[redacted]

AN ORTHOGONAL PARAMETER ESTIMATION ALGORITHM FOR NONLINEAR STOCHASTIC SYSTEMS

by

[†]M Korenberg

[*]S A Billings, PhD,BEng,CEng,MIEE,AFIMA,MInstMC

[*]Y P Liu

[†]Department of Electrical Engineering
Queens University, Ontario, Canada

[*]Department of Control Engineering
University of Sheffield
Sheffield S1 3JD, UK

Research Report No 307

March 1987

Abstract

An orthogonal parameter estimation algorithm is derived which allows each parameter in a nonlinear difference equation model to be estimated recursively and quite independently of the other parameters in the model. The algorithm can be applied for any persistently exciting input and provides both unbiased estimates in the presence of correlated noise and an indication of which terms to include in the model. Several simulated examples are included to demonstrate the effectiveness of the algorithm.

1. Introduction

The successful development of identification and controller design procedures for nonlinear systems critically depends upon the model which is used to represent the system under investigation. Traditionally the functional series descriptions of Volterra and Wiener have been used and an extensive literature describing the identification and analysis of such models exists [Billings, Gray and Owens 1984; Marmarelis and Marmarelis 1978; Schetzen 1980] . Unfortunately, functional series models require an excessive parameter set, often extending to over 500 kernel values, to describe even simple nonlinear systems and consequently few practical applications of the identification algorithms and virtually no controller design studies have been reported. However, by expanding the system output in terms of past inputs and outputs using a NARMAX model [Leontaritis and Billings 1985a,b] (Nonlinear AutoRegressive Moving Average Model with eXogenous inputs) a very concise representation for a wide class of nonlinear systems can be obtained which allevaites many of the problems associated with functional series methods. Several parameter estimation algorithms have been derived for the NARMAX model [Billings and Voon 1984, 1986a; Korenberg 1985] and it has been shown that provided the significant terms in the model can be detected models with less than ten terms are usually sufficient to capture the dynamics of highly nonlinear processes.

In the present study a new orthogonal parameter estimation algorithm is derived for stochastic nonlinear systems which can be represented by a NARMAX model. By introducing an auxiliary model defined such that the terms in the model are orthogonal over the data set [Korenberg 1985] for any input it is shown that each coefficient can be estimated recursively and quite independently of the other terms in the model in the presence of correlated measurement noise. Repeated application of this simple algorithm not only provides unbiased estimates of each coefficient in turn but also provides an indication of the contribution that each term makes to the output variance and this assists the user to detect the structure of the system under investigation and yields a parsimonious system model. Details of implementation including pretreatment of data, the input sensitivity problem, data segmentation and the interactive application of the orthogonal algorithm with nonlinear model validity tests are included together with numerous simulated examples for both linear and nonlinear systems.

2. The NARMAX Model

A wide class of nonlinear systems can be represented by the NARMAX model [Leontaritis and Billings 1985a,b] (Nonlinear AutoRegressive Moving Average model with eXogenous inputs)

$$y(t) = F^{\ell}[y(t-1),\dots y(t-N_y),u(t),\dots u(t-N_u),$$
$$\varepsilon(t-1),\dots\varepsilon(t-N_\varepsilon)] + \varepsilon(t) \qquad (1)$$

where u(t) and y(t) represent the measured input and output respectively, $\varepsilon(t)$ is the prediction error defined as

$$\varepsilon(t) = y(t) - \hat{y}(t)$$

where $E[\varepsilon(t)|y^{t-1},u^{t}] = 0$

$$y^{t-1} = (y(t-1),y(t-2),\dots y(1))^T \qquad (2)$$

$$u^{t} = (u(t),u(t-1),\dots u(1))^T$$

$$\hat{y}(t) = E[y(t)|y^{t-1},u^{t}]$$

N_u,N_y and N_ε represent the number of lags in the input, output and prediction error respectively, and $F^{\ell}[\cdot]$ is some nonlinear function. A time delay in the input and a dc level can easily be accommodated by rewriting eqn (1) as

$$y(t) = dc+F^{\ell}[y(t-1),\ldots y(t-N_y),u(t-d),\ldots u(t-d-N_u+1),\\ \varepsilon(t-1),\ldots\varepsilon(t-N_\varepsilon)] + \varepsilon(t) \qquad (3)$$

The NARMAX model of eqn (3) will be used in the present study and this can be shown to represent a large class of finite dimensional nonlinear systems [Leontaritis and Billings 1985a,b].

Expanding eqn (3) by defining the function $F^{\ell}[\cdot]$ to be a polynomial of degree ℓ gives the representation

$$y(t) = \sum_{m=0}^{M} \theta_m p_m(t) + \varepsilon(t) \qquad (4)$$

where $p_o(t) = 1$

$$p_m(t) = y(t-n_{y1})\ldots y(t-n_{yk})u(t-d+1n_{u1})\ldots u(t-d+1n_{uj})\\ \varepsilon(t-n_{\varepsilon 1})\ldots\varepsilon(t-n_{\varepsilon q})$$

for $m = 1,2,\ldots M$

$k \geq 0,\ j \geq 0,\ q \geq 0$

$$\begin{array}{l} 1 \leq n_{y1} \leq N_y \ \ldots\ 1 \leq n_{yk} \leq N_y \\ 0 \leq n_{u1} \leq N_u \ \ldots\ 0 \leq n_{uj} \leq N_u \\ 1 \leq n_{\varepsilon 1} \leq N_\varepsilon \ \ldots\ 1 \leq n_{\varepsilon q} \leq N_\varepsilon \end{array} \qquad (5)$$

and

$k = 0$ indicates that $p_m(t)$ contains no $y(\cdot)$ terms

$j = 0$ indicates that $p_m(t)$ contains no $u(\cdot)$ terms

$q = 0$ indicates that $p_m(t)$ contains no $\varepsilon(\cdot)$ terms

Notice that θ_o represents the dc value.

For example the NARMAX model

$$y(t) = dc+\theta_1 y(t-1)+\theta_2 u(t-1)+\theta_3 u(t-1)y(t-1)$$

$$+\theta_4 u(t-1)\varepsilon(t-1)+\theta_5\varepsilon(t-1)+\varepsilon(t) \qquad (6)$$

could be described by eqn (4) by defining

$$p_1(t) = y(t-1), p_2(t) = u(t-1), p_3(t) = u(t-1)y(t-1),$$

$$p_4(t) = u(t-1)\varepsilon(t-1),\ p_5(t) = \varepsilon(t-1),\ p_o(t) = 1,$$

$$\theta_o = dc$$

If N measurements of the input and output are available eqn (4) can be expressed in matrix form as

$$Y = P\theta + \underline{\varepsilon} \qquad (7)$$

where

$$Y^T = [y(1), y(2), \ldots y(N)]$$

$$\theta^T = [\theta_o, \theta_1, \ldots \theta_M]$$

$$\underline{\varepsilon}^T = [\varepsilon(1), \varepsilon(2), \ldots \varepsilon(N)]$$

$$P = \begin{pmatrix} P_o(1) & P_1(1) & \ldots & P_M(1) \\ P_o(2) & P_1(2) & \ldots & P_M(2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ P_o(N) & P_1(N) & & P_M(N) \end{pmatrix}$$

The parameter vector θ in eqn (7) could now be estimated using a least squares based or a prediction error routine [Billings and Voon 1986a]. However, by reformulating the problem such that each parameter in θ can be estimated independently using an orthogonal algorithm considerable advantages can be achieved [Korenberg 1985].

