Monograph:
Billings, S.A. and Zhu, Q.M. (1992) A Structure Detection Algorithm for Nonlinear Dynamic Rational Models. Research Report. Acse Report 439 . Dept of Automatic Control and System Engineering. University of Sheffield

A Structure Detection Algorithm for Nonlinear Dynamic Rational Models

S.A. Billings and Q.M. Zhu

Department of Automatic Control
and Systems Engineering

University of Sheffield

Mappin Street

Sheffield S1 4DU

U. K.

Research Report No. 439

March 1992

A Structure Detection Algorithm for Nonlinear Dynamic Rational Models

S.A. Billings, Q.M. Zhu

Department of Automatic Control and Systems Engineering,

University of Sheffield, Sheffield S1 4DU, UK

Abstract:

A general structure detection and parameter estimation algorithm is derived for the identification of nonlinear systems which can be approximated by a stochastic nonlinear rational model defined as the ratio of two polynomial expansions of past inputs, outputs, and noise sequences. The algorithm includes an intelligent structure detection module which learns the structure of the model from the input output data. Simulation results are included to illustrate the application of the new algorithms.

1 Introduction

Determining the model structure or which terms to include in a model is an important part of any system identification procedure. Parameter estimation becomes straightforward once the terms to regress upon are known. But in practice the model structure is seldom given a priori. This is not a major problem if the system under investigation is linear because the total number of possible model terms is manageable. For example the total number of terms in a linear ARMAX (AutoRegressive Moving Average model with eXogenous inputs) is just the sum of the number of lags in the input, output, and noise terms and the common practice is to increase the number of lagged variables in the model until an adequate fit is obtained. This can be tedious but because the search space is relatively small an acceptable combination of model terms can usually be found without too much difficulty.

If the system is nonlinear the increase in dimensionality suggests that this brute force approach is no longer practical. Just a simple polynomial NARMAX (Nonlinear ARMAX) model with four lagged inputs, outputs, and noise model terms expanded as a third degree polynomial will for example contain 455 terms. The number of terms in nonlinear models therefore increases rapidly as the lag and degree of nonlinear expansion are increased. This is a severe problem because in practical identification the experimenter is forced to search over a wide range of lags and degrees of nonlinear expansion to ensure that the best model is identified. Identification algorithms which

detect the model structure as an integral part of the parameter estimation procedures are therefore vital if concise accurate models of nonlinear systems are sought.

The forward regression orthogonal estimation algorithm developed by Billings, Korenberg, and Chen (1988) offers a possible solution to this problem. The algorithm which was developed for polynomial NARMAX models breaks the m-dimensional estimation problem down into m one-dimensional stages. This ensures that the method is computationally simple and can be applied to very complex models without inducing numerical problems. Structure detection is included by incorporating an error reduction ratio (ERR) as an inherent part of the procedure to provide a measure of the importance of each candidate model term to the overall model fit. Combining these ideas yields an algorithm which can optimally search through a library of model terms applying the same procedure to each and which sorts the terms in the order of importance to the output variance and provides estimates of the unknown model parameters in the presence of coloured possibly nonlinear noise.

The objective of the present paper is to extend these ideas to nonlinear rational models. The superior approximation properties of rational functions are well known in static curve fitting and approximation theory and it is therefore natural to try and exploit these advantages in dynamic nonlinear system identification. The main problem with this idea is that the dynamic rational model is highly nonlinear in the parameters and consequently only a few authors have considered parameter estimation based on this description (Marquardt 1963, Billings and Chen 1989a). All these algorithms however are complex and difficult to apply in practice. These disadvantages can be avoided if a linear in the parameter expansion is adopted but this involves a totally new formulation of the orthogonal estimator to overcome the bias which this induces even for white additive noise. Furthermore unless the model structure is known a priori, very unlikely, the impracticallity of searching over many model sets which was discussed above for the polynomial NARMAX model remains

These problems must be resolved if the superior approximation potential of nonlinear dynamic rational models is to be exploited and the present study introduces a new parameter estimation and intelligent structure detection algorithm as one possible solution. It is shown that the new algorithm provides both unbiased estimates of the unknown parameters and a measure of the significance of each candidate model term to the output variance in the presence of unknown additive and multiplicative coloured noise. The concept of intelligent structure detection is used to determine the optimal cut off for term inclusion. This is a vital part of the algorithm because the cut off is a function of the data set and must therefore be learnt during the identification. The

combined algorithm therefore adjusts both the terms in the model and the parameter estimates to yield a concise model description. Simulated results are included to demonstrate the performance of the new algorithms.

2 Rational model

A parameterized stochastic nonlinear rational model can be expressed as a ratio of two polynomials,

$$y(t) = \frac{a(y(t-1), \cdots, y(t-r), u(t-1), \cdots, u(t-r), e(t-1), \cdots, e(t-r))}{b(y(t-1), \cdots, y(t-r), u(t-1), \cdots, u(t-r), e(t-1), \cdots, e(t-r))} + e(t)$$

$$= \frac{a(t)}{b(t)} + e(t) = \frac{\sum_{j=1}^{num} p_{nj}(t)\theta_{nj}}{\sum_{j=1}^{den} p_{dj}(t)\theta_{dj}} + e(t) \qquad (2.1)$$

where $u(t)$ and $y(t)$ represent the input and output at time t, $e(t)$ is an unobservable independent noise sequence with zero mean and finite variance σ_e^2, and $a(t)$ and $b(t)$ are defined as the numerator polynomial and denominator polynomial respectively. Eqn (2.1) is a strictly causal expression which relates the past inputs and outputs to the current output. The rational model can be classified as a subset of the NARMAX model set. The rational model is distinguished from the polynomial NARMAX model because it is nonlinear in the parameters (Zhu and Billings 1991b).

The rational model has been recognised as an excellent model to approximate a wide range of linear and nonlinear systems with an arbitrary accuracy (Sontag 1979, Braess 1980). It has been proved that the rational model structure is a very general representation (Chen and Billings 1989) which includes many varieties of models as subsets such as the ARMAX model, polynomial NARMAX model, Volterra series, and output affine model. The rational model is mathematically easy to manipulate and provides concise model descriptions.

Studies of rational models have been mainly divided into two aspects, model structure characteristics and model identification. Sontag's (1979) and Braess' (1986) results concentrate on the model characteristics such as convergence of the approximation, observability, realizability, and minimality. Billings and Chen's (1989a) contribution was to expand the model into a linear in the parameters expression to facilitate identification of the related output affine model and Zhu and Billings (1990) gave a criterion to test rational model stability. Identification based on the rational model is a

much more difficult problem than for other types of nonlinear models which are naturally linear in the parameters such as the Volterra and polynomial NARMAX models. The nonlinear least squares algorithm of (Marquardt (1963) and the prediction error algorithm of Billings and Chen (1989a) are available, but the former assumes noise free data and both are highly computational demanding, thus inhabiting applications to realistic identification problems. The main difficulty is that the rational model is nonlinear in the inputs, outputs, noise, and especially parameters (Zhu and Billings 1991b).

Alternatively the rational model structure can be multiplied out to yield a linear in the parameters expression. This appears to significantly simplify the identification problem. For the unrealistic noise free case, the algorithms developed for the polynomial NARMAX model can then be applied directly. However measured data from real systems will always involve some noise contamination and this will induce inherent errors in the denominator terms which will cause severe bias in the parameter estimates. Billings and Zhu (1991) and Zhu and Billings (1991a, b) have shown how this bias can be removed and this forms the starting point for the current study.

Ford, Titterington, and Kitsos (1989) recently surveyed the application of static nonlinear rational models in chemical kinetics and other related fields. A rational model was used by Box and Hunter (1965) to describe a gaseous methane and absorbed oxygen chemical reaction of the form

$$\frac{\theta_1 \theta_3 x_1}{1 + \theta_1 x_1 + \theta_2 x_2} \tag{2.2}$$

The famous Brunauer-Emmet-Teller equation in catalysis also takes the form of a rational model (Hill and Hunter 1974, Khuri 1984)

$$\frac{\theta_1 \theta_2 x(1 - x)}{1 + (\theta_2 - 1)x} \tag{2.3}$$

Pritchard and Bacon (1977) used a rational model to approximate the process of isomerization of n-pentane

$$\frac{\theta_1 \theta_3 (x_2 - x_3/1.632)}{1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_3} \tag{2.4}$$

Other rational models were given by Behnken (1964) for the reactivity ratio in copolymers and by Currie (1982) for the Michaelis-Menten enzyme kinetic function. In function approximation, Braess (1986) used static rational models to approximate certain typical nonlinear functions such as e^x, e^{-x}, $\sqrt{x}$, and $|x|$. All these models however are static and the application of rational models with dynamic terms such as eqn (2.1) is

rare despite the obvious advantages in terms of approximating complex phenomena.

3 Parameter estimation

The rational model can be expanded as a linear in the parameters expression by multiplying $b(t)$ on both sides of eqn (2.1) and then moving all the terms except $y(t)p_{d1}(t)\theta_{d1}$ to the right hand side to give

$$\begin{aligned} Y(t) &= a(t) - y(t)\sum_{j=2}^{den} p_{dj}(t)\theta_{dj} + b(t)e(t) \\ &= \sum_{j=1}^{num} p_{nj}(t)\theta_{nj} - \sum_{j=2}^{den} y(t)p_{dj}(t)\theta_{dj} + \zeta(t) \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} Y(t) &= y(t)p_{d1}(t)|_{\theta_{d1}=1} \\ &= p_{d1}(t)\frac{a(t)}{b(t)} + p_{d1}(t)e(t) \end{aligned} \tag{3.2}$$

Alternatively divide all the right hand side terms by θ_{d1} and redefine symbols to give essentially $\theta_{d1} = 1$. It should be noticed that the strictly causality of eqn (2.1) is lost when it is expanded as a linear in the parameters expression of eqn (3.1), because the denominator terms on the right hand side in eqn (3.1) include the current output $y(t)$ as factors. The third term on the right hand side in eqn (3.1) is given by

$$\begin{aligned} \zeta(t) &= b(t)e(t) \\ &= (\sum_{j=1}^{den} p_{dj}(t)\theta_{dj})e(t) \\ &= p_{d1}(t)e(t) + (\sum_{j=2}^{den} p_{dj}(t)\theta_{dj})e(t) \end{aligned} \tag{3.3}$$

where

$$E[\zeta(t)] = E[b(t)]E[e(t)] = 0 \tag{3.4}$$

providing $e(t)$ has been reduced to an uncorrelated sequence.

