A Prediction Error and Stepwise Regression Estimation Algorithm for Nonlinear Systems

S. A. Billings, PhD,BEng,CEng,MIEE,AFIMA,MInstMC

W. S. F. Voon, BEng,PhD

Department of Control Engineering
University of Sheffield
Mappin Street, Sheffield

September 1985

Research Report No. 281

Abstract

The identification of nonlinear systems based on a NARMAX (Nonlinear AutoRegressive Moving Average model with eXogenous inputs) model representation is considered and a combined stepwise regression/ prediction error estimation algorithm is derived. The stepwise regression routine determines the model structure by detecting significant terms in the model whilst the prediction error algorithm provides optimal estimates of the final model parameters. Implementation of the algorithms is discussed in detail and several simulated examples and industrial applications are included to illustrate that parsimonious models of nonlinear systems can be identified using the algorithm.

1. Introduction

The traditional representation for nonlinear systems of unknown structure has been the Volterra and Wiener functional series. Whilst these descriptions can represent a wide class of nonlinear systems they do so at the expense of introducing an excessive number of unknown coefficients which have to be estimated [Billings 1980, Marmarelis and Marmarelis 1978]. This not only complicates the identification of systems based on these descriptions but also makes the development of controller design procedures very complex. It is perhaps for these reasons that few practical systems have been identified using the Volterra or Wiener approach and that the design of controllers based on these models has received virtually no attention.

In contrast the rapid expansion of linear controller design techniques has been largely based on the difference equation or pulse transfer function model [Harris,Billings 1985]. The linear system impulse response, which represents the first term in a Volterra expansion, requires a large number of parameters to characterise a system compared with a difference equation model and consequently has seldom been used to formulate controller design algorithms. Both the linear impulse response model and the functional series of Volterra and Wiener map the past system input into the present output. This provides a very redundant system description since the dynamic information in the lagged system outputs is ignored. It is precisely this information which is included in the linear difference equation model which maps the past input and output into the current output and provides a parsimonious system description.

In a similar manner the nonlinear difference equation model known as the NARMAX model [Leontaritis and Billings 1985] (Nonlinear AutoRegressive Moving Average model with eXogenous input) can, by including information from both lagged inputs and outputs, provide a very concise representation for nonlinear systems. It is this model which in the present study is used as a basis for the development of a prediction error parameter estimation algorithm for nonlinear systems. The algorithm has been designed to include a stepwise regression routine which is used to detect significant terms in the model prior to final estimation. It is shown that implementation of the prediction error/stepwise regression algorithm produces parsimonious models, usually with between six and ten terms of complex nonlinear systems which are suitable as a basis for controller design studies [Billings,Tsang,Voon 1985].

2. The NARMAX Model

A nonlinear system can be represented by the nonlinear difference equation model [Leontaritis and Billings 1985]

$$y(t) = F^*[y(t-1),\ldots y(t-n_y),u(t-d),\ldots u(t-d-n_u+1)] \qquad (1)$$

where $F^*[\cdot]$ is some nonlinear function of $u(\cdot)$ and $y(\cdot)$ providing

(i) the state-space of the Neroole realization does not have infinite dimensions (i.e. we exclude distributed parameter systems), and

(ii) the linearized system around the origin has a Hankel matrix of maximum rank (i.e. a linearized model would exist if the system were operated close to an equilibrium point).

Equation (1) represents the single-input single-output case but the results have been extended to the multivariable case. The Hammerstein, Wiener, bilinear, Volterra and other nonlinear models can be shown to be special cases of eqn (1).

An equivalent representation for nonlinear stochastic systems can be derived [Leontaritis and Billings 1985] by considering input-output maps based on conditional probability density functions to yield the model

$$z(t) = F^{\ell}[z(t-1),\ldots z(t-n_z),u(t-d),\ldots u(t-d-n_u+1),\varepsilon(t-1)\ldots \ldots\varepsilon(t-n_\varepsilon)] + \varepsilon(t) \qquad (2)$$

where $F^{\ell}[\cdot]$ is a nonlinear map of degree ℓ nonlinearity, d is the time delay and $\varepsilon(t)$ the prediction error. This model is referred to as the Nonlinear AutoRegressive Moving Average model with eXogenous inputs or NARMAX model.

3. Maximum Likelihood Estimation

Consider initially the development of a maximum likelihood estimator [Goodwin and Payne 1977] assuming that the probability distribution of the input/output data is known.

Define the vectors

$$Z_t = [z(t),\ldots z(1)]^T$$
$$U_t = [u(t),\ldots u(1)]^T \qquad (3)$$

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

$$p(z(t) \mid Z_{t-1},U_t) \qquad (4)$$

The function eqn (4) can be put in its innovation form

$$z(t) = f(Z_{t-1},U_t) + \varepsilon(t) \qquad (5)$$

where $\varepsilon(t)$, the prediction error or innovation sequence is the stochastic

process defined as

$$\varepsilon(t) = z(t) - E\left[z(t) \mid Z_{t-1}, U_t\right] \tag{6}$$

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

$$\hat{z}(t) = E\left[z(t) \mid Z_{t-1}, U_t\right] = f(Z_{t-1}, U_t) \tag{7}$$

and thus the innovation form eqn (5) separates the output that can be predicted from the past as $f(Z_{t-1}, U_t)$ and the unpredictable part as the innovation $\varepsilon(t)$.

The likelihood function is then given by Bayes rule [Papoulis 1965]

$$L(\theta;\ Z_N, U_N) = p(Z_N \mid U_N, \theta) = \prod_{t=1}^{N} p(z(t) \mid Z_{t-1}, U_t; \theta) \tag{8}$$

From eqn (5) and the transformation of random variables rule the conditional distribution of $z(t)$ can be related to that of $\varepsilon(t)$ as

$$p(z(t) \mid Z_{t-1}, U_t; \theta) = p_{\varepsilon(t)}(\hat{\varepsilon}(t) \mid \theta) \left| \det\left(\frac{\partial \varepsilon(t)}{\partial z(t)}\right) \right| \tag{9}$$

where

$$\hat{\varepsilon}(t) = z(t) - f(Z_{t-1}, U_t) \tag{10}$$

the Jacobian

$$\left| \det\left[\frac{\partial \varepsilon(t)}{\partial z(t)}\right] \right| = 1 \tag{11}$$

and $p_{\varepsilon(t)}(\hat{\varepsilon}(t) \mid \theta)$ is the conditional probability density function of $\varepsilon(t)$ given Z_{t-1} and U_t.

Substituting eqn (9) into eqn (8) using the result of eqn (11) gives the likelihood function

$$L(\theta; Z_N, U_N) = \prod_{t=1}^{N} p_{\varepsilon(t)}\left[\hat{\varepsilon}(t) \mid \theta\right] \tag{12}$$

The maximum likelihood estimate is obtained by maximising the likelihood function eqn (12) with respect to θ the unknown parameters.

If $\hat{\varepsilon}(t)$ is an independent normally distributed sequence with common covariance R, then the likelihood function of eqn (12) can be expressed as

$$L(\theta; Z_N, U_N) = \prod_{t=1}^{N} \left[(2\pi)^m \det R\right]^{-\frac{1}{2}} \cdot \exp\{ -\tfrac{1}{2}\hat{\varepsilon}^T(t) R^{-1} \hat{\varepsilon}(t) \} \tag{13}$$

The negative log likelihood function per sample then is

$$\mathcal{L}(\theta; Z_N, U_N) = \frac{1}{N} \log_e L(\cdot) = \tfrac{1}{2} m \log(2\pi) + \tfrac{1}{2} \log \det R - \tfrac{1}{2N} \sum_{t=1}^{N} \hat{\varepsilon}(t)^T R^{-1} \hat{\varepsilon}(t) \tag{14}$$

