Monograph:

Wei, H.L., Billings, S.A. and Liu, J. (2003) Term and Variable Selection for Nonlinear System Identification. Research Report. ACSE Research Report 837 . Department of Automatic Control and Systems Engineering

Term and Variable Selection for Nonlinear System Identification

H. L. Wei, S. A. Billings, J. Liu

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

Research Report No. 837

June 2003

Term and Variable Selection for Nonlinear System Identification

H.L. Wei , S.A. Billings and J. Liu
Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK

The purpose of variable selection is to pre-select a subset consisting of the significant variables or to eliminate the redundant variables from all the candidate variables of a system under study prior to model term detection. It is required that the selected significant variables alone should sufficiently represent the system. Generally, not all the model terms, which are produced by combining different variables, make an equal contribution to the system output and terms, which make little contribution, can be omitted. A parsimonious representation, which contains only the significant terms, can often be obtained without the loss of representational accuracy by eliminating the redundant terms. Based on these observations, a new variable and term selection algorithm is proposed in this paper. The term detection algorithm can be applied to the general class of nonlinear modelling problems which can be expressed as a linear-in-the-parameters form. The variable selection procedure is based on locally linear and cross-bilinear models, which are used together with the forward orthogonal least squares (OLS) and error reduction ratio (ERR) approach to determine the significant terms and to pre-select the important variables for both time series and input-output systems. Several numerical examples are provided to illustrate the applicability and effectiveness of the new approach.

1. Introduction

It is well known that system identification, which can be used construct a model to represent a given system, plays an important role in system analysis, control and prediction. The mathematical modelling of an engineering system generally consists of two aspects, model structure detection and parameter estimation. For a given system, assume that a dependent variable y (output) is affected by n candidate predictor variables (the inputs), say, x_1, x_2, $\cdots$, x_n. Assume that there exists a functional relationship between y and the input $x=[x_1,x_2,\cdots,x_n]^T$ such that $y=f_0(x)$. The objective of system identification is to construct a model to approximate the underlying relationship f_0 using a set of input and output observations.

Several types of model structures are available in nonlinear approximation, including parametric models such as polynomial models, rational models, neural networks, neurofuzzy networks etc., and nonparametric models such as radial basis function networks, wavelet networks and other basis function expansion methods. But the first problem encountered in modelling is how to determine which variables should be included in the model whatever kind of model form is used. It is often found in practice that some of the variables x_1, x_2, $\cdots$, x_n are redundant and only a subset of these variables are significant. Inclusion of redundant variables might result in a much more complex model since the number of model terms increases dramatically with the number of variables. Furthermore, including redundant variables might lead to a large number of free parameters in the model, and as a consequence the model may become oversensitive to training data and is likely to exhibit poor generalisation properties. Therefore, it is important to determine which variables should be included in the model.

For a linear regression model, the model terms and the variables are exactly the same, they are the regressors. However, variables and terms are generally distinct in a typical nonlinear model. The distinction between variables and terms is important and this can be illustrated using the simple nonlinear model below

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_1^2 + a_5x_2^2 + a_6x_3^2 + a_7x_1x_2 + a_8x_2x_3 + a_9x_1x_3 \tag{1}$$

Here there are only 3 variables: x_1, x_2 and x_3, but there are 10 terms, that is, *const*, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_2x_3 and x_1x_3. Ideally the selection of variables and the determination of terms should be separated. If the significant variables can be determined initially from a candidate variable set, then the number of model terms produced by combining these selected significant variables will be drastically decreased compared with that generated from the whole candidate variables, and the model structure detection procedure will therefore be simplified. This is even more important in dynamic nonlinear modelling where typically the terms are formed by the lagged inputs and outputs.

Variable selection and term determination are generic problems in nonlinear system identification. Once the significant variables have been selected and the candidate model structures have been determined, the model terms can be determined by performing some term searching procedures over the selected variables. A parsimonious model structure can then be detected from the candidate model set, and finally the parameters can be estimated based on the parsimonious model structure.

Several methods have been developed for selecting variables in linear regression analysis, including hypothesis tests, forward selection, backward elimination (Miller 1990), and principle component analysis (Oja 1992). Unfortunately, these linear methods cannot be easily extended to the nonlinear case. The main approaches developed for selecting variables for nonlinear systems include conditional probability analysis (Savit and Green 1991), Bayesian methods (George and McCulloch 1993), mutual information (Battiti 1994, Zheng and Billings 1996) and piecewise linearization (Mao and Billings 1999, Gomm and Yu 2000). As to the term detection methods, although some of the aforementioned variable selection approaches for linear models can be applied for term detection, these methods are time consuming and therefore some efficient approaches have been investigated, including the orthogonal least squares(OLS) algorithms (Billings et al. 1988, Korenberg et al. 1988, Billings et al.1989, Chen et al. 1989, Billings and Zhu 1994).

In the present paper, a new algorithm which pre-selects the variables and then detects the significant model terms is introduced for a wide class of nonlinear modelling problems. The method is based on the determination of the significant variables in locally linear and cross-bilinear models using a sum of error reduction ratios criteria. The significant nonlinear model terms are then determined by searching over terms formed from the pre-selected variable set. It is shown that the new procedures have significant advantages compared to existing methods because the new algorithms are robust with respect to noise, and are very efficient computationally and can be applied to relatively short data lengths.

The paper is organized as follows. Section 2 briefly describes the input-output model structures, and where linear-in-the-parameters regression is emphasized since many parametric and nonparametric models can be converted into this type of structure. In Section 3, a model term detection approach based on the forward OLS algorithm is described. In Section 4, a novel variable selection approach based on a sum of the error reduction ratios (SERR) and the final prediction errors (FPE) is introduced and described in detail. Several numerical examples are given in Section 5.

2. Model Structures

Assume that the system response y is determined by a set of predictor variables (inputs), say, $x_1, x_2, \cdots, x_n$. It is usually assumed that the response y and the input $x = [x_1, x_2, \cdots, x_n]^T$ are related by an unknown mapping $f(\cdot)$, such that

$$y(t) = f(x_1(t), x_2(t), \cdots, x_n(t)) + e(t) \tag{2}$$

where $e(t)$ is a noise signal which is often assumed to be an independent identically distributed random variable. Several types of model structures are available to approximate the unknown mapping $f(\cdot)$, including parametric, non-parametric and semi-parametric models. Parametric structures include polynomial models such as ARMAX and polynomial NARMAX, rational models. Non-parametric models include neural networks, fuzzy logic based models, wavelet expansions, radial basis function networks and other basis function expansion based models. Semi-parametric models are hybrid structures, in which one part is parametric and another is nonparametric. It has been proved that most nonlinear dynamic systems can be described using the NARMAX (N*onlinear* A*uto*R*egressive* M*oving* A*verage with* eX*ogenous* inputs) model (Leontaritis and Billings 1985)

$$y(t) = f(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u), e(t-1), \cdots, e(t-n_e)) + e(t) \tag{3}$$

where in the MIMO case $y(t) = [y_1(t) \ \ y_2(t) \ \ \cdots \ \ y_m(t)]^T$, $u(t) = [u_1(t) \ \ u_2(t) \ \ \cdots \ \ u_r(t)]^T$ and $e(t) = [e_1(t) \ \ e_2(t) \ \ \cdots \ \ e_m(t)]^T$ are the system output, input and noise sequences, respectively; $f(\cdot)$ is a nonlinear mapping vector; n_y, n_u and n_e are the maximum lags in the output, input and noise; the noise variable $e(t)$ is a zero mean independent sequence which accommodates the effects of measurement noise, modelling errors and unmeasured disturbances; $e(t)$ is sometimes called the prediction error which is defined as $e(t) = y(t) - \hat{y}(t)$, where $\hat{y}(t)$ is the one-step-ahead prediction.

The objective of system identification is to find a suitable model to approximate the underlying relationship $f(\cdot)$ using a set of input and output observations. Most of the model structures considered so far can be expressed using the form below (Sjoberg et al. 1995):

$$y(t) = f(x(t); \theta) = \sum_{i \in I} \gamma_i B_i(x(t); \alpha_i) \tag{4}$$

which can be parametrized using fixed basis functions such that

$$y(t) = f(x(t); \theta) = \sum_{i \in I} \gamma_i B_i(x(t)) \tag{5}$$

or, parametrized using the basis functions in terms of dilation and translation parameters α and β such that

$$y(t) = f(x(t); \theta) = \sum_{i \in I} \gamma_i B_i(\alpha_i x(t) + \beta_i) \tag{6}$$

where $x \in R^n$, $\theta = \{\gamma_i\}_{i\in I}$ or $\theta = \{\alpha_i, \beta_i, \gamma_i\}_{i\in I}$ are the parameters to be estimated, I is a subset of the integrals, B_i is a family of some basis functions belonging to the following three types: i) local basis functions with bounded support which vanish rapidly at points far from the centre, such as radial basis functions and wavelets; ii) semi-global basis functions, such as ridge basis functions (Friedman and Stuetzle 1981); iii) global basis functions such as polynomials and Fourier series.