Alternatively eqn (3.1) can be expressed as

$$\begin{aligned} Y(t) &= \sum_{j=1}^{num} p_{nj}(t)\theta_{nj} - \sum_{j=2}^{den} y(t)p_{dj}(t)\theta_{dj} + b(t)e(t) \\ &= \sum_{j=1}^{num} p_{nj}(t)\theta_{nj} - \sum_{j=2}^{den} \frac{a(t)}{b(t)} p_{dj}(t)\theta_{dj} + p_{d1}(t)e(t) \end{aligned} \tag{3.5}$$

Although the term $\frac{a(t)}{b(t)}p_{dj}(t)$ in eqn (3.5) cannot be obtained directly the expression is very useful in the analysis of bias and the derivation of the new estimator.

Eqn (3.1) may be written in vector notation as

$$\begin{aligned} Y(t) &= \phi(t)\Theta + \zeta(t) \\ &= \hat{\phi}(t)\Theta + p_{d1}(t)e(t) \end{aligned} \tag{3.6}$$

where

$$\phi(t) = [\phi_n(t) \;\; \phi_d(t)]$$

$$= [p_{n1}(t) \cdots p_{nnum}(t) \; -p_{d2}(t)y(t) \cdots -p_{dden}(t)y(t)]$$

$$= [p_{n1}(t) \cdots p_{nnum}(t) \; -p_{d2}(t)(\frac{a(t)}{b(t)} + e(t)) \cdots -p_{dden}(t)(\frac{a(t)}{b(t)} + e(t))] \tag{3.7}$$

$$\Theta^T = [\Theta_n \; \Theta_d]$$

$$= [\theta_{n1} \cdots \theta_{nnum} \; \theta_{d2} \cdots \theta_{dden}] \tag{3.8}$$

and

$$\hat{\phi}(t) = [\phi_n(t) \;\; \hat{\phi}_d(t)]$$

$$= [p_{n1}(t) \cdots p_{nnum}(t) \; -p_{d2}(t)\frac{a(t)}{b(t)} \cdots -p_{dden}(t)\frac{a(t)}{b(t)}] \tag{3.9}$$

Notice that the matrix $\hat{\phi}(t)$ cannot be obtained directly because $\frac{a(t)}{b(t)}$ cannot be measured.

3.1 Bias problem

When a rational model is multiplied out as a linear in the parameters expression as in eqn (3.1), the term $y(t)p_{dj}(t)$ which is the product of the current output $y(t)$ and the denominator term $p_{dj}(t)$ will contain an element of the current noise $e(t)$. From eqn (2.1)

$$y(t)p_{dj}(t) = p_{dj}(t)\,\frac{a(t)}{b(t)} + p_{dj}(t)e(t) \tag{3.1.1}$$

where $p_{dj}(t)\,\frac{a(t)}{b(t)}$ represents the elements which are independent of $e(t)$ and $p_{dj}(t)e(t)$ represents the elements which involve the current noise term $e(t)$. Here $p_{dj}(t)e(t)$ will be referred to as an inherent error which can not be removed from $y(t)p_{dj}(t)$. This is the source of bias which results if linear least squares type estimators are applied directly. Unlike linear and polynomial nonlinear models this bias exists even for the

case of white noise corruption. For example a polynomial NARMAX model with no denominator terms, (ie $b(t) = 1$), is a linear in the parameter model but in this case there are no inherent error terms.

To show the bias problem associated with the rational model eqn (3.1), consider eqn (3.6). The least squares parameter estimate is given by

$$\hat{\Theta} = [\Phi^T\Phi]^{-1}\,\Phi^T\vec{Y} \tag{3.1.2}$$

where both the normal matrix $\Phi^T\Phi$ and the correlation vector $\Phi^T\vec{Y}$ are biased. An unbiased rational model estimator (RME) was proposed by Billings and Zhu (1991) as

$$\begin{aligned}\Theta &= [\Phi^T\Phi - \sigma_e^2\,\Psi_{ls}]^{-1}\,[\Phi^T\vec{Y} - \sigma_e^2\,\psi_{ls}] \\ &= [[\Phi^T\Phi]_{(t-1)}]^{-1}\,[\Phi^T\vec{Y}]_{(t-1)}\end{aligned} \tag{3.1.3}$$

See Appendix I for further details.

3.2 Orthogonal parameter estimation

The RME algorithm presented in section 3.1 solves the unbiased parameter estimation problem, but it does not give an efficient routine for model structure detection. The derivation of an orthogonal rational model estimation (ORME) routine which provides the basis for both structure detection and parameter estimation is briefly summarised below

Consider an orthogonal transformation of the original model eqn (3.6)

$$\begin{aligned}Y(t) &= w(t)G + \zeta(t) \\ &= w(t)G + b(t)e(t) \\ &= \sum_{j=1}^{num} w_{nj}(t)g_{nj} + \sum_{j=2}^{den} w_{dj}(t)g_{dj} + b(t)e(t)\end{aligned} \tag{3.2.1}$$

where

$$\begin{aligned}G &= [G_n \;\; G_d] \\ &= [g_{n1}, \cdots, g_{nnum}, g_{d2}, \cdots, g_{dden}]\end{aligned} \tag{3.2.2}$$

and

$$\begin{aligned}w(t) &= [w_n(t) \;\; w_d(t)] \\ &= [ww_n(t) \;\; ww_d(t)] + [e_n(t) \;\; e_d(t)]e(t) \\ &[ww_{n1}(t), \cdots, ww_{nnum}(t), ww_{d2}(t), \cdots, ww_{dden}(t)] \\ &+ [e_{n1}(t), \cdots, e_{nnum}(t), e_{d2}(t), \cdots, e_{dden}(t)]e(t)\end{aligned} \tag{3.2.3}$$

where $[e_n(t) \;\; e_d(t)]e(t)$, the inherent error in the orthogonal transformation, represents all terms which include the factor $e(t)$ and $[ww_n(t) \;\; ww_d(t)]$ represents all other terms which may include lagged noise terms $e(t-j)$, $j > 0$.

The orthogonal regression matrix W is defined

$$W = \Phi \, T^{-1} \tag{3.2.4}$$

Where Φ is given in eqn (3.1.3) and

$$W^T = [w^T(1) \cdots w^T(N)]$$

$$= \begin{bmatrix} w_n^T(1) & . \; . \; . & w_n^T(N) \\ w_d^T(1) & . \; . \; . & w_d^T(N) \end{bmatrix} \tag{3.2.5}$$

The orthogonal transform matrix T is unit upper triangular

$$T = \begin{bmatrix} 1 & t_{1\,2} & . & . & . & t_{1\,n} \\ 0 & 1 & . & . & . & . \\ 0 & 0 & 1 & . & . & . \\ . & . & 0 & 1 & . & . \\ . & . & . & 0 & 1 & t_{n-1\,n} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.2.6}$$

There are several methods of computing the elements of T, such as Gram-Schmidt, Householder, and Givens, transformations.

The orthogonality of the matrix W yields

$$W^T \, W = D \tag{3.2.7}$$

where D is a positive diagonal matrix

$$D = diag \; \{d_{n1}, \cdots, d_{nnum}, d_{d2}, \cdots, d_{dden}\} \tag{3.2.8}$$

and

$$d_{*j} = \sum_{t=1}^{N} w_{*j}(t) \; w_{*j}(t) \tag{3.2.9}$$

where from eqn (3.2.3)

$$\frac{d_{*j}}{N} = \frac{\sum_{t=1}^{N} w_{*j}(t) \; w_{*j}(t)}{N} = \overline{ww_{*j}^2(t)} + \overline{e_{*j}^2(t)} \; \sigma_e^2 \tag{3.2.10}$$

where the over bar denotes time averaging and * denotes either n or d.

The unbiased parameter vector G can now be obtained from

$$\begin{aligned} G &= [W^T W - \sigma_e^2 \Psi_{orth}]^{-1} [W^T \vec{Y} - \sigma_e^2 \psi_{orth}] \\ &= [D - \sigma_e^2 \Psi_{orth}]^{-1} [W^T \vec{Y} - \sigma_e^2 \psi_{orth}] \\ &= [[W^T W]_{(t-1)}]^{-1} [W^T \vec{Y}]_{(t-1)} \end{aligned} \tag{3.2.11}$$

where $\vec{Y}$ is defined in eqn (3.1.3) and

$$\begin{aligned} [W^T W]_{(t-1)} &= [W^T W - \sigma_e^2 \Psi_{orth}] \\ [W^T \vec{Y}]_{(t-1)} &= [W^T \vec{Y} - \sigma_e^2 \psi_{orth}] \end{aligned} \tag{3.2.12}$$

All terms involving $e(t)$ appear in $\sigma_e^2 \Psi_{orth}$ and $\sigma_e^2 \psi_{orth}$ which are called error terms and the subscript $(t-1)$ indicates that only lagged noise terms (eg $e(t-j)\ j \geq 1$) are present.

$$\Psi_{orth} = diag\ \{\overline{e_{n1}^2}, \cdots, \overline{e_{nnum}^2}, \overline{e_{d2}^2}, \cdots, \overline{e_{dden}^2}\}$$

$$\psi_{orth} = [\overline{p_{d1}(t)e_{n1}(t)}, \cdots, \overline{p_{d1}(t)e_{nnum}(t)}, \overline{p_{d1}(t)e_{d2}(t)}, \cdots, \overline{p_{d1}(t)e_{dden}(t)}]^T \tag{3.2.13}$$

Notice that the orthogonal estimator of eqn (3.2.11) is related to the least squares estimator given in eqn (3.1.3). Inspection of the orthogonal estimator shows that the noise variance σ_e^2 is needed a priori. An estimate of σ_e^2 can be obtained by an iterative procedure (Billings and Zhu 1991, Zhu and Billings 1991b) in which parameter estimation and noise variance prediction are recursively updated.