3. The Orthogonal Estimation Algorithm

The orthogonal estimation algorithm can be derived in two ways. Initially a simple derivation which is very transparent and amenable to computer simulation is formulated. This is then augmented by rederiving the algorithm using matrix notation.

Although the objective is to estimate the parameters θ_i, $i = 0,\ldots M$ in eqn (4) or eqn (7) the algorithm is formulated for the auxiliary model

$$y(t) = \sum_{m=o}^{M} g_m w_m(t) + \varepsilon(t) \qquad (8)$$

where $w_i(n)$, $i = 0,\ldots M$ are constructed to be orthogonal over the data record such that

$$\sum_{t=1}^{N} w_j(t) w_{k+1}(t) = 0 \qquad j = 0,1,\ldots k \qquad (9)$$

A family of orthogonal polynomials could be constructed by applying the Gram-Schmidt procedure but this can be shown to be very sensitive to rounding errors [Blum 1972]. A simpler scheme which makes use of the algebraic polynomial structure is the three-term recurrance method which can be adapted to the dynamic model of eqn (8) by defining

$$w_o(t) = p_o(t) = 1$$

$$w_m(t) = p_m(t) - \sum_{r=o}^{m-1} \alpha_{rm} w_r(t) \qquad m = 1,\ldots M \qquad (10)$$

and $$\alpha_{rm} = \sum_{t=1}^{N} p_m(t) w_r(t) / \sum_{t=1}^{N} w_r^2(t) \qquad 0 \leq r \leq m-1 \qquad (11)$$

Setting

$$\hat{g}_o = \frac{1}{N} \sum_{t=1}^{N} y(t) \qquad (12)$$

and using eqn (8) and the orthogonality of the $w_m(t)$ gives the parameter estimates

$$\hat{g}_m = \sum_{t=1}^{N} y(t)w_m(t) / \sum_{t=1}^{N} w_m^2(t) \tag{13}$$

Once the parameters g_m, $m = 0,1,\ldots M$ have been estimated using eqn (13) the parameters θ_m, $m = 0,1,\ldots M$ in the NARMAX model eqn (7) can be computed as

$$\hat{\theta}_m = \sum_{i=m}^{M} \hat{g}_i v_i \tag{14}$$

where

$$v_m = 1$$

$$v_i = \sum_{r=m}^{i-1} \alpha_{ri} v_r \qquad m < i \leq M \tag{15}$$

Notice that the prediction errors $\varepsilon(t)$ are not known a priori and must be estimated from eqn (8) as

$$\hat{\varepsilon}(t) = y(t) - \sum_{m=o}^{N} \hat{g}_m w_m(t) \tag{16}$$

The algorithm consists of the following steps

(i) Assume the prediction errors are zero and estimate all the parameters which do not include $\varepsilon(\cdot)$ terms using eqn's (10) to (13)

(ii) Estimate the prediction errors using eqn (16)

(iii) Using $\hat{\varepsilon}(t)$ estimate the parameters associated with prediction error terms using eqn's (10) to (13)

(iv) Go to (ii) and continue until convergence

(v) Estimate the NARMAX model coefficients using eqn's (14) and (15).

Convergence of the algorithm can be detected by monitoring parameter change

$$\sum_{m=o}^{M} \frac{|\hat{g}_m^{s+1} - \hat{g}_m^{s}|}{|g_m^{s+1}|}$$

at the (s+1)th iteration. The test value is typically chosen to be $10^{-3} - 10^{-5}$ and simulation has shown that convergence is achieved in typically ten iterations.

The orthogonal estimation is easy to implement yet offers considerable advantages compared with existing parameter estimation routines. Because of the orthogonality property the parameter vector can be estimated by computing each parameter g_j, $j = 0,...M$ one at a time. This simplifies the implementation, allows additional terms to be added to the model without the need to re-estimate the parameter vector, and allows the estimation of the process and noise parameters to be decoupled. Consequently, the estimates in step (i) will not change or be affected by the inclusion of noise terms in step (iii).

Notice that for any non-negative integer $Q<M$ the mean square error

$$\sum_{t=1}^{N} (y(t) - \sum_{m=0}^{Q} g_m w_m(t))^2 \qquad (17)$$

is minimised by the estimates in eqn (13). This shows that if the right hand side of eqn (8) is truncated by including only terms m = 0 to m = Q, the NARMAX model terms corresponding to $\sum_{m=0}^{Q} g_m w_m(t)$ will similarly minimise the mean square error for this value of Q.

3.1. Matrix Formulation

Although it is probably easier to implement the algorithm using the formulation given above the derivation of the results can be tidied up considerably by using a matrix notation.

The matrix decomposition theorem [Fox 1964] states that a positive definite square matrix A can be decomposed as

$$A = LDU \tag{18}$$

in which L and U are unit lower and upper triangular, D is diagonal with all positive elements and the expression is unique. If the additional constraint that A is symmetric is imposed then it is easily shown that

$$L = U^T \tag{19}$$

and hence eqn (18) becomes

$$A = U^T DU \tag{20}$$

Since the correlation matrix $P^T P$ associated with the model eqn (7) is symmetric and positive definite the matrix decomposition theorem shows that this can be expressed as

$$P^T P = T^T DT \tag{21}$$

where T is a unit upper triangular matrix and D is diagonal with all positive elements. Since $T^{-1}T = I$ the model of eqn (7) can now be written in the form of the auxiliary model

$$Y = P(T^{-1}T)\theta + \underline{\varepsilon}$$

or
$$Y = Wg + \underline{\varepsilon} \tag{22}$$

where
$$W = PT^{-1} \quad , \quad g = T\theta \tag{23}$$

and W is an orthogonal matrix

$$\begin{aligned} W^T W &= (PT^{-1})^T (PT^{-1}) = (T^{-1})^T (P^T P) T^{-1} \\ &= (T^T)^{-1} T^T D^T T T^{-1} = D \end{aligned} \tag{24}$$

