An On-line Algorithm for Approximate Maximum Likelihood Identificatioin of Linear Dynamic Systems

Söderström, Torsten

1973

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

AN ON-LINE ALGORITHM FOR APPROXIMATE MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR DYNAMIC SYSTEMS

T. SÖDERSTRÖM

Report 7308, March 1973
Lund Institute of Technology
Division of Automatic Control

AN ON-LINE ALGORITHM FOR APPROXIMATE MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR DYNAMIC SYSTEMS

T. Söderström

ABSTRACT

A recursive algorithm for maximum likelihood estimation of parameters in a linear dynamic system is presented. The basic idea in the algorithm is a recursive optimization of the likelihood function. Different approximations are used. With special simplifications the algorithm becomes identical to methods earlier proposed. The properties of the algorithm are illustrated by application to data from simulated systems as well as plant measurements.

TABLE OF CONTENTS

I. INTRODUCTION

In the field of the identification of dynamic systems special interest has been given to on-line methods. It may be desirable to proceed the identification until a specified accuracy is achieved. An on-line identification method also is necessary for adaptive control.

Several on-line identification methods have been proposed. In Åström-Eykhoff (1971) a short description of different methods is given. The algorithms described in Young (1970), Young-Shellswell-Nethling (1971) seem to work quite satisfactorily.

When off-line methods are considered it is known that the maximum likelihood method is a powerful one and in most cases gives the "best" estimates, Åström-Bohlin (1966), Gustavsson (1969b). The purpose of this report is to describe an approximative recursive version of this method using ideas due to Åström, who has made an outline of the algorithm.

It is well-known, Åström (1968), Åström-Eykhoff (1971) that the least squares (LS) method easily can be computed recursivly. The recursive version can be interpreted as a Kalman filter. The ML method can be considered as an extension of the LS method. One way to construct a recursive ML algorithm is to generalize the Kalman filter of the LS case. This approach has been taken by Young for the estimation of parameters of time series.

Panuška (1968) gives a similar algorithm based on stochastic approximation. A comparison of Panuška's algorithm and the off-line ML method is given in Valis-Gustavsson (1969).

In this report an estimation algorithm will be derived via a recursive minimization of a time varying loss function. When different approximations and simplifications are made

the algorithm is the same as the one used by Young or the one used by Panuška.

The approach of minimizing a loss function can be applied to different models. Several well-known methods as least squares, generalized least squares, Clarke (1967), Söderström (1972), the "ordinary" ML, Åström-Bohlin (1966) and the method used by Bohlin (1970) can be interpreted as maximum likelihood models when appropriate assumptions of the structure of the systems are made. All the methods can be expressed as a minimization of a loss function of the form

$$V_N(\hat{\theta}) = \frac{1}{2} \sum_{t=1}^{N} \varepsilon^2(t;\hat{\theta}) \tag{1.1}$$

N is the number of samples and $\varepsilon(t;\hat{\theta})$ the residual at time t. The vector $\hat{\theta}$ is an estimate of θ, a vector containing parameters which describe the system. The elements of $\hat{\theta}$ will be called the model parameters. The explicit expression of $\varepsilon(t;\hat{\theta})$ as a function of $\hat{\theta}$ differs between the different methods. The variances of the residuals can be estimated by

$$\hat{\lambda}^2 = \frac{2}{N} \min_{\hat{\theta}} V_N(\hat{\theta}) \tag{1.2}$$

Let $\hat{\theta}_N$ minimize $V_N(\hat{\theta})$. A recursive algorithm must give $\hat{\theta}_{N+1}$ from $\hat{\theta}_N$, the measurements at time N+1 and a reasonably small amount of collected information of the system. In the recursive LS method this is done exactly but for the other methods approximations have to be used.

Another way of discussing the properties of a reasonable algorithm it to use the concept of sufficient statistics. When the disturbances are gaussian it is well-known that there is a sufficient statistic in the LS case, namely

$V_N''(\hat{\theta}_{N-1})$, $\hat{\theta}_{N-1}$ and a few of the latest measurements. In the general case a sufficient statistic must include all old measurements explicitly. Thus it is suitable to base an algorithm on **an approximate sufficient statistic.**

In the next chapter an algorithm for the recursive minimization of V_N is developed. Different approximations are discussed. In chapter III the Kalman filter approach is taken into consideration and some comparisons are made. Possible limits to which the estimates may converge are analysed in chapter IV. The fifth chapter contains some examples and discussions about how to implement the algorithm. Finally examples using plant measurements are presented in chapter VI.

II. A RECURSIVE MAXIMUM LIKELIHOOD ESTIMATOR

In this chapter the recursive algorithm is developed. The first part deals with the recursive minimization of the loss function

$$V_N(\hat{\theta}) = \frac{1}{2} \sum_{t=1}^{N} \varepsilon^2(t;\hat{\theta}) \tag{1.1}$$

in general. In the second part the algorithm will be applied to the specific model, Åström-Bohlin (1966)

$$\hat{A}(q^{-1})y(t) = \hat{B}(q^{-1})u(t) + \hat{C}(q^{-1})\varepsilon(t;\hat{\theta}) \tag{2.1}$$

where **y(t)** is the output and u(t) the input at time t. The polynomial operators are

$$\hat{A}(q^{-1}) = 1 + \hat{a}_1 q^{-1} + \ldots + \hat{a}_n q^{-n}$$

$$\hat{B}(q^{-1}) = \hat{b}_1 q^{-1} + \ldots + \hat{b}_n q^{-n}$$

$$\hat{C}(q^{-1}) = 1 + \hat{c}_1 q^{-1} + \ldots + \hat{c}_n q^{-n}$$

q^{-1} is the backward shift operator and

$$\hat{\theta} = [\hat{a}_1 \ldots \hat{a}_n \hat{b}_1 \ldots \hat{b}_n \hat{c}_1 \ldots \hat{c}_n]^T$$

Let $\hat{\theta}_N$ be the minimum point of $V_N(\hat{\theta})$. The estimate $\hat{\theta}_{N+1}$ will be computed from a Taylor expansion of $V_{N+1}(\hat{\theta})$ around $\hat{\theta}_N$. Suppose that an expansion including second order terms is accurate enough.

$$V_{N+1}(\hat{\theta}) \approx V_{N+1}(\hat{\theta}_N) + V'_{N+1}(\hat{\theta}_N)(\hat{\theta}-\hat{\theta}_N) +$$

$$+ \frac{1}{2}(\hat{\theta}-\hat{\theta}_N)^T V''_{N+1}(\hat{\theta}_N)(\hat{\theta}-\hat{\theta}_N)$$

Minimization gives

$$\hat{\theta}_{N+1} = \hat{\theta}_N - V''_{N+1}(\hat{\theta}_N)^{-1} V'_{N+1}(\hat{\theta}_N)^T \tag{2.2}$$

which is the first iteration of a Newton Raphson algorithm applied to the equation $V'_{N+1}(\hat{\theta}) = 0$.

The estimated minimum value of $V_{N+1}(\hat{\theta})$ is

$$V_{N+1}(\hat{\theta}_{N+1}) = V_{N+1}(\hat{\theta}_N) - \frac{1}{2} V'_{N+1}(\hat{\theta}_N) V''_{N+1}(\hat{\theta}_N)^{-1} V'_{N+1}(\hat{\theta}_N)^T \tag{2.3}$$

To form a recursive estimator the relation between $V_{N+1}(\hat{\theta})$ and $V_N(\hat{\theta})$ must be utilized. By definition

$$V_{N+1}(\hat{\theta}) = V_N(\hat{\theta}) + \frac{1}{2} \varepsilon^2(N+1;\hat{\theta}) \tag{2.4}$$

$$V'_{N+1}(\hat{\theta}) = V'_N(\hat{\theta}) + \varepsilon(N+1;\hat{\theta})\varepsilon'(N+1;\hat{\theta}) \tag{2.5}$$

$$V''_{N+1}(\hat{\theta}) = V''_N(\hat{\theta}) + \varepsilon'(N+1;\hat{\theta})^T \varepsilon'(N+1;\hat{\theta}) + \varepsilon(N+1;\hat{\theta})\varepsilon''(N+1;\hat{\theta}) \tag{2.6}$$

The following approximations are made now

$$V'_N(\hat{\theta}_N) = 0 \tag{2.7}$$

$$\varepsilon(N+1;\hat{\theta}_N)\varepsilon''(N+1;\hat{\theta}_N) = 0 \tag{2.8}$$

$$V''_N(\hat{\theta}_N) = V''_N(\hat{\theta}_{N-1}) \tag{2.9}$$

The assumption (2.7) can be assumed to hold since $\hat{\theta}_N$ is assumed to minimize $V_N(\hat{\theta})$. For off-line ML the term $\sum_{t=1}^{N} \varepsilon(t;\hat{\theta})\varepsilon''(t;\hat{\theta})$ has little influence on the minimization, Gustavsson (1969b). The equation (2.9) is motivated if $\hat{\theta}_N$ is close to $\hat{\theta}_{N-1}$. Also notice that (2.7) - (2.9) as well as the Taylor expansion hold exactly in the LS case.

