[redacted]

A Prediction Error Estimator for Nonlinear Stochastic Systems

I. J. Leontaritis B.Sc., M.Sc., Ph.D.

S. A. Billings Ph.D., B.Eng., C.Eng., MIEE., AFIMA., M.Inst.,MC.

Department of Control Engineering
University of Sheffield
Sheffield
S1 3JD

March 1986

RESEARCH REPORT NO, 295.

Abstract

A prediction error estimation algorithm incorporating model selection and validation techniques is developed for a class of multivariable discrete-time stochastic nonlinear systems which can be represented by the NARMAX model (Nonlinear AutoRegressive Moving Average Model with eXogenous inputs).

1. Introduction

The development of microprocessor based control algorithms for nonlinear systems would be simplified if identification techniques could be developed to yield representative models of such systems in a stochastic environment.

The present paper introduces a prediction error estimation algorithm for systems which can be described by the NARMAX model (Nonlinear AutoRegressive Moving Average model with eXogenous inputs) |Leontaritis and Billings 1985|. An algorithm for the estimation of the parameters in a NARMAX model is developed using a prediction error estimator based on Newton's method with a line search at every step. The precise numerical implementation using square root methods is described in detail. Two basic methods of model selection, the model reduction and the model expansion method are described as an integral part of the estimation. The Householder orthogonal transformation is employed in the model reduction to perform the selection of the model in an efficient manner. Model validation |Billings and Voon 1983, 1986a, Leontaritis and Billings 1986| is also briefly described and simulated examples are included. The algorithms presented represent an alternative to the prediction error/stepwise regression method described in an earlier publication |Billings and Voon 1986b|.

2. The NARMAX Model

Assume that the system that generated the data which is to be analysed is a general stochastic discrete time system with input space on r-dimensional vector space and output space on m-dimensional vector space. Define the input and output of the system at time t as the r-dimensional and m-dimensional column vectors u(t) and y(t) respectively. If the observation of the system is assumed to start from time 1 to time t this can be denoted by y^t

$$y^t=(y(t)^\tau\ ,\ y(t-1)^\tau,\ \ldots\ ,\ y(1)^\tau)^\tau \tag{1}$$

Similarly for u(t)

$$y^t=\ (u(t)^\tau,\ u(t-1)^\tau,\ \ldots\ u(1)^\tau)^\tau \tag{2}$$

A general stochastic, discrete-time, dynamical system can now be described by the conditional probability density function of y(t) given all past inputs and outputs y^{t-1} and y^t

$$p(y(t) \mid y^{t-1}, u^t) \tag{3}$$

The function eqn.(3) can be put into innovation form

$$y(t) = f(y^{t-1}, u^t) + \varepsilon(t) \tag{4}$$

Where $\varepsilon(t)$, the prediction error or innovation process is the stochastic process defined as

$$\varepsilon(t) = y(t) - E[y(t) \mid y^{t-1}, u^t] \tag{5}$$

The mean square error estimate of the output y(t) given all past inputs and outputs, is the vector $\hat{y}(t)$

$$\hat{y}(t) = E[y(t) \mid y^{t-1}, u^t] = f(y^{t-1}, u^t) \tag{6}$$

and thus the innovation form eqn. (4) separates the output that can be predicted from the past as $f(y^{t-1}, u^t)$ and the unpredictable part as the innovation $\varepsilon(t)$.

Equation (4) can be expressed in expanded form as [Leontaritis and Billings 1985]

$$\begin{aligned} y_i(t+p) = q_i[& y_1(t+n_1-1),\ y_1(t+n_1-2),\ \ldots\ y_1(t), \\ & y_2(t+n_2-1),\ y_2(t+n_2-2),\ \ldots\ y_2(t), \\ & \vdots \\ & y_m(t+n_m-1),\ y_m(t+n_m-2),\ \ldots\ y_m(t), \\ & u_1(t+p),\ u_1(t+p-1),\ \ldots\ u_1(t) \\ & u_2(t+p), u_2(t+p-1), \ldots,\ u_2(t) \\ & \vdots \\ & u_r(t+p),\ u_r(t+p-1), \ldots,\ u_r(t) \\ & \varepsilon_1(t+p-1),\ \ldots, \qquad \varepsilon_1(t) \\ & \varepsilon_m(t+p-1),\ \ldots, \qquad \varepsilon_m(t)] + \varepsilon_i(t+p) \end{aligned} \tag{7}$$

where i=1,2,...m and p=max $(n_1, n_2, \ldots n_m)$. The usagers n_i are the observability indices of the system, and the function f(·) eqn(4) can be found by recursive use of the functions q_i.

The representation in eqn.(7) is referred to as a multistructural input-output prediction error or innovation model. For single-input single-output systems this reduces to the NARMAX model (Nonlinear AutoRegressive Moving Average model with eXogenous inputs).

$$y(t) = q[y(t-1), \ldots y(t-n_y), u(t-d), \ldots u(t-d-n_u+1), \varepsilon(t-1), \ldots \varepsilon(t+n_\varepsilon)] + \varepsilon(t) \quad (8)$$

A rigorous derivation of these results together with numerous examples are available in the literature |Leontaritis and Billings 1985|.

3. Parameter estimation

The prediction error and maximum likelihood estimation techniques both minimize a loss function |Goodwin and Payne 1977, Ljung and Soderstrom 1983|. The maximum likelihood method is an asymptotically optimum method but the probability density function of the innovations must be known. The prediction error method does not require any density function to be known and is equivalent to the maximum likelihood method for the case of Gaussian innovations. It can be shown that the performance of the prediction error method is only slightly inferior to the maximum likelihood method for bell-shaped probability density functions of the innovations. The estimate of the parameter vector θ given by the prediction error method is the one that minimizes the loss function

$$J_2(\theta) = \frac{1}{2} \log \det Q(\theta) \quad (9)$$

where

$$Q(\theta) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon(t,\theta) \quad \varepsilon^T(t,\theta) \quad (10)$$

$\varepsilon(t,\theta)$ are the residuals

$$\varepsilon(t,\theta) = y(t) - f(y^{t-1}, u^t, \theta) \quad (11)$$

and the function f(·) eqn.(4) is now expressed as a function of θ to show the

dependence on the parameter vector. The gradient and the approximate Hessian of the loss function $J_2(\theta)$ are given by

$$\frac{\partial J_2}{\partial \theta_i} = \frac{1}{N} \sum_{t=1}^{N} \varepsilon(t,\theta)^T Q(\theta)^{-1} \frac{\partial \varepsilon(t,\theta)}{\partial \theta_i} \qquad i = 1,2,\ldots,n_\theta \tag{12}$$

$$\frac{\partial^2 J_2}{\partial \theta_i \partial \theta_j} = \frac{1}{N} \sum_{t=1}^{N} \frac{\partial \varepsilon(t,\varepsilon)^T}{\partial \theta_i} Q(\theta)^{-1} \frac{\partial \varepsilon(t,\theta)}{\partial \theta_j} \quad i,j=1,2,\ldots,n_\theta \tag{13}$$

where the derivatives of the residuals can be calculated from the equation

$$\frac{\partial \varepsilon(t,\theta)}{\partial \theta_i} = - \frac{\partial}{\partial \theta_i} f(y^{t-1}, u^t \theta) \qquad i=1,2,\ldots,n_\theta \tag{14}$$

The derivatives of the function f of the model have thus to be calculated with respect to the parameters.

