On Consistency and Identifiability

Ljung, Lennart

1975

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

ON CONSISTENCY AND IDENTIFIABILITY

L. LJUNG

Report 7521(C) June 1975
Department of Automatic Control
Lund Institute of Technology

ON CONSISTENCY AND IDENTIFIABILITY *

Lennart Ljung +

ABSTRACT

The convergence, with probability one, of the parameter estimates obtained from prediction error identification methods, such as the maximum likelihood method, is analysed in this paper. It is shown that under quite weak assumptions on the actual system, that has generated the data, the expected value of the identification criterion can be used for the asymptotic analysis of the estimates. In particular, does not the true system have to belong to the set of models over which the search for optimum is made. The implications of this result for consistency analysis and for questions of identifiability, as well as for other related problems are discussed.

---

* This work was supported in part by the Air Force Office of Scientific Research, AF Systems Command, under contract AF 44-620-69-C-0101, the Joint Services Electronics Program under contract N-00014-67-A-0112-0044, and the Industrial Affiliates Program at the Information Systems Laboratory, Stanford University.

+ Division of Automatic Control, Lund Institute of Tecnology, S-220 07 LUND, Sweden. At present with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305.

1. INTRODUCTION

The identification problem is to determine a dynamic model that (according to some criterion) as well as possible describes the input-output data measured from some process. Once a certain method to solve this problem has been deviced it is natural to test its performance in various ways. The tests can be numerical, like when the method is applied to data simulated on a computer. A particularly common analytical test is to study the asymptotic behaviour of the method (and of the estimates that it produces) as the number of measured data tends to infinity. Since the data often are considered to be random processes, such analysis has to be performed using probabilistic methods. The concepts of consistency and identifiability are closely related to such analysis and to the limits of the estimates (if they exist), as is further explained in Section 3.

The class of identification methods to be studied here are defined as procedures that minimize the prediction error of the model, when applied to the recorded data. This class contains min-max entropy methods, and under certain assumptions on the statistics, the Maximum-likelihood method. These methods have attracted much interest and have shown good performance in practical applications, [1], [7].

In Section 2 we shall define the class of identification methods and the set of models formally, while in Section 3 the concepts of identifiability and consistency are discussed. Section 4 reviews some results on consistency of these methods and in Section 5 a general result on the asymptotic behaviour of them is proved. The implications of this result are discussed in Section 6.

2. MODELS AND IDENTIFICATION CRITERION

Loosely speaking, the identification setup is entirely determined by three entities: the data, the set of models, and the identification criterion. The identification procedure, then, is to determine that (those) element in the set of models that gives the best fit to the measured data according to the chosen criterion. In this section we shall discuss some different set of models to be used throughout the paper and also define a class of identification criteria.

2.1 Models

We shall generally denote a specific model by $\mathcal{M}(\theta)$ where θ is some parameter vector belonging to a given set $D_{\mathcal{M}}$. As θ varies over $D_{\mathcal{M}}$, $\mathcal{M}(\theta)$ describes a set of models, which will be denoted by $\mathcal{M}$:

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

In this paper we shall only consider linear models. Some results valid for more general models are given in [2] and [3].

Example 1. - Linear Models in State Space Representation.

The state space representation is a common and convenient way of describing linear, time-varying systems. The input-output relation for the model $\mathcal{M}(\theta)$ is then defined by

$$\begin{aligned} x_\theta(t+1) &= F_\theta(t)\, x_\theta(t) + G_\theta(t)\, u(t) + e(t) \\ y(t) &= H_\theta(t)\, x_\theta(t) + v(t) \end{aligned} \qquad \theta \in D_{\mathcal{M}} \qquad (1)$$

where $e(\cdot)$ and $v(\cdot)$ are random processes with zero means and covariances

$$E\, e(t)e'(s) = Q_\theta(t)\, \delta_{ts} \quad ; \quad E\, e(t)v'(s) = R^c_\theta(t)\, \delta_{ts} \quad ;$$

$$E\, v(t)v'(s) = R_\theta(t)\, \delta_{ts} .$$

We shall have reason to be interested in the linear least squares estimate of $y(t+1)$ given $y(s)$, $u(s)$, $s=0,\ldots,t$ and some initial estimate $\hat{x}(0|\theta)$ with error covariance $P(0|\theta)$, and given that the model $\mathcal{M}(\theta)$, (1), is a true description of the data. This estimate will be denoted by $\hat{y}(t+1|\theta)$ and is obtained from standard Kalman filtering,

$$\hat{y}(t+1|\theta) = H_\theta(t+1)\,\hat{x}(t+1|\theta) \tag{2a}$$

where

$$\hat{x}(t+1|\theta) = F_\theta(t)\,\hat{x}(t|\theta) + G_\theta(t)\,u(t) + K_\theta(t)\,[y(t) - H_\theta(t)\,\hat{x}(t|\theta)] \tag{2b}$$

$K_\theta(t)$ is the Kalman gain matrix, determined as

$$K_\theta(t) = [F_\theta(t)P(t|\theta)H_\theta'(t) + R_\theta^c(t)]\,[H_\theta(t)P(t|\theta)H_\theta'(t) + R_\theta(t)]^{-1} \tag{2c}$$

$$P(t+1|\theta) = F_\theta(t)P(t|\theta)F_\theta'(t) - K_\theta(t)\,[H_\theta(t)P(t|\theta)H_\theta'(t) + R_\theta(t)]^{-1}\,K_\theta'(t) + Q_\theta(t) \tag{2d}$$

The initial values $\hat{x}(0|\theta)$ and $P(0|\theta)$ can either be known or parameterized in an arbitrary way by θ. We shall not be much concerned with the way in which the matrices F,G,H,Q,R,R^c, $\hat{x}(0|\theta)$ and $P(0|\theta)$ are determined from θ, but we shall assume that the matrix-elements are continuously differentiable functions of θ. Apart from this assumption, the unknown elements may enter quite arbitrarily in the matrices. Some elements may be known from basic physical laws, or a priori fixed like in canonical representations. Other elements may be related to each other etc. The important thing is that θ is a finite dimensional, time-invariant parameter that determines all the matrices for all $t \geq 0$.

Example 2. - General Linear, Time-Invariant Models

A linear time-invariant model can be described as

$$y(t+1) = \mathcal{G}_\theta(q^{-1})\, u(t) + \mathcal{H}_\theta(q^{-1})\, e(t+1) \qquad (3)$$

where q^{-1} is the backward shift operator: $q^{-1}\, u(t) = u(t-1)$ and $\mathcal{G}_\theta(z)$ and $\mathcal{H}_\theta(z)$ are matrix functions of z (z replaces q^{-1}). The variables $e(\cdot)$ are assumed to be independent random variables with zero mean values and covariance matrices $E\, e(t)e'(t) = \Lambda_\theta$ (which actually may be time-varying). It will be assumed that $\mathcal{G}_\theta(z)$ and $\mathcal{H}_\theta(z)$ are matrices with rational functions of z as entries and that $\mathcal{H}_\theta(0) = I$. The latter assumption implies that $e(t)$ has the same dimension as $y(t)$, but this is no loss of generality.

