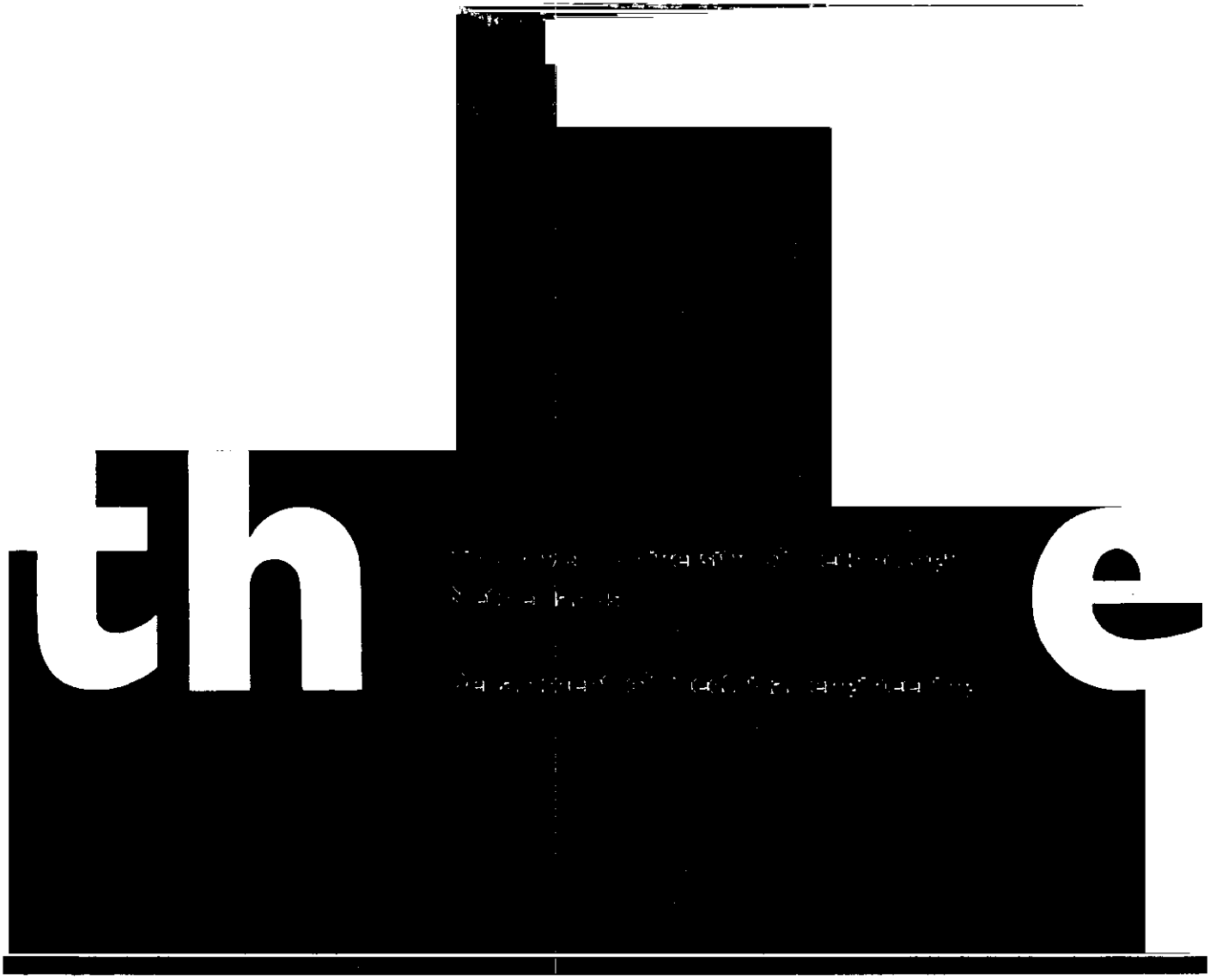
Statistical regression and dispersion ratios in nonlinear system identification



STATISTICAL REGRESSION AND DISPERSION RATIOS IN NONLINEAR SYSTEM IDENTIFICATION

by

ir. A.A. Jongbloed

Group Measurement and Control
Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, the Netherlands

STATISTICAL REGRESSION AND DISPERSION RATIOS IN NONLINEAR SYSTEM IDENTIFICATION

by

ir. A.A. Jongbloed

TH-Report 77-E-70

March 1977

ISBN 90 6144 070X

Summary.

In this report is described the identification of a nonlinear system with a second-degree nonlinear model. The representation is a discrete-time one. After an introduction tot system identification there is a discussion of a measure of identity between process and identification model using dispersion ratios.
Then a statistical analysis shows the possibilities of how to apply dispersion ratios for judging adequacy, when sequential least squares regression procedures are used for identification purposes.

In two examples with second-degree nonlinear systems the discriminative power and the quantified risk in decision making are illustrated.
It is concluded that - especially in case of a small set of observations - the sequential regression methods with adequacy criteria have advantegeous properties in comparison with ordinary correlation methods.

Contents

List of symbols

$\underline{y}$ random output variable; event: $\underline{y} = y$

$\underline{Y}$ random output vector column; event: $\underline{Y} = Y$

$\underline{x}$ random input variable

$\underline{X}$ random vector column

$\mathcal{E}$ mathematical expectation operator

$\mathcal{D}$ dispersion operator (2-11)

L lag operator (section 2.1)

c memory depth (section 1.3)

N number of samples

A process operator (2-2)

A^* estimate for process operator (2-5)

$\underline{z}$ residual (2-7)

Q measure of identity (2-14)

R system-model correlation (2-16)

η statistical regression model (3-6)

$\mathcal{f}$ regression model error (3-12)

V loss function

B regression model parameter vector (3-10)

$\hat{B}$ estimation for B (3-16)

V_R *residual sum of squares* (3-23)

S_R estimate for model error variance (3-24)

ESQ explained sum of squares, corrected for mean (3-28)

DSQ differential sum of squares (3-36)

$\mathcal{f}_m$ dispersion ratio in overall F-test (model adequacy) (3-31)

$\mathcal{f}_c$ dispersion ratio in partial F-test (contribution adequacy) (3-40)

FO Fisher distributed statistic for overall F-test (3-29)

FC Fisher distributed statistic for partial F-test (3-38)

$\hat{Q}$ multiple determination coefficient (3-35)

1. Introduction

1.1 Some remarks about system identification using a model

The dynamic behaviour of a physical system can be represented by a mathematical equation. Mostly such equations contain a set of distinctive parameters, and often the purpose of identification is the determination of those parameters. In most cases we only have the physical reality, i.e. the measured data at inputs and outputs; therefore identification is a matter of recognition:

> "is the system under test an element of a specified class of systems?"

This specified class of systems must be chosen, and it obviously is the formulation of a physical or mathematical hypothesis, based on a priori information ad hoc, that describes the class of systems.
The system under test mostly is called "process"; the elements of the specified class of systems to be compared are called "models".
The formulation of the basic mathematical properties is the fundamental step in identification problems, leading to the models in a particular case.
Then we meet with some important questions to be answered:

a) how to carry out experiments for identification purposes,
b) under what conditions is the process identifiable with one of the specified models, i.e. which criterion for equivalence has to be used.

Starting with the second question, we will formulate such a criterion in terms of a loss function that has to be minimized. Hence we may decide to formulate the identification problem as an optimization problem with a diagnostic character:

> "the purpose to find the best model from the specified set is to obtain a determination of the topology and the estimation of parameter values such as to have minimized the used loss function".(lit.1)

The first question will be answered in the next section.

1.2 Estimation during normal operation of the process

Suppose we deal with a system of which we know or postulate that the behaviour can be described by a second order differential equation:

$$\ddot{x} + a_1\dot{x} + a_2 = a_3 y$$

Now we could compose a set of models with different parameter sets $\{a_1, a_2, a_3\}$ and compare the process and the models by applying adequate test signals like step functions at the input.
The model, the output of which approaches the process output in the best way, i.e. in terms of the minimum loss function, gives the optimizing parameters values $(\hat{a}_1, \hat{a}_2, \hat{a}_3)$. This method with stepfunctions, however, cannot be used in situations where it is not permitted or possible to disturb the normal operation of the process. These situations occur very often in medicine, chemical production etc.
In all following discussions we will use other detection methods, taking care that during the experiments the perturbations must be small.
With respect to the input and output data we only consider stochastic signals, which are sampled; the analysis will be restricted to the stationary case.