In this paper, the linear-in-the-parameters model structure will be considered since many parametric and nonparametric models can be expressed in this form. The generic form of a linear-in-the-parameters model, taking the MISO case as an example, is

$$y(t) = \sum_{i=1}^{M} \theta_i p_i(x(t)) + \varepsilon(t),\ t = 1,2,\cdots,N \tag{7}$$

where N is the data length, $p_i(\cdot)$ are model terms which are formed by combining some of the input variables ($p_1(\cdot) \equiv 1$ corresponds to a constant term), M is the number of all the distinct terms, $\varepsilon(t)$ is the modelling error, and θ_i are the unknown parameters to be estimated. A compact matrix form corresponding to (7) is

$$Y = P\Theta + \Xi \tag{8}$$

where $Y = [y(1), y(2), \cdots, y(N)]^T$, $P = [p_1, p_2, \cdots, p_M]$, $p_i = [p_i(x(1)), p_i(x(2)), \cdots, p_i(x(N))]^T$, $\Theta = [\theta_1, \theta_2, \cdots, \theta_M]^T$, $\Xi = [\varepsilon(1), \varepsilon(2), \cdots, \varepsilon(N)]^T$.

3. Model Term Detection

In reality, not all the terms in the model (7) will be significant, some may be redundant and can be removed from the model. The forward orthogonal least squares (OLS) algorithm and the error reduction ratio (ERR) were originally introduced to determine which terms should be included in the model (Billings et al. 1988, Korenberg et al. 1988, Billings et al. 1989, Chen et al. 1989, Billings and Zhu 1994). This approach has been extensively studied in the past decade and widely applied in nonlinear system identification (see, for example, Chen et al. 1991, Wang and Mendel 1992, Zhu and Billings 1996, Chng et al. 1996, Hong and Harris 2001a). The forward OLS algorithm and the ERR approach for term detection will be briefly described below because this forms the basis of the new algorithm which pre-selects variables and the terms which is described in Section 4.

Assume that the regression matrix P is full rank in columns and can be orthogonally decomposed as

$$P = WA \tag{9}$$

where A is an $M \times M$ unit upper triangular matrix and W is an $N \times M$ matrix with orthogonal columns $w_1, w_2, \cdots, w_M$ in the sense that $W^T W = D = diag[d_1, d_2, \cdots, d_M]$ with $d_i = <w_i, w_i> = \sum_{t=1}^{N} w_i(t) w_i(t)$, and the symbol $<\cdot,\cdot>$ denotes the inner product of two vectors. The space spanned by the orthogonal basis $w_1, w_2, \cdots, w_M$ is the same as that spanned by the basis set $p_1, p_2, \cdots, p_M$, and (8) can be expressed as

$$Y = (PA^{-1})(A\Theta) + \Xi = WG + \Xi \tag{10}$$

where $G=[g_1,g_2,\cdots,g_M]^T$ is an auxiliary parameter vector, which can be calculated directly from Y and W by means of the property of orthogonality as

$$G=D^{-1}W^TY \tag{11}$$

or

$$g_i=\frac{<Y,w_i>}{<w_i,w_i>}, \quad i=1,2,\cdots,M \tag{12}$$

The parameter vector Θ is related by the equation $A\Theta=G$, and this can be solved using either a classical or modified Gram-Schmidt algorithm (Chen et al. 1989).

The number M of all the candidate terms in model (7) is often very large. Some of these terms may be redundant and should be removed to give a parsimonious model with only M_0 terms ($M_0 << M$). Detection of the significant model terms can be achieved using the OLS procedures described below.

Assume that the residual signal $\varepsilon(t)$ in the model (7) is uncorrelated with the past outputs of the system, then the output variance can be expressed as

$$\frac{1}{N}Y^TY=\frac{1}{N}\sum_{i=1}^{M}g_i^2w_i^Tw_i+\frac{1}{N}\Xi^T\Xi \tag{13}$$

Note that the output variance consists of two parts, one is the desired output, $(1/N)\sum_{i=1}^{M}g_i^2w_i^Tw_i$, which can be explained by the regressors, and the other part, $(1/N)\Xi^T\Xi$, represents the unexplained variance. Thus $(1/N)\sum_{i=1}^{M}g_i^2w_i^Tw_i$ is the increment to the explained desired output variance brought by w_i, and the i th error reduction ratio, ERR_i, introduced by w_i, can be defined as

$$ERR_i=\frac{g_i^2<w_i,w_i>}{<Y,Y>}\times100\%=\frac{<Y,w_i>^2}{<Y,Y><w_i,w_i>}\times100\%, \quad i=1,2,\cdots,M, \tag{14}$$

This ratio provides a simple but effective means for seeking a subset of significant regressors. The significant terms can be selected in a forward-regression manner according to the value of ERR_i. Several orthogonalization procedures, such as Gram-Schmidt, modified Gram-Schmidt and Householder transformation (Chen et al. 1989) can be applied to implement the orthogonal decomposition. Take the classical Gram-Schmidt algorithm as an example, the orthogonalization procedure can be implemented in a stepwise manner and this is described as follows.

```
**Step 1:** Set $I_1=\{1,2,\cdots,M\}$; $\sigma=Y^TY$;
    for i=1 to M
        $w_i=p_i$;
        $ERR_i=\frac{<Y,w_i>^2}{\sigma<w_i,w_i>}\times100\%$;
        $a_{11}=1$;
    end for
    $\ell_1=\arg\max_{i\in I_1}\{ERR_i\}$;
    $w_1^0=w_{\ell_1}$;  $g_1^0=<Y,w_1^0>/<w_1^0,w_1^0>$;
```

```
Step $j$, $j \ge 2$:
    For $j$=2 to $M$
        $I_j = I_{j-1} \setminus \{\ell_{j-1}\}$;
        for all $i \in I_j$
            $w_i = p_i - \sum_{k=1}^{j-1} \frac{< p_i, w_k^0 >}{< w_k^0, w_k^0 >} w_k^0$;
            $ERR_i = \frac{< Y, w_i >^2}{\sigma < w_i, w_i >} \times 100\%$;
        end for ( end loop for $i$ )
        $J_j = \{\arg_{i \in I_j}(w_i^T w_i < \tau)\}$; $I_j = I_j \setminus J_j$;          (15)
        $\ell_j = \arg\max_{i \in I_j}\{ERR_i\}$;
        $w_j^0 = w_{l_j}$; $g_j^0 = < Y, w_j^0 > / < w_j^0, w_j^0 >$;          (16)
        $a_{jj} = 1$;
        for $k$=1 to $j$-1
            $a_{kj} = < w_k^0, p_{l\,j} > / < w_k^0, w_k^0 >$;
        end for (end loop for $k$ )
    end for (end loop for $j$ )
```

The improved version of this algorithm (Zhu and Billings 1996) provides significant reductions in computation and is advantageous compared to the classical Gram-Schmidt algorithm when dealing with high order MIMO systems. Other recent studies by Hong and Harris (2001b) have proposed other improvements to this procedure.

***Remark* 1:** The candidate terms that are not chosen in the first step are orthogonalized with respect to all previously selected basis functions. Because of the orthogonality the j th term can be selected in the same way as in the first step. In Eq. (16), w_j^0 is the j th selected orthogonal term and g_j^0 is the corresponding parameter. Any numerical ill conditioning can be avoided by eliminating the candidate basis functions for which $w_i^T w_i$ are less than a predetermined threshold τ in Eq. (15), for example, $\tau = 10^{-r}$ and $r \ge 10$.

***Remark* 2:** The assumption that the regression matrix P is full rank in columns is unnecessary in the iterative forward OLS algorithm. In fact, if the M columns of the matrix P are linearly dependent, and assume that the rank in columns is M_1 ($<M$) , then the algorithm will stop at the M_1 -th step.

***Remark* 3:** If required, the procedure can be terminated at the M_0 -th step ($M_0 \le M_1$) when $1 - \sum_{i=1}^{M_0} ERR_i < \rho$, where ρ is a desired error tolerance, which can be learnt during the regression procedure. The final model is the linear combination of the M_0 significant terms selected from the M candidate terms $\{p_i\}_{i=1}^{M}$

$$y(t) = \sum_{i=1}^{M_0} g_i^0 w_i^0(t) + \varepsilon(t) \tag{17}$$

which is equivalent to

$$y(t) = \sum_{i=1}^{M_0} \theta_{\ell i} p_{\ell i}(x(t)) + \varepsilon(t) \tag{18}$$

where the parameters $\Theta^{(OLS)} = [\theta_{\ell_1}, \theta_{\ell_2}, \cdots, \theta_{\ell_{M_0}}]^T$ are calculated from the triangular equation $AG^{(0)} = \Theta^{(OLS)}$ with $G^{(0)} = [g_1^{(0)}, g_2^{(0)}, \cdots, g_{M_0}^{(0)}]^T$ 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 $a_{ij}\,(1 \le i < j \le M_0)$ are given in the above OLS algorithm.