With the use of the approximations

$$V_{N+1}(\hat{\theta}_{N+1}) = V_N(\hat{\theta}_N) + \frac{1}{2}\varepsilon^2(N+1;\hat{\theta}_N) - \frac{1}{2}V'_{N+1}(\hat{\theta}_N)V''_{N+1}(\hat{\theta}_N)^{-1}V'_{N+1}(\hat{\theta}_N)^T \tag{2.10}$$

$$V'_{N+1}(\hat{\theta}_N) = \varepsilon(N+1;\hat{\theta}_N)\,\varepsilon'(N+1;\hat{\theta}_N) \tag{2.11}$$

$$V''_{N+1}(\hat{\theta}_N) = V''_N(\hat{\theta}_{N-1}) + \varepsilon'(N+1;\hat{\theta}_N)^T\varepsilon'(N+1;\hat{\theta}_N) \tag{2.12}$$

Introduce the notations

$$P_N = V''_N(\hat{\theta}_{N-1})^{-1} \tag{2.13}$$

$$\varphi_N = \varepsilon'(N;\hat{\theta}_{N-1})^T \tag{2.14}$$

$$\varepsilon_N = \varepsilon(N;\hat{\theta}_{N-1}) \tag{2.15}$$

$$\gamma_{N+1} = 1 + \varphi_{N+1}^T P_N \varphi_{N+1} \tag{2.16}$$

Then (2.2) can be written as

$$\hat{\theta}_{N+1} = \hat{\theta}_N - P_{N+1}\varphi_{N+1}\varepsilon_{N+1} \tag{2.17}$$

The well-known matrix lemma

$$[M+bb^T]^{-1} = M^{-1} - M^{-1}b[1 + b^TM^{-1}b]^{-1}b^TM^{-1}$$

applied to (2.12) gives

$$P_{N+1} = P_N - \frac{1}{\gamma_{N+1}} P_N\varphi_{N+1}\varphi_{N+1}^T P_N \tag{2.18}$$

Finally (2.10) can be rewritten after some trivial calculations as

$$V_{N+1}(\hat{\theta}_{N+1}) = V_N(\hat{\theta}_N) + \frac{1}{2}\frac{1}{\gamma_{N+1}}\varepsilon_{N+1}^2 \tag{2.19}$$

In the general case it now remains to develop recursive equations for ε_N and φ_N. For the LS case this is very simple since $\varepsilon(t;\hat{\theta})$ is linear in $\hat{\theta}$. The derived algorithm coincides with the well-known one in the LS case. The expression (2.19) can be found in Wieslander (1971) where it is derived using a Kalman filter representation.

In the derivation of recursive equations for ε_N and φ_N specialization will be made to the model

$$\hat{A}(q^{-1})y(t) = \hat{B}(q^{-1})u(t) + \hat{C}(q^{-1})\varepsilon(t;\hat{\theta}) \tag{2.1}$$

which can be written in state space form as

$$x(t+1) = \begin{bmatrix} -\hat{c}_1 & 1 & & 0 \\ \vdots & & \ddots & \\ & & & 1 \\ -\hat{c}_n & 0 & & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 & \hat{a}_1 & -\hat{b}_1 \\ 0 & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & \hat{a}_n & -\hat{b}_n \end{bmatrix} \begin{bmatrix} y(t+1) \\ y(t) \\ u(t) \end{bmatrix} \tag{2.20}$$

$$\varepsilon(t;\hat{\theta}) = x_1(t)$$

The derivatives $\varepsilon'(t;\hat{\theta})$ are given by

$$\begin{aligned} \hat{C}(q^{-1})\,\frac{\partial \varepsilon}{\partial \hat{a}_i}(t;\hat{\theta}) &= y(t-i) \\ \hat{C}(q^{-1})\,\frac{\partial \varepsilon}{\partial \hat{b}_i}(t;\hat{\theta}) &= -u(t-i) \\ \hat{C}(q^{-1})\,\frac{\partial \varepsilon}{\partial \hat{c}_i}(t;\hat{\theta}) &= -\varepsilon(t-i;\hat{\theta}) \end{aligned} \tag{2.21}$$

A state space form of (2.21) is

$$\varepsilon'(t+1;\hat{\theta}) = \begin{bmatrix} -\hat{c}_1 & \cdots & -\hat{c}_n & & & \\ 1 & & & & 0 & \\ & \ddots & & & & \\ & 1 & 0 & & & \\ & & -\hat{c}_1 & \cdots & -\hat{c}_n & \\ & & 1 & & & \\ & & & \ddots & & \\ & 0 & & 1 & 0 & \\ & & & & -\hat{c}_1 & \cdots & -\hat{c}_n \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & 1 & 0 \end{bmatrix} \varepsilon'(t;\hat{\theta}) + \begin{bmatrix} y(t) \\ 0 \\ \vdots \\ 0 \\ -u(t) \\ 0 \\ \vdots \\ 0 \\ -\varepsilon(t;\hat{\theta}) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (2.22)$$

The initial values are $x(0) = 0$, $\varepsilon'(0;\hat{\theta}) = 0$.

In order to compute $\varepsilon(t;\hat{\theta})$ and $\varepsilon'(t;\hat{\theta})$ (2.20) and (2.22) have to be solved from t=0. Since ε and ε' must be computed for new arguments at every time step this means an unreasonable lot of calculations. Moreover, all old measurements must be saved. Note that no matrix multiplications have to be done explicitly. E.g. all but three components of $\varepsilon(t+1;\hat{\theta})$ can be computed by shift.

One way of reducing the computational work is the following. (2.20) and (2.22) are solved only once and with time variable matrices. When $x(t)$ and $\varepsilon'(t;\hat{\theta}_{t-1})$ are computed from $x(t-1)$ respectively $\varepsilon'(t-1;\hat{\theta}_{t-2})$ the components of $\hat{\theta}_{t-1}$ are used in the matrices. If $\hat{\theta}_t$ does not change very much with t this approximation can be assumed to be good. The resulting values of the residual will be denoted $\hat{\varepsilon}_t$.

A further simplification would be to substitute $\hat{C}(q^{-1})$ in (2.21) with 1. This does not reduce the computations very much but it has a nice interpretation which will be shown in the next chapter.

There are other possibilities to compute approximative values of the residuals. One is the following which is used by Young (1970) and Panuška (1968). The equation (2.1) can be written as

$$\varepsilon(t) = y(t) - [-y(t-1)..-y(t-n)\ u(t-1)..u(t-n)\ \varepsilon(t-1)...\varepsilon(t-n)]\hat{\theta} \quad (2.23)$$

An exact computation of $\varepsilon(t;\hat{\theta})$ requires the solution of (2.23) from t=0 with constant $\hat{\theta}$. Similarly to the method previously described $\varepsilon(t;\hat{\theta}_{t-1})$ can be approximated by

$$\varepsilon(t;\hat{\theta}_{t-1}) = y(t) -$$

$$-[-y(t-1)..-y(t-n)\ \ u(t-1)..u(t-n)\ \varepsilon(t-1;\hat{\theta}_{t-2})..\varepsilon(t-n;\hat{\theta}_{t-n-1}]\hat{\theta}_{t-1} \quad (2.24)$$

The algorithm used by Young is obtained if (2.24) is used for computations of ε_N and (2.21) with $\hat{C}(q^{-1})$ substituted by 1 for computations of φ_N.

Panuška's algorithm uses a gradient method for the minimization. In (2.17) P_N is substituted by $\frac{K}{N}$ I where K is a suitable constant. ε_N and φ_N are computed as in Young's algorithm.

The general algorithm and Young's version are compared using simulated data in chapter V. For these examples both the methods may give bad estimates if they are applied straight-forward. Suitable modifications are discussed in chapter V. Further it turns out that after these modifications both the methods seem to work well in the present simulated systems but Young's algorithm gives larger variances of the parameter estimates. For both the methods the convergence of the $\hat{A}$ and the $\hat{B}$-parameters are considerably faster than the convergence of the $\hat{C}$-parameters.

In Valis-Gustavsson (1969) a comparison is made between Panuška's method and the off-line ML method. The comparison shows not

unexpectedly that the off-line ML method is superior. Especially the C-parameters seem to be difficult to estimate accurately with Panuška's method.

III. COMPARISON WITH KALMAN FILTERING

It is well-known that the recursive least squares method can be interpreted as a Kalman filter, Åström (1968), Åström-Eykhoff (1971). Using some approximations this idea can be used for the model (2.1) as well. It turns out that Young's algorithm is very "natural" from this point of view.

The system corresponding to the model (2.1) can be written as

$$\begin{aligned} \theta(t+1) &= \theta(t) \\ y(t) &= C(t)\theta(t) + e(t) \end{aligned} \tag{3.1}$$

where $e(t)$ is white noise with variance λ^2 and

$$C(t) = [-y(t-1)...-y(t-n) \quad u(t-1)..u(t-n) \quad e(t-1)...e(t-n)]$$

$$\theta(t) = [a_1,\ldots, a_n, b_1,\ldots,b_n, c_1,\ldots, c_n]^T$$

If $C(t)$ were known a Kalman filter for estimating the state $\theta(t)$ would be

$$\hat{\theta}(t+1) = \hat{\theta}(t) + K(t+1)[y(t+1) - C(t+1)\hat{\theta}(t)]$$

$$K(t) = \frac{1}{\lambda^2} P(t)C(t)^T \tag{3.2}$$

$$P(t) = P(t-1)-P(t-1)C(t)^T[\lambda^2+C(t)P(t-1)C(t)^T]^{-1}C(t)P(t-1)$$

A way to overcome the difficulty of $C(t)$ being partly unknown is to replace $e(t-1)...e(t-n)$ in $C(t)$ by $\varepsilon(t-1)...\varepsilon(t-n)$. The residuals $\{\varepsilon(t)\}$ are defined recursively through

$$\varepsilon(t+1) = y(t+1) - C(t+1)\hat{\theta}(t)$$

The algorithm obtained is exactly Young's method.

IV. ANALYSIS

To establish convergence of the algorithm, i.e. to prove that $\hat{\theta}_k \to \theta$, $k \to \infty$ is a very hard task. The purpose of the following analysis only is to determine possible limits of $\{\hat{\theta}_k\}$.

First it is observed that the recursive algorithm given by (2.17) - (2.19) formally can be interpreted as a recursive least squares solution of the system of equations

$$\begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \end{bmatrix} \hat{\theta} = \begin{bmatrix} -\varepsilon_1 + \varphi_1^T \hat{\theta}_o \\ -\varepsilon_2 + \varphi_2^T \hat{\theta}_1 \\ \vdots \end{bmatrix} \qquad (4.1)$$

This is true only formally since the right-hand side involves $\hat{\theta}_o$, $\hat{\theta}_1$,...

Assume that $\hat{\theta}_t$ tends to θ^* with probability one when the number of samples tends to infinity. Assume that θ^* corresponds to a model for which $A^*(z)$ and $C^*(z)$ have all zeros outside the unit circle.