1.3 Nonlinear processes

In some cases nonlinear processes can easily be approximated by linear ones. Such cases are not the topic of this discussion.
Because a general and simple classification of nonlinear systems is not available, the identification problem for nonlinear processes can only be solved in an ad hoc way. Several authors distinguish two topological classes of nonlinear systems (lit. 2,3): the Hammerstein and the Wiener structure.
The first one is the sequence: NLNM-LWM; a nonlinear part without memory influence (NLNM) is followed by a linear part with memory (LWM).
It is not possible to measure between the two parts of the total system; the situation is shown in fig. 1.
The Wiener sequence is the reverse: LWM-NLNM (fig. 1).
In this discussion we will restrict the investigations to nonlinearities

of the second degree.

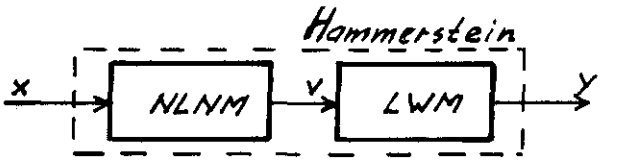
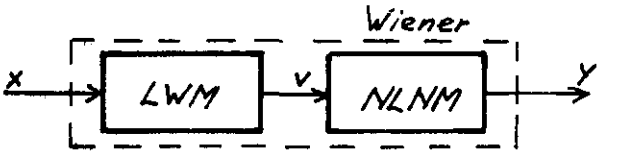


fig. 1



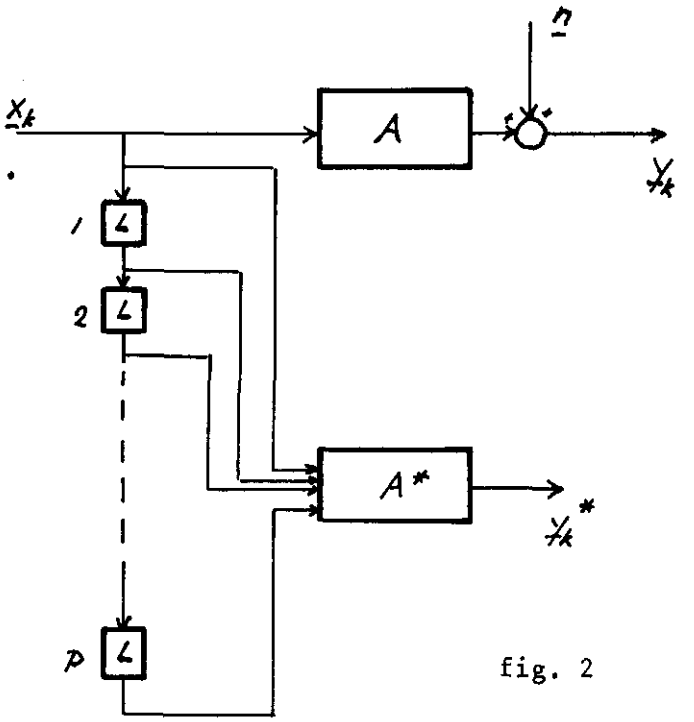


fig. 2

For the dynamic part of the process we consider only the "finite memory" linear behaviour. Remembering the discrete-time situation we have for the Wiener configuration:

$$v_k = \sum_{j=0}^{c} g_j \, x_{k-j} \qquad (1\text{-}1)$$

$$y_k = a v_k^2 + b v_k + d \qquad (1\text{-}2)$$

The Hammerstein structure gives:

$$v_k = ax_k^2 + bx_k + d \qquad (1\text{-}3)$$

$$y_k = \sum_{j=0}^{c} g_j \cdot v_{k-j} \qquad (1\text{-}4)$$

We call g_j the convolution weights; the memory parameter c is called the memory depth.

Now identification of such processes shows two main characteristerics:
- dynamic aspect: the memory depth and convolution weights have to be found,
- topological sequence: a method must be developed in order to decide with which structure we deal, Wiener or Hammerstein.

Details of the answers to these questions are explained in chapter 3.
It will be shown that in this case several "Dispersion Ratios" are useful.
The next chapter gives some mathematical properties of dispersion techniques using the mean squared error, as was described earlier by Rajbman (lit.4).

2. Dispersion Analysis in Identification Procedures

2.1 The mean squared error

The criterion to be minimized will be the mean squared error.
We want to know the best possible estimate of the operator A that describes the process. Here we choose for the process a single input - single output configuration, see fig.2, but the results mentioned are also valid for other structures. It is important to notice that we deal with input variables measured without error; at the output an error is admitted due to several reasons which will be explained presently.
Based on previous investigations, a multiple input - single output structure is used for the model; every model input is obtained from his preceding one by a delay over the sampling interval.
We use the lag operator L: $Lx_n = x_{n-1}$

$$L^k x_n = x_{n-k}$$

The purpose of identification now is to obtain the estimate A^* of the process operator A such as to minimize the criterion of mean squared error:

$$V(y,y^*) = \mathcal{E}\{(\underline{y} - \underline{y}^*)^2\} \tag{2-1}$$

The random variable $\underline{y}$ is the process output, the random variable $\underline{y}^*$ is the output of the model used, fed with the delayed input as shown in fig. 2. Postulating the input-output dependency of the process with the equation:

$$\underline{y}_k = A\{\underline{x}_k, \underline{x}_{k-1}, \cdots, \underline{x}_{k-p}\} + \underline{n}_k \tag{2-2}$$

we can omit the indices k for the sampling moments, referring to the stationarity. Using the notation:

$$\underline{X}_k(p) = \begin{pmatrix} \underline{x}_k \\ \underline{x}_{k-1} \\ \cdot \\ \cdot \\ \underline{x}_{k-p} \end{pmatrix} \tag{2-3}$$

we obtain the postulation equation:

$$\underline{y} = A\{\underline{X};\ p\} + \underline{n} \tag{2-4}$$

As before the indices are omitted.
Now we like to find the estimate A^* of the process operator A in such a way that criterion (2-1) is minimized through:

$$\underline{y}^* = A^*\{\underline{X};\ p\} \tag{2-5}$$

As is known from probability theory (lit. 4,5,6), criterion (2-1) is minimized by:

$$\underline{y}^* = \mathcal{E}\{\underline{y}|X;\ p\} \tag{2-6}$$

So we see: $\underline{y}^* = \underline{y}^*(X; p)$

Finally we consider the meaning of the additive noise $\underline{n}$.
As said before, equation (2-4) has been a postulate. The p delays of the process input fed to the model at any sample moment are based on this postulate. It is not sure, however, that this postulating equation (2-4) is correct. So the noise $\underline{n}$ is not only the measurement error proper, but also the representation of the input information not taken into account. A method to distinguish both parts of the noise is given by Rajbman (lit. 4) in his dispersion analysis.

