An Adaptive Orthogonal Least Squares Algorithm for Model Subset Selection and Nonlinear System Identification

H. L. Wei and S. A. Billings

Research Report No. 884

Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street, Sheffield,
S1 3JD, UK

Feb. 2005

An Adaptive Orthogonal Least Squares Algorithm for Model Subset Selection and Nonlinear System Identification

H. L. Wei and S. A. Billings
Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK
S.Billings@Sheffield.ac.uk, W.Hualiang@Sheffield.ac.uk

Abstract: A new adaptive orthogonal least squares (AOLS) algorithm is proposed for model subset selection and nonlinear system identification. Model subset selection, or model structure detection, is a key step in any identification procedure and consists of detecting and selecting significant model terms from a redundant candidate model term set to determine a parsimonious final model. In the proposed new AOLS algorithm, a new indicator called the error-to-signal ratio (ESR) and a new R^2-like statistic, the adjustable prediction error sum of squares (R^2-APRESS), are introduced and combined with the widely used orthogonal least squares algorithms. The new AOLS algorithm integrates the selection of significant model terms, with the automatic determination of the optimal number of model terms, and the estimation of the unknown model parameters.

Keywords: information criteria, model subset selection, model structure detection nonlinear system identification, orthogonal least squares, prediction error sum of squares

1. Introduction

A wide class of input-output nonlinear dynamical systems can be represented by the NARX (*N*onlinear *A*uto*R*egressive with e*X*ogenous inputs) model of the form

$$y(t) = f(y(t-1),\cdots,y(t-n_y),u(t),u(t-1),\cdots,u(t-n_u)) + e(t) \tag{1}$$

where the nonlinear mapping f is often unknown and needs to be identified from given observational data of the input $u(t)$ and the output $y(t)$; n_u and n_y are the maximum input and output lags; $e(t)$ is the modelling error. The nonlinear mapping f can be constructed using a variety of local or global basis functions including polynomials, kernel functions, splines, radial basis functions, neural networks and wavelets. Such a NARX model constructed on the basis of a set of known basis functions with a specified form can often be expressed using a linear-in-the-parameters form

$$y(t) = \sum_{m=1}^{M} \theta_m \phi_m(t) + e(t) \tag{2}$$

where $\phi_m(t) = \phi_m(\varphi(t))$ are model terms generated in some way from the regression vector $\varphi(t)$ $= [y(t-1),\cdots, y(t-n_y), u(t),\cdots, u(t-n_u)]^T$, θ_m are unknown parameters, and M is the number of total potential model terms involved. One of the most popular representations is the polynomial model, where the

candidate model terms $\phi_m(t)$ are of the form $x_1^{i_1}(t)\cdots x_\ell^{i_\ell}(t)$, where $x_j^{i_j}(t) \in \{y(t-1),\cdots, y(t-n_y),\ u(t),\cdots, u(t-n_u)\}$ for j=1, ..., ℓ, with $0 \le i_j \le \ell$ and $0 \le i_1 + \cdots + i_\ell \le \ell$. The order of such a polynomial model is determined by n_y and n_u, and the nonlinear degree of such a model is referred to as ℓ.

The linear ARX model, which is a special case of the NARX model, has been extensively studied in the literature and several approaches have been developed for order and variable selection of the ARX and similar models. The conventional Akaike information criterion (AIC), Bayesian information criterion (BIC) and final prediction error (FPE), and the minimum description length (MDL) are among the most commonly used techniques for model order selection in linear system models(Akaike 1969, 1970, 1974, Schwartz 1978, Rissanen 1978, 1983, Wax and Klaith 1985). These criteria were developed fully or partly on the basis of the maximum likelihood principle with some assumptions on the signals involved. These well-established criteria can be used to select the model order and significant variables for data sets that can be well characterized by linear AR and ARX models, where model terms and variables are the same, they are all regressors. These criteria, however, cannot be used directly for subset selection of nonlinear models, where model terms and variables are typically distinguished. The distinction between variables and terms is important and can be illustrated using the simple nonlinear polynomial model below

$$\begin{aligned} y &= f(x_1, x_2, x_3) \\ &= a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_1^2 + a_5 x_2^2 + a_6 x_3^2 + a_7 x_1 x_2 + a_8 x_1 x_3 + a_9 x_2 x_3 \end{aligned} \tag{3}$$

There are only 3 variables: x_1, x_2 and x_3, but there are 10 terms: a_0 (a *const* term), $x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3$ and x_2x_3.

The initial linear-in-the-parameters model (2) may involve a great number of candidate model terms whatever basis functions are employed to approximate the unknown nonlinear mapping f, especially when the maximum lags n_u and n_y are large. Experience shows that in most cases only a small number of significant model terms are necessary in the final model to represent given observational data. Most candidate model terms are either redundant or make very little contribution to the system output and can therefore be removed from the model. An efficient model structure determination approach has been developed based on the forward orthogonal least squares (OLS) algorithm and the error reduction ratio (ERR) criterion, which was originally introduced to determine which terms should be included in a model (Korenberg *et al.* 1988, Billings *et al.*, 1989; Chen *et al.*, 1989). This approach has been extensively studied and widely applied in nonlinear system identification (Chen *et al.* 1991, Wang and Mendel 1992, Henrique et al. 2000, Hong and Harris 2001, Yong *et al.* 2001, Wei and Billings 2004). The OLS-ERR algorithm provides a powerful tool to effectively select significant model terms step by step, one at a time, by orthogonalizing the associated regressors in a forward stepwise way based on the ERR criterion, an index indicating the significance of each model term. Existing OLS algorithms, however, do not provide information on how many significant model terms should be selected and included in the final model. An additional separate procedure is therefore often needed to aid the determination of the optimal number of significant model terms.

A commonly used approach to evaluate the performance of an identified model is to check the extrapolation ability, or the predictive capability, by applying the identified model to some fresh data, which was not used for

model estimation. With this motivation, several cross-validation methods have been proposed (Allen 1971, 1974, Stone 1974, Snee 1977, Li 1986, 1987).

Motivated by the successful applications of the OLS algorithms and cross-validation, this study aims to develop an adaptive orthogonal least squares (AOLS) scheme that can be used to select not only the significant model terms but also the optimal number of model terms to arrive at a good balance for the bias-variance trade-off. In the new AOLS algorithm, a new criterion, the error-to-signal ratio (ESR), and a new R^2-like cross-validation statistic, the adjustable prediction error sum of squares (R^2-APRESS), are introduced and combined with the traditional OLS algorithms. The new AOLS scheme has been developed to achieve the following objectives: i) to detect significant model terms and put the selected terms in order of significance and contribution made to the system output; ii) to determine the optimal number of model terms to arrive at a good balance between the bias-variance trade-off and, iii) to estimate the unknown model parameters.

This paper is organized as follows. In Section 2, the basic idea of orthogonal transformation methods for model term selection is summarised. In Section 3, a new criterion for selecting the optimal number of model terms is proposed. The new AOLS algorithm is described in detail in Section 4. In Section 5, several examples are given to illustrate the efficiency of the new AOLS algorithm. In Section 6, several modified information criteria are proposed. Noise modelling is also discussed in this section. The work is concluded in Section 7.

2. Model term selection and the orthogonal transformation

Consider the term selection problem for the linear-in-the-parameters model (2). Let $\mathbf{y} = [y(1), \cdots, y(N)]^T$ be a vector of measured outputs at N time instants, and $\alpha_m = [\phi_m(1), \cdots, \phi_m(N)]^T$ be a vector associated with the mth candidate model term, where m=1,2, ..., M. From the viewpoint of practical modelling and identification, the finite dimensional set $\Gamma = \{\alpha_1, \cdots, \alpha_M\}$ is often redundant. The model term selection problem is equivalent to finding a full dimensional subset $\Gamma_n = \{\beta_1, \cdots, \beta_n\} = \{\alpha_{i_1}, \cdots, \alpha_{i_n}\} \subseteq \Gamma$, where $\beta_k = \alpha_{i_k}$, $i_m \in \{1,2,\cdots,M\}$ and m=1,2, ..., n, so that $\mathbf{y}$ can be satisfactorily approximated using a linear combination of $\beta_1, \cdots, \beta_n$ as below

$$\mathbf{y} = \theta_1\beta_1 + \cdots + \theta_n\beta_n + \mathbf{e} \tag{4}$$

or in a compact matrix form

$$\mathbf{y} = P\theta + \mathbf{e} \tag{5}$$

where the matrix $P = [\beta_1, \cdots, \beta_n]$ is of full column rank, $\theta = [\theta_1, \cdots, \theta_n]^T$ is a parameter vector, and $\mathbf{e}$ is an approximation error. From matrix theory, the full rank matrix P can be orthogonally decomposed as

$$P = QR \tag{6}$$

where R is an $n \times n$ unit upper triangular matrix and Q is an $n \times n$ matrix with orthogonal columns $q_1, q_2, \cdots, q_n$. Substituting (6) into (5), yields

$$\mathbf{y} = (PR^{-1})(R\theta) + \mathbf{e} = Qg + \mathbf{e} \tag{7}$$

where $g=[g_1,\cdots,g_n]^T=R\theta$ is an auxiliary parameter vector. Using the orthogonal property of Q, g_i can be directly calculated from $\mathbf{y}$ and Q as $g_i=(\mathbf{y}^T q_i)/(q_i^T q_i)$ for i=1,2, ..., n. The unknown parameter vector θ can then be easily calculated from g and R by substitution using the special structure of R.