If the initial values of the recursive least squares solution are neglected then $\hat{\theta}_N$ must fulfil the normal equations

$$\frac{1}{N}\sum_{t=1}^{N}(\varphi_t\varphi_t^T)\hat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N}\varphi_t(-\varepsilon_t + \varphi_t^T\hat{\theta}_{t-1}) \qquad (4.2)$$

It is shown in the appendix that $\hat{\theta}_N$ and $\hat{\theta}_{t-1}$ asymptotically can be replaced by θ^*. Further ε_t and φ_t (asymptotically) can be replaced by $\varepsilon(t;\theta^*)$ and $\varepsilon'(t;\theta^*)$ respectively. Thus (4.2) implies

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t;\theta^*)\varepsilon'(t;\theta^*) = 0 \qquad (4.3)$$

Using standard ergodic theory it is possible to show, see Söderström (1972), that (4.3) can be substituted by

$$E\, \varepsilon(t;\theta^*)\varepsilon'(t;\theta^*) = 0 \tag{4.4}$$

However, (4.4) is exactly the equation for the stationary points of

$$W(\hat{\theta}) = E[\varepsilon(t;\hat{\theta})]^2 \tag{4.5}$$

In Söderström (1973) an analysis of the number of local minimum points of $W(\hat{\theta})$ is given. It is shown that θ is always a minimum point and conditions are given which guarantee that θ is a unique local minimum point of $W(\hat{\theta})$.

Thus if these condition are fulfilled and $\hat{\theta}_t$ converges a.s. it must converge to the correct values.

V. NUMERICAL EXAMPLES

In this chapter some numerical examples will be given. It has appeared to the author by practical experience that the algorithm cannot be successfully applied in a straightforward way, but suitable tricks make it work rather well. Several tricks and modifications have been tried by the author but only the best one is used in the examples presented. At the end of the chapter a brief discussion of other tricks is given.

Two different versions of the algorithm are used in the examples. One is called RMLE1 (Recursive Maximum Likelihood Estimation, version 1) and the other RMLE2. Both the versions include the basic algorithm given by (2.17) - - (2.19). The estimate of λ is taken as

$$\hat{\lambda} = \sqrt{\frac{2}{N} V_N(\hat{\theta}_N)}$$

In RMLE1 the residuals are computed from (2.22). RMLE2 is the version used by Young (who calls it AML, Approximate Maximum Likelihood).

The initial values of all variables involved were all chosen as 0 with the exception of P_0 which was chosen $100 \cdot I$. In the off-line version of the ML algorithm a test of stability of $\hat{C}(z)$ is made at every iteration and the estimates are modified to give stability, Gustavsson (1969b) This trick was tried in the recursive algorithm as well and it improved the result.

It would be valuable to have one number giving the accuracy of the result. For instance, one can use $||\hat{\theta}-\theta||^2$ or more generally $(\hat{\theta}-\theta)^T Q(\hat{\theta}-\theta)$ where Q is some symmetric positive definite matrix.
In the following examples an asymptotic loss function was used, namely,

$$W(\hat{\theta};\theta) = \frac{1}{\lambda^2} E\, \varepsilon^2(t)$$

where

$$\hat{C}(q^{-1})\varepsilon(t) = \hat{A}(q^{-1})y(t) - \hat{B}(q^{-1})u(t)$$

and the process is described by

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t), \quad E\, e^2(t) = \lambda^2$$

with $\{e(t)\}$ white noise. Thus

$$W(\hat{\theta};\theta) = \frac{1}{\lambda^2} E\left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})} u(t) + \right.$$

$$\left. + \frac{\hat{A}(q^{-1})C(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})} e(t)\right]^2$$

Assume that the input is independent of the noise. If the spectral density of the input is known (in the examples the input is treated as white noise) $W(\hat{\theta};\theta)$ can easily be computed from integrals.

Clearly, Åström-Söderström (1973), $W(\hat{\theta};\theta) \geq 1$ for all $\hat{\theta}$ where equality implies $\hat{\theta} = \theta$. Further $W'_{\hat{\theta}}(\theta;\theta) = 0$.

An expected asymptotic value of $W(\hat{\theta};\theta)$ can be calculated. Assume that $\hat{\theta}$ is asymptotically gaussian distributed with mean θ and variance equal to the Cramér-Rao lower bound, i.e.

$$P_{\hat{\theta}} = \frac{2}{N} W(\theta;\theta)\, W''_{\hat{\theta}\hat{\theta}}(\theta;\theta)^{-1}$$

This assumption is valid for the off-line ML estimates, Åström-Bohlin (1966). For large values of N, $W(\hat{\theta};\theta)$ then can be approximated by $1 + \frac{1}{N} x$ where

$$x = (\hat{\theta}-\theta)^T P_{\hat{\theta}}^{-1}(\hat{\theta}-\theta)$$

is asymptotically $\chi^2(3n)$ distributed. Especially, under these assumptions $E\, W(\hat{\theta};\theta) = 1 + \frac{3n}{N}$.

In order to analyse the properties of the methods, the algorithms were applied to data from simulated systems. A number of realizations was used. The average values and the RMS errors of $\hat{\theta}$ were computed. The RMS errors are

$$\left(\frac{1}{k}\sum_{j=1}^{k}(\hat{\theta}_i(j) - \theta_i)^2\right)^{1/2}$$

where $\hat{\theta}_i(j)$ denotes the i:th component of $\hat{\theta}$ obtained at the identification of the j:th realization. The average values and the RMS errors are compared with their theoretically expected values based on the Cramér-Rao lower bound.

In all examples the number of samples was 2 000. The input signal was a PRBS with amplitude 1.0. 11 different realizations were used.

For the first order system the algorithms applied straightforward work rather satisfactorily. RMLE1 produces a considerably lower variance of $\hat{c}_1$ than RMLE2. The results are given in table 5.1.

		a_1	b_1	c_1	λ	W
Expected	mean	-0.8	1.0	0.7	1.0	1.0015
	RMS error	0.012	0.017	0.017	0.032	0.0019
RMLE1	mean	-0.796	1.005	0.695	1.009	1.0023
	RMS error	0.019	0.020	0.014	0.037	0.0028
RMLE2	mean	-0.796	1.005	0.675	1.019	1.0056
	RMS error	0.023	0.027	0.038	0.042	0.0067

Table 5.1. Results for a first order system. RMLE1 is the general algorithm given in chapter II. RMLE2 is Young's algorithm.

For a second order system, however, the results are considerably inferior than in the first order case. The results of a straight-forward application of the algorithms are given in table 5.2.

	a_1	a_2	b_1	b_2	c_1	c_2	λ	W
Expected mean	-1.5	0.7	1.0	0.5	-1.0	0.2	1.0	1.0030
RMS error	0.007	0.006	0.022	0.029	0.023	0.022	0.032	0.0032
RMLE1 mean	-1.418	0.624	0.990	0.513	-0.747	0.103	1.267	1.164
RMS error	0.259	0.242	0.073	0.072	0.517	0.116	0.591	0.185
RMLE2 mean	-1.490	0.688	1.009	0.487	-0.867	0.044	1.112	1.054
RMS error	0.029	0.027	0.028	0.072	0.180	0.161	0.140	0.076

Table 5.2. Results for a second order system. Straight-forward application of the algorithms. RMLE1 is the general algorithm given in chapter II. RMLE2 is Young's algorithm.

In Figures 5.1 and 5.2 the estimates of one of the realizations (RMLE1 is used) are plotted versus time. A comparison with table 2 shows that the result of the identification of this realization is among the best ones. From Figure 5.1 a general tendency of the algorithm can be seen. It loses its "gain" after some hundred samples and most often the estimates of the C-parameters then are not close to the correct values. This fact indicates that some kind of restarts would be valuable.

This idea will be combined with another. In the computations of the derivatives of the loss function $\hat{\varepsilon}_t$ is used as an approximation of $\varepsilon(t;\hat{\theta})$ for various values of $\hat{\theta}$. If $\hat{\theta}$ is fixe for a number of samples, the approximation $\hat{\varepsilon}_t \approx \varepsilon(t;\hat{\theta})$ probably

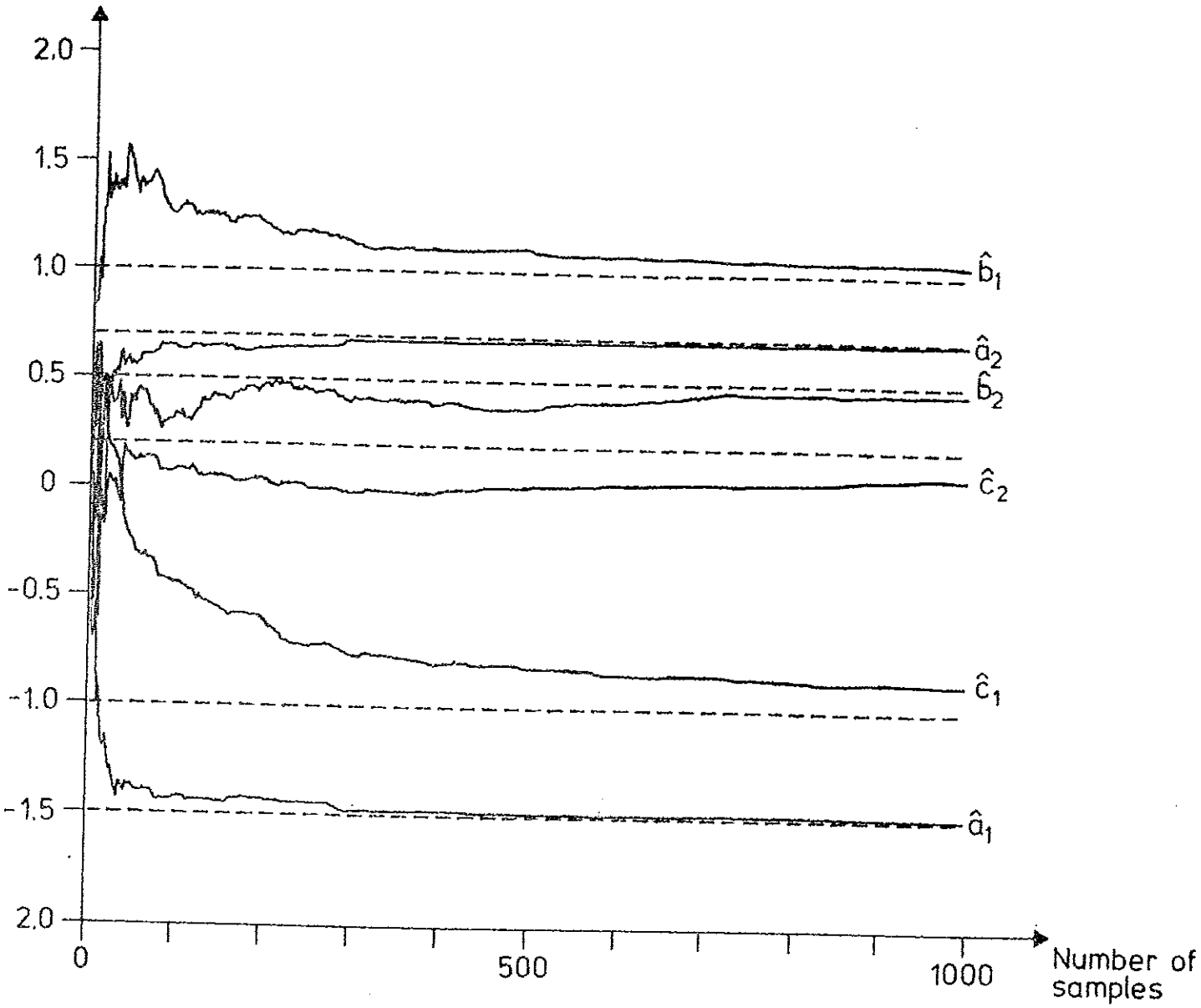

Figure 5.1 Parameter estimates of a second order system. Straightforward application of the algorithm is done. The dashed lines give the true values of the parameters.