If the covariance matrix R of the innovation $\varepsilon(t)$ is known then the maximisation of $\mathcal{L}(\cdot)$ is equivalent to the minimisation of

$$J_1(\theta) = \frac{1}{2N} \sum_{t=1}^{N} \hat{\varepsilon}(t)^T R^{-1} \hat{\varepsilon}(t) = \tfrac{1}{2} \text{ trace } R^{-1} Q(\theta) \tag{15}$$

where Q is the sample covariance matrix of the residuals $\hat{\varepsilon}(t)$

$$Q(\theta) = \frac{1}{N} \sum_{t=1}^{N} \hat{\varepsilon}(t) \hat{\varepsilon}^T(t) \tag{16}$$

and the property $x^T R x = \text{trace } R x x^T$ has been used. Generally, the covariance R will be unknown and will therefore need to be estimated. Differentiating eqn (14) with respect to R gives

$$\frac{\partial \mathcal{L}(\cdot)}{\partial R} = \tfrac{1}{2} R^{-1} - \frac{1}{2N} R^{-1} \sum_{t=1}^{N} \hat{\varepsilon}(t) \hat{\varepsilon}(t)^T R^{-1} \tag{17}$$

where the identities $\partial / \partial R \log \det R = R^{-T}$ and $\partial / \partial R (\text{trace } W R^{-1}) = -(R^{-1} W R^{-1})^{-T}$ have been used. Setting $\partial(\cdot)/\partial R$ equal to zero gives the maximum

likelihood estimate of the covariance matrix of the residuals

$$\hat{R} = \frac{1}{N} \sum_{t=1}^{N} \hat{\varepsilon}(t)\hat{\varepsilon}(t)^T = Q(\theta) \tag{18}$$

and substituting this value of the covariance matrix in $\mathcal{L}(\cdot)$ gives

$$\mathcal{L}(\theta; Z_N, U_N) = \tfrac{1}{2}m(\log 2\pi + 1) + \tfrac{1}{2}\log \det Q(\theta) \tag{19}$$

Hence the maximization of $\mathcal{L}(\cdot)$ is equivalent to the minimization of

$$J_2(\theta) = \log \det Q(\theta) \tag{20}$$

The maximum likelihood estimator is consistent, asymptotically normally distributed and asymptotically efficient [Goodwin and Payne 1977]. The covariance matrix of the maximum likelihood estimator reaches the Cramer-Rao bound asymptotically.

4. Prediction Error Estimation for the NARMAX Model

When it is known that a system is linear it is often possible by the central limit theorem to assume that the innovations have a Gaussian distribution and the maximum likelihood method can be applied directly. When a system is nonlinear however the distribution of the prediction error will seldom be Gaussian and will usually be unknown. Maximum likelihood estimation cannot in general be applied when the system is nonlinear therefore and alternative algorithms must be developed. The least squares algorithms described in Billings and Voon [1984] are one possibility but in the present study prediction error algorithms based either on the criterion $J_1(\theta)$ eqn (15) or $J_2(\theta)$ eqn (20) will be considered. Minimisation of either of these criterions with respect to the unknown parameter vector θ produces prediction error estimates with very similar asymptotic properties as the maximum likelihood estimator [Goodwin and Payne 1977]. Because both of these criteria were derived

from the log likelihood function it is not surprising that for Gaussian innovations the prediction error method is equivalent to the maximum likelihood method. In the case of non-Gaussian innovations the prediction error method can be applied without any knowledge of the distribution of the innovations.

The prediction error estimates obtained by minimising either $J_1(\theta)$ or $J_2(\theta)$ are strongly consistent and asymptotically normally distributed. In the case where the criterion $J_2(\theta)$ is used the inverse of the Hessian of the loss function at its minimum approaches $\bar{P}_2$, the per sample asymptotic covariance matrix of the estimator. This is an equivalent result to the maximum likelihood estimator except that here $\bar{P}_2$ does not equal the Cramer-Rao bound but takes a somewhat larger value since the prediction error method is not in general asymptotically efficient. For this reason only $J_2(\theta)$ eqn (20) will be considered in the present analysis. These results indicate that the maximum likelihood algorithm derived for Gaussian innovations can be applied to general distributions without any of the essential properties being lost. Moreover the asymptotic normality results for the prediction error methods means that statistical tests such as the t-test or F-ratio test can be applied to determine significant parameters in the estimated model.

Minimisation of the criterion $J_2(\theta)$, eqn (20) to yield the parameter estimates must be done using numerical methods. Differentiating $J_2(\theta)$ eqn (20) with respect to θ gives the gradient

$$\frac{\partial J_2}{\partial \theta_i} = \frac{2}{N} \sum_{t=1}^{N} \hat{\varepsilon}^T(t) Q(\theta)^{-1} \frac{\partial \hat{\varepsilon}(t)}{\partial \theta_i} \qquad i = 1 \ldots n_\theta \tag{21}$$

The derivative of the residuals $\partial \hat{\varepsilon}(t)/\partial \theta_i$ are calculated from

$$\frac{\partial \hat{\varepsilon}(t)}{\partial \theta_i} = -\frac{\partial}{\partial \theta_i} f(Z_{t-1}, U_t) \qquad i = 1,\ldots n_\theta \tag{22}$$

The Hessian of the function $J_2(\theta)$ is given by

$$\frac{\partial^2 J_2}{\partial \theta_i \partial \theta_j} = \frac{2}{N} \sum_{t=1}^{N} \frac{\partial \hat{\varepsilon}(t)^T}{\partial \theta_i} Q(\theta)^{-1} \frac{\partial \hat{\varepsilon}(t)}{\partial \theta_j}$$

$$+ \frac{2}{N} \sum_{t=1}^{N} \hat{\varepsilon}(t)^T Q(\theta)^{-1} \frac{\partial^2 \hat{\varepsilon}(t)}{\partial \theta_i \partial \theta_j}$$

$$- \frac{2}{N^2} \sum_{t=1}^{N} \sum_{k=1}^{N} \hat{\varepsilon}(t)^T Q(\theta)^{-1} \left[\hat{\varepsilon}(k) \frac{\partial \hat{\varepsilon}(k)^T}{\partial \theta_i} + \frac{\partial \hat{\varepsilon}(k)}{\partial \theta_i} \hat{\varepsilon}(k) \right]$$

$$\cdot Q(\theta)^{-1} \frac{\partial \hat{\varepsilon}(t)}{\partial \theta_j}$$

$$\text{for } i,j = 1,2,\ldots n_\theta \tag{23}$$

The second derivatives of the residuals are calculated from

$$\frac{\partial^2 \hat{\varepsilon}(t)}{\partial \theta_i \partial \theta_j} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(Z_{t-1}, U_t) \qquad i,j = 1,2\ldots n_\theta \tag{24}$$

The second term of eqn (23), for θ approaching the true parameter value, tends to the expected value of zero. The third term of eqn (23) also approaches zero for $N \to \infty$. This suggests that the first term only should be used for an approximate calculation of the Hessian. This also guarantees that the approximate Hessian matrix is always positive definite and thus no measures for the contrary need to be taken in the numerical optimization procedure. Hence the Hessian matrices are calculated as

$$\frac{\partial^2 J_2^*}{\partial\theta_i \partial\theta_j} = \frac{2}{N} \sum_{t=1}^{N} \frac{\partial\hat{\varepsilon}(t)^T}{\partial\theta_i} Q(\theta)^{-1} \frac{\partial\hat{\varepsilon}(t)}{\partial\theta_j} \tag{25}$$

The prediction error estimates are then given by

$$\hat{\theta}_i(k+1) = \hat{\theta}_i(k) - \alpha \left(\frac{\partial^2 J_2^*}{\partial\theta_i \partial\theta_j} \right)^{-1} \frac{\partial J_2}{\partial\theta_i(k)} \quad i,j = 1,2\ldots n_\theta \tag{26}$$

where α is a line search constant. Notice that on substituting eqn (21) and eqn (25) into eqn (26) $Q(\theta)^{-1}$ cancels. The algorithm can now be summarised as

(i) Initially, estimate $\theta(k)$, for $k = 0$ using a simple least squares estimator

(ii) Calculate $\partial J_2/\partial\theta_i$, $\partial^2 J_2^*/\partial\theta_i\partial\theta_j$ at $\theta(k)$ for $i,j = 1,2\ldots n_\theta$.

(iii) Use a line search algorithm, such as the Golden section method [Bazaraa and Shetty 1979], to estimate α such that the loss function $J_2(\theta)$ is minimised.
{Incorporating a line search speeds up the convergence of the estimates}.

(iv) Evaluate the estimates eqn (26).

(v) If $J_2(\hat{\theta}(k+1)) - J_2(\hat{\theta}(k)) \leq$ a suitable convergence limit then stop, otherwise increase k to k+1 and go to (ii).