It is possible to calculate the derivatives of the residuals recursively. In fact differentiating eqn.(7) with respect to the parameters, the following recursions are found

$$\begin{aligned} \frac{\partial \varepsilon_i(t+p)}{\partial \theta_j} = & - \frac{\partial q_i}{\partial \varepsilon_j} - \frac{\partial q_i}{\partial \varepsilon_1(t+n_1-1)} \frac{\partial \varepsilon_1(t+n_1-1)}{\partial \theta_j} - \ldots - \frac{\partial q_1}{\partial \varepsilon_1(t)} \frac{\partial \varepsilon_1(t)}{\partial \theta_j} \\ & - \frac{\partial q_i}{\partial \varepsilon_2(t+n_2-1)} \frac{\partial \varepsilon_2(t+n_2-1)}{\partial \theta_j} - \ldots - \frac{\partial q_1}{\partial \varepsilon_2(t)} \frac{\partial \varepsilon_2(t)}{\partial \theta_j} \\ & \cdots \cdots \cdots \cdots \\ & - \frac{\partial q_i}{\partial \varepsilon_m(t+n_m-1)} \frac{\partial \varepsilon_m(t+n_m-1)}{\partial \theta_j} - \ldots - \frac{\partial q_1}{\partial \varepsilon_m(t)} \frac{\partial \varepsilon_m(t)}{\partial \theta_j} \end{aligned} \tag{15}$$

where $i=1,2,\ldots,m \qquad j=1,2,\ldots,n_\theta$

and $p=\max(n_1,n_2,\ldots,n_m)$

The $m*n_\theta$ recursions eqn.(15) can be used to find the derivatives of the residuals with respect to the parameters [Billings and Voon 1986b].
The recursions need initial values for the derivatives of the residuals. Since the residuals do not depend on the parameters for $t \leq 0$, it is clear that the initial conditions are $(\partial \varepsilon(t)/\partial \theta_i)=0$ for $t \leq 0$.

If the function f is to be derived from the functions q_i certain values of the inputs, outputs and residuals before t=1 have to be known. The first p inputs and outputs of the data can be used as initial values of the inputs and outputs. The rest of the data are available for the estimation of the parameters. The initial values of the residuals present a more difficult problem. One usually acceptable solution is to assume that they are zero.

The gradient and the Hessian of the loss function are thus evaluated for a specific θ by iterating the recursions eqn.(15) to find the matrix $\partial\varepsilon(t,\theta)/\partial\theta$ for t=1,...,N and substituting them in equations (12) and (13).

The minimization of the loss function $J_2(\theta)$ can be done very efficiently by using Newton's method with a line search at every step. The method always converges to at least a local minimum. The algorithm consists of the following steps:

(i) Select an initial value of the parameter vector θ_0 using least squares. Set k=0.

(ii) Evaluate $\partial J_2(\theta)/\partial\theta$ and $\partial^2 J_2(\theta)/\partial\theta^2$ at θ_k.

(iii) Calculate the direction vector $d_k = -(\partial^2 J_2/\partial\theta^2)^{-1}(\partial J_2/\partial\theta)$.

(iv) Find the scalar a_k for which $J_2(\theta_k + a_k d_k)$ becomes minimum.

(v) Set $\theta_{k+1}=\theta_k + a_k d_k$.

(vi) If $J_2(\theta_{k+1})-J_2(\theta_k)$ is smaller than some small number, stop, otherwise set k=k+1 and go to (ii).

Notice that the calculation of the direction vector d_k in step (iii) does not require the inversion of the Hessian since the inverted Hessian has only to be multiplied with the gradient vector. The problem is thus equivalent to one of solving a set of linear equations. The approximate Hessian is always symmetric positive definite matrix and thus special methods can be used that take advantage of this property. The square root decomposition methods provide an efficient solution to the problem.

The Hessian can be factorized as

$$\frac{\partial^2 J_2}{\partial \theta^2} = U^T U \tag{16}$$

where U is an upper triangular square matrix and U^T is a lower triangular matrix. Once the Hessian has been decomposed, the direction vector d_k can be easily found. Let the gradient,vector be called g_k. It is then

$$U^T(Ud_k) = - g_k \tag{17}$$

The vector Ud_k can be calculated easily since U^T is a triangular matrix and the back substitution algorithm |Bierman (1977)| can be used to compute the elements of Ud_k iteratively. When the vector Ud_k has been computed, back substitution can again be used to give the vector d_k.

When the model is over-parametrized, the Hessian tends to become almost singular and the square root of the Hessian is very difficult to calculate numerically. In this case, a diagonal matrix μI, where μ is a small scalar, is added to the Hessian before the square root factorization is attempted. Thus

$$\frac{\partial^2 J_2}{\partial \theta^2} + \mu 1 = U^T U \tag{18}$$

This alters the direction vector d_k only slightly, which is unimportant, since d_k is only used as a direction along which a line search for the minimum can be done.

4. Model selection methods

One of the major problems in system identification is the selection of the model that can be used to identify the system S. If the model is not general enough, however well the parameters are estimated, the final model cannot behave like the true system S. On the other hand a very complicated model might be much more complex than the one that is actually needed. The theory of model selection can easily be applied to select the most appropriate model |Billings and Voon 1986b|. One is a model reduction method and the other a

model expansion method. For both methods the most complicated model to be considered must be chosen first. This model is called the full model. All the other models to be considered are special cases of the full model with some of the parameters of the full model equal to zero or some other constant value. When the non-linear input-output model eqn.(7) is used, the maximum values of the observability indices must be selected first. The parametric expansion of the functions q_i has to be decided upon next. In the case of a polynomial expansion, the highest degree of the approximating polynomials must be chosen.

4.1 The model reduction method

The estimation of the parameter of the full model and the Hessian of the loss function at the minimum are calculated first using the methods in section 3. If the full model is heavily over-parametrized, the square root decomposition of the Hessian might need the addition of a small diagonal matrix, μI, in order to be numerically evaluated. The comparison between the full model and any of the reduced models can be done by evaluating the log determinant ratio test statistic |Leontaritis and Billings 1986| d(y) and comparing it with the critical value 4s, where s is the number of the reduced coefficients. The statistic d(y) is given by

$$d(y) = (b^* - \hat{b}_1)^T \left[H_{bb} - H_{ab}^T \quad H_{aa}^{-1} H_{ab} \right] (b^* - \hat{b}_1) \tag{19}$$

where H is the Hessian of $NJ_2(\theta)$ at the minimum $\hat{\theta}$, H has been partitioned as

$$H = \begin{bmatrix} H_{aa} & H_{ab} \\ H_{ab}^T & H_{bb} \end{bmatrix} \tag{20}$$

to correspond to the partitioned parameter vector $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$ and b is a column vector of dimension s and a is a column vector of dimension n_θ-s. Thus $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$ corresponds to the full model and θ with b set to some specific vector b*, usually a zero vector, corresponds to the reduced model. The purpose of the test is to investigate if there is significant statistical evidence that the more complicated full model gives a better explanation of

the data than the simpler reduced model where $\hat{b}_1$ represents the estimate of the assumed true value b*. The numerical evaluation of the statistic d(y) can be done more efficiently using the Householder orthogonal transformation than using equation (19), for the following reason. The Hessian of the full model H has already been decomposed as U^TU where U is the upper triangular square root matrix of H. The Hessian of the reduced model H_{aa} must also be decomposed as $U_a^TU_a$ if equation (19) is to be used, so that the inverse of the matrix H_{aa} does not need to be calculated. If the matrix H_{aa} is the upper left partition of the matrix H, the square root U_a is also the upper left partition of the matrix U. The problem is that the Hessian of the reduced model H_{aa} is the upper left partition of the matrix H only if the parameters to be eliminated are the last ones in the parameter vector. Since this is not generally the case, the decomposition of the full matrix H cannot be used to find the decomposition of the reduced matrix H_{aa}. The Householder orthogonal transformation may be applied though since it can make full use of the already calculated square root matrix U.

