M}$ is an open set containing $D_\mathcal{M}$, satisfies the following condition

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{1}^{N} z(t)^2 < \infty \quad \text{w.p.1} \tag{4.3}$$

Assume further that

$$\limsup_{N\to\infty} \operatorname{tr} \frac{1}{N} Q_N(\mathcal{M}(\theta)) < \infty \quad \text{w.p.1 for any fixed } \theta\varepsilon D_\mathcal{M} \tag{4.4}$$

and that

$$E\{\varepsilon(t+1,\mathcal{Y}_t,S)\varepsilon(t+1,\mathcal{Y}_t,S)^T|\mathcal{Y}_t\}\geq\delta I \quad \text{all } t \tag{4.5}$$

(for h = tr this assumption (4.5) is not necessary).

Then the estimate

$$\theta(N,\omega) \to D_I(\omega) \quad \text{a.e. as } N\to\infty$$

The proof of the lemma is given in the appendix.

To apply Lemma 4.1 conditions (4.3) and (4.4) have to be checked. This requires some analysis of the model structures.

Remark. If the search is restricted to a finite number of models, conditions (4.3) and (4.4) can be removed and as the set D_I the smaller set

$$\{ \theta | \ \theta \varepsilon D_M \quad \sum_{1}^{\infty} \ |\hat{y}(t+1|S) - \hat{y}(t+1|\theta)|^2_{R(t)} < \infty \}$$

can be chosen, see Ljung (1974).

4.2 Linear systems.

For the linear, time-invariant model described in Example 2.2 it is relatively easy to find the variable z(t) defined in Lemma 4.1. If the search for models is restricted to those corresponding to stable predictors, (4.3) will be satisfied if the input and output are subject to similar conditions. This is shown in Theorem 4.1.

Theorem 4.1 (Linear, time-invariant systems)
Consider the system (2.7)

$$y(t+1) = G_S(q^{-1}) \ u(t) + H_S(q^{-1}) \ e(t+1)$$

where $E|e(t)|^4 < C$ (and, if the general criterion (2.17) is used $E \ e(t)e(t)^T > \delta I$), and let the model set be described by

$$y(t+1) = G_{M(\theta)}(q^{-1}) \ u(t) + H_{M(\theta)}(q^{-1}) \ \varepsilon(t+1) \qquad ; \ \theta \varepsilon D_M$$

where D_M is compact and $G_{M(\theta)}(z)$ and $H_{M(\theta)}(z)$ are matrices with rational functions as entries. Assume

o for VDE-parametrizations (2.12), that det $\overline{C}_{M(\theta)}(z)$ has all zeros outside the unit circle for $\theta \varepsilon D_M$.

- for state space realizations (2.13), that $A_{M(\theta)} - K_{M(\theta)}C_{M(\theta)}$ has all eigenvalues inside the unit circle for $\theta \varepsilon D_M$.

- for the general case, that det $H_{M(\theta)}(z)$ as well as the least common denominator to the denominator polynomials in $G_{M(\theta)}(z)$ and $H_{M(\theta)}(z)$ have all zeros outside the unit circle for $\theta \varepsilon D_M$.

Suppose also that $D_T(S,M)$ (defined in (3.2)) is non-empty.

Any feedback relationships between u and y may exist, but assume that

$$\lim_{N\to\infty} \sup \frac{1}{N} \sum_{1}^{N} \left(y(t)^T y(t) + u(t)^T u(t) \right) < \infty \quad \text{w.p.1} \tag{4.7}$$

Then the identification estimate $\theta(N)$ converges into D_I w.p.1 as N tends to infinity. In this case D_I can be expressed as

$$D_I = \{\theta \mid \theta \varepsilon D_M, \ \lim_{N\to\infty} \inf \frac{1}{N} \sum_{1}^{N} \left| \left(H_{M(\theta)}^{-1} - H_S^{-1}\right) y(t+1) + \right.$$

$$\left. + \left(H_S^{-1} G_S - H_{M(\theta)}^{-1} G_{M(\theta)}\right) u(t) \right|^2 = 0 \}$$

■

The proof is given in the appendix. It uses the fact that if det $H_{M(\theta)}(z)$ has all zeros outside the unit circle, then the linear filters that determine $\hat{y}(t|\theta)$ and $\frac{d}{d\theta}\hat{y}(t|\theta)$ from u and y are exponentially stable.

