Monograph:
Zhu, Q.M. and Billings, S.A. (1994) Fast Orthogonal Identification of Nonlinear Stochastic Models and Radial Basis Function Neural Networks. Research Report. ACSE Research Report 526 . Department of Automatic Control and Systems Engineering

Fast orthogonal identification of nonlinear stochastic models and radial basis function neural networks

Q.M. Zhu and S.A. Billings

Department of Automatic Control and Systems Engineering,
University of Sheffield, Sheffield S1 4DU, UK

Abstract:

A new fast orthogonal estimation algorithm is derived for a wide class of nonlinear stochastic models including training radial basis function neural networks. The selection of significant regressors and the estimation of unknown parameters in the presence of nonlinear noise sources are considered and simulated examples are included to demonstrate the efficiency of the new procedure.

Research Report No. 526

September 1994

Fast orthogonal identification of nonlinear stochastic models and radial basis function neural networks

Q.M. Zhu and S.A. Billings

Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 4DU, UK

Abstract:

A new fast orthogonal estimation algorithm is derived for a wide class of nonlinear stochastic models including training radial basis function neural networks. The selection of significant regressors and the estimation of unknown parameters in the presence of nonlinear noise sources are considered and simulated examples are included to demonstrate the efficiency of the new procedure.

1.0 Introduction

Modelling of nonlinear stochastic systems based on the Nonlinear Auto-Regressive Moving Average with eXogenous inputs (NARMAX) model has been extensively studied and model identification algorithms have evolved from extended least squares, prediction error to orthogonal estimators based on polynomial NARMAX expansions, rational model forms and the training of neural networks (Billings and Voon 1984, Korenberg, Billings, Liu and McIlroy 1988, Billings, Korenberg and Chen 1988, Billings and Chen 1989, Billings and Chen 1989, Billings and Zhu 1991, Zhu and Billings 1993). All the algorithms have been designed to provide unbiased estimates in the presence of nonlinear correlated noise but the orthogonal based procedures offer added functionality, are upwardly extendable and can be applied to very complex model types.

The orthogonal routines can be used to sort through a library of possible model terms to rank these in order of importance and to provide unbiased parameter estimates. If the model form is a polynomial or extended model set expansion the term ranking can be completed prior to noise modelling and this provides a very powerful procedure for determining concise model forms from large data sets. Estimation of the structure and the parameters of rational models, defined as the ratio of two stochastic polynomial expansions, is much more complex and an iterative procedure must be employed to remove bias terms which are induced even when the noise is uncorrelated and white. Although the exact formulation of the algorithm is different for each model form the basis of the orthogonal property ensures that model term can be processed one at a time based on a core estimation routine. Hence m dimensional estimation problems can be broken down into m one dimensional problems and this means that numerical ill conditioning can be avoided and the algorithms are upwardly extendable to very complex model forms.

While previous studies have concentrated on formulations of the algorithm for different

model forms, the need to determine model structure and avoid estimation bias very little attention has been given to the properties of the core orthogonal routines. But since the basis of the whole formulation is to re-enter the orthogonal module for each model term, and this can amount to hundreds and often thousands of calls even for relatively simple models, optimization of these routines should have considerable benefits to the overall performance of the algorithms and this forms the basis of present study.

Previous work by Korenberg (1988), and Korenberg and Paarmann (1991) has considered fast orthogonal algorithms but unfortunately these methods are restricted to model forms with special properties. Because the NARMAX model and variants of this form include time delayed nonlinear terms and transcendental terms which are frequently ordered or processed in an irregular order Korenberg's methods cannot be applied and alternatives must be investigated. There are also a number of fast algorithms which are used in linear system identification and signal processing. For example fast recursive least squares (Ljung and Soderstrom 1983) and the fast transversal filter algorithm (Haykin 1986), which effectively retain the advantages of ordinary recursive least squares but the computational complexity is reduced to a level comparable to that of the least mean squares algorithm, are used in system identification. Another typical example is the fast computation algorithms used in digital signal processing (Blahut 1985) which deal with the computational complexity of convolutions and transformations. However all these algorithms are oriented to computations based on a prefixed linear model structure and cannot be directly applied to select an optimal model from a large number of candidate models and to estimate the associated unknown parameters.

The aim of the present paper therefore is to derive a generalized fast orthogonal algorithm which maintains all the properties of the methods described above, which can be configured to work for polynomial, extended model set, rational models and in the training of neural networks. Two simulated examples are included to demonstrate the efficiency of the new routine.

2.0 System identification

In this section a general model structure will be introduced as a basis to approximate a wide range of systems and then the techniques of parameter estimation, structure detection and model validity tests will be briefly presented. These results provide a fundamental basis for the derivation of the new fast orthogonal algorithm and other studies in the following sections.

2.1 Model description

A wide class of linear or nonlinear stochastic systems can be described by a NARMAX model defined as

$$y(t) = f(y^{t-1}, u^{t-1}, \varepsilon^{t-1}) + \varepsilon(t) \tag{2.1}$$

where t (t=1, 2, ...) is a time index, $y(t)$, $u(t)$ and $\varepsilon(t)$ denote the output, input and residual

sequences respectively, $f(.)$ is a linear or nonlinear function and

$$y^{t-1} = [y(t-1), \ldots, y(t-n_y)] \qquad u^{t-1} = [u(t-1), \ldots, u(t-n_u)]$$
$$\varepsilon^{t-1} = [\varepsilon(t-1), \ldots, \varepsilon(t-n_\varepsilon)]$$

(2.2)

are output, input and residual vectors with delayed elements from 1 to n_y, n_u and n_ε respectively. Notice that the form of the model can be very wide and can include the linear model, the Nonlinear AutoRegressive Moving Average with eXogenous input (NARMAX) model (Billings and Chen 1989), a neural network expansion (Zhu and Billings 1994) etc.