Assume that the error $\mathbf{e}$ in model (7) is uncorrelated with vectors β_j for j=1,2, ..., n, the total sum of squares of the output from the origin can then be expressed as

$$\mathbf{y}^T\mathbf{y}=\sum_{i=1}^{n} g_i^2 q_i^T q_i+\mathbf{e}^T\mathbf{e} \tag{8}$$

Note that the total sum of squares $\mathbf{y}^T\mathbf{y}$ consists of two parts, the desired output $\sum_{i=1}^{n} g_i^2 q_i^T q_i$, which can be explained by the selected regressors (model terms), and the part $\mathbf{e}^T\mathbf{e}$, which represents the residual sum of squares. Thus, $g_i^2 q_i^T q_i$ is the increment to the desired total sum of squares of the output brought by q_i. The ith error reduction ratio (ERR) introduced by q_i (or equally by including β_i), is defined as

$$\mathrm{ERR}[i]=\frac{g_i^2(q_i^T q_i)}{\mathbf{y}^T\mathbf{y}}\times 100\%=\frac{(\mathbf{y}^T q_i)^2}{(\mathbf{y}^T\mathbf{y})(q_i^T q_i)}\times 100\%,\quad i=1,2,\ldots,n, \tag{9}$$

This ratio provides a simple but an effective index to indicate the significance of adding the ith term into the model. The orthogonalization procedure for model term selection is usually implemented in a stepwise way, one term at a time. The sum of error reduction ratio (SERR) and error-to-signal ratio (ESR) due to $q_1,\cdots,q_j$ (or equally due to $\beta_1,\cdots,\beta_j$) are defined as

$$\mathrm{SERR}[j]=\sum_{i=1}^{j}\mathrm{ERR}[i] \tag{10}$$

$$\mathrm{ESR}[j]=\frac{\mathbf{e}^T\mathbf{e}}{\mathbf{y}^T\mathbf{y}}=1-\sum_{i=1}^{j}\frac{g_i^2 q_i^T q_i}{\mathbf{y}^T\mathbf{y}}=1-\sum_{i=1}^{j}\mathrm{ERR}[i]=1-\mathrm{SERR}[j] \tag{11}$$

The selection procedure will be terminated when the ESR of an identified model satisfies some specified conditions. Several orthogonal transforms including Gram-Schmidt, modified Gram-Schmidt and Householder transformations can be applied to implement the orthogonal decomposition (Billings *et al.* 1989, Chen *et al.* 1989) and a detailed algorithm will be given in Section 4.

3. The determination of the optimal number of model terms

The determination of the optimal number of model terms is critical in dynamical modelling. Neither an over-fitting nor an under-fitting model is desirable in practical identification. In practice, however, the true number of terms is generally unknown and needs to be estimated during model identification. Several approaches have been developed for model order and variable selection in the literature including the AIC, BIC, MDL (Akaike 1969, 1970, 1974, Schwartz 1978, Rissanen 1983a, 1983b, Wax and Klaith 1985) and many variants (Miller 1990, Chap. 6). In this study, a R^2-like statistic, the adjustable prediction error sum of squares (R^2-PRESS) proposed by Allen (1971, 1974), is modified and will be used to solve the term selection problem.

The R^2 and the adjustable R^2-statistics are respectively defined as

$$R^2 = 1 - \text{NMSE} \tag{12}$$

and

$$R_a^2 = 1 - \frac{N-1}{N-n}\text{NMSE} \tag{13}$$

where N is the data length, n is the number of model terms included in the identified model, NMSE is the normalised-mean-square-error defined as

$$\text{NMSE} = \frac{\text{SSE}}{\text{SST}} = \frac{\sum_{i=1}^{N}[y(i)-\hat{y}(i)]^2}{\sum_{i=1}^{N}[y(i)-\bar{y}]^2} \tag{14}$$

where $\text{SST} = \sum_{i=1}^{N}[y(i)-\bar{y}]^2$ denotes the total sum of squared deviations in $\mathbf{y}$ from the mean $\bar{\mathbf{y}}$, $\text{SSE} = \sum_{i=1}^{N}[y(i)-\hat{y}(i)]^2$ denotes the sum of the squared errors (residuals), $\{\hat{y}(i)\}_{i=1}^{N}$ is the one-step-ahead prediction sequence from the identified model with n terms. While for two models with the same number of model terms, the model with the higher value of R^2 is preferred, for two models containing different numbers of model terms, the model with the higher R_a^2 is often preferred (Chatterjee and Hadi 1988).

The prediction error sum of squares (PRESS) proposed by Allen (1971, 1974) provides a useful residual scaling, which can be used as a form of cross validation by leaving one point out at a time (Myers 1990). The prediction error sum of squares is defined as

$$\text{PRESS} = \sum_{i=1}^{N}[y(i)-\hat{y}_{-i}(i)]^2 = \sum_{i=1}^{N}[\varepsilon_{-i}(i)]^2 \tag{15}$$

where $\hat{y}_{-i}(i)$ are one-step-ahead predicted values from a model fitted using a data set consisting of N-1 observational data point pairs, which are obtained by leaving the ith point pair out, $\varepsilon_{-i}(i)$ are the PRESS predicted residuals evaluated at the ith point. Let $\varepsilon(i)$ be the normally defined residuals of a model fitted using the total N data points, it can be shown that the relationship between $\varepsilon_{-i}(i)$ and $\varepsilon(i)$ is

$$\varepsilon_{-i}(i) = \frac{y(i)-\hat{y}(i)}{1-\beta_i^T(P^TP)^{-1}\beta_i} = \frac{\varepsilon(i)}{1-h(i,i)} \tag{16}$$

where β_i and P are defined as in (4). Thus PRESS can be reduced to

$$\text{PRESS} = \sum_{i=1}^{N}\left(\frac{\varepsilon(i)}{1-h(i,i)}\right)^2 \tag{17}$$

This shows that the PRESS statistic can be calculated by fitting only one model using the total N data points, but N "leave-one-out" matrices are still required. It can be proved (Miller 1990) that if $N >> n$, PRESS can be approximated as

$$\text{PRESS} \approx \left(\frac{N}{N-n}\right)^2 \text{SSE} \tag{18}$$

Statistic (18) gives some indication of the predictive capability of the regression model. This will be used to define the adjustable R^2-PRESS statistic given below

$$R^2_{\text{press}} = 1 - \frac{\text{PRESS}}{\text{SST}} = 1 - \left(\frac{N}{N-n}\right)^2 \frac{\text{SSE}}{\text{SST}} \tag{19}$$

Note, however, that sometimes the data length N may be long, say $N \geq 2000$. In this case, the effect of n in the denominator of (19) is minimal due to the fact that $(N/(N-n))^2 \approx 1 + 2n/N \approx 1$ for $n << N/2$. One way to avoid the tendency that small n's are mitigated by a large N is to replace the number n by λn, where λ is an adjustable coefficient. Experience shows that a typical choice for λ is to set $\lambda = \max\{1, \rho N\}$ with $0.002 \leq \rho \leq 0.01$. The adjustable R^2-PRESS can then be defined as

$$R^2_{\text{apress}} = 1 - \left(\frac{N}{N-\lambda n}\right)^2 \text{NMSE} \tag{20}$$

Note that the R^2-APRESS statistic (20) is in formulation similar to the adjustable R^2-statistic given by (13). In the next section, the R^2-APRESS statistic will be combined with the criterion ESR (error-to-signal ratio) and will then be incorporated into the orthogonal least squares algorithm.

4. The adaptive orthogonal least squares (AOLS) algorithm

At first sight, the calculation of the R^2-APRESS statistic defined by (20) requires an initial calculation of the value of NMSE, which involves the calculation of the one-step-ahead prediction, $\hat{\mathbf{y}}$. By noting the definition of ESR in (11), however, the calculation of NMSE is not necessary. In fact, from (11) and (20), the R^2-APRESS for an identified model with p model terms can be calculated as

$$\begin{aligned} R^2_{\text{apress}}[p] &= 1 - \left(\frac{N}{N-\lambda p}\right)^2 \text{NMSE}[p] \\ &= 1 - \left(\frac{N}{N-\lambda p}\right)^2 \left(\frac{\mathbf{e}^T\mathbf{e}}{\text{SST}}\right)_{[p]} \\ &= 1 - \left(\frac{\text{SST}_0}{\text{SST}}\right)\left(\frac{N}{N-\lambda p}\right)^2 \left(\frac{\mathbf{e}^T\mathbf{e}}{\text{SST}_0}\right)_{[p]} \\ &= 1 - \left(\frac{\text{SST}_0}{\text{SST}}\right)\left(\frac{N}{N-\lambda p}\right)^2 \text{ESR}[p] \end{aligned} \tag{21}$$

where $\text{SST}_0 = \mathbf{y}^T\mathbf{y} = \sum_{i=1}^{N} y^2(i)$ is the total sum of squared deviations in $\mathbf{y}$ from the origin, and SST is defined in (14), and the index or subscript $[p]$ indicates that the associated items are calculated from an identified model

with p terms. Note that $\mathrm{ESR}[p]$ (p=1,2, ...) in (21) are available as a by-product of the orthogonalization procedure.