Defining W as

$$W = \begin{pmatrix} w_o(1) & w_1(1) & \cdots & w_M(1) \\ w_o(2) & w_1(2) & & w_M(2) \\ \vdots & \vdots & & \vdots \\ w_o(N) & w_1(N) & & w_M(N) \end{pmatrix} \tag{25}$$

specifies D eqn (24) as

$$D = \begin{bmatrix} \sum_{t=1}^{N} w_o^{\,2}(t) & & & \bigcirc \\ & \sum_{t=1}^{N} w_1^{\,2}(t) & & \\ & & \ddots & \\ \bigcirc & & & \sum_{t=1}^{N} w_M^{\,2}(t) \end{bmatrix} \qquad (26)$$

and clearly $D^{-1} = \mathrm{diag} \dfrac{1}{\sum_{t=1}^{N} w_i^{\,2}(t)} \qquad i = 0,1,\ldots M.$

From eqn (23) $\qquad W = PT^{-1}$

and premultiplying by W^T and postmultiplying by T gives

$$W^TWT = W^TP$$

$$\text{or} \quad T = (W^TW)^{-1}W^TP \qquad (27)$$

$$= D^{-1}W^TP$$

Using the definitions of P eqn (7), W eqn (25) and D eqn (26) shows that eqn (27) can be expressed as

$$T = \begin{bmatrix} \dfrac{\sum_{t=1}^{N} w_o(t)p_o(t)}{\sum_{t=1}^{N} w_o^{\,2}(t)} & \dfrac{\sum_{t=1}^{N} w_o(t)p_1(t)}{\sum_{t=1}^{N} w_o^{\,2}(t)} & \cdots & \dfrac{\sum_{t=1}^{N} w_o(t)p_M(t)}{\sum_{t=1}^{N} w_o^{\,2}(t)} \\ \vdots & & & \vdots \\ \dfrac{\sum_{t=1}^{N} w_M(t)p_o(t)}{\sum_{t=1}^{N} w_M^{\,2}(t)} & \dfrac{\sum_{t=1}^{N} w_M(t)p_1(t)}{\sum_{t=1}^{N} w_M^{\,2}(t)} & \cdots & \dfrac{\sum_{t=1}^{N} w_M(t)p_M(t)}{\sum_{t=1}^{N} w_M^{\,2}(t)} \end{bmatrix}$$

$$= \begin{pmatrix} \alpha_{00} & \alpha_{01} & \cdots & \alpha_{0M} \\ \alpha_{10} & \alpha_{11} & & \alpha_{1M} \\ \vdots & \vdots & & \vdots \\ \alpha_{M0} & \alpha_{M1} & \cdots & \alpha_{MM} \end{pmatrix} \qquad (28)$$

To satisfy the requirements of eqn (21) T should be unit upper triangular, and this can be achieved by defining

$$\alpha_{ij} = \begin{cases} 0 & \forall \; i > j \\ 1 & \forall \; i = j \\ \dfrac{\sum\limits_{t=1}^{N} w_i(t) p_j(t)}{\sum\limits_{t=1}^{N} w_i^2(t)} & \forall \; i < j \end{cases} \qquad (29)$$

Substituting eqn (29) in eqn (28) gives T the required form

$$T = \begin{pmatrix} 1 & \alpha_{01} & \alpha_{02} & \cdots & \alpha_{0M} \\ & 1 & \alpha_{12} & & \alpha_{1M} \\ & & 1 & & \\ & \bigcirc & & & \\ & & & & 1 \end{pmatrix} \qquad (30)$$

and the definitions of α_{ij} in eqn's (29) and (11) coincide. The elements of the W matrix eqn (25) can now be determined from eqn (23)

$$WT = P \qquad (31)$$

or $\quad W = P - W(T-I)$

which can be expressed as

$$\begin{pmatrix} w_o(1)\ldots w_M(1) \\ w_o(2) \qquad \cdot \\ \cdot \qquad\qquad \cdot \\ \cdot\ \cdot \qquad\qquad \cdot \\ w_o(N)\ldots w_M(N) \end{pmatrix} = \begin{pmatrix} p_o(1)\ldots p_M(1) \\ p_o(2)\ldots p_M(2) \\ \cdot \qquad\qquad \cdot \\ \cdot \qquad\qquad \cdot \\ p_o(N)\ldots p_M(N) \end{pmatrix} - \begin{pmatrix} w_o(1)\ldots w_M(1) \\ w_o(2)\ldots\ldots\ldots \\ \cdot \qquad\qquad \cdot \\ \cdot \qquad\qquad \cdot \\ w_o(N)\ldots w_M(N) \end{pmatrix} \begin{pmatrix} 0 & \alpha_{01}\ldots\alpha_{0M} \\ & 0 \qquad \alpha_{1M} \\ & \qquad\quad \cdot \\ \bigcirc & \qquad\quad \cdot \\ & \qquad\quad 0 \end{pmatrix} \tag{32}$$

Multiplying out eqn (32) yields the definition of $w_j(t)$, $j = 0,\ldots M$ given in eqn (10).

An estimate of the auxiliary model parameter vector g in eqn (22) can now be determined by minimising the mean squared error to yield

$$\begin{aligned} \hat{g} &= (W^TW)^{-1}W^TY \\ &= D^{-1}W^TY \end{aligned} \tag{33}$$

or

$$\begin{pmatrix} \hat{g}_o \\ \hat{g}_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \hat{g}_M \end{pmatrix} = \begin{pmatrix} \dfrac{\sum\limits_{t=1}^{N} w_o(t)y(t)}{\sum\limits_{t=1}^{N} w_o^2(t)} \\ \vdots \\ \dfrac{\sum\limits_{t=1}^{N} w_M(t)y(t)}{\sum\limits_{t=1}^{N} w_M^2(t)} \end{pmatrix}$$

The relationship between the NARMAX model parameter vector of eqn (7) and the auxiliary model parameter vector can be obtained from eqn (23)

$$\hat{g} = T\hat{\theta} \tag{23}$$

$$\hat{\theta} = \hat{g} - (T-I)\hat{\theta}$$

or

$$\begin{pmatrix}\hat\theta_o\\ \hat\theta_1\\ \cdot\\ \cdot\\ \cdot\\ \hat\theta_M\end{pmatrix} = \begin{pmatrix}\hat g_o\\ \hat g_1\\ \cdot\\ \cdot\\ \cdot\\ \hat g_M\end{pmatrix} - \begin{pmatrix} 0 & \alpha_{01} & \alpha_{02} & \cdots & \alpha_{0M}\\ & 0 & \alpha_{12} & & \alpha_{1M}\\ & & 0 & & \cdot\\ & & & \ddots & \cdot\\ & \text{\huge 0} & & & \cdot\\ & & & & 0\end{pmatrix}\begin{pmatrix}\hat\theta_o\\ \hat\theta_1\\ \\ \\ \\ \hat\theta_M\end{pmatrix} \tag{34}$$

which is equivalent to eqn's (14) and (15).