4.1 Example

For example, consider a NARMAX model with first order dynamics and second degree nonlinearity

$$z(t) = F^2[z(t-1), u(t-1), e(t-1)] + e(t)$$

$$= \psi^T(t)\theta + e(t)$$

$$z(t) = \theta_1 z(t-1) + \theta_2 u(t-1) + \theta_3 e(t-1) + \theta_4 z^2(t-1) + \theta_5 z(t-1)u(t-1) + \theta_6 z(t-1)e(t-1) + \theta_7 u^2(t-1) + \theta_8 u(t-1)e(t-1) + \theta_9 e^2(t-1) + e(t) \tag{27}$$

where the only assumption made about the noise e(t) is that it is independent of the input. The prediction error, $\hat{\varepsilon}(t)$ can be computed from

$$\hat{\varepsilon}(t) = z(t)-\theta_1 z(t-1)-\theta_2 u(t-1)-\theta_3\hat{\varepsilon}(t-1)-\theta_4 z^2(t-1)-\theta_5 z(t-1)u(t-1)-$$
$$\theta_6 z(t-1)\hat{\varepsilon}(t-1)-\theta_7 u^2(t-1)-\theta_8 u(t-1)\hat{\varepsilon}(t-1)-\theta_9\hat{\varepsilon}^2(t-1) \quad (28)$$

once initial estimates of θ_i are available. Notice that because the NARMAX model maps the past of the input and output into the present output multiplicative noise terms are introduced. For the estimates to be unbiased, all the coefficients θ_i including the coefficients of the noise terms must be included in the parameter vector. The gradient matrix is then given by

$$\frac{\partial J_2}{\partial \theta} = \begin{bmatrix} \frac{\partial J_2}{\partial \theta_1} \\ \vdots \\ \frac{\partial J_2}{\partial \theta_9} \end{bmatrix} = \frac{2}{N} Q(\theta)^{-1} \begin{bmatrix} \sum_{t=1}^{N} \hat{\varepsilon}(t) \frac{\partial \hat{\varepsilon}(t)}{\partial \theta_1} \\ \vdots \\ \sum_{t=1}^{N} \hat{\varepsilon}(t) \frac{\partial \hat{\varepsilon}(t)}{\partial \theta_9} \end{bmatrix} \quad (29)$$

where $\frac{\partial \hat{\varepsilon}(t)}{\partial \theta_i}$ are obtained by differentiating equation (28)

$$
\begin{aligned}
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_1} &= -\,z(t-1) - [\theta_3+\theta_6 z(t-1)+\theta_8 u(t-1)+2\theta_9\hat{\varepsilon}(t-1)]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_1}\\
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_2} &= -\,u(t-1) - [\quad\text{"}\quad]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_2}\\
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_3} &= -\,\hat{\varepsilon}(t-1) - [\quad\text{"}\quad]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_3}\\
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_4} &= -\,z^2(t-1) - [\quad\text{"}\quad]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_4}\\
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_5} &= -\,z(t-1)u(t-1) - [\quad\text{"}\quad]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_5}\\
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_6} &= -\,z(t-1)\hat{\varepsilon}(t-1) - [\quad\text{"}\quad]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_6}\\
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_7} &= -\,u^2(t-1) - [\quad\text{"}\quad]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_7}\\
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_8} &= -\,u(t-1)\hat{\varepsilon}(t-1) - [\quad\text{"}\quad]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_8}\\
\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_9} &= -\,\hat{\varepsilon}^2(t-1) - [\quad\text{"}\quad]\,\frac{\partial\hat{\varepsilon}(t-1)}{\partial\theta_9}
\end{aligned}
\qquad (30)
$$

(1) (2)

Column (1) has the same form as the data vector matrix and column (2) represents the coefficients of the differentiated NARMAX noise model. Equation (30) can then be substituted into equation (29) to evaluate the gradient of the loss function.

The approximated Hessian matrix is represented as

$$\begin{bmatrix} \frac{\partial^2 J_2^*}{\partial\theta_1\partial\theta_1} & \cdots & \frac{\partial^2 J_2^*}{\partial\theta_1\partial\theta_9} \\ \vdots & & \\ \frac{\partial^2 J_2^*}{\partial\theta_9\partial\theta_1} & \cdots & \frac{\partial^2 J_2^*}{\partial\theta_9\partial\theta_9} \end{bmatrix}$$

$$= \frac{2}{N} Q(\theta)^{-1} \begin{bmatrix} \sum_{t=1}^{N} \frac{\partial\hat{\varepsilon}(t)}{\partial\theta_1}\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_1} & \cdots & \sum_{t=1}^{N} \frac{\partial\hat{\varepsilon}(t)}{\partial\theta_1}\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_9} \\ \vdots & & \\ \sum_{t=1}^{N} \frac{\partial\hat{\varepsilon}(t)}{\partial\theta_9}\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_1} & \cdots & \sum_{t=1}^{N} \frac{\partial\hat{\varepsilon}(t)}{\partial\theta_9}\frac{\partial\hat{\varepsilon}(t)}{\partial\theta_9} \end{bmatrix} \qquad (31)$$

and is computed by substituting equation (30) into equation (31). The estimates are then calculated using eqn (26) and the deterministic predicted output can be evaluated as

$$\hat{z}(t) = F^{\ell}[\hat{z}(t-1),\ldots,\hat{z}(t-n_z),u(t-d),\ldots,u(t-d-n_u+1)]$$

or

$$\hat{z}(t) = \psi_{\hat{z}u}^{T}(t)\hat{\theta}_{\hat{z}u} \qquad (32)$$

5. Structure Determination

The determination of the structure or which terms to include in the model is essential in nonlinear parameter estimation since the NARMAX model can easily become overparameterised. Direct estimation based on say a polynomial expansion of eqn (2) may involve an excessive number of terms. Simply increasing the order of the dynamic terms $(n_u, n_z, n_\varepsilon)$ and the order of the polynomial expansion ℓ to achieve the desired prediction accuracy will in general result in an excessively complex model and possibly numerical ill-conditioning. The maximum number of coefficients in the

nonlinear model equation (2) is given by

$$n = \sum_{i=1}^{\ell} n_i$$

$$n_i = \left[n_{i-1}(n_z+n_u+n_\varepsilon+i-1)\right]/i \ , \quad \text{where } n_o = 1 \tag{33}$$

For example, a first order dynamic model for input, output and noise with third degree nonlinearity would have 19 coefficients whilst a fifth degree nonlinearity would have 55 coefficients. Simulation has shown [Billings and Fadzil 1985] that usually less than ten of the terms in the NARMAX model are significant and the remainder can be discarded with little deterioration in the prediction accuracy of the model.

The Akaike [1974,1977] information criteria which indicates the optimum number of coefficients that are required to characterize a model is given by

$$AIC = N \log_e \hat{R} + 2 \text{ (no. of parameters)} \tag{34}$$

where N represents the number of data points and $\hat{R}$ is the variance of the prediction error.

The optimum number of coefficients is chosen when any further increase in the number of parameters does not increase the AIC criteria. This has the disadvantage that the experimenter has to guess the best set of significant coefficients from the various permutations available within the total number of coefficients and this can be computationally expensive when the system is nonlinear.

Alternative methods of detecting model structure can be based on the forward, backward or stepwise regression methods.

The forward regression [Draper and Smith 1981] method selects each coefficient in the model in turn until the regression is satisfactory. The order of including the coefficients is determined by using the partial correlation coefficient as a measure of the significance of the

coefficients that have not yet been included in the equation. The coefficient that is most highly correlated with the output is chosen as the next term to be included in the equation. This process is repeated until the most recently selected coefficient has an F-ratio test that indicates insignificant partial correlation and the process is terminated. This forward regression method does not examine the effect of a newly selected coefficient which may cause any of those coefficients entered in the earlier stages to become insignficiant.

The backward regression [Draper and Smith 1981, Smillie 1966] method initially estimates all the coefficients that are included in the model and then calculates the partial F-ratio for each individual term. The lowest value of the partial F-ratio is selected and tested for its significance level. If this partial F-ratio is insignificant as compared to a preselected significance F-value, the corresponding term is removed from the regression and the coefficients that remain in the model are re-estimated. This process is repeated until the lowest partial F-ratio is significant and the process is terminated. This determines the final selected model required to represent the system. However, this method of including all the coefficients into the regression in the initial stage may sometimes cause the covariance or information matrix to become ill-conditioned and possibly marginally singular which often yields erroneous estimates.