2.2 Residuals and dispersion; a measure of identity

The residual $\underline{z}$ is defined to be the difference between the outputs of process and model:

$$\underline{z} = \underline{y} - \underline{y}^* \tag{2-7}$$

Selecting $\underline{y}^* = \mathcal{E}\{\underline{y}|X; p\}$ yields:

$$\mathcal{E}\{\underline{z}|X; p\} = 0 \tag{2-8}$$

$$\mathcal{E}\{\underline{z}\} = 0 \tag{2-9}$$

$$\mathrm{Cov}\{\underline{z},\underline{y}^*\} = 0 \tag{2-10}$$

For proofs cf. lit.6.
As can be seen from (2-10), the residual $\underline{z}$ and the model output $\underline{y}^*$, that is optimal in the mean squared sense, are statistically orthogonal.
Defining the operator $\mathcal{D}$ that gives the variance of a random variable $\underline{a}$:

$$\mathcal{D}\{\underline{a}\} = \mathcal{E}\{(\underline{a} - \mathcal{E}\{\underline{a}\})^2\} \tag{2-11}$$

we conclude now:

$$\underline{y} = \mathcal{E}\{\underline{y}|X; p\} + \underline{z} \tag{2-12}$$

and with equation (2-10):

$$\mathcal{D}\{\underline{y}\} = \mathcal{D}\{\mathcal{E}\{\underline{y}|X;\ p\}\} + \mathcal{D}\{\underline{z}\} \qquad (2\text{-}13)$$

Note: equation (2-12) is somewhat misleading: after $y^* = \mathcal{E}\{\underline{y}|X;\ p\}$ has been found we can state (2-12) a posteriori. The correct expression, however, is the definition mentioned, i.e. (2-7)
Equation (2-13) is known as the Decomposition Formula.
Obviously we see the evidence of minimizing $\mathcal{D}\{\underline{z}\} = \mathcal{E}\{\underline{z}^2\}$.
In accordance with Rajbman we define the measure of identity based on a dispersion ratio:

$$Q\{y;p\} = \frac{\mathcal{D}\{\mathcal{E}\{\underline{y}|X;\ p\}\}}{\mathcal{D}\{\underline{y}\}} \qquad (2\text{-}14)$$

The possible interval for $Q\{y;p\}$ is: $0 \leq Q \leq 1$, cf. lit.4.
The purpose is to maximize $Q\{y;p\}$, i.e. to find the optimal value for p.
It does not matter whether the process is a linear or a nonlinear system: in both cases we see, dealing with a memory depth c:

$$\begin{aligned} Q\{y;p\} &< Q\{y;c\} \quad \text{if} \quad p<c \\ Q\{y;p\} &= Q\{y;c\} \quad \text{if} \quad p\geq c \end{aligned} \qquad (2\text{-}15)$$

In order to find the right memory depth c we have to look for that smallest value of p that maximizes $Q\{y;p\}$.
We finally mention the relation between $Q\{y;p\}$ and the model-process correlation $R\{y, y^*;p\}$:

$$\text{with } R\{y, y^*;p\} = \frac{\mathrm{Cov}\{\underline{y}, \underline{y}^*(X;\ p)\}}{\sqrt{\mathcal{D}\{\underline{y}\}}\sqrt{\mathcal{D}\{\underline{y}^*(X;\ p)\}}} \qquad (2\text{-}16)$$

is obtained:

$$Q\{y;p\} = R^2\{y,y^*;p\} \qquad (2\text{-}17)$$

2.3 The lack of knowledge with respect to probability

In the analysis using dispersion operators mentioned before, the complete knowledge of the probability density functions is required, every time we use the operators $\mathcal{E}$ and $\mathcal{D}$. We often can 't say anything especially of the probability density of the process output.
This practical requirement has been the reason of some investigations with respect to estimates. For instance estimates for R, see (2-17), have been tried out, but the results are not satisfying (lit.4).
Well-known estimation methods are available that can well be used in the absence of probability density information, like least squares methods and stochastic approximation (lit. 7,8).
We will investigate now the effectiveness of the application of estimates for dispersion rations like Q, using the theory of quadratic forms.

3. Dispersion Ratios for Judging Adequacy of Regression Models

3.1 Least squares criterion

We now explain a suitable regression method as the solution of the problem posed in the preceding chapter. We take p samples at the input (the choice of the sampling interval will not be discussed here). Hereafter, using identical sampling intervals, N samples are taken both at input and output at the same moments. We form the column vector Y, containing all N samples y_j of the process output. The random character is given by the notation $\underline{Y}$, and an event is indicated by:

$$\underline{Y} = Y$$

We will distinguish two kinds of models:

the identification model, and

the regression model.

The identification model has been described in chapter 1.
The regression model will be defined presently: it represents the postulate that earlier was called the "fundamental step" in identification (section 1.1).

The output of the model gives also N samples, see fig. 2, using the same sampling interval as for sampling the process.
From the N samples $\hat{y}_j$ of the model we compose the column vector $\hat{Y}$.

Remembering equation (2-3) we compose N column vectors $X_j(p)$ with the N+p input samples, using a chain of p delays (fig.2).
We form a matrix of all those columns $X_j(p)$:

$$D_x(p) = \left[X_1(p), X_2(p), \ldots\ldots, X_N(p)\right]^T \tag{3-1}$$

$D_x(p)$ is a N rows - p+1 columns matrix.
The least squares criterion to be minimized is:

$$V(Y,\hat{Y}) = \|Y - \hat{Y}\|^2 = \sum_{j=1}^{N} (y_j - \hat{y}_j)^2 \tag{3-2}$$

It will be pointed out that there exists a special relation between the model column vector $\hat{Y}$ and the data matrix $D_x(p)$: $\hat{Y} = \hat{Y}(D_x(p))$ (3-3)

3.2 Sequential regression model linear in parameters

From equation (1-1), ..., (1-4) we derive:
Hammerstein structure:

$$y_k = d\sum_{t=0}^{c} g_t + b\sum_{t=0}^{c} g_t x_{k-t} + a\sum_{t=0}^{c} g_t x_{k-t}^2 \tag{3-4}$$

Wiener structure

$$y_k = d + b\sum_{t=0}^{c} g_t x_{k-t} + a\sum_{t=0}^{c} g_t^2 x_{k-t}^2 + 2a\sum_{t=0}^{c}\sum_{v=t+1}^{c} g_t g_v x_{k-t} x_{k-v} \tag{3-5}$$

k = 1, 2, ..., N.
A regression model linear in parameters is composed after the structure of both equations (3-4) and (3-5):

$$\eta_k(p+1) = b_0 u_{k,0} + \sum_{t=0}^{P} (b_{1t} u_{k,1t} + b_{2t} u_{k,2t}) \qquad (3\text{-}6)$$

with

$$u_{k,o} = 1; \qquad \text{(dummy)} \qquad (3\text{-}7)$$

$$u_{k,1t} = L^t x_k; \qquad (3\text{-}8)$$

$$u_{k,2t} = (L^t x_k)^2; \qquad (3\text{-}9)$$

$$k = 1,2,\ldots,N.$$

(L is the lag operator defined in section 2.1)

The relation between the parameters of the regression model for p delays and the parameters of the identification model will be pointed out later.
From (3-6) we obtain:

$$H(p+1) = UB(p) \qquad (3\text{-}10)$$

with:

$$H(p+1) = \begin{pmatrix} \eta_1(p+1) \\ \eta_2(p+1) \\ \cdot \\ \eta_N(p+1) \end{pmatrix} .$$

$$B(p) = (b_0, b_{10}, b_{11}, \ldots, b_{1p}, b_{20}, b_{21}, \ldots, b_{2p})^T \qquad (3\text{-}11)$$

$$U(p) = (U_0, U_{10}, U_{11}, \ldots, U_{1p}, U_{20}, U_{21}, \ldots, U_{2p}) = U(D_x(p))$$

We note the dimensions:

$H(p+1)$: $N.1$

$B(p)$: $(2p+3).1$

$U(p)$: $N.(2p+3)$

With the regression equation obtained we postulate the dependency:

$$\underline{Y} = H(p+1) + \underline{\varepsilon} = UB(p) + \underline{\varepsilon} \qquad (3\text{-}12)$$

This is the usual form of the statistical regression model (lit.9). The purpose of least squares regression analysis is, to find

$$\hat{Y} = U\hat{B},$$

or

$$\text{find } \min_{\hat{B}(p)} \left\| Y - U\hat{B}(p) \right\|^2 \qquad (3\text{-}13)$$

In our particular case we compare several methods H(p+1) for increasing values of p. In other words, cf. fig.2, the number of entries of the identification model is increased step by step until a satisfying result has been obtained. The correct description of "satisfying result" is the main aspect of this chapter.

A sequential regression method using eq.(3-6) is useful in our case: without delay (p=0) the identification model leads to the regression model:

$$\eta_k(1) = b_0 u_{k,0} + (b_{10} u_{k,10} + b_{20} u_{k,20}) \qquad (3\text{-}14)$$

And again with (3-6):

$$\eta_k(p+1) = \eta_k(p) + (b_{1p} u_{k,1p} + b_{2p} u_{k,2p}) \quad , \quad p \geq 1 \qquad (3\text{-}15)$$

By means of (3-15) an answer for (3-13) can be found: increasing the number p of delays used, we can find the memory depth c, since more delays than c are not statistically explanatory.