To find the linear least squares estimate $\hat{y}(t+1|\,\theta)$ from (3) requires some caution regarding the initial values. In general the filter determining the estimate $\hat{y}(t+1|\,\theta)$ will be time-varying, even though (3) is time-invariant. In such a case a state space representation can be used. A simpler approach is to assume that information , equivalent to knowing all previous $y(t)$, $u(t)$, $t<0$, is available and that hence the prediction filter has reached stationarity. It will follow from the analysis in the following sections that this assumption is quite relevant for identification problems.

From (3) we have

$$\mathcal{H}_\theta^{-1}(q^{-1})\, y(t+1) = \mathcal{H}_\theta^{-1}(q^{-1})\, \mathcal{G}_\theta(q^{-1})\, u(t) + e(t+1)$$

and

$$y(t+1) = [I - \mathcal{H}_\theta^{-1}(q^{-1})]\, y(t+1) + \mathcal{H}_\theta^{-1}(q^{-1})\, \mathcal{G}_\theta(q^{-1})\, u(t) + e(t+1). \qquad (4)$$

Since $\mathcal{H}_\theta^{-1}(0) = I$, the right hand side of (4) 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(t) is determined from output feedback. Hence, if we assume that all previous y(s) and u(s) are known, we have

$$\hat{y}(t+1|\theta) = [I - H_\theta^{-1}(q^{-1})]\, y(t+1) + H_\theta^{-1}(q^{-1})\, G_\theta(q^{-1})\, u(t) \tag{5}$$

which in this case also equals the conditional mean.

Now, linear systems are often not modelled directly in terms of the impulse response functions $G_\theta(z)$ and $H_\theta(z)$. A frequently used representation is the vector difference equation (VDE):

$$A_\theta(q^{-1})\, y(t+1) = B_\theta(q^{-1})\, u(t) + C_\theta(q^{-1})\, e(t+1). \tag{6}$$

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

$$\begin{aligned} x_\theta(t+1) &= F_\theta x_\theta(t) + G_\theta\, u(t) + K_\theta\, e(t) \\ y(t) &= H_\theta\, x_\theta(t) + e(t) \end{aligned} \tag{7}$$

It is easily seen that these two representations correspond to

$$G_\theta(z) = A_\theta^{-1}(z)\, B_\theta(z) \quad ; \qquad H_\theta(z) = A_\theta^{-1}(z)\, C_\theta(z) \tag{8}$$

and

$$G_\theta(z) = H_\theta[I - zF_\theta]^{-1}\, G_\theta \qquad H_\theta(z) = z\, H_\theta[I - zF_\theta]^{-1}\, K_\theta + I \tag{9}$$

respectively.

Inserting (8) into (5) it is seen that $\hat{y}(t+1|\theta)$ is found as the solution of

$$C_\theta(q^{-1})\, \hat{y}(t+1|\theta) = [C_\theta(q^{-1}) - A_\theta(q^{-1})]\, y(t+1) + B_\theta(q^{-1})\, u(t) \tag{10}$$

for the case of a VDE-model. For the state space model (7), $\hat{y}(t+1|\theta)$ is found from

$$\hat{x}_\theta(t+1) = F_\theta \hat{x}_\theta(t) + G_\theta u(t) + K_\theta[y(t) - H_\theta \hat{x}_\theta(t)]$$
$$\hat{y}(t+1|\theta) = H_\theta \hat{x}_\theta(t+1) \qquad (11)$$

We shall also in this case assume that the matrix elements of $\mathcal{G}_\theta$ and $\mathcal{H}_\theta$ (and $A_\theta(z)$, $B_\theta(z)$, $C_\theta(z)$, F_θ, G_θ, H_θ and K_θ) are continuously differentiable with respect to θ, but apart from that the parameter vector θ may enter arbitrarily in the matrices. ■

Remark. From (10) and (11) it is seen that certain initial information is required to start up the algorithms, namely for (10) $\{y(0), \ldots, y(-N), u(0), \ldots, u(-N), \hat{y}(0|\theta), \ldots, \hat{y}(-N|\theta)\}$ and for (11) $\hat{x}_\theta(0)$. In many cases it is not feasible to assume that these are known. Therefore they should be parametrized by the parameter vector θ. However, nothing prevents us from taking trivial parametrizations, like $\hat{x}_\theta(0) = 0$ for all $\theta \in D_{\mathcal{M}}$, etc., since we shall not introduce the requirement that there is a θ in $D_{\mathcal{M}}$ that corresponds to a "true" description of the data. We shall often, for notational convenience, also suppress the initial values in explicit formulas (i.e. suppose that we have the above "trivial" parameterization).

In these two examples the predicted value $\hat{y}(t+1|\theta)$ is obtained by linear filtering operations on $y(\cdot)$ and $u(\cdot)$,

$$\hat{y}(t+1|\theta) = \sum_{s=0}^{t} [h_{t,s}(\theta)\, y(s) + f_{t,s}(\theta)\, u(s)]. \qquad (12)$$

Since the coefficients of this filter are continuously differentiable with respect to the system matrix parameters, we have

$$\frac{d}{d\theta}\, \hat{y}(t+1|\theta) = \sum_{s=0}^{t} \left[\frac{d}{dt} h_{t,s}(\theta)\, y(s) + \frac{d}{dt} f_{t,s}(\theta)\, u(s) \right] \qquad (13)$$

We shall be particularly interested in the case where the linear filters (12) as well as (13) are exponentially stable. The set of those θ yielding this property will be denoted by $D_S(\mathcal{M})$. It is easy to see that for models described by (6)

$$D_S(\mathcal{M}) = \{\theta \mid \det C_\theta(z) = 0 \Rightarrow |z| > 1 \} \tag{14}$$

and for models described by (7),

$$D_S(\mathcal{M}) = \{\theta \mid F_\theta - K_\theta H_\theta \text{ has all eigenvalues in } |z| < 1 \}. \tag{15}$$

Moreover, for the general time-varying model (1), the well known stability properties of the Kalman filter, see, e.g., Jazwinski[4,Thm 7.4], implies that

$$D_S(\mathcal{M}) = \{\theta \mid [F_\theta(\cdot), Q_\theta(\cdot), H_\theta(\cdot)] \text{ is completely uniformly controllable and observable}\} . \tag{16}$$

Furthermore, it follows that over compact subsets of these D_S, the base of the exponential decay of the filter coefficients, is uniformly bounded by a constant strictly less than 1.

2.2 A Class of Identification Criteria

From the linear least squares predictions $\hat{y}(t+1|\theta)$ and the data we can form the following matrix

$$Q_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} [y(t) - \hat{y}(t|\theta)][y(t) - \hat{y}(t|\theta)]' \tag{17}$$

This matrix is a measure of how well the model $\mathcal{M}(\theta)$ is able to describe the recorded data.