3.2. Properties of the Algorithm

An analysis of the properties of the estimates is almost identical to the analysis for an extended least squares algorithm [Norton 1986].

Assuming that sufficient noise terms are included in the model to account for correlated noise and assuming that the estimates converge then the prediction error sequence $\varepsilon(\cdot)$ in eqn's (7) and (8) will be reduced to a white noise sequence. Thus from eqn (33)

$$\hat{g} = D^{-1}W^TY$$

substituting for Y from eqn (22)

$$\hat{g} = D^{-1}W^T(Wg + \underline{\varepsilon})$$

$$= g + D^{-1}W^T\underline{\varepsilon} \tag{35}$$

Rearranging eqn (35)

$$D(\hat{g} - g) = W^T\underline{\varepsilon}$$

and taking expected value

$$E\{D(\hat{g}-g)\} = E\{W^T\underline{\varepsilon}\} \tag{36}$$

Providing $\varepsilon(t)$ is a zero mean white noise sequence which is independent of the input then expansion of eqn (36) shows that $E\{W^T\underline{\varepsilon}\} = 0$ and consequently the estimates will be unbiased $E[\hat{g}] = g$.

The covariance of the auxiliary model parameter vector is given by

$$\mathrm{Cov}(\hat{g}) = E\{(\hat{g}-g)(\hat{g}-g)^T\} \tag{37}$$

Substituting from eqn (35) and using the result $E\{\varepsilon\varepsilon^T\} = \sigma^2 I$ yields

$$\mathrm{Cov}(\hat{g}) = \sigma^2 D^{-1} \tag{38}$$

and hence $$\mathrm{Cov}(\hat{\theta}) = T^{-1}\mathrm{Cov}(\hat{g})(T^{-1})^T \tag{39}$$

Since T is a unity upper triangular matrix the inverse $T^{-1} = \{t_{ij}\}$ can be computed directly from the elements of $T = \{\alpha_{ij}\}$ eqn (30) by

$$t_{ij} = \begin{cases} -\sum_{k=i+1}^{j} \alpha_{ik} t_{kj} & i < j \\ 1 & i = j \\ 0 & i > j \end{cases} \tag{40}$$

The i'th diagonal element of $\mathrm{Cov}(\hat{\theta})$ defines the variance of $\hat{\theta}_i$, $i = 0,1,\ldots M$ which can be evaluated by substituting eqn (40) in (39) to give

$$\mathrm{Var}\{\hat{\theta}_i\} = \sigma^2 (d_i + \sum_{j=i+1}^{M} t_{ij}^2 d_j) \tag{41}$$

$$= \sigma^2 (\sum_{j=i}^{M} t_{ij}^2 d_j)$$

where from (26)

$$d_i = 1/\sum_{t=1}^{N} w_i^2(t) \tag{42}$$

and σ can be estimated from eqn (8)

$$\hat{\sigma} = \sqrt{\frac{1}{N-(M+1)-N_y} \sum_{t=N_y+1}^{N} (y(t) - \sum_{m=o}^{M} \hat{g}_m w_m(t))^2} \tag{43}$$

4. Selection of Variables

The determination of the model structure or which variables to include in the NARMAX model expansion eqn (3) is vital if a parsimonious representation of the system is to be identified. Simply increasing the order of the dynamic terms $(N_y, N_u, N_\varepsilon)$ in eqn (3) 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 ill-conditioned computations.

The maximum number of terms in the NARMAX model of eqn (3) is given by

$$n = M + 1$$

where

$$M = \sum_{i=1}^{\ell} n_i \tag{44}$$

$$n_i = \{n_{i-1}(N_y+N_u+N_\varepsilon+i-1)\}/i \quad , \qquad n_o = 1$$

For example a first order dynamic process model $(N_y = N_u = 1)$ with a third order noise model $(N_\varepsilon = 3)$ expanded as a cubic polynomial $(\ell = 3)$ would contain 56 possible terms. This is clearly excessive but fortunately simulation has shown that only a few of these terms (typically less than ten including the noise model) are significant and the remainder can be discarded with little deterioration in prediction accuracy.

There are several possible ways to determine which are the significant terms which should be included in the model. Most of these including a stepwise regression algorithm [Billings and Voon 1986a] and a log determinant ratio test [Leontaritis and Billings 1987a] have been described in earlier publications. Whilst these methods are extremely effective an alternative and much simpler method can be derived as a by-product of the orthogonal estimation algorithm.

Multiplying the auxiliary model eqn (8) by itself and taking the time average gives

$$\frac{1}{N}\sum_{t=1}^{N} y^2(t) = \frac{1}{N}\sum_{t=1}^{N}\left\{\sum_{m=o}^{M} g_m^2 w_m^2(t)\right\} + \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t) \qquad (45)$$

assuming that $\varepsilon(t)$ is a zero mean white noise sequence and the orthogonality property of eqn (9) holds. The maximum mean squared prediction error is achieved when no terms are included in the model (M = O) to give

$$\left\{\frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t)\right\}_{M=O} = \frac{1}{N}\sum_{t=1}^{N} y^2(t) \qquad (46)$$

From eqn's (8) and (45) the reduction in mean squared error by including a term $\theta_i p_i(t)$ in the model will be equal to

$$\frac{1}{N}\sum_{t=1}^{N} g_i^2 w_i^2(t) \qquad (47)$$

Thus the reduction in the mean squared error eqn (47) as a result of including the term $\theta_i p_i(t)$ can be expressed as a percentage reduction in the total mean squared error eqn (46) by defining

$$\frac{\sum_{t=1}^{N} g_i^2 w_i^2(t)}{\sum_{t=1}^{N} y^2(t)} \times 100 \qquad (48)$$

for i = O,1,2,...M. In practice a constant or dc level is always included in the model eqn (3), where from (12)

$$g_o = \frac{1}{N}\sum_{t=1}^{N} y(t) \qquad (49)$$

The influence of g_o can be removed by rewriting eqn (45) as

$$\frac{1}{N}\sum_{t=1}^{N} y^2(t) - \left\{\frac{1}{N}\sum_{t=1}^{N} y(t)\right\}^2 = \frac{1}{N}\sum_{t=1}^{N}\left\{\sum_{m=1}^{M} g_m^2 w_m^2(t)\right\} + \frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t) \quad (50)$$

and redefining eqn (48) with the effects of the dc term eliminated to yield

$$ERR_i = \frac{\sum_{t=1}^{N} g_i^2 w_i^2(t)}{\sum_{t=1}^{N} y^2(t) - \frac{1}{N}\left\{\sum_{t=1}^{N} y(t)\right\}^2} \times 100 \quad (51)$$

for i = 1,2...M. The quantity ERR_i which will be called the Error Reduction Ratio provides an indication of which terms to include in the model. Insignificant terms can then be discarded by defining a value of ERR_i below which terms are considered to contribute a negligible reduction in the mean squared error.