However, the forward and backward regression methods can be combined together to form the so-called stepwise regression method.

5.1 Stepwise Regression

The stepwise regression [Hall, Gupta and Tyler 1974, Efroymson 1962, Draper and Smith 1981, Smillie 1966, Klein, Batterson and Murphy 1981] technique is used to detect the set of coefficients that can best

characterize a model. In fact, stepwise regression is computationally efficient because it gives intermediate statistical information at each stage of the calculation which is used to select the most appropriate coefficients to be added into the model. The model is increased by adding one coefficient at a time during the intermediate stage, such that the added coefficient is the one which contributes the greatest improvement in the 'goodness of fit' to the model. If a coefficient which was significant at an earlier stage later becomes insignificant, after several other coefficients are included in the model, this coefficient is then deleted before adding another significant coefficient. Hence by adding and deleting the appropriate coefficients, the best model should be determined.

Consider the NARMAX model

$$z(t) = F^{\ell}[z(t-1),\ldots,z(t-n_z),u(t-d),\ldots,u(t-d-n_u+1), \varepsilon(t-1),\ldots,\varepsilon(t-n_\varepsilon)] + \varepsilon(t) \qquad (35)$$

Expanding equation (35) as a polynomial expansion gives

$$\begin{aligned} x_z(t) = {} & \theta_1 x_1(t)+\ldots+\theta_{n_z} x_{n_z}(t)+\theta_{n_z+1} x_{n_z+1}(t)+\ldots \\ & +\theta_{n_z+n_u} x_{n_z+n_u}(t)+\theta_{n_z+n_u+1} x_{n_z+n_u+1}(t)+\ldots \\ & +\theta_{n_z+n_u+n_\varepsilon} x_{n_z+n_u+n_\varepsilon}(t)+\theta_{n_z+n_u+n_\varepsilon+1} x_{n_z+n_u+n_\varepsilon+1}(t)+\ldots \\ & .....+\theta_n x_n(t)+\varepsilon(t) \end{aligned} \qquad (36)$$

where $x_z(t) = z(t)$

$x_1(t) = z(t-1)$

. .

$$x_{n_z}(t) = z(t-n_z)$$

$$x_{n_z+1}(t) = u(t-d)$$

$$x_{n_z+n_u}(t) = u(t-d-n_u+1)$$

$$x_{n_z+n_u+1}(t) = \varepsilon(t-1)$$

$$x_{n_z+n_u+n_\varepsilon}(t) = \varepsilon(t-n_\varepsilon)$$

$$x_{n_z+n_u+n_\varepsilon+1}(t) = z^2(t-1)$$

$$x_{n_z+n_u+n_\varepsilon+2}(t) = z(t-1)z(t-2)$$

$$x_{n_z+n_u+n_\varepsilon+3}(t) = z(t-1)z(t-3)$$

$$x_{n_z+n_u+n_\varepsilon+4}(t) = z(t-1)z(t-4)$$

$$x_k(t) = \text{polynomial terms of } z(t), u(t) \text{ and } \varepsilon(t)$$

$$x_n(t) = \varepsilon^{\ell}(t-n_\varepsilon) \tag{37}$$

where n is given by equation (33).

Equation (36) can now be expressed as

$$x_z(t) = \sum_{i=1}^{n} \theta_i x_i(t) + \varepsilon(t) \tag{38}$$

The data for each term are then usually normalised by removing the mean for each of the corresponding terms and hence equation (38) becomes

$$x_z(t)-\bar{x}_z = \sum_{i=1}^{n} \theta_i (x_i(t)-\bar{x}_i) + \varepsilon(t) \qquad (39)$$

where the bar denotes average value.

The error in the estimates at the t-th observation for N data points is

$$\varepsilon(t) = (x_z(t)-\bar{x}_z) - \sum_{i=1}^{n} \theta_i (x_i(t)-\bar{x}_i)$$

so that minimizing the sum of the squares of the vector $\varepsilon(t)$

$$\|\varepsilon\|^2 = \sum_{t=1}^{N} \left\{ (x_z(t)-\bar{x}_z) - \sum_{i=1}^{n} \theta_i (x_i(t)-\bar{x}_i) \right\}^2 \qquad (40)$$

with respect to θ_i yields

$$\sum_{j=1}^{n} \left\{ \sum_{t=1}^{N} (x_i(t)-\bar{x}_i)(x_j(t)-\bar{x}_j) \right\} \theta_j = \sum_{t=1}^{N} (x_i(t)-\bar{x}_i)(x_z(t)-\bar{x}_z) \qquad (41)$$

$$\text{for } i = 1,2,\ldots,n$$

These are just a set of n simultaneous linear algebraic equations in θ_i which can now be expressed in the correlation coefficient r_{ij} form

$$\underbrace{\begin{pmatrix} r_{11} & r_{12} & \cdots\cdots & r_{1n} \\ r_{21} & r_{22} & \cdots\cdots & r_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ r_{n1} & r_{n2} & \cdots\cdots & r_{nn} \end{pmatrix}}_{A} \begin{pmatrix} \theta_1^* \\ \theta_2^* \\ \cdot \\ \cdot \\ \theta_n^* \end{pmatrix} = \begin{pmatrix} r_{z1} \\ r_{z2} \\ \cdot \\ \cdot \\ r_{zn} \end{pmatrix} \qquad (42)$$

where

$$r_{ij} = \frac{\sum_{t=1}^{N} (x_i(t)-\bar{x}_i)(x_j(t)-\bar{x}_j)}{\sigma_i \sigma_j}$$

$$r_{zi} = \frac{\sum_{t=1}^{N} (x_i(t)-\bar{x}_i)(x_z(t)-\bar{x}_z)}{\sigma_i \sigma_z}$$

$$\sigma_i = \sqrt{\sum_{t=1}^{N} (x_i(t) - \bar{x}_i)^2}$$

$$\sigma_z = \sqrt{\sum_{t=1}^{N} (x_z(t) - \bar{x}_z)^2}$$

and

$$\theta_i = \frac{\sigma_z}{\sigma_i} \theta_i^* \qquad (43)$$

The stepwise regression procedure begins by selecting the term which is most closely correlated with the output $x_z(t)$, that is the variable $x_i(t)$ say whose r_{zi} is the largest of all the r_{zk}, $k = 1,2,...n$. The model

$$x_z(t) = \theta_i x_i(t) + \varepsilon(t) \qquad (44)$$

is then used to fit the data. The next term to be included in the model is that particular term which gives the largest partial correlation between the output $x_z(t)$ and the j-th term, $x_j(t)$ say, from the remaining terms not in the regression after removing the effect of the i-th term. This partial correlation is defined as $r_{zj.i}$ and is interpretted as the partial correlation between $x_z(t)$ and the j-th term with the effect of the i-th term removed. The resulting model is then

$$x_z(t) = \theta_i x_i(t) + \theta_j x_j(t) + \varepsilon(t) \qquad (45)$$

Within each intermediate stage, the partial correlation between the output and the k-th term of the remaining coefficients not in the model can be calculated from previous calculated correlations using the general recurrence relation [Smillie 1966]

$$r_{zk.ij...k-1} = \frac{r_{zk.ij...k-2} - r_{zk-1.ij...k-2} r_{kk-1.ij...k-2}}{\sqrt{(1-r^2_{zk-1.ij...k-2})(1-r^2_{kk-1.ij...k-2})}} \qquad (46)$$

Before each of the selected terms with the largest partial correlation are included in the model, those terms that are already in the model are tested for their significance using the statistical t-test. Since a prediction error estimator is used to estimate the coefficients within each stage, the estimates have an estimated prediction error variance, $\hat{R}$ eqn (18) and the estimated coefficient $\hat{\theta}_i$ is normally distributed about the mean and has a variance, $\hat{R}P_{2,ii}$, where $P_{2,ii}$ is the diagonal term of the i-th element of the prediction error parameter covariance matrix P_2. The t-test is given by

$$t_{N-n^*} = \frac{\hat{\theta}_i - \theta_i}{\sqrt{\hat{R}P_{2,ii}}} \quad , \quad i = 1,2,...n^* \qquad (47)$$

with N-n* degrees of freedom, where n* is the number of coefficients already in the model. It is desirable to test the hypothesis that $\theta_i = 0$ (i.e. $x_z(t)$ does not depend on θ_i) and the statistic

$$t_{N-n^*} = \frac{\hat{\theta}_i}{\sqrt{\hat{R}P_{2,ii}}} \quad , \quad i = 1,2,...n^* \qquad (48)$$

is used. However, the F-distribution with 1 and N-n* degrees of freedom is equivalent to the t^2 distribution with N-n* degrees of freedom and hence the significance of the individual coefficient θ_i can be determined

from the F-ratio test

$$F_i = \frac{\hat{\theta}_i^2}{\hat{RP}_{2,ii}} \quad , \quad i = 1,2,\ldots,n^* \tag{49}$$

Normally, a 95% confidence interval for the F-ratio test, F_{test} is chosen and when N-n*>100 is large, the typical F_{test} value is ≈4. If the smallest $F_i \geq F_{test}$ then the coefficient is significant and no coefficient is deleted, on the other hand, if the smallest $F_i < F_{test}$, then that particular corresponding coefficient is deleted.