3.3 Parameter estimation and confidence

The parameters $\hat{B}$ that minimize (3-2), see eq. (3-13), are obtained in a well-known way (lit. 7,9):

$$\nabla_{\hat{B}} V(Y, U\hat{B}) = 0 \qquad (3\text{-}16)$$

yields:

$$\underline{\hat{B}} = (U'U)^{-1}U'\underline{Y} \tag{3-17}$$

and

$$\|\underline{Y}\|^2 = \|U\underline{\hat{B}}\|^2 + \|\min_{\hat{B}} \underline{V}(Y,U\hat{B})\|^2 \tag{3-18}$$

This minimum value for the loss criterion V is called V_R:

$$\min_{\hat{B}} V(Y,\hat{Y}) = V_R, \text{ the } \underline{\text{residual}} \text{ value.} \tag{3-19}$$

We investigate some properties of this residual sum of squares V_R, using the standard Gauss-Markov model for (3-12), i.e. (lit. 9):

$$\underline{\varepsilon} \sim \underline{n}_N(0, I_N\sigma^2) \tag{3-20}$$

Hence:

$$\underline{Y} \sim \underline{n}_N(UB, I_N\sigma^2) \tag{3-21}$$

Suppose the regression model to have k parameters. Obviously k is a function of the number of delays p:

$$k = k(p) \tag{3-22}$$

From the theory of mathematical statistics it is known that under this condition $V_R = V_R(k)$ is a central chi-square variate (lit. 10,12):

$$\underline{V}_R(k) \sim \sigma^2 \chi^2(N-k) \tag{3-23}$$

Hereafter we mention some results of the theory of quadratic forms. For proofs the reader is referred to the literature (lit. 6,10,11).
Let

$$\underline{S}_R^2(k) = \frac{\underline{V}_R(k)}{N-k} \tag{3-24}$$

then follows:

$$\frac{\|U(\underline{\hat{B}} - B)\|^2}{k\ \underline{S}_R^2(k)} \sim \underline{F}\ (k,\ N-k) \tag{3-25}$$

This Fisher distribution yield a confidence region for the parameters B after the event $\underline{\hat{B}} = \hat{B}$ has been realized.
With the probability

$$P\{\underline{F}(k, N-k) \leq F_o(\alpha)\} = \alpha$$

the parameter region (3-26)

$$\|U(\hat{B} - B)\|^2 \leq kS_R^2 F_o(\alpha)$$

has a confidence α.

Using projections of the region we obtain confidence intervals for the parameters B; the intervals are based on Student's-t variates (lit. 9).

3.4 Overall hypothesis for the postulated model

In this section a method is exposed how to test the postulated model (3-10). The method can indicate whether increasing the number p of delays might be worthwile or not. We derive some dispersion ratios as criteria for judging the adequacy of the postulation. We use some supplementary definitions:

$$\underline{\bar{Y}} = \frac{1}{N} \begin{pmatrix} 1 \\ 1 \\ . \\ 1 \end{pmatrix} \sum_{j=1}^{N} \underline{y}_j \qquad (3\text{-}27)$$

$$\underline{ESQ}(k) = \left\| \underline{Y} - \underline{\bar{Y}} \right\|^2 \qquad \text{(k parameters in the model)} \quad (3\text{-}28)$$

ESQ(k) is a quadratic form with (k-1) degrees of freedom.
We define the dispersion ratio FO:

$$\underline{FO} = \frac{\underline{ESQ}(k)}{(k-1)\underline{S}_R^2(k)} \qquad (3\text{-}29)$$

Based on a Fisher distribution the following hypothesis can be derived, see (lit. 6,9):

H_o : "all parameters in the regression model except b_o are zero"

against

H_1 : "not all parameters are zero".

H_o is accepted with confidence α if (see (3-26))

$$FO \leq F_o(\alpha) \tag{3-30}$$

If FO exceeds $F_o(\alpha)$, H_o is rejected: the model is found to be explanatory to some extent. In order to judge the extent of adequacy of the model used, the non-central F-variate is used.
As shown by Box and Wetz (lit. 14) non-central variates could indicate that the regression model is not only well-fitting, but is a good tool for predictions too (within the experimental region).
As is derived in previous work (lit. 6)

$$\gamma_m = \sqrt{\frac{\|H - \bar{H}\|^2}{(k-1)\,\sigma^2}} \tag{3-31}$$

is a dispersion ratio giving the ratio of the average of variance in the regression model H and the overall amount of variance in the output of the identification model. We find ESQ(k) to be a non-central chi-square variate with this dispersion ratio γ_m as the non-centrality parameter:

$$\underline{ESQ}(k) \sim \sigma^2 \chi^2(k-1;\gamma_m) \tag{3-32}$$

The overall hypothesis using (3-29) shows that the F-variate involved becomes non-central too.
It seems reasonable to measure the adequacy of the postulated model with the value of γ_m. This criterion, however, is arbitrary to some extent.
The ratio γ_m is included into a hypothesis.
Approximating the non-central variates (lit. 6,13,14) we have:

$$\underline{F}\ (\nu_1,\nu_2;\gamma_m) \approx (1+\gamma_m^2)\ \underline{F}\ (\nu_1^*,\nu_2) \tag{3-33}$$

$$\nu_1^* = \nu_1 \frac{(1+\gamma_m^2)^2}{1+2\gamma_m^2}$$

We see:

$$\nu_1^* > \nu_1 \quad \forall \gamma_m > 1$$

It can be shown that

$$\underline{F}\ (\nu_1, \nu_2;\ \gamma_m = 1) \approx 4\ \underline{F}\ (\nu_1, \nu_2) \tag{3-34}$$

Hereafter we form the adequacy hypothesis:

$$H_o : \gamma_m = 1$$

against

$$H_1 : \gamma_m > 1$$

In order to judge adequacy now, the overall hypothesis of (3-29) and (3-30) is extended as follows:

(i) $F_o(\alpha) < FO < 4F_o(\alpha)$: not all parameters in the model are zero, "$\gamma_m = 1$" is accepted;

(ii) $4F_o(\alpha) < FO$: "$\gamma_m = 1$" is contradicted, the model is considered to be sufficiently adequate.

In this way both hypotheses give a "three-area test":

- the lower area leads to a complete rejection of the model;
- the upper one affirms the adequacy of the model used;
- the middle area give the indication of a well-fitting model, but the adequacy is not shown sufficiently by this result.

Finally we mention the multiple determination coefficient $\hat{\underline{Q}}(k)$:

$$\hat{\underline{Q}}(k) = \frac{\left\|\hat{\underline{Y}} - \bar{\underline{Y}}\right\|^2}{\left\|\underline{Y} - \bar{\underline{Y}}\right\|^2} \tag{3-35}$$

This estimate based on dispersions is a proper analogue of the measure of identity defined in chapter 2. We see again:

$$0 \leq \hat{Q} \leq 1$$

3.5 Model extension and test of contribution; a partial test

The extension of the model with two new parameters (3-15) asks for another

test: has it been useful to add new parameters?
This partial test is important since it gives a criterion to stop the further extension of the model.
Let

$$DSQ(\lambda) = ESQ(k+\lambda) - ESQ(k) \tag{3-36}$$

i.e. a model extension from k to k+λ parameters.
Under the hypothesis that the λ added parameters are zero it can be proved:

$$\frac{DSQ(\lambda)}{\lambda \, S_R^2(k+\lambda)} \sim F(\lambda, N-k-\lambda) \tag{3-37}$$

Consequently this hypothesis is accepted with confidence α if

$$FC = \frac{DSQ(\lambda)}{\lambda S_R^2(k+\lambda)} < F_o(\alpha) \tag{3-38}$$

If FC exceeds, however, this value $F_o(\alpha)$, it is concluded that not all λ new parameters are zero, so the model has been improved to some extent and the addition of the λ parameters has been worthwile.
Again the non-central F-variate will be used to judge the adequacy of the contribution.
Since it can be derived that

$$\mathcal{E}\{DSQ(\lambda)\} = \delta_D + \lambda\sigma^2 \tag{3-39}$$

analogous to (3-31) we form the following ratio (lit. 6,14):

$$\gamma_c = \sqrt{\frac{\delta_D}{\lambda \sigma^2}} \tag{3-40}$$

Note: the amount δ_D is that part of the mathematical expectation of DSQ, that is independent of the model error variance σ^2. In this discussion a further analysis of δ_D is not important.
In accordance with the preceding case, we combine the ordinary test of eq. (3-38) with the hypothesis:

$$H_o : \gamma_c = 1$$

against

$$H_1 : \gamma_c > 1$$

The"three-area test" shows now:

(i) if FC exceeds the significant F_o, the "all zero" hypothesis for the λ new parameters is contradicted; adding the parameters has been worthwile.