The threshold value of ERR for process model terms called C_d (all terms which do not include $\varepsilon(\cdot)$) is typically set to 0.05 to 0.5. For terms which involve the prediction errors the threshold value of ERR is called C_{de} and a lower value of 0.0001 to 0.001 is selected to ensure that sufficient noise terms are included so that the prediction errors are reduced to a white noise sequence.

Notice that ERR only gives an indication of which terms should be included in the model. The ERR value is however dependent upon the order in which the term enters the equation. To overcome this problem a search algorithm which provides a type of ERR test which is independent of the order of inclusion of terms is required. This is currently under development.

5. Implementing the Algorithm

The orthogonal estimation algorithm given in section 3 can now be extended to include the error reduction ratio test and several other minor modifications to give a combined parameter estimation and structure detection algorithm. The algorithm is given for the summation notation of section 3 with the corresponding matrix notation of section 3.1 given in brackets:-

(i) Select values for N_y, N_u, N_ε, d, ℓ in eqn (3) and set $\varepsilon(t) = 0$, $t = 1 \ldots N$. Select C_d and C_{de}.

(ii) Estimate all the parameters which do not include $\varepsilon(\cdot)$ terms using eqn's (10) and (13) {compute the elements of the matrices W eqn (25), T eqn (30) and $\hat{g}$ eqn (33)}

(iii) If $\hat{\varepsilon}(t) = 0$, $t = 1,2 \ldots N$ go to (iv) otherwise use $\hat{\varepsilon}(t)$ to estimate the parameters associated with the prediction error terms using eqn's (10) to (13) {compute W, T and $\hat{g}$}

(iv) Compute ERR_i eqn (51), test against the thresholds C_d, C_{de} and delete insignificant terms

(v) Estimate the prediction errors using eqn (16)

(vi) If any process model terms were deleted in (iv) then go to (ii), otherwise go to (iii) and repeat until convergence.

(vii) Estimate the NARMAX model coefficients using eqn's (14) and (15) {compute $\hat{\theta}$ from eqn (34)}.

In step (i) suitably large values of N_y, N_u, N_ε and ℓ should be selected to ensure that the class of models which this choice defines is large enough to include the model of the process under investigation. If N_u is selected large enough the value of d will not be crucial and it is often set to d = 1.

Estimates of the $\hat{g}_i$ terms in steps (ii) and (iii) can become ill-conditioned if the numerator in eqn (13) becomes very small. This can be avoided by including an additional criterion in steps (ii) and (iii) such that if for any specific i

$$\sum_{t=istart}^{ifin} w_i^2(t) \leq 10^{-5} \qquad (52)$$

then set $\alpha_{ij} = 0 \;\forall\; j \geq i$; (this is equivalent to setting elements of the i'th row of T eqn (30) to zero) $g_i = 0$ and hence $ERR_i = 0$, where istart = max $\{(N_u+d-1), N_y, N_\varepsilon\} + 1$ and ifin = record length of the estimation set. It is also important to emphasise that the next term $w_{i+1}(t)$ should not be orthogonalised against the previous offending $w_i(t)$ term since this would only compound the problem. The solution is to go back and choose a new $p_j(t)$ term in eqn (4). It is also important to ensure that identical $p_m(n)$ terms are not formed for two different values of m.

In situations where no process model terms (terms which do not involve $\varepsilon(t-j)$, $j>0$) are deleted using ERR there will be no need to re-estimate the process parameters in step (ii). However, the terms in the model are orthogonalised in the order that they are introduced and this means that if any process model terms are deleted as insignificant in step (iv) then all the subsequent terms will need to be re-orthogonalised. Hence when this situation arises, usually on the first iteration only, it is necessary to return to step (ii) following step (vi).

Although the algorithm above identifies or detects the structure of the model and then estimates the unknown parameters the model obtained should only be accepted after model validity tests have confirmed that the fit is adequate. Although the ERR test provides an indication of which terms to include in the model situations do arise where terms which are insignificant according to ERR will induce bias if they are excluded from the final model. Simulation has shown that this situation usually only occurs with the noise model or prediction error terms. The majority of the prediction error terms often have a very low ERR value, if these terms are deleted however $\varepsilon(t)$ may become a correlated sequence instead of a white noise sequence and this will induce bias in the model parameters. Fortunately, this situation can be readily detected because the prediction errors will have the required properties iff the following results hold [Billings and Voon 1986b]

$$\begin{aligned}
&\phi_{\xi\xi}(\tau) = \delta(\tau) \\
&\phi_{u\xi}(\tau) = 0 \; \forall \; \tau \\
&\phi_{u^{2'}\xi^{2}}(\tau) = 0 \forall \tau \\
&\phi_{u^{2'}\xi}(\tau) = 0 \forall \tau \\
&\phi_{\xi\xi u}(\tau) = \tau \geq 0
\end{aligned} \qquad (53)$$

where $\xi(t)$ represents an estimate of the prediction error sequence and the ' indicates that the mean has been removed from a signal. The model validity tests of eqn (53) often indicate which, if any, terms have been omitted from the model [Billings and Voon 1986b]. These terms can then be forced into the model to rectify the discrepancy.

Whenever nonlinear models are to be estimated mean levels should not be removed from the data, the data should be split into an estimation and a testing set and input excitation signals must be carefully chosen [Leontaritis and Billings 1987b]. Removing mean levels can induce input sensitivity when estimating nonlinear models [Billings and Voon 1984] and it is therefore preferable to include a dc level as part of the parameter vector eqn (3). Splitting the data into an estimation set which is used to estimate the parameters, and a testing set which is used to judge the predictive capability of the fitted model can reveal severe model deficiencies which would otherwise go undetected.

6. Simulation Results

Several linear and nonlinear models were simulated to illustrate the orthogonal parameter estimation algorithm.