Remark. In some cases there might be reason to study a weighted version of (17),

$$Q_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} [\sqrt{R(t)}(y(t) - \hat{y}(t|\theta))][\sqrt{R(t)}(y(t) - \hat{y}(t|\theta))]' \qquad (18)$$

where $R(\cdot)$ is a sequence of positive semidefinite matrices. However, this can also be seen as a rescaling of the output, and we shall confine ourselves, for reasons of notational convenience, to the case (17).

It is reasonable to take as the identification criterion some continuous function $h(\cdot)$ of $Q_N(\theta)$:

$$V_N(\theta) = h[Q_N(\theta)] \qquad (19)$$

The parameter estimate based on N measurements, $\hat{\theta}(N)$, is thus taken as the θ that minimizes $V_N(\theta)$ over $D_{\mathcal{M}}$, and the corresponding model is taken as $\mathcal{M}(\hat{\theta}(N))$.

For the minimization to make sense, some simple properties of $h(\cdot)$ should be required, essentially that $h(\cdot)$ retains an ordering property among the matrices, see [3].

The identification criterion (19) has in itself a good physical interpretation: To choose that model that has the best prediction performance when applied to the data. Moreover, cf. [28], if the innovations of the models in Examples 1 and 2 are Gaussian with covariances $\Lambda(\cdot)$, then it is well known that the log likelihood function for the problem is

$$-\frac{1}{2} \sum_{t=1}^{N} [y(t) - \hat{y}(t|\theta)]' \Lambda^{-1}(t) [y(t) - \hat{y}(t|\theta)] - \frac{N}{2} \log 2\pi -$$

$$- \frac{1}{2} \sum_{t=1}^{N} \log \det \Lambda(t) \qquad (20)$$

This holds even if there is non-linear output feedback present in the system. If $\Lambda(\cdot)$ are known, then maximizing (20) with respect to θ

is equivalent to minimizing tr $\tilde{Q}_N(\theta)$ $[R(t) = \Lambda^{-1}(t)]$. If Λ does not depend on t, but is unknown, then the maximization of (20) with respect to Λ can be performed analytically, see Eaton [5], and θ is found by minimizing

$$\det Q_N(\theta). \tag{21}$$

In case the distribution of the innovations is unknown, (21) is the θ-dependent term in

$$\max \; H[y(\cdot) - \hat{y}(\cdot|\,\theta)]$$

where H is the entropy of the prediction error and where the maximization is over all possible distributions, with the constraint that the covariance equals the sample covariance $Q_N(\theta)$, see, e.g., Rissanen [6].

Consequently, the class of prediction error identification methods, defined by minimization of (19), contains the maximum-likelihood method as well as min-max entropy methods. The criterion (19) was first suggested and applied to system identification problems in [7], and has after that successfully been applied to numerous practical identification problems.

3. IDENTIFIABILITY AND CONSISTENCY

The concept of identifiability has been given several different definitions in the literature, and we shall here briefly discuss a few of them.

We may distinguish two major approaches. The apparently most common approach is to relate the identifiability property to consistency of the parameter estimate $\hat{\theta}_N$. We shall label this approach as "consistency-oriented identifiability definitions". The "true" parameter θ_0 is then said to be identifiable if the sequence of estimates $\hat{\theta}_N$ converges to θ_0 in some stochastic sense. This is the path followed e.g. in Åström-Bohlin [7] (convergence with probability one), Staley-Yue [8] (convergence in the mean square sense) and in Tse-Anton [9] (convergence in probability). A somewhat different definition is used in Ljung et.al. [10] and in Ljung [3]. There a set

$$D_T(\mathcal{S},\mathcal{M}) = \left\{ \theta \;\middle|\; \theta \in D_{\mathcal{M}},\ \lim_{N\to} \frac{1}{N} \sum_{1}^{N} |\hat{y}(t|\mathcal{S}) - \hat{y}(t|\theta)|^2 = 0 \right. \text{ for all bounded inputs } u(\cdot) \left. \right\} \tag{22}$$

is defined, where $\hat{y}(t|\mathcal{S})$ is the true prediction of the system $\mathcal{S}$. Then $\mathcal{S}$ is said to be System Identifiable if $\hat{\theta}_N \to D_T(\mathcal{S},\mathcal{M})$ with probability one as $N \to \infty$ and to be Parameter Identifiable if, in addition, $D_T(\mathcal{S},\mathcal{M})$ consists of only one point. Although this definition makes no reference to any "true" parameter value θ_0, it should be regarded as "consistency-oriented", since the requirement that $D_T(\mathcal{S},\mathcal{M})$ is non-empty implies that there is a "very good model" available among the set of models $\mathcal{M}$. Indeed, if $D_{\mathcal{M}}$ contains a "true" parameter θ_0, then this definition of Parameter Identifiability is equivalent to the one first given.

These definitions require that the true system allows an exact description within the model set. In practice this is usually not a very realistic assumption, since almost any real-life process is more complex than we would allow our model to be. However, even if the set of models

does not contain the true system, questions of identifiability of the model parameters are still relevant. One could think of a state space model like (1) where all the matrices are filled with parameters. Even if the data are furnished by an infinitely complex system, it will not be possible to identify the parameters of the model, simply because several models give exactly the same fit, i.e., the identification criterion $V_N(\theta)$ does not have a unique minimum.

This leads us to "uniqueness-oriented identifiability definitions", like in Bellman-Åström [11], where a model set is said to be (globally) identifiable, if the identification criterion used as a unique global minimum. A complication in the present context is that the identification criterion is a random function and a bit awkward to handle. We would be much better off if $V_N(\theta)$ converges (with probability one) to a deterministic function (or asymptotically behaves like one). Let us already here remark that such convergence must be uniform in θ, in order to enable us to relate minima of $V_N(\theta)$ to minima of the deterministic function. We shall have occasion to return to this point below.

In addition to the references mentioned above, interesting results can also be found in, e.g., [29], [30] and [31].

4. SOME CONSISTENCY RESULTS

The consistency problem for the maximum likelihood method has been quite widely studied. For independent observations the consistency has been studied by, e.g., Cramer [12], Wald [13] and Kendall-Stuart [14]. The application of the maximum likelihood method to system identification (for single input - single output models on a difference equation form) was introduced in Åström-Bohlin [7], where it also is shown how the assumption on independent observations can be relaxed. Applications to other (linear) model choices have been considered in, e.g., Caines [15], Aoki-Yue [16], Balakrishnan [17], Spain [18], Tse-Anton [9] and Caines-Rissanen [19].

However, it should be remarked that several of the proofs on strong consistency (convergence with probability one to the true parameter value) are not complete, a fact that can be traced back to a shortcoming in the proof in [14]. The first complete strong consistency proofs for applications to system identification seem to be given in [2] and [20].