4. Variable Selection Using the Forward OLS Algorithm

Detecting the significant model terms in nonlinear model estimation can be very computationally intensive because the number of potential terms can become very large. For example, the expression of Eq. (3) using a polynomial, radial basis function or wavelet basis can easily produce thousands of potential model terms because each expression is over many lagged variables. But if the significant variables could be determined as a first step, the problem is considerably simplified because now the model terms can be formed just from the pre-selected variables. A new algorithm is introduced below as one solution to this important problem. The method consists of pre-selecting the variables by searching over a set of locally linear or cross-bilinear modes initially and using a new sum of error reduction ratio criteria.

4.1 Linearization

Piecewise linear modelling identifies a series of local linear models that approximate the nonlinear behaviour under study over defined operating ranges of the system. Assume that the function f in model (2) is smooth enough so that the Taylor expansion is valid at least to second order for a small domain around a given operating point. Let $x^{(0)} = [x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)}]^T \in D$ be one of the operating points of the system, and D is the operating space of the system under study. Expand f at a small domain $\Delta x^{(0)}$ around $x^{(0)}$ to the first and second order respectively, then the operating point dependent linear and cross-bilinear models can be obtained as below

$$y\,|_{x^{(0)}} = f(x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)}) + \sum_{i=1}^{n} \left[\frac{\partial f}{\partial x_i} \right]_{x=x^{(0)}} (x_i - x_i^{(0)}) + \varepsilon(t) \tag{19}$$

and

$$y\,|_{x^{(0)}} = f(x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)}) + \sum_{i=1}^{n} \left[\frac{\partial f}{\partial x_i} \right]_{x=x^{(0)}} (x_i - x_i^{(0)}) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{x=x^{(0)}} (x_i - x_i^{(0)})(x_j - x_j^{(0)}) + \varepsilon(t) \tag{20}$$

Assume that $\partial f / \partial x_i$ and $\partial^2 f / \partial x_i \partial x_j$ are invariant in the domain $\Delta x^{(0)}$ around the operating point $x^{(0)}$, then (19) and (20) can be re-arranged into the forms

$$y(t) = a_0 + \sum_{i=1}^{n} a_i x_i + \varepsilon(t) \tag{21}$$

and

$$y(t) = c_0 + \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_i x_j + \varepsilon(t), \quad x = [x_1, x_2, \cdots, x_n]^T \in x^{(0)} + \Delta x^{(0)} \tag{22}$$

Clearly, if the variable x_i is significant with respect to the output y of the original nonlinear system, then this variable should also make a contribution to the linearized models (21) and (22). Thus the variable selection problem for the nonlinear system is simplified to the detection of the significant variables in the linearized models. For the linear model (21), the variable selection problem is the same as the term detection problem, and this can easily be solved using the forward OLS algorithm in the previous section. Note that the parameters of the linearized models are operating region dependent and therefore the significance of the variables will also be operating region dependent. The overall significance of certain variables will therefore have to be evaluated based on their significance in several operating regions.

Based on the linearized model (21), some variable selection approaches have been proposed, for example, the linear subset selection approach based on an exhaustive search and genetic algorithms (Mao and Billings 1999), the order and delay selection method based on the linearized model and a final prediction error (FPE) criterion (Gomm and Yu 2000), and the variable selection approach based on local models and the forward OLS algorithm (Hong and Harris 2001a).

The main advantage of system identification based on the piecewise linear approach is that well-known linear algorithms can be used to identify the model and develop control strategies in a manner which utilizes the wealth of knowledge and experience that is available for linear systems. There are several ways in which nonlinear systems can be approximated by local linear models (e.g. Billings and Voon 1987). The spatial piecewise linear models will, providing that the nonlinearities are smooth, provide an adequate representation of a nonlinear system. This however, can only be achieved at the expense of fitting a very large number of linear models, each valid in a small region of operation. This number increases exponentially as the number of intervals (operating regions) for each independent variable is increased. For example, a 10-variable function where each variable has 10 intervals results in 10^{10} linear models. Another problem is the selection of optimal operating regions for the locally piecewise linear models and this can be computationally time consuming. Improperly selected operating regions can result in a deterioration of the model accuracy. However, in the present study, the linearization is used as an initial step in the variable and term selection prior to fitting a nonlinear model and not as a method of constructing a linearizd model representation.

4.2 Variable Selection Using the Forward OLS Algorithm

In this paper, the linear and cross-bilinear models are used as an initial step for selecting significant variables for a nonlinear system. This is because most types of model structures applied in engineering are whole processoriented, therefore the global significant variables defined on the whole operating region D of the system under study are more suitable for the purpose of representing the system using a global model.

Consider the case of single input and single output dynamic systems. Assume that (21) and (22) are valid over the whole operating region D of the system. This results in the ARX and cross-bilinear models

$$y(t) = a_0 + \sum_{i=1}^{n_y} a_i y(t-i) + \sum_{j=1}^{n_u} b_j u(t-j) + \eta(t) \tag{23}$$

$$y(t) = c_0 + \sum_{i=1}^{n} c_i x_i(t) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_i(t) x_j(t) + \eta(t) \tag{24}$$

where $\eta(t)$ is a noise which accommodates the effects of measurement noise, unmeasured disturbances and modelling errors and which is assumed to be uncorrelated with the past input and output observations. $x_i(t) = y(t-i)$ for $i = 1,2,\cdots,n_y$ and $x_j(t) = u(t-j)$ for $j = n_y+1, n_y+2,\cdots,n$ ($n = n_y + n_u$), n_y, n_u are the maximum lags which are unknown and should be determined first. Eqs (23) and (24) are special cases of the NARMAX model (3). Clearly, if some of the lagged outputs $y(t-i)$ and/or lagged inputs $u(t-j)$ are important to the output $y(t)$ in the original system, then they must make significant contributions to the linear and cross-bilinear models (23) and (24). The model order n_y and n_u, and the significant lagged variables can be identified simultaneously by applying the forward OLS procedure to these models. In the following, the model order determination problem is initially considered and then the variable selection approach will be described.

Given the measured input and output data $\{u(t)\}_{t=1}^{N}$ and $\{y(t)\}_{t=1}^{N}$, let $u^t = [u(1), u(2),\cdots,u(t)]^T$ and $y^t = [y(1), y(2),\cdots, y(t)]^T$. Apply the forward OLS procedure described in Section 3 on the input and output data $\{u(t)\}_{t=1}^{N}$ and $\{y(t)\}_{t=1}^{N}$ to fit a set of linear models

$$\mathrm{M_L}: \quad y(t) = \hat{f}(y^{t-1}, u^{t-1}) = a_0 + \sum_{i=1}^{p} a_i y(t-i) + \sum_{j=1}^{q} b_j u(t-i) + \eta(t) \tag{25}$$

or a set of cross-bilinear models

$$\begin{aligned}\mathrm{M_{BL}}: \quad y(t) = \hat{f}(y^{t-1}, u^{t-1}) = a_0 + \sum_{i=1}^{p} a_i y(t-i) + \sum_{j=1}^{q} b_i u(t-j) + \sum_{i=1}^{p}\sum_{j=i+1}^{p} c_{ij} y(t-i) y(t-j) \\ + \sum_{i=1}^{q}\sum_{j=i+1}^{q} d_{ij} u(t-i) u(t-j) + \sum_{i=1}^{p}\sum_{j=1}^{q} e_{ij} y(t-i) u(t-j) + \eta(t)\end{aligned} \tag{26}$$

for $p = 1,2,\cdots,N_y$ and $q = 1,2,\cdots,N_u$, where N_y and N_u are the maximum output and input lags to be considered, that is, the order range to be tested is $1 \le p \le N_y$ and $1 \le q \le N_u$. Note that the real orders of the system under study might be unknown, therefore, it is required that N_y and N_u be sufficiently large. From the OLS algorithm described in Section 3, the corresponding orthogonalized model for (25) and (26) can be expressed as

$$y(t) = \sum_{i=1}^{m[p,q]} g_i^{[p,q]} w_i^{[p,q]}(t) + \eta(t) \tag{27}$$

where $m[p,q] = p + q + 1$ for the linear model set M_{L} and $m[p,q] = (p+q+1)(p+q+2)/2$ for the cross-bilinear model set M_{BL}. Corresponding to (27), the error reduction ratio (ERR) is defined as

$$ERR_i^{[p,q]} = \frac{\{g_i^{[p,q]}\}^2 < w_i^{[p,q]}, w_i^{[p,q]} >}{< Y, Y >}, \quad i=1,2,\ldots,m[p,q], \tag{28}$$

where the upper script $[p,q]$ in $ERR_i^{[p,q]}$ is used to emphasize that the value of the error reduction ratio depends on p and q, and $Y = [y(1), y(2), \cdots, y(N)]^T$.