A short introduction to the orthogonal transformation and in particular to the Householder orthogonal transformation follows.

A square matrix T is orthogonal if $T^TT=I$ (I is the identity matrix). Orthogonal matrices play a very important role when square root decomposition of matrices are used. This can be explained by the properties of the orthogonal matrices.

(i) If T_1 and T_2 are orthogonal matrices, then so is T_1T_2.

(ii) If U is a square root of the matrix H so is TU, where T is some orthogonal matrix.

This property can be exploited to create a square root of a matrix that has desirable properties, when some other square root is given. The usually desirable property is triangularity.

(iii) For any vector y

$$||Ty|| = ||y|| \tag{21}$$

where

$$||y|| = (y^{\tau}y)^{1/2}$$

So an orthogonal matrix transforms a vector in a way that preserves distance. It also preserves the inner product of the two vectors so it also preserves angles.

The Householder orthogonal transformation is the transformation that corresponds to the geometric notion of reflection on a plane perpendicular to a vector u. The orthogonal matrix of the Householder transformation is

$$T_u = 1 - \frac{2}{u^{\tau}u} uu^{\tau} \tag{22}$$

where I is the identity matrix. The transformed vector $T_u y$ is given by

$$T_u y = y - 2 \frac{y^{\tau}u}{u^{\tau}u} u \tag{23}$$

The transformed vector $T_u y$ can then be evaluated easily if the vector u that defines T_u is known and the actual transformation matrix T_u is not needed at all.

The Householder transformation can be used to triangulize a matrix. Initially a specific Householder transformation that transforms a vector y into a vector that has all components zero except the first one is developed. Let

$$T_u y = (-\sigma, 0, \ldots, 0)^{\tau} = -\sigma e_1 \tag{24}$$

where e_1 is the vector $(1,0,\ldots,0)^{\tau}$. Property (iii) of the orthogonal matrices gives

$$||T_u y|| = (y^{\tau}y)^{1/2} = |\sigma| \tag{25}$$

The direction of u is suggested by eqn.(23)

$$u = \text{const}(y + \sigma e_1) \tag{26}$$

The constant in eqn.(26) can be any scalar and it is taken equal to 1. The sign of σ is also arbitrary. For the choice

$$\sigma = \text{sign}(y_1)(y^{\tau}y)^{1/2} \tag{27}$$

the elements of the vector u are

$$u_1 = y_1 + \sigma$$

$$u_i = y_i \qquad \text{for } i > 1 \tag{28}$$

and

$$2/u^\tau u = 1/(\sigma u_1) \tag{29}$$

The sign of σ was chosen to maximize $|u_1|$ so that the term $1/(\sigma u_1)$ needed in the transformation of other vectors is as numerically well defined as possible. The transformation of other vectors is given by eqn.(23).

The following procedure can be used to triangulize a matrix. The Householder transformation that transforms the first column of the matrix to a vector with all elements equal to zero except the first one is applied to all the columns of the matrix. The lower right partition of the transformed matrix without the first column and row is transformed again so that the first column of the partitioned matrix has all elements zero except for the first one. Continuing in this way a triangular matrix is eventually created if the original matrix is square. All the partial orthogonal transformation matrices are not actually calculated and are not needed. If the original matrix is not square the final one has all the elements below the 45 degree diagonal that starts at the top left corner of the matrix, equal to zero. The Householder triangulizaion of a matrix can now be employed to calculate the statistic d(y) efficiently.

Let the Hessian H at the minimum point of the loss function $NJ_2(\theta)$ be decomposed as $U^\tau U$. The loss function around the unrestricted minimum θ_1 then is

$$NJ_2(\theta) = NJ_2(\hat{\theta}_1) + \frac{1}{2}(\theta - \hat{\theta}_1)^\tau U^\tau U(\theta - \hat{\theta}_1)$$

$$= NJ_2(\hat{\theta}_1) + \frac{1}{2}(U\theta - U\hat{\theta}_1)^\tau (U\theta - U\hat{\theta}_1) \tag{30}$$

The matrix $U\theta$ can be written as

$$U\theta = \bar{U}_a a + \bar{U}_b b \tag{31}$$

where the matrix $\bar{U}_a$ is the $n_\theta x(n_\theta - s)$ matrix which consists of the columns of the matrix U that correspond to the parameters of the vector a, and $\bar{U}_b$ is the $n_\theta xs$ matrix which consists of the columns of U that correspond to the vector b. For the vector b restricted to being equal to b*, the loss function of the

reduced model becomes

$$NJ_2(\theta)=NJ_2(\hat{\theta}_1)+\frac{1}{2}(\bar{U}_a a+\bar{U}_b b^*-U\hat{\theta}_1)^\tau(\bar{U}_a a+\bar{U}_b b-U\hat{\theta}_1) \tag{32}$$

Let the constant vector c be

$$c = U\hat{\theta}_1 - U_b b^* \tag{33}$$

Usually, since $b^* = 0$, the vector c is $c = U\hat{\theta}_1$. The loss function becomes

$$NJ_2(\theta) = NJ_2(\hat{\theta}_1) + \frac{1}{2}(\bar{U}_a a - c)^\tau(\bar{U}_a a - c) \tag{34}$$

The function eqn.(34) must be minimized to find the restricted minimum $\hat{\theta}_o$. The matrix $\bar{U}_a$ is no longer triangular since it is actually the triangular matrix U with several columns removed. If these columns are not the last ones, the matrix $\bar{U}_a$ is not triangular. The orthogonal matrix T that triangularizes $\bar{U}_a$ can be found using the Householder transformation. Since $T^\tau T = I$, the function eqn. (34) becomes

$$\begin{aligned} NJ_2(\theta) &= NJ_2(\hat{\theta}_1) + \frac{1}{2}(\bar{U}_a a - c)^\tau T^\tau T(\bar{U}_a a - c) \\ &= NJ_2(\hat{\theta}_1) + \frac{1}{2}(T\bar{U}_a a - Tc)^\tau(T\bar{U}_a a - Tc) \end{aligned} \tag{35}$$

The matrix $T\bar{U}_a$ is an $n_\theta x(n_\theta - s)$ matrix where the top square matrix is triangular and the bottom s rows are zero. Let the $(n_\theta - s)x(n_\theta - s)$ top square triangular matrix be called U_a. Also let the top $(n_\theta - s)$ elements of the vector Tc be the vector z and the bottom z elements, the vector e. Then

$$T\bar{U}_a = \begin{bmatrix} U_a \\ 0 \end{bmatrix} \quad \text{and} \quad Tc = \begin{bmatrix} z \\ e \end{bmatrix} \tag{36}$$

The loss function becomes

$$NJ_2(\theta) = NJ_2(\hat{\theta}_1) + \frac{1}{2}(U_a a - z)^\tau(U_a a - z) + \frac{1}{2}e^\tau e \tag{37}$$

The value of a that minimizes $NJ_2(\theta)$ is $\hat{a}_o$ and it is obviously given by the solution of the equation

$$U_a \hat{a}_o = z \tag{38}$$

and the value of the statistic d(y) is given by

$$d(y) = 2NJ_2(\hat{\theta}_o) - 2NJ_2(\hat{\theta}_1) = e^\tau e \tag{39}$$

This approach provides not only the statistic d(y) but the square root of the Hessian of the reduced model U_a and a simple back-substitution in eqn.(38)

provides the parameters of the reduced model $\hat{a}_o$. The actual matrix T does not need to be calculated at all since every partial transformation used to triangularize $\bar{U}_a$ can also be applied to the vector c, the only other vector needed to be transformed.