Let us cite, for future discussion, the following consistency result from [3] (Theorem 4.2 and lemma 5.1),

Theorem 1. Consider the set of models described in Example 1. Assume that $D_{\mathcal{M}}$, over which the search in θ is performed, is a compact subset of $D_s(\mathcal{M})$ (cf. (16)), and is such that $D_T(\delta,\mathcal{M})$ defined by (22) is non-empty. Assume that the actual system (with possible feedback terms) is exponentially stable and that the innovations of its output have bounded variance and are of full rank. Then, the identification estimate $\hat{\theta}_N$ that minimizes the criterion (19) converges into

$$D_I = \Big\{ \theta \;|\; \theta \in D_{\mathcal{M}};\ \liminf_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} E\,|\hat{y}(t|\delta) - \hat{y}(t|\theta)|^2 = 0 \text{ for the actual input to the process} \Big\} \tag{23}$$

with probability one as N tends to infinity.

This result is rather general and is not based on any ergodicity assumptions.

To ensure parameter consistency, it should be required first that the actual input during the identification experiment was sufficiently general so that

$$D_I \subset D_T(\mathcal{S},\mathcal{M})$$

holds (which implies "System Identifiability"),and secondly, that the model is suitably paramterized so that

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

holds. It is convenient to study these conditions separately.

The restrictive assumption in the theorem apparently is that $D_T(\mathcal{S},\mathcal{M})$ be non-empty. This requires the true system to be "not too complex" and is rarely met for real life processes. However, the philosophy of consistency results should be viewed as a test of the method: If the method is unable to recognize the true system in a family of models, then it is probably not a good method. The same philosophy clearly lies behind testing identification methods on simulated data.

It should however be noted, that from such consistency results, stricly speaking nothing can be stated about the performance of the method when applied to a system that does not have en exact description within the set of models.

5. A LIMIT RESULT FOR THE CRITERION FUNCTION

In this section we shall give results for determining the asymptotic behaviour of the estimates $\hat{\theta}_N$ that minimize the criterion function (19), $V_N(\theta)$, also in the case where the true system is more complex than can be described within the set of models. We shall do that by giving conditions under which

$$\bar{V}_N(\theta) = h[EQ_N(\theta)]$$

can be used for the asymptotic analysis. Thereby "the stochastic part of the problem" is removed and the analysis can proceed with the deterministic loss function $\bar{V}_N(\theta)$.

In order to make the analysis as general as possible, we would like to impose as weak conditions as possible upon the actual system. The important property we need is a stability property, but in order to state it, we shall assume that the true system with (possibly adaptive) feedback admits a description as follows,

$$\begin{aligned} x(t+1) &= f[t;x(t),u(t),e(t)] \\ y(t) &= g[t;x(t),e(t)] \\ u(t) &= h[t;x(t), \ldots ,x(0), u_R(t)] \end{aligned} \tag{24}$$

where $y(\cdot)$ is the output, $u(\cdot)$ the actual input to the process, $u_R(\cdot)$ a reference (extra) input, or noise in the feedback and $e(\cdot)$ is a sequence of independent random variables. The over-all stability property which we shall require is the following,

Define $y_s^0(\cdot)$ and $u_s^0(\cdot)$ through

$$\begin{aligned} x_s^0(t+1) &= f[t;x_s^0(t),u_s^0(t), e(t)] \\ y_s^0(t) &= g[t;x_s^0(t),e(t)] \\ u_s^0(t) &= h[t;x_s^0(t), \ldots , x_s^0(s),0, \ldots ,0, u_R(t)] \\ x_s^0(s) &= 0 \text{ (or any value independent of } e(r),\ r< s\ ) \end{aligned} \tag{25a}$$

Then the property is

$$E\,|y(t) - y_s^0(t)|^4 < C\,\lambda^{t-s} \quad , \quad E\,|u(t) - u_s^0(t)|^4 < C\,\lambda^{t-s} \quad ; \lambda<1,\ t<s \tag{25b}$$

and

$$E\,|y(t)|^4 < C \quad , \quad E\,|u(t)|^4 < C \tag{25c}$$

All expectations are over $e(\cdot)$.

The assumptions (24) and (25) are quite weak, in particular as we shall not need to specify the description (24).

We now have the following result.

Lemma. Let the set of models be defined by (1) or (3) (which includes in particular (6) and (7)), and assume that the actual process is subject to (24) and (25). Then

$$\sup_{\theta \in \bar{D}_s} |Q_N(\theta) - E\,Q_N(\theta)| \to 0 \quad \text{with probability one as } N \to \infty \tag{26}$$

where $\bar{D}_s$ is a compact subset of $D_s(\mathcal{M})$ (cf. (14)-(16)), and $Q_N(\theta)$ is defined by (17). The expectation is over $e(\cdot)$ in (24).

The proof of the lemma is given in the appendix.

The lemma implies that, if $h(\cdot)$ is continuous, then $\bar{V}_N(\theta) = h[EQ_N(\theta)]$ will be arbitrarily close to $V_N(\theta)$ in the sup-norm, w.p.1, and hence that the local and global minima of $V_N(\theta)$ are arbitrarily close to those of $\bar{V}_N(\theta)$. In particular will the globally minimizing points of $\bar{V}_N(\theta)$ and $V_N(\theta)$ be arbitrarily close, and if

$$\bar{V}(\theta) = \lim_{N\to\infty} \bar{V}_N(\theta) \tag{27}$$

exists, then $\hat{\theta}_N$ will converge w.p.1 to the globally minimizing point(s) of $\bar{V}(\theta)$.

It is important to notice that these properties follow only since (26) holds uniformly in θ. If it is known only that $Q_N(\theta) - E\, Q_N(\theta) \to 0$ w.p.1 as $N \to \infty$ for all $\theta \in \bar{D}_s$ (as has been proved by several authors, although under more restrictive assumptions on the system), then the minimizing points of $V_N(\theta)$ and $\bar{V}_N(\theta)$ do not have to be close.

Moreover, in the lemma no assumptions on stationarity neither on the system nor on the model are introduced and $E\, Q_N(\theta)$ does not have to converge. However, if it does converge, then the asymptotic analysis can most conveniently be performed on $\bar{V}(\theta)$.

As the following simple example shows, the lemma does not hold without some kind of stability assumption on the true system.