Clearly, fitting model (23) or (24) to a nonlinear system will be an approximation and normally this would lead to biased estimates because the $\eta(t)$ term would include both noise and neglected or unmodelled nonlinear effects. Lumping all these effects together and denoting them as a coloured noise term demonstrates that this will usually induce bias. However, in the present algorithm, only the ERR values, which are used to pre-select the model terms, are required and it is proved in the Appendix that while the parameter estimates may be biased because of the neglected nonlinear terms described above, the relative values of ERR are always preserved. This is a significant result which both demonstrates an important property of the ERR but which is directly applicable in the current approximation.

Introduce the following criteria as an indicator for the optimal model order and significant variables:

- *Sum of error reduction ratios(SERR)*

$$SERR(p,q) = \sum_{i=1}^{m} ERR_i^{[p,q]} \tag{29}$$

This is a bounded function satisfying $0 < SERR(p,q) \le 1$ for any p and q according to the definition of the error reduction ratio (ERR). For a fixed p, this is a non-decreasing function with respect to the index q, and vice versa. Since the most significant variables are initially selected according to the ERR values, they make greater contributions compared with the later selected variables. In fact most of the later selected variables make very little contribution to the system output and therefore can be ignored. This means that the function $SERR(p,q)$ will become flat from a certain point (p_1, q_1), based on which the model order can be determined. Since for a fixed p, $SERR(p,q)$ is a non-decreasing function with respect to the index q, the best maximum input lag (model order of the input) can be determined by inspecting the plot of the function $SERR(p,q)$ versus q with a fixed p. The point q_1, from which the function $SERR(p,q)$ becomes stable, can be chosen as the best maximum input lag. Similarly, the best maximum output lag (model order of the output) can be determined by inspecting the plot of the function $SERR(p,q)$ versus p with a fixed q. The point p_1, from which the function $SERR(p,q)$ becomes stable, can be chosen as the best maximum output lag.

- *The variance of the kth-step –ahead-prediction errors*

The *k*th-step –ahead-prediction errors are defined as

$$\xi^{[p,q]}(t) = y(t) - \hat{y}(t \mid t-k) \tag{30}$$

where $\hat{y}(t \mid t-k)$ is the *k*th-step-ahead prediction. Again, the upper script [*p*,*q*] in $\xi^{[p,q]}$ is used to emphasize that the prediction errors are evaluated based on a model with the order *p* and *q* for the lagged output and input, respectively. Let $VPE\ (p,q) = \text{var}(\xi^{[p,q]})$ be the variance of the above *k*th-step –ahead-prediction errors. For a fixed p, this is a non-decreasing function with respect to the index q, and vice versa. Similar to the function $SERR\ (p,q)$, $VPE\ (p,q)$ will also become flat from a certain point (p_2, q_2), based on which the model order can be determined.

- *The final prediction errors (FPE)*

The FPE is defined as (Akaike 1969)

$$FPE(p,q) = \frac{N + m\gamma}{N - m\gamma} VPE(p,q) \tag{31}$$

where *N* is the data length, m is the number of model terms, $VPE(p,q)$ is the covariance of the *k*th-step – ahead-prediction errors, and γ is an adjustable parameter. Similar to the function $SERR\ (p,q)$, $FPE\ (p,q)$ will also becomes flat from a certain point (p_3, q_3), based on which the model order can be determined.

The variable selection approach based on the forward OLS algorithm can now be summarized as below:

(*i*) Perform the forward OLS procedure on the system input and output data $\{u(t)\}_{t=1}^{N}$ and $\{y(t)\}_{t=1}^{N}$ to fit a set of linear models M_{L} ($p = 1,2,\cdots,N_y$, $q = 1,2,\cdots,N_u$) with the form of (25).

(*ii*) Determine the model order n_y and n_u (the maximum input and output lags) using the above criteria, a best choice is to set $n_y = \max\{p_1, p_2, p_3\}$ and $n_u = \max\{q_1, q_2, q_3\}$, where p_1, p_2, p_3 and q_1, q_2, q_3 are defined in the definitions of SEER, VPE and FPE.

(*iii*) For the chosen model order n_y and n_u, fit a linear model with the form of (23). Orthogonize this model using the forward OLS algorithm and determine the ERR values for each corresponding variable. The significant variables can then be easily found by inspecting the ERR values of the corresponding variables since a significant variable usually possesses a far greater value of ERR than that of an insignificant variable.

(*iv*) In order that the significant variables are selected properly and sufficiently, check the values of $SERR(p,q)$, if they are too small, for example, if $SERR(p,q) \leq 0.8$ for large p and q, then it follows that a linearised model in step (*iii*) is insufficient to represent the original system. In this case, a series of cross-bilinear models or polynomial models with a higher degree should be fitted in step (*i*), and steps (*ii*) and (*iii*) should be repeated to find suitable model orders n_y and n_u. The significant terms for these models can be determined in accordance with the ERR values of the corresponding terms. Note that the related significant variables are contained in the significant terms and can easily be found by inspecting the significant terms.

***Remark* 4:** The variable selection procedure described above was developed for input-output system identification, but it can also be applied to find the embedding dimension (model order) for time series (autonomous systems). Although the procedure was described for SISO system modelling, it can easily be extended to the MIMO case. Once the global significant variables are successfully determined, they can be used to form model terms in polynomials, radial basis function networks, neural networks, wavelet networks and other modelling structures.

***Remark* 5:** For short data sets(e.g., 300-500 samples), the variable selection algorithm can be applied to all the data at one time to determine the significant variables based on the whole data set. For longer input-output data sets (e.g., 1000-2000 samples), it is recommended that the data set is divided into overlapped data segments initially before applying the variable selection algorithm on each data segment to determine the significant variables, and then finally to collect all the individual significant variables together. For example, let the whole input-output data set be $D=\{(x(t),y(t)):x\in R^n, y\in R, 1\leq t\leq N\}$. Divide D into different data segments, say, $D_k=\{(x(t),y(t)\}_{t=1+(k-1)N/s}^{t=(k+1)N/s}$, where *s* is the number to be divided and *k*=1,2,...,*s*-1.

***Remark* 6:** The above procedure for determining model terms and selecting significant variables is automatically performed. The inputs required by the software are the system input-output data, and the maximum output and input lags N_y and N_u. The outputs of the software are the values of SEER(*p*,*q*), VPE(*p*,*q*) and FPE(*p*,*q*) versus *p* and *q*, with $p=1,2,\cdots,N_y$ and $q=1,2,\cdots,N_u$. Based on the values of SEER(*p*,*q*), VPE(*p*,*q*) and FPE(*p*,*q*), the best model order, the significant terms along with the significant variables can easily be determined.

***Remark* 7:** Notice that, for an input-output system, once N_y and N_u are chosen, the number of linear or cross-bilinear models to estimate in the model set M_L or M_BL with the form of (25) or (26) is $n_{\text{model}}=N_yN_u$, which means that a large number of models of different orders will be involved for large N_y and N_u. At first sight this would appear to be time consuming because of the need to fit a great number of models. However, in reality this is often not the case because fitting a model set M_L or M_BL with reasonably large N_y and N_u takes a matter of a few minutes on a modern computer. Tables 1(a)-(c) list the average CPU time on a Sun workstation, required to fit a model set with different input-output data lengths and different maximum lags N_y and N_u.

Table 1(a) Time required to fit a model set with different model orders(input-output data length N=500)

The order range to be tested for the output (N_y)	The order range to be tested for the input (N_u)	Number of models to be fitted in the model set M_L or M_{BL} ($N_y N_u$)	Average CPU time required to fit the model set	
			Linear model set M_L	Bilinear model set M_{BL}
5	5	25	< 2 sec	< 4 sec
5	10	50	< 5 sec	< 8 sec
10	5	50	< 5 sec	< 8 sec
10	10	100	< 6 sec	< 30 sec
20	20	400	< 30 sec	< 20 min

Table 1(b) Time required to fit a model set with different model orders(input-output data length N=1000)

The order range to be tested for the output (N_y)	The order range to be tested for the input (N_u)	Number of models to be fitted in the model set M_L or M_{BL} ($N_y N_u$)	Average CPU time required to fit the model set	
			Linear model set M_L	Bilinear model set M_{BL}
5	5	25	< 5 sec	< 4 sec
5	10	50	< 5 sec	< 10 sec
10	5	50	< 5 sec	< 10 sec
10	10	100	< 10 sec	< 60 sec
20	20	400	< 40 sec	< 60 min

Table 1(c) Time required to fit a model set with different model orders(input-output data length N=2000)

The order range to be tested for the output (N_y)	The order range to be tested for the input (N_u)	Number of models to be fitted in the model set M_L or M_{BL} ($N_y N_u$)	Average CPU time required to fit the model set	
			Linear model set M_L	Bilinear model set M_{BL}
5	5	25	< 5 sec	< 5 sec
5	10	50	< 5 sec	< 20 sec
10	5	50	< 5 sec	< 20 sec
10	10	100	< 15 sec	< 100 sec
20	20	400	< 50 sec	< 150 min

5. Examples and Applications

In this section, several examples are provided to illustrate the efficiency of the new algorithm introduced above for model order determination, variable selection and term detection. For all the examples, the one-step-ahead prediction was considered, that is, k=1 in Eq (30). As to the adjustable parameter γ in Eq (31), this was set to γ =1 for Examples 1 and 2, and γ =2 for the other examples.