It has been mentioned that the Hessian H might be nearly singular that, in order to calculate the square root U, a small diagonal matrix μI might need to be added to H. Although this does not affect the estimation of the parameters of the full model it does affect the evaluation of the statistic d(y). A more accurate expression for the statistic d(y) can be given. It is based on the fact that the evaluation of the parameters of the reduced model given by eqn. (38) is not affected as much as the statistic d(y) by the addition of the diagonal matrix μI given in eqn. (39). It is

$$U^T U = H + \mu I \tag{40}$$

Thus a more accurate expression for the statistic d(y) is

$$d(y) = (\hat{\theta}_o - \hat{\theta}_1)^T \; H \; (\hat{\theta}_o - \hat{\theta}_1) \tag{41}$$

where as usual $\hat{\theta}_o = [\hat{a}_o^T, b^{*T}]^T$ and the original Hessian is used. Another expression that does not need the original Hessian and is quicker to evaluate can be derived from eqns. (40) and (41). It is

$$d(y) = e^T e - \mu(\hat{\theta}_o - \hat{\theta}_1)^T (\hat{\theta}_o - \hat{\theta}_1) \tag{42}$$

Both eqns (41) and (42) require the solution of equation (33) to provide the reduced model parameters. The above method for evaluating the statistic d(y) was found to be quick and numerically very robust.

The selection of one model from several competing models can be determined by computing the criterion C that is derived from the likelihood ratio test |Leontaritis and Billings, 1986|. Let two models have parameter vectors θ_1 and θ_2 with dimensions n_{θ_1} and n_{θ_2}. Assume that $n_{\theta_1} < n_{\theta_2}$ and $s = n_{\theta_2} - n_{\theta_1}$. The model with parameter vector θ_1 is selected according to the likelihood ratio test if

$$2L(\theta_1) - 2L(\theta_2) < k(s) = sk(1) = (n_{\theta_2} - n_{\theta_1})\, k(1)$$
$$= n_{\theta_2} k(1) - n_{\theta_1} k(1) \qquad (43)$$

or if

$$2L(\theta_1) + n_{\theta_1} k(1) < 2L(\theta_2) + n_{\theta_2} k(1) \qquad (44)$$

The model that is selected amongst all the several competing models is the one that minimises the criterion

$$C = 2L(\theta) + n_\theta k(1)$$
$$= N\log\det Q(\theta) + n_\theta k(1) \qquad (45)$$

where θ is the parameter vector and n_θ is its dimension.

If k(1) is set equal to 2 the above criterion becomes equal to Akaike's Information Criterion (AIC)

$$AIC = 2L(\theta) + 2n_\theta \qquad (46)$$

It is well known that the AIC criterion may overestimate the true parameter vector but it has recently been shown that the use of the criterion C with k(1) = 4 reduces the probability of selecting a model with one more parameter than the true model to an insignificant level [Leontaritis and Billings 1986].

A full model with even a relatively small number of parameters can generate a prohibitively large number of reduced models. It is thus of great importance to have the full model with as few parameters as possible. One way of achieving such a reduction is to use a full model with the correct observability indices. A practical way of estimating the observability indices is to assume that the system is linear, fit a full linear model and find the best reduced one. The observability indices of the best linear model can thus be found. It can be argued that the observability indices of the best linear model should not be different from the observability indices of the best non-linear model, for mild non-linearities and for long data sets. This method also works well in practice so it can be quite safely used. It is very probable that the full non-linear model with the correct observability indices will still have a large number of parameters. The only feasible solution in that case is to apply the Stepwise Backward Elimination (SBE) method or the Stepwise Forward Inclusion

(SFI) method |Draper and Smith 1981| or to use a combined prediction error stepwise regression algorithm |Billings and Voon 1986b|. The finally selected model by these methods is not always the best one but it is greatly reduced compared to the full model. Both methods should be employed so that the final models they select can be compared and the best one of the two chosen. Simulation has shown that the models the SBE and the SFI methods select are, if not the correct ones, very near to the correct ones.

4.2 The model expansion method

The basic difference between the model expansion method and the model reduction method is that in the former, the parameters for the full model are not estimated at all.

A basic model is first chosen so that it contains parameters known to belong to the final model. If no such parameters are known, the model with no parameters is the basic model. Another obvious choice is the best linear model. The parameters of the basic model are estimated and the Hessian of the loss function at the minimum is calculated. If the basic model has no parameters obviously no estimation needs to be done. The basic model is too simple to explain the data and should be expanded. All the expanded models with just one more parameter are considered. The statistic d(y) is evaluated for every one of the expanded models and the model that gives the statistic d(y) its maximum value is selected as the best expanded model with one more parameter. The parameters of the expanded model have now to be estimated using the original data y. Newton's method can be used to minimize the loss function of this expanded model. The Hessian of the loss function at the minimum should also be calculated. This expanded model can now be expanded again by adding just one more parameter. The statistic d(y) for all such models is calculated and the one with the maximum value of the statistic d(y) is chosen and its parameters estimated using the original data. The expansion process continues until the maximum value of the statistic d(y) is found to be less than the critical value k(1) (chosen to be equal to 4) for all the expanded models.

Then, since none of the expanded models can explain the data significantly better than the non-expanded one, there are no more parameters that can be included.

The statistic d(y) can be shown to be given by

$$d(y) = \left[\frac{\partial NJ_2}{\partial b}\right]_{\hat{\theta}_o} \left[H_{bb} - H_{ab}^{\tau} H_{aa}^{-1} H_{ab}\right]^{-1} \left[\frac{\partial NJ_2}{\partial b}\right]_{\hat{\theta}_c}^{\tau} \tag{47}$$

where H_{aa} is the Hessian at the minimum $\hat{\theta}_o$ of the non-expanded model, b is the extra parameter of the expanded model and

$$H_{ab} = \left[\frac{\partial^2 NJ_2}{\partial a \partial b}\right]_{\hat{\theta}_o} \tag{48}$$

$$H_{bb} = \left[\frac{\partial^2 NJ_2}{\partial b^2}\right]_{\hat{\theta}_o} \tag{49}$$

The model expansion is structurally the same as the Stepwise Forward Inclusion (SFI) method of the general model reduction method. The difference between the two lies is the way the statistic d(y) is evaluated.