When the function $f(.)$ takes the form of a polynomial $y(t)$ can be expressed as

$$y(t) = \sum_{j=1}^{m} p_j(t)\,\theta_j + \varepsilon(t)$$

(2.3)

where $p_j(t) = p_j(y^{t-1}, u^{t-1}, \varepsilon^{t-1})$ is defined as a term which is a linear or nonlinear function of past outputs, inputs and residual sequences, θ_j is the associated unknown parameter. Notice that for nonlinear models there may be cross product terms involving y, u and ε.

2.2 Parameter estimation

First consider an extended least squares parameter estimation (Ljung 1987) of (2.3)

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

(2.4)

where

$$\hat{\Theta} = [\hat{\theta}_1 \ldots \hat{\theta}_m]^T$$
$$\Phi = \begin{bmatrix} p_1(1) & \ldots & p_m(1) \\ \ldots & \ldots & \ldots \\ p_1(N) & \ldots & p_m(N) \end{bmatrix}$$
$$Y = [y(1) \ldots y(N)]$$

(2.5)

Given a known model structure the extended least squares algorithm can be readily

applied and delivers unbiased parameter estimates. However in practice the terms which should be included in the model are seldom known a priori and so it becomes necessary to consider how to determine those terms or the model structure. This is critically important in the case of nonlinear systems because the number of candidate model terms is often very large. Orthogonal least squares algorithms can deal with this problem by transforming the cross correlated normal matrix $[\Phi^T\Phi]$ into an orthogonal matrix so that the solution of m coupled equations becomes equivalent to solving m independent equations. An optimal model term selection procedure can then be developed to exploit the orthogonal property and to determine if each term is significant or not.

The orthogonal parameter estimation algorithm for the model of (2.3) is summarized below. Consider an orthogonal transformation of (2.3) to yield

$$y(t) = \sum_{j=1}^{m} w_j(t)\, g_j + \varepsilon(t) \tag{2.6}$$

where the parameters can be estimated by

$$\hat{G} = [\hat{g}_1 \ldots \hat{g}_m]^T = [W^T W]^{-1} W^T Y \tag{2.7}$$

and

$$W = \Phi A^{-1} = \begin{bmatrix} w_1(1) & \ldots & w_m(1) \\ \ldots & \ldots & \ldots \\ w_1(N) & \ldots & w_m(N) \end{bmatrix} \tag{2.8}$$

is an orthogonal regression matrix with the properties

$$\frac{1}{N} W^T W = diag\,\{\overline{w_1^2(t)} \ldots \overline{w_m^2(t)}\}$$

$$\overline{w_k^2(t)} = \frac{1}{N}\sum_{t=1}^{N} w_k^2(t) \qquad \overline{w_k(t)\, w_i(t)} = \frac{1}{N}\sum_{t=1}^{N} w_k(t)\,{}_i w(t) = 0, \qquad k \neq i \tag{2.9}$$

The parameter estimates in the model of (2.3) can be recovered by computing

$$\hat{\Theta} = A^{-1}\hat{G} \tag{2.10}$$

where Φ is defined in (2.5) and A is the orthogonal transform expressed as

$$A = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{1m} \\ 0 & \dots & \alpha_{(m-1)m} \\ 0 & 0 & 1 \end{bmatrix}$$

(2.11)

which is a unit upper triangular matrix. There are several approaches of computing the elements of A such as Gram-Schmidt, modified Gram-Schmidt, Householder or Givens transformations (Chen, Billings and Luo 1989).

A model is said to have been satisfactorily identified when the model residual is reduced to an unpredictable sequence so that the parameter estimate is unbiased

$$E[\hat{G} - G] = 0$$

(2.12)

where G is the true parameter vector of (2.6) and $E[.]$ denotes expected value. The covariance of the parameter estimate is given by

$$Cov(\hat{G}) = E[(\hat{G} - G)(\hat{G} - G)^T] = \sigma_\varepsilon^2 [W^T W]^{-1}$$

(2.13)

where σ_ε^2 is the residual variance and throughout all sequences are assumed to be ergodic. From (2.10) the covariance of the parameter estimate $\hat{\Theta}$ is given by

$$cov(\hat{\Theta}) = A^{-1} Cov(\hat{G}) A^{-T}$$

(2.14)

2.3 Structure detection

The identification based on the model of (2.3) includes the selection of m model terms $p_j(t)$ from a full model set of M ($\gg m$) terms (typically several hundreds or even thousands of terms) and the estimation of the parameters θ_j. It has been shown previously that the orthogonal algorithms can be employed to select the model structure and estimate the parameters simultaneously (Billings, Korenberg and Chen 1988).

The orthogonal term selection is formulated using the error reduction ration (ERR) defined as

$$ERR_m = [err_1 \dots err_m]^T = \frac{\hat{G}^T W^T W \hat{G}}{Y^T Y}$$

(2.15)

To find m optimal model terms a stepwise procedure is applied to the full model set. At each step the model term with the maximum err_j value from all of the full model terms excluding previously selected terms is selected. The selection procedure is terminated at the mth step when

$$1 - \sum_{j=1}^{m} err_j \tag{2.16}$$

is less than a desired tolerance. The final orthogonal model is selected as (2.6) and the parameter estimation associated with this model was described in section 2.2. The term selection procedure can be recognized as a series of steps to reduce the model residual or residual to output ratio. The justification for this can be seen by taking time average of the square of (2.6) and utilizing the orthogonality properties to yield

$$\frac{1}{N}\sum_{t=1}^{N} y^2(t) = \frac{1}{N}\sum_{t=1}^{N} w_j^2(t)\,\hat{g}_j^2 + \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t) \tag{2.17}$$

Defining

$$\sigma_y^2 = \frac{1}{N}\sum_{t=1}^{N} y^2(t) = \overline{y^2(t)} \qquad \sigma_\varepsilon^2 = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t) = \overline{\varepsilon^2(t)} \tag{2.18}$$

yields the error reduction ratio or err value

$$err_j = \frac{\sum_{t=1}^{N} w_j^2(t)\,\hat{g}_j^2}{\sum_{t=1}^{N} y^2(t)} = \frac{\hat{g}_j^2\,\overline{w^2(t)}}{\overline{y^2(t)}} \tag{2.19}$$

So that from (2.17)

$$1 - \sum_{j=1}^{m} err_j = \frac{\sigma_\varepsilon^2}{\sigma_y^2} \tag{2.20}$$

The larger the value of the err term the larger the reduction in residual variance as that term is included in the model.