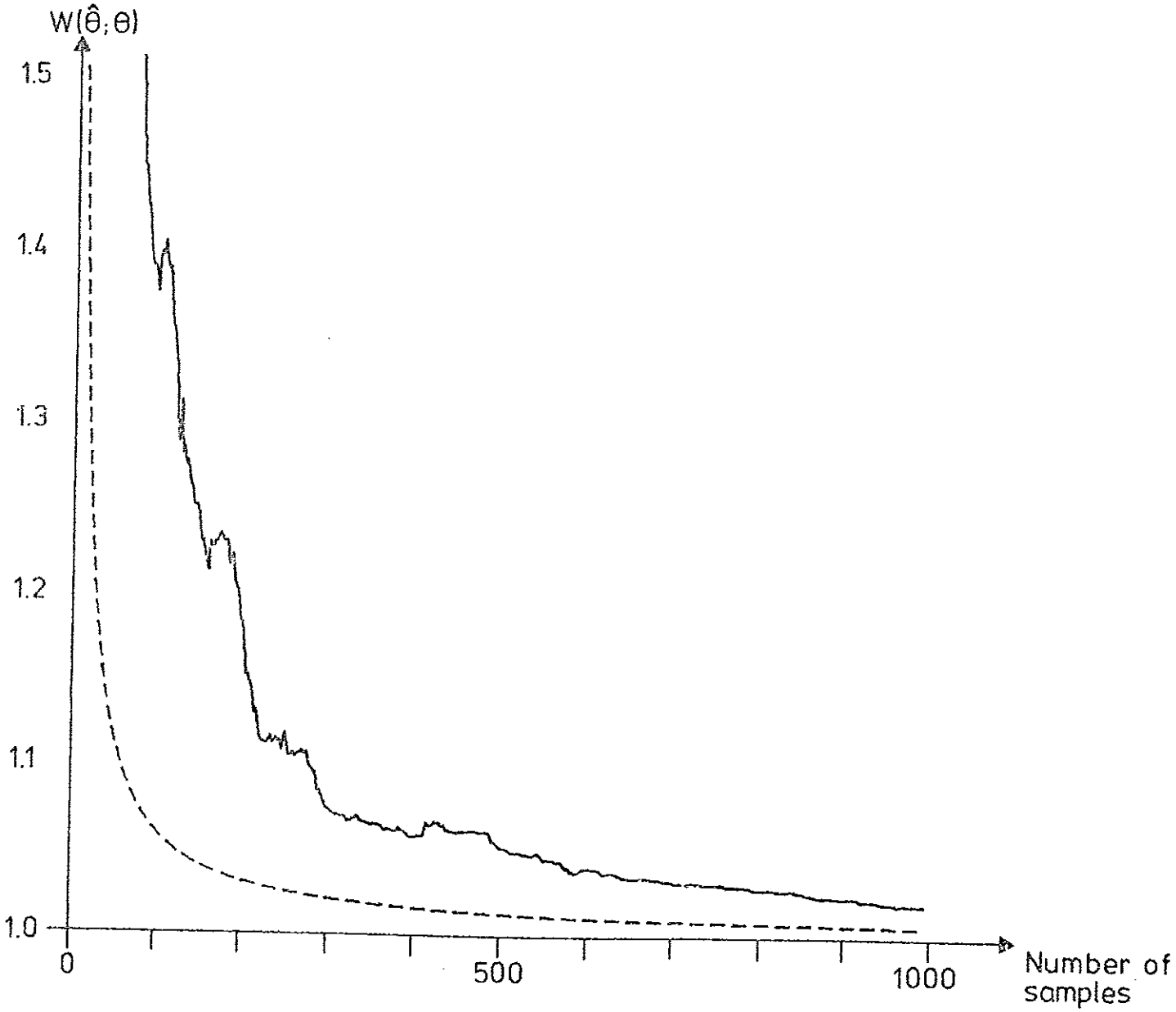

Figure 5.2. Loss function of a second order system. Straight-forward application of the algorithm is done. The dashed line gives the asymptotically expected loss.

will be considerably better. If this idea would have any practical value $\hat{\theta}$ must not change too much in the rest of the identification.

The second idea has been examined to some extent by simulations. The particular system given in table 2 was used. In some simulations $\hat{\theta}_t$ was fixed to the correct values for the first 100 samples and after that $\hat{\theta}$ was estimated according to the algorithm. Good results were obtained. In other simulations the first 100 samples were used at an identification with the off-line ML algorithm. These off-line identifications produced good initial values of the recursive algorithm, which produced satisfactory results in this case.

The algorithm has been modified in the following way. It is applied straight-forward in N_1 steps. Then a test of "convergence" is performed. If "convergence" has occurred, the algorithm is continued straight-forward. If no convergence has occurred a restart is made with $\hat{\theta}_t$ keeping its value and the other variables as their ordinary start values. $\hat{\theta}_t$ is constrained to be constant for the next N_2 steps. After another N_1 steps a new test of "convergence" is made. The estimate $\hat{\lambda}$ is modified in an obvious way with regard to the latest restart. This prodedure of successive restarts is continued until "convergence" has occurred.

A suitable test of "convergence" would be to use $W(\hat{\theta};\theta)$. If this quantity is small (close to 1) "convergence" may be considered to have occurred. However, this test quantity cannot be used in practice, since it requires knowledge of θ and λ^2. Instead $W(\hat{\theta}_t;\hat{\theta}_{t-N_1-N_2})$ is used with λ^2 substituted by $\hat{\lambda}_t^2$. This means that θ is substituted by the estimate $\hat{\theta}$ which was present when the latest test of "convergence" was made. If the test quantity is smaller than VTEST no more restarts are made.

Simulations were made using the same realizations as before. The values of the variables used were VTEST 1.05, N_1 = 300, and N_2 = 50. It is the author's experience that the method is not very sensitive to the values of the parameters VTEST, N_1 and N_2. The results are good for RMLE1 and a bit inferior, but yet satisfactory for RMLE2, see table 5.3.

	a_1	a_2	b_1	b_2	c_1	c_2	λ	W
Expected mean	-1.5	0.7	1.0	0.5	-1.0	0.2	1.0	1.0030
RMS error	0.007	0.006	0.022	0.029	0.023	0.022	0.032	0.0032
RMLE1 mean	-1.498	0.699	0.998	0.500	-0.987	0.180	0.994	1.0057
RMS error	0.008	0.008	0.022	0.025	0.034	0.042	0.048	0.0065
RMLE2 mean	-1.505	0.702	1.002	0.473	-0.966	0.160	1.009	1.017
RMS error	0.014	0.014	0.032	0.062	0.062	0.080	0.060	0.020

Table 5.3. Results for a second order system. The trick with restarts is used. RMLE1 is the general algorithm given in chapter II. RMLE2 is Young's algorithm.

In Figures 5.3 and 5.4 it is shown how the modified algorithm RMLE1 works on the same data as were used in Figures 5.1 and 5.2.

It can be seen that the restarts give the algorithm larger "gain" than before which causes a jerkiness of the estimates. The long range effect, however, is that the estimates are considerably closer to the correct values than without restarts.

Now a brief discussion of other tricks and approaches tried by the author is given. His experience is that these tricks do not give a satisfactory improvement of the algorithm.