5.1 Time series

(A) Example 1— a time series generated from a continuous time system

Consider a second order continuous linear system

$$\ddot{y}+\omega^2 y=0 \tag{32}$$

with $\omega=1$, $y(0)=0$, $\dot{y}(0)=0.9998$. This system was simulated using a Runge-Kutta integral routine to obtain 315 even-sampled data points with the sampling interval $T=0.02$. A series of linear models of the form (25) with q=0 and different values of p were fitted. The values obtained for the criteria SERR and FPE are shown in Figure 1, which clearly shows that the model order should be 2. Setting $n_y=2$ and $n_u=0$, a linear model (23) was estimated. The detected model terms, estimated parameters and the corresponding ERR values are listed in Table 2, in which the constant term can obviously be omitted. Notice that using the approximations $\dot{y}(t)\approx[y(t+T)-y(t)]/T$ and $\ddot{y}(t)\approx[\dot{y}(t+T)-\dot{y}(t)]/T$, the difference equation for the system (32) is approximately

$$y[kT]=2y[(k-1)T]-(1+T^2\omega^2)y[(k-2)T] \tag{33}$$

and this can be considered as the true time discrete model for (32) with the initial condition $y(0)=0$, $y(1)=\omega\cos(\omega T)$. Clearly the identified terms and parameters in Table 2 coincide with the true model almost perfectly.

Table 2 The selected variables, estimated parameters and the corresponding ERR values for Example 1

Search Steps	Terms	Estimates of Parameters	ERRs
1	y(t-1)	0.19996E+01	0.99960E+00
2	y(t-2)	-0.10000E+01	0.39700E-03
3	constant	-0.54742E-16	0.31362E-31

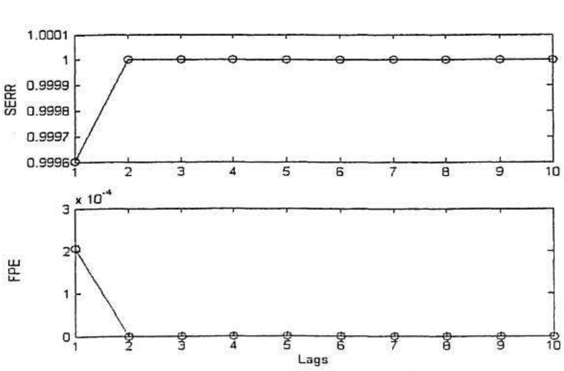


Figure 1 The criteria SERR and FPE for Example 1

(B) Example 2 — the fourth-order Henon map

The following Henon map was used

$$y(t) = 1 - 1.4y^2(t-2) + 0.3y(t-4) \tag{34}$$

to generate a time sequence of 500 points with the initial condition $y(1) = 0$, $y(2) = 1$, $y(3) = 0.5$, $y(4) = -0.5$.

A series of linear models of the form (25) with q=0 and different values of p were initially fitted and the criteria SERR and FPE were calculated and shown in Figure 2(a) and (b). Both criteria indicate that the linear model order should be 10. However, the small SERR values suggest that linear models are insufficient to represent this data set. Therefore, cross-bilinear models or polynomial models with a higher degree should be fitted and tested.

A series of cross-bilinear models with q=0 and different values of p were then fitted and the two criteria SERR and FPE were calculated and are shown in Figure 2(c) and (d). Clearly, the model order n_y should be 4. Therefore a cross-bilinear model with $n_y = 4$ and $n_u = 0$ was estimated. The detected model terms, estimated parameters and the corresponding ERR values based on the cross-bilinear models were listed in Table 3, in which only the first 3 terms are significant, giving the significant variables as *y*(*t*-2), *const*, and *y*(*t*-4). Note that 85.65% of the system output variance can be explained by the first three terms. In fact the other terms can be removed since their ERR values or parameters are almost zero and can be omitted. If the 4th term is forced out of the model (this term possesses a very small ERR value), then almost 100% of the system output variance can be explained by the first three terms.

Table 3 The selected variables, estimated parameters and the corresponding ERR values for the Henon map in Example 2

Search Steps	Terms	Estimates of Parameters	ERRs
1	y(t-2)y(t-2)	-0.14000E+01	0.39090E+00
2	Constant	0.10000E+01	0.39048E+00
3	y(t-4)	0.30000E+00	0.75099E-01
4	y(t-4)y(t-4)	0.70664E-17	0.14353E+00
5	y(t-2)	-0.47299E-16	0.10111E-28
6	y(t-1)y(t-1)	0.91522E-16	0.23411E-30
7	y(t-3)y(t-3)	0.31982E-16	0.15082E-30
8	y(t-1)	0.10450E-16	0.11481E-29
9	y(t-2)y(t-3)	0.23968E-16	0.37141E-31
10	y(t-2)y(t-4)	0.52989E-16	0.15286E-31
11	y(t-3)	0.26969E-16	0.82793E-32
12	y(t-1)y(t-3)	0.10238E-15	0.29971E-32
13	y(t-1)y(t-4)	0.35372E-16	0.42335E-33
14	y(t-3)y(t-4)	0.41892E-16	0.41299E-33
15	y(t-1)y(t-2)	0.37652E-16	0.54070E-16

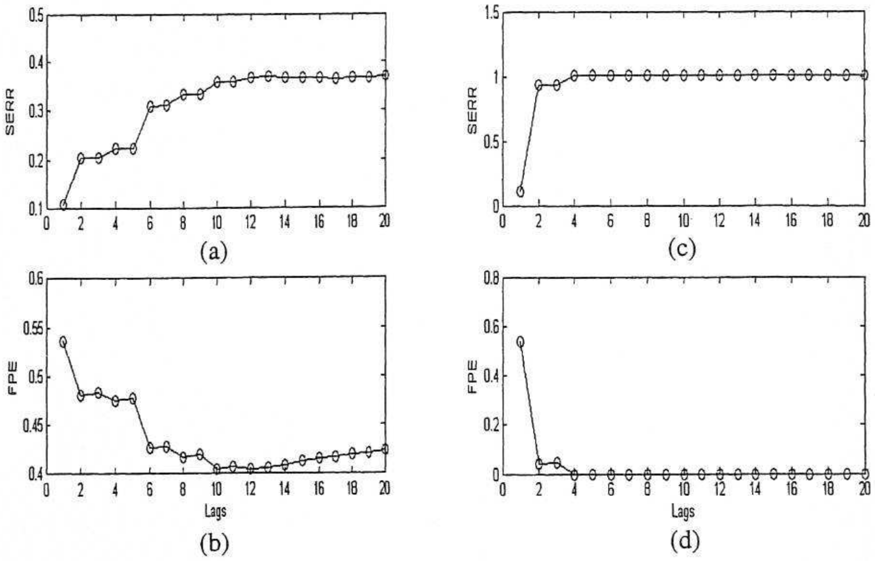


Figure 2 The criteria SERR and FPE for the Henon map in Example 2 described by Eq. (34). The left two plots (a) and (b) were calculated based on linear models; the right two plots (c) and (d) were calculated based on cross-bilinear models.

(C) Example 3 —the sunspot time series

This example uses the Wolf sunspot data series recording the annual sunspot indices from 1700 to 1999 as shown in Figure 3. A series of linear models of the form (25) with q=0 and different p were initially fitted. The two criteria SERR and FPE were calculated and are shown in Figure 4. Both criteria indicate that the model order should be 9. Setting n_y=9 and n_u=0 in the linear model (23), the detected model terms, estimated parameters and the corresponding ERR values are listed in Table 4, which clearly shows that 94.37% of the system output variance can be explained by the first four terms (variables), and 94.47% of the system output variance can be explained by all ten terms. Therefore the significant variables were chosen as {$y(t$-1), $y(t$-9), $y(t$-2), $const$}.