For the time-varying model described in Example 2.1 it is known that the Kalman filter (see e.g. Jazwinski (1970), theorem 7.4) is exponentially stable if the pair $(A(t),C(t))$ is completely uniformly observable and the pair $(A(t), \sqrt{R(t)})$ is completely uniformly controllable. Furthermore, the basis for the exponential depends only on the bounds on the observability and controllability Gramians. Hence we have the following theorem:

Theorem 4.2 (Time-varying linear systems on state space form). Consider the linear system described in Example 2.1 and suppose that the covariance matrices are uniformly bounded from above. (For the general criterion (2.17) assume also that $C_S(t)\ P_S(t)\ C_S(t)^T + Q(t) > \delta I$.) Let the set of models be defined by

$$x(t+1) = A_{M(\theta)}(t)\ x(t) + B_{M(\theta)}(t)\ u(t) + \varepsilon(t) \qquad \theta \varepsilon D_M$$
$$y(t) = C_{M(\theta)}(t)\ x(t) + w(t)$$

where $\{\varepsilon(t)\}$ and $\{w(t)\}$ are sequences of independent gaussian random variables with zero mean values and $E\ \varepsilon(t)\ \varepsilon(t)^T = R_{M(\theta)}(t)$, $E\ \varepsilon(t)\ w(t)^T = R^c_{M(\theta)}(t)$ and $E\ w(t)\ w(t)^T = Q_{M(\theta)}(t)$. D_M is a compact set such that $D_T(S,M)$ defined by (3.1) is non-empty and such that (A,C) is unformly (in t and in $\theta\varepsilon D_M$) completely observable and $(A,\sqrt{R})$ is uniformly (in t and in $\theta\varepsilon D_M$) completely controllable.

Any feedback relationships between u and y may exist but assume that (4.7) is satisfied. Then the identification estimate $\theta(N)$ converges into D_I with probability one as N tends to infinity.

Theorems 4.1 and 4.2 determine the set D_I under quite general and weak conditions. Actually, the imposed conditions: bounded fourth moments of the innovations, model search over stable predictors and the condition on the overall system behaviour (4.7) are believed to be satisfied in almost all conceivable applications.

For actual applications it is of interest to study D_I more closely: When is it possible to find a set D_I satisfying (4.2) and what additional assumtions on the input generation must be imposed in order to assure $D_I \subset D_T$, i.e. system identifiability. These questions are discussed in the next section.

5. IDENTIFIABILITY RESULTS.

As outlined in Section 3, the identifiability and consistency questions can be separated into three problems. The first problem, to determine a set D_I was solved in the previous section. The second problem, investigation of the set D_I and in particular the relation $D_I \subset D_T$ will be the subject of the present section.

5.1 A deterministic set D_I.

D_I as defined in (4.1) is a random variable. However in most applications

$$D_I = \bar{D}_I \quad \text{w.p.1} \tag{5.1}$$

where

$$\bar{D}_I = \{\theta |\ \theta \varepsilon D_M, \ \liminf_{N\to\infty} \ \frac{1}{N}\sum_{1}^{N} \ E|\hat{y}(t+1|S) - \hat{y}(t+1|\theta)|^2_{R(t)} = 0\} \ (5.2)$$

where the expectation is with respect to the sequence of innovations. This deterministic set $\bar{D}_I$ may be easier to handle.

For linear systems the relation (5.1) will hold if the system is in open loop and is stable, or contains linear feedback which makes the closed loop system exponentially stable. To include also non-linear feedback terms, which makes the closed loop system non-linear, the concept of exponential stability has to be extended to stochastic, non-linear systems.

Definition 5.1. Consider the linear system

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

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

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

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

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

For linear feedback this definition is consistent with that the closed loop poles be inside the unit circle.

It turns out that exponential stability assures not only (5.1) but also (4.7). Hence we have the following lemma:

Lemma 5.1. Consider the linear systems of example 2.1 or example 2.2. Let the input have the general form

$$u(t) = f_t(y(t),\ldots.,y(0);u(t-1),\ldots,u(0)) + u_r(t) + w(t)$$

where $u_r(t)$ is a signal that is independent of $y(s),u(s),s< t$ and such that

$$\lim_{N\to\infty} \sup \frac{1}{N} \sum |u_r(t)|^2 < \infty.$$

$\{w(t)\}$ is a sequence of disturbances of a filtered white noise character, say, which is independent of $\{e(t)\}$ and such that $E\ |w(t)|^4 < C$. The function f_t may be unknown to the experiment designer. Assume that the input is such that the closed loop system is exponentially stable (Def 5.1) and that D_M satisfies the assumptions of theorem 4.2 or 4.1 respectively. Suppose that $e(t)$, $y(t)$ and $u(t)$ have uniformly bounded fourth moments. Then (4.7) and (5.1) hold.