To bring in another variable into the model, the partial correlation coefficients of all other parameters are examined and the F-ratio test for the coefficients is formed

$$F_k = \frac{r^2_{zk.ij\ldots k-1}(N-n^*)}{(1-r^2_{zk.ij\ldots k-1})} \tag{50}$$

The above procedure is repeated until no term is either deleted or included in the determined model. The final model with n* number of significant coefficients is then expressed as

$$x_z(t) = \sum_{i=1}^{n^*} \theta_i x_i(t) + \varepsilon(t) \tag{51}$$

where θ_i are the significant coefficients.

The F-ratio test as described above, which is included in the stepwise regression routine requires considerable computation for every intermediate stage. Fortunately, it is not necessary to calculate the covariance P_2, the prediction error covariance $\hat{R}$ and the various partial correlation coefficients as described above, in order to detect the significant coefficients in the model. The stepwise regression technique is simply adapted using a Gauss-Jordan [Efroymson 1961, Smillie 1966]

elimination algorithm based on an augmented form of the transformed correlation matrix of equation (42) and not on the original matrices of equation (41). The Gauss-Jordan elimination technique uses the pivot method and is well documented elsewhere and will not be detailed here.

The pivot elements in the G-J elimination method are calculated along the principal diagonal of the matrix A eqn (42) which has the effect of including that particular coefficient corresponding to the pivot element which is selected. Suppose that n* coefficients are significantly selected then the G-J elimination method merely transforms the matrix A of equation (42) to a unity diagonal matrix of dimension (n*xn*)

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_n^* \end{bmatrix} = \begin{bmatrix} r_{z1}^{tr}(n^*) \\ r_{z2}^{tr}(n^*) \\ \vdots \\ r_{zn}^{tr}(n^*) \end{bmatrix} \qquad (52)$$

where $r_{zi}^{tr}(n^*)$ represents the transformed value at the n* stage of the G-J elimination and θ_i^* are the estimates. The $r_{zz}^{tr}(n^*)$ term is the variance of the predicted error at the n* stage. From equation (52), equation (43) can be expressed as

$$\theta_i = \frac{\sigma_z r_{zi}^{tr}(n^*)}{\sigma_i} \quad , \qquad i = 1,2,\ldots,n^* \qquad (53)$$

Therefore, the unity matrix of equation (52) resulting from the transformation of the correlation matrix A of equation (42) is used as

a working matrix to store the transformed value calculated during the elimination process. This minimizes the storage required in the computation. The matrix transformation [Smillie 1966] used at the k-th intermediate stage for the m-th selected coefficient is given by the iterative relations below

$$r_{ij}(k) = \frac{r_{mm}(k-1)\,r_{ij}(k-1) - r_{im}(k-1)\,r_{mj}(k-1)}{r_{mm}(k-1)} \quad \text{for } i \neq m,\ j \neq m$$

$$r_{ij}(k) = \frac{-\,r_{im}(k-1)}{r_{mm}(k-1)} \quad \text{for } i \neq m,\ j = m \tag{54}$$

$$r_{mj}(k) = \frac{r_{mj}(k-1)}{r_{mm}(k-1)} \quad \text{for } i = m,\ j \neq m$$

$$r_{mm}(k) = \frac{1}{r_{mm}(k-1)} \quad \text{for } i = m,\ j = m$$

The values of the matrix A at the k-th stage of the elimination for the m-th selected coefficient is equal to the inverse of that particular m-th coefficient. This matrix A also provides statistical information for the F-ratio test for each individual term. The F-ratio for those terms that are in the model at the k-th stage of the elimination process is evaluated as

$$F_{di} = \frac{(N-n^*)\,r_{zi}^2(k)}{r_{ii}(k)\,r_{zz}(k)} \quad , \quad i = 1,2,\ldots,n^* \tag{55}$$

Equation (55) is equivalent to equation (49), since

$$\hat{R} = \frac{r_{zz}(k)}{(N-n^*)} \quad , \quad P_{2,ii} = r_{ii}(k) \quad \text{and} \quad \theta_i^* = r_{zi}(k) \ .$$

If the minimum value of F_d, F_{min} is insignificant compared to the 95% significance level then that corresponding m-th coefficient is deleted

by dividing that particular term by its inverse using equation (54). In order to include the next most significant term that is not in the regression, the F-ratio is calculated as

$$F_{aj} = \frac{(N-n^*)V_j}{(r_{zz}(k)-V_j)} \qquad j = 1+n^*,\ldots,n-n^* \tag{56}$$

If the maximum value of the F_a, F_{max} is significant compared to the 95% significance level, then that particular m-th term is included in the model which can be estimated using equation (54). In fact equation (56) is equivalent to equation (50), since

$$V_j = \frac{r_{jz}^2(k)}{r_{jj}(k)}$$

and

$$\frac{V_j}{r_{zz}(k)} = \frac{r_{jz}^2(k)}{r_{zz}(k)r_{jj}(k)} = r_{jz.ik\ldots n-n^*}^2 \tag{57}$$

where the ik...N-n* are terms already in the model.

In the above computations the correlation coefficients were normalised by removing the mean from each individual term and dividing by the standard deviations. Whilst this is highly desirable when dealing with linear systems mean levels must not be removed when the model to be fitted is nonlinear. Removing mean levels from signals when estimating nonlinear models can change the structure of the model and will almost always induce input sensitivity [Billings and Voon 1984]. This means that the model parameters become a function of the statistics of the input signal. A model estimated for one particular input would therefore not be valid for prediction based on any other input with different statistics. This problem can be avoided by operating on the raw data and including a constant term in the model eqn (2) and (38).

These problems can easily be avoided when using the stepwise regression algorithm above for the NARMAX model if all mean levels are set to zero.

5.1.1. Implementing Stepwise Regression

Implementation of the combined algorithm begins by postulating the terms which might enter the model and specifying the parameters d, n_u, n_z, n_ε and ℓ in eqn (35). Application of the algorithm to numerous simulated and industrial processes [Billings and Fadzil 1984,1985] has shown that good initial estimates of d, n_u, n_z and n_ε can be obtained by fitting a linear model to the data initially. This can be achieved either by implementing the prediction error algorithm with ℓ set to unity or by applying any appropriate linear parameter estimation routine. Appropriate values for d, n_u, n_z, n_ε can then be determined by initially setting each of these parameters to some small value and estimating models over a range of values. The best linear model is then selected by applying standard linear methods of model order and time delay selection and examining the predicted model output and analysing the residuals [Ljung and Soderstrom 1983]. Note that it may not be possible to satisfy the normal linear validation tests of white residuals ($\phi_{\hat{\varepsilon}\hat{\varepsilon}}(\tau) = \delta(\tau)$) and residuals which are uncorrelated with the input ($\phi_{u\hat{\varepsilon}}(\tau) = 0$) because of the presence of nonlinear effects in the data set. However, this initial analysis does seem to provide excellent initial estimates for d, n_u, n_z, n_ε and the prediction error sequence $\hat{\varepsilon}(t)$.

Estimation of the nonlinear model begins by selecting ℓ the degree of nonlinearity in eqn (35) (typically $\ell \leq 3$) and entering the values of d, n_u, n_z, n_ε and $\hat{\varepsilon}(t)$ into the stepwise regression routine. To simplify the computations the prediction error sequence $\hat{\varepsilon}(t)$ is not regenerated for every inclusion or deletion of coefficients in the model. The final estimates from the stepwise regression routine may well therefore be

slightly biased. Experience has shown that improved prediction accuracy is obtained if the linear terms found to be significant in the initial linear model estimation are forced into the model so that the stepwise regression routine detects significant nonlinear terms only [Billings and Fadzil 1984,1985].