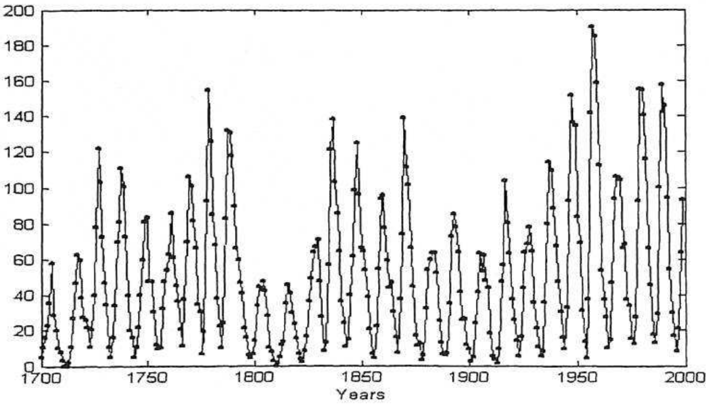


Figure 3 The Wolf sunspot data series in Example 3

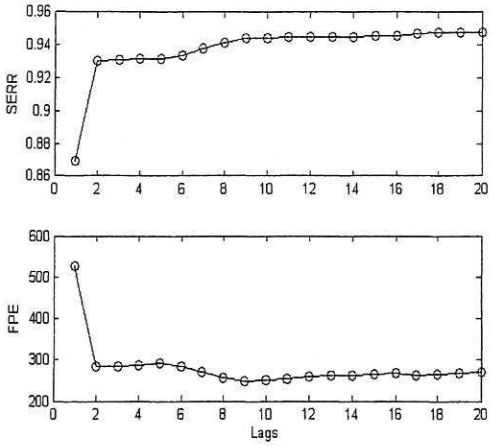


Figure 4 The two criteria SERR and FPE for the sunspot data in Example 3

Table 4 The selected variables, estimated parameters and the corresponding ERR values for the sunspot data in Example 3

Search Steps	Terms	Estimates of Parameters	ERRs
1	y(t-1)	0.11913E+01	0.86185E+00
2	y(t-9)	0.22402E+00	0.26600E-01
3	y(t-2)	-0.43154E+00	0.26600E-01
4	constant	0.62705E+01	0.13907E-02
5	y(t-3)	-0.16673E+00	0.24383E-03
6	y(t-4)	0.18215E+00	0.18311E-03
7	y(t-5)	-0.13313E+00	0.50612E-03
8	y(t-6)	0.41561E-01	0.45281E-04
9	y(t-8)	-0.29072E-01	0.28019E-04
10	y(t-7)	0.57414E-02	0.70913E-06

The objective here was to identify a hybrid nonlinear model a combing polynomial representation with a wavelet expansion of the form

$$\begin{aligned} y(t) &= f(y(t-1), y(t-2), \cdots, y(t-9)) + e(t) \\ &= a_0 + \sum_{i=1}^{9} a_i x_i(t) + \sum_{i=1}^{8} \sum_{j=i+1}^{9} b_{ij} x_i(t) x_j(t) + \sum_{p=1}^{3} f_p(z_p(t)) + e(t) \end{aligned} \tag{35}$$

where $x_i(t) = y(t-i)$ for i=1,2, ..., 9 and $z_1(t) = y(t-1)$, $z_2(t) = y(t-2)$, $z_3(t) = y(t-9)$. Each $f_p(z_p(t))$ in Eq. (35) was approximated using a multiresolution wavelet decomposition

$$f_p(z_p(t)) = \sum_{k \in K^0} \alpha_{j_0,k}^{(p)} \phi_{j_0,k}(z_p(t)) + \sum_{j=j_0}^{J} \sum_{k \in K_j} \beta_{j,k}^{(p)} \varphi_{j,k}(z_p(t)), \; p = 1,2,3. \tag{36}$$

In this example, the 4th-order B-spline wavelet and scaling functions were used with an initial resolution scale level j_0=0 and the highest resolution scale level J=4. Inserting (36) into (35), a linear-in-the-parameters equation is obtained, and the parameters in this model can be estimated using the forward OLS algorithm in Section 3. Notice that the data points were initially normalized and the modelling procedure was performed on the standard interval $[0,1]$. The model outputs can then be recovered to the system original operating domain by inverse transforms. The first 280 data points were used for model identification. The final identified model involved 15 B-spline wavelet regressors selected from 241 candidate terms. The terms, parameters and corresponding ERR values are listed in Table 5. The 2-step-ahead predictions are shown in Figure 5.

Table 5 The selected model terms, estimated parameters and the corresponding ERR values for the Wolf sunspot data in Example 3

Number	Terms	Parameters	ERRs [×100%]
1	$y(t-1)$	8.85324e-001	86.18723
2	$y(t-9)$	2.06379e-001	5.38071
3	$y(t-2)$	-2.08325e-001	2.65944
4	$\varphi_{0,-1}(y(t-1))$	-1.04111e+000	0.53544
5	$\phi_{0,0}(y(t-2))$	2.96762e+000	0.46384
6	$\varphi_{4,11}(y(t-9))$	4.53625e-002	0.19253
7	$\varphi_{4,6}(y(t-9))$	2.28958e-002	0.15577
8	$y(t-1)\ y(t-9)$	1.10663e+000	0.12643
9	$y(t-2)\ y(t-9)$	-1.16830e+000	0.11431
10	$y(t-3)$	-1.84325e-001	0.18242
11	$\varphi_{0,-1}(y(t-9))$	-4.27452e-001	0.19795
12	$\varphi_{4,11}(y(t-2))$	3.35518e-002	0.10391
13	$\varphi_{4,-3}(y(t-1))$	-1.77625e-002	0.09915
14	$\varphi_{4,5}(y(t-9))$	-1.70824e-002	0.10854
15	$\varphi_{3,-1}(y(t-9))$	9.74233e-003	0.06713
Note:	$\phi_{j,k}(x)=2^{j/2}\phi(2^j x-k)$ — the 4th-order B-splne functions; $\varphi_{j,k}(x)=2^{j/2}\varphi(2^j x-k)$ —the 4th-order B-splne wavelets.		

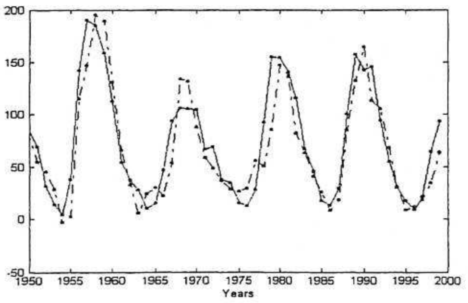


Figure 5 The 2-step-ahead predictions using the identified model listed in Table 5 for Example 3. (Dashed line denotes the predictions)

5.2 Input-output systems

(A) Example 4—a high-order bilinear system

Consider the fourth-order bilinear system (Roy et al. 1996)

$$\begin{aligned} y(t) = \;& 0.8833u(t-1)+0.0393u(t-2)+0.8546u(t-3)+0.8528u^2(t-1) \\ & +0.7582u(t-1)u(t-2)+0.1750u(t-1)u(t-3)+0.0864u^2(t-2) \\ & +0.4916u(t-2)u(t-3)+0.0711u^2(t-3)-0.0375y(t-1) \\ & -0.0598y(t-2)-0.0370y(t-3)-0.0468y(t-4)-0.0476y^2(t-1) \\ & -0.0781y(t-1)y(t-2)-0.0189y(t-1)y(t-3)-0.0626y(t-1)y(t-4) \\ & -0.0221y^2(t-2)-0.0617y(t-2)y(t-3)-0.0378y(t-2)y(t-4) \\ & -0.0041y^2(t-3)-0.0543y(t-3)y(t-4)-0.0603y^2(t-4)+\varepsilon(t) \end{aligned} \tag{37}$$

where $u(t)$~$N(0,\sigma^2)$ with σ=1, and $\varepsilon(t)$ ~$N(0,\sigma_\varepsilon^2)$ with σ_ε=0.05. This model was used to generate 1000 input and output data shown in Figure 6.

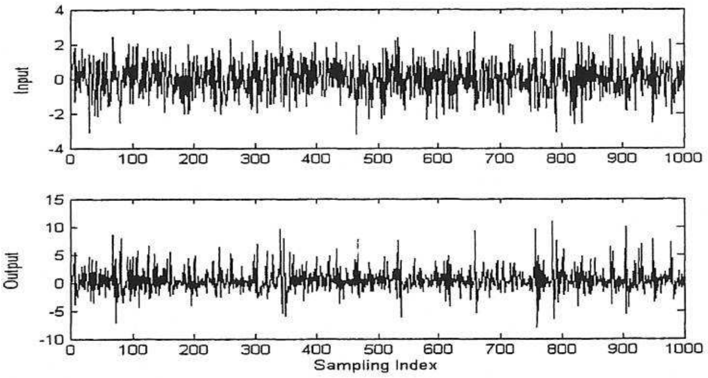


Figure 6 Input and output data for the system described by (37) in Example 4

A set of linear models with the form of (25) for different values of p and q were initially fitted and the two criteria SERR and FPE were calculated and are shown in Figure 7. The two three-dimensional plots for SERR(Figure 7(a)) and FPE (Figure 7(e))were decomposed into two-dimensional graphs shown in Figure 7(c) and (d) and Figure 7(e) and (f) respectively. Although the values of SERR and FPE clearly indicate that the model order is n_y =4 and n_u =3, the very small values of SERR also suggest that the linear model approximations are insufficient for this data set. Therefore, cross-bilinear models or polynomial models should be fitted and tested.