2.4 Model validity tests

Model validity tests are applied to check if the residual of the identified model has been reduced to an unpredictable sequence. A general algorithm for both linear and nonlinear model validation which has been derived by Billings and Zhu (1994a, b) consists of computing the following correlation functions

$$\phi_{\alpha\varepsilon^2}(\tau) = \frac{\sum_{t=1}^{N-\tau} (\alpha(t) - \bar{\alpha})(\varepsilon^2(t-\tau) - \overline{\varepsilon^2})}{\sqrt{\left(\sum_{t=1}^{N} (\alpha(t) - \bar{\alpha})^2\right)\left(\sum_{t=1}^{N} (\varepsilon^2(t) - \overline{\varepsilon^2})^2\right)}}$$

$$\phi_{\alpha u^2}(\tau) = \frac{\sum_{t=1}^{N-\tau} (\alpha(t) - \bar{\alpha})(u^2(t-\tau) - \overline{u^2})}{\sqrt{\left(\sum_{t=1}^{N} (\alpha(t) - \bar{\alpha})^2\right)\left(\sum_{t=1}^{N} (u^2(t) - \overline{u^2})^2\right)}}$$

(2.21)

where

$$\alpha(t) = y(t)\,\varepsilon(t)$$

$$\bar{\alpha} = \overline{y\varepsilon} = \frac{1}{N}\sum_{t=1}^{N} y(t)\,\varepsilon(t)$$

(2.22)

In the ideal case where the residuals are zero mean and uncorrelated with all linear and nonlinear combinations of past inputs and outputs these tests yield

$$\phi_{\alpha\varepsilon^2}(\tau) = \begin{cases} 1, \tau = 0 \\ 0, \textit{otherwise} \end{cases}$$

$$\phi_{\alpha u^2}(\tau) = 0, \forall\tau$$

(2.23)

Other alternative model validity test procedures (Billings and Woon 1986, Leontaritis and Billings 1987) also can be used to check the quality of the model residual.

3.0 Fast orthogonal technique

3.1 Parameter estimation

Since the orthogonal estimation algorithm will be used to sort through typically thousands of candidate model terms it is important to study the computational efficiency of the pro-

cedures involved and if possible to reformulate these to yield fast versions. To derive a fast algorithm multiply out the ordinary orthogonal algorithm presented in section 2 based on the Gram-Schmidt transformation (Korenberg, Billings, Liu and McIlroy 1988) to give

$$
\begin{aligned}
w_k(t) &= p_k(t) - \sum_{i=1}^{k-1} \alpha_{ik} w_i(t) \\
\alpha_{ik} &= \frac{\overline{p_k(t)\, w_i(t)}}{\overline{w_i^2(t)}} \\
\hat{g}_k &= \frac{\overline{y(t)\, w_k(t)}}{\overline{w_k^2(t)}} \\
err_k &= \frac{\hat{g}_k^2 \overline{w_k^2(t)}}{\overline{y^2(t)}} = \frac{(\overline{y(t)\, w_k(t)})^2}{\overline{w_k^2(t)}\ \overline{y^2(t)}}
\end{aligned}
\tag{3.1}
$$

The parameters in the ordinary model of (2.3) can be computed using transformation

$$
\hat{\theta}_i = \hat{g}_i - \sum_{k=i+1}^{m} \alpha_{ik} \hat{\theta}_k \ , \qquad \hat{\theta}_m = \hat{g}_m
\tag{3.2}
$$

Inspection of (3.1) shows that the order of processing in the ordinary orthogonal algorithm involves computing the orthogonal term $w_k(t)$ first and then all the transformations α_{ik}, estimates g_k and err_k. A faster algorithm can be derived based on computing the estimates using the correlation computations instead of the orthogonal terms themselves so that the computation of the orthogonal terms becomes unnecessary. The detailed derivations, which exploit the orthogonality, are given below by considering the component terms in (3.1)

$$
\begin{aligned}
\overline{p_k(t)\, w_i(t)} &= \overline{p_k(t) \left(p_i(t) - \sum_{j=1}^{i-1} \alpha_{ji} w_j(t) \right)} \\
&= \overline{p_k(t)\, p_i(t)} - \sum_{j=1}^{i-1} \alpha_{ji} \overline{p_k(t)\, w_j(t)}
\end{aligned}
$$

$$\overline{w_k^2(t)} = \overline{\left(p_k(t) - \sum_{i=1}^{k-1} \alpha_{ik} w_i(t)\right)^2}$$
$$= \overline{p_k^2(t)} - \sum_{i=1}^{k-1} \alpha_{ik}^2 \overline{w_i^2(t)}$$

$$\overline{y(t)\, w_k(t)} = \overline{y(t)\left(p_k(t) - \sum_{i=1}^{k-1} \alpha_{ik} w_i(t)\right)}$$
$$= \overline{y(t)\, p_k(t)} - \sum_{i=1}^{k-1} \alpha_{ik} \overline{y(t)\, w_i(t)}$$

(3.3)

Define

$$R(k,i) = \overline{p_k(t)\, w_i(t)} \qquad R(k,k) = \overline{w_k^2(t)} \qquad C(k) = \overline{y(t)\, w_k(t)}$$

(3.4)

then (3.3) becomes

$$R(k,i) = \overline{p_k(t)\, p_i(t)} - \sum_{j=1}^{i-1} \alpha_{ji} R(k,j) \qquad k > i,\ i = 1, \ldots, k-1$$
$$R(k,k) = \overline{p_k^2(t)} - \sum_{j=1}^{k-1} \alpha_{jk}^2 R(j,j) \qquad C(k) = \overline{y(t)\, p_k(t)} - \sum_{j=1}^{k-1} \alpha_{jk} C(j)$$