Proof. The proof is based on the following theorem due to Cramer and Leadbetter (1967):

$$\text{If } |\mathrm{Cov}(\ \xi(s),\ \xi(t))| \le K\ \frac{s^p + t^p}{1 + |t-s|^q} \qquad 0 \le 2p < q < 1 \tag{5.3}$$

then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{s=1}^{N} (\ \xi(s) - E\ \xi(s)) = 0 \quad \text{with probability one.}$$

It follows by straightforward calculations from the assumptions on exponential stability and on D_M that

$$\xi(t) = |\hat{y}(t|\mathcal{S}) - \hat{y}(t|\theta)|^2_{R(t)} \quad \text{and}$$

$$\eta(t) = |y(t)|^2 + |u(t)|^2$$

satisfy (5.3). (For details, see Ljung(1974), Lemma 5.2.) This proves the lemma.

5.2 Linear Time-invariant Systems

Let us now study in more detail linear time-invariant systems as treated in Example 2.2 and Theorem 4.1. Since this class includes any parametrization of vector difference equations or state space realizations or any other parametrization of a linear time-invariant system, it is believed that such analysis is sufficient for most applications.

From Theorem 4.1 and Lemma 5.1 it follows that the estimates tend to the set

$$\bar{D}_I = \{\theta | \liminf_{N\to\infty} \frac{1}{N}\sum_1^N E|(H_{M(\theta)}^{-1} - H_{\mathcal{S}}^{-1})y(t+1) + \\ + (H_{\mathcal{S}}^{-1}G_{\mathcal{S}} - H_{M(\theta)}^{-1}G_{M(\theta)})u(t)|^2 = 0\}$$

This set clearly depends on the input signal. If the input is not sufficiently general, the set may contain parameters corresponding to models that describe the system well for the used input, but fail to describe it for other inputs. This is the case if the input contains too few frequences or if it has certain relationships with the output. Then $\bar{D}_I$ is not contained in D_T and the system is not System Identifiable for this input (experiment condition).

The set $\bar{D}_I$ has been analysed in Ljung-Gustavsson-Soderstrom(1974) in detail for the case of time-varying feedback. Here we will consider a case with linear feedabck and an extra input signal (or noise).

Let the input be

$$u(t) = F(q^{-1})\ y(t) + u_R(t)$$

where $\{u_R(t)\}$ is a sequence that does not depend on $\{e(t)\}$. Suppose that $F(z)$ is a matrix with rational functions as entries, such that the closed loop system is stable.

The closed loop system is

$$y(t+1) = \left(I - q^{-1}G_S(q^{-1})\ F(q^{-1})\right)^{-1} H_S(q^{-1})\ e(t+1) +$$

$$+\left(I - q^{-1}G_S(q^{-1})\ F(q^{-1})\right)^{-1} G_S(q^{-1})\ u_R(t)$$

Introduce

$$\tilde{e}(t+1) = \left(I - q^{-1}G_S(q^{-1})\ F(q^{-1})\right)^{-1} H_S(q^{-1})\ e(t+1)$$

$$\tilde{u}_R(t) = \left(I - q^{-1}G_S(q^{-1})\ F(q^{-1})\right)^{-1} G_S(q^{-1})\ u_R(t)$$

$$K_\theta(q^{-1}) = H^{-1}_{M(\theta)}(q^{-1}) - H_S^{-1}(q^{-1})$$

$$L_\theta(q^{-1}) = H_S^{-1}(q^{-1})G_S(q^{-1}) - H^{-1}_{M(\theta)}(q^{-1})G_{M(\theta)}(q^{-1})$$

Then

$$\overline{D}_I = \{\theta |\ \liminf_{N\to\infty} \frac{1}{N}\sum_1^N E|K_\theta(q^{-1})\left(\tilde{e}(t+1) + \tilde{u}_R(t)\right) -$$

$$- L_\theta(q^{-1})\ \left(F(q^{-1})\ \tilde{e}(t+1) + F(q^{-1})\ \tilde{u}_R(t) + u_R(t)\right)|^2 = 0\ \}$$

Since $\tilde{e}$ is independent of $\tilde{u}_R$ and u_R, the expectation can be written

$$E\,|(K_\theta(q^{-1}) - L_\theta(q^{-1})F(q^{-1}))\tilde{e}(t+1)|^2 +$$

$$+ E|(K_\theta(q^{-1}) - L_\theta(q^{-1})F(q^{-1}))\tilde{u}_R(t) + L_\theta(q^{-1})u_R(t)|^2$$

If $E\,e(t)e(t)^T > \delta I$, then it follows that

$$K_\theta(q^{-1}) - L_\theta(q^{-1})F(q^{-1}) = 0 \quad \text{for} \quad \theta\varepsilon\overline{D}_I \tag{5.4}$$

since the first term has to be arbitrarily close to zero infinetely often for $\theta\varepsilon\overline{D}_I$.