Assume that there exists a number p_0, at which the function $R^2_{\text{apress}}[p]$ with respect to p is a maximum. At the maximum of $R^2_{\text{apress}}[p]$, the following relationships hold

$$R^2_{\text{apress}}[p_0] > R^2_{\text{apress}}[p_0 - 1] \tag{22a}$$

$$R^2_{\text{apress}}[p_0] \geq R^2_{\text{apress}}[p_0 + 1] \tag{22b}$$

A little rearrangement of (22a) and (22b) gives

$$\frac{\mathrm{ESR}(p_0)}{\mathrm{ESR}(p_0 - 1)} < \left(\frac{N - \lambda p_0}{N - \lambda(p_0 - 1)} \right)^2 \tag{23a}$$

$$\frac{\mathrm{ESR}(p_0 + 1)}{\mathrm{ESR}(p_0)} \leq \left(\frac{N - \lambda(p_0 + 1)}{N - \lambda p_0} \right)^2 \tag{23b}$$

Define two functions

$$\chi_1(p) = \frac{\mathrm{ESR}(p+1)}{\mathrm{ESR}(p)} \tag{24}$$

$$\chi_2(p) = \left(\frac{N - \lambda(p+1)}{N - \lambda p} \right)^2 \tag{25}$$

From (23a) and (23b), χ_1 and χ_2 have the following property: $\chi_1(p) < \chi_2(p)$ for $p < p_0$, and $\chi_1(p) \geq \chi_2(p)$ for $p = p_0$. The two functions defined by (24) and (25) will be used as an indicator to find the optimal model term number p_0, where the two indicating functions intersect. In fact, the optimal number p_0 can be chosen as the point where χ_1 enters into the 90% confidence interval of $\chi_2(\cdot)$ for the first time. The 90% confidence interval is defined as $\chi_2(\cdot) \pm 1.28/\sqrt{N}$.

The new adaptive orthogonal least squares algorithm (AOLS) can now be described below, where $\alpha_1, \cdots, \alpha_M$ are the vectors associated with the M candidate model terms.

The ALOS algorithm:

Step 1: Set $I_1 = \{1, 2, \cdots, M\}$; $s_0 = \mathbf{y}^T \mathbf{y}$; $s_1 = (\mathbf{y} - \overline{\mathbf{y}})^T (\mathbf{y} - \overline{\mathbf{y}})$;

for i=1 to M

$\beta_i^{(1)} = \alpha_i$;

$err^{(1)}[i] = \dfrac{(\mathbf{y}^T \beta_i^{(1)})^2}{s_0 (\beta_i^{(1)})^T \beta_i^{(1)}}$; {if $(\beta_i^{(1)})^T \beta_i^{(1)} \approx 0$, set $err^{(1)}[i] = 0$};

$a_{ii} = 1$;

end for

$\ell_1 = \arg\max_{i \in I_1} \{err^{(1)}[i]\}$; $err[1] = err^{(1)}[\ell_1]$;

$serr[1] = err[1]$; $esr[1] = 1 - serr[1]$;

$$q_1 = \beta_{\ell_1}^{(1)}; \quad g_1 = \frac{\mathbf{y}^T q_1}{q_1^T q_1};$$

Step j, $j \geq 2$:

For j=2 to M

$$I_j = I_{j-1} \setminus \{\ell_{j-1}\};$$

for $i \in I_j$

$$\beta_i^{(j)} = \beta_i^{(j-1)} - \frac{\alpha_i^T q_{j-1}^T}{q_{j-1}^T q_{j-1}} q_{j-1}^T; \tag{26}$$

$$err^{(j)}[i] = \frac{(\mathbf{y}^T \beta_i^{(j)})^2}{s_0 (\beta_i^{(j)})^T \beta_i^{(j)}}; \quad \{\text{if } (\beta_i^{(j)})^T \beta_i^{(j)} \approx 0, \text{ set } err^{(j)}[i] = 0\}; \tag{27}$$

end for (end loop for i)

$$J_j = \{\arg_{i \in I_j}((\beta_i^{(j)})^T \beta_i^{(j)} < \delta)\}; \quad I_j = I_j \setminus J_j; \tag{28}$$

$$\ell_j = \arg\max_{i \in I_j} \{err^{(j)}[i]\}; \; err[j] = err^{(j)}[\ell_j];$$

$$serr[j] = \sum_{k=1}^{j} err[k]; \; esr[j] = 1 - serr[j]$$

$$R^2_{\text{apress}}[j] = 1 - \left(\frac{N}{N - \lambda j}\right)^2 \frac{s_0}{s} esr[j];$$

$$\chi_1[j] = \frac{esr[j]}{esr[j-1]};$$

$$\chi_2[j] = \left(\frac{N - \lambda j}{N - \lambda(j-1)}\right)^2$$

$$q_j = \beta_{\ell_j}^{(j)}; \quad g_j = \frac{\mathbf{y}^T q_j}{q_j^T q_j};$$

$$a_{jj} = 1;$$

for k=1 to j-1

$$a_{kj} = \frac{\alpha_{\ell_j}^T q_k}{q_k^T q_k};$$

end for (end loop for k)

end for (end loop for j)

***Remark* 1:** The AOLS algorithm provides an effective tool for selecting significant model terms in an iterative stepwise way. Terms are selected step by step, one term at a time. Note that when selecting the jth significant term (regressor), this only involves the (j-1)th significant term (see Eq (26)). This modified version of the Gram-Schmidt orthogonal transform is therefore not only effective for selecting significant model terms but is also much more efficient in computation.

***Remark* 2:** Most numerical ill conditioning can be avoided by eliminating the candidate regressors for which $(\beta_i^{(j)})^T \beta_i^{(j)}$ are less than a predetermined threshold δ, say $\delta = 10^{-\tau}$ with $\tau \geq 10$ (see Eqs. (27), (28)).

***Remark* 3:** The assumption that the initial candidate regression vector set $\Gamma = \{\alpha_1, \cdots, \alpha_M\}$ is of full dimensionality is unnecessary in the iterative forward AOLS algorithm. In fact, if the M vectors P are linearly dependent, and assuming that the dimension of Γ is n ($<M$), the algorithm will stop at the n-th step.

***Remark* 4:** If required, the selection procedure can be terminated at step M_0 (generally $M_0 << M$), the optimal number of model terms, at which point the function $R^2_{\text{apress}}[m]$ with respect to m will be a maximum that satisfies $\chi_1(M_0) \geq \chi_2(M_0)$. The system output can be expressed as a linear combination of the M_0 selected significant regressrs

$$\mathbf{y} = g_1 q_1 + \cdots + g_{M_0} q_{M_0} + \varepsilon \tag{29}$$

which is equivalent to

$$y(t) = \sum_{i=1}^{M_0} \theta_{\ell_i} \phi_{\ell_i}(t) + \varepsilon(t) \tag{30}$$

where the parameters $\theta^{(AOLS)} = [\theta_{\ell_1}, \theta_{\ell_2}, \cdots, \theta_{\ell_{M_0}}]^T$ are calculated from the triangular equation $Ag = \theta^{(AOLS)}$ with $g = [g_1, g_2, \cdots, g_{M_0}]^T$ and

$$A = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1M_0} \\ 0 & 1 & \cdots & a_{2M_0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & a_{M_0-1,M_0} \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

The entries $a_{ij}\ (1 \leq i < j \leq M_0)$ are calculated during the orthogonalization procedure.

5. The performance of the new AOLS algorithm—simulation studies

Four examples are provided to illustrate the performance of the new AOLS algorithm. In all the four examples, the adjustable coefficient λ in Eq.(21) was set to $\lambda = \max\{1, 0.0025N\}$. It was assumed that the true model structure was completely unknown once the models had been simulated and the input-output data were obtained. In each example, noise models were estimated to ensure unbiased model parameters but the noise model estimates are not shown to save space (see Section 6.2).

Example 1—a high order linear model

Consider a high order linear system described by the model

$$x(t) = 0.0627x(t-1) + 0.3068x(t-2) - 0.0539x(t-3) + 1.0015u(t-9) + 0.6332u(t-10) + \xi(t) \tag{31a}$$

$$y(t) = x(t) + \eta(t) \tag{31b}$$

where

$$\xi(t) = \varepsilon(t) + 0.3\varepsilon(t-1) - 0.6\varepsilon(t-2) \tag{32}$$

$$\eta(t) = \varepsilon(t) - 0.4\varepsilon(t-1) + 0.8\varepsilon(t-2) \tag{33}$$

and $\varepsilon(t)$ was Gaussian white noise with zero mean and standard deviation σ_ε=0.075. By setting the input $u(t)$ as a random sequence that was uniformly distributed in [-2,2], the model (31) was simulated and 600 input-output data points were collected. This data set was used for model identification with an assumption that the true model structure was unknown.