(3.5)

with initial settings

$$R(k,1) = \overline{p_k(t)\, p_1(t)} \qquad R(1,1) = \overline{p_1^2(t)} \qquad C(1) = \overline{y(t)\, p_1(t)}$$

(3.6)

The estimates of the orthogonal regression term $w_k(t)$ in (3.1) can now be obtained much more efficiently as

$$\alpha_{jk} = \frac{R(k,j)}{R(j,j)} \qquad \hat{g}_k = \frac{C(k)}{R(k,k)} \qquad err_k = \frac{C^2(k)}{R^2(k,k)\ \overline{y^2(t)}}$$

(3.7)

Under the fast formulation the parameters in the ordinary model of (2.3) are computed as

$$\hat{\theta}_i = \hat{g}_i - \sum_{k=i+1}^{m} \frac{R(k,i)}{R(i,i)} \hat{\theta}_k , \qquad \hat{\theta}_m = \hat{g}_m$$

(3.8)

3.2 Computational comparisons

The computational requirements for the two algorithms are given in Table 1

computation	Ordinary orthogonal	Fast orthogonal
Φ matrix	$N*M$	$N*m$
W matrix	$1/2*N*(2M-m)$	0
A matrix	$1/2*N*(2M-m)$	$1/2*N*(2M-m)$
G and ERR vectors	$1/2*(2M-m)$	$1/2*(2M-m)$

Table 1

where N is the data length, M is the number of candidate terms in the full model and m is the number of terms in the selected model. Each number listed denotes a number of basic unit computations. Two obvious improvements of the fast algorithm are to get rid of directly forming the orthogonal matrix W and only to build up the selected original term matrix instead of the full original model term matrix.

The bias and covariance of the parameter estimates can be shown to be the same as for the ordinary orthogonal algorithm because all the statistical properties are maintained by the new algorithm.

4.0 Training of Radial Basis Functions neural networks

The fast orthogonal algorithm can also be used to train Radial Basis Function (RBF) neural networks. The radial basis function technique consists of choosing a function f_r with n inputs and m outputs, which has the following form (Chen, Billings, Cowan and Grant 1990)

$$f_r(x) = \Gamma_0 + \sum_{j=1}^{M_c} \Gamma_j \phi(\|x - c_j\|)$$

(4.1)

where ϕ (.) is the radial basis function, x is the input vector, $\|.\|$ denotes the euclidean norm, Γ_j is the weight vector, Γ_0 is a constant vector and c_j is the radial basis function centre. Let

$$\Gamma_j = [\gamma_{1j} \cdots \gamma_{mj}]^T, \qquad j = 0, \ldots, M_c$$

(4.2)

then (4.1) may be decomposed as

$$f_{r_i}(x) = \gamma_{i0} + \sum_{j=1}^{M_c} \gamma_{ij}\phi(\|x - c_j\|), \qquad j = 0, \ldots, M_c \tag{4.3}$$

which can be interpreted in terms of the NARMAX model formulation of (2.3).

The network can be trained using the following procedure

(i) Select an appropriate radial basis function. Typically choices of the radial basis function. are the thin-plate-spline

$$\phi(v) = v^2 \log(v) \tag{4.4}$$

the gaussian function

$$\phi(v) = exp\left(-\frac{v^2}{\beta^2}\right) \tag{4.5}$$

where β is a real constant, and the multiquadric function

$$\phi(v) = \sqrt{v^2 + \beta^2} \tag{4.6}$$

(ii) Select the centres. The functions $f_{r_i}(x)$ and $\phi(.)$ of (4.3) are equivalent to the NARMAX model output $y(t)$ and term $p_j(t)$ of (2.3) respectively, that is

$$f_{r_i}(x) = y(t) \qquad \phi(\|x - c_j\|) = p_j(t) \tag{4.7}$$

The selection of a subset of radial basis function centres from a larger number of candidate centres can thus be regarded as an example of the selection of significant regression terms or model structure detection for the NARMAX model.

(iii) Estimation of weights. The weights γ_{ij} of (4.3) equivalent to the NARMAX model parameters θ_j of (2.3), that is

$$\gamma_{ij} = \theta_j \tag{4.8}$$

Estimation of the weights therefore corresponds to the associated parameter estimation of

a NARMAX model.

The new fast orthogonal algorithm can therefore be used as a basis for structure detection and parameter estimation of polynomial NARMAX models and to determine the topology and train radial basis function neural networks.

5.0 Identification of nonlinear rational modes

Rational models are widely used in static function approximation because they provide parsimonious models of complex phenomena and have excellent extrapolation properties. Recently dynamic rational models have been introduced into nonlinear system identification (Billings and Zhu 1991, 1994c, Zhu and Billings 1991, 1993) and because of the excellent properties of these models it is important to consider how the fast orthogonal algorithm can be adapted to both detect the structure and estimate the unknown parameters for this important class of models. Expanding (2.1) as a rational function gives

$$y(t) = \frac{a(t)}{b(t)} + \varepsilon(t) = \frac{\sum_{j=1}^{num} p_{nj}(t)\,\theta_{nj}}{\sum_{j=1}^{den} p_{dj}(t)\,\theta_{dj}} + \varepsilon(t) \tag{5.1}$$

where $a(t)$ and $b(t)$ are polynomial NARMAX models defined in (2.3). In order to apply the orthogonal least squares techniques. (5.1) is expanded into a linear in the parameters expression by multiplying $b(t)$ on both sides of (5.1) and moving all the terms except