The parameter vector Θ of the original model eqn (3.6) can then be calculated by

$$T\ \Theta = G \tag{3.2.14}$$

or

$$\Theta = T^{-1}\ G \tag{3.2.15}$$

Full details of the ORME algorithm were given in Zhu and Billings (1991b).

4 Structure detection

A major advantage of the estimation algorithm presented in section 3 is the orthogonal property and the capability of detecting the model structure. The polynomial NARMAX model is a subset of the rational model and both are specific realizations of the NARMAX model. Therefore a new formulation to detect model structure should be developed for the rational model such that this simplifies to the polynomial NARMAX model (Korenberg, Billings, Liu, and McIlroy 1988) as a special case. The error reduction ratio (ERR), which is computed as a by-product of the orthogonal

estimation algorithm, can be used as a criterion to select model terms. However like parameter estimation for the rational model, the problem of inherent bias must be addressed by starting from first principles and devising a means of eliminating the bias terms. Consider eqn (3.2.1), squaring this with the assumption that the signals are egodic gives

$$\begin{aligned}\vec{Y}^T \vec{Y} &= G^T W^T W G + 2 W G \vec{\zeta}^T + \vec{\zeta}^T \vec{\zeta} \\ &= G^T G D + 2 W G \vec{\zeta}^T + \vec{\zeta}^T \vec{\zeta}\end{aligned} \tag{4.1}$$

where

$$\vec{\zeta} = [\zeta(1), \cdots, \zeta(N)]^T = [b(1)e(1), \cdots, b(N)e(N)]^T \tag{4.2}$$

$$\frac{\vec{Y}^T \vec{Y}}{N} = \overline{Y^2(t)} = \sigma_Y^2 \tag{4.3}$$

$$\begin{aligned}\frac{G^T G D}{N} &= \sum_{j=1}^{num} g_{nj}^2 \, \overline{w_{nj}^2(t)} + \sum_{j=2}^{den} g_{dj}^2 \, \overline{w_{dj}^2(t)} \\ &= \sum_{j=1}^{num} g_{nj}^2 \, \overline{(ww_{nj}(t) + e_{nj}(t)e(t))^2} + \sum_{j=2}^{den} g_{dj}^2 \, \overline{(ww_{dj}(t) + e_{dj}(t)e(t))^2} \\ &= \sum_{j=1}^{num} g_{nj}^2 \, (\overline{ww_{nj}^2(t)} + \overline{e_{nj}^2(t)} \, \sigma_e^2) + \sum_{j=2}^{den} g_{dj}^2 \, (\overline{ww_{dj}^2(t)} + \overline{e_{dj}^2(t)} \, \sigma_e^2)\end{aligned} \tag{4.4}$$

$$\begin{aligned}\frac{W G \vec{\zeta}^T}{N} &= \sum_{j=1}^{num} g_{nj} \, \overline{w_{nj}(t)b(t)e(t)} + \sum_{j=2}^{den} g_{dj} \, \overline{w_{dj}(t)b(t)e(t)} \\ &= \sum_{j=1}^{num} g_{nj} \, \overline{(ww_{nj}(t) + e_{nj}e(t))b(t)e(t)} + \sum_{j=2}^{den} g_{dj} \, \overline{(ww_{dj}(t) + e_{dj}e(t))b(t)e(t)} \\ &= \sum_{j=1}^{num} g_{nj} \, \overline{e_{nj}(t)b(t)} \, \sigma_e^2 + \sum_{j=2}^{den} g_{dj} \, \overline{e_{dj}(t)b(t)} \, \sigma_e^2\end{aligned} \tag{4.5}$$

$$\frac{\vec{\zeta}^T \vec{\zeta}}{N} = \overline{b^2(t)} \;\; \overline{e^2(t)} = \sigma_b^2 \, \sigma_e^2 \tag{4.6}$$

Substituting eqns (4.3) to (4.6) into eqn (4.1) and then multiplying by $\frac{1}{\sigma_Y^2\ \sigma_b^2}$ on both sides gives

$$\frac{1}{\sigma_b^2} = \sum_{j=1}^{num} \frac{g_{nj}^2\ \overline{ww_{nj}^2(t)} + g_{nj}^2\ \overline{e_{nj}^2(t)}\ \sigma_e^2 + 2g_{nj}\ \overline{e_{nj}(t)b(t)}\ \sigma_e^2}{\sigma_Y^2\ \sigma_b^2}$$

$$+ \sum_{j=2}^{den} \frac{g_{dj}^2\ \overline{ww_{dj}^2(t)} + g_{dj}^2\ \overline{e_{dj}^2(t)}\ \sigma_e^2 + 2g_{dj}\ \overline{e_{dj}(t)b(t)}\ \sigma_e^2}{\sigma_Y^2\ \sigma_b^2} + \frac{\sigma_e^2}{\sigma_Y^2} \qquad (4.7)$$

Define the error reduction ratio (ERR) as

$$\hat{err}_{nj} = \frac{g_{nj}^2\ \overline{ww_{nj}^2(t)} + g_{nj}^2\ \overline{e_{nj}^2(t)}\ \sigma_e^2 + 2g_{nj}\ \overline{e_{nj}(t)b(t)}\ \sigma_e^2}{\sigma_Y^2\ \sigma_b^2}$$

$$\hat{err}_{dj} = \frac{g_{dj}^2\ \overline{ww_{dj}^2(t)} + g_{dj}^2\ \overline{e_{dj}^2(t)}\ \sigma_e^2 + 2g_{dj}\ \overline{e_{dj}(t)b(t)}\ \sigma_e^2}{\sigma_Y^2\ \sigma_b^2} \qquad (4.8)$$

Introduce

$$err_{nj} = \frac{g_{nj}^2\ \overline{ww_{nj}^2(t)}}{\sigma_Y^2\ \sigma_b^2}$$

$$err_{dj} = \frac{g_{dj}^2\ \overline{ww_{dj}^2(t)}}{\sigma_Y^2\ \sigma_b^2} \qquad (4.9)$$

as the ERR estimates that would arise if $e(t) = 0$, and

$$Bias\ [\ err_{nj}\ ] = \frac{g_{nj}^2\ \overline{e_{nj}^2(t)}\ \sigma_e^2 + 2g_{nj}\ \overline{e_{nj}(t)b(t)}\ \sigma_e^2}{\sigma_Y^2\ \sigma_b^2}$$

$$Bias\ [\ err_{dj}\ ] = \frac{g_{dj}^2\ \overline{e_{dj}^2(t)}\ \sigma_e^2 + 2g_{dj}\ \overline{e_{dj}(t)b(t)}\ \sigma_e^2}{\sigma_Y^2\ \sigma_b^2} \qquad (4.10)$$

as the biases which are induced in the ERR estimates for the realistic case of $e(t) \neq 0$.

An unbiased estimate of ERR for the rational model can therefore be estimated using

$$err_{nj} = \hat{err}_{nj} - Bias\ [\ err_{nj}\ ]$$

$$err_{dj} = \hat{err}_{dj} - Bias\ [\ err_{dj}\ ] \qquad (4.11)$$

where err_{nj}, $Bias\ [\ \hat{err}_{nj}\ ]$, err_{dj}, and $Bias\ [\ \hat{err}_{dj}\ ]$ are obtained directly from the computations.

With reference to the definitions in eqns (4.8) to (4.10), eqn (4.7) can alternatively be written as

$$\frac{1}{\sigma_b^2} = \sum_{j=1}^{num} \hat{err}_{nj} + \sum_{j=2}^{den} \hat{err}_{dj} + \frac{\sigma_e^2}{\sigma_Y^2}$$

$$= \sum_{j=1}^{num} err_{nj} + \sum_{j=2}^{den} err_{dj} + \sum_{j=1}^{num} bias\ [\ err_{nj}\ ] + \sum_{j=2}^{den} bias\ [\ err_{dj}\ ] + \frac{\sigma_e^2}{\sigma_Y^2} \qquad (4.12)$$

Enq (4.12) can be used as a criterion for determining the number of terms to be included in the model, it therefore determines the model structure. The larger the value of ERR associated with a specific term the more the ratio $\frac{\sigma_e^2}{\sigma_Y^2}$ would be reduced if that term were included in the model. Hence terms can be ordered based upon that ERR value. Insignificant terms can be rejected by defining a cut off value of $1 - \sum err_{*j}$ below which terms are deemed to be negligible. As a criterion ERR attempts to balance the prediction accuracy and complexity of the final model.

There are several alternative term selection methods, most of these including a stepwise regression algorithm and a log determinant ratio test have been studied respectively by Billings and Voon (1984) and Leontaritis and Billings (1987).

4.1 Properties of ERR for polynomial NARMAX models

It is informative at this stage to consider the properties of the simplified case of ERR term selection for the polynomial NARMAX model before extending these ideas to the more complex rational model. Notice that the NARMAX model is just a subset of eqn (2.1) with $b(t)$ the denominator set to 1.

4.1.1 Bias and ERR

ERR estimated directly for the polynomial NARMAX model is unbiased.