The initial starting model orders for the input and output were deliberately set to n_y=10 and n_u=20 in this example. The model term selection procedure started from a candidate polynomial model with a nonlinear degree ℓ =2, which contained 528 candidate model terms with the form $z_1^{i_1}(t)z_2^{i_2}(t)$, where $z_j^{i_j}(t) \in \{y(t-1),\cdots, y(t-10), u(t),\cdots,u(t-20)\}$ for j=1,2, $0 \le i_j \le 2$ and $0 \le i_1 + i_2 \le 2$. The values of the two indicating functions $\chi_1(\cdot)$ and $\chi_2(\cdot)$ are shown in figure 1, which clearly indicates that the optimal number of model terms is 5. The 5 selected model terms and associated parameters are listed in Table 1, which clearly shows that both the model structure and the parameters have been correctly estimated. All the nonlinear model terms have been correctly deleted by the algorithm.

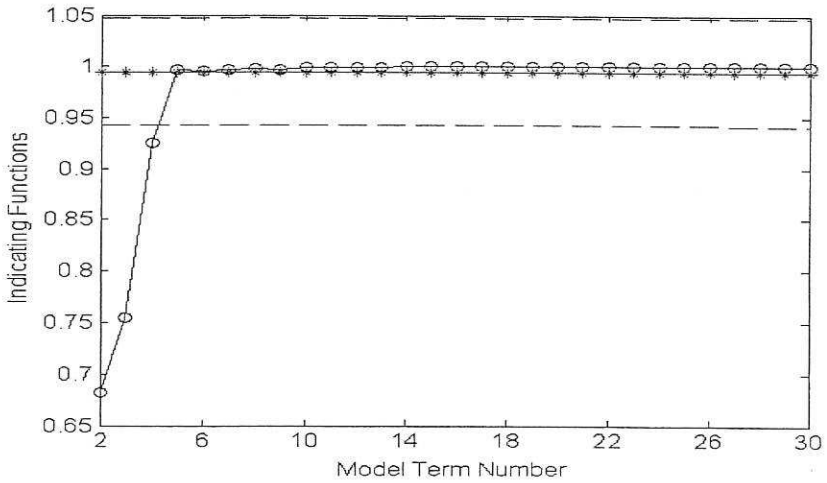

Figure 1. The indicating functions $\chi_1(\cdot)$ and $\chi_2(\cdot)$ defined by (24) and (25) versus the number of model terms for the linear system (31). The circled-line 'o-' indicates $\chi_1(\cdot)$, the stared-line '*-' indicates $\chi_2(\cdot)$, and the two dashed lines indicate the 90% confidence interval of $\chi_2(\cdot)$.

TABLE 1
THE SELECTED MODEL TERMS, ESTIMATED PARAMETERS AND ASSOCIATED ERR VALUES FOR THE MODEL (31)

No.	Terms $\phi_k(t)$	Parameters θ_k	$ERR_k \times 100\%$
1	$y(t\text{-}1)$	6.09561567e-002	9.31394063e+001
2	$u(t\text{-}9)$	1.00223919e+000	6.57627708e+000
3	$u(t\text{-}10)$	6.19325151e-001	8.30910081e-002
4	$y(t\text{-}2)$	3.04174816e-001	3.85142613e-002
5	$y(t\text{-}3)$	-4.28406321e-002	9.97412512e-003

Example 2—a second order, second degree nonlinear model

Consider a second order nonilinear system described by the model

$$x(t) = -0.605x(t-1) - 0.163x^2(t-2) + 0.588u(t-1) - 0.240u(t-2) + \xi(t) \tag{34a}$$

$$y(t) = x(t) + \eta(t) \tag{34b}$$

where

$$\xi(t) = \varepsilon(t) + 0.2\varepsilon(t-1) - 0.5\varepsilon(t-2) \tag{35}$$

$$\eta(t) = \varepsilon(t) - 0.3\varepsilon(t-1) + 0.6\varepsilon(t-2) \tag{36}$$

and $\varepsilon(t)$ was Gaussian white noise with zero mean and standard deviation σ_ε=0.04. By setting the input $u(t)$ as a random sequence that was uniformly distributed in [-1,1], the model (34) was simulated and 500 input-output data points were collected. This data set was used for model identification with an assumption that the true model structure was unknown.

Although the model order and significant variables can be correctly selected using a variable selection algorithm (Wei *et al.* 2004), the initial model orders for the input and output was deliberately set to n_y=5 and n_u=5 in this example. The model term selection procedure started from a candidate polynomial model with a nonlinear degree ℓ=3, which contained 364 candidate model terms with the form $z_1^{i_1}(t)z_2^{i_2}(t)z_3^{i_3}(t)$, where $z_j^{i_j}(t) \in \{y(t-1),\cdots, y(t-5), u(t),\cdots,u(t-5)\}$ for j=1,2,3, $0 \le i_j \le 3$ and $0 \le i_1 + i_2 + i_3 \le 3$. The values of the two indicating functions $\chi_1(\cdot)$ and $\chi_2(\cdot)$ are shown in figure 2, which clearly indicates that the optimal number of model terms is 4. The 4 selected model terms and associated parameters are listed in Table 2, which clearly shows that both the model structure and the parameters have been correctly estimated.

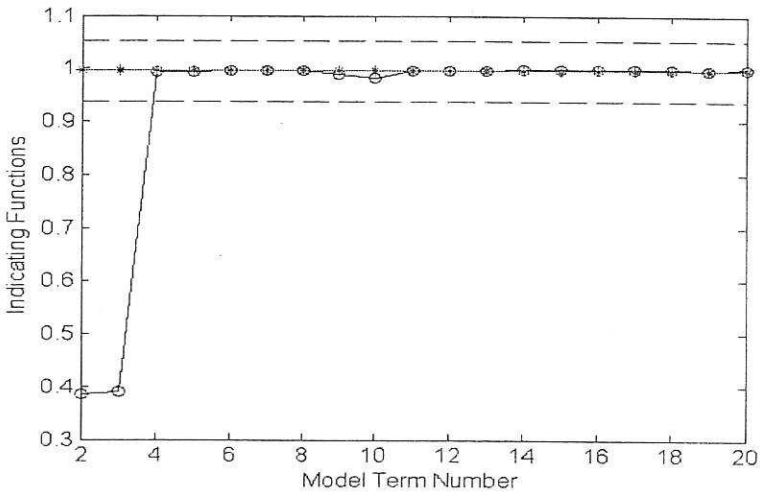

Figure 2. The indicating functions $\chi_1(\cdot)$ and $\chi_2(\cdot)$ defined by (24) and (25) versus the number of model terms for the nonlinear system (34). The circled-line 'o-' indicates $\chi_1(\cdot)$, the stared-line '*-' indicates $\chi_2(\cdot)$, and the two dashed lines indicate the 90% confidence interval of $\chi_2(\cdot)$.

TABLE 2
THE SELECTED MODEL TERMS, ESTIMATED PARAMETERS AND ASSOCIATED ERR VALUES FOR THE MODEL (34)

No.	Terms $\phi_k(t)$	Parameters θ_k	$ERR_k \times 100\%$
1	$y(t\text{-}1)$	-5.97683257e-001	6.50487433e+001
2	$u(t\text{-}1)$	5.91092947e-001	2.91410631e+001
3	$y^2(t\text{-}2)$	-1.64778127e-001	2.70441508e+000
4	$u(t\text{-}2)$	-2.40698404e-001	2.44104449e+000

TABLE 3
THE SELECTED MODEL TERMS, ESTIMATED PARAMETERS AND ASSOCIATED ERR VALUES FOR THE MODEL (37)

No.	Terms $\phi_k(t)$	Parameters θ_k	$ERR_k \times 100\%$
1	$u(t\text{-}1)$	1.00251598e+000	8.30902332e+001
2	$y(t\text{-}1)$	9.86141811e-002	1.44668436e+001
3	$y\,(t\text{-}1)\,y\,(t\text{-}2)\,u\,(t\text{-}2)$	9.59277528e-004	4.49680879e-001
4	$u(t\text{-}2)$	-5.05326462e-001	1.12701859e+000
5	$y(t\text{-}2)$	1.63353923e-001	1.58338992e-001
6	$y^3(t\text{-}2)$	-1.48927094e-003	2.24371699e-001
7	$y^2(t\text{-}2)\,u\,(t\text{-}2)$	1.41548044e-003	2.08489012e-002
8	$y^2(t\text{-}1)\,u\,(t\text{-}2)$	2.48017967e-003	1.92876073e-002
9	$y^3(t\text{-}1)$	5.23491349e-004	1.99220679e-002

Example 3—a rational model

Consider a system described by the nonlinear rational model

$$x(t)=\frac{0.5x^2(t-1)x(t-2)+0.2x(t-1)x^2(t-2)}{1+x^2(t-1)+x^2(t-2)}+u(t-1)-0.4u(t-2)\ +\xi(t) \tag{37a}$$

$$y(t)=x(t)+\eta(t) \tag{37b}$$

where

$$\xi(t)=\varepsilon(t)+0.4\varepsilon(t-1)+0.8\varepsilon(t-2) \tag{38}$$

$$\eta(t)=\varepsilon(t)+0.3\varepsilon(t-1)+0.6\varepsilon(t-2) \tag{39}$$

and $\varepsilon(t)$ was Gaussian white noise with zero mean and standard deviation σ_ε=0.2. By setting the input $u(t)$ as a random sequence that was uniformly distributed in [-10,10], the model (37) was simulated and 400 input-output data points were collected. This data set was used for model identification with an assumption that the true model structure was unknown.