(ii) if FC exceeds even $4F_o$, the alternative is affirmed that not only are some of the added parameters explanatory, but that an improvement of adequacy for the whole model is obtained by adding λ parameters.

3.6 Topology decision by testing the cross term contributions

We consider again equations (3-4) and (3-5). After the contribution of two added parameters has shown to give no better statistical explanation, we have found the memory depth c. In order to determine the topological structure of the system, $\frac{1}{2}c(c+1)$ cross terms of eq.(3-5) are added to the regression model (3-6).

We now have:

$$p = c$$

$$u_{k,2tv} = (L^t x_k)(L^v x_k) \tag{3-41}$$

$$\eta_k(c+2) = \eta_k(c+1) + \sum_{t+0}^{c} \sum_{v=t+1}^{c} b_{2tv} u_{k,2tv} \tag{3-42}$$

We deal here with the regression model number (c+2).

Is the contribution of this tail of cross terms statistically significant? This question can be answered by the procedure of the preceding section: we test the contribution after the model number (c+1) is extended with $\lambda = \frac{1}{2}c(c+1)$ parameters. We conclude now:

$$FC = \frac{DSQ\ (\ \frac{1}{2}c(c+1)\ )}{\frac{1}{2}c(c+1)\ \ S_R^2\ (\ \frac{1}{2}(c+2)(c+3)\ )} \tag{3-43}$$

- if $FC \leq F_o(\alpha)$, the tail contribution is not significant (confidence α) the system structure is a Hammerstein one,
- if $FC > 4F_o(\alpha)$, the tail contribution is adequately significant; we have a Wiener configuration.

3.7 Parameter relations

After the decision of memory depth and topological structure the parameter values of the regression model can be determined.
Depending on which topological structure was found, the parameter values of the identification model follow from the regression parameters, see the equations (3-4) and (3-5). These relations give nonlinear statistical models, and this will not be analysed here.
It can be derived that a simple way to determine the convolution weights g_t is to assign the value 1 or 0 to the identificarion model parameter b. Which one is correct can be figured out easily (lit. 6).
This may appear from the examples given in the next chapter.

4. Results with Simulated Systems

4.1 Simulation of a second-degree nonlinear system

We carry out the procedure of sequential regression analysis as pointed out in the preceding chapter. The process to which this nonlinear system identification is applied will be simulated by a data generating structure. The two examples to be discussed in this chapter will contain the list of process parameters (convolution weights, memory depth, parameters of the second-degree polynomial) and the topological structure.
We specify the character of the input data, and the noise added at the process output. Attention is paid to the number of samples.
Both cases indicate the important function of the use of the model hypotheses: dispersion ratios appear to be a worthwile tool for decision making with a strong discriminatory character.
The two samples give some of the most important features that occur when sequential regression models are used; the discussion is, however, not exhaustive.

4.2 Results with a simulated Wiener configuration

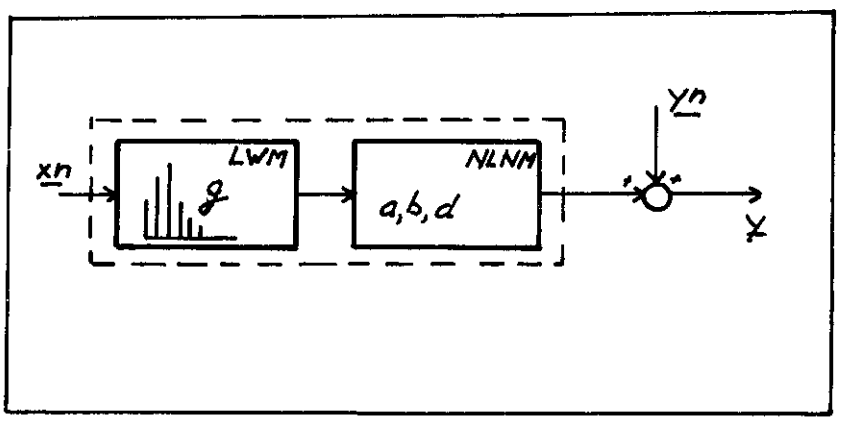


fig. 4.1

We give the simulation characteristics, see fig. 4.1 and eq. (1-1),...,(1-4):

topological structure: Wiener (LWM - NLNM)

LWM: memory depth : c = 6

convolution weights : g_0 = 0.20
g_1 = 0.50
g_2 = 0.70
g_3 = 0.80
g_4 = 0.90
g_5 = 0.95
g_6 = 0.70
g_7 = 0.0

NLNM: parameters polynomial: a = 0.50
b = 1.0
d = 3.0

input source: xn has a uniform distribution over (-4,8), the noise is white;
additive noise: yn is uniform over (-5,5), white noise;
xn and yn are uncorrelated;
number of samples: N = 100.

The sequential extension of regression models with 2 parameters (see section 3.4 and 3.5) gives results as tabulated on this page.
We see that model 7 is the best one: extensions give only non-significant contributions. So is concluded:

$$c+1=7, \quad \text{so } c = 6$$

This conclusion corresponds with the simulated data generating process.
Now adding 21 cross terms we find with equation (3-42) for the model 8:

$$\hat{Q} = 99.9\ \%$$
$$F_o(\alpha) = 1.72 \qquad (\alpha = 95\%)$$
$$4F_o(\alpha) = 6.89$$
$$FC = 611$$

Conclusion: with FC>> $4F(\alpha)$ we find with 95% confidence the Wiener structure to be adequately significant.
The spread of the additive noise, based on this result, is:

$$\hat{\sigma}_{yn} = 2.66, \text{ confidence (95\%) interval: } 2.27 < \sigma_{yn} < 3.22$$

Since the additive noise at the output is uniform over (-5,5) we expect an average noise power $P_{yn} = 8.33$; hence the spread must be: $\sigma_{yn} = 2.86$.
With this result a signal-to-noise ratio of 29 dB is found.

Table of results with the two-parameter extension:

model	m.det.c.	overall F-test (95%)			partial F-test (95%)		
-	$\hat{Q}$ (%)	F_o	$4F_o$	FO	F_o	$4F_o$	FC
1	8	3.1	12.4	4.2	-	-	-
2	25	2.5	10	7.9	3.09	12.37	10.75
3	39.7	2.2	8.8	10.2	3.09	12.37	11.34
4	50.8	2.07	8.3	11.7	3.10	12.39	10.27
5	61.8	1.95	7.8	14.4	3.10	12.40	12.76
6	74.5	1.87	7.5	21.2	3.10	12.41	21.78
7	83.3	1.80	7.2	30.3	3.10	12.42	22.40
8	83.6	1.75	7.0	26.4	3.11	12.43	0.67
9	83.7	1.70	6.8	23.1	3.11	12.44	0.29
10	83.9	1.7	6.8	20.7	3.11	12.45	0.59
11	84.6	1.7	6.8	19.2	3.12	12.46	1.62
12	84.8	1.7	6.8	17.4	3.12	12.47	0.39
13	84.9	1.7	6.6	15.8	3.12	12.49	0.36
14	85.3	1.6	6.5	14.7	3.13	12.50	0.99
15	86.1	1.6	6.5	14.2	3.13	12.52	1.88