To show this, consider the definition given in eqn (4.8) derived for the rational model. Setting $b(t) = 1$ to yield a polynomial NARMAX model shows that $e_{*j}(t) = 0$, and eqn (4.8) becomes

$$\hat{err}_{nj} = \frac{g_{nj}^2\ \overline{ww_{nj}^2(t)} + g_{nj}^2\ \overline{e_{nj}^2(t)}\ \sigma_e^2 + 2g_{nj}\ \overline{e_{nj}(t)b(t)}\ \sigma_e^2}{\sigma_Y^2\ \ \sigma_b^2}$$

$$= \frac{g_{nj}^2\ \overline{ww_{nj}^2(t)}}{\sigma_Y^2}$$

$$= \frac{g_{nj}^2\ \overline{w_{nj}^2(t)}}{\sigma_Y^2} = err_{nj} \qquad (4.1.1)$$

which is unbiased according to the definition given in eqn (4.9).

4.1.2 Independence of ERR values for process model and noise model terms

The selection of process model terms is independent of the selection of noise model terms when $b(t) = 1$.

To show this, consider a polynomial NARMAX model, $b(t) = 1$,

$$y(t) = a(t) + e(t) = \sum_{j=1}^{num} p_j(t)\theta_j + e(t)$$

$$= \sum_{j=1}^{m_p} p_j(t)\theta_j + \sum_{i=m_p+1}^{m_n} p_i(t)\theta_i + e(t)$$

$$= \sum_{j=1}^{m_p} p_j(t)\theta_j + \eta(t) \tag{4.1.2}$$

where

$$\sum_{process} = \sum_{j=1}^{m_p} p_j(t)\theta_j \tag{4.1.3}$$

is defined as the process model and

$$\sum_{noise} = \sum_{i=m_p+1}^{m_n} p_i(t)\theta_i \tag{4.1.4}$$

which includes all terms that involve $e(t-j)$, $j \geq 0$ is defined as the noise model. The terms in eqn (4.1.4) will in general be correlated with the terms in the process model and may not be zero mean. From eqn (4.1.2) and eqn (4.1.4)

$$\eta(t) = \sum_{i=m_p+1}^{m_n} p_i(t)\theta_i + e(t) \tag{4.1.5}$$

Squaring eqn (4.1.2) and taking expected values gives

$$\overline{y^2(t)} = \sum_{j=1}^{m_p} g_j^2 \overline{w_j^2(t)} + 2 \sum_{j=1}^{m_p} g_j \overline{w_j(t)\eta(t)} + \overline{\eta^2(t)} \tag{4.1.6}$$

or alternatively

$$1 = \frac{\sum_{j=1}^{m_p} g_j^2 \overline{w_j^2(t)} + 2 \sum_{j=1}^{m_p} g_j \overline{w_j(t)\eta(t)} + \overline{\eta^2(t)}}{\overline{y^2(t)}} \tag{4.1.7}$$

That is

$$1 - \sum_{process} err_j = \frac{2 \sum_{j=1}^{m_p} g_j \overline{w_j(t)\eta(t)} + \overline{\eta^2(t)}}{\overline{y^2(t)}} \tag{4.1.8}$$

where

$$err_j = \frac{g_j^2 \overline{w_j^2(t)}}{\overline{y^2(t)}} = \frac{g_j^2 \overline{w_j^2(t)}}{\sigma_y^2} \tag{4.1.9}$$

This result shows that err_j values computed for the process model is unbiased and is not affected by the noise $\eta(t)$.

The parameter estimates are also unbiased because

$$\begin{aligned} g_j &= \frac{\overline{w_j(t)y(t)}}{\overline{w_j^2(t)}} \\ &= \frac{\overline{w_j(t)(\,a(t) + e(t)\,)}}{\overline{w_j^2(t)}} \\ &= \frac{\overline{w_j(t)a(t)} + \overline{w_j(t)e(t)}}{\overline{w_j^2(t)}} \\ &= \frac{\overline{w_j(t)a(t)}}{\overline{w_j^2(t)}} \end{aligned} \tag{4.1.10}$$

4.1.3 Forward independence of the ERR computation

Another important property of ERR term selection is that the computation of err_i values is independent of err_j's when $j>i$.

To show this, consider err_i and err_j, $j > i$, and where

$$err_i = \frac{g_i^2\ \overline{w_i(t)^2}}{\overline{Y^2(t)}}, \quad err_j = \frac{g_j^2\ \overline{w_j(t)^2}}{\overline{Y^2(t)}} \tag{4.1.11}$$

The computations for g_i and $w_i(t)$ are independent of g_j and $w_j(t)$ because at the *ith* step $w_j(t)$ has not been selected. Therefore the computation of err_i is independent of err_j's when $j>i$. However the computation of err_j is dependent on the err_i's when $j>i$ because the computations for g_j and $w_j(t)$ depend on g_i and $w_i(t)$. Consequently the computation of ERR is forward independent and backward dependent.

4.1.4 Convergence of the ERR sequence

This states that the ERR sequence is convergent in terms of summation and

$$0 < err_j \leq 1 - \frac{\sigma_e^2}{\sigma_y^2} \tag{4.1.12}$$

To show this, consider the first term on the right hand of eqn (4.12). Simplifying the notations as $\hat{efr}_{nj} = err_{nj} = err_j$, $\sigma_b^2 = 1$ and $\hat{efr}_{dj} = 0$ ($b(t) = 1$), the summation of the err sequence is given by

$$\sum_{j=1}^{num} err_j = 1 - \frac{\sigma_e^2}{\sigma_y^2}, \quad \frac{\sigma_e^2}{\sigma_y^2} < 1 \tag{4.1.13}$$

Hence the summation of a number of finite ERR values converges to $1 - \frac{\sigma_e^2}{\sigma_y^2}$. For the noise free case the summation equals to unity

$$\sum_{j=1}^{num} err_j = 1 \tag{4.1.14}$$

From eqn (4.1.1)

$$0 < err_j \tag{4.1.15}$$

and eqn (4.1.13)

$$err_j \leq 1 - \frac{\sigma_e^2}{\sigma_y^2} \tag{4.1.16}$$

So that

$$0 < err_j \leq 1 - \frac{\sigma_e^2}{\sigma_y^2} \tag{4.1.17}$$

where the equality holds only for the model with one term.

In summary, the ERR technique can be directly applied for polynomial NARMAX model term selection in the presence of arbitrary noise. The ERR values are unbiased irrespective of whether the noise model is known or not. The computation of ERR in the process model is totally independent of the computation for the noise model. These results show that term selection for process model and noise model terms can be decoupled.

4.2 Properties of ERR for the rational NARMAX model

For the rational model, ERR for the process and noise terms can not be independent because the ERR computations for the process terms depends on a knowledge of the noise sequence. Properties of ERR term selection for the rational model are considered below.

4.2.1 Independence of the numerator model from the denominator model

The selection of the numerator model terms is independent of the denominator model terms but only if all of the former are selected first.

This follows because numerator terms do not include the inherent error and so bias is not induced provided no denominator terms are selected. The parameter estimates associated with the numerator terms are also unbiased for the same reason. Forcing the algorithm to select numerator terms only however is suboptimal and hence decoupling the numerator and denominator estimation is not recommended. Correct estimates can only be obtained if the full algorithm is implemented.

4.2.2 Forward independence of the ERR computation

The computation of err_i is independent of err_j's when $j>i$.

This is a direct corollary from the property proved for the polynomial NARMAX model. The ERR computation for the rational model is also forward independent and backward dependent.

4.2.3 Convergence of err sequence

The summation of the ERR values for the rational model is given by

$$\sum err = \frac{1}{\sigma_b^2} - \sum bias[\ err\ ] - \frac{\sigma_e^2}{\sigma_Y^2} \tag{4.2.1}$$

Hence the summation of a finite number of ERR values converges to $\frac{1}{\sigma_b^2} - \sum bias[\ err\ ] - \frac{\sigma_e^2}{\sigma_Y^2}$. For the polynomial NARMAX model the summation reduces to

$$\sum err = 1 - \frac{\sigma_e^2}{\sigma_y^2} \tag{4.2.2}$$

as given in eqn (4.1.14)

Also from ERR definition of eqn (4.9)

$$0 < err_{*j} \tag{4.2.3}$$

and eqn (4.2.1)

$$err_{*j} \leq \frac{1}{\sigma_b^2} - \sum bias[\ err\ ] - \frac{\sigma_e^2}{\sigma_Y^2} \tag{4.2.4}$$

So that

$$0 < err_{*j} \leq \frac{1}{\sigma_b^2} - \sum bias[\ err\ ] - \frac{\sigma_e^2}{\sigma_Y^2} \tag{4.2.5}$$

where the equality holds only for a model with one term.

In summary the new ERR technique can be applied to the rational model in the presence of arbitrary noise, but now each ERR value depends on the removal of the bias terms according to eqn (4.11).

4.3 Intelligent structure detection (ISD)

Now consider the choice of the cut off point for the ERR computation. The effect of selecting a term is to reduce the noise variance σ_e^2, therefore the cut off point may be expressed as a function of σ_e^2. Let the cut off point be *COP* then

$$COP = f(\ \sigma_e^2\ ) \tag{4.3.1}$$

where $f(.)$ denotes some function. The cut off point is very important because it affects both the prediction accuracy and complexity of the final model. When the *COP* is chosen to be too large the model will be inadequate because of missing terms and poor accuracy. On the other hand if the *COP* is chosen to be too small the model will include many redundant terms and numerical problems may be induced.

Because the cut off point is a function of the noise variance or noise floor the cut off value used for one data set will be inappropriate for another. But based on the analysis at the beginning of this section the optimal cut off points can be derived.