A variable selection algorithm (Wei *et al.* 2004) was applied to the simulation data set and 4 significant variables were selected as $\{y(t\text{-}1), y(t\text{-}2), u(t\text{-}1\}, u(t\text{-}2)\}$. The 4 significant variables were used to form an initial candidate polynomial model with a nonlinear degree ℓ =4, which contains 70 candidate model terms. The values of the two indicating functions $\chi_1(\cdot)$ and $\chi_2(\cdot)$ are shown in figure 3, which clearly indicates that the optimal

number of model terms is 8 or 9. Although both models with 8 and 9 terms meet the model validity tests based on the residual correlation analysis (Billings and Voon 1986), it follows that the model with 9 terms possesses better long term predictive capability according to the adjustable R^2-statistic given by (13). The identified model of 9 terms is listed in Table 3.

To inspect the performance of the identified models listed in Table 3, two input signals, $u(t) = w(t)$, with $w(t)$ a Gaussian white noise sequence of zero mean and standard deviation σ =2.5 and $u(t) = 2.5\sin(20 \times 2\pi t / 400) + 5\sin(30 \times 2\pi t / 500)$ were separately used to drive the model (37), where the noise $\xi(t)$ and $\eta(t)$ were set to be zero. The model predicted outputs, $\hat{y}_{mpo}(t) = \hat{f}(\hat{y}_{mpo}(t-1), \hat{y}_{mpo}(t-2), u(t-1), u(t-2))$, were compared with the noise free output from the original model (37) over the range $3 \le t \le 1000$. The data points corresponding to the two input signals over the range from 500 to 600 and from 400 to 600 are shown in figures 4 and 5, respectively. It is clear from figures 5 and 6 that the identified model provides excellent approximation for the original model (37) even when driven by input that are totally different to the input used in the identification.

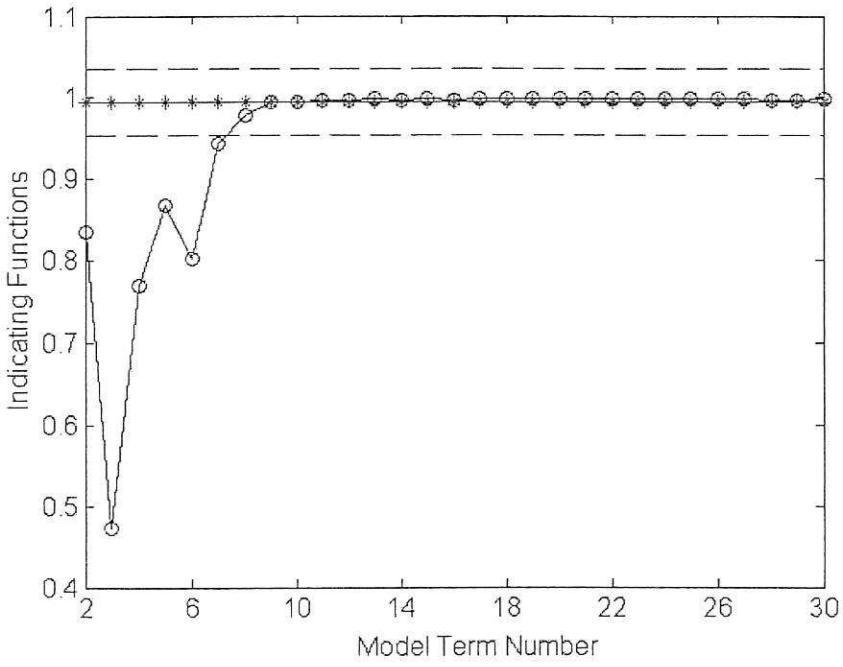

Figure 3. The indicating functions $\chi_1(\cdot)$ and $\chi_2(\cdot)$ defined by (24) and (25) versus the number of model terms for the nonlinear system (37). The circled-line 'o-' indicates $\chi_1(\cdot)$, the stared-line '*-' indicates $\chi_2(\cdot)$, and the two dashed lines indicate the 90% confidence interval of $\chi_2(\cdot)$.

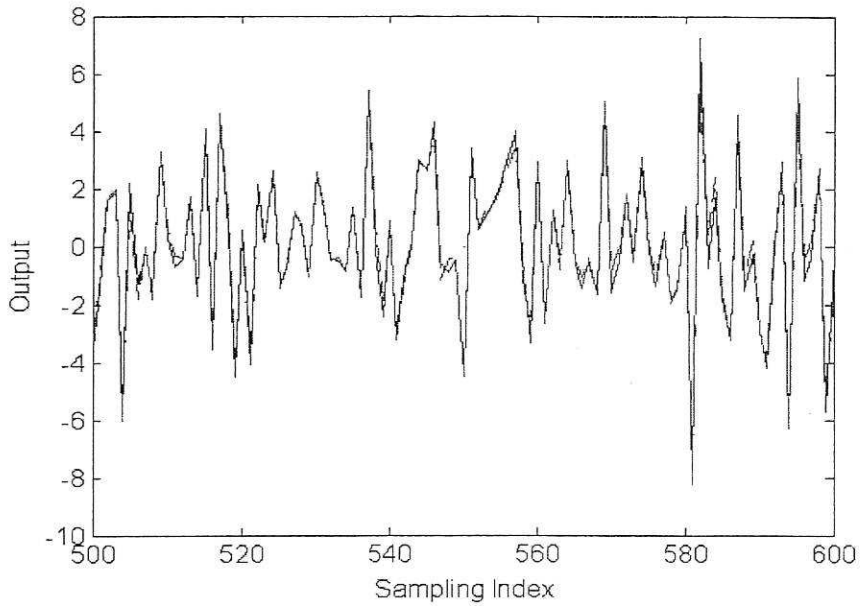

Figure 4. A comparison between the noise free measurements from the original model (37) and the output from the identified model driven by a white noise sequence. The solid line indicates the noise free measurements from the original model (37) , the dashed line indicates the model predicted output from the identified model.

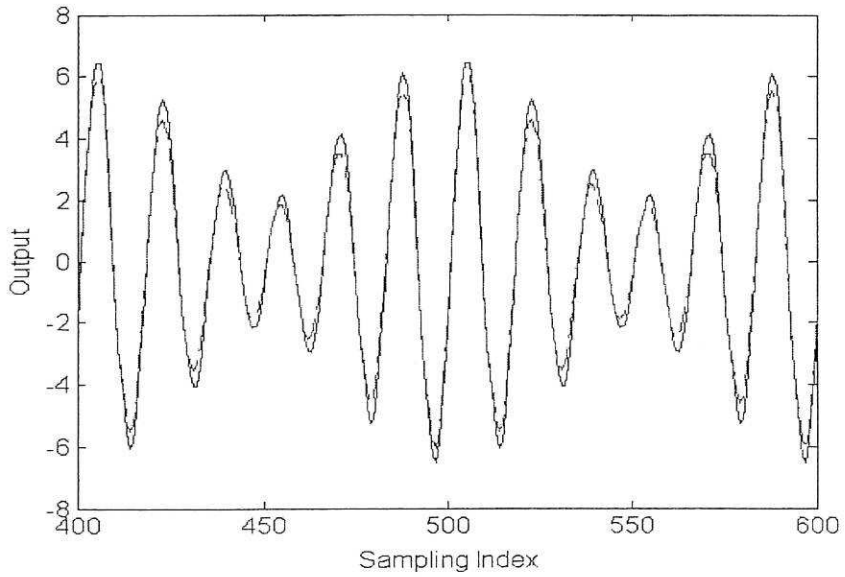

Figure 5. A comparison between the noise free measurements from the original model (37) and the output from the identified model driven by a sine wave. The solid line indicates the noise free measurements from the original model (37) , the dashed line indicates the model predicted output from the identified model.

Example 4—a MIMO nonlinear model

Consider a system described by the nonlinear rational model

$$x_1(t) = -0.5x_1(t-1) + \frac{x_1(t-2)x_2(t-1)}{1+[x_1^2(t-1)+x_2^2(t-1)]^{1/2}} + u_1(t-1) \tag{40a}$$

$$x_2(t) = -0.25x_2(t-1) + \frac{0.5x_1(t-1)x_2(t-2)}{1+[x_1^2(t-1)+x_2^2(t-1)]^{1/2}} + 0.2u_2(t-1) \tag{40b}$$

$$y_1(t) = x_1(t) + w_1(t) \tag{40c}$$

$$y_2(t) = x_2(t) + w_2(t) \tag{40d}$$

where $w_1(t)$ and $w_2(t)$ were Gaussian white noise sequences with zero mean and standard deviation σ_1 =0.1 and $\sigma_2 = 0.01$, respectively. By setting the two input signals $u_1(t)$ and $u_2(t)$ as random sequences that were uniformly distributed in [-1,1], the model (40) was simulated and 1000 input-output data points were collected. This data set was used for model identification with an assumption that the true model structure was unknown.