Example 3. Consider the system

$$y(t+1) = b\, u(t) + e(t+1) \qquad \text{where} \quad b<1 \text{ and } e(t) \in N(0,1)$$

with (adaptive) feedback

$$u(t) = -\frac{1}{2}\,[1 + \operatorname{sign} y(1)\,]\, y(t) \quad ; \quad u(0) = 0.$$

Let the model be

$$y(t+1) = \hat{b}\, u(t) + e(t+1)$$

and the criterion is

$$V_N(\hat{b}) = \frac{1}{N} \sum [y(t+1) - \hat{b}\, u(t)]^2$$

Then

$$E\, V_N(\hat{b}) = \frac{1}{2}\left[\, 1 + \frac{1 + \hat{b}^2 - 2\, b\, \hat{b}}{1 - b^2}\,\right].$$

However, for realizations such that $y(1) > 0$

$$V_N(\hat{b}) \to \frac{1 + \hat{b}^2 - 2\, b\, \hat{b}}{1 - b^2} \qquad \text{w.p.1 as } N \to \infty$$

and

$$V_N(\hat{b}) \to 1 \qquad \text{w.p.1 as } N \to \infty$$

for $y(1) < 0$.

Clearly, this adaptive regulator does not yield the over-all stability property (25), since the effect of $y(1)$ lingers forever. This simple example could of course easily be handled by conditioning with respect to $y(1)$, but it illustrates the difficulties that may arise with adaptive regulators. For such applications it is sometimes helpful to avoid taking the expectation of the criterion function, cf. Ljung [3].

6. SOME APPLICATIONS OF THE LEMMA

We may consider the lemma of Section 5 as a basic tool for the convergence analysis of the estimates $\hat{\theta}(\cdot)$, and we shall in this section point out some potential applications of it.

6.1 Properties of the asymptotic estimates.

For the sake of definiteness, let us take $h[\cdot] = \mathrm{tr}$ and assume that

$$\bar{V}_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} E\, |y(t) - \hat{y}(t|\theta)|^2$$

converges to $\bar{V}(\theta)$ as N tends to infinity. It then folows that the estimate(s) $\hat{\theta}_N$ that globally minimizes $V_N(\theta)$, will w.p.1 tend to

$$D_L = \left\{ \vartheta \;:\; \bar{V}(\vartheta) = \inf_{\theta^* \in D_{\mathcal{M}}} V(\theta^*) \right\}$$

if $D_{\mathcal{M}} \subset \bar{D}_s$. Moreover, let

$$\hat{y}(t+1|\delta) = E[\, y(t+1) \,|\, \mathcal{Y}_t] \;, \text{ where } \mathcal{Y}_t = \left\{ y(s),\, u(s);\; s \le t \right\}$$

be the true prediction (which of course may be non-linear in $\mathcal{Y}_t$). Then

$$\hat{y}(t+1) = \hat{y}(t+1|\delta) + v(t+1)$$

where $v(\cdot)$ are the innovations, obeying

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

We now have

$$E\, |y(t+1) - \hat{y}(t+1|\theta)|^2 = E\, v'(t+1)\, v(t+1) + E\, 2v'(t+1)[\hat{y}(t+1|\delta) - \hat{y}(t+1|\theta)]$$

$$+ E\, |\hat{y}(t+1|\delta) - \hat{y}(t+1|\theta)|^2.$$

The second term of the right hand side obviously is zero, and hence the global minima of $\bar{V}(\theta)$ are also the global minima of

$$W(\theta) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} E \left| \hat{y}(t+1|\mathcal{S}) - \hat{y}(t+1|\theta) \right|^2 \qquad (28)$$

Consequently we have proved that $\hat{\theta}_N$ will tend to the global minimum of $W(\theta)$, no matter if this is unique or not. In other words, the limiting estimate will give a model that is the best approximation of the true system (in the sense of (28)) for the particular input used in the identification experiment. This is by no means a surprising result, but it has here been established under quite general conditions.

It can be remarked that it is this property that makes the identification method powerful in applications, rather than the consistency properties.

While these results do not follow from consistency analysis, it is of course possible to deduce consistency properties from the lemma. Hence, if $D_T(\mathcal{S},\mathcal{M})$ is non-empty, then $W(\theta)$ assumes the value 0 for some θ^*, which has to be its global minimum, and from this the theorem of Section 4 follows.

6.2 Identifiability Properties.

The limit function $\bar{V}(\theta)$ or $W(\theta)$ of the previous subsection can be used for determining the identifiability properties of a certain model set (parameterization) without reference to any true parameter values. Hence, a model parameterization, $\mathcal{M}$, can be said to be Parameter Identifiable, under given experimental conditions (input signal properties), for a given system $\mathcal{S}$, if $\bar{V}(\theta)$ [or $W(\theta)$] has a unique global minimum.

It is clear that this concept is appropriate and that the parameterization problem for multi-output structures is equally important even if the true system is "very complex".

The same holds for the identifiability properties under output feedback, cf.,e.g., Ljung et. al. [10] and Söderström et. al. [21], as is indicated in this simple example.

Example 4. Consider the model

$$y(t+1) + \hat{a}\, y(t) = \hat{b}\, u(t) + e(t+1) \tag{29}$$

with output feedback

$$u(t) = g\, y(t) \tag{30}$$

and the criterion

$$V_N(\hat{a},\hat{b}) = \frac{1}{N} \sum_{t=1}^{N} [y(t+1) + \hat{a}\, y(t) - \hat{b}\, u(t)]^2.$$

It is clear that, regardless of the true system,

$$\bar{V}_N(\hat{a},\hat{b}) = \bar{V}_N(\hat{a} + kg,\ \hat{b} + k) \qquad -\infty < k < \infty$$

and consequently $\bar{V}_N(\theta)$ cannot have a unique minimum. Hence, the model set (29) under the experimental condition (30) is never Parameter Identifiable, no matter what the true system might be. ∎

6.3 Local Minima of the Criterion Function.

If the numerical minimization of $V_N(\theta)$ is performed using a gradient method, the "false" local minima of $V_N(\theta)$ are potential traps for the algorithm,and it is a most interesting problem to analyse the conditions under which such local minima may or may not exist. Since $V_N(\theta)$ converges uniformly to $\bar{V}(\theta)$, it follows that a local minimum of $\bar{V}(\theta)$ will, w.p.1, for sufficiently large N correspond to a local minimum of $V_N(\theta)$ and vice versa. Therefore the analysis for local minima can be performed in terms of $\bar{V}(\theta)$ instead of $V_N(\theta)$, which of course is a great

simplification. In [22] and [23] several interesting results of this kind are given.

6.4 Certain Difficulties With Adaptive Regulators.

When the true system cannot be modelled exactly, the identification method will still make the best possible out of the situation by minimizing (28), as explained in Section 6.1. However, it should be realized that the minimum of $\bar{V}_N(\theta)$ in general depends on the actual input during the identfication experiment. If the input is determined as output feedback, the minimizing element $\hat{\theta}_N$ will depend on the feedback law. If now the regulator is adaptive, and the feedback law is determined from the current estimate, cf. [24], the analysis of the loss function $\bar{V}_N(\theta)$ becomes cumbersome. Let us consider the following simple example.