For the polynomial model

$$COP \geq \frac{\sigma_e^2}{\sigma_Y^2} \tag{4.3.2}$$

and for the rational model, from eqn (4.12)

$$COP \geq 1 - \frac{1}{\sigma_b^2} + \sum bias[\ err\ ] + \frac{\sigma_e^2}{\sigma_Y^2} \tag{4.3.3}$$

Obviously eqn (4.3.2) is a subset of eqn (4.3.3) obtained by setting $b(t)=1$ so that *bias*[*err*]=0. Intelligent structure detection which was originally introduced for the polynomial NARMAX model (Billings and Chen 1989b) uses the expression (4.3.2) to learn the optimal cut off for each separate data set. Eqn (4.3.3) shows that to implement this idea for the rational model an iterative procedure is required to estimate the noise variance σ_e^2 and σ_b^2 both of which are unknown. The following algorithm is

proposed for the rational model.

(1) Fit a deterministic rational model (i.e. no noise model initially) and compute $\hat{\sigma}_e^2$ as an initial estimate. Predict the residual sequence and hence obtain an estimate of the noise sequence $e(t)$.

(2) Weight the sequence $\hat{\sigma}_e^2$ by 0.1, 0.25, 0.5, 0.75, 0.95 and set a constant *COP* which can be determined by trial in advance at iterations 2, 3, 4, 5, and 6 respectively. All subsequent iterations use a weight of 1.0. This is necessary because $\hat{\sigma}_e^2$ and *COP* tend in the early iterations to be an over estimate of the noise and are therefore incorrect because initially the model is biased. Choose *COP* using eqn (4.3.3) when the weighting sequence for $\hat{\sigma}_e^2$ approaches 1.

(3) Do a search over the specified full stochastic rational model to select significant terms according to the ERR values and estimate the unknown parameters based on the residual sequence, $\hat{\sigma}_e^2$, and $\hat{\sigma}_b^2$ obtained from the last iteration.

(4) Repeat steps 2 and 3 until the computations converge, the preset maximum number of the iterations are exceeded, or a specified number of terms are included in the model.

5 Simulation studies

Three simulation examples were selected to demonstrate the structure detection and parameter estimation algorithm.

Example S_1 consisted of the rational model

$$y(t) = \frac{a(t)}{b(t)} + e(t) = \frac{y(t-1)y(t-2) + u^2(t-1) + u(t-1)e(t-2)}{1 + y^2(t-1) + y^2(t-2)} + e(t) \qquad (5.1)$$

where the model outputs were generated based upon an independent and uniformly distributed input with amplitude range ± 1 (variance $\sigma_u^2 = 0.33$) and an independent zero mean Gaussian noise with variance $\sigma_e^2 = 0.01$. The input and output data length of 1000 are shown in Figure 1.1, and the states for $t<1$ were set to zero.

The linear in the parameters expression for this model is

$$\begin{aligned} Y(t) = {} & y(t-1)y(t-2) + u^2(t-1) + u(t-1)e(t-2) \\ & - y(t)y^2(t-1) - y(t)y^2(t-2) + b(t)e(t) \end{aligned} \qquad (5.2)$$

where the dependent variable is

$$Y(t) = y(t) \qquad (5.3)$$

An overspecified full model consisting of 56 terms was used to start the search with *numerator degree* = *denominator degree* = 2 and *input lag* = *output lag* = *noise lag* = 2. With reference to Table 1.1, during each iteration, a search over the full model set was performed to select the terms with significant ERR values. At the first iteration the input and output data were modelled using a deterministic model because no knowledge of the noise was assumed.

From the second iteration to the seventh iteration, the sequence $\hat{\sigma}_e^2$ as weighted by 0.1, 0.25, 0.5, 0.75, 0.95, and 1.0 respectively (see step 2 of the algorithm). A fixed cut off point with a value of 0.001 was used at this stage because the optimal cut off eqn (4.3.3) will be not accurate until $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$ have converged. From iterations 8 to 10, the *COP* of eqn (4.3.3) was used and the model shows that the algorithm has, from no a priori information, correctly determined the model structure and the unknown parameters. The one step ahead predictions and residuals are illustrated in Figure 1.2. The model validity tests are shown in Figure 1.3 and because all the tests are within the relevant 95% confidence bands this confirms that the model is adequate and the parameters are unbiased.

Example S_2 consisted of the complicated output affine model

$$y(t) = \frac{a(t)}{b(t)} + e(t)$$

$$= \frac{0.5y^3(t-1) - 0.1u^2(t-1)u(t-2) + u(t-2) + e(t-2)}{1 + u^2(t-1) + 0.7u^3(t-1)} + e(t) \tag{5.4}$$

where the model outputs were generated using the same specifications as given for example S_1. The input and output sequences are shown in Figure 2.1.

The linear in the parameters expression for this model is

$$Y(t) = 0.5y^3(t-1) - 0.1u^2(t-1)u(t-2) + u(t-2) + e(t-2)$$
$$- y(t)u^2(t-1) - 0.7y(t)u^3(t-1) + b(t)e(t) \tag{5.5}$$

with the dependent variable

$$Y(t) = y(t) \tag{5.6}$$

The full model consisting of 168 terms was specified with *numerator degree* = *denominator degree* = 3 and *input lag* = *output lag* = *noise lag* = 2. Once again, the algorithm successfully identified the model with correct term selection and unbiased parameter estimates. The specifications were a maximum of seven model terms, 0.001

for the initial cut off point for iterations 1 to 7 with the *COP* computed using eqn (4.3.3) thereafter, and 0.1, 0.25, 0.5, 0.75, 0.95, and 1.0 were the weights for $\hat{\sigma}_e^2$ for the iterations 2 to 7. The final model obtained at iteration ten is shown in Table 2.1. The one step ahead predictions and residuals are illustrated in Figure 2.2. The model validations are shown in Figure 2.3.

Example S_3 consisted of the polynomial model (Billings and Chen 1989b)

$$y(t) = a(t) + e(t)$$

$$= 0.5y(t-1) + u(t-2) + 0.1u^2(t-1) + 0.5e(t-1) + 0.2u(t-1)e(t-2) + e(t) \qquad (5.7)$$

where the model outputs were generated by an independent and uniformly distributed input with amplitude range $\pm\sqrt{3}$ (variance $\sigma_u^2 = 1.0$) and an independent zero mean Gaussian noise with variance $\sigma_e^2 = 0.04$. The data sets of length 500 are shown in Figure 3.1.

It has been proved that the polynomial model is a subset of the rational model and the orthogonal rational model estimator (ORME) will reduce to the orthogonal polynomial model estimator (Zhu and Billings 1992). The purpose of selecting this example is to demonstrate that the structure detection does correctly discard all the rational model denominator terms.

The full model consisting of 56 terms was used to start the search with *numerator degree* = *denominator degree* = 2 and *input lag* = *output lag* = *noise lag* = 2. The same specifications were set as for the first two examples. An initial cut off point of 0.032 was used based on the experimental results given in Billings and Chen (1989b). With eight iterations, a polynomial model was detected as shown in Table 3.1. The one step ahead predictions and residuals are illustrated in Figure 3.2 and the model validations are shown in Figure 3.3. All these results confirm the theories presented in Billings and Zhu (1991) and Zhu and Billings (1992) and are consistent with those obtained by Billings and Chen (1989b).

6 Conclusions

Fitting models to data is easy if you know which terms to regress upon. This information is rarely available to the experimenter and a combined intelligent structure detection and parameter estimation algorithm has been presented as one possible solution to this problem for the class of nonlinear rational models.

Acknowledgements

The authors gratefully acknowledge that this work is supported by SERC under grant GR/F2417.7.

Appendix I

To show the bias problem associated with the rational model eqn (3.1), consider the vector expression eqn (3.6)

$$Y(t) = \phi(t)\Theta + \zeta(t) \tag{A.1}$$

The least squares parameter estimate is given by

$$\hat{\Theta} = [\Phi^T\Phi]^{-1}\,\Phi^T\vec{Y} \tag{A.2}$$

where

$$\Phi^T = [\phi^T(1) \cdots \phi^T(N)]$$

$$= \begin{bmatrix} \phi_n^T(1) & \cdots & \phi_n^T(N) \\ \phi_d^T(1) & \cdots & \phi_d^T(N) \end{bmatrix}$$

$$= \begin{bmatrix} p_{n1}(1) & \cdot & p_{n1}(N) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ p_{nnum}(1) & \cdot & p_{nnum}(N) \\ -p_{d2}(1)(\dfrac{a(1)}{b(1)} + e(1)) & \cdot & -p_{d2}(N)(\dfrac{a(N)}{b(N)} + e(N)) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ -p_{dden}(1)(\dfrac{a(1)}{b(1)} + e(1)) & \cdot & -p_{dden}(N)(\dfrac{a(N)}{b(N)} + e(N)) \end{bmatrix}$$

$$\vec{Y} = [Y(1) \cdots Y(N)]^T \tag{A.3}$$

It is clear from eqns (2.1), (3.1), and (3.6) that Φ may include lagged noise model terms where N is the data length. The normal matrix $\Phi^T\Phi$ and the correlation vector $\Phi^T\vec{Y}$ can be expressed as

$$\Phi^T\Phi = \begin{bmatrix} \sum_{t=1}^{N}\phi_n^T(t)\phi_n(t) & \sum_{t=1}^{N}\phi_n^T(t)\phi_d(t) \\ \sum_{t=1}^{N}\phi_d^T(t)\phi_n(t) & \sum_{t=1}^{N}\phi_d^T(t)\phi_d(t) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{t=1}^{N}\phi_n^T(t)\phi_n(t) & \sum_{t=1}^{N}\phi_n^T(t)\hat{\phi}_d(t) \\ \sum_{t=1}^{N}\hat{\phi}_d^T(t)\phi_n(t) & \sum_{t=1}^{N}\hat{\phi}_d^T(t)\hat{\phi}_d(t) \end{bmatrix} + \begin{bmatrix} 0 & \sum_{t=1}^{N} 0 \\ 0 & \sigma_e^2\sum_{t=1}^{N}p_d^T(t)p_d(t) \end{bmatrix}$$

and

$$\Phi^T\vec{Y} = \begin{bmatrix} \sum_{t=1}^{N}\phi_n^T(t)p_{d1}(t)\frac{a(t)}{b(t)} \\ \sum_{t=1}^{N}\hat{\phi}_d^T(t)p_{d1}(t)\frac{a(t)}{b(t)} \end{bmatrix} + \begin{bmatrix} \sum_{t=1}^{N} 0 \sum_{t=1}^{N} \\ \sigma_e^2\sum_{t=1}^{N}p_d^T(t)p_{d1}(t) \end{bmatrix} \qquad (A.4)$$

where

$$p_d(t) = [p_{d2}(t) \quad \cdots \quad p_{dden}(t)] \qquad (A.5)$$

and $\hat{\phi}_d(t)$ is defined in eqn (3.9)