$y(t)\,p_{d1}(t)\,\theta_{d1}$ to the right hand side to give

$$Y(t) = \sum_{j=1}^{num} p_{nj}(t)\,\theta_{nj} - \sum_{j=2}^{den} y(t)\,p_{dj}(t)\,\theta_{dj} + \zeta(t) \tag{5.2}$$

where $Y(t) = y(t)\,p_{d1}(t)|_{\theta_{d1}=1}$ and $\zeta(t) = b(t)\,\varepsilon(t)$. The orthogonal transform of (5.2) can then be expressed as

$$Y(t) = \sum_{j=1}^{num+den-1} w_j(t)\,g_j + \zeta(t) \tag{5.3}$$

Because the right hand side of (5.3) contains $\varepsilon(t)$ terms multiplied by the elements of $b(t)$ the direct application of the orthogonal least squares algorithm will yield biased estimates

even when $\varepsilon(t)$ is white. this effect which does not occur for polynomial or RBF expansions is a consequence of multiplying out (5.1) to make the model linear in the parameters. Failure to properly accommodate these effects leads to severe bias in the parameter estimates and consequently the orthogonal algorithm (Zhu and Billings 1993, Billings and Zhu 1994c) must be modified as follows

$$w_k(t) = p_k(t) - \sum_{i=1}^{k-1} \alpha_{ik} w_i(t)$$

$$e_k(t) = \Delta_k(t) - \sum_{i=1}^{k-1} \alpha_{ik} e_i(t), \qquad \Delta_k(t) = \begin{cases} 0 & \text{numerator term} \\ p_{dk}(t) & \text{denominator term} \end{cases}$$

$$\alpha_{ik} = \frac{\overline{p_k(t)\, w_i(t)} - \overline{\Delta_k(t)\, e_i(t)}\, \sigma_\varepsilon^2}{\overline{w_i^2(t)} - \overline{e_i^2(t)}\, \sigma_\varepsilon^2}$$

$$\hat{g}_k = \frac{\overline{y(t)\, w_k(t)} - \overline{p_{d1}(t)\, e_k(t)}\, \sigma_\varepsilon^2}{\overline{w_k^2(t)} - \overline{e_k^2(t)}\, \sigma_\varepsilon^2}$$

$$err_k = \frac{\hat{g}_k^2 (\overline{w_k^2(t)} - \overline{e_k^2(t)}\, \sigma_\varepsilon^2) - 2\hat{g}_k \overline{b(t)\, e_k(t)}\, \sigma_\varepsilon^2}{\overline{y^2(t)}\ \ \overline{b^2(t)}}$$

$$\hat{\theta}_k = \hat{g}_k - \sum_{j=k+1}^{m-1} \alpha_{kj} \hat{\theta}_j, \qquad \hat{\theta}_m = \hat{g}_m$$

$$b(t) = \sum_{j=1}^{den} p_{jd}(t)\, \hat{\theta}_{jd}$$

$$\sigma_\varepsilon^2 = \frac{1}{N - md} \sum_{t=md+1}^{N} \left(y(t) - \frac{a(t)}{b(t)}\right)^2$$

(5.4)

The elements which multiply residual variance σ_ε^2 are called bias correction components. A fast implementation of (5.4) can also be derived because $\overline{y(t)\, w_k(t)}$, $\overline{w_k^2(t)}$ and $\overline{y(t)\, w_k(t)}$ can be reformulated based upon the fast orthogonal computations in section

3.1. The implementation of the bias correction components, from (5.4), are given by

$$\overline{\Delta_k(t)\,e_i(t)} = \overline{\Delta_k(t)\left(\Delta_i(t) - \sum_{j=1}^{i-1}\alpha_{ji}e_j(t)\right)}$$

$$= \overline{\Delta_k(t)\,\Delta_i(t)} - \sum_{j=1}^{i-1}\alpha_{ji}\overline{\Delta_k(t)\,e_j(t)}$$

$$\overline{e_k^2(t)} = \overline{\left(\Delta_k(t) - \sum_{i=1}^{k-1}\alpha_{ik}e_i(t)\right)^2}$$

$$= \overline{\Delta_k^2(t)} - \sum_{i=1}^{k-1}\alpha_{ik}^2\overline{e_i^2(t)}$$

$$\overline{p_{d1}(t)\,e_k(t)} = \overline{p_{d1}(t)\left(\Delta_k(t) - \sum_{i=1}^{k-1}\alpha_{ik}e_i(t)\right)}$$

$$= \overline{p_{d1}(t)\,\Delta_k(t)} - \sum_{i=1}^{k-1}\alpha_{ik}\overline{p_{d1}(t)\,e_i(t)}$$

$$\overline{b(t)\,e_k(t)} = \overline{b(t)\left(\Delta_k(t) - \sum_{i=1}^{k-1}\alpha_{ik}e_i(t)\right)}$$

$$= \overline{b(t)\,\Delta_k(t)} - \sum_{i=1}^{k-1}\alpha_{ik}\overline{b(t)\,e_i(t)}$$

(5.5)

Define

$$R_e(k,i) = \overline{\Delta_k(t)\,e_i(t)} \qquad R_e(k,k) = \overline{e_k^2(t)}$$

$$C_e(k) = \overline{p_{d1}(t)\,e_k(t)} \qquad B_e(k) = \overline{b(t)\,e_k(t)}$$

(5.6)

Such that (5.5) becomes

$$R_e(k,i) = \overline{\Delta_k(t)\,\Delta_i(t)} - \sum_{j=1}^{i-1}\alpha_{ji}R_e(k,j) \qquad k > i,\ i = 1,\ldots,k-1$$

$$R_e(k,k) = \overline{\Delta_k^2(t)} - \sum_{j=1}^{k-1} \alpha_{jk}^2 R_e(j,j)$$

$$C_e(k) = \overline{p_{d1}(t)\Delta_k(t)} - \sum_{j=1}^{k-1} \alpha_{jk} C_e(j)$$

$$B_e(k) = \overline{b(t)\Delta_k(t)} - \sum_{j=1}^{k-1} \alpha_{jk} C_e(j)$$

(5.7)

with initial settings

$$R_e(k,1) = \overline{\Delta_k(t)\Delta_1(t)} \qquad R_e(1,1) = \overline{\Delta_1^2(t)}$$

$$C_e(1) = \overline{p_{d1}(t)\Delta_1(t)} \qquad B_e(1) = \overline{b(t)\Delta_1(t)}$$