A variable selection algorithm (Wei *et al.* 2004) was applied to the simulation data set and 6 significant variables were selected as $\{y_1(t\text{-}1), y_1(t\text{-}2), y_2(t\text{-}1), y_2(t\text{-}2), u_1(t\text{-}1), u_2(t\text{-}1)\}$. The 6 significant variables were used to form an initial candidate polynomial model with a nonlinear degree ℓ =3, which contained 84 candidate model terms. The values of the two indicating functions $\chi_1(\cdot)$ and $\chi_2(\cdot)$ for the two subsystems are shown in figure 6, which clearly indicates that the optimal number of model terms for both the two submodels is 4. The identified model terms and associated parameters are listed in Table 4.

To inspect the performance of the identified models, two input signals, $u_1(t) = \sin(\pi t/20)$, $u_2(t) = \cos(\pi t/30)$ were used to drive both the identified model and the original noise free model (40). The model predicted outputs were compared with the noise free output from the original model (40) over the range $3 \le t \le 1000$, and these are shown in figures 7 and 8. It is clear from figure 8 that the identified model provides excellent approximation for the original model (40).

TABLE 4
THE SELECTED MODEL TERMS, ESTIMATED PARAMETERS AND ASSOCIATED ERR VALUES FOR THE MODEL (40)

No.	Submodel I			Submodel II		
	Terms $\phi_k(t)$	Parameters θ_k	$ERR_k \times 100\%$	Terms $\phi_k(t)$	Parameters θ_k	$ERR_k \times 100\%$
1	$u_1(t\text{-}1)$	1.00021201e+000	7.43228544e+001	$u_2(t\text{-}1)$	2.00076326e-001	9.08018786e+001
2	$y_1(t\text{-}1)$	-5.00336484e-001	2.50339132e+001	$y_2(t\text{-}1)$	-2.49135624e-001	5.99431358e+000
3	$y_1(t\text{-}2)\ y_2(t\text{-}1)$	6.74439026e-001	6.07735064e-001	$y_1(t\text{-}1)\ y_2(t\text{-}2)$	2.65222930e-001	3.10508291e+000
4	$y_1(t\text{-}1)\ y_2(t\text{-}1)$	1.05823417e-001	1.29427954e-002	$y_1(t\text{-}2)\ y_2(t\text{-}2)$	1.30978626e-002	6.03751117e-003

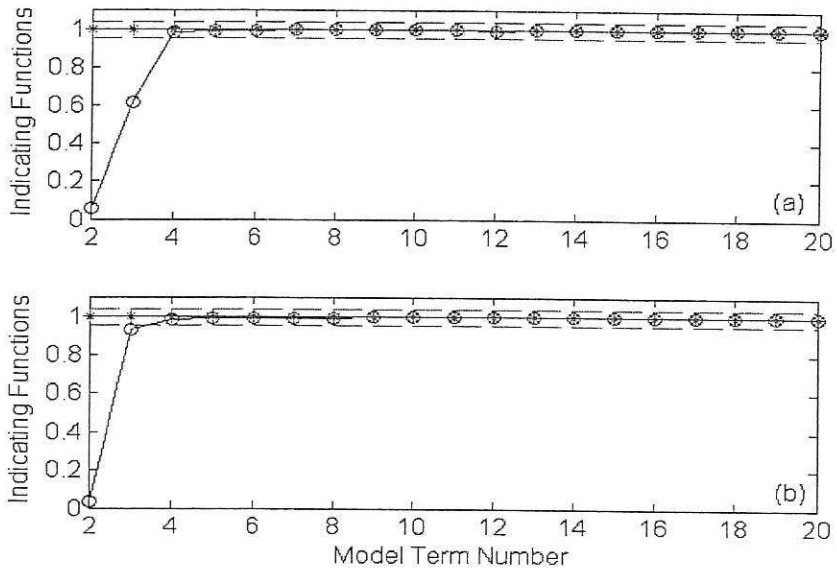

Figure 6. The indicating functions $\chi_1(\cdot)$ and $\chi_2(\cdot)$ defined by (24) and (25) versus the number of model terms for the nonlinear system (40). (a) for submodel I, (b) for submodel II. The circled-line 'o-' indicates $\chi_1(\cdot)$, the stared-line '*-' indicates $\chi_2(\cdot)$, and the two dashed lines indicate the 90% confidence interval of $\chi_2(\cdot)$.

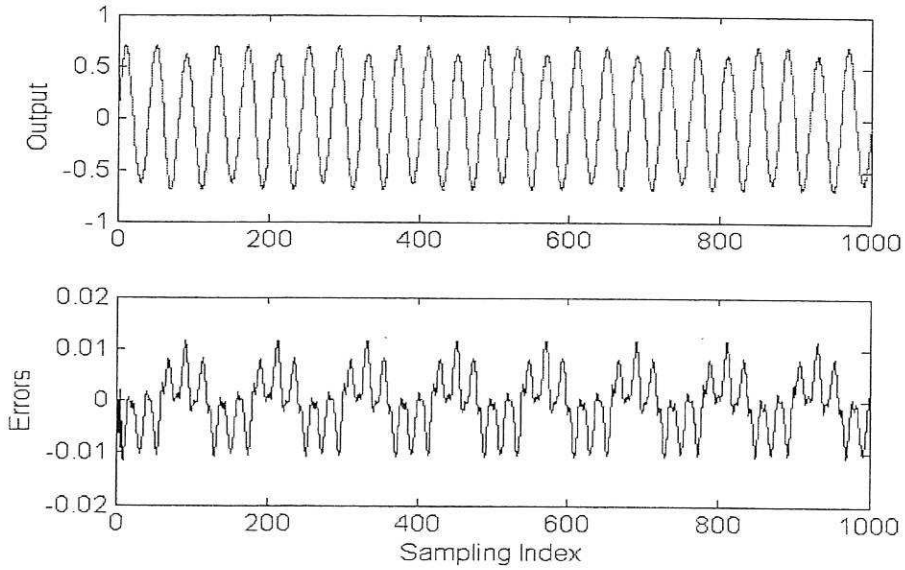

Figure 7. An overlap of the output from the identified submodel I and the corresponding noise free output from the original model (40) (the top figure), and the errors (the bottom figure), with the given sinusoidal inputs.

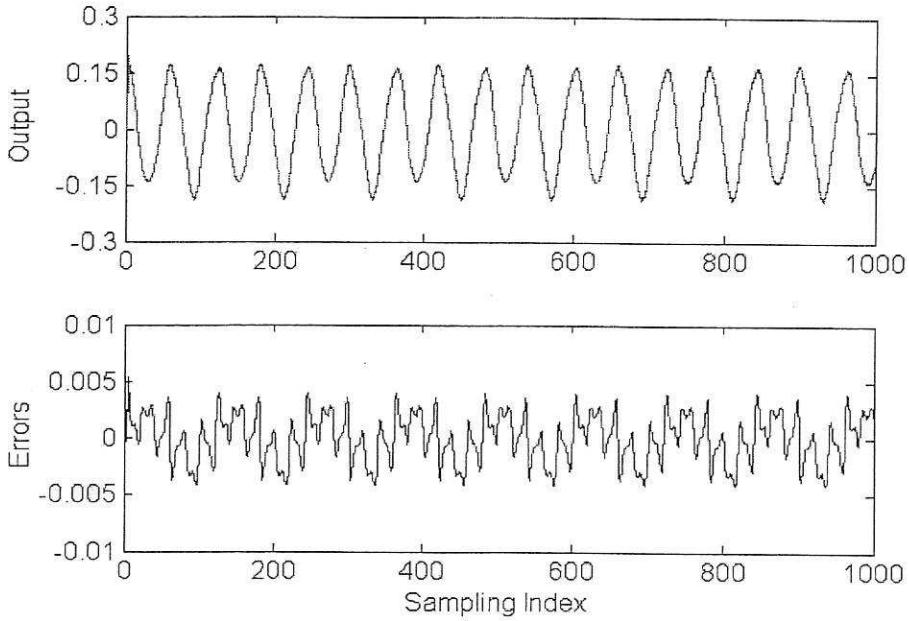

Figure 8. An overlap of the output from the identified submodel II and the corresponding noise free output from the original model (40) (the top figure), and the errors (the bottom figure), with the given sinusoidal inputs.

6. Modifications to classical information criteria and noise modelling

This section will demonstrate that other criteria may also be used to determine the number of model terms in a way similar to the use of the R^2-APRESS statistic based AOLS algorithm. This section will also give some information on correlated noise modelling when a NARX model cannot provide sufficient description for a system or a given data set.