Rewritting eqn (A.4) gives

$$\Phi^T\Phi = [\Phi^T\Phi]_{(t-1)} + \sigma_e^2\,\Psi_{ls}$$

$$\Phi^T\vec{Y} = [\Phi^T\vec{Y}]_{(t-1)} + \sigma_e^2\,\psi_{ls} \qquad (A.6)$$

where the definition of terms follows directly and

$$\Psi_{ls} = \begin{bmatrix} 0 & 0 \\ 0 & \sum_{t=1}^{N}p_d^T(t)p_d(t) \end{bmatrix} = \sum_{t=1}^{N}\rho^T(t)\rho(t) = \sum_{t=1}^{N}\Psi_{ls}(t)$$

$$\psi_{ls} = \begin{bmatrix} 0 \\ \sum_{t=1}^{N}p_d^T(t)p_{d1}(t) \end{bmatrix} = \sum_{t=1}^{N}\rho^T(t)p_{d1}(t) = \sum_{t=1}^{N}\psi_{ls}(t) \qquad (A.7)$$

and

$$\rho(t) = [0 \; p_d(t)] \qquad (A.8)$$

All terms involving $e(t)$ appear in $\sigma_e^2 \Psi_{ls}$ and $\sigma_e^2 \psi_{ls}$ which are called error terms and the subscript $(t-1)$ indicates that only lagged noise terms (eg $e(t-j)$ $j \geq 1$) are present.

Hence the estimate given in eqn (A.2) can be written as

$$\hat{\Theta} = [\Phi^T\Phi]^{-1} \Phi^T\vec{Y}$$

$$= [[\Phi^T\Phi]_{(t-1)} + \sigma_e^2 \Psi_{ls}]^{-1} [[\Phi^T\vec{Y}]_{(t-1)} + \sigma_e^2 \psi_{ls}] \qquad \text{(A.9)}$$

The two terms $\sigma_e^2 \Psi_{ls}$ and $\sigma_e^2 \psi_{ls}$ will cause bias even if the additive noise is white. An unbiased rational model estimator (RME) (Billings and Zhu 1991) can be developed by removing these terms

$$\Theta = [\Phi^T\Phi - \sigma_e^2 \Psi_{ls}]^{-1} [\Phi^T\vec{Y} - \sigma_e^2 \psi_{ls}]$$

$$= [[\Phi^T\Phi]_{(t-1)}]^{-1} [\Phi^T\vec{Y}]_{(t-1)} \qquad \text{(A.10)}$$

References

Behnken, D.W., "Estimation of copolymer reactivity ratios: an example of nonlinear estimation," Journal of polymer science, vol. 2, 1964.

Billings, S.A. and W.S.F. Woon, "Least squares parameter estimation algorithms for nonlinear systems," Int. J. Systems Sci., vol. 19, 1984.

Billings, S.A., M. Korenberg, and S. Chen, "Identification of nonlinear output affine systems using orthogonal least squares algorithm," Int. J. systems science, vol. 19, 1988.

Billings, S.A. and S. Chen, "Identification of nonlinear rational systems using a prediction error estimation algorithm," Int. J. Systems Sci., vol. 20, 1989a.

Billings, S.A. and S. Chen, "Extended model set, global data and threshold model identification of severely nonlinear systems," Int. J. control, vol. 50, 1989b.

Billings, S.A. and Q.M. Zhu, "Rational model identification using an extended least squares algorithm," Int. J. Control, vol. 54, 1991.

Box, G.E.P. and W.G. Hunter, "The experimental study of physical mechanisms," Technomatrics, vol. 7, 1965.

Braess, D., Nonlinear approximation theory, Springer Verlag, 1986.

Chen, S. and S.A. Billings, "Represntations of nonlinear systems: the NARMAX model," Int. J. Control, vol. 48, 1989.

Currie, D.J., "Estimating Michaelis-menten parameters: bias, variance and experimental design," Biometrics, vol. 38, 1982.

Ford, I., D.M. Titterington, and C.P. Kitsos, "Recent advances in nonlinear experimental design," Technometrics, vol. 31, 1989.

Hill, W.J. and W.G. Hunter, "Design of experiments for subsets of parameters," Technomatrics, vol. 16, 1974.

Khuri, A.I., "A note on D-optimal design for partially nonlinear regression models," Technomatric, vol. 26, 1984.

Korenberg, M., S.A. Billings, Y.P. Liu, and P.J. McIlroy, "Orthogonal parameter estimation algorithm for nonlinear stochastic systems," Int. J. Control, vol. 48, 1988.

Leontaritis, I.J. and S.A. Billings, "Model selection and validation methods for nonlinear systems," Int. J. Control, vol. 45, 1987.

Marquardt, D.W., "An algorithm for least squares estimation of nonlinear parameters," Journal of the society for industrial and applied mathematics, vol. 11, 1963.

Pritchard, D.J. and D.W. Bacon, "Accounting for heteroscedasticity in experiment design," Technomatrics, vol. 19, 1977.

Sontag, E.D., Polynomial response maps--Lecture notes in control and information sciences 13, Springer - Verlag, Berlin, 1979.

Zhu, Q.M. and S.A. Billings, "Stability of a class of nonlinear systems," Reseach report 385, Department of control engineering, University of Sheffiled, 1990.

Zhu, Q.M. and S.A. Billings, "Recursive parameter estimation for nonlinear rational models," Journal of Systems Engineering , vol. 1, 1991a.

Zhu, Q.M. and S.A. Billings, "Parameter estimator for stochastic nonlinear rational models," to be published in Int. J. Control, 1992.