6.1. Linear Systems

The linear system S_1 described by

$$y(t) = \frac{z^{-1}}{1-0.5z^{-1}} u(t) + (1+0.4z^{-1})\xi(t) \tag{54}$$

was simulated where u(t) was a sixth order prbs and $\xi(t)$ a zero mean white noise sequence. The estimation set was defined over the points (1,700) and the testing set over the points (701,1000). The orthogonal estimation algorithm of section 5 with initial values $N_y = N_u = N_\varepsilon = 4$, $\ell = 1$, $d = 1$ $C_d = 0.5$, $C_{de} = 0.0$ produced the following model after ten iterations

$$\begin{aligned} y(t) = {} & 0.0097 + \underset{\{22.1\}}{0.49y(t-1)} + \underset{\{71.4\}}{0.9967u(t-1)} \\ & + \varepsilon(t) - \underset{\{0.06\}}{0.0384\varepsilon(t-1)} - \underset{\{0.21\}}{0.187\varepsilon(t-2)} \\ & - \underset{\{0.007\}}{0.0352\varepsilon(t-3)} - \underset{\{0.03\}}{0.07462\varepsilon(t-4)} \end{aligned} \tag{55}$$

where the number in {} under each term is the error reduction ratio associated with that term. The model validity tests of eqn (53) were all satisfactory and clearly the ERR test has detected the correct model structure. Notice that setting $C_{de} = 0$ forced all four noise terms into the model. Even though ERR the percentage contribution of the noise terms to the output mean squared error was very small they would induce bias if they were omitted. In contrast ERR for the y(t-1) and u(t-1) terms shows that they contribute 22.1% and 71.4% respectively to the output mse.

A second linear system S_2 described by

$$\cdot y(t) = \frac{z^{-1}+0.5z^{-2}}{1-1.5z^{-1}+0.7z^{-2}}u(t) + \frac{1+0.6z^{-1}}{1-1.5z^{-1}+0.7z^{-2}}\xi(t) \tag{56}$$

was simulated with u(t), ξ(t) and the estimation and testing sets defined as for S_1. The orthogonal estimation algorithm with initial values $N_y = N_u = N = 4$, $d = 1$, $\ell = 1$, $C_d = 0.5$, $C_{de} = 0.0$ produced the following model after ten iterations

$$\begin{aligned} y(t) &= 0.010 + \underset{\{78.6\}}{1.937y(t-1)} - \underset{\{14.7\}}{1.426y(t-2)} + \underset{\{0.838\}}{0.3699y(t-3)} \\ &+ \underset{\{5.10\}}{0.99u(t-1)} + \varepsilon(t) - \underset{\{0.00004\}}{0.012\varepsilon(t-1)} - \underset{\{0.0065\}}{0.099\varepsilon(t-2)} \\ &+ \underset{\{0.00058\}}{0.029\varepsilon(t-3)} - \underset{\{0.0001\}}{0.013\varepsilon(t-4)} \end{aligned} \tag{57}$$

Inspection of eqn (57) shows that ERR has incorrectly selected three lags in y(t-i), i = 1,2,3 and only one lag in the input u(t-1). However the model validity tests illustrated in Fig.1 clearly show that the model eqn (57) is not of the correct structure. A u(t-2) term was therefore forced into the model to yield

$$\begin{aligned} y(t) &= 0.02 + \underset{\{78.6\}}{1.513y(t-1)} - \underset{\{14.7\}}{0.7385y(t-2)} + \underset{\{0.838\}}{0.02421y(t-3)} \\ &+ \underset{\{5.1\}}{0.9906u(t-1)} + \underset{\{0.114\}}{0.4857u(t-2)} + \text{noise} \end{aligned} \tag{58}$$

Clearly the u(t-2) term was deleted in the initial estimate eqn (57) because ERR for this term is 0.114% below the threshold of $C_d = 0.5$. The model validity tests eqn (53) were now all within the 95% confidence bounds.

Although ERR for y(t-3) is above the threshold and greater than ERR for u(t-2) the coefficient of y(t-3) has now become very small and consequently the latter was deleted from the model to see what effect this would have. With this additional constraint the estimated model was

$$\begin{aligned} y(t) &= 0.02 + 1.485y(t-1) - 0.6929y(t-2) \\ &\quad + 0.9913u(t-1) + 0.5119u(t-2) + \text{noise} \end{aligned} \tag{59}$$

The model validity tests associated with the model of eqn (59) were all within the 95% confidence bands and this was accepted as the final model, Thus by an interactive use of the orthogonal estimator together with the model validity tests the correct model structure has been determined. Other simulations have indicated that when fitting linear models ($\ell = 1$) it is often appropriate to force an equal number of lagged u(•) and y(•) terms into the model and to set $C_{de} = 0$ to ensure sufficient noise model terms are included.

6.2. Nonlinear Systems

The nonlinear system S_3 described by

$$\begin{aligned} x(t) &= 0.8x(t-1) + 0.4\{u(t-1) + u(t-1)**2 + u(t-1)**3\} \\ e(t) &= \xi(t) + 0.6\,\xi(t-1) \\ y(t) &= x(t) + e(t) \end{aligned} \tag{60}$$

was simulated where u(t) was a zero mean uniform random sequence (-3,3) and $\xi(t)$ was a uniformly distributed zero mean white noise sequence (-0.3,0.3) The estimation set was defined over the points (1,700) and the testing set over the points (701,1000). The orthogonal estimation algorithm with initial values $N_y = N_u = N_\varepsilon = 2$, $d = 1$, $\ell = 3$, $C_d = 0.5$, $C_{de} = 0.001$ produced the following model after ten iterations

$$\begin{aligned} y(t) = {} & -0.015 + \underset{\{51.1\}}{0.8y(t-1)} + \underset{\{35.5\}}{0.399u(t-1)} + \underset{\{2.4\}}{0.40u(t-1)**2} \\ & + \underset{\{4.8\}}{0.40u(t-1)**3} + \varepsilon(t) - \underset{\{0.0027\}}{0.229\varepsilon(t-1)} \end{aligned} \qquad (61)$$

The model validity tests which are illustrated in Fig.2 show that $\phi_{\xi\xi}(\tau) \neq \delta(\tau)$. This indicates a deficient noise model so additional linear noise terms were forced into the model by setting $N_\varepsilon = 3$, $C_{de} = 0$ with a polynomial order for the noise of one to yield after ten iterations

$$\begin{aligned} y(t) = {} & -0.011 + \underset{\{57.13\}}{0.80y(t-1)} + \underset{\{35.5\}}{0.39u(t-1)} + \underset{\{2.4\}}{0.4u(t-1)**2} \\ & + \underset{\{4.8\}}{0.4u(t-1)**3} + \varepsilon(t) - \underset{\{0.003\}}{0.25\varepsilon(t-1)} - \underset{\{0.0009\}}{0.135\varepsilon(t-2)} - \underset{\{0.00003\}}{0.024\varepsilon(t-3)} \end{aligned} \qquad (62)$$

All the model validity tests were now within the confidence bands and the model was accepted. Notice that if C_{de} had been set to zero at the initial estimation stage a large number of both linear and nonlinear terms in $\varepsilon(\cdot)$ would have been forced into the model. It is preferable therefore whenever $\ell>1$ to set $C_{de} = 0.001$ initially to determine if there are significant nonlinear noise terms. If the model validity tests then indicate that the noise model is incorrect, as in the above example, linear noise terms can be forced into the model.