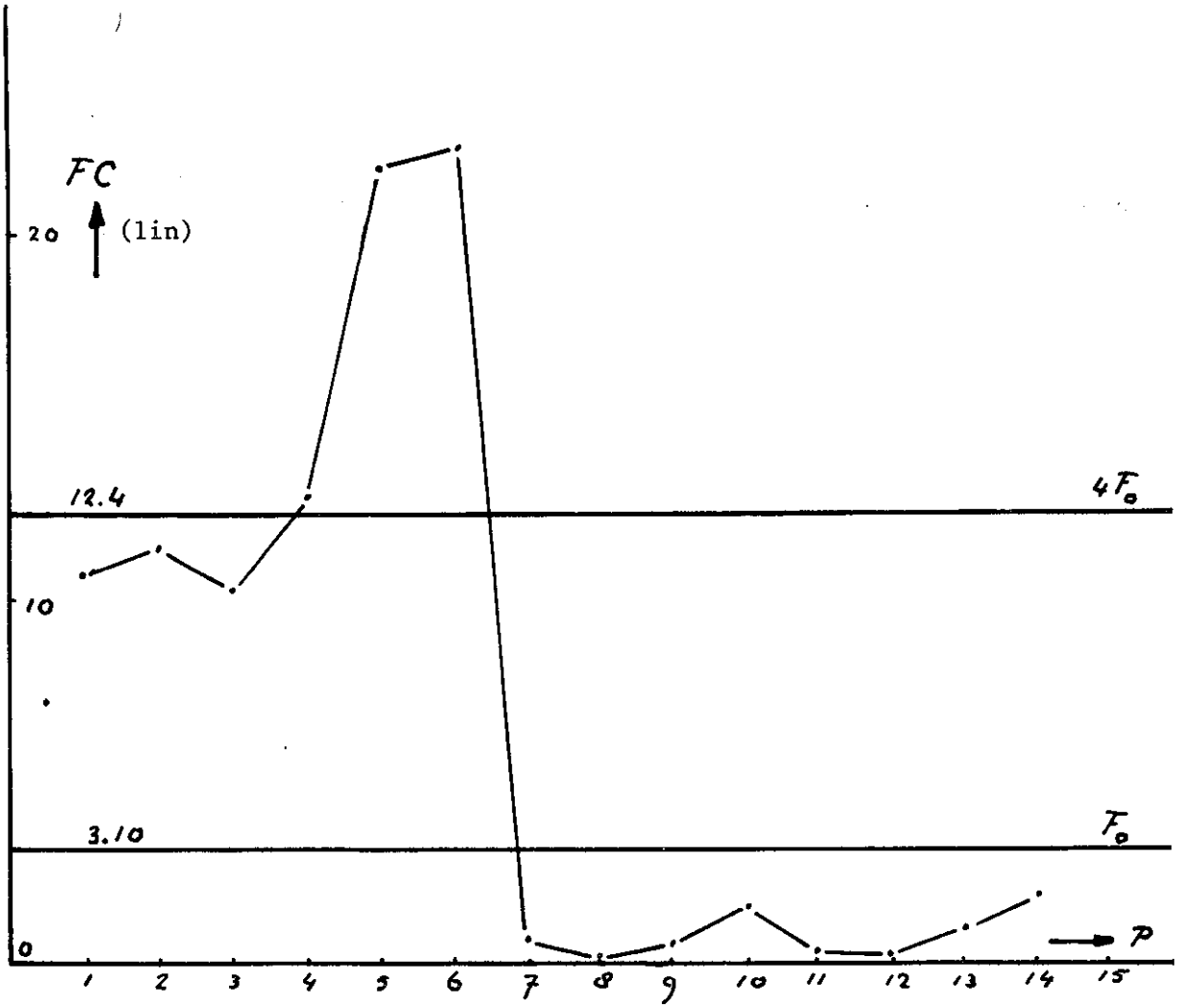

FC
20
(lin)
12.4
4 F₀
10
3.10
F₀
0
P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

fig. 4.2

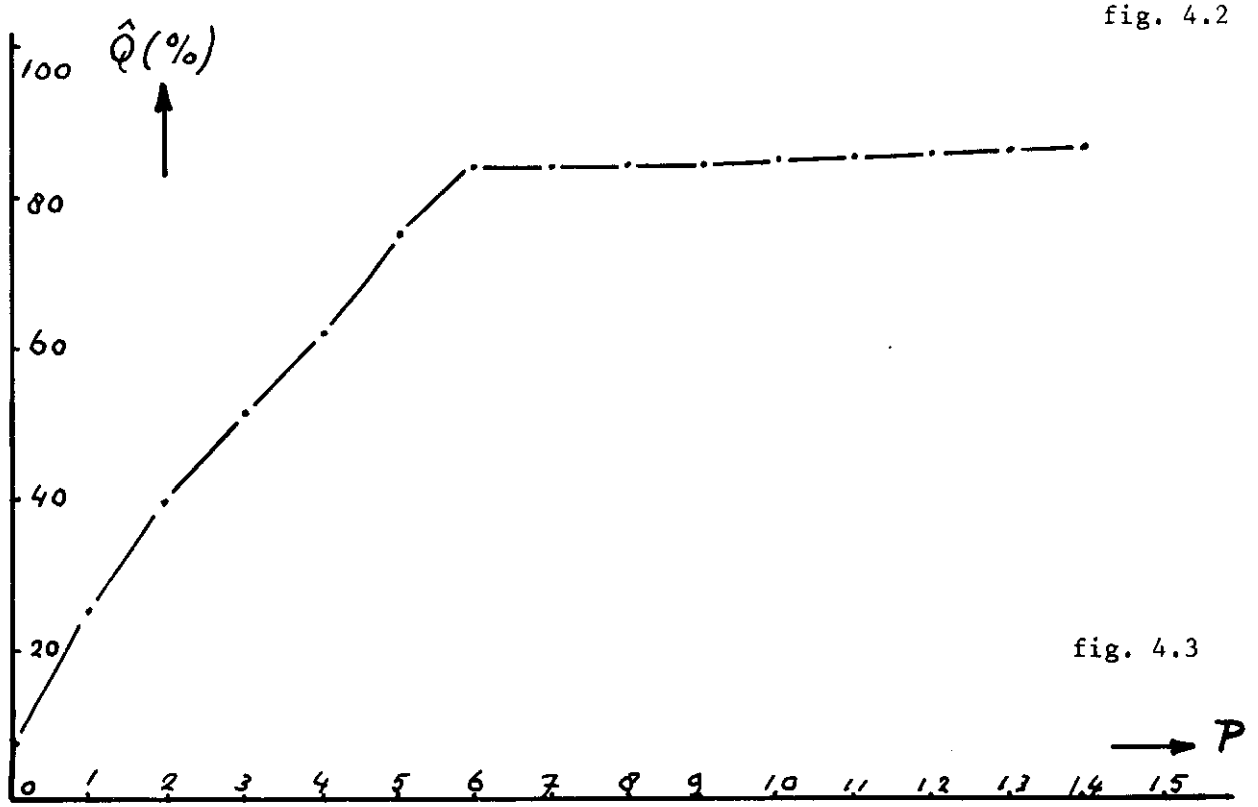

Q̂(%)
100
80
60
40
20
0
P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

fig. 4.3

In fig. 4.2 the path FC = FC(p) is shown; it gives the behaviour of the contribution with increasing memory depth of the identification model. It obviously shows the meaning of the "three-area test".
Fig. 4.3 shows the behaviour of the multiple determination coefficient $\hat{Q}$. This estimate for the system-model correlation shows a misleading well-discriminative power: as appears in the next example, decisions based on this quantity are not safe and sometimes risky.
In many experiments carried out the discriminative power using partial F-tests is as strong as shown in fig. 4.2
The conclusions drawn from these hypotheses give a quantified risk (5%); therefore for decision making FC-events are preferable.
The estimate $\hat{Q}$ for $Q\{y; p\}$, see chapter 2, affirms the statement in eq. (2-15).
We don't mention the estimate results for the parameters in this example: parameter results are the main aspect in the discussion of the next example.