As an alternative to the above procedure the detailed estimation of a linear model can be omitted and a recursive least squares routine can be used to provide initial estimates of the prediction error sequence for entry into the stepwise regression algorithm. Application of the algorithm has shown however that this approach produces inferior models.

The output from the stepwise regression routine defines the NARMAX model structure and the estimates are used as start values in the prediction error algorithm. The prediction error routine re-estimates the significant coefficients and the model is tested using an F-ratio test to ensure that the coefficients are significant. Any coefficients found to be insignificant are deleted and the reduced model is once again optimised in the prediction error algorithm. This procedure is repeated until all the coefficients are significant. The F-ratio test for the prediction error estimates is defined by

$$F_i = \frac{\theta_i^2}{\hat{RP}_{2,ii}} \qquad i = 1,2,\ldots n^* \tag{58}$$

A flow chart of the procedure is illustrated in Fig.1.

The stepwise regression algorithm can if required be combined with any parameter estimation algorithm. The prediction error routine would then for example be replaced by an extended least squares, modified instrumental variables or suboptimal least squares routine [Billings and Voon 1984].

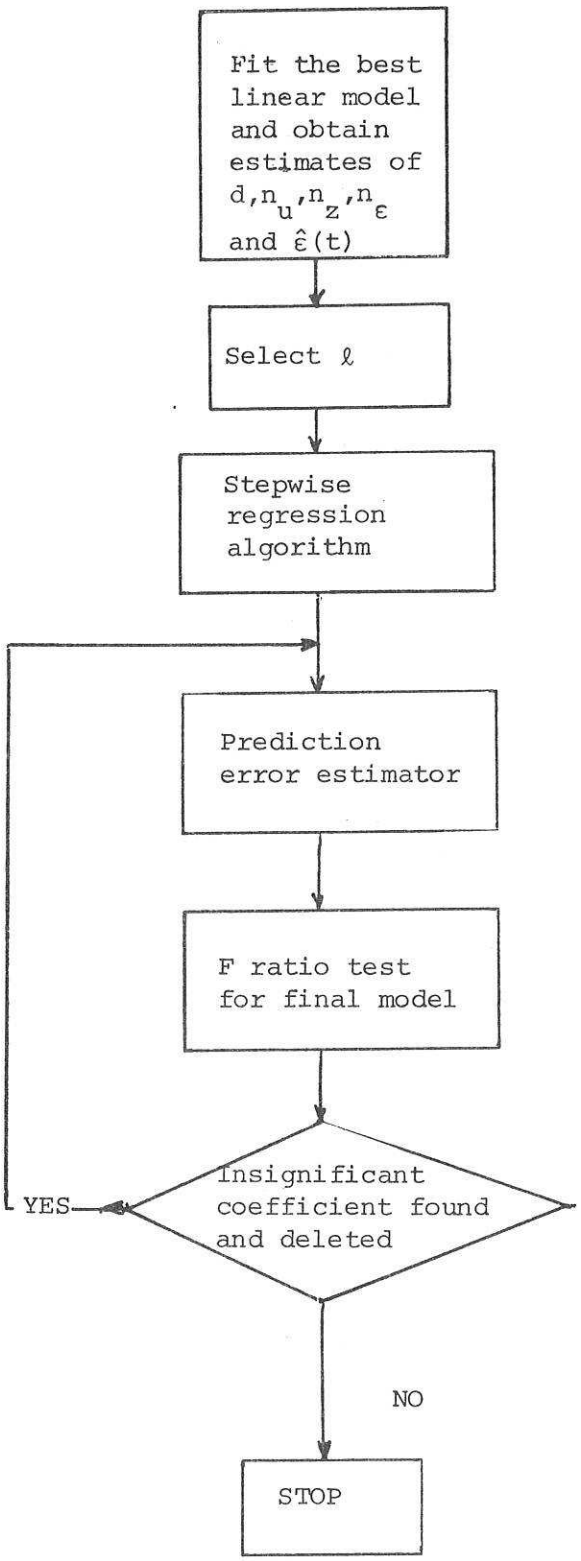
Fit the best linear model and obtain estimates of d, n_u, n_z, n_ε and $\hat{\varepsilon}(t)$
Select ℓ
Stepwise regression algorithm
Prediction error estimator
F ratio test for final model
Insignificant coefficient found and deleted
YES
NO
STOP

Fig. 1

Whatever parameter estimation algorithm is used to optimise the final model the data set should be split to form an estimation set and a testing set. The estimation set is used to estimate all the models and the testing set it used to provide a comparison between the plant and model output. Experience has shown that linear models often produce a predicted output which follows the data in the estimation set reasonably closely. However, prediction over a different piece of data, the testing set, usually reveals severe deficiencies in the model when nonlinear effects are present in the data.

5.2 Model Validation

When the parameters are estimated, there is no guarantee that the significant coefficients that are selected in the NARMAX model represent the true model. Irrespective of which parameter estimation routine was implemented, if the model structure and parameter estimates are correct then the prediction error sequence $\hat{\varepsilon}(t)$ should be unpredictable from all linear and nonlinear combinations of past inputs and outputs and this condition will hold iff [Billings and Voon 1983]

$$\begin{aligned} \phi_{\hat{\varepsilon}\hat{\varepsilon}}(\tau) &= \delta(\tau) \\ \phi_{u\hat{\varepsilon}}(\tau) &= 0 \;\forall\; \tau \\ \phi_{\hat{\varepsilon}(\hat{\varepsilon}u)}(\tau) &= 0 \;\forall\; \tau \end{aligned} \tag{59}$$

If instrumental variables or suboptimal least squares algorithms are used the prediction errors may be coloured. In this case the prediction error $\hat{\varepsilon}(t)$ should be independent of all linear and nonlinear combinations of inputs and this condition will hold iff [Billings and Voon 1983]

$$\begin{aligned} \phi_{u^2\hat{\varepsilon}^2}(\tau) &= 0 \;\forall\; \tau \\ \phi_{u^2\hat{\varepsilon}}(\tau) &= 0 \;\forall\; \tau \\ \phi_{u\hat{\varepsilon}}(\tau) &= 0 \;\forall\; \tau \end{aligned} \tag{60}$$

Notice that for nonlinear systems the traditional linear covariance tests $\phi_{\hat{\varepsilon}\hat{\varepsilon}}(\tau)$, $\phi_{u\hat{\varepsilon}}(\tau)$ are not sufficient.

Experience has shown that when using a prediction error algorithm the tests in both eqns (59) and (60) give the experimenter a great deal of information regarding deficiencies in the fitted model and can indicate which terms should be included in the model to improve the fit.

If a linear model is fitted at the initial stage of the analysis the higher order model validity checks in eqns (59) and (60) can be used to indicate if the prediction accuracy of the model could be improved by inserting nonlinear terms into the model. If the tests indicate that no significant nonlinear terms remain in the residuals the analysis would terminate at this stage. Tests which detect the presence of nonlinearities in data prior to parameter estimation are also available [Billings and Voon 1983] and can be used to augment the model validation tests to indicate if it is worthwhile fitting a nonlinear model.

6. Simulation Results

The identification of several simulated systems and industrial processes are described below to illustrate the effectiveness of the algorithm.

6.1 Simulated Examples

Several simulated examples are described below. In each case the stepwise regression routine was entered directly, and linear models were not fitted at the preliminary stage as recommended above. This was possible because the order of the linear dynamics were known in each case.

A Hammerstein model described as

$$y(t) = 1.3y(t-1) - 0.42y(t-2) + 0.2\{u(t-1) + u^3(t-1)\}$$
$$+ 0.4\{u(t-2) + u^3(t-2)\} \qquad (61)$$
$$z(t) = y(t) + e(t)$$

was simulated using a uniformly random distributed input of amplitude ±1.0 superimposed on an operating point b = 0.2. The output y(t) was corrupted by an additive Gaussian noise sequence of zero mean with 0.1 standard deviation, $e(t) \sim \mathcal{N}(0,0.1)$. The estimation set consisted of 500 data points.