(5.8)

The estimates computed without directly forming the orthogonal regressor $w_k(t)$ or the error term $e_k(t)$ can be obtained based upon the fast implementation

$$\alpha_{jk} = \frac{R(k,j) - R_e(k,j)\sigma_\varepsilon^2}{R(j,j) - R_e(k,k)\sigma_\varepsilon^2} \qquad \hat{g}_k = \frac{C(k) - C_e(k)\sigma_\varepsilon^2}{R(k,k) - R_e(k,k)\sigma_\varepsilon^2}$$

$$err_k = \frac{\hat{g}_k^2 (R(k,k) - R_e(k,k)\sigma_\varepsilon^2) - 2\hat{g}_k^2 B_e(k)\sigma_\varepsilon^2}{\overline{Y^2(t)}\ \ \overline{b^2(t)}}$$

(5.9)

The parameter estimates in the rational model of (5.1) can then be recovered using

$$\hat{\theta}_i = \hat{g}_i - \sum_{k=i+1}^{m} \frac{R(k,i) - R_e(k,i)\sigma_\varepsilon^2}{R(i,i) - R_e(i,i)\sigma_\varepsilon^2}\hat{\theta}_k\ , \qquad \hat{\theta}_m = \hat{g}_m$$

(5.10)

Obviously the fast orthogonal rational model identification algorithm given in (5.9) reduces to the algorithm presented in (3.7) when the denominator in the rational model is set to unity ($b(t)$=1) so that all the components in (5.9) which are multiplied by $e_k(t)$

become zero because of $e_k(t)=0$.

6.0 Simulated examples

Two simulated systems were selected to demonstrate the effectiveness of the new fast algorithm. The fast algorithm was implemented using the rational model formulation to provide generality since this is applicable to polynomial, RBF and rational models. For each of the simulated systems 1000 pairs of input and output data were used with the same choice of input and noise signals in each case. The input $u(t)$ was a uniformly distributed random excitation sequence with zero mean and amplitude ± 1 (variance $\sigma_u^2 = 0.33$) and the noise $e(t)$ was a normally distributed disturbance sequence with zero mean and variance $\sigma_e^2 = 0.01$. An initial full model of 56 terms was specified with numerator degree = denominator degree = 2 and input lag = output lag = noise lag = 2.

Example S_1

The first simulated nonlinear rational system was defined as

$$y(t) = \frac{0.25y(t-1) - 0.85y(t-2)e(t-2) + u(t-1)u(t-2)}{1 + 0.68y^2(t-1) + y^2(t-2)} + e(t) \tag{6.1}$$

In this form the model is a highly nonlinear with nonlinearity in the parameters, input, output and noise. The input and noise corrupted output data sequences are shown in Fig. 1. Applying the fast orthogonal algorithm to search through all the 56 possible candidate terms to detect the structure or significant model terms and to estimate the unknown parameters produced the results shown in Table 1. The final estimated model contains just five terms and the one step ahead predictions and residuals are illustrated in Fig. 2. The model validity tests are shown in Fig. 3.

Numerator polynomial	Parameter estimates
y(t-1)	0.269
y(t-2)e(t-2)	-0.613
u(t-1)u(t-2)	1.012
denominator polynomial	
y^2(t-1)	0.654
y^2(t-2)	0.963
Residual variance	
σ_ε^2	0.0994

Table 1 Identified model for S_1

Example S_2

A second simulated nonlinear system consisting of the polynomial model

$$y(t) = 0.54y(t-1) + 0.95y(t-2)u(t-1) + u(t-1) + 0.77u^2(t-2) + u(t-1)e(t-2) + e(t) \tag{6.2}$$

was simulated. This system was used to demonstrate the generality of the fast algorithm, which reduces to the fast orthogonal polynomial model algorithm when the denominator in the rational model is set to one, $b(t)=1$. The input and noise corrupted output data sequences are shown in Fig. 4. The identified model which was obtained by searching over 56 candidate model terms is shown in Table2. The term selection and parameter estimation procedures of the fast algorithm produce a final model with just five terms. The one step ahead predictions and residuals are illustrated in Fig. 5, and the model validity tests are illustrated in Fig. 6.

Numerator	Parameter estimates
y(t-1)	0.541
y(t-2)u(t-1)	0.948
u(t-1)	1.000
u^2(t-2)	0.776
u(t-1)e(t-2)	0.900
Residual variance	
σ_ε^2	0.0989

Table 2 Identified model for S_2

7.0 Conclusions

A new class of fast orthogonal estimation routines have been introduced for polynomial

models, radial basis function neural networks and nonlinear rational models. The new algorithms provide a computationally efficient implementation which can be used to rapidly search through a large class of candidate model terms, to order the terms according to their significance and hence to determine the structures of the model. Unbiased parameter estimations can be obtained in the presence of both additive and multiplicative correlated noise. The fast algorithms are typically two to three times faster than the original formulations making this approach computationally efficient while preserving the flexibility that the orthogonal properties provided.

Acknowledgment

The authors gratefully acknowledge that this work was supported by EPSRC under grant GR/H3528.6.

References

Billings, S. A., Chen, S., 1989, Extended model set, global data and threshold model identification of severely nonlinear systems. *Int. J. Control*, **50**, 1897-1923.

Billings, S. A., Korenberg, M. J. and Chen, S., 1988, Identification of nonlinear output affine systems using an orthogonal least squares algorithm. *Int. J.* Systems Sci., **19**, 1159-1568.

Billings, S. A. and Voon, W. S. F., 1984, Least squares parameter estimation algorithms for nonlinear systems. *Int. J. Systems Sci.*, **15**, 601-615; 1986, Correlation based model validity tests for nonlinear models. *Int. J. Control*, **44**, 235-244.