This in turn implies that

$$\liminf_{N\to\infty} \frac{1}{N}\sum_1^N |L_\theta(q^{-1})u_R(t)|^2 = 0 \quad \text{for } \theta\varepsilon\overline{D}_I.$$

If u_R is persistenly exciting (see e.g. Mayne(1972)) of sufficiently high order then this implies that

$$L_\theta(q^{-1}) \equiv 0 \quad \text{for } \theta\varepsilon\overline{D}_I$$

which, via (5.4) implies that

$$K_\theta(q^{-1}) \equiv 0 \quad \text{for } \theta\varepsilon\overline{D}_I.$$

That is, $\overline{D}_I = D_T$.

<u>Remark</u>. Let $U_M(t) = \text{col}\ (u_R(t), \ldots ,u_R(t-M))$.
Then it is sufficient to assume that

$$\delta I < \frac{1}{N}\sum_1^N U_M(t)U_M(t)^T < \frac{1}{\delta} I \quad ; N > N_0 \tag{5.5}$$

The limit of the sum does not have to exist, as in the definition of persistent excitation in Mayne(1972).

The number M for which (5.5) has to be satisfied depends on $\mathcal{S}$ and on the parametrization of $\mathcal{M}$. For state space representations M can be related to the orders of the system and model, see e.g. Mayne(1972).

For the unspecified models, which we deal with here, we can require that (5.5) holds for any M.

Summing up this discussion, Lemma 5.1 and Theorem 4.1, we have the following theorem.

Theorem 5.1. Consider the system (2.7), S,

$$y(t+1) = G_S(q^{-1})\ u(t) + H_S(q^{-1})\ e(t+1)$$

where $\{e(t)\}$ is a sequence of independent random variables such that $E|e(t)|^4 < C$ and $E\ e(t)e(t)^T > \delta I$

The input is

$$u(t) = F(q^{-1})\ y(t) + u_R(t)$$

where $\{u_R(t)\}$ is independent of $\{e(t)\}$ and satisfies (5.5) for any M. Assume that F is such that the closed loop system is exponetially stable.

Let the model set, M , be described by

$$y(t+1) = G_{M(\theta)}(q^{-1})\ u(t) + H_{M(\theta)}(q^{-1})\ e(t+1) \quad ; \quad \theta \varepsilon D_M$$

where D_M is compact and such that $H_{M(\theta)}(z)$ satisfies the same conditions as in Theorem 41for $\theta \varepsilon D_M$. Assume that $D_T(S,M)$ is non-empty. Let $\theta(N)$ be the estimate of θ based on N data points, obtained by minimizing the general criterion (2.17). Then

$$\theta(N) \to D_T(S,M) \text{ with probability one as } N\to\infty$$

where

$$D_T(S,M) = \{\theta | G_S(z) = G_{M(\theta)}(z)\ ;\ H_S(z) = H_{M(\theta)}(z) \quad \text{a.e. } z.\}$$

That is, S is System Identifiable.

Remark. Notice that, when evaluating the criterion (2.17), the predictor $\hat{y}(t|\theta)$ does not have to be based on the true initial data. As remarked several times above, it is most suitably chosen as the time-invariant,steady state predictor (2.9) initialized with zero initial state.

6. CONCLUSIONS.

In this contribution consistency and identifiability properties of prediction error identification methods have been investigated. A separation of the problem into three different tasks has been introduced, and results of varying generality and on varying levels have been given. The results can be used as a kit for "doing-your-own consistency theorem" in a specific application. They also solve the identifiability and consistency problem for linear, time-invariant systems under quite general assumptions, as shown in theorem 5.1.

The hard part in consistency theorems is believed to be to determine the set into which the estimates converge (Problem I in the formulation of Section 3). This has been solved for quite general (Lemma 4.1) as well as for linear systems (Theorems 4.1 and 4.2). Due to the very weak conditions imposed, these results are applicable to adaptive systems of almost any kind, in addition to the more straightforward cases treated in Section 5.