A series of cross-bilinear models or the form (26) with different values p and q were then fitted and the two criteria SERR and FPE were calculated and are shown in Figure 8. Now 100% of the system output variance can be explained by the terms produced by the seven variables {y(t-1), y(t-2), y(t-3), y(t-4), u(t-1), u(t-2), u(t-3)}. A cross-bilinear model was therefore fitted by setting n_y=4 and n_u=3, and the detected model terms, estimated parameters and the corresponding ERR values obtained are listed in Table 6.

Setting the SERR threshold value at ρ=0.0001 so that when $1-\text{SERR} < \rho$ the search procedure terminates automatically, results in 23 terms being selected from the 36 candidate terms. It can be seen from Table 6 that the selected terms are identical with the original system, and the estimated parameters are very close to the true values. Simulation results show that, if the noise sequence $\varepsilon(t)$ in (37) was set to be zero, then both the detected terms and the estimated parameters are identical to the real system.

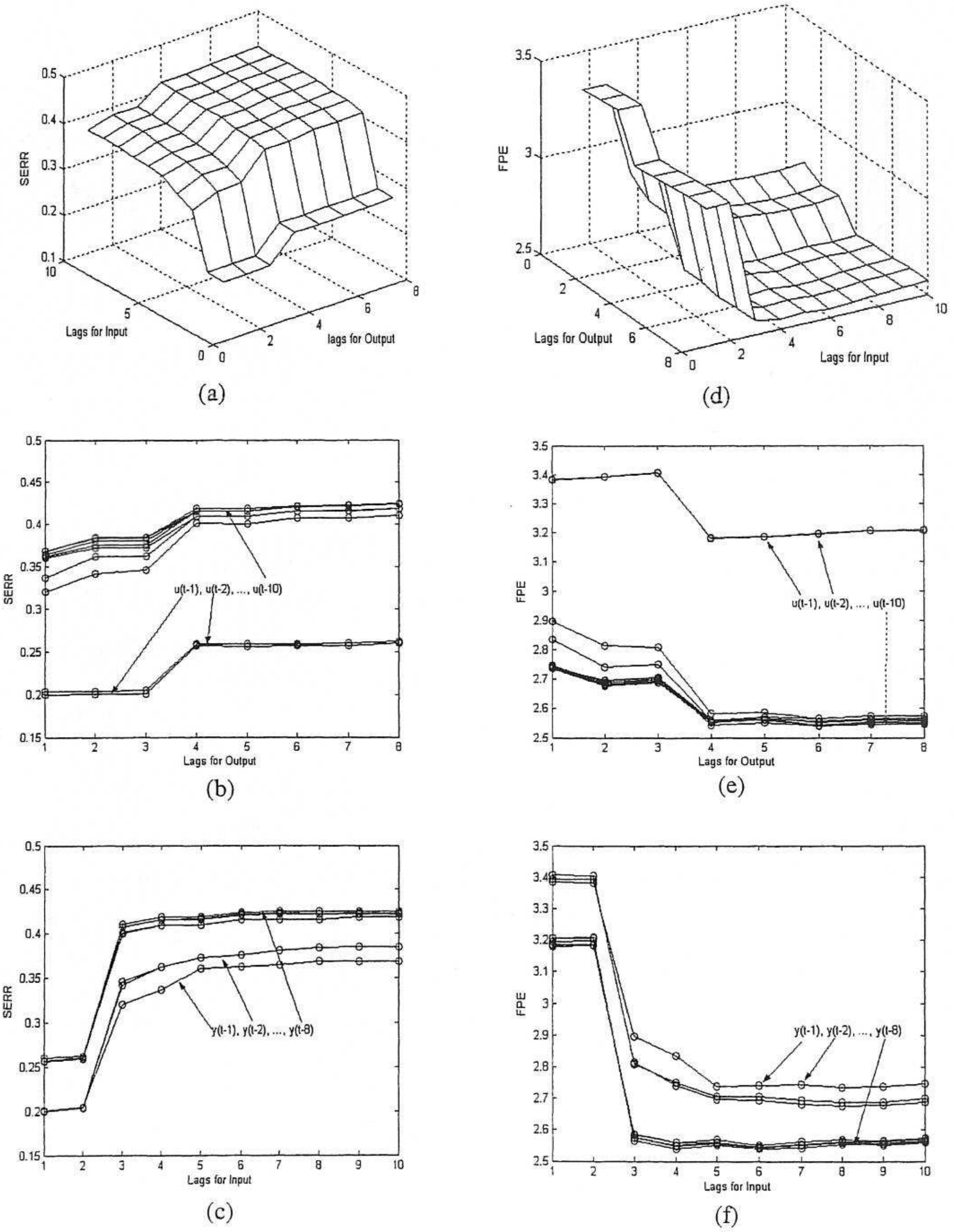


Figure 7 The criteria SERR and FPE calculated from a set of linear models for the system described by Eq. (37) in Example 4. (a) SERR; (b) SERR vs n_y; (c) SERR vs n_u; (d) FPE; (e) FPE vs n_y; (f) FPE vs n_u.

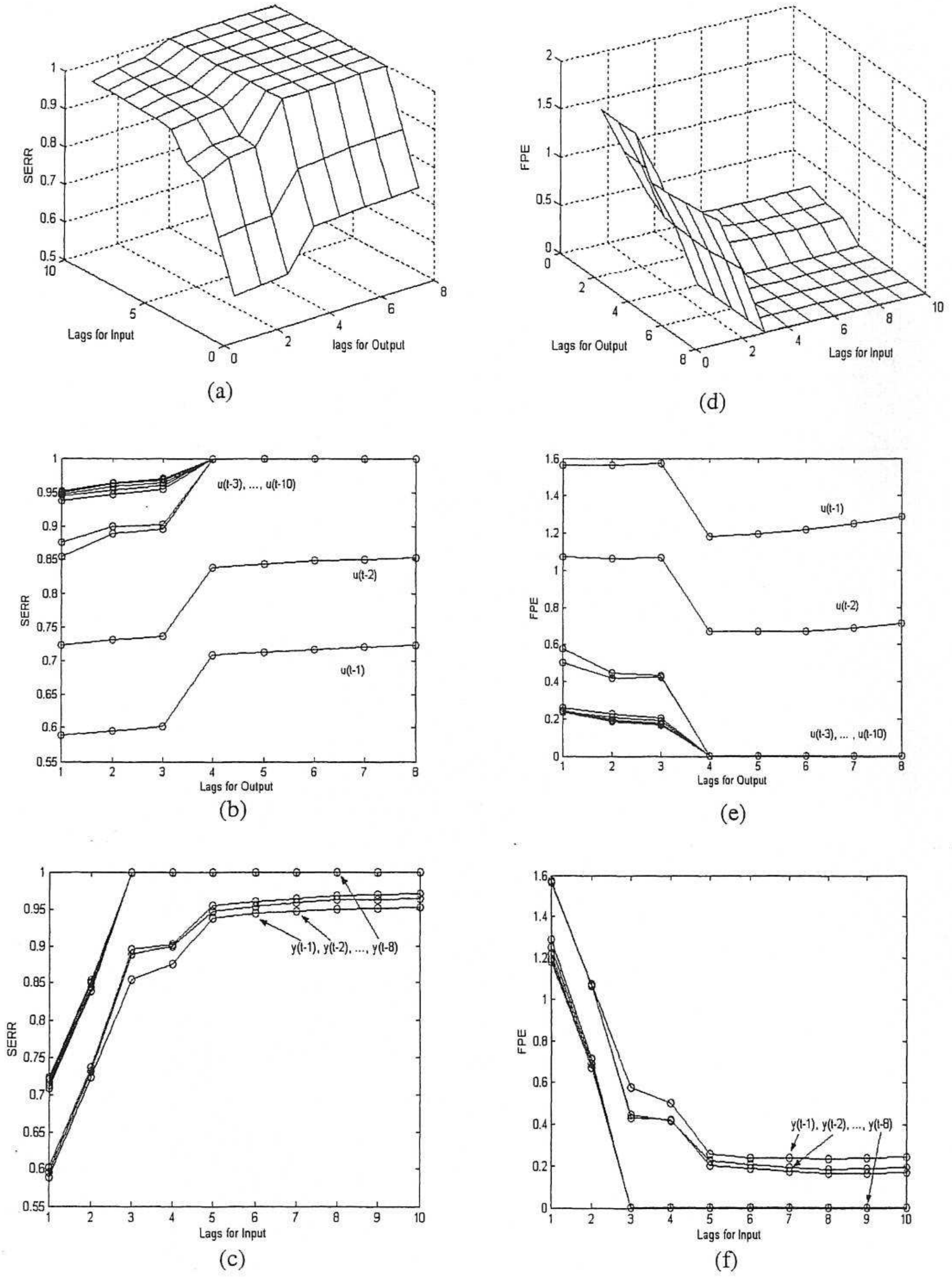


Figure 8 The criteria SERR and FPE calculated from a set of cross-bilinear models for the system described by Eq. (37) in Example 4.