The numerical evaluation of the statistic d(y) in eqn.(47) can be done very efficiently using the already calculated decomposition of the Hessian $H_{aa} = U_a^{\tau} U_a$. The vector $H_{aa}^{-1} H_{ab}$ can be evaluated, as it has been done before, using the back substitution procedure for solving a triangular set of equations twice. The rest of the calculations are trivial. The Householder orthogonal transformation could also be employed here but without any advantage. This is because the extra parameter of the extended model is added at the bottom of the parameter vector and the decomposition of H_{aa} can be used without any problem. The evaluation of d(y) in eqn.(47) requires the derivatives $[\partial NJ_2/\partial b]_{\hat{\theta}_o}$, $[\partial^2 NJ_2/\partial a \partial b]_{\hat{\theta}_o}$ and $[\partial^2 NJ_2/\partial b^2]_{\hat{\theta}_o}$. These should be evaluated using the original data and equations (12), (13) and (15). The evaluation of these derivatives is the most time consuming part in the calculation of the statistic d(y).

The model expansion method has advantages and disadvantages compared with the model reduction method. The advantages are:

(i) There is no need to estimate the parameters of the full model. The full model is likely to have a large number of parameters and the minimization of a loss function with so many variables might be very time consuming. The full model is also over-parametrized, so the loss function may have almost singular Hessian. The minimization of such a loss function has to be done numerically with great care. However the use of Newton's method coupled with square root decomposition of the Hessian and line search at every step has proved that it can estimate the parameters of highly complicated and over-parametrized full models quite successfully. However it can still be a rather time consuming task.

(ii) The Hessian of the models considered in the model expansion method are never singular and thus numerically well conditioned. This happens because every parameter that is included in the model contributes significantly to the loss function and thus cannot cause singularity of the Hessian.

The disadvantages of the model expansion method are:

(i) The calculation of the statistic d(y) for every candidate extended model is quite time consuming since derivatives of the loss function using the original data have to be calculated.

(ii) Every time the model is extended with one extra parameter, the loss function has to be minimized. This minimization is however numerically well defined since the Hessian of the loss function is never singular.

(iii) The model expansion method is similar to the stepwise forward inclusion method of model reduction. The stepwise backward elimination and the optimal combinatorial methods in the model reduction cannot be easily extended to the case of model expansion.

The model expansion method is preferable in practice if a small number of parameters need to be added to the basic model and the full model is actually very complicated. In such a case the model reduction method is too wasteful.

5. Model validation methods

Model validation can be achieved using either parametric or non-parametric methods, |Billings and Voon 1983, 1986a, Leontaritis and Billings 1986|. The parametric validation method tests if some extension of the chosen model explains the data significantly better. Comparison between the final model and models which are reductions of the full model have already been done in the model selection part of the identification. The final model should thus be validated against models which are extensions of the full model. The parametric validation method can thus be regarded as a variation of the model expansion method with an expanded full model. Parametric validation has the advantage that it has maximum power and it does not need the estimation of the parameters of the expanded model. Since parametric validation is actually the same as the model expansion method nothing more needs to be added.

The non-parametric validation methods are the correlation techniques described in detail in Leontaritis and Billings |1986|. Chi-square correlation tests between residuals and monomials of past inputs, residuals and outputs have to be performed to validate a non-linear model.

The number of correlation tests needed to be evaluated can be extensive since a wide variety of monomials of past inputs, residuals and outputs should be used. The fact that a non-linear model satisfied the correlation tests for a few monomials does not give much confidence that the chosen model is correct and a wide range of monomials should be used. Alternatively, the simple correlation tests developed by Billings and Voon |1983, 1986a|, can be applied to alleviate this difficulty. Cross-validation methods are correlation methods with the difference that the chosen model is tested using a completely different set of data to the one used for parameter estimation and model selection. If a cross-validation test fails for a particular model which passed all the normal validation tests, something very wrong has been done in the identification process. Cross-validation is the final and ultimate test that confirms that a

model has been correctly created.

A situation where a different set of data can detect a mistake in the selection of the correct model is the following.

Suppose that a completely deterministic model is used to fit some input-output data that are actually created by a stochastic system. The output of the deterministic model will be closer to the real output than the output of the deterministic part of the true stochastic system for the set of input-output data used for the estimation of the parameters. Thus the closeness of fit of the output of the deterministic part of the model to the actual output is not a good measure of the correctness of a model. If however such a wrong model is chosen, the mistake becomes immediately apparent if the same comparison is done for a set of data different from the one used for the estimation of the parameters.

6. Simulation Results

This simulation was carried out to demonstrate the application of the prediction error method in closed loop operation. In order to show the robustness of the prediction error method, the system to be identified is an unstable non-linear system. The system is called S_2 and it is given by

$$y(t) = 1.2y(t-1) + 0.2u(t-1) - 0.8e(t-1) + 0.1y^3(t-1) - 0.05y(t-1)u^2(t-1) - 0.2y(t-1)u(t-1)e(t-1) + e(t) \quad (50)$$

where $u(t)$ is the input, $y(t)$ is the output and $e(t)$ is a Gaussian white sequence of standard deviation equal to 0.05. The system S_2 is unstable and it can only be operated in closed loop. The feedback law used is given by

$$u(t) = w(t) - 2.0y(t) \quad (51)$$

where $w(t)$ is a set point disturbance signal so that the identification can be done efficiently. The disturbance signal was chosen to be an independent Gaussian sequence of standard deviation equal to 1.15. An input-output data sequence of 500 points was generated and used for the identification of system

S_2. The first 100 points are given in figure 1. The model used first to identify the system S_2 is the correct one with 6 parameters. The estimates of the parameters and their standard deviations are

1	y(t-1)**1 =	0.1195E+01 (±0.3841E-02)
2	u(t-1)**1 =	0.1989E+00 (±0.1230E-02)
3	e(t-1)**1 =	-0.7452E+00 (±0.3865E-01)
4	y(t-1)**3 =	0.1054E+00 (±0.4812E-02)
5	y(t-1)**1*u(t-1)**2 =	-0.5045E-01 (±0.9678E-03)
6	y(t-1)**1*u(t-1)**1*e(t-1)**1 =	-0.1840E+00 (±0.3000E-01)

The estimates of the parameters are not biased as expected. The output of the model and the residuals are given in figure 1.

The linear part of the system with only 3 parameters was used next as a model to identify the system S_2. The estimates of the parameters and their standard deviations are

1	y(t-1)**1 =	0.1161E+01 (±0.7740E-02)
2	u(t-1)**1 =	0.2326E+00 (±0.3022E-02)
3	e(t-1)**1 =	-0.3756E+00 (±0.4234E-01)

The estimates are biased as expected. The output of this model for the first 100 points is given in figure 2. The correlation tests detect that this model is not the correct one. They are given in figure 3. It can be noted that here the input sequence u(t) is not white because of the effect of the feedback.