The nonlinear system S_4 described by

$$\begin{aligned} x(t) &= 1.4x(t-1) - 0.1x(t-1)**2 - 0.4x(t-2) \\ &\qquad + 0.2u(t-1) + 0.4u(t-1)**2 \qquad (63) \\ y(t) &= x(t) + \xi(t) \end{aligned}$$

was simulated where u(t) was a zero mean uniform random sequence with variance 2.5 and $\xi(t)$ was a Gaussian sequence N(0,0.1). The estimation and testing sets were defined as in S_3 above.

The orthogonal estimation algorithm with $N_y = N_u = N_\varepsilon = 3$, $d = 1$, $\ell = 3$, $C_d = 0.5$, $C_{de} = 0.001$ produced the following model after ten iterations

$$\begin{aligned}
y(t) = 0.0055 &+ \underset{\{31.5\}}{1.398y(t-1)} - \underset{\{3.43\}}{0.099y(t-1)**2} \\
&- \underset{\{6.8\}}{0.40y(t-2)} + \underset{\{10.4\}}{0.199u(t-1)} + \underset{\{46.7\}}{0.398u(t-1)**2} \\
&+ \varepsilon(t) - \underset{\{0.33\}}{1.198\varepsilon(t-1)} + \underset{\{0.36\}}{0.245\varepsilon(t-2)} + \underset{\{0.013\}}{0.142\varepsilon(t-3)} \\
&+ \underset{\{0.026\}}{0.1594\varepsilon(t-1)y(t-1)}
\end{aligned} \tag{64}$$

All the model validity tests were inside the 95% confidence bands and the model was accepted. A comparison of eqn (64) with the model eqn (63) shows that of the 220 possible terms in the model, eqn (44), the algorithm has selected the correct model structure.

In all the examples, S_1' to S_4, the predicted output of the estimated model over both the estimation and testing set was virtually coincident with the measured process output.

7. Conclusions

An orthogonal parameter estimation and structure detection routine has been derived for both linear and nonlinear stochastic systems. The orthogonal property of the algorithm allows each parameter in the auxiliary model to be estimated one at a time by repeated application of a very simple formula. Additional terms can be added to the model without the need to re-estimate all the previous coefficients and the percentage reduction that each term makes to the output mean squared error, the ERR test, provides an extremely simple indication of the significance of each term in the model. Other advantages of the algorithm are that the estimation of the process and noise model parameters can be decoupled and that implementation on a microprocessor should be straightforward.

The extension of the algorithm to other linear least squares based estimators and to multivariable linear and nonlinear systems is now complete and will appear in forthcoming publications.

8. References

Billings, S.A., Gray, J.O., Owens, D.H. (Eds) (1984): Nonlinear system design; P. Peregrinus.

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

Billings, S.A., Voon, W.S.F. (1986a): A prediction error and stepwise regression estimation algorithm for nonlinear systems; Int. J. Control, 44, 803-822.

Billings, S.A., Voon, W.S.F. (1986b): Correlation based model validity tests for nonlinear models; Int. J. Control, 44, 235-244.

Blum, E.K. (1972): Numerical Analysis and Computation; Addison Wesley.

Fox, L. (1964): An Introduction to Numerical Linear Algebra; Clarendon Press.

Korenberg, M.J. (1985): Orthogonal identification of nonlinear difference equation models; Mid West Symp. on Cts and Systems, Louisville.

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

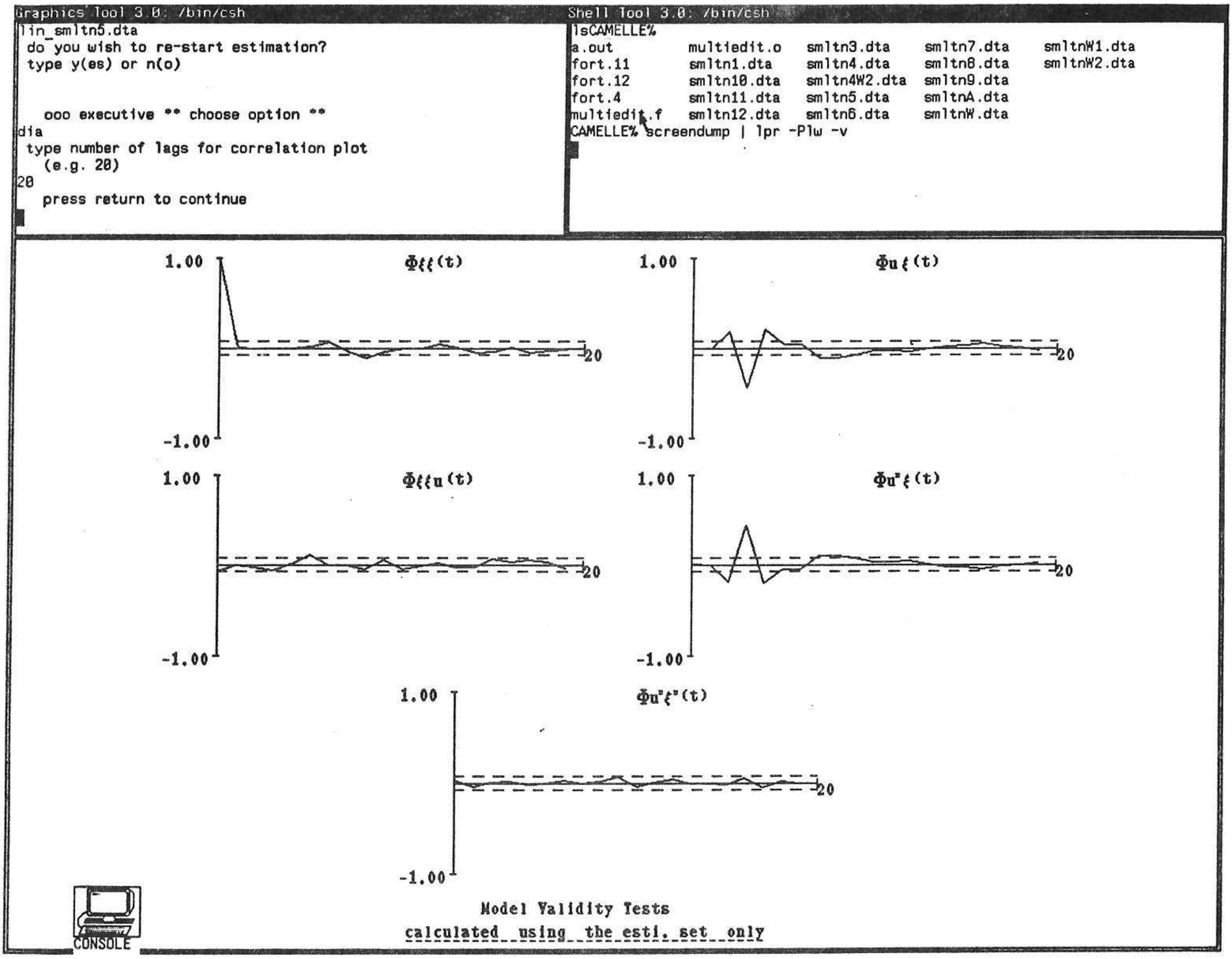
Leontaritis, I.J., Billings, S.A. (1987a): Model selection and validation methods for nonlinear systems; Int. J. Control (to appear).