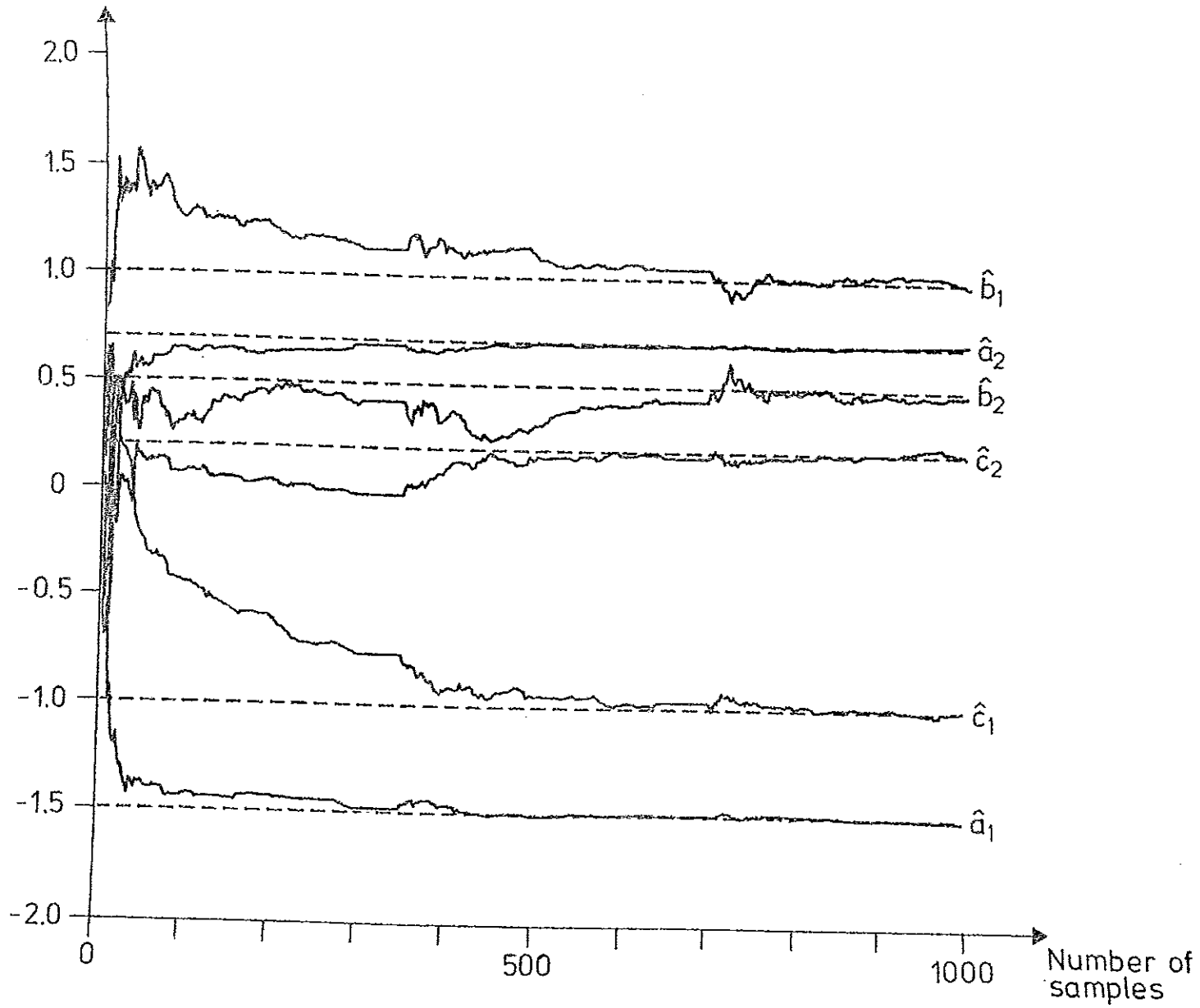

Figure 5.3 Parameter estimates of a second order system. The modified algorithm is used. The dashed lines give the true values of the parameters.

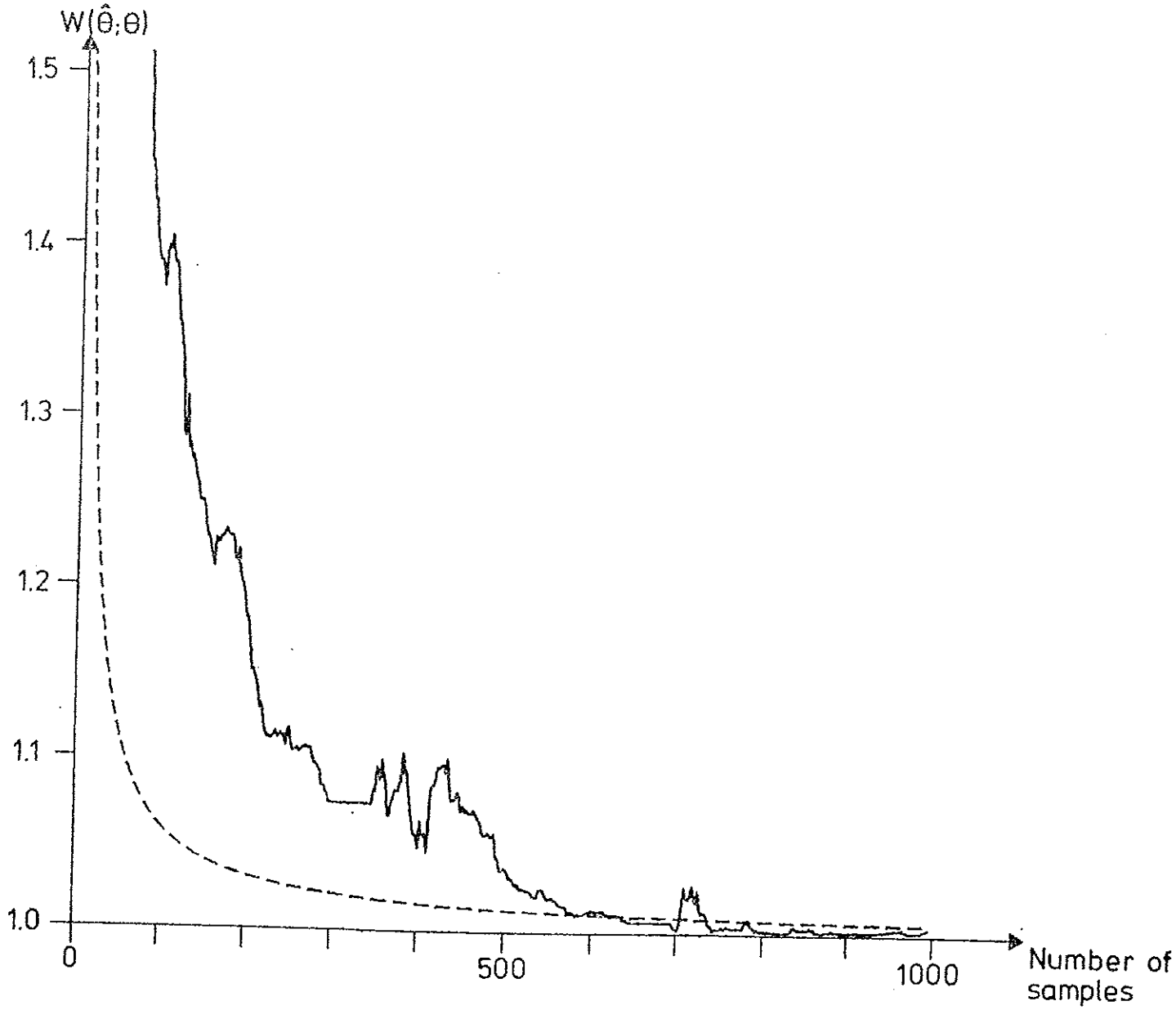

Figure 5.4 Loss function of a second order system. The modified algorithm is applied. The dashed line gives the asymptotically expected loss.

- The inverse $V_N''(\hat{\theta}_{N-1})^{-1}$ was computed by inversion of $V_N''(\hat{\theta}_{N-1})$, i.e. (2.12) was used together with inversion instead of (2.18). Since $V_N''(\hat{\theta})$ is <u>not</u> independent of $\hat{\theta}$ this may change the result of the algorithm.

- The term $\varepsilon(N+1;\hat{\theta}_N)\varepsilon''(N+1;\hat{\theta}_N)$ was not dropped in the computation of $V''_{N+1}(\hat{\theta}_N)$.

- If the algorithm does not really minimize V_N the approximation $V_N'(\hat{\theta}_N) = 0$ may not be accurate. The equation (2.11) of the gradient was changed to

$$V'_{N+1}(\hat{\theta}_N) = \alpha V_N'(\hat{\theta}_N) + \varepsilon(N+1;\hat{\theta}_N)\varepsilon'(N+1;\hat{\theta}_N)$$

  The parameter α was chosen in the interval [0, 1]. When $\alpha = 0$ the previous algorithm is obtained. The choice of $\alpha = 1$ caused very large changes in the parameter estimates and was very unsatisfactory. The choice $\alpha = 0.6$ gave some improvements of the convergence but it was not satisfactory enough.

- In order to speed up the convergence it may be appropriate to change (2.17) to

$$\hat{\theta}_{N+1} = \hat{\theta}_N - N^{\beta}P_{N+1}\varphi_{N+1}\varepsilon_{N+1}$$

  where $0 \leq \beta < 1$. This attempt gave no improvement in a few simulated examples.

- The normalized loss function $\frac{1}{N}V_N(\hat{\theta})$ was minimized instead of $V_N(\hat{\theta})$. No significantly improvements occured.

VI. APPLICATION TO PLANT MEASUREMENTS

It was shown in chapter V that the recursive ML method worked well on the simulated data. In order to examine the properties of the algorithm when it is applied to real data, plant measurments were used. Measurements from a nuclear reactor and from a laboratory heat diffusion process were tried. Identification using an off-line ML algorithm on the same data have been made by others. Comparisons are made between the results of the different methods.

It turned out that it is much more difficult to get the algorithm to work satisfactorily on real data. There are probably several reasons for that, for example that the structure and the order of the process is not known.

Different values of the parameters N_1, N_2, and VTEST were tried. The results of the identifications were not very sensitive to the choice of these values. However, it cannot be excluded that better results may be possible to obtain by other choices or by a suitable combination of the tricks mentioned in chapter V.

To illustrate the on-line identification procedure the estimates $\hat{\theta}_t$ and the residuals $\hat{\varepsilon}_t$ are plotted versus the time t. The comparison between the results of the off-line and the on-line algorithms are illustrated with plots of the following signals:

1. the input u(t)
2. the output y(t)
3. the model output $y_m(t) = \frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})} u(t)$
4. the model error $e_m(t) = y(t) - y_m(t)$
5. the residuals $\varepsilon(t;\hat{\theta})$

The model outputs of the on-line models were computed using the parameter estimates obtained at the last sampling interval in the identification.

Example 1

The system is a nuclear power reactor in Ågesta, Sweden. The data were supplied to the Division of Automatic Control by AB Atomenergi, Studsvik, Sweden. The system is described briefly in Gustavsson (1969a) where also ML identifications are reported. The number of data is 1700 and the measurements are called AR 60. The input is control rod position and the output is the nuclear power. An idealized input signal was used both for on-line and off-line identification. The sampling interval is 1 second. Using an F-test it is concluded in Gustavsson (1969a) that the system is of third order.

When recursive ML identification was performed for a model of third order several problems arose. The parameter estimates did not converge. At no time their values were close to the parameter values obtained in Gustavsson (1969a). However, the model outputs of the two models did not differ significantly. A possible explanation of these phenomena is that the order of the model was chosen too high. An indication of this is that both the model in Gustavsson (1969a) and the model obtained by on-line identification have approximately one pole and zero in common.

The results of the on-line identification of a second order model were more satisfactory. The parameters N_1, N_2, and VTEST were chosen as 300, 50 and 1.05 respectively. The parameter estimates obtained are given in Table 6.1. In Gustavsson (1969a) 95 % confidence intervals of the parameter estimates are given. Only the parameters $\hat{a}_1$ and $\hat{c}_2$ of the model obtained on-line are inside these intervals.