Comparison of $\bar{z}_b = 1.02786$ and $\bar{z} = 2.01382$ indicates that the model is nonlinear (because $\bar{z}_b \neq \bar{z}$) and this is confirmed in figure 2 where $\phi_{z'z'^2}(\tau)$ and $\phi_{z_b'z_b'^2}(\tau) \neq 0$.

The stepwise regression routine in association with a prediction error algorithm was used for structure determination and parameter estimation. A nonlinear polynomial model with second order dynamics ($d = 1$, $n_u = n_z = n_\varepsilon = 2$) and third degree nonlinearity ($\ell = 3$) which has 83 terms was initially used to estimate the model. The structure determination algorithm selects the significant terms which are then optimised by the prediction error algorithm to yield the final model with eight significant terms

$$z(t) = 1.302z(t-1) - 0.4219z(t-2) + 0.2074u(t-1) + 0.3835u(t-2)$$
$$+ 0.203u^3(t-1) + 0.4025u^3(t-2) - 1.304\varepsilon(t-1) \qquad (62)$$
$$+ 0.4043\varepsilon(t-2) + \varepsilon(t)$$

The model validity tests eqns (59) and (60) illustrated in figure 3 shows that $\phi_{\xi\xi}(\tau) = \delta(\tau)$, $\phi_{u\xi}(\tau)$, $\phi_{\xi(\xi u)}(\tau)$, $\phi_{u^2{}'\xi^2}(\tau)$, $\phi_{u^2{}'\xi}(\tau)$ and $\phi_{u\xi}(\tau) = 0$ which indicates that the residuals $\xi(t) = \hat{\varepsilon}(t)$ contain no further information and that the estimates are unbiased.

A Wiener model described as

$$y(t) = 0.8y(t-1) + 0.4u(t-1)$$
$$z(t) = y(t) + y^3(t) + e(t) \qquad (63)$$

was excited by a uniformly distributed input with amplitude range ±1.0 superimposed on an operating point b = 0.2. The output y(t) was corrupted by an additive Gaussian noise sequence of zero mean with standard deviation 0.1, $e(t) \sim \mathcal{N}(0,0.1)$. The estimation set consisted of 500 data points.

Comparison of $\bar{z}_b = 0.4588$ and $\bar{z} = 0.6822$ clearly shows that the model is nonlinear (because $\bar{z}_b \neq \bar{z}$) and this is confirmed in figure 4 where $\phi_{z'z'^2}(\tau)$ and $\phi_{z_b'z_b'^2}(\tau) \neq 0$.

A nonlinear polynomial with first order dynamics (d = 1, $n_u = n_z = n_\varepsilon = 1$) and third degree nonlinearity ($\ell = 3$) which has 19 terms was initially used to estimate the model. The algorithm produced a final model with 9 significant coefficients

$$\begin{aligned} z(t) = {} & 0.7578z(t-1) + 0.3891u(t-1) - 0.739\varepsilon(t-1) \\ & - 0.03723z^2(t-1) + 0.3794z(t-1)u(t-1) + 0.0684u^2(t-1) \\ & - 0.368u(t-1)\varepsilon(t-1) + 0.1216z(t-1)u^2(t-1) \\ & + 0.0633u^3(t-1) + \varepsilon(t) \end{aligned} \qquad (64)$$

The model validity tests are illustrated in figure 5 where $\phi_{\xi\xi}(\tau) = \delta(\tau)$, $\phi_{\xi(\xi u)}(\tau)$, $\phi_{u^2'\xi^2}(\tau)$, $\phi_{u^2'\xi}(\tau)$ and $\phi_{u\xi}(\tau) = 0$ clearly indicating that the model is adequate.

6.2 Industrial Examples

Whilst the use of simulated examples was a necessary stage in the development of the stepwise regression/prediction error method it does not provide a realistic test for the algorithms. To achieve the latter objective several sets of data were recorded and sampled from pilot scale industrial processes.

A liquid level system [Billings, Tsang and Voon 1985] consisting of a series of interconnected tanks one of which has a conical section and induces nonlinearities was studied. The model fitted between a perturbation on the input volume flowrate and the level of liquid in the conical tank was (for a sampling interval of 9.6 secs)

$$\begin{aligned} z(t) = {} & 0.436z(t-1) + 0.681z(t-2) - 0.149z(t-3) \\ & + 0.396u(t-1) + 0.014u(t-2) - 0.071u(t-3) \\ & - 0.351z(t-1)u(t-1) - 0.03z^2(t-2) - 0.135z(t-2)u(t-2) \\ & - 0.027z^3(t-2) - 0.108z^2(t-2)u(t-2) \\ & - 0.099u^3(t-2) + \varepsilon(t) + 0.344\varepsilon(t-1) - 0.201\varepsilon(t-2) \end{aligned} \tag{65}$$

All the model validity tests were well within the required confidence intervals indicating that the model is adequate.

A heat exchanger [Billings and Fadzil 1985] consisting of a radiator through which heated water is passed and a fan which blows air across the radiator was studied. The system is a two input (heater and fan controls), two output (drop in temperature across the radiator, air flow rate) system. Models have been fitted to all the loops only one of which is nonlinear. The model for the nonlinear fan/air flow loop was identified as (sampling interval 0.3 secs)

$$\begin{aligned} z(t) = {} & 2.301 + 0.9173z(t-1) + 0.449u(t-1) \\ & + 0.04577u(t-2) - 0.01889z^2(t-1) \\ & - 0.0099u^2(t-1) - 0.002099z^2(t-1)u(t-1) \\ & - 0.00243u^3(t-1) + \varepsilon(t) - 0.004\varepsilon(t-1) \\ & + 0.0380\varepsilon(t-2) + 0.2745\varepsilon(t-3) + 0.1037\varepsilon(t-4) \end{aligned} \tag{66}$$

All the model validity tests eqns (59), (60) were well within the 95% confidence bands.