Example 5. Let the system be given by

$$y(t+1) + a\, y(t) = u(t) + e(t+1) + c\, e(t) \tag{31}$$

where $e(\cdot)$ is a sequence of independent random variables with unit variance, and let the model set be given by

$$y(t+1) + \hat{a}\, y(t) = u(t) + v(t+1)$$

where $v(\cdot)$ are assumed to be independent. Then the model set does not contain a true description of the system (31). Let the identification criterion be as in Example 3. The input to the system is determined as

$$u(t) = \hat{a}(t)\, y(t) \tag{32}$$

where $\hat{a}(t)$ is the value that minimizes $V_t(\hat{a})$. With a constant feedback (30), we would have as the asymptotic estimate

$$\hat{a} = a - \frac{c\,[1 - (a-g)^2]}{1 + c^2 - 2(a-g)c}$$

which clearly depends on the actual feedback coefficient g. When taking into account the adaptive feedback (32), the determination of $E\ V_N(\hat{a})$ becomes difficult, and it is even impossible to easily decide whether $E\ V_N(\hat{a})$ will converge or not as N increases. In fact, for adaptive regulators of this kind, $E\ V_N(\theta)$ may really fail to converge (without tending to infinity), cf. [24].

Our conclusion of this example is that although the lemma provides the tool for analysing the asymptotic behaviour of the estimate (regardless of convergence of $E\ V_N(\theta)$) even for these more complex problems, it may not be so easy to use.

An idea that, for these problems, seems more appropriate than to determine $E\ V_N(\theta)$, is to consider the conditional change in the criterion

$$E\left[V_{N+1}(\hat{\theta}) - V_N(\theta) \mid V_N(\cdot)\right]$$

since this quantity reflects the significance of the present control action for future estimates. By instead considering the expected change in the minimizing point $\hat{\theta}_N$ of $V_N(\theta)$, given $\hat{\theta}_N$, it is possible to track the estimate and analyse its asymptotic properties. This is, essentially, the approach that has been taken in [26], [25] and [24].

7. CONCLUSIONS

A particular aspect of the asymptotic analysis of the estimates obtained by minimization of the prediction error identification criterion (19) is the question of identifiability. By this can be meant that the identification method has the ability of recovering the true parameter values if they belong to the set over which the criterion is minimized, (or to yield a model that has an equivalent input-output mapping, "System Identifiability", see Section 3). It can also be meant that the estimate will converge (w.p.1, say) to a unique value, which, however, does not have to be related to any "true system". We have labeled these approaches as "consistency-oriented" and "uniqueness-oriented", respectively, and given some general results for either approach.

The lemma of section 5 should be regarded as the main contribution of this paper. It states that, under quite weak assumptions on the true system, the expected value of the loss function, which is a deterministic function, can be used for the asymptotic analysis. We have also indicated how this result can be used for identifiability and convergence studies, as well as for analysis of other related problems.

APPENDIX: PROOF OF THE LEMMA OF SECTION 5.

The idea of the proof is to show that

$$\sup \frac{1}{N} \sum_{t=1}^{N} [y(t)-\hat{y}(t|\theta)]_i[y(t)-\hat{y}(t|\theta)]_j - E[y(t)-\hat{y}(t|\theta)]_i[y(t)-\hat{y}(t|\theta)]_j$$

taken over a "small" open set, is arbitrarily small for large N, and then to extend this result to $\bar{D}_s$, using the Heine-Borel theorem. We shall need some properties of $\hat{y}_i(t|\theta)$ ["i" denotes the i:th component], and let us state these results as a lemma.

Lemma A. Let $B = B(\theta^*,\rho) = \{ \theta \mid |\theta-\theta^*|<\rho \}$. Then

$$E \sup \left|\frac{d}{d\theta} \hat{y}_i(t|\theta)\right|^2 < C(\theta^*) \qquad \text{(A1)}$$

Let

$$\eta_{ij}(t,\theta^*,\rho) = \sup_{\theta\in B} [y_i(t)-\hat{y}_i(t|\theta)][y_j(t)-\hat{y}_j(t|\theta)]$$

$$\zeta_{ij}(t,\theta^*,\rho) = \inf_{\theta\in B} [y_i(t)-\hat{y}_i(t|\theta)][y_j(t)-\hat{y}_j(t|\theta)]$$

Then

$$\frac{1}{N} \sum_{t=1}^{N} [\eta_{ij}(t,\theta^*,\rho) - E\,\eta_{ij}(t,\theta^*,\rho)] \to 0 \text{ w.p.1 as } N \to \infty \text{ for all } \theta^*$$

and ρ such that $B(\theta^*,\rho) \subset \bar{D}_s$.

[That is, for all θ^*, ρ and ϵ, there exists a $N_1(\theta^*,\rho,\epsilon,\omega)$ such that

$$\left| \frac{1}{N} \sum_{t=1}^{N} \eta_{ij}(t,\theta^*,\rho,\omega) - E\,\eta_{ij}(t,\theta^*,\rho,\omega) \right| < \epsilon \text{ for } N>N_1(\theta^*,\rho,\epsilon,\omega) \text{ for all } \omega\in\Omega(\theta^*), \text{ where } P[\Omega(\theta^*)]=1. \qquad \text{(A2)}$$

and similarly for $\zeta_{ij}(t,\theta^*,\rho)$.

Proof of lemma A: We shall throughout the proof use C and λ for constants, where $\lambda<1$, that do not need to be the same. We shall also allow ourselves to suppress arguments and subscripts freely, when there is no risk of confusion.

Let the linear filter determining $\hat{y}(t|\theta)$ be given by (12). For $B(\theta^*,\rho)\subset \bar{D}_s$ we then have

$$\sup_{\theta\in B} [\ |h_{t,k}(\theta)| + |f_{t,k}(\theta)|\ ] < C\ \lambda^{t-k} \tag{A3a}$$

and

$$\sup_{\theta\in B} [\ |\frac{d}{d\theta}h_{t,k}(\theta)| + |\frac{d}{d\theta}f_{t,k}(\theta)|\ ] < C\ \lambda^{t-k} \tag{A3b}$$

Hence

$$\sup\ |\frac{d}{d\theta}\hat{y}(t|\theta)| = \sup\left|\sum_{k=0}^{t} \frac{d}{d\theta}h_{t,k}(\theta)\ y(k) + \frac{d}{d\theta}f_{t,k}(\theta)\ u(k)\right| \leq$$

$$\leq \sum_{k=0}^{t} \sup|\frac{d}{d\theta}h_{t,k}(\theta)|\ |y(k)| + \sup|\frac{d}{d\theta}f_{t,k}(\theta)|\ |u(k)| < C\sum_{k=0}^{t}\lambda^{t-k}[|y(k)| + |u(k)|]$$

and

$$E\ \sup|\frac{d}{d\theta}\hat{y}(t|\theta)|^2 < C\sum_{k,s=0}^{t}\lambda^{2t-k-s}\ E[\ |y(k)| + |u(k)|\ ][\ |y(s)| + |u(s)|\ ] <$$

$$< C\sum_{k,s=0}^{t}\lambda^{2t-k-s}\ [\ \sqrt{E}\,|y(k)|^2 + \sqrt{E}\,u(k)^2][\ \sqrt{E}\,|y(s)|^2 + \sqrt{E}\,|u(s)|^2] \leq$$

$$< C\left(\sum_{k=0}^{t}\lambda^{t-k}\right)^2 < C. \tag{A4}$$

which proves (A1).