	On-line algorithm used	Off-line algorithm used
$\hat{a}_1$	-0.95	-1.08
$\hat{a}_2$	0.14	0.20
$\hat{b}_1$	1.69	1.69
$\hat{b}_2$	-1.12	-1.31
$\hat{c}_1$	-0.76	-0.92
$\hat{c}_2$	0.23	0.27
$\hat{\lambda}$	0.18	0.17

Table 6.1 Results of identification of the nuclear reactor data.

Figure 6.1 shows how the parameter estimates $\hat{\theta}_t$ and the estimated residuals $\hat{\varepsilon}_t$ vary with time. The large values of $\hat{\varepsilon}_t$ at t = 300, 650, 1000 and 1350 are due to the restarts. The large residuals at t = 41, 143, 1233, 1291, 1517, and 1597 are explained by large measurement errors at these points. The measurement errors can be seen clearly from plots of the data.

In Figures 6.2 the model outputs and the residuals are shown for different models. When the second order models are compared it is clear from Table 6.1, Figures 6.1 and 6.2 that there are only small differences between the results of on-line identification and the result of off-line identification. The model output for a third order model computed by on-line identification was very similar to the

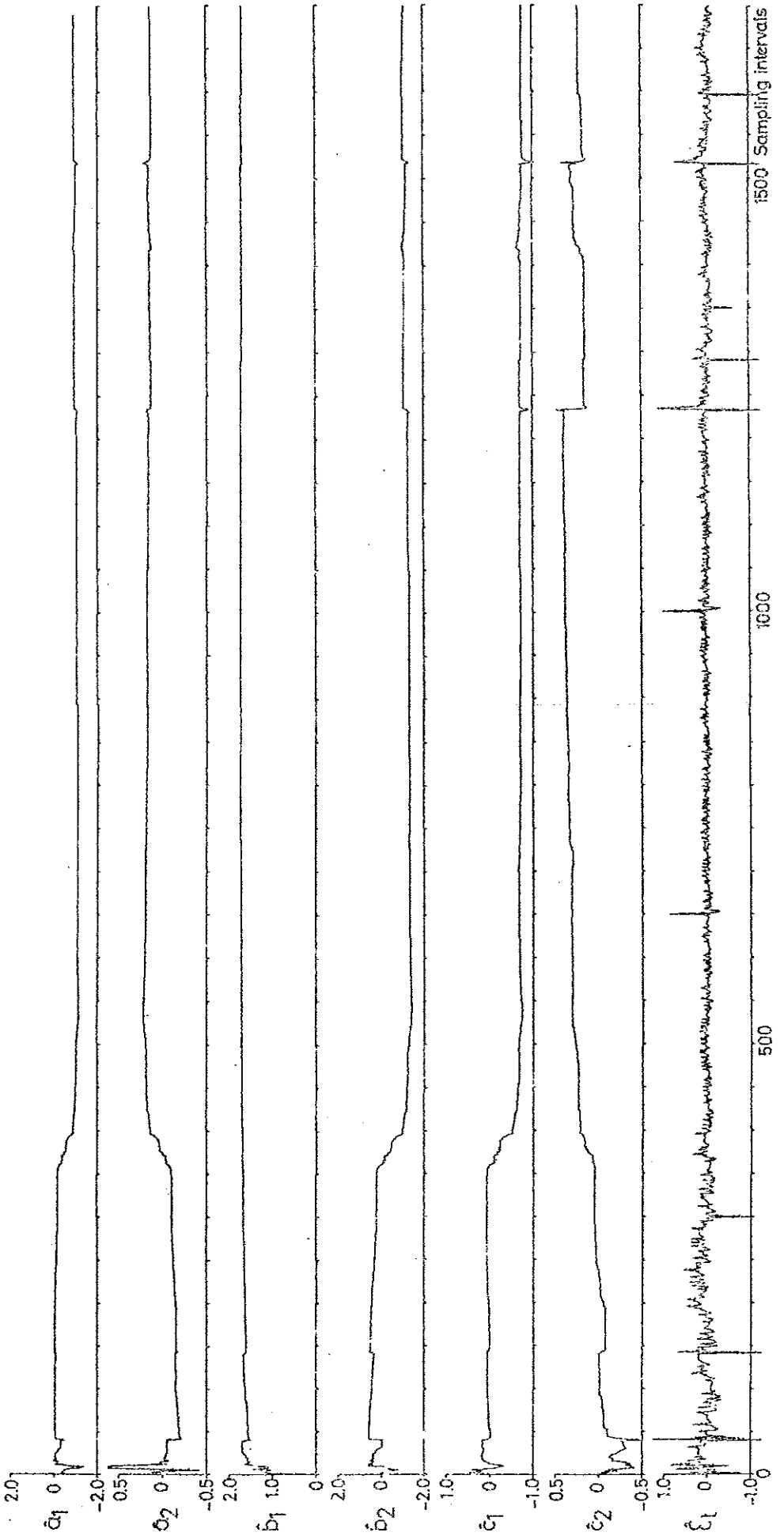

Figure 6.1 The parameter estimates and the residuals estimated for the nuclear reactor data. The sampling interval is 1 second.

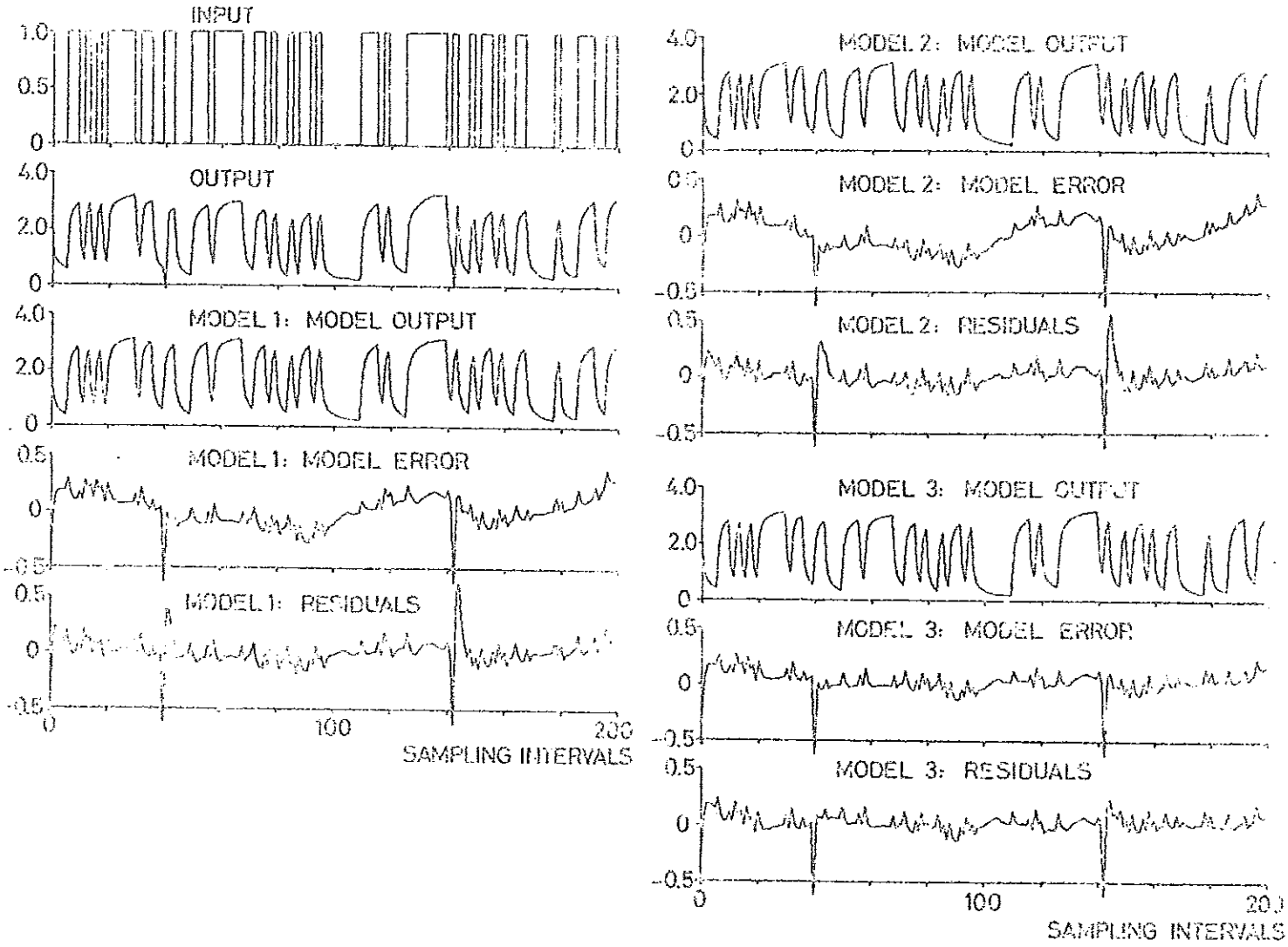

Figure 6.2 Three models of the nuclear reactor.
Model 1 is of second order and is obtained by on-line identification.
Model 2 is of second order and is obtained by off-line identification.
Model 3 is of third order and is obtained by off-line identification.
Digital units are used. The sampling interval is 1 second.
Notice the different scales.

model outputs of the second order models. The best result is obtained with a third order model obtained by off-line identification. However, the improvements are not very great as can be seen in Figure 6.2. The slow oscillation with small amplitude in the model error disappears, however.

Example 2

The system is a laboratory heat diffusion process at the Division of Automatic Control, Lund Institute of Technology. The process consists of a long copper rod. The end temperatures can be controlled using Peltier elements. Identification results of the system using the off-line ML method as well as a short description of the process is given in Leden (1971). The data used here are called series S1. The input is the temperature of one of the end points of the rod. The other end point temperature was kept constant. The output of the process is the temperature in the middle of the rod. The number of data is 862 and the sampling interval is 10 seconds. Leden (1971) found that a model of fourth order was appropriate.

Recursive identification was performed with $N_1 = 200$, $N_2 = 50$, and VTEST = 1.05. The resulting parameter estimates are given in Table 6.2. They differ very much from the estimates obtained with off-line identification. In Figure 6.3 it is shown how the estimates vary with time. The large values of the residuals at t = 200 and 450 are due to the restarts.

The model identified off-line is obtained by a straight-forward application of the ML algorithm. In Leden (1971) also a considerably better model is obtained by inclusion of estimation of initial values and constant errors and by limiting the residuals. This improved model has four real-valued poles and the model error is much smaller than before.