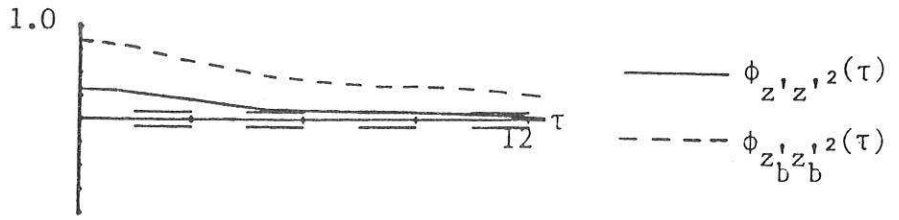

Figure 2. Structure detection for the Hammerstein model

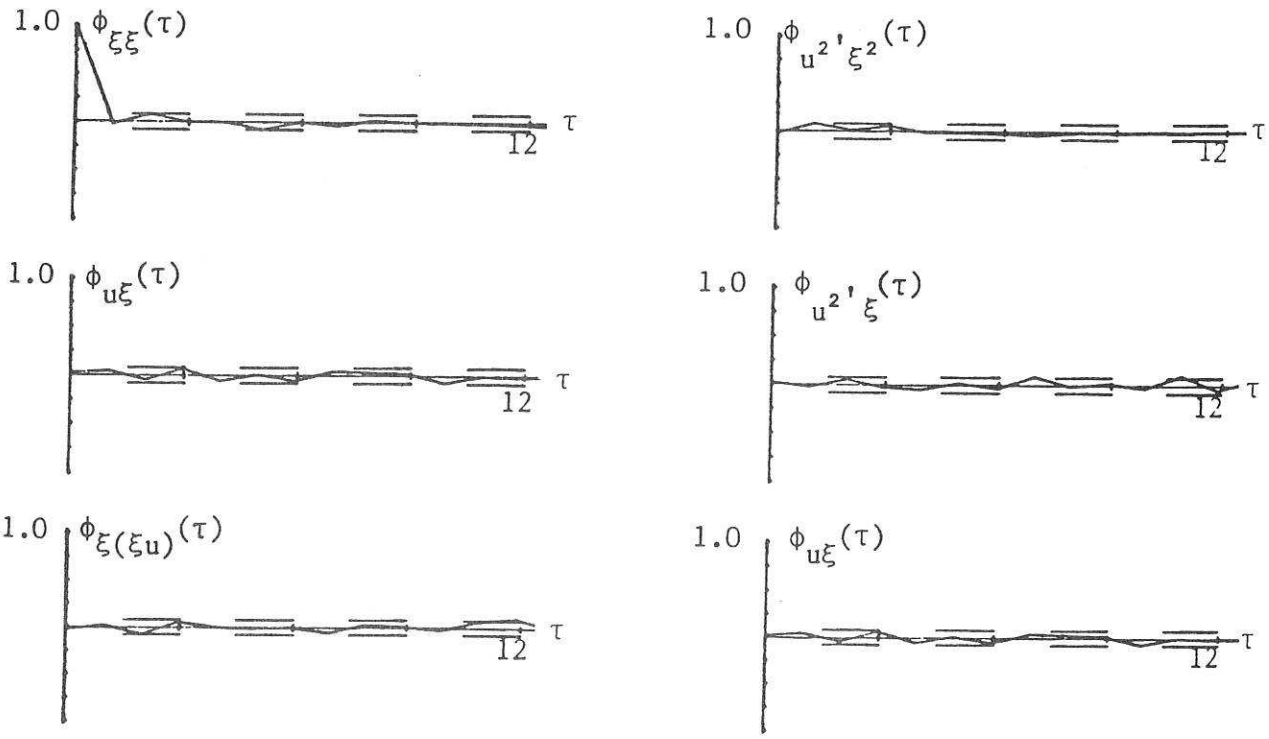

Figure 3. Model validity tests for the Hammerstein model

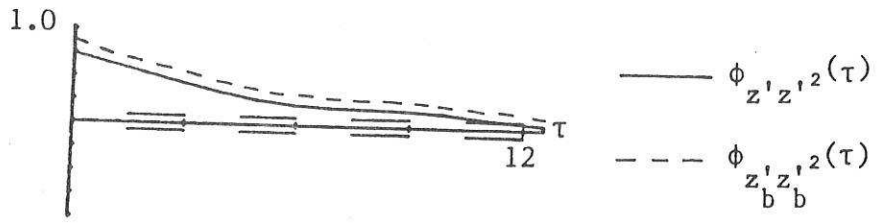

Figure 4. Structure detection for the Wiener model

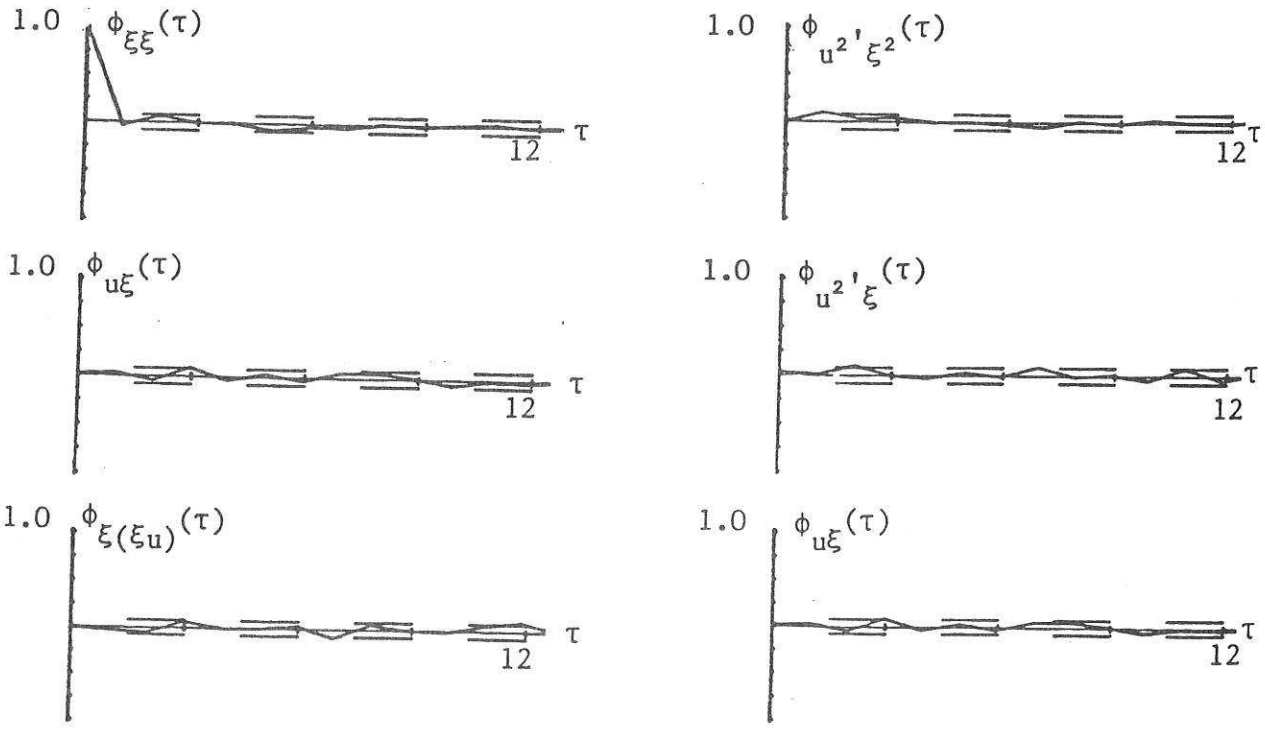

Figure 5. Model validation for the Wiener model

A full analysis of the liquid level system and the heat exchanger is available in the literature [Billings, Tsang, Voon 1985, Billings and Fadzil 1985]. In the identification of both these systems linear models were estimated prior to entry into the stepwise regression routine.

7. Conclusions

A combined stepwise regression and prediction error algorithm for nonlinear systems has been described. The stepwise regression routine detects the significant terms in a NARMAX model description whilst the prediction error algorithm provides optimised estimates of the model parameters which have properties very similar to maximum likelihood estimates. The algorithm when combined with model validation tests provides a powerful procedure for fitting parsimonious models to non-linear systems.

8. Acknowledgements

One of the authors (SAB) gratefully acknowledges financial support for the work presented above from SERC grant GR/B/31163.

9. References

Akaike, H. (1974). A new look at statistical model identification; IEEE Trans. Auto. Cont., AC-19, pp.716-723.

Akaike, H. (1977). On the entropy maximisation principle; Proc. Symp. on Applications of Statistics. Editted by P.R.Krishnalah, North Holland, pp.27-41.

Bazaraa, M.S., Snetty, C.M. (1979). Nonlinear programming theory and algorithms; Wiley.

Billings, S.A. (1980). Identification of nonlinear systems - a survey; Proc.IEE, Part D, 127, pp.272-285.

Billings, S.A., Fadzil, M.B. (1984). Identification of nonlinear difference equation model of an industrial diesel generator; Research Report 246, Univ. of Sheffield.

Billings, S.A., Fadzil, M.B. (1985). The practical identification of nonlinear systems; 7th IFAC Symp. Ident. & Syst. Par. Est., York, pp.155-160.

Billings, S.A., Tsang, K.M., Voon, W.S.F. (1985). Identification and control of a nonlinear liquid level system (in preparation).

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, pp.193-199.

Billings, S.A., Voon, W.S.F. (1984). Least squares parameter estimation algorithms for nonlinear systems; Int. J. Systems Sci., 15, pp.601-615.

Draper, N.R., Smith, H. (1981). Applied Regression Analysis, Wiley.

Efroymson, M.A. (1961). Multiple regression analysis; in "Mathematical Methods for Digital Computers", Ed. by A. Ralston, H. S. Wilf, Wiley.

Goodwin, G.C., Payne, R.L. (1977). Dynamic system identification; Experiment Design and Data Analysis, Academic Press.

Hall, W.E., Narendra, K.G., Tyler, J.S. (1975). Model structure determination and parameter identification for nonlinear aerodynamic flight regimes; Acard Conf. Proc. No. 178, Methods for Aircraft State and Parameter Identification.

Harris, C.J., Billings, S.A. (1985). Self tuning and adaptive control: theory and applications; Peter Peregrinus, 2nd Edition.

Klein, V., Batterson, J.G., Murphy, P.C. (1981). Determination of airplane model structure from flight data by using modified stepwise regression; NASA Tech. Paper 1916.

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. Contr., 41, pp.303-344.

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

Papoulis, A. (1965). Probability, Random Variables and Stochastic Processes; McGraw-Hill.

Marmarelis, P.Z., Marmarelis, V.Z. (1978). Analysis of physiological systems - the white noise approach; Plenum Press.

Smillie, K.W. (1966). An introduction to regression and correlation; Academic Press.