Let the variables $y_s^0(\cdot)$ and $u_s^0(\cdot)$ be defined as in (25a) and let the prediction based on these variables be denoted by

$$\hat{y}_s^0(t|\theta) = \sum_{k=s}^{t}\ [h_{t,k}(\theta)\ y_s^0(k) + f_{t,k}(\theta)\ u_s^0(k)]$$

The components of $\hat{y}_s^0(t|\theta)$ will be denoted by $\hat{y}_{i:s}^0(t|\theta)$. Let us also introduce

$$\eta_{ij:s}^0(t,\theta^*,\rho) = \sup_{\theta\in B}\ [y_{i:s}^0(t) - \hat{y}_{i:s}^0(t|\theta)][y_{j:s}^0(t) - \hat{y}_{j:s}^0(t|\theta)].$$

Notice that $\eta_s^0(t)$ is by definition independent of $\eta(r)$ for $r<s$.

Consider

$$\mu_s(t) = \eta(t) - \eta_s^0(t).$$

After elementary calculations we find

$$E\,\mu_s(t)^2 \leq$$

$$\leq 4\Big[E\sup|y-\hat{y}|^4\cdot[\,E\sup|y-y^0|^4 + E\sup|\hat{y}-\hat{y}^0|^4] + E\sup|y^0-\hat{y}^0|^4\cdot[\,E\sup|y^0-\hat{y}^0|^4 + E\sup|\hat{y}-\hat{y}^0|^4]\Big] \qquad (A5)$$

As above, (A4), we readily find that

$$E\sup|y-\hat{y}|^4 < C \quad \text{and} \quad E\sup|y^0-\hat{y}^0|^4 < C \qquad (A6)$$

and by (25b) we have $E\,|y(t)-y_s^0(t)|^4 < C\,\lambda^{t-s}$. Consider now, (A7)
(for subscript i or j)

$$\sup|\hat{y}(t|\theta) - \hat{y}_s^0(t|\theta)| = \sup\Big|\sum_{k=0}^{s}[h_{t,k}(\theta)y(k) + f_{t,k}(\theta)u(k)] + \sum_{k=s+1}^{t}[h_{t,k}(\theta)[y(k)-y_s^0(k)] + f_{t,k}(\theta)[u(k)-u_s^0(k)]]\Big| .$$

Using (A3a), we find after some straighforward algebra, and use of Schwarz' inequality (a detailed account is given in [2;pp94-95]),

$$E\sup|\hat{y}(t|\theta) - \hat{y}_s^0(t|\theta)|^4 < C\,\lambda^{t-s} \qquad (A8)$$

Collecting (A5), (A6), (A7) and (A8) we find,

$$E\,\mu_s(t)^2 < C\,\lambda^{t-s} . \qquad (A9)$$

Consider now for $t>s$,

$$\mathrm{Cov}\,(\,\eta(t),\,\eta(s)\,) = \mathrm{Cov}\,[\,\eta_s^0(t) + \mu_s(t),\,\eta(s)] = \mathrm{Cov}\,[\mu_s(t),\eta(s)] \leq$$

$$\leq \left(E\,\mu_s(t)^2 \cdot E\,\eta(s)^2\right)^{1/2} < C\,\lambda^{t-s} \qquad (A10)$$

where the second equality follows since $\eta_s^0(t)$ is independent of $\{e(r), r<s\}$, and hence of $\eta(s)$. Boundedness of $E\,\eta(s)^2$ follows as above, (A4).

In [27], the following convergence theorem is given : Let f_i be a sequence of random variables, with zero mean values and with

$$|E\,f_i\,f_j| < K\,\frac{i^p + j^p}{1+|i-j|^q} \qquad 0 \leq 2p < q < 1 \qquad (A11)$$

Then

$$\frac{1}{N}\sum_{i=1}^{N} f_i \to 0 \quad \text{w.p.1 as } N \to \infty \tag{A12}$$

We can now apply this result to

$$\frac{1}{N}\sum_{t=1}^{N} [\eta_{ij}(t,\theta^*,\rho) - E\,\eta_{ij}(t,\theta^*,\rho)]$$

since (A10) well implies (A11). This concludes the proof of lemma A.

Proof of Main Lemma: Let

$$z_j(t,\theta) = y_j(t) - \hat{y}_j(t|\theta)$$

and consider

$$r(\theta) = Q_N^{(ij)}(\theta) - E\,Q_N^{(ij)}(\theta^*) = \frac{1}{N}\sum_{t=1}^{N} [z_i(t,\theta)z_j(t,\theta) - E\,z_i(t,\theta^*)z_j(t,\theta^*)],$$

$$\sup_{\theta\in B}[r(\theta)] \le \frac{1}{N}\sum \sup[z_i(t,\theta)z_j(t,\theta)] - E\,z_i(t,\theta^*)z_j(t,\theta^*) = \tag{A13}$$

$$= \frac{1}{N}\sum [\eta_{ij}(t,\theta^*,\rho) - E\eta_{ij}(t,\theta^*,\rho)] + \frac{1}{N}\sum E[\eta_{ij}(t,\theta^*,\rho) - z_j(t,\theta^*)z_i(t,\theta^*)]$$

Using the mean value theorem and (A1), we readily obtain (cf. p.85-86 in [2]),

$$E\,|\eta_{ij}(t,\theta^*,\rho) - z_i(t,\theta^*)z_j(t,\theta^*)| < \rho\cdot C(\theta^*) \tag{A14}$$

Similarly,

$$E\,|z_i(t,\theta^*)z_j(t,\theta^*) - z_i(t,\theta)z_j(t,\theta)| < \rho\cdot C(\theta^*) \quad \text{for} \quad \theta\in B(\theta^*,\rho),$$

which implies

$$E\,|Q_N^{(ij)}(\theta) - E\,Q_N^{(ij)}(\theta^*)| < \rho\cdot C(\theta^*) \quad \text{for } \theta \in B(\theta^*,\rho). \tag{A15}$$

Collecting (A13), (A14) and (A15), we obtain

$$\sup_{\theta\in B}[Q_N^{(ij)}(\theta) - E\,Q_N^{(ij)}(\theta)] < \frac{1}{N}\sum_{t=1}^{N} [\eta_{ij}(t,\theta^*,\rho) - E\eta_{ij}(t,\theta^*,\rho)] + 2\rho C(\theta^*) \tag{A15}$$