The difficult and vital problem of choosing a parametrization (model set) that gives identifiable parameters and consistent parameter estimates has not been considered here. However, this problem is most conveniently treated as a problem of its own, and it does not depend on the identification method chosen.

APPENDIX.

A.1 Proof of Lemma 4.1

The idea of the proof is to show that

$$\inf V_N(\theta) > V_N(\tilde{\theta}) \quad \text{for} \quad N > N_0(\theta^\times, \rho, \omega)$$

where the infimum is taken over an open sphere around $\theta^\times$ with radius ρ, and where $\theta \varepsilon D_T$. Then the minimizing point $\theta(N)$ cannot belong to this sphere for $N > N_0(\theta^\times, \rho, \omega)$. This result is then extended to hold for the complement of any open region containing D_I, by applying the Heine-Borel theorem.

Let without loss of generality $R(t) = I$. Introduce, for short,

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

and consider

$$Q_N(\mathcal{S}) = \sum_1^N \; e(t)e(t)^T$$

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

$$\frac{1}{N} Q_N^{(1)} - \frac{1}{N} \sum_1^N S_t \to 0 \qquad \text{w.p. 1 as } N\to\infty$$

and

$$2/\delta' \geq \frac{1}{N} \Omega_N^{(1)} \geq \frac{\delta'}{2} I \quad \text{for } n > N_1(\omega) \ , \ \omega\varepsilon\Omega_1 \text{ where } P(\Omega_1)=1$$

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

(The argument ω will as a rule be suppressed in the variables $y, e, \hat{y}$ etc, but used explicitly in bounds.)

Introduce also

$$\beta(t) = \hat{y}(t|S) - \hat{y}(t|\theta^{\times})$$

Then it follows from (4.4) and (A.1) that

$$\frac{1}{N} \sum_{1}^{N} |\beta(t)|^2 < C_1(\omega,\theta^{\times}) \quad ; \quad \omega\varepsilon\Omega_2(\theta^{\times}) \ ,P(\Omega_2(\theta^{\times}))=1. \qquad (A.2)$$

Now take a fixed element $\tilde{\theta}\varepsilon D_T$ and consider

$$Q_N(\tilde{\theta}) = \sum_{1}^{N} \left(y(t) - \hat{y}(t|\tilde{\theta})\right) \left(y(t) - \hat{y}(t|\tilde{\theta})\right)^T$$

Introduce

$$\alpha(t) = \hat{y}(t|S) - \hat{y}(t|\tilde{\theta})$$

Then

$$Q_N(\tilde{\theta}) = Q_N(S) + \sum_{1}^{N} \left(e(t)\alpha(t)^T + \alpha(t)e(t)^T + \alpha(t)\alpha(t)^T \right)$$

Since $\tilde{\theta}\varepsilon D_T$, by definition

$$\frac{1}{N} \sum_{1}^{N} \alpha(t)^T\alpha(t) \rightarrow 0 \text{ w.p.1 as } N\rightarrow\infty$$

and

$$\left| \frac{1}{N} \sum_{1}^{N} e(t)\alpha(t)^T + \alpha(t)e(t)^T \right|^2 \leq 4\left(\frac{1}{N} \sum_{1}^{N} |e(t)|^2 \cdot \frac{1}{N} \sum_{1}^{N} |\alpha(t)|^2 \right)$$

But from (A.1)

$$\frac{1}{N} \sum_{1}^{N} |e(t)|^2 < 2n_y/\delta \quad \text{for} \quad N > N_1(\omega) \quad (n_y = \text{number of outputs})$$

Hence

$$Q_N(\tilde{\theta}) - Q_N(S) \to 0 \text{ w.p.1 as } N\to\infty$$

and, since h is continuous,

$$V_N(\tilde{\theta}) < V_N(S) + \varepsilon \quad \text{for } N > N_2(\omega,\varepsilon) \text{ and } \omega\varepsilon\Omega_3,\ P(\Omega_3)=1 \qquad \text{(A.3)}$$