Billings, S. A. and Zhu, Q. M., 1991, Rational model identification using an extended least squares algorithm. *Int. J. Control*, **54**, 529-546; 1994a, Nonlinear model validation using correlation tests. *Int. J. Control*, (to appear); 1994b, Model validity tests for multivariable nonlinear models including

neural networks. *Int. J. Control*, (to appear); 1994c, Structure detection algorithm for nonlinear rational models. *Int. J. Control*, **59**, 1439-1463.

Blahut, R.E., 1985, Fast algorithms for signal processing. *Addison-Wesley Publishing Company, Reading.*

Broomhead, D.S. and Lowe, D., 1988, Multivariable functional interpolation and adaptive networks. *Complex Systems*, **2**, 321-355.

Chen, S., Billings, S. A. and Luo, W., 1989, Orthogonal least squares methods and their application to nonlinear system identification. *Int. J. Control*, **50**, 1873-1896.

Chen, S., Billings, S. A., Cowan, C.F.N. and Grant, P. W., 1990, Nonlinear systems identification using radial basis functions. *Int. J. Systems Sci.*, **21**, 2513-2539.

Haykin, S., 1986, Adaptive filter theory. *Prentice-Hall, Englewood Cliffs.*

Korenberg, M. J., 1988, Identifying nonlinear difference equation and functional expression representations: the fast orthogonal algorithm. *Ann. Biomed. Eng.*, **16**, 123-142.

Korenberg, M. J. and Paarmann, L.D., 1991, Orthogonal approaches to time series analysis and system identification. *IEEE SP Magazine*, **7**, 29-43.

Leontaritis, I. J., and Billings, S. A., 1987, Model selection and validation methods for nonlinear systems. *Int. J. Control*, **45**, 311-341.

Ljung, L., 1987, System identification--Theory for the user. *Prentice Hall Englewood Cliffs, New Jersey.*

Ljung, L. and Soderstrom, T., 1983, Theory and practice of recursive identification. *MIT Press, Cambridge.*

Micchelli, C.A., 1986, Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, **2**, 11-22.

Powell, M.J.D., 1985, Radial basis functions for multivariable interpolation: a review. *IMA Conference on Algorithms for the Approximation of Functions and Data, RMCS Shrivenham.*

Soderstrom, T. and Stoica, P., 1990, On covariance function tests used in system identification. *Automatica*, **26**, 125-133.

Zhu, Q. M. and Billings, S. A., 1991, Recursive parameter estimation for nonlinear rational models. *J. Sys. Eng.*, **1**, 63-67; 1993, Parameter estimation for stochastic nonlinear rational models. *Int. J. Control*, **57**, 309-333; 1994, Modelling of complex systems using nonlinear rational approximations. *International Symposium of Young Investigators on Information, Computer and Control, Beijing, China.*