4.3 Hammerstein simulation with memory depth c=7

The data generating structure was a sequence NLNM - LWM, as is shown in fig. 4.4. The linear part with the influence of memory has a discrete convolution with the parameters:

c = 7	g:	0	1.0
		1	0.70
		2	0.49
		3	0.343
		4	0.240
		5	0.168
		6	0.118
		7	0.082
		8	0.0

The nonlinear part is characterized by:

a	1.31
b	0
d	0.70

Input data: xn uniform over (-1.8, 5.2), white noise.
Added noise: yn is a normal white variate, no correlation with xn; $yn \sim n(0, 0.25)$

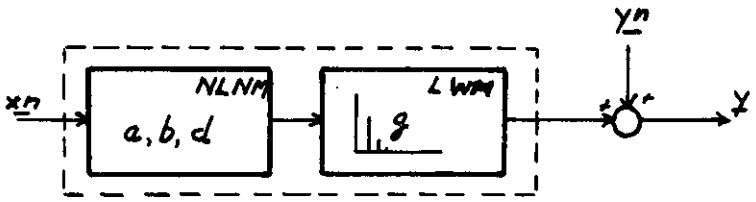


fig. 4.4

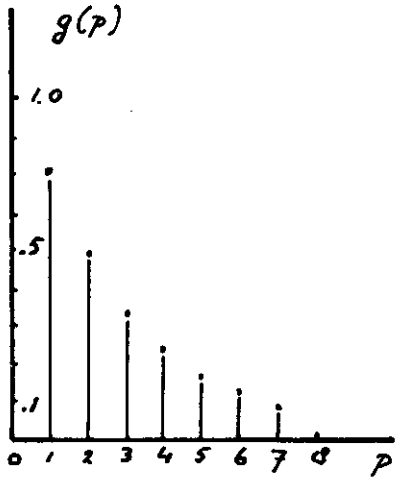


fig. 4.5

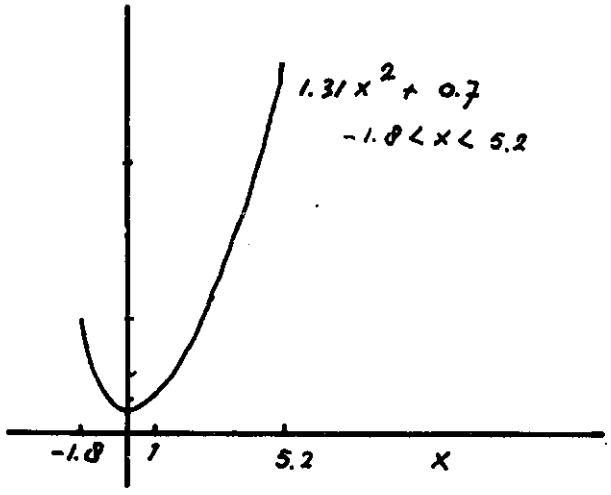


fig. 4.6

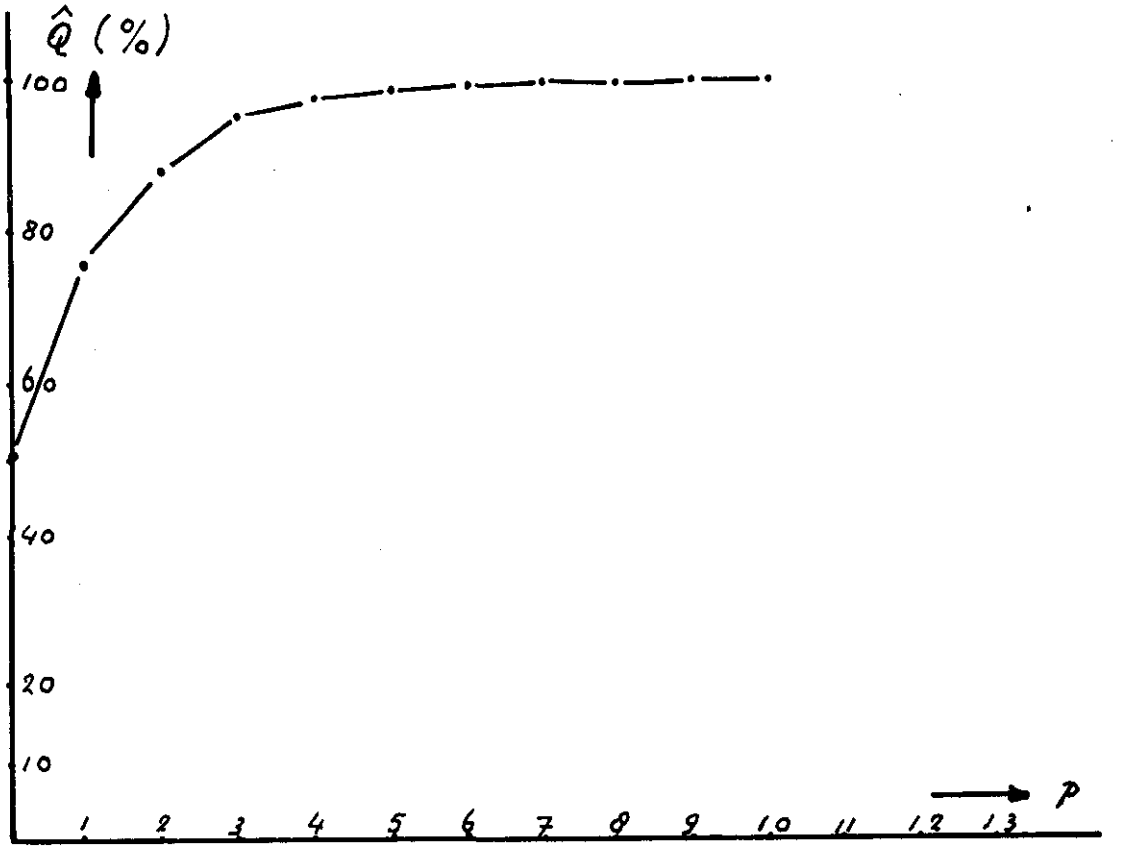

$\hat{Q}$ (%)
100
80
60
40
20
10
1
2
3
4
5
6
7
8
9
10
11
12
13
P

fig. 4.7

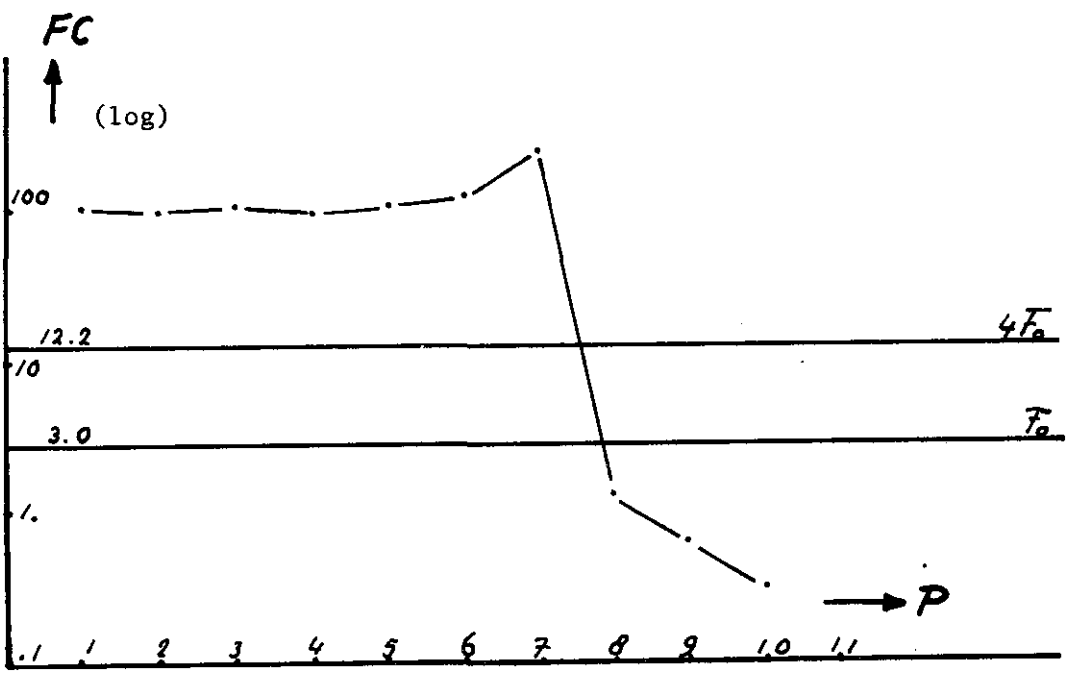

FC
(log)
100
12.2
10
3.0
1.
.1
1
2
3
4
5
6
7
8
9
10
11
$4F_o$
F_o
P

fig. 4.8

Fig. 4.6 shows the nonlinear character in the process, so an adequate regression function will have more terms than only linear ones.
With the set-up demonstrated in fig. 2 we analyse the process from 200 samples. Sequential regression analysis is carried out now, and the results are given in fig. 4.7 and 4.8. We mention only the partial hypotheses results:

memory depth found: c = 7

topology found : Hammerstein (no significant cross term contribution)

The result of the partial FC-test has again a discriminative character, see fig. 4.8. The path of $\hat{Q}$, however, gives no reason for decision: as mentioned in the preceding example, the discriminative power of $\hat{Q} = \hat{Q}(p)$ is not as strong to be useful for decision purposes.