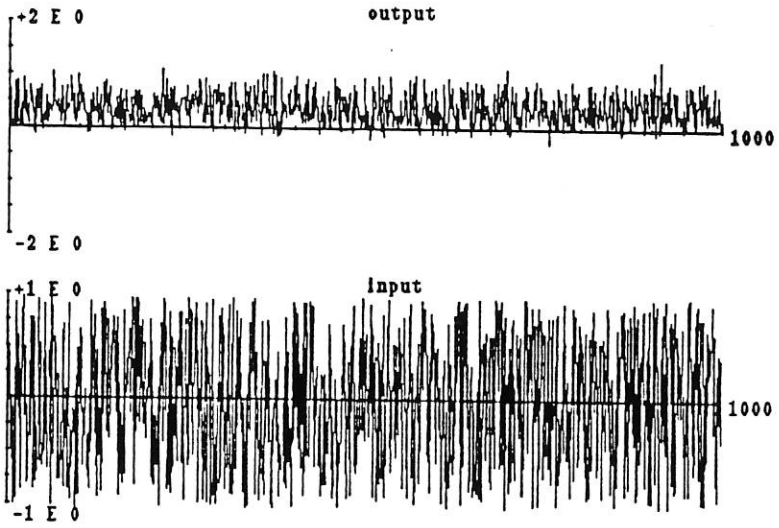

Figure 1.1 Input & output for example S_1

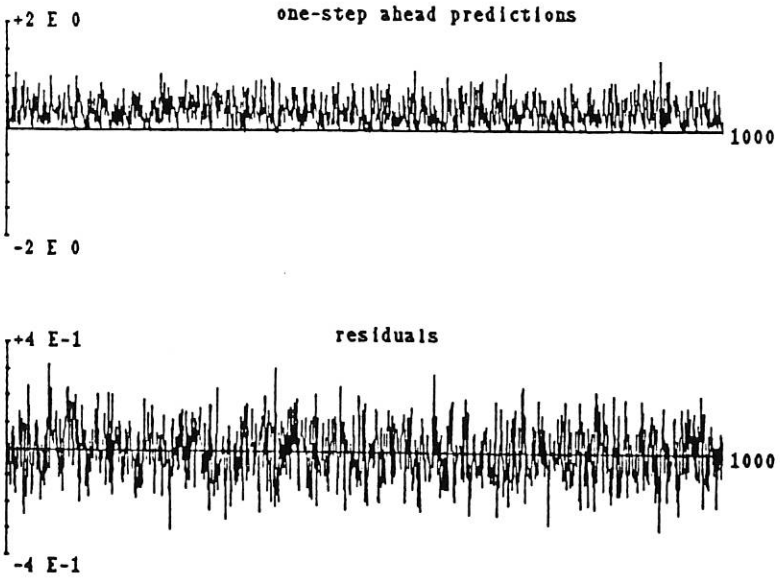

Figure 1.2 One step ahead predictions & residuals for example S_1

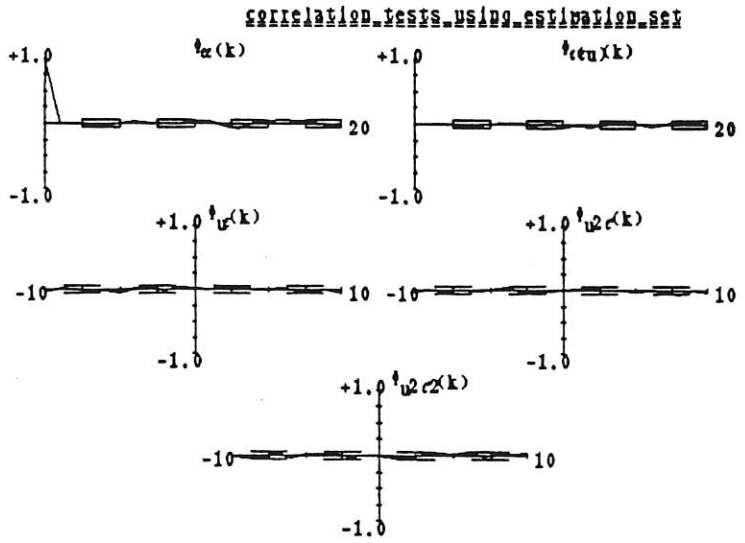

Figure 1.3 Model validation for example S_1 using correlation test

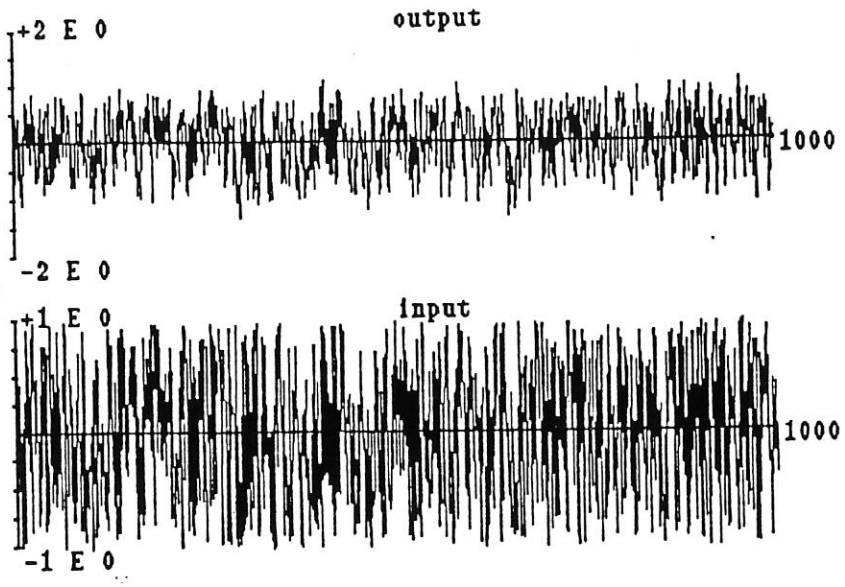

Figure 2.1 Input & output for example S_2

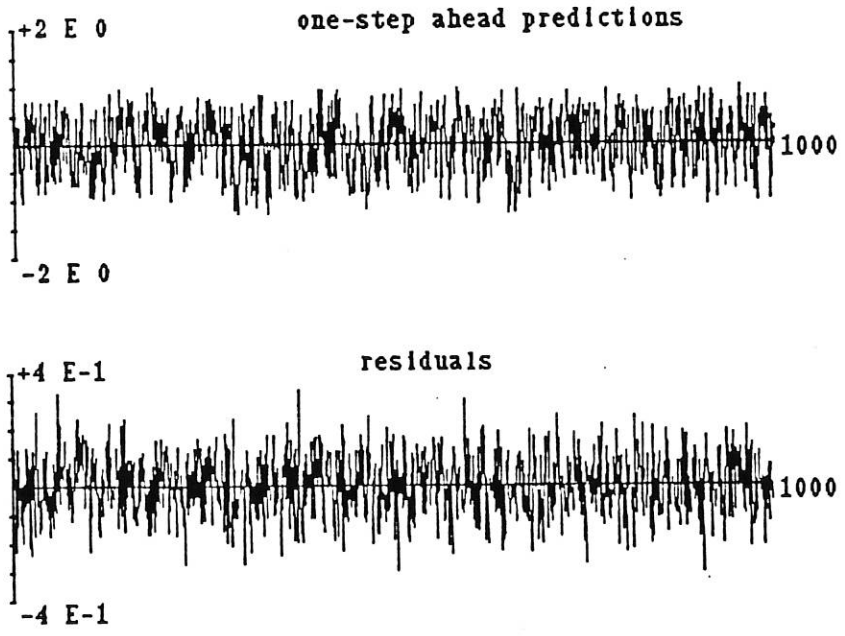

Figure 2.2 One step ahead predictions & residuals for example S_2

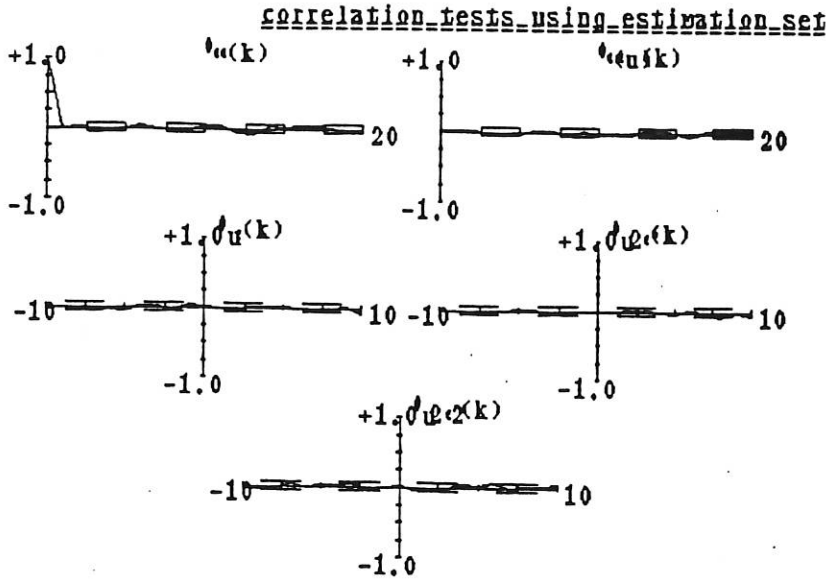

Figure 2.3 Model validation for example S_2 using correlation test

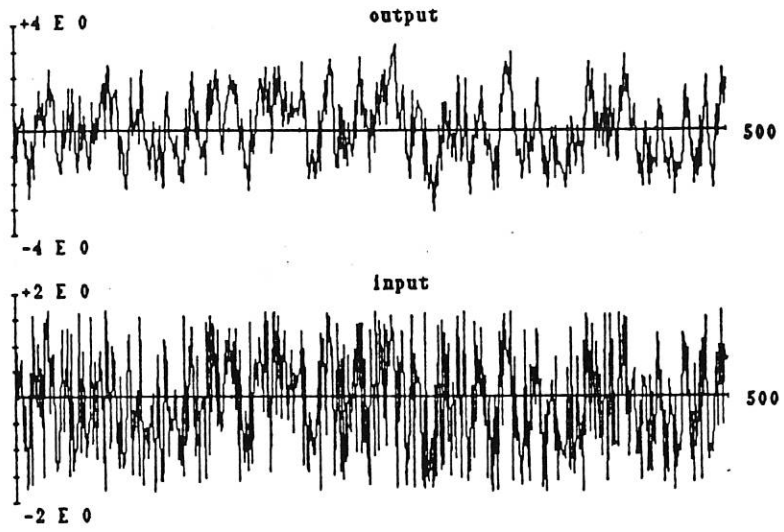

Figure 3.1 Input & output for example S_3

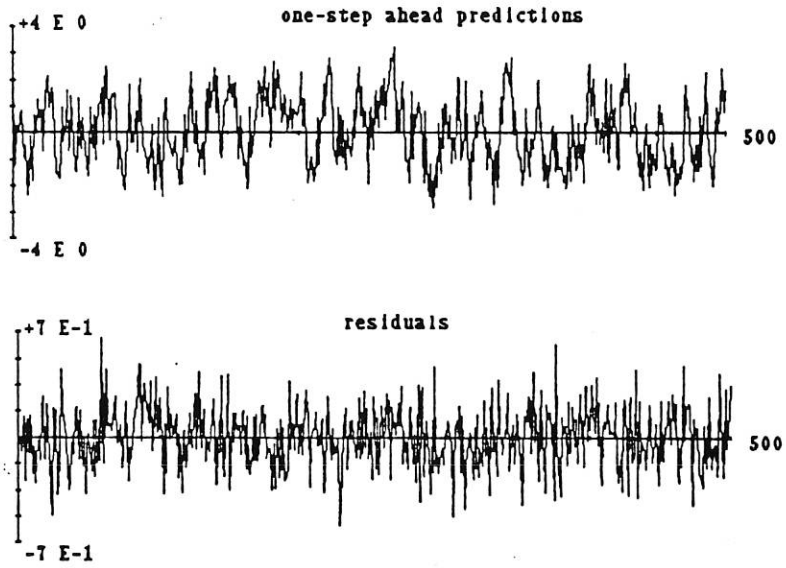

Figure 3.2 One step ahead predictions & residuals for example S_3

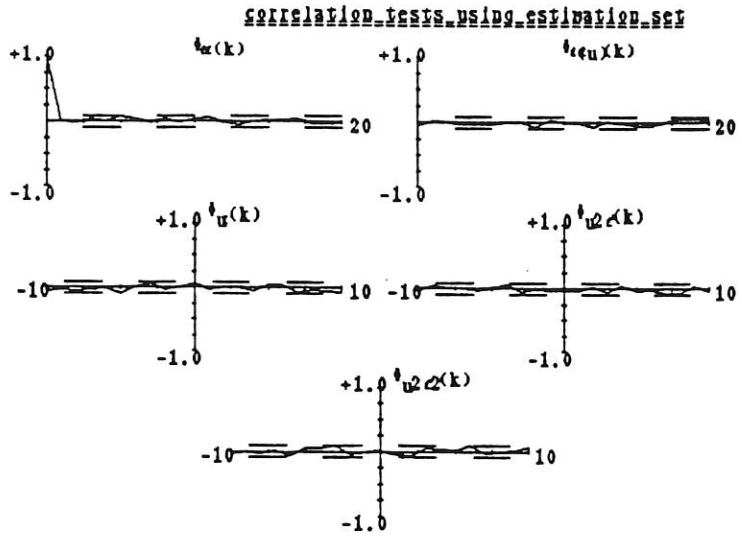

Figure 3.3 Model validation for example S_3 using correlation test

iterations	terms	estimates	e.r.r.s	st.de.s	o.s.	σ_e^2, σ_y^2, suggested *COP*, *COP*
1	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (1) constant= (3) y(t- 2)**1= denominator: (47) u(t- 1)**2*y(t)= (31) y(t- 2)**1*y(t)= (42) y(t- 2)**2*y(t)=	0.1202E+00 0.1809E+00 0.1302E+00 -0.1524E+00 0.6827E+00 0.9417E+00 -0.5969E+00	0.8942E+00 0.1688E-01 0.1328E-01 0.3822E-02 0.1268E-01 0.3071E-02 0.1295E-02	0.2611E-01 0.3097E-01 0.7230E-02 0.1905E-01 0.3310E-01 0.6963E-01 0.8782E-01	(1) (2) (4) (6) (3) (5) (7)	σ_e^2= 0.629319E-01 σ_y^2= 0.16100E+01 s. *COP*= 0.549715996 *COP*= 0.001000000
2	numerator: (19) u(t- 1)**2= denominator: (37) y(t- 1)**1*y(t- 2)**1*y(t)=	0.1108E+01 -0.2371E+01	0.8942E+00 0.1949E+00	0.2312E-01 0.1782E+00	(1) (2)	σ_e^2= 0.280715E-01 σ_y^2= 0.62702E+00 s. *COP*= 0.084426917 *COP*= 0.001000000
3	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (22) u(t- 1)**1*e(t- 2)**1= denominator: (36) y(t- 1)**2*y(t)= (42) y(t- 2)**2*y(t)= (46) y(t- 2)**1*e(t- 2)**1*y(t)= (35) e(t- 2)**1*y(t)=	0.9438E+00 0.8172E+00 0.2393E+00 -0.6979E+00 -0.6691E+00 0.8272E+00 -0.2855E+00	0.8942E+00 0.1688E-01 0.4452E-02 0.1162E-01 0.1910E-01 0.3999E-03 0.4992E-03	0.1457E-01 0.4870E-01 0.4367E-01 0.6519E-01 0.7448E-01 0.2796E+00 0.1231E+00	(1) (2) (5) (3) (4) (6) (7)	σ_e^2= 0.111684E-01 σ_y^2= 0.17539E+01 s. *COP*= 0.084717892 *COP*= 0.001000000
4	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (22) u(t- 1)**1*e(t- 2)**1= (10) y(t- 1)**1*u(t- 1)**1= denominator: (36) y(t- 1)**2*y(t)= (42) y(t- 2)**2*y(t)= (45) y(t- 2)**1*e(t- 1)**1*y(t)=	0.9332E+00 0.7748E+00 0.6991E+00 -0.2799E-01 -0.6654E+00 -0.5729E+00 0.2947E+00	0.8942E+00 0.1688E-01 0.1254E-01 0.1770E-03 0.9289E-02 0.1345E-01 0.1876E-03	0.1385E-01 0.4620E-01 0.6818E-01 0.1589E-01 0.6224E-01 0.6939E-01 0.2284E+00	(1) (2) (5) (6) (3) (4) (7)	σ_e^2= 0.104697E-01 σ_y^2= 0.10711E+01 s. *COP*= 0.046939328 *COP*= 0.001000000
5	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (22) u(t- 1)**1*e(t- 2)**1= (24) u(t- 2)**1*e(t- 1)**1= denominator: (36) y(t- 1)**2*y(t)= (42) y(t- 2)**2*y(t)= (52) u(t- 2)**1*e(t- 1)**1*y(t)=	0.9631E+00 0.9074E+00 0.7687E+00 -0.2541E+00 -0.8296E+00 -0.7906E+00 0.6982E+00	0.8942E+00 0.1688E-01 0.1551E-01 0.7341E-03 0.1290E-01 0.2241E-01 0.1643E-03	0.1469E-01 0.4937E-01 0.7302E-01 0.1301E+00 0.6650E-01 0.7468E-01 0.3197E+00	(1) (2) (5) (7) (3) (4) (6)	σ_e^2= 0.101248E-01 σ_y^2= 0.10816E+01 s. *COP*= 0.038297378 *COP*= 0.001000000