	On-line algorithm used	Off-line algorithm used
$\hat{a}_1$	-0.88	-2.03
$\hat{a}_2$	-0.35	1.40
$\hat{a}_3$	-0.01	-0.40
$\hat{a}_4$	0.27	0.04
$\hat{b}_1 \cdot 10^3$	1.08	0.02
$\hat{b}_2 \cdot 10^3$	1.22	0.46
$\hat{b}_3 \cdot 10^3$	4.43	3.90
$\hat{b}_4 \cdot 10^3$	8.87	2.30
$\hat{c}_1$	0.44	-0.86
$\hat{c}_2$	0.32	0.54
$\hat{c}_3$	0.26	-0.15
$\hat{c}_4$	-0.03	0.24
$\hat{\lambda} \cdot 10^3$	2.53	0.36

Table 6.2 Results of identification of the heat rod data.

In Figure 6.4 the model identified on-line and the model identified by a straight-forward off-line ML algorithm are compared. These two models differ very much in the parameter values. It can be seen from Figure 6.4, however, that the model obtained by on-line identification describes the slowest modes of the process well. When the input is constant for a longer period the residuals are small. The fast modes of the process are badly estimated.

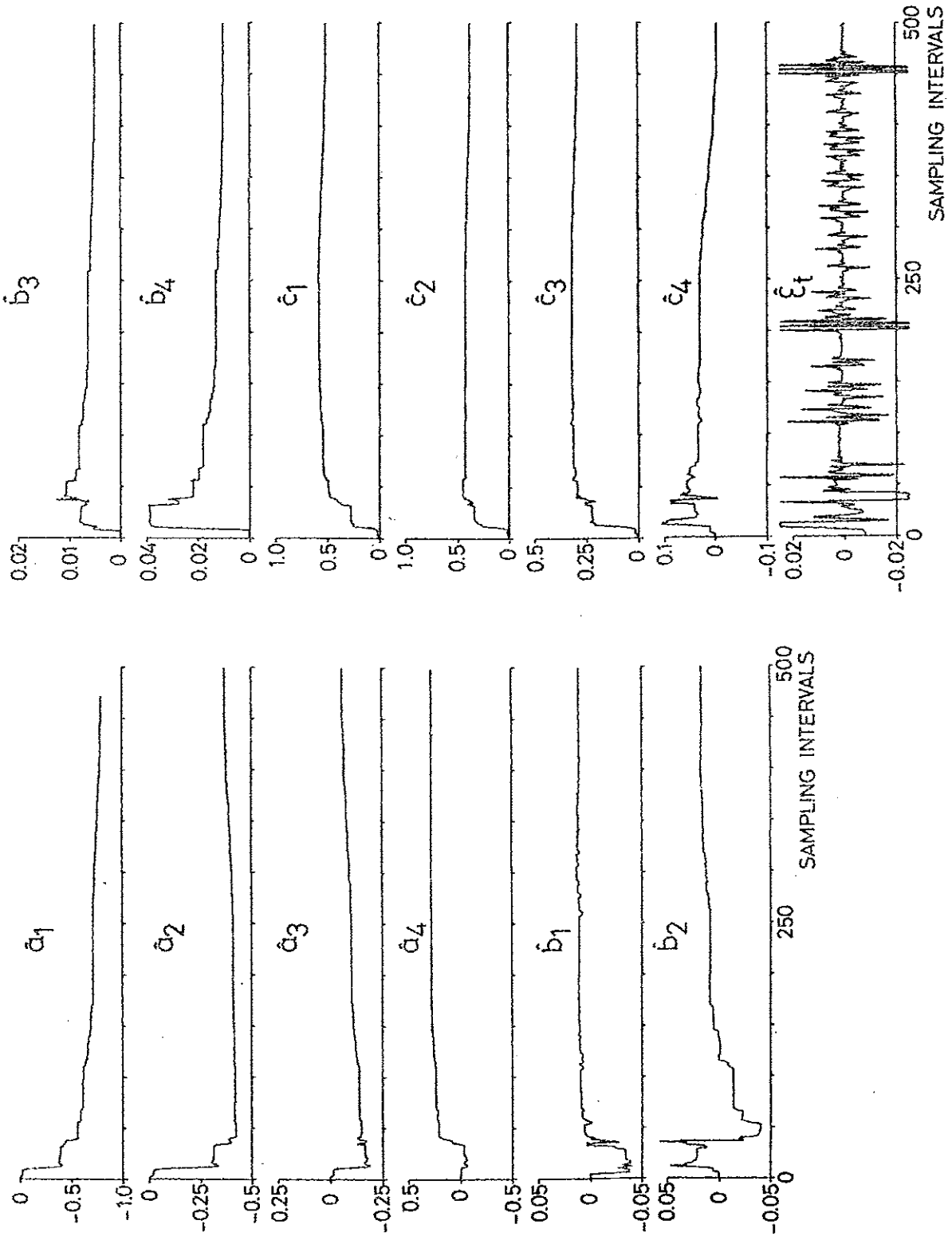

Figure 6.3 The parameter estimates and the residuals estimated for the heat rod data. The sampling interval is 10 seconds.

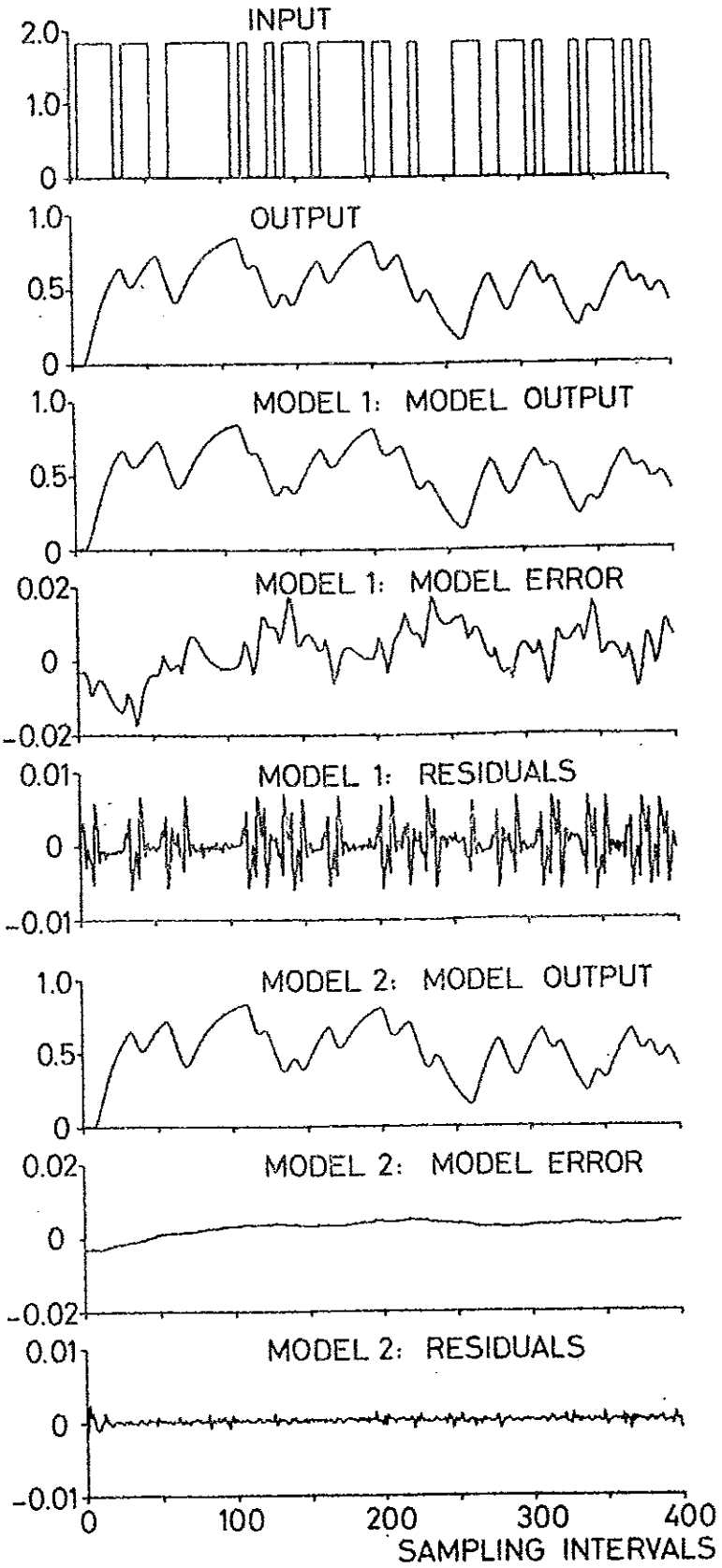

Figure 6.4 Models of the heat diffusion process.

Model 1 is obtained by on-line identification.
Model 2 is obtained by off-line identification.

All variables are given in ^{o}C. Constant levels are added to the input, the output and the model outputs. The sampling interval is 10 seconds. Notice the different scales.

CONCLUSIONS

The following conclusions are based on the examples given.

- The algorithm must be applied with caution. It does not give as good estimates as the off-line ML algorithm.

- Applied to simulated data the algorithm works quite well when suitable tricks are used.

- Applied to real data it is difficult to get the algorithm working satisfactorily. An improper choice of the order of the model may cause considerable difficulties. The most dominating modes of the process are well estimated. It is probably often appropriate to use a low order model.

- The choice of the values of N_1, N_2, and VTEST is not very crucial.

ACKNOWLEDGEMENTS

The author wants to express his great gratitude to Professor K J Åström. He has done a first examination of the algorithm. His guidance was very valuable. Tekn lic Ivar Gustavsson has continued professor Åström's preliminary work. I owe him my sincere thanks for a good cooperation and enlightening discussions.

It is also a pleasure to thank Miss Kerstin Palmqvist who typed the manuscript and Mr Bengt Lander who prepared the figures.

The author is grateful for the data which were supplied to the Division of Automatic Control by AB Atomenergi and Tekn lic Bo Leden, Lund.

This project was partially supported by the Swedish Board of Technical Development under contract 72-202/U137.

REFERENCES

Åström, K.J. (1968).
Lectures on the Identification Problem - the Least Squares Method. Report 6806, Division of Automatic Control, Lund Institute of Technology.