We notice that the time-discrete convolution starts with its maximum value, $g_0 = 1 = g_{max}$; in consequence we find a high starting point for $\hat{Q}(p=0)$ in fig. 4.7. This first model, that is explanatory for 50%, is also adequate.

We must conclude that, due to the maximum value for the non-delayed input sample, an identification model without any memory could be regarded as an acceptable solution (only for purposes in the observed region).

We now consider the estimation results.

Estimation results and confidence intervals for the parameters of the regression model are tabulated below.

We mention the round noise level: at the output the signal-to-noise-ratio is 29 dB (N = 200).

The hypothesis H_o : "the parameter is zero"

against H_1 : "the parameter is significantly different from zero"

is tested now by the resulting value of the parameter: if this event lies in the interval formed by the minus value and the plus value of the number of the column "hypothesis", then H_o is accepted under 95% confidence.

In other caseswe find its contradiction.

parameter	value	confidence (95%)		hypothesis
b_0	2.25	2.06	2.46	0.21
b_{10}	0.009	-0.072	0.091	0.081
b_{20}	1.31	1.29	1.33	0.021
b_{11}	0.02	-0.061	0.101	0.081
b_{21}	0.91	0.892	0.934	0.021
b_{12}	-0.075	-0.160	0.006	0.082
b_{22}	0.66	0.637	0.679	0.021
b_{13}	0.01	-0.070	0.094	0.082
b_{23}	0.45	0.428	0.471	0.021
b_{14}	0.01	-0.071	0.092	0.081
b_{24}	0.31	0.284	0.327	0.021
b_{15}	-0.067	-0.148	0.014	0.081
b_{25}	0.24	0.214	0.256	0.021
b_{16}	-0.082	-0.163	-0.0	0.082
b_{26}	0.17	0.149	0.192	0.021
b_{17}	-0.002	-0.08	0.08	0.082
b_{27}	0.11	0.09	0.13	0.021

All linear parameters b_{1j}, j=0,1,...,7, have values falling within the acceptance region of the hypotheses. Together with the earlier found conclusion that we deal with a Hammerstein configuration, we must conclude that there is a systematic cause for this hypothesis acceptation: considering the present effect of memory influence in the other part of the regression function, we conclude:

$$b = 0$$

Now no limitations are involved in the assignment: a := 1.

The last part of the discussion of the results is to compare the realized and the expected results.
In this situation we have generated the data with the nonlinear coefficient a = 1.31, and in the identification we stated: a = 1.
So the original values of g_p, p=0,1,....,7 must be multiplied with 1.31 in order to compare the expected results with the obtained estimations.
Though we have to mention also the confidence intervals, the two columns are:

	EXPECTED	FOUND
d_H	2.1987	2.25
ag_p, p=0	1.31	1.31
1	0.92	0.91
2	0.64	0.66
3	0.45	0.45
4	0.31	0.31
5	0.22	0.24
6	0.16	0.17
7	0.11	0.11

The result is very satisfying.
From the result $ag_0 = 1.31$ we can also conclude:

$$\hat{g}_0 := 1 \quad \text{and} \quad \hat{a} = 1.31$$

Hence:

$$\hat{g}_j := \hat{g}_j/1.31, \qquad j=1,2,\ldots,7.$$

These manipulations are, however, not essential for the identification result.

5. Conclusions

The main aspect in using dispersion ratios for judging adequacy and parameter significancy in statistical regression methods with sequential character for nonlinear discrete -time system identification is the possibility to make decisions with a quantified risk level.
Based on the confidence analysis by testing several hypotheses, conclusions can be formulated for the two important aspects:

- the effectiveness of memory (memory depth), and
- the topological structure.

The statistical approach can be useful in linear cases too.

As is illustrated with the two examples described in the preceding chapter, we see that decisions made by using partial Fisher-tests have more discriminative background than they should have with the application of system-model correlation estimates. If estimates for this correlation were used, see eq. (2-17) and (3-35), only the hypothesis that outputs of model and process are uncorrelated could be tested under a safe accuracy level (Pearson correlation test). If this correlation differs, however, significantly from zero, estimates for the system-model correlation are biased.
Under circumstances of small numbers of samples (lit. 6) estimation results may lead to very misleading conclusions about the presence of correlation.
A confidence interval is hard to give; mostly the impression about such confidence given by some authors is not correct.
Especially with small numbers of samples the method of least squares regression with dispersion ratios as the critera for judging adequacy is now very advantageous. Another good aspect is, that the successful identification procedure result does not require any information of probability density of the stochastical input or output data sequences.
As indicated in the examples, the character of the additive noise does not influence the developed decision strategy. Though we mentioned only white noise input situations, this sequence can be a coloured one as well.
In that case the analysis could be extended by distinguishing "pure errors" and "lack of fit", see lit. 9.
Finally is suggested that dispersion ratio techniques could be applied in other more complicated cases of nonlinear system identification; decision strategy and regression functions must be developed in an ad hoc approach.

6. Literature

1. Åstrøm K.J. and Eykhoff P.
"System Identification, a survey"
Automatica, vol.7, pp. 123-162. (1971)

2. Haber, R.
"Identification of nonlinear discrete processes"
Master thesis, Univ. Budapest, dep. Automatic Control (1972)

3. Haber R. and Keviczky L.
"The identification of the discrete-time Hammerstein model"
Separatum Periodica Polytechnica, EE, vol. 18 no. 1 (1974)

4. Rajbman N.S.
"What is Identification" (in Russian)
Nauka Moscow (1970)

5. Papoulis A.
"Probability, Random Variables, and Stochastic Processes"
MacGrawHill (1966)

6. Jongbloed A.A.
"The Identification of Discrete-time Nonlinear Systems with Finite Memory; Applications of Dispersion Ratios"
Master Thesis, Eindhoven, Univ.of Technology, dep. EE/ER (1976)

7. Eykhoff P.
"System Identification"
John Wiley (1974)

8. Richalet J.
"Identification des processus par la méthode du modèle"
Gordon/Breach Paris (1975)

9. Draper N. and Smith H.
"Applied Regression Analysis"
John Wiley (1966)

10. Rao C.R.
"Linear Statistical Inference and its Applications"
John Wiley (1973)

11. Graybill F.A.
"An introduction to Linear Statistical Models"
MacGraw-Hill (1961)

12. Craig A.T. and Hogg R.V.
"Mathematical Statistics"
Collier Mc Millan, ed. 3 (1972)

13. Abramowitz M. and Stegun I.
"Handbook of Mathematical Functions"
Dover Publications (1962)

14. Box G.E.P. and Wetz J.
"Criteria for judging adequacy of estimation by an approximating response function"
Technical Report no.9, Univ. of Wisconsin-Madison, dep. of Statistics (1973)

Acknowledgements

No book is written in a vacuum.
I was given the opportunity to write this text after a period of research in the "stochastic section" of the group Measurement and Control of prof.dr.ir. P. Eykhoff, and I am grateful for his advices in preparing the manuscript.

I wish to acknowledge the encouragements and skilful help of Mrs.J. Kregting-Jansen.
I am grateful to Mr. F. Gerretsen and to Dr. J.F. Barrett for the comments and suggestions and to Mrs. A. Vermeulen for typing the manuscripts.