Now choose an $\epsilon > 0$, and take for every $\theta^* \in \bar{D}_s$, the radius $\rho = \rho^* = \rho(\theta^*)$ to be the minimum of $\epsilon/4C(\theta^*)$ and the distance from θ^* to the boundary of D_s. Then for $N > N_1(\theta^*, \rho^*, \epsilon/2, \omega)$, the first term in the right hand side of (A15) is less than $\epsilon/2$ according to (A2), and hence

$$\sup_{\theta \in B(\theta^*, \rho^*)} [Q_N^{(ij)}(\theta) - E\, Q_N^{(ij)}(\theta)] < \epsilon \qquad \text{(A16)}$$

We shall now extend (20) to hold over $\bar{D}_s$, by applying Heine-Borel's theorem. Clearly, $\{B(\theta^*, \rho^*),\ \theta^* \in \bar{D}_s\}$ is a family of open sets covering $\bar{D}_s$. Select a finite family of sets $\{B(\theta_i, \rho_i),\ i=1, \ldots, M\}$ that also covers $\bar{D}_s$ and let

$$\bar{N}_1(\epsilon, \omega) = \max_{1 \le i \le M} N_1(\theta_i, \rho(\theta_i), \epsilon/2, \omega)$$

Then

$$\sup_{\theta \in \bar{D}_s} [Q_N^{(ij)}(\theta) - E\, Q_N^{(ij)}(\theta)] < \epsilon \quad \text{for } N > \bar{N}_1(\epsilon, \omega)$$

$$\text{for all } \omega \in \Omega' = \bigcap_{i=1}^{M} \Omega(\theta_i), \text{ where } P(\Omega') = 1.$$

Similarily,

$$\sup [EQ_N^{(ij)}(\theta) - Q_N^{(ij)}(\theta)] < \epsilon \quad \text{for } N > \bar{N}_2(\epsilon, \omega)$$

which concludes the proof of the main lemma. ∎

REFERENCES

[1] K.J. Åström and P. Eykhoff, "System Identification - A Survey", Automatica, Vol 7 (1971), pp.123-162.

[2] L. Ljung, "On Consistency for Prediction Error Identification Methods", Report 7405, Division of Automatic Control, Lund Institute of Technology, Lund, Sweden, March 1974.

[3] L. Ljung, "On the Consistency of Prediction Error Identification Methods", to appear in System Identification: Advances and Case Studies (D.G. Lainiotis and R.K. Mehra, Eds), Marcel Dekker, Inc.,N.Y.

[4] A.H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press 1970, New York.

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

[6] J. Rissanen, "Minmax Entropy Estimation of Models for Vector Processes", to appear in System Identification: Advances and Case Studies (D.G. Lainiotis and R.K. Mehra, Eds), Marcel Dekker, Inc.,N.Y.

[7] K.J. Åström and T. Bohlin, "Numerical Identification of Linear Dynamic Systems from Normal Operating Records", IFAC Symposium on Self-Adaptive Systems, Teddington, England 1965. Also in Theory of Self-Adaptive Control Systems (P.H. Hammond, Ed), Plenum Press, New York 1965.

[8] R.M. Staley and P.C. Yue, "On System Parameter Identifiability", Information Sciences, Vol 2, No 2 (1970), pp.127-138.

[9] E. Tse and J. Anton, "On the Identifiability of Parameters", IEEE Trans. Autom. Control, Vol AC-17, No 5 (1972).

[10] L. Ljung, I. Gustavsson and T. Söderström, "Identification of Linear Multivariable Systems Operating Under Linear Feedback Control", IEEE Trans. Autom.Control, Vol AC-19,No 6(1974),pp.836-841.

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

[12] H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton 1946.

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

[14] M.G. Kendall and A. Stuart, The Advanced Theory of Statistics, Vol.2. Hafner Publishing Co., New York, 1967.

[15] P.E. Caines, "The Parameter Estimation of State Variable Models of Multivariable Linear Systems", Proc. U.K.A.C. Conf. on Multivariable Systems, Manchester, England 1971.

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

[17] A.V. Balakrishnan, "Stochastic System Identification Techniques", In Stochastic Optimization and Control, (M.F. Karreman, Ed), Wiley, New York, 1968.

[18] D.S. Spain, Identification and Modelling of Discrete, Stohastic Linear Systems, Tech. Report No 6302-10, 1971, Stanford University.

[19] P.E. Caines and J. Rissanen, "Maximum Likelihood Estimation in Multivariable Gaussian Stochastic Processes", IEEE Trans. Info. Theory, Vol IT-20, No 1 (1974), pp. 102-104.

[20] J. Rissanen and P.E. Caines, "Consistency of Maximum Likelihood Estimators for ARMA Processes", Control Systems Report No. 7424, Department of Electrical Engineering, University of Toronto, Toronto, Canada, December 1974.

[21] T. Söderström, I. Gustavsson and L. Ljung, "Identifiability Conditions for Linear Systems Operating in Closed Loop", Int. J. Control, Vol 21, No 2 (1975), pp. 243-255.

[22] K.J. Åström and T. Söderström, "Uniqueness of the Maximum Likelihood Estimates of the Parameters of An ARMA Model", IEEE Trans. Autom. Control, Vol AC-19, No 6 (1974), pp.769-774.

[23] T. Söderström, "On the Uniqueness of Maximum Likelihood Identification for Different Structures", To appear in Automatica, Vol 11, No 2,(March 1975).

[24] K.J. Åström, U. Borisson, L. Ljung and B. Wittenmark, "Theory and Applications of Adaptive Regulators Based on Recursive Parameter Estimation", To appear in Proc. of the 5th IFAC World Congress, Boston, Mass., August 1975.

[25] L. Ljung and B. Wittenmark, "Analysis of a Class of Adaptive Regulators", Proc. of the IFAC Symposium on Stochastic Control Theory, Budapest, Hungary, September 1974.

[26] L. Ljung, "Convergence of Recursive Stochastic Algorithms", Proc. of the IFAC Symposium on Stochastic Control Theory, Budapest, Hungary, September 1974.

[27] H. Cramer and M.R. Leadbetter, Stationary and Related Stochastic Processes , Wiley, New York 1967.

[28] T. Kailath, "The Innovations Approach to Detection and Estimation Theory", Proc. IEEE, Vol 58, No 5 (1970), pp.680-695.

[29] K. Glover and J.C. Willems, "Parameterizations of Linear Dynamical Systems: Canonical Forms and Identifiability", IEEE Trans. Autom. Control, Vol AC-19, No 6 (1974), pp.640-646.

[30] H.E. Berntsen and J.G. Balchen, "Identifiability of Linear Dynamic Systems", Proc. of the 3rd IFAC Symposium on Identification and System Parameter Estimation, (P. Eykhoff, Ed.), The Hague, Holland, June 1973.

[31] E. Tse, "Information Matrix and Local Identifiability of Parameters", 1973 Joint Automatic Control Conf., Preprints.