Åström, K.J. - Bohlin, T. (1966).
Numerical Identification of Linear Dynamic Systems from Normal Operating Records. Paper, IFAC Symposium on Theory of Self-Adaptive Systems, Teddington, England. Also in Theory of Self-Adaptive Control Systems (Ed. P.H. Hammond), Plenum Press, New York.

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

Åström, K.J. - Söderström, T. (1973).
Uniqueness of the Maximum Likelihood Estimates of the Parameters of a Mixed Autoregressive Moving Average Process, Report 7306, Division of Automatic Control, Lund Institute of Technology.

Bohlin, T. (1970).
On the Maximum Likelihood Method of Identification. IBM J. Res. and Dev., 14, No 1, 41 - 51.

Clarke, D.W. (1967).
Generalized Least Squares Estimation of the Parameters of a Dynamic Model. 1st IFAC Symposium on Identification in Automatic Control Systems. Prague.

Gustavsson, I. (1969a).
Maximum Likelihood Identification of Dynamics of the Ågesta Reactor and Comparison with Results of Spectral Analysis. Report 6903, Division of Automatic Control, Lund Institute of Technology.

Gustavsson, I. (1969b).
Parametric Identification on Multiple Input, Single Output Linear Dynamic Systems. Report 6907, Division of Automatic Control, Lund Institute of Technology.

Leden, B. (1971).
Identification of Dynamics of a One Dimensional Heat Diffusion Process. Report 7121, Division of Automatic Control, Lund Institute of Technology.

Ljung, L. (1973).
New Convergence Criteria for Stochastic Approximation Algorithms. Forthcoming report. Division of Automatic Control, Lund Institute of Technology.

Ljung, L. - Wittenmark, B. (1973).
Asymptotic Properties of Self-Tuning Regulators Based on Least Squares Identification. Forthcoming report. Division of Automatic Control, Lund Institute of Technology.

Panuška, V. (1968).
A Stochastic Approximation Method for Identification of Linear Systems Using Adaptive Filtering. 1968 JACC, Ann Arbor, Michigan.

Söderström, T. (1972).
On the Convergence Properties of the Generalized Least Squares Identification Method. Report 7228, Division of Automatic Control, Lund Institute of Technology.

Söderström, T. (1973).
On the Uniqueness of Maximum Likelihood Identification for Different Structures. Report 7307, Division of Automatic Control, Lund Institute of Technology.

Valis, J. - Gustavsson, I. (1969).
Some Computational Results Obtained by Panuška's Method of Stochastic Approximations for Identification of Discrete Time Systems. Report 6915, Division of Automatic Control, Lund Institute of Technology.

Wieslander, J. (1971).
Real Time Identification, Part I. Report 7111, Division of Automatic Control, Lund Institute of Technology.

Young, P.C. (1970).
An Extension to the Instrumental Variable Method for Identification of a Noisy Dynamic Process. Univ. of Cambridge, Dep of Eng, Technical note CN/70/1.

Young, P.C. - Shellswell, S.H. - Neethling, C.G. (1971).
A Recursive Approach to Time Series Analysis. Univ. of Cambridge, Dep of Eng, CUED/B - Control/TR16.

APPENDIX

The purpose of this appendix is to show that the equation (4.2) can be substituted by (4.3). The basic tool is the following lemma which is taken from Ljung (1973).

Lemma. Let $\{f_n\}$ be a strictly stationary process such that $E|f_n|$ exists. Assume that the sequence $\{a_n\}$ fulfils

$$a_n \to 0 \quad \text{a.s.} \quad n \to \infty$$

Then

$$\frac{1}{N} \sum_{i=1}^{N} a_i f_i \to 0 \quad \text{a.s.} \quad N \to \infty$$

Corr 1. Let $\{\theta_n\}$ be a sequence of stochastic variables such that

$$\theta_n \to \theta^* \quad \text{a.s.} \quad n \to \infty$$

Further let $\{f_n(\theta)\}$ be a strictly stationary process, which depends on the parameter θ such that $f_n(\theta)$ is (continuously) differentiable with respect to θ a.s. and that $E|f_n'(\theta)|^2$ exists if θ belongs to some neighbourhood of θ^*.

Assume that the sequence $\{a_n\}$ is bounded a.s. Then

$$\frac{1}{N} \sum_{i=1}^{N} [f_i(\theta_i) - f_i(\theta^*)]a_i \to 0 \quad \text{a.s.} \quad N \to \infty$$

Proof The assumptions imply

$$|f_i(\theta_i) - f_i(\theta^*)| \leq M_i|\theta_i - \theta^*| \qquad i \geq N_0$$

for some N_0, where $\{M_i\}$ is a strictly stationary process such that $E|M_i|$ exists. Thus

$$\left|\frac{1}{N}\sum_{i=1}^{N}[f_i(\theta_i) - f_i(\theta^*)]a_i\right| \leq$$

$$\leq \frac{1}{N}\sum_{i=1}^{N_0}|f_i(\theta_i) - f_i(\theta^*)|\,|a_i| + \frac{1}{N}\sum_{i=N_0+1}^{N}|f_i(\theta_i) - f_i(\theta^*)||a_i|$$

$$\leq \frac{1}{N}\sum_{i=1}^{N_0}|f_i(\theta_i) - f_i(\theta^*)||a_i| + \frac{1}{N}\sum_{i=1}^{N}M_i|\theta_i - \theta^*||a_i|$$

The first term trivially tends to zero as N tends to infinity. It follows from the lemma that the second term tends to zero as well.

Corr 2. Let the assumption of $\{a_n\}$ in Corr 1 be changed. Assume instead that $\{a_n\}$ is a strictly stationary process such that $E|a_n|^2$ exists. Then the result of Corr 1 remains true.

In the present algorithm ε_t and φ_t are computed in an approximate way as discussed in chapter II. To simplify the calculations it will be assumed here that they are computed exactly. The results of Ljung-Wittenmark (1973) indicate that it may be possible to extend the calculations to the actual ε_t and φ_t.

Assumptions on the distribution of the noise will be made indirectly. It will be assumed that the expectations

$$E|\varepsilon'(t;\hat{\theta})|^2 \quad \text{and} \quad E|\varepsilon''(t;\hat{\theta})|^2$$

exist for all $\hat{\theta}$ such that the corresponding polynomials

$\hat{A}(z)$ and $\hat{C}(z)$ have all zeros outside the unit circle.

The residuals $\varepsilon(t;\hat{\theta})$ and the gradient $\varepsilon'(t;\hat{\theta})$ are strictly stationary processes if the initial values are chosen properly. However, the effect of the initial values do not affect the result and it will be assumed generally that they are chosen in a proper way.

To simplify, the following notations will be used

$$\varphi_t = \varepsilon'(t;\hat{\theta}_{t-1}) \qquad \varphi_t^* = \varepsilon'(t;\theta^*)$$

$$\varepsilon_t = \varepsilon(t;\hat{\theta}_{t-1}) \qquad \varepsilon_t^* = \varepsilon(t;\theta^*)$$

The calculations are organized as proofs of three assertions.

Assertion 1 $$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \varphi_t \varphi_t^T \hat{\theta}_N = (\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \varphi_t \varphi_t^{*T})\theta^* \quad \text{a.s.}$$

Proof After a decomposition the sum of the left hand side is written as

$$\frac{1}{N} \sum_{t=1}^{N} \varphi_t \varphi_t^T \hat{\theta}_N = \frac{1}{N} \sum_{t=1}^{N} (\varphi_t^* \varphi_t^{*T})\theta^*$$

$$+ \frac{1}{N} \sum_{t=1}^{N} (\varphi_t \varphi_t^T - \varphi_t^* \varphi_t^{*T})\hat{\theta}_N + \frac{1}{N} \sum_{t=1}^{N} \varphi_t^* \varphi_t^{*T}(\hat{\theta}_N - \theta^*)$$

It follows from Corr 1 that the second term tends to zero and from the lemma that the third term tends to zero.

□

Assertion 2 $$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varphi_t\varphi_t^T\hat{\theta}_{t-1} = (\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varphi_t^*\varphi_t^{*T})\theta^* \quad \text{a.s.}$$

Proof A decomposition gives

$$\frac{1}{N}\sum_{t=1}^{N}\varphi_t\varphi_t^T\hat{\theta}_{t-1} = \frac{1}{N}\sum_{t=1}^{N}(\varphi_t^*\varphi_t^{*T})\theta^*$$

$$+\frac{1}{N}\sum_{t=1}^{N}(\varphi_t\varphi_t^T - \varphi_t^*\varphi_t^{*T})\hat{\theta}_{t-1} + \frac{1}{N}\sum_{t=1}^{N}\varphi_t^*\varphi_t^{*T}(\hat{\theta}_{t-1}-\theta^*)$$

Using the same type of arguments as in the preceding proof the assertion follows.

□

Assertion 3 $$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t\varphi_t = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^*\varphi_t^* \quad \text{a.s.}$$

Proof Using a similar decomposition as before

$$\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t\varphi_t = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^*\varphi_t^* + \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^*(\varphi_t-\varphi_t^*)$$

$$+\frac{1}{N}\sum_{t=1}^{N}(\varepsilon_t-\varepsilon_t^*)\varphi_t^* + \frac{1}{N}\sum_{t=1}^{N}(\varepsilon_t-\varepsilon_t^*)(\varphi_t-\varphi_t^*)$$

The second and the third terms tend to zero according to Corr 2. It follows from the lemma (put $f_n\equiv 1$) that the fourth term tends to zero.

□

It can be shown, see e.g. Söderström (1972), that the right hand sides of the assertions really exist under mild conditions.

CORRECTIONS

The abbreviation pa.b denotes page a, line b.

p1.2 Delete "the"

p7.5 Read "derived for the LS case"

p8.17 Read "residuals"

p16.Table5.1 The theoretical RMS error of λ is 0.016

p17.Table5.2 and p21.Table5.3 The theoretical RMS error of λ is 0.016 and the theoretical RMS error of W is 0.035

p17.7 and p20.5 Read "table 5.2"

p19 The scale on the W-axis is incomplete. Figure 5.4 shows the correct scale.

p25.14 Read "are not known"

p26.31 Replace "$\hat{a}_1$" with "$\hat{b}_1$"

p41.11 Replace "φ_t" in the right hand side with "φ_t^*"