6	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (22) u(t- 1)**1*e(t- 2)**1= (1) constant= (8) y(t- 1)**2= denominator: (36) y(t- 1)**2*y(t)= (42) y(t- 2)**2*y(t)=	0.1025E+01 0.1110E+01 0.8682E+00 -0.1182E-01 0.3137E-01 -0.1147E+01 -0.1079E+01	0.8942E+00 0.1688E-01 0.2057E-01 0.9589E-03 0.1283E-02 0.1877E-01 0.3887E-01	0.2490E-01 0.6920E-01 0.8085E-01 0.1021E-01 0.3864E-01 0.1210E+00 0.9075E-01	(1) (2) (5) (6) (7) (3) (4)	σ_e^2= 0.994566E-02 σ_y^2= 0.11901E+01 s. *COP*= 0.026616283 *COP*= 0.001000000
7	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (22) u(t- 1)**1*e(t- 2)**1= (1) constant= denominator: (36) y(t- 1)**2*y(t)= (42) y(t- 2)**2*y(t)= (47) u(t- 1)**2*y(t)=	0.1169E+01 0.1264E+01 0.9847E+00 -0.2778E-01 -0.1254E+01 -0.1280E+01 -0.1455E+00	0.8942E+00 0.1688E-01 0.2218E-01 0.1702E-02 0.2027E-01 0.4320E-01 0.8500E-02	0.7671E-01 0.8842E-01 0.9851E-01 0.1274E-01 0.1039E+00 0.1146E+00 0.7994E-01	(1) (2) (5) (6) (3) (4) (7)	σ_e^2= 0.997107E-02 σ_y^2= 0.11286E+01 s. *COP*= 0.014265623 *COP*= 0.001000000
8	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (22) u(t- 1)**1*e(t- 2)**1= denominator: (36) y(t- 1)**2*y(t)= (42) y(t- 2)**2*y(t)=	0.1015E+01 0.1123E+01 0.8903E+00 -0.1112E+01 -0.1128E+01	0.8942E+00 0.1688E-01 0.2232E-01 0.2033E-01 0.4340E-01	0.1633E-01 0.5571E-01 0.8011E-01 0.7488E-01 0.8485E-01	(1) (2) (5) (3) (4)	σ_e^2= 0.992984E-02 σ_y^2= 0.84487E+00 s. *COP*= 0.019668126 *COP*= 0.015000000
9	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (22) u(t- 1)**1*e(t- 2)**1= denominator: (36) y(t- 1)**2*y(t)= (42) y(t- 2)**2*y(t)=	0.1013E+01 0.1122E+01 0.8994E+00 -0.1111E+01 -0.1121E+01	0.8942E+00 0.1688E-01 0.2271E-01 0.2020E-01 0.4301E-01	0.1627E-01 0.5555E-01 0.7990E-01 0.7467E-01 0.8451E-01	(1) (2) (5) (3) (4)	σ_e^2= 0.989752E-02 σ_y^2= 0.84422E+00 s. *COP*= 0.019605691 *COP*= 0.020000000
10	numerator: (19) u(t- 1)**2= (9) y(t- 1)**1*y(t- 2)**1= (22) u(t- 1)**1*e(t- 2)**1= denominator: (36) y(t- 1)**2*y(t)= (42) y(t- 2)**2*y(t)=	0.1012E+01 0.1118E+01 0.8977E+00 -0.1107E+01 -0.1115E+01	0.8942E+00 0.1688E-01 0.2254E-01 0.2010E-01 0.4271E-01	0.1624E-01 0.5544E-01 0.7991E-01 0.7453E-01 0.8433E-01	(1) (2) (5) (3) (4)	σ_e^2= 0.989646E-02 σ_y^2= 0.84426E+00 s. *COP*= 0.019720621 *COP*= 0.020000000

Table 1.1 Iterative model selection for example S_1

terms	estimates	e.r.r.s	st.de.s	o.s.
numerator:				
(5) u(t- 2)**1=	0.1016E+01	0.8063E+00	0.1276E-01	(1)
(29) y(t- 1)**3=	0.5145E+00	0.7801E-01	0.1905E-01	(2)
(66) u(t- 1)**2*u(t- 2)**1=	-0.7114E-01	0.3319E-01	0.6468E-01	(3)
(7) e(t- 2)**1=	0.9403E+00	0.2336E-01	0.5084E-01	(4)
denominator:				
(149) u(t- 1)**3*y(t)=	-0.7723E+00	0.1054E-01	0.5145E-01	(5)
(103) u(t- 1)**2*y(t)=	-0.1089E+01	0.2958E-01	0.1209E+00	(6)

Table 2.1 Selected model for example S_2

terms	estimates	e.r.r.s	st.de.s	o.s.
numerator:				
(5) u(t- 2)**1=	0.1004E+01	0.6829E+00	0.8980E-02	(1)
(2) y(t- 1)**1=	0.5063E+00	0.2717E+00	0.7222E-02	(2)
(19) u(t- 1)**2=	0.1054E+00	0.1248E-01	0.6623E-02	(3)
(6) e(t- 1)**1=	0.4900E+00	0.5522E-02	0.4585E-01	(4)
(22) u(t- 1)**1*e(t- 2)**1=	0.1713E+00	0.7272E-03	0.4501E-01	(5)

Table 3.1 Selected model for example S_3