Now consider $Q_N(\theta)$ and decompose

$$y(t) - \hat{y}(t|\theta) = y(t) - \hat{y}(t|S) + \hat{y}(t|S) - \hat{y}(t|\theta^{\times}) + \hat{y}(t|\theta^{\times}) - y(t|\theta)$$

where $\theta^{\times}\varepsilon D_M$ is a fixed point and $\theta\varepsilon B(\theta^{\times},\rho) =\{\theta|\ |\theta - \theta^{\times}|<\rho \}$
Introduce for short

$$\gamma(t) = \hat{y}(t|\theta^{\times}) - \hat{y}(t|\theta)$$

From the mean value theorem

$$|\gamma(t)| \leq \rho\, z(t) \qquad \text{(A.4)}$$

if ρ is sufficiently small. Then

$$Q_N(\theta) = Q_N(S) + \sum_{1}^{N} \beta(t)\beta(t)^T + \sum_{1}^{N} \gamma(t)\gamma(t)^T + \sum_{1}^{N} \left(e(t)\beta(t)^T + \beta(t)e(t)^T\right) +$$

$$+ \sum_{1}^{N} \left(e(t)\gamma(t)^T + \gamma(t)e(t)^T\right) + \sum_{1}^{N} \left(\beta(t)\gamma(t)^T + \gamma(t)\beta(t)^T \right) \qquad \text{(A.5)}$$

We will first show that each element of the matrix

$$\frac{1}{N} Q_N^{(1)}(\theta^\times) = \frac{1}{N} \sum_1^N \left(e(t)\beta(t)^T + \beta(t)e(t)^T\right)$$

tends to zero w.p.1 as N tends to infinity:

Let $r(t) = \max(t, \sum_1^t \beta_i^2(k))$ (β_i = i:th component of β)

It is easy to see that

$$m_N = \sum_1^N e_j(t)\beta_i(t)/\, r(t)$$

is a martingale with respect to $\{\mathcal{Y}_N\}$.

Furthermore,

$$E\, m_N^2 = E \sum_1^N E\left(e_j^2(t) | \mathcal{Y}_{t-1}\right) \cdot \beta_i^2(t)/r^2(t) \le E \sum_1^N \frac{C\, \beta_i^2(t)}{\sum_1^t \beta_i^2(k)} \le C$$

Hence m_N converges with probability one (for $\omega\varepsilon\Omega_4(\theta^\times)$, $P(\Omega_4(\theta^\times))=1$) according to the martingale convergence theorem, Doob(1953). It now follows from Kronecker's lemma (see e.g. Chung(1968)) that, since $r(N)\to\infty$,

$$\frac{1}{r(N)} \sum_1^N e_j(t)\, \beta_i(t) \to 0 \quad \text{as } N\to\infty \quad \text{for } \omega\varepsilon\Omega_4(\theta^\times)$$

But $1 \le r(N)/N \le C_1(\omega,\theta^\times)$ for $\omega\varepsilon\Omega_2$ according to (A.2). Hence

$$\frac{1}{N} \sum_1^N e_j(t)\, \beta_i(t) \to 0 \quad \text{as } N\to\infty \quad \text{for } \omega\varepsilon\Omega_4(\theta^\times)\cap\Omega_2$$

and so

$$\mathrm{tr}\ \frac{1}{N} Q_N^{(1)}(\theta^\times)\, \frac{1}{N} Q_N^{(1)}(\theta^\times)^T < \varepsilon \quad \text{for } N>N_3(\varepsilon,\omega,\theta^\times) \text{ and } \omega\varepsilon\Omega_2\cap\Omega_4(\theta^\times) \qquad (A.6)$$

From (A.4),(A.1), (A.2) and (4.3) it follows that

$$\mathrm{tr}\ \frac{1}{N} Q_N^{(2)}(\theta,\theta^\times)\, \frac{1}{N} Q_N^{(2)}(\theta,\theta^\times)^T < C_2(\omega,\theta^\times)\cdot\rho \qquad (A.7)$$

where

$$Q_N^{(2)}(\theta,\theta^\times) = \sum_1^N \gamma(t)\gamma(t)^T + e(t)\gamma(t)^T + \gamma(t)e(t)^T + \beta(t)\gamma(t)^T + \gamma(t)\beta(t)^T$$

Property (2.16) can now be applied to (A.5) with

$$A = \frac{1}{N} Q_N(S) \quad , \quad B = \frac{1}{N}\sum_1^N \beta(t)\beta(t)^T \quad \text{and} \quad C_\varepsilon = \frac{1}{N}Q_N^{(1)}(\theta^\times) + \frac{1}{N}Q_N^{(2)}(\theta,\theta^\times)$$