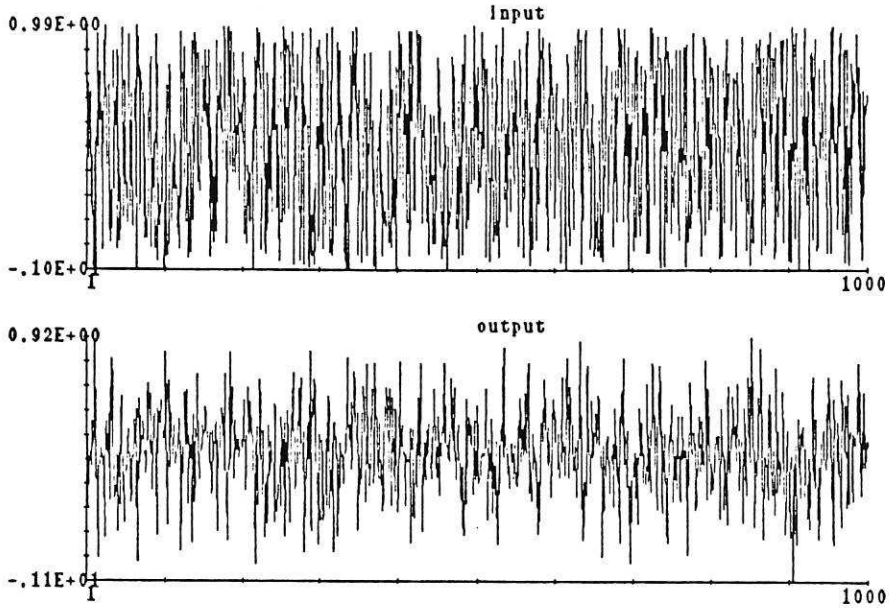

Fig 1 Input and output data sequences for Example S_1

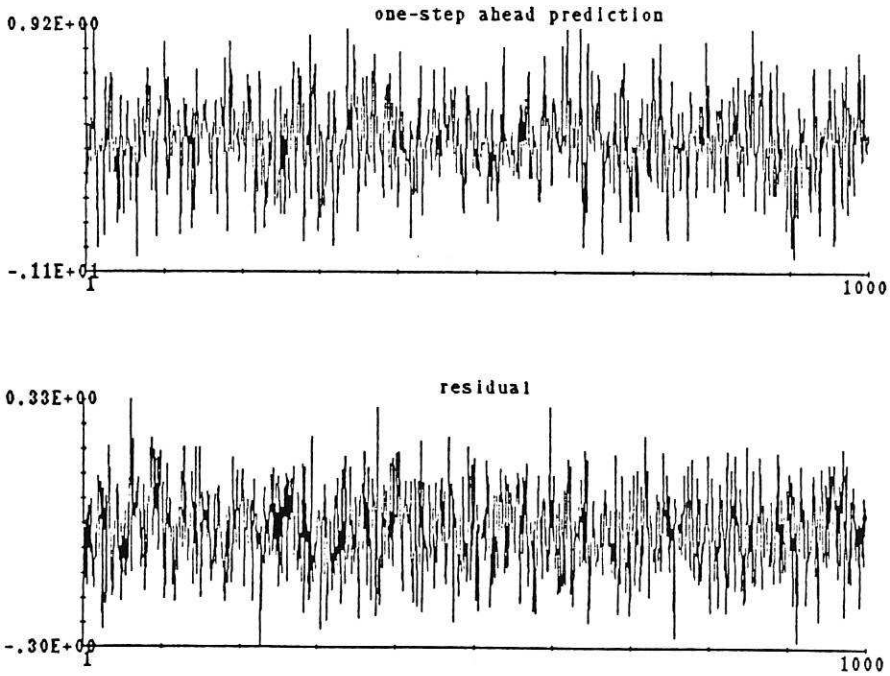

Fig 2 One step ahead predictions and residuals for Example S_1

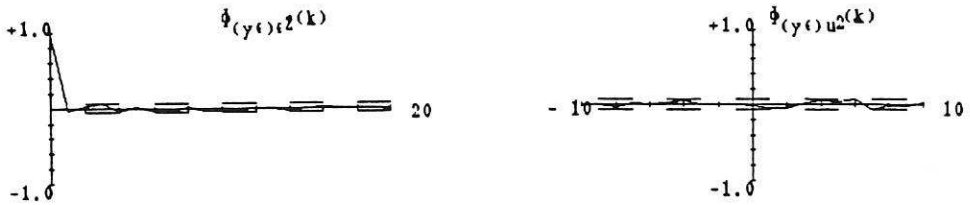

Fig 3 Model validity tests for Example S_1

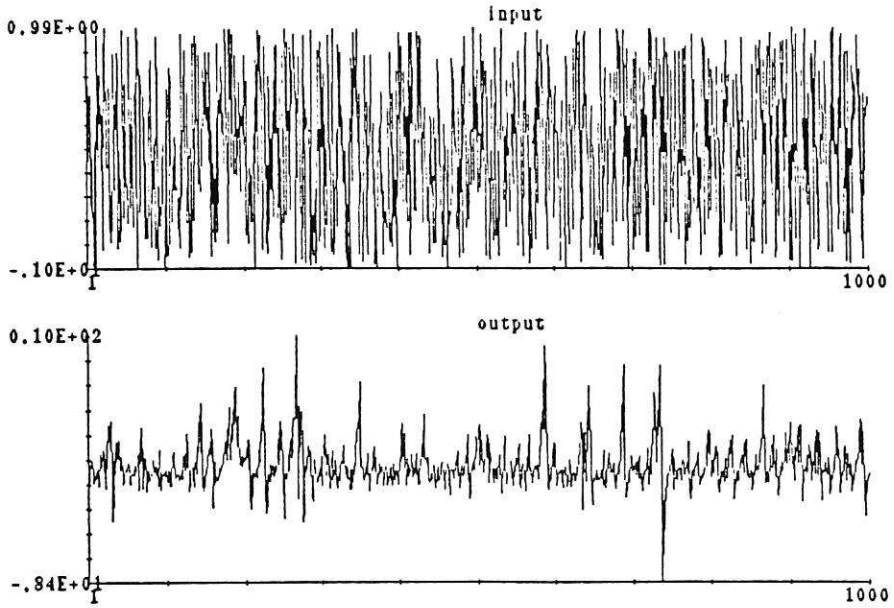

Fig 4 Input and output data sequences for Example S_2

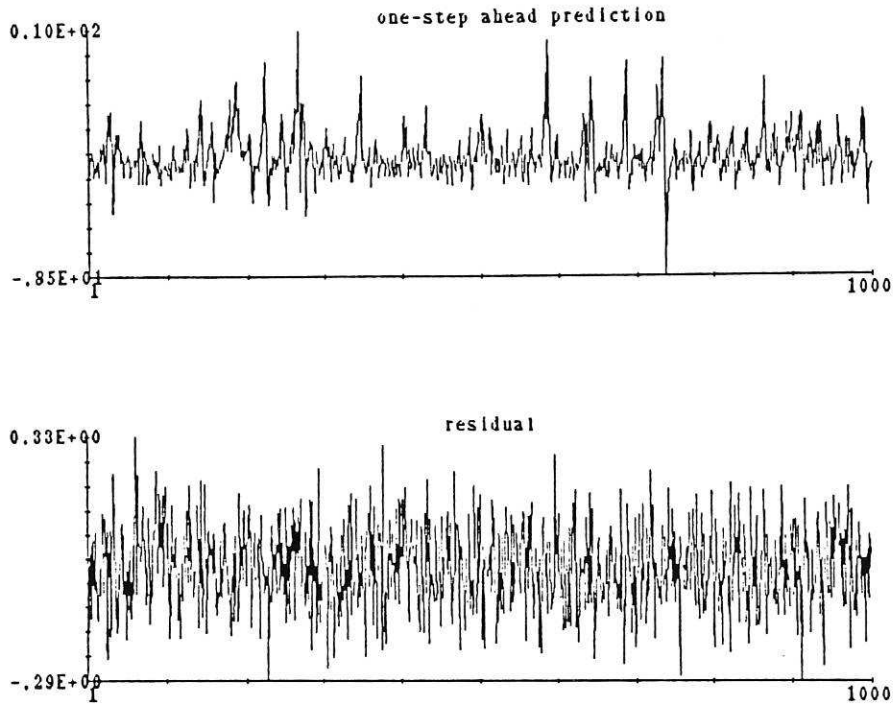

Fig 5 One step ahead predictions and residuals for Example S_2

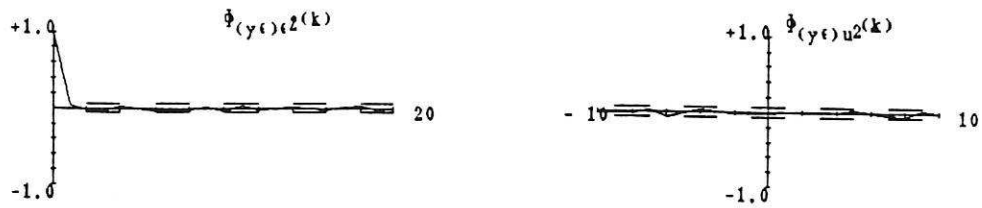

Fig 6 Model validity tests for Example S_2