An over-parametrized model with 13 parameters was also used to identify the system S_2. The output and the residuals of this model for the first 100 points are given in figure 4. The estimates of the parameters and their standard deviations are

1	y(t-1)**1 =	0.1194E+01 (±0.5645E-02)
2	u(t-1)**1 =	0.1978E+00 (±0.2127E-02)
3	e(t-1)**1 =	-0.7885E+00 (±0.5812E-01)

4	y(t-1)**3 =	0.1071E+00 (±0.7913E-02)
5	y(t-1)**2*u(t-1)**1 =	0.2974E-02 (±0.6782E-02)
6	y(t-1)**2*e(t-1)**1 =	-0.9730E-01 (±0.1364E+00)
7	y(t-1)**1*u(t-1)**2 =	-0.4870E-01 (±0.2135E-02)
8	y(t-1)**1*u(t-1)**1*e(t-1)**1 =	-0.2336E+00 (±0.6984E-01)
9	y(t-1)**1*e(t-1)**2 =	-0.3318E+00 (±0.1258E.01)
10	u(t-1)**3 =	0.2135E-03 (±0.3004E-03)
11	u(t-1)**2*e(t-1)**1 =	-0.8173E-02 (±0.1197E-01)
12	u(t-1)**1*e(t-1)**2 =	0.1221E+00 (±0.3624E+00)
13	e(t-1)**3 =	0.8456E+01 (±0.5695E+01)

The over-parametrized model can now be reduced. The SBE process is used first. The value of the AIC criterion and the criterion C for k(1) = 4 for the reduction of every term is

Total Number of eliminated parameters	No of eliminated parameter	AIC of reduced model - AIC of full model	C of reduced model - C of full model	Standard Deviation of the residuals
1	9	-0.1930E+01	-0.3930E+01	0.4913E-01
2	5	-0.3718E+01	-0.7718E+01	0.4914E-01
3	11	-0.5331E+01	-0.1133E+02	0.4916E-01
4	6	-0.6910E+01	-0.1491E+02	0.4918E-01
5	10	-0.8593E+01	-0.1859E+02	0.4920E-01
6	12	-0.9449E+01	-0.2150E+02	0.4925E-01
7	13	-0.8035E+01	-0.2204E+02	0.4942E-01
8	8	0.4014E+02	0.2414E+02	0.5197E-01

Akaike's criterion over-estimates the number of required parameters and it keeps term No 13 while the criterion C finds the correct model. The SFI process is also used to reduce the over-parametrized model. The value of the AIC criterion and of criterion C for k(1) = 4, for the inclusion of every term is

Total Number of included parameters	No of included parameter	AIC of reduced model - AIC of full model	C of reduced model - C of full model	Standard Deviation of the residuals
1	1	0.6958E+05	0.6956E+05	0.1000E+20
2	2	0.3468E+04	0.3446E+04	0.1621E+01
3	7	0.1020E+04	0.9999E+03	0.1393E+00
4	4	0.4225E+03	0.4054E+03	0.7639E-01
5	3	0.4014E+02	0.2414E+02	0.5197E-01
6	8	-0.8035E+01	-0.2204E+02	0.4942E-01
7	13	-0.9449E+01	-0.2150E+02	0.4925E-01
8	12	-0.8593E+01	-0.1859E+02	0.4920E-01

The SFI process gives the same results as the SBE process.

7. Conclusions

The numerical implementation of a prediction error estimation algorithm and associated model selection techniques have been preserved for nonlinear systems which can be represented by a NARMAX model. The square root method of decomposing a positive definite symmetric matrix was used to ensure an efficient numerical minimisation of the loss function. Two basic methods of selecting the correct model were discussed, the model expansion and the model reduction methods, and the Householder orthogonal transformation was employed to evaluate the criterion the selected model must minimise.

8. Acknowledgements

One of the authors (SAB) gratefully acknowledges that this work was supported by SERC.

References

Billings S.A., Voon W.S.F. (1983) Structure detection and model validity tests in the identification of nonlinear systems: Proc IEE, Part D, 130, 193-199.

Billings S.A., Voon W.S.F. (1986a) Correlation based model validity tests for nonlinear models: Int.J.Control (to appear).

Billings S.A., Voon W.S.F. (1986b): A prediction error and stepwise regression estimation algorithm for nonlinear systems: Int.J.Control (to appear)

Bierman G.J. (1977): Factorization methods for discrete sequential estimation: Academic Press, New York.

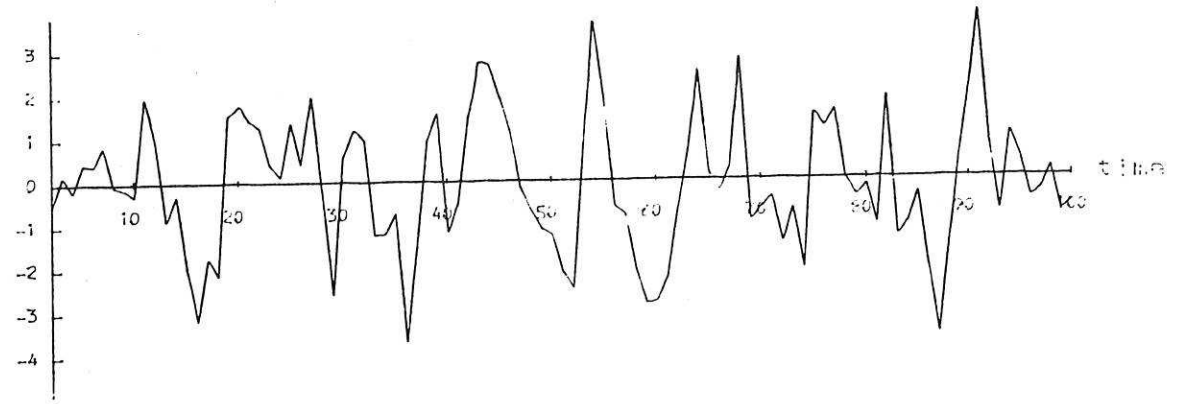
Draper N.R., Smith L. (1981). Applied regression analysis: Witey.

Goodwin G.C, Payne R.L. (1977) Dynamic System Identification; Experiment design and data analysis: Academic Press.

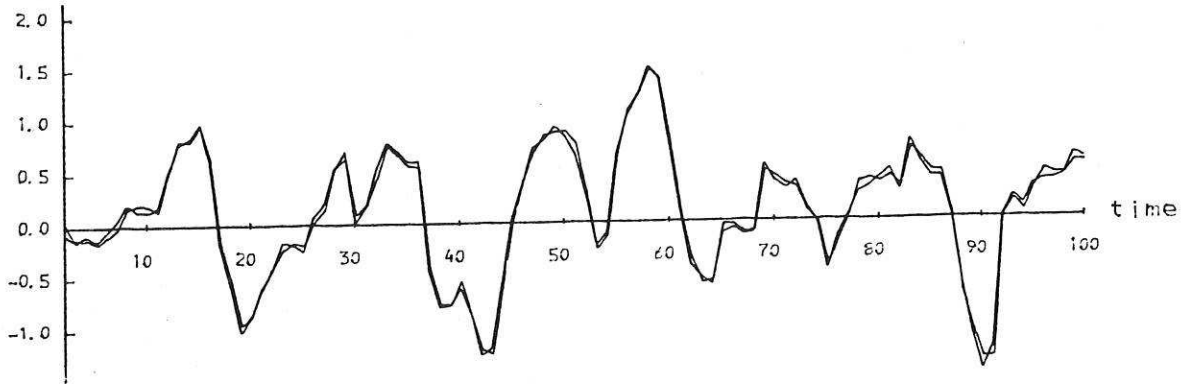
Leontaritis I.J., Billings S.A. (1985): Input-output parametric models for nonlinear systems, Part I Deterministic nonlinear systems, Part II Stochastic nonlinear systems, Int.J. Control, 41, 303-344.

Leontaritis I.J., Billings S.A. (1986): Model Selection and Validation methods for nonlinear systems: Research Report 292, Department of Control Engineering, University of Sheffield (submitted for publication).