6.1 Modified classical information criteria

It can be seen from the results given in the previous section that the AOLS scheme, which combines the efficient orthogonal least squares algorithm and a R^2-APRESS statistic, provides an effective tool to find not only the correct model terms but also the correct number of model terms. There are some similarities between the R^2-APRESS statistic and the classical information criteria AIC, BIC and their derivatives. For example, using the fact that

$$\frac{1}{(1-x)^2} \approx 1+2x+3x^2+4x^3+\cdots, \ |x|<1, \tag{41}$$

The PRESS statistic given by (18) can be approximated when $p << N$ by

$$\text{PRESS}[p] \approx \left(1+\frac{2\lambda p}{N}\right)\text{NMSE}[p] = \left(\frac{\text{SST}_0}{\text{SST}}\right)\left(1+\frac{2\lambda p}{N}\right)\text{ESR}[p] \tag{42}$$

This is a modified version of the generalized cross-validation (GCV). Other traditional information criteria including the AIC, BIC, FPE and similar versions can also be incorporated in the AOLS algorithm after some necessary modifications. One way to modify and improve these information criteria is to relate them to ESR, and at the same time to drop some insignificant constant terms and then to introduce some necessary penalty terms into the initial formulation of these criteria. The modified information criteria can then be incorporated in the AOLS algorithm. Consider the AIC and BIC as an example. The classical AIC and BIC criteria can be modified as follows:

$$\mathrm{AIC}(p) = \log(\mathrm{ESR}(p)) + \frac{2(\lambda p + 1)}{N} \tag{43}$$

$$\mathrm{BIC}(p) = \log(\mathrm{ESR}(p)) + (\lambda p + 1)\frac{\log N}{N} \tag{44}$$

where $\lambda = \max\{1,\ \rho N\}$ with $0.002 \le \rho \le 0.01$. Similar criteria were used for model selection by Harvich and Tsai (1995). Note that the BIC given by (44) is similar to the Schwarz-Rissanen information criterion, SIC (Schwarz 1978, Rissanen 1978). Following the discussion in Section 4, the indicating functions for the modified AIC and BIC in (43) and (44) can be chosen as

$$\chi_{1,\mathrm{AIC}}(p) = \chi_{1,\mathrm{BIC}}(p) = \frac{\mathrm{ESR}(p+1)}{\mathrm{ESR}(p)} \tag{45}$$

$$\chi_{2,\mathrm{AIC}}(p) = e^{-2\lambda/N} \approx 1 - \frac{2\lambda}{N} \tag{46}$$

$$\chi_{2,\mathrm{BIC}}(p) = e^{-(\lambda/N)\log N} \tag{47}$$

Example 5—a nonlinear time series

Consider the following nonlinear time series

$$\begin{aligned} y(t) = {} & (0.8 - 0.5e^{-y^2(t-1)})y(t-1) - (0.3 + 0.9e^{-y^2(t-1)})y(t-2) \\ & + 0.1\sin(\pi y(t-1)) \end{aligned} \tag{48}$$

This model has been used as a benchmark example by several authors (Billings and Chen 1998). In this example, this model was simulated starting with the initial condition $y(-1)=0.25$, $y(0)=0$ and 5000 data points were sampled. The first 400 data appoints were used for model estimation and the remaining points for model testing. Significant variables were chosen as $\{y(t\text{-}1), y(t\text{-}2)\}$, and these were used to construct a polynomial model for this time series, starting from an initial candidate polynomial model with a nonlinear degree $\ell = 6$, which contained 28 candidate model terms. The values of the indicating functions $\chi_{1,\mathrm{AIC}}(\cdot)$, $\chi_{1,\mathrm{BIC}}(\cdot)$, $\chi_{2,\mathrm{AIC}}(\cdot)$ and $\chi_{2,\mathrm{BIC}}(\cdot)$ defined by (45), (46) and (47) were calculated and these are shown in figure 9, which clearly indicates that the optimal number of model terms is 9. The selected model terms are listed in Table 5. The results identified using the modified AIC and BIC based AOLS algorithm are identical to those produced by the R^2-APRESS statistic based AOLS algorithm for the noise free data generated from the time series (48).

To inspect the performance of the identified model, the first return map reconstructed from the data points (model predicted outputs) generated by the identified model was compared with the map produced by the noise

free data generated from the original model (48). These are shown in figure 10, which clearly shows that the identified model provides an excellent approximation for the original time series (48). In fact, if the two maps in figure 10(a) and (b) were put together, they would overlap with each other and are indistinguishable.

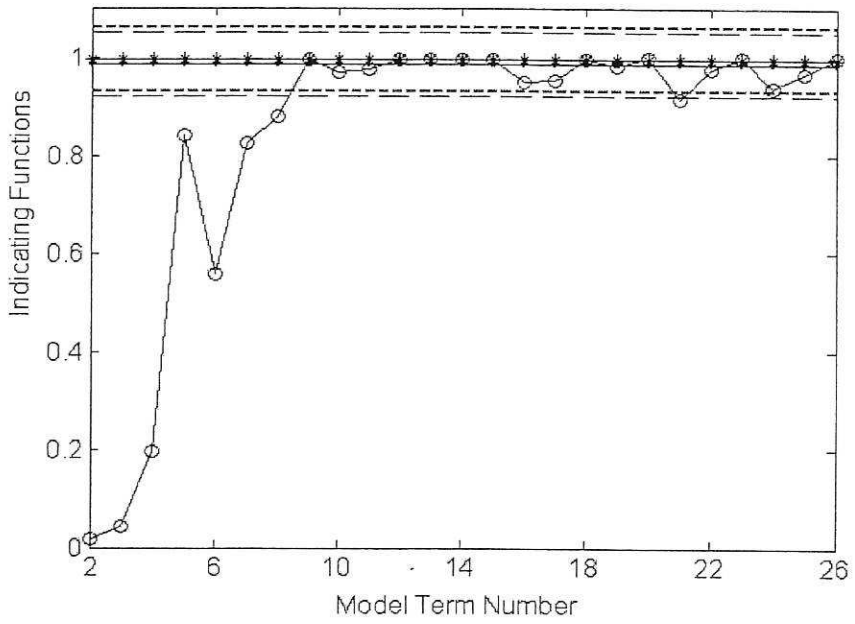

Figure 9. The $\chi_{1,\mathrm{AIC}}(\cdot)$, $\chi_{1,\mathrm{BIC}}(\cdot)$, $\chi_{2,\mathrm{AIC}}(\cdot)$ and $\chi_{2,\mathrm{BIC}}(\cdot)$ defined by (45), (46) and (47) versus the number of model terms for the time series (48). The circled-line 'o-' indicates $\chi_{1,\mathrm{AIC}}(\cdot)$ and $\chi_{1,\mathrm{BIC}}(\cdot)$, the stared-line '*-' indicates $\chi_{2,\mathrm{AIC}}(\cdot)$, the dotted-line indicates $\chi_{2,\mathrm{BIC}}(\cdot)$, the two dashed lines indicate the 90% confidence interval of $\chi_{2,\mathrm{AIC}}(\cdot)$, and the two short dashed-line in bald indicate the 90% confidence interval of $\chi_{2,\mathrm{BIC}}(\cdot)$.

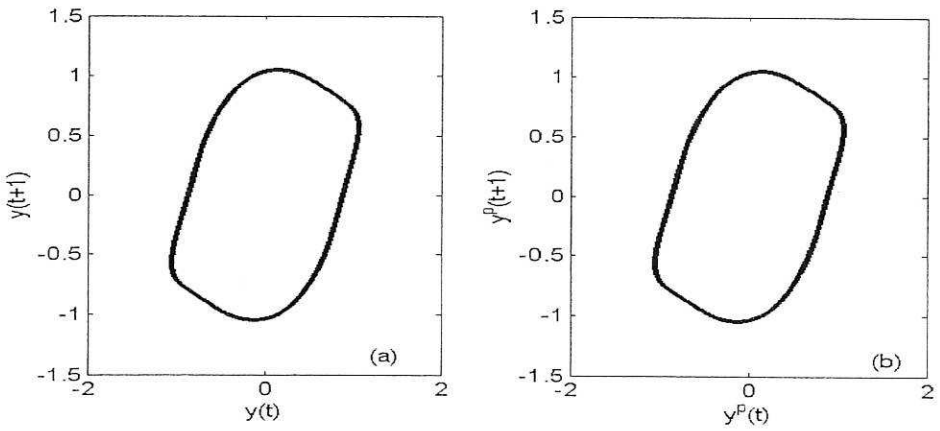

Figure 10. The first return map for the time series described by (48). (a) constructed from 4600 points generated by the original model (48); (b) reconstructed from 4600 data points generated by the identified model.