First introduce a countable subset $\tilde{D}_M$ of D_M that is dense in D_M. Also introduce a subset $\Omega^\times$ of the sample space

$$\Omega^\times = \Omega_1 \cap \prod_{\theta^\times \in \tilde{D}_M} \Omega_2(\theta^\times) \cap \Omega_3 \cap \prod_{\theta^\times \in \tilde{D}_M} \Omega_4(\theta^\times)$$

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

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

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

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

Since $\theta^\times \notin D_I$,

$$\text{tr } B = \frac{1}{N}\sum \beta(t)^T\beta(t) > \delta_2(\theta^\times) \quad \text{for} \quad N > N_4(\omega)$$

According to (2.16) there is an $\varepsilon_0 = \varepsilon_0(\text{tr}B, \delta) = \varepsilon_0(\delta_2(\theta^\times), \delta)$. Choose $N > N_0(\theta^\times, \omega) = \max\{N_4(\omega), N_3(\varepsilon_0/2, \omega, \theta^\times), N_2(\omega, p(\delta)\ \delta_2(\theta^\times)/2)\}$ and choose

$$\rho < \rho^\times(\theta^\times) = \varepsilon_0 / 2C_2(\omega, \theta^\times)$$

Then

$$V_N(\theta) = h\Big(\frac{1}{N} Q_N(S) + \frac{1}{N} \sum_1^N \beta(t)\beta(t)^T + \frac{1}{N}Q_N^{(1)}(\theta^\times) + \frac{1}{N}Q_N^{(2)}(\theta,\theta^\times) \Big) \geq$$

$$\geq h\Big(\frac{1}{N} Q_N(S)\Big) + p(\delta)\, \delta_2(\theta^\times) = V_N(S) + p(\delta)\delta_2(\theta^\times) > V_N(\tilde{\theta}) + p(\delta)\delta_2(\theta^\times)/2$$

for $N > N_0(\theta^\times,\omega)$ and $\theta\, \varepsilon B(\theta^\times,\rho^\times(\theta^\times))$.

Hence

$$\inf_{\theta\varepsilon B(\theta^\times,\rho^\times)} V_N(\theta) \geq V_N(\tilde{\theta}) + p(\delta)\, \delta_2(\theta^\times)/2 \quad \text{for } N > N_0(\theta^\times,\omega) \tag{A.8}$$

The family of open sets

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

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

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

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

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

It then follows from (A.8) 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.

$$|\theta(N) - D_I| < \varepsilon \quad \text{for } N > N_0(\omega,\varepsilon)$$

which, since ε is an arbitrary small number, is the conclusion of the theorem.

A.2 Proof of Theorem 4.1

The theorem follows from Lemma 4.1 if (4.3) and (4.4) can be shown to be satisfied.

We will consider the general case, and let $\alpha_{M(\theta)}(z)$ be the least common denominator to the denominator polynomials in $G_{M(\theta)}(z)$ and $H_{M(\theta)}(z)$. Introduce the matrix polynomials

$$C_{M(\theta)}(z) = \alpha_{M(\theta)}(z)\ H_{M(\theta)}(z)$$

$$B_{M(\theta)}(z) = \alpha_{M(\theta)}(z)\ G_{M(\theta)}(z)$$

Due to the assumptions, det $C_{M(\theta)}(z)$ has all zeroes outside the unit circle for $\theta \varepsilon D_M$, i.e. $C^{-1}_{M(\theta)}(z)$ is an exponentially stable filter. (There are several pole-zero cancellations between α and det H. Therefore the requirement that α has all zeros outside the unit circle is sufficient, but not necessary. Stability of C^{-1}_{M} is the only thing that matters.)

From (2.14)

$$C_{M(\theta)}(q^{-1})\ \hat{y}(t|\theta) = \left(C_{M(\theta)}(q^{-1}) - \alpha_{M(\theta)}(q^{-1})I\right)\ y(t) + B_{M(\theta)}(q^{-1})\ u(t-1) \qquad (A.9)$$

and, with $C'(q^{-1}) = \frac{d}{d\theta} C_{M(\theta)}(q^{-1})$ etc, and suppressed q^{-1},

$$C_{M(\theta)} \frac{d}{d\theta} \hat{y}(t|\theta) = (C' - \alpha' I)\ y(t) + B'\ u(t-1) - C'\ \hat{y}(t|\theta) \qquad (A.10)$$