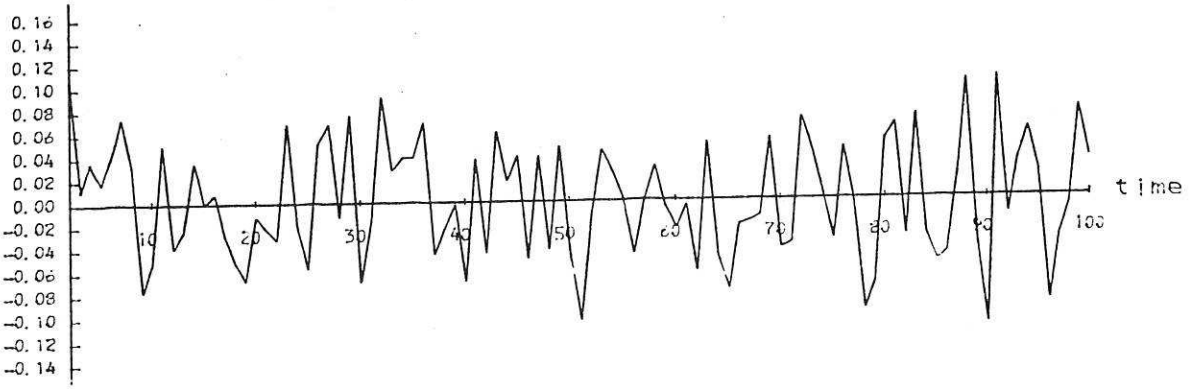
Ljung L., Soderstrom T. (1983): Theory and Practice of Recursive Identification; MIT Press.



Input

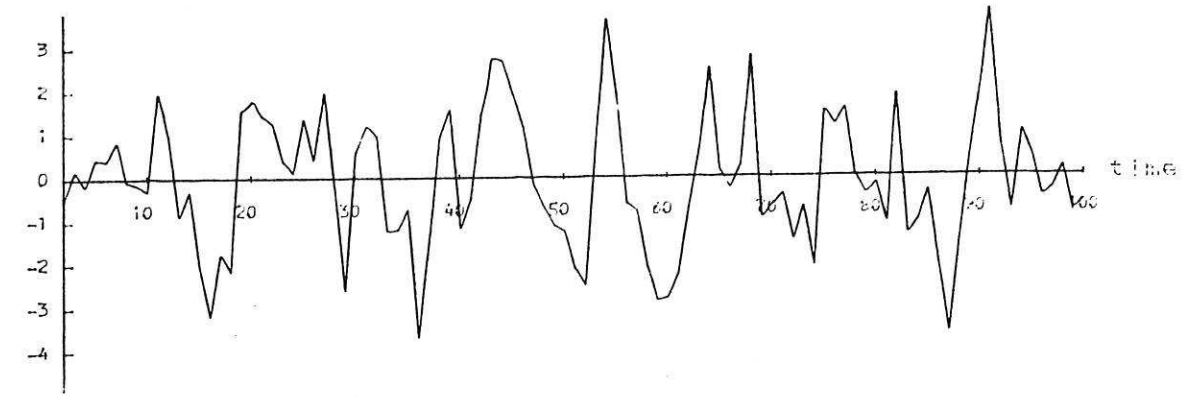


—— Output of the system
—— Output of the model

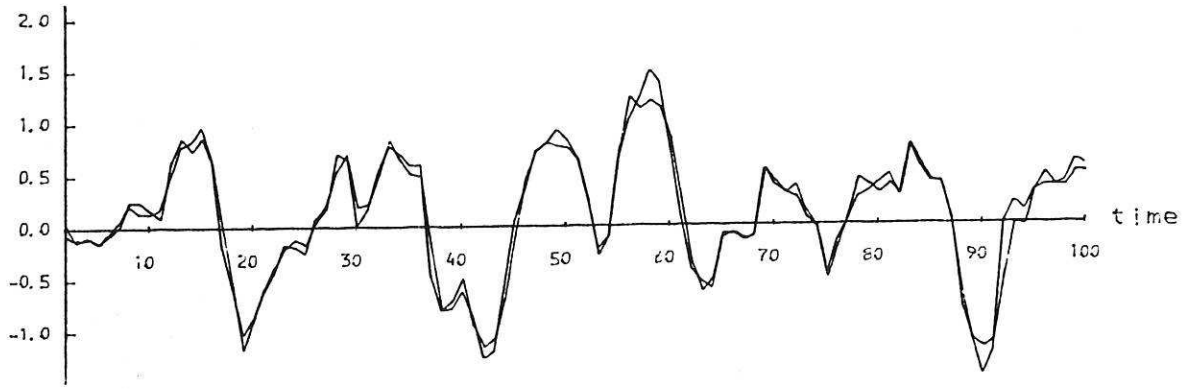


Residuals

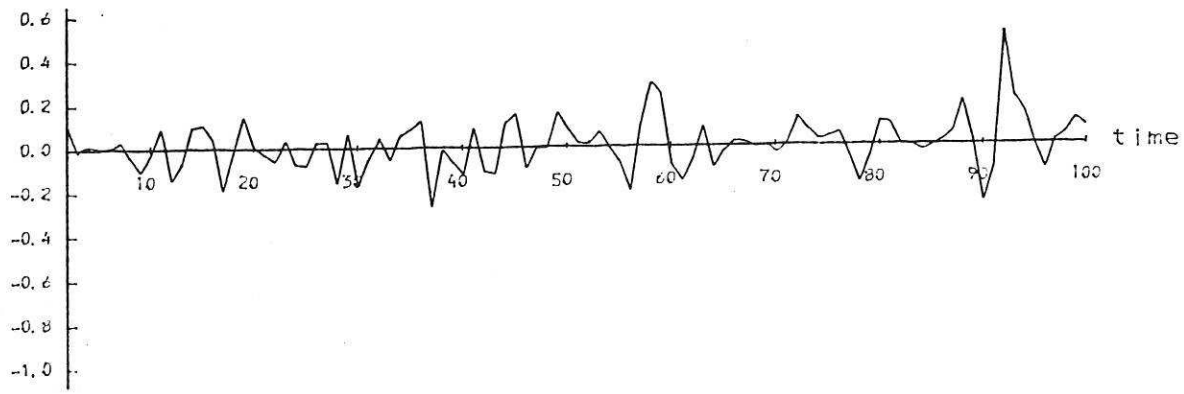
Fig.1. Nonlinear Identification of System S_2.



Input



—— Output of the system
—— Output of the model



Residuals

Fig.2. Linear Identification of S_2.