Leontaritis, I.J., Billings, S.A. (1987b): Experimental design and identifiability for nonlinear systems; Int. J. Systems Science (to appear).

Marmarelis, P.Z., Marmarelis, V.Z. (1978): Analysis of Physiological Systems. The White Noise Approach; Plenum Press.

Norton, J.P. (1986): An Introduction to System Identification; Academic Press.

Schetzen, M. (1980): The Volterra and Wiener Theories of Nonlinear Systems; Wiley.



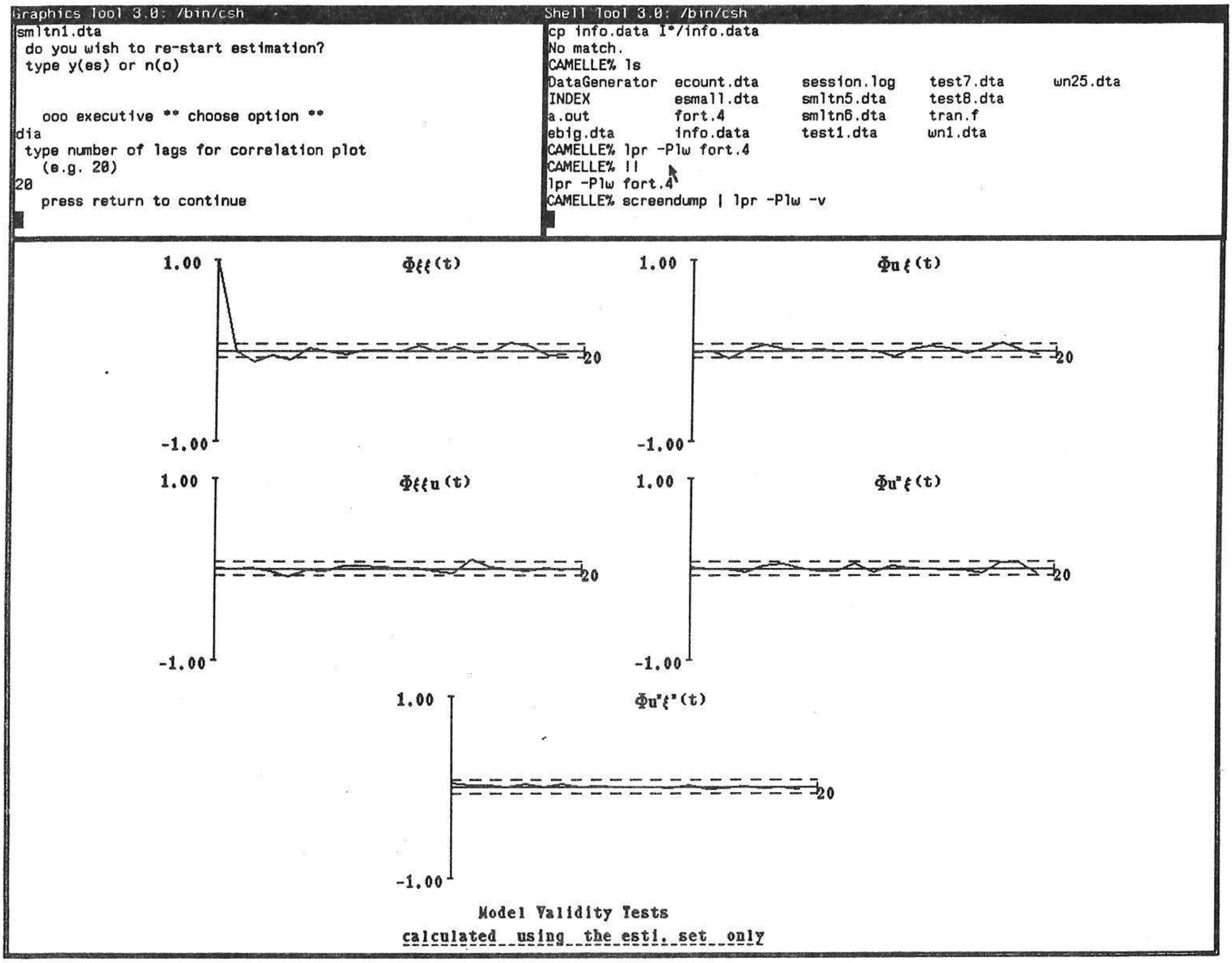
```
lin_smltn5.dta
 do you wish to re-start estimation?
 type y(es) or n(o)


   ooo executive ** choose option **
dia
 type number of lags for correlation plot
   (e.g. 20)
20
  press return to continue
```

```
lsCAMELLE%
a.out        multiedit.o   smltn3.dta     smltn7.dta    smltnW1.dta
fort.11      smltn1.dta    smltn4.dta     smltn8.dta    smltnW2.dta
fort.12      smltn10.dta   smltn4W2.dta   smltn9.dta
fort.4       smltn11.dta   smltn5.dta     smltnA.dta
multiedit.f  smltn12.dta   smltn6.dta     smltnW.dta
CAMELLE% screendump | lpr -Plw -v
```

Model Validity Tests

calculated using the esti. set only

Fig.1. Model Validity Tests for S_2



```
smltn1.dta
 do you wish to re-start estimation?
 type y(es) or n(o)


   ooo executive ** choose option **
dia
 type number of lags for correlation plot
   (e.g. 20)
20
   press return to continue
```

```
cp info.data I*/info.data
No match.
CAMELLE% ls
DataGenerator  ecount.dta    session.log   test7.dta    wn25.dta
INDEX          esmall.dta    smltn5.dta    test8.dta
a.out          fort.4        smltn6.dta    tran.f
ebig.dta       info.data     test1.dta     wn1.dta
CAMELLE% lpr -Plw fort.4
CAMELLE% !!
lpr -Plw fort.4
CAMELLE% screendump | lpr -Plw -v
```

Fig.2. Model Validity Tests for S_3