From (A.9) and (A.10) it follows that

$$\hat{y}(t|\theta) = \sum_{k=0}^{t} \left(h^{(1)}_{t,k}(\theta)\ y(t-k) + h^{(2)}_{t,k}(\theta)\ u(t-k)\right)$$

and

$$\frac{d}{d\theta} \hat{y}(t|\theta) = \sum_{k=0}^{t} \left(h^{(3)}_{t,k}(\theta)\ y(t-k) + h^{(4)}_{t,k}(\theta)\ u(t-k) + h^{(5)}_{t,k}(\theta)\ \hat{y}(t-k|\theta)\right)$$

Since $C_{M(\theta)}^{-1}(q^{-1})$ is exponentially stable for all $\theta\varepsilon D_M$, and D_M is compact

$$\sup_{\substack{\theta\varepsilon D_M \\ 1\leq i\leq 5}} |h_{t,k}^{(i)}(\theta)| < C_1\cdot\lambda_1^{\;k} \quad ; \quad \lambda_1 < 1. \tag{A.11}$$

and hence

$$\sup_{\theta\varepsilon D_M} |\hat{y}(t|\theta)| < C_1 \sum_{k=0}^{t} \lambda_1^{\;k} \{|y(t-k)| + |u(t-k)|\} \tag{A.11}$$

and

$$\sup_{\theta\varepsilon D_M} |\frac{d}{d\theta}\,\hat{y}(t|\theta)| < C_1 \sum_{k=0}^{t} (\lambda_1^{\;k} + k\lambda_1^{\;k})\{|y(t-k)|+ |u(t-k)|\} \tag{A.12}$$

Let C and λ ; $\lambda<1$, be such that

$$C_1(\lambda_1^{\;k} + k\lambda_1^{\;k}) < C\,\lambda^k$$

Introduce for brevity the notation

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

and define

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

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

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

or

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

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

$$\frac{1}{N}\sum_{1}^{N} \tilde{z}(t)^2 \leq \lambda^2 \frac{1}{N}\sum_{1}^{N} \tilde{z}(t)^2 + \frac{1}{N}\sum_{0}^{N} \eta(t)^2 + $$

$$+ 2\lambda \frac{1}{N}\sum_{0}^{N} z(t)\eta(t)$$

or

$$(1-\lambda^2)\frac{1}{N}\sum_{1}^{N} \tilde{z}(t)^2 \leq \frac{1}{N}\sum_{0}^{N} \eta(t)^2 + 2\lambda\left[\frac{1}{N}\sum_{0}^{N} \eta(t)^2 \cdot \frac{1}{N}\sum_{0}^{N} \tilde{z}(t)^2\right]^{1/2}$$

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 (4.3). Condition (4.4)follows from (A.12) and (4.7), in the same way since

$$\mathrm{tr}\, \frac{1}{N} Q_N = \frac{1}{N}\sum_{1}^{N} |y(t) - \hat{y}(t|\theta)|^2_{R(t)} \leq \frac{2}{N}\sum_{1}^{N} |y(t)|^2_{R(t)} + \frac{2}{N}\sum_{1}^{N} |\hat{y}(t|\theta)|^2_{R(t)}$$

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.

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

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. (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 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.

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., and Rissanen, J. (1974)
Maximum Likelihood Estimation of Parameters in Multivariable Gaussian Stochastic Processes.
IEEE Trans. IT-<u>20</u>, 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.

Jazwinski, A.H. (1970)
Stochastic Processes and Filtering Theory.
Academic Press, New York.

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

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

Ljung, L. (1974)
On Consisten cy for Prediction Error Identification Methods.
Report 7405, Division of Automatic Control, Lund Institute of Technology.

Ljung, L., Gustavsson, I., and Soderstrom, T. (1974)
Indentification of Linear Multivariable Systems Operating Under Linear Feedback Control.
IEEE Trans., AC-19, No 6 (Special Issue on System Identification and Time series analysis) pp 836-841.

Luders, G., and Narendra, K.S. (1974)
Stable Adaptive Schemes for State Estimation and Identification of Linear Systems.
IEEE Trans., AC-19 No 6 (Special Issue on System Identification and Time Series Analysis) pp 841-848.

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.

Mendel, J.M. (1973)
Discrete Techniques of Parameter Estimation: The Equation Error Formulation.
Marcel Dekker, New York.

Rissanen, J. (1974)
Basis of Invariants and Canonical Forms for Linear Dynamic Systems.
Automatica, Vol 10, No. 2 pp.175-182.

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

Soudack, A.C., Suryanarayanan, K.L., and Rao, S.G. (1971)
A Unified Approach to Discrete-Time System Identification.
Int. J. of Control, Vol 14, No 6, pp 1009-1029.

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

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.