TABLE 5
THE SELECTED MODEL TERMS, ESTIMATED PARAMETERS AND ASSOCIATED ERR VALUES FOR THE MODEL (48)

No.	Terms $\phi_k(t)$	Parameters θ_k	$ERR_k \times 100\%$
1	$y(t\text{-}2)$	-1.19591378e+000	5.70039032e+001
2	$y(t\text{-}1)$	6.08163570e-001	4.06119954e+001
3	$y^2(t\text{-}1)\,y(t\text{-}2)$	8.69722510e-001	2.34052691e+000
4	$y^4(t\text{-}1)\,y(t\text{-}2)$	-3.02237294e-001	4.16885795e-002
5	$y^4(t\text{-}1)\,y^2(t\text{-}2)$	-2.41415155e-003	1.51410216e-003
6	$y^5(t\text{-}2)$	1.99825511e-002	1.38005940e-004
7	$y^5(t\text{-}1)$	9.28908169e-003	5.86269331e-005
8	$y^3(t\text{-}2)$	-1.87355282e-002	3.07525656e-005
9	$y^2(t\text{-}1)\,y^3(t\text{-}2)$	-2.08950788e-002	1.75082048e-005

6.2 Noise modelling

In many cases the noise signal $e(t)$ in Eq. (1) may be a correlated or coloured noise sequence. This is likely to be the case for most real data sets. The NARX model (1) may fail to give a sufficient description due to the bias in the parameter estimates. In this case, the effects of the noise must be taken into account and some model terms relating to the noise should be included in the model. This is achieved by extending the regressor vector defined in (2) to $\varphi^{\{n\}}(t) = [y(t-1),\cdots, y(t-n_y), u(t),\cdots, u(t-n_u), e(t-1),\cdots, e(t-n_e)]^T$. This results in the celebrated NARMAX (*Nonlinear AutoRegressive Moving Average* with *eXogenous* inputs) model (Leontaritis and Billings 1985)

$$y(t) = f(y(t-1),\cdots, y(t-n_y), u(t),\cdots, u(t-n_u), e(t-1),\cdots, e(t-n_e)) + e(t) \tag{49}$$

Model (49) is obviously more general than the NARX model (1) and includes as special cases several linear and nonlinear representations (Pearson 1999). Note that the noise signal $e(t)$ in model (49) is generally unobserved and is often replaced by the model residual sequence. Let $\hat{f}(\cdot)$ represent an estimator for the model $f(\cdot)$, the residuals $\varepsilon(t)$ can then be estimated as

$$\begin{aligned}\varepsilon(t) &= y(t) - \hat{y}(t) \\ &= y(t) - \hat{f}(y(t-1),\cdots, y(t-n_y), u(t),\cdots, u(t-n_u), \varepsilon(t-1),\cdots, \varepsilon(t-n_e))\end{aligned} \tag{50}$$

In practice, a NARMAX model can be implemented using the new AOLS algorithm as follows:

- Identify a NARX model. Assume M_0 model terms are included in the NARX model.
- Add noise-related candidate model terms into the NARX model. Assume N_0 noise-related terms are added.
- Calculate the prediction errors $\varepsilon(t)$ recursively as

$$\begin{aligned}\varepsilon^{(k)}(t) &= y(t) - \hat{y}(t) \\ &= y(t) - \hat{f}(y(t-1),\cdots, y(t-n_y), u(t),\cdots, u(t-n_u), \varepsilon^{(k-1)}(t-1),\cdots, \varepsilon^{(k-1)}(t-n_e))\end{aligned}$$

$$= y(t) - \sum_{i=1}^{M_k} \theta_i^{(k)} \pi_i(\varphi(t)) - \sum_{i=1}^{N_k} \theta_i^{(k)\{n\}} \pi_i^{\{n\}}(\varphi^{\{n\}}(t)) \tag{51}$$

where $\varepsilon^{(0)}(t-j)=0$ for $j=1,2,\cdots,n_e$, $\pi_i(\varphi(t))$ for $i=1,2,\cdots,M_k$ are the selected model terms during the kth step, $\pi_i^{\{n\}}(\varphi^{\{n\}}(t))$ $i=1,2,\cdots,N_k$ are the selected noise-related model terms at the kth step. Typically, this iteration can be terminated in a few steps.

- Form a NARMX model by including all the selected terms associated with the process and noise.

Practical identification experience shows that the bias on the parameter estimates can be virtually eliminated by including the noise signals $e(t-1),\cdots,e(t-n_e)$ in the model. Readers are referred to Billings *et al.* (1989), Chen *et al.* (1989), and Billings and Chen (1998) for details about the NARMAX modelling methodology.

7. Conclusions

An efficient fast adaptive orthogonal least squares (AOLS) algorithm has been developed for subset selection and nonlinear system identification. In the new AOLS algorithm, a new indicator, the error-to-signal ratio (ESR), and a new R^2-like statistic, the adjustable prediction error sum of squares (R^2-APRESS), have been introduced. The new AOLS algorithm was developed from a combination of ESR, R^2-APRESS and an efficient forward orthogonal least squares algorithm. Combining ESR with R^2-APRESS, and especially by unifying these into a fast orthogonal least squares algorithm, has produced the AOLS algorithm, which is multifunctional and can be used for model term selection, optimal term number determination, and parameter estimation. Modified versions of several classical information criteria have also been given. The new AOLS scheme, accompanied with our previous results on model term and variable selection methods (Wei *et al.* 2004), provides an efficient tool which can be really applied to a wide class of nonlinear system identification problems. The only limitation of the new ALOS algorithm is that, it requires that the dynamical behaviour of the system under study be approximated using a nonlinear model of a specified form, which can be expressed using a linear-in-the-parameters representation with respect to given basis functions including polynomials, radial basis functions, splines, and wavelets. Several illustrative examples have been described to show the effectiveness of the new AOLS algorithm for nonlinear system identification.

Acknowledgment

The authors gratefully acknowledge that this work was supported by EPSRC (UK).

References

H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Statist. Math.*, 1969, 21, pp. 243-247.

H. Akaike, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, 1970, 22, pp. 203-217.

H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, 1974, AC-19(6), pp. 716-723.

D. M. Allen, "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, 1971, 13, pp. 469-475.

D. M. Allen, "The relationship between variable selection and data augmentation and a method for prediction," *Technometrics*, 1974, 16, pp. 125-127.

S. A. Billings and S. Chen, "The determination of multivariable nonlinear models for dynamic systems using neural networks," in *Neural Network Systems Techniques and Applications*, C.T. Leondes, Ed. San Diego: Academic Press, 1998, pp. 231-278.

S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems suing a forward regression orthogonal estimator," *Int. J. Control*, 1989, 49(6), pp. 2157-2189.

S. A. Billings and W. S. F. Voon, "Correlation based model validity tests for nonlinear models," *Int. J. Control*, 1986, 44(1), pp. 235-244.

S. Chatterjee and A. Hadi, *Sensitivity Analysis in Linear Regression*. New York: John Wiley & Sons, 1988.

S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, 1989, 50(5), pp. 1873-1896.

S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least-squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Networks*, 1991, 2(2), pp. 302-309.

C. M. Harvich and C.-L. Tsai, "Model selection for extended quasi-likelihood models in small samples," *Biometrics*, 1995, 51(3), pp. 1077-1084.

H. M. Henrique, E. L. Lima, and D. E. Seborg, "Model structure determination in neural network models," *Chem. Eng. Sci.*, 2000, 55(22), pp. 5457-5469.

X. Hong and C. J. Harris, "Nonlinear model structure detection using optimum experimental design and othogonal least squares," *IEEE Trans. Neural Networks*, 2001, 12(2), pp. 435-439.

M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," *Int. J. Control*, 1988, 48(1), pp. 93-210.

I. J. Leontaritis. and S. A. Billings, "Input-output parametric models for non-linear systems (part I: deterministic non-linear systems; part II: stochastic non-linear systems)," *Int. J. Control*, 1985, 41(2), pp. 303-344.

K. C. Li, "Asymptotic optimality of c_L and generalized cross-validation in ridge regression and application to spline smoothing," *Ann. Statist.*, 1986, 14, pp. 1101-1112.

K. C. Li, "Asymptotic optimality of c_L and generalized cross-validation: discrete index set," *Ann. Statist.*, 1987, 15, pp. 958-975.

A. J. Miller, *Subset Selection in Regression*. London: Chapman and Hall, 1990.

R. Myers, *Classical and Modern Regression with Applications* (2^{nd} Ed.). Boston: PWS-KENT Publishing Company, 1990

R. K. Pearson, *Descrete-Time Dynamic Models*. Oxford: Oxford University Press, 1999.

J. Rissanene, "Modelling by shortest data description," *Automatica*, 1978, 14(5), pp. 465-471.

J. Rissanene, "A universal prior for integers and estimation by minimum description length," *Ann. Stat.*, 1983, 11(2), pp. 416-431.

G. Schwartz, " Estimating the dimension of a model," *Ann. Stat.*, 1978, 6(2), pp. 461-464.

R. D. Snee, "Validation of regression models methods and examples," *Technometrics*, 1977, 19, pp. 415-428.

M. Stone, "Cross-validity choice and assessment of statistical predictor," *J. Roy. Statist. Soc.*, 1974, 36, pp. 111-147.

L. X. Wang and J. M. Mendel, "Fuzzy basis functions, universal approximations, and orthogonal least squares learning," *IEEE Trans. Neural Networks*, 3(5), pp. 807-814.

M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust. Speech Signal Processing*, 1985, ASSP-33(2), pp. 387-392.

H. L. Wei and S. A. Billings, "Identification and reconstruction of chaotic systems using multiresolution wavelet models," *Int. J. Syst. Sci.*, 2004, 35(9), pp. 511-526.

H. L. Wei, S. A. Billings, and J. Liu, "Term and variable selection for nonlinear system identification," *Int. J. Control*, 2004, 77(1), pp. 86-110.

A. Yong, M. J. Fuente, and W. Colmenares, "Learning discrete linear systems with the orthogonal learning algorithm," *Eng. Appl. Artif. Intel.*, 2001, 14(4), pp. 435-439.