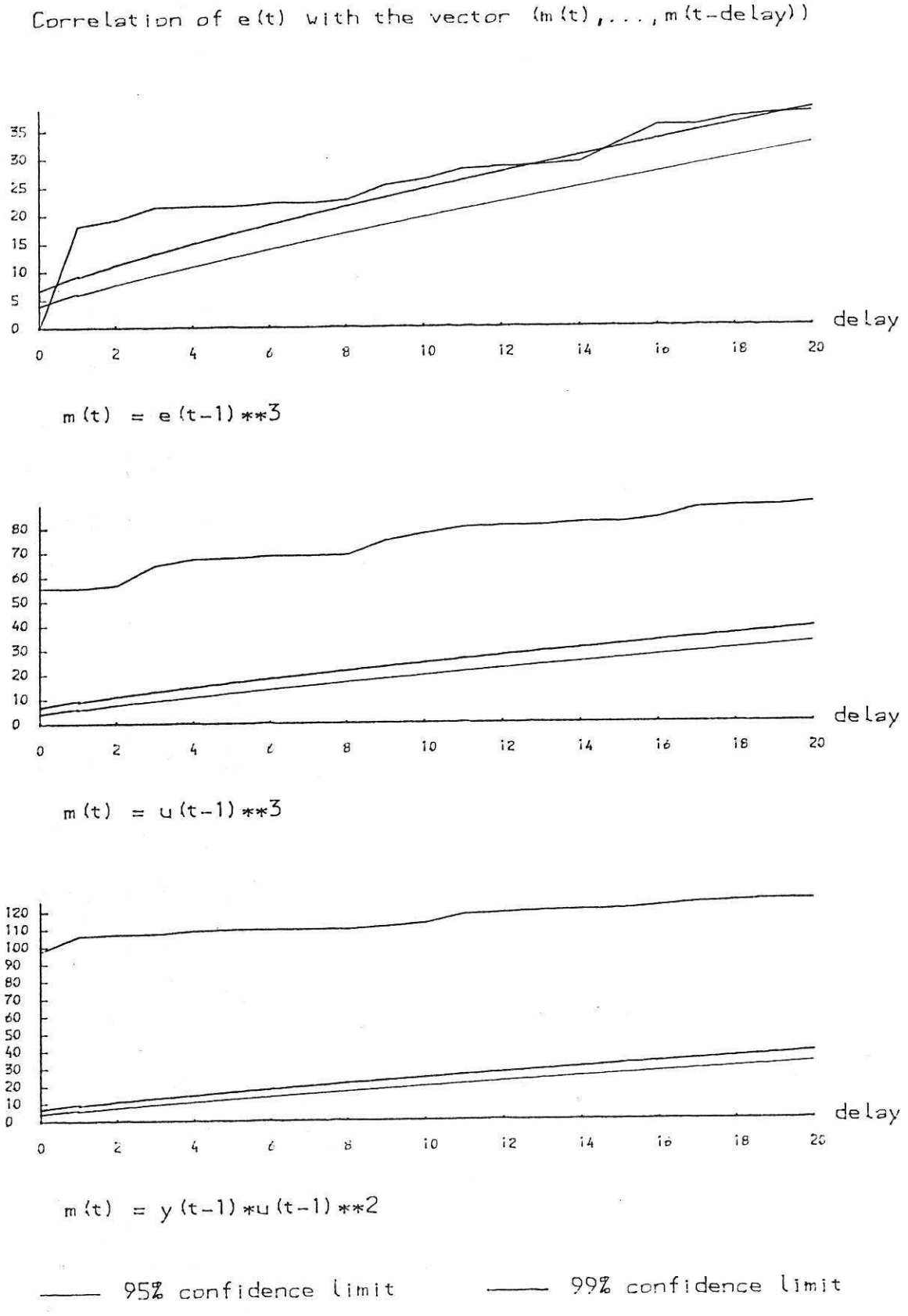
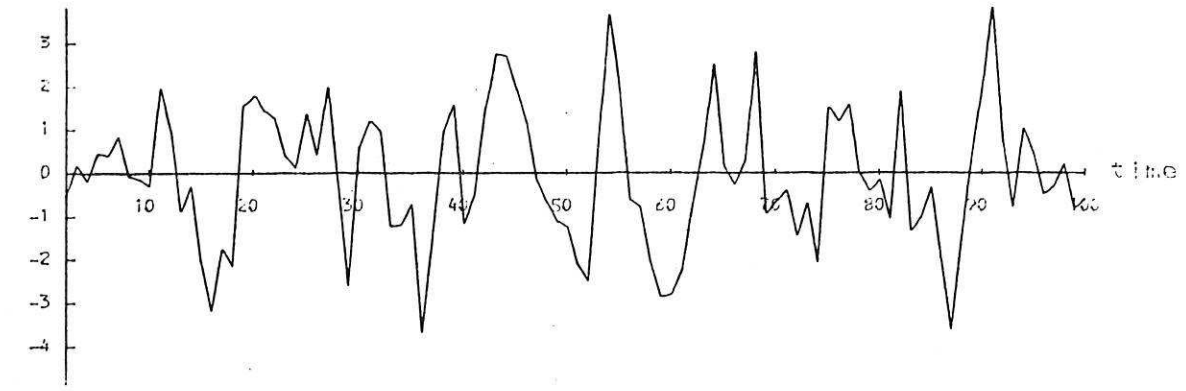
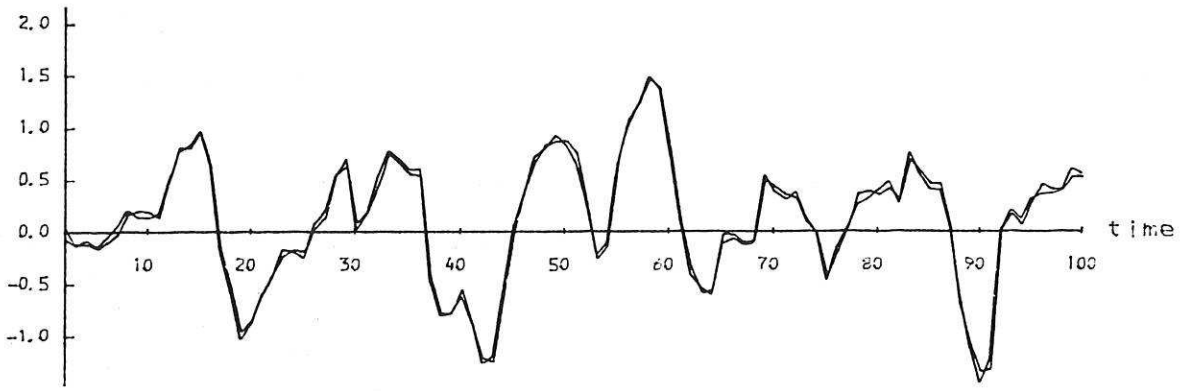


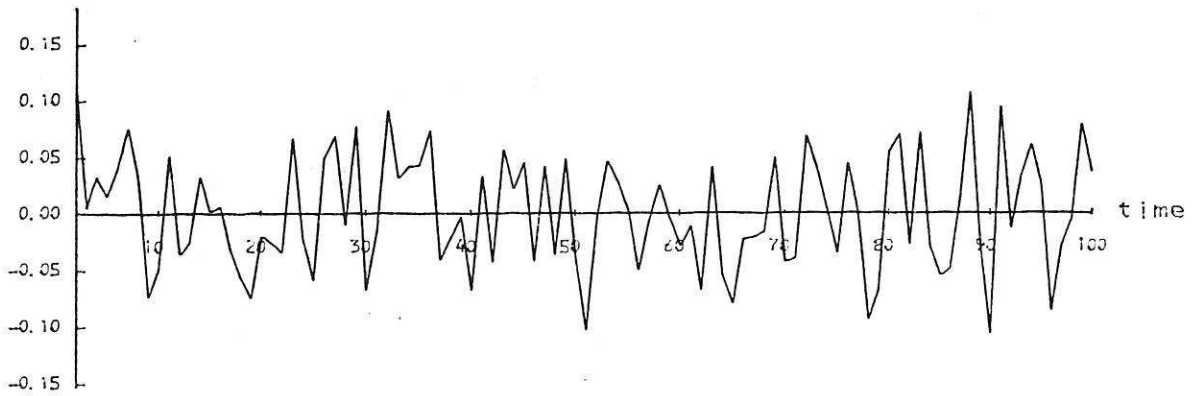
Fig.3. Chi-square Correlation Tests for the Linear Identification of S_2.



Input



—— Output of the system
—— Output of the model



Residuals

Fig.4, Over parameterized nonlinear identification of S_2.