(a) SERR; (b) SERR vs n_y; (c) SERR vs n_u; (d) FPE; (e) FPE vs n_y; (f) FPE vs n_u.

Table 6 The selected variables, estimated parameters and the corresponding ERR values for the cross-bilinear system (37) in Example 4

Search Steps	Terms	Estimates of Parameters	ERRs
1	u(t-1)u(t-1)	0.85282E+00	0.32594E+00
2	u(t-1)	0.88332E-01	0.16882E+00
3	u(t-1)u(t-2)	0.75819E+00	0.15570E+00
4	u(t-3)	0.85519E+00	0.12282E+00
5	y(t-4)y(t-4)	-0.60268E-01	0.65865E-01
6	y(t-1)y(t-1)	-0.47546E-01	0.34552E-01
7	u(t-2)u(t-3)	0.49224E+00	0.20894E-01
8	y(t-1)y(t-2)	-0.78124E-01	0.32585E-01
9	y(t-2)y(t-3)	-0.61620E-01	0.15576E-01
10	y(t-3)y(t-4)	-0.54154E-01	0.12528E-01
11	y(t-1)y(t-4)	-0.62702E-01	0.11607E-01
12	y(t-2)y(t-4)	-0.37560E-01	0.10921E-01
13	u(t-1)u(t-3)	0.17442E+00	0.76008E-02
14	y(t-2)y(t-2)	-0.22009E-01	0.57029E-02
15	y(t-1)y(t-3)	-0.18924E-01	0.24348E-02
16	u(t-2)u(t-2)	0.84913E-01	0.15185E-02
17	y(t-4)	-0.46877E-01	0.12276E-02
18	y(t-3)	-0.37336E-01	0.15254E-02
19	u(t-3)u(t-3)	0.70752E-01	0.49042E-03
20	y(t-2)	-0.59823E-01	0.11875E-02
21	y(t-1)	-0.37849E-01	0.13413E-03
22	u(t-2)	0.40205E-01	0.17583E-03
23	y(t-3)y(t-3)	-0.40757E-02	0.16986E-03
24	y(t-2)u(t-3)	-0.40556E-03	0.96717E-08
25	yt-1)u(t-3)	-0.45907E-03	0.10518E-08
26	y(t-4)u(t-1)	0.26991E-03	0.59791E-08
27	y(t-4)u(t-2)	0.25174E-03	0.67826E-08
28	y(t-3)u(t-2)	-0.23528E-03	0.29796E-08
29	y(t-1)u(t-1)	0.16047E-03	0.21019E-08
30	Constant	-0.42955E-03	0.20320E-08
31	y(t-4)u(t-3)	-0.18180E-03	0.14525E-08
32	y(t-3)u(t-3)	-0.14895E-03	0.11996E-08
33	y(t-1)u(t-2)	-0.10528E-03	0.11102E-08
34	y(t-2)u(t-2)	0.96455E-04	0.61201E-09
35	y(t-3)u(t-1)	-0.83313E-04	0.51567E-09
36	y(t-2)u(t-1)	0.13999E-05	0.14056E-12

(B) Example 5— a discrete nonlinear system

Consider the following nonlnear input-output system

$$y(t)=\frac{P(t)}{Q(t)}+R(t)+\varepsilon(t) \tag{38a}$$

where

$$P(t)=y(t-1)y(t-2)+y(t-1)y(t-3)+y(t-2)y(t-3)$$

$$+\,y(t-1)u(t-1)+y(t-2)u(t-2)+y^2(t-3)u(t-2)+u(t-1)u(t-2) \tag{38b}$$

$$Q(t)=1+y^2(t-1)+y^2(t-2)+y^2(t-3) \tag{38c}$$

$$R(t)=\sin[y(t-1)+1]+\sin[y(t-2)+2]+\sin[y(t-3)+3] \tag{38d}$$

The input $u(t)$ was selected to be a random variable uniformly distributed in the range [-1, 1], and the noise sequence $\varepsilon(t)\sim N(0,0.1^2)$.

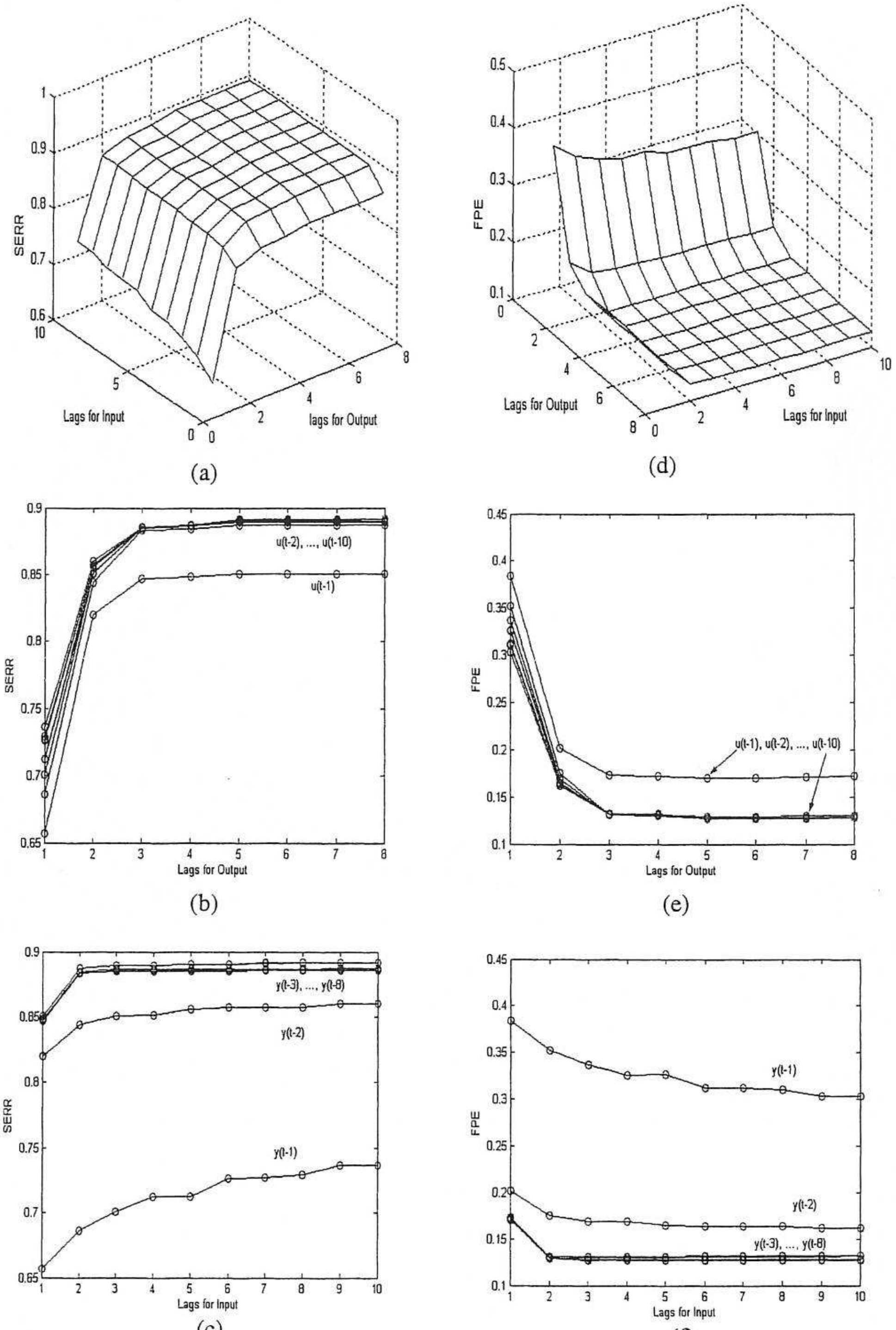
A data set of 1000 samples of the input and output was generated. A set of linear models with the form of (25) for different values of p and q were initially fitted and the two criteria SERR and FPE were calculated and are shown in Figure 9. The values of SERR and FPE clearly indicate that the model order is n_y=3 and n_u=2. Note that 88.01% of the system output variance can be explained by the terms produced by the seven variables {*y*(t-1), *y*(t-2), *y*(t-3), *u*(t-1), *u*(t-2)}

A set of cross-bilinear models with different values p and q were also fitted and the two criteria SERR and FPE were calculated and shown in Figure 10. Now 98.73% of the system output variance can be explained by the terms combined by the five variables {*y*(t-1), *y*(t-2), *y*(t-3), *u*(t-1), *u*(t-2)}.

An NARMAX model of degree 3 was then fitted using the selected variables {*y*(t-1), *y*(t-2), *y*(t-3), *u*(t-1), *u*(t-2)}. The SERR threshold value was set at ρ=0.05 so that when $1-\text{SERR}<\rho$ the search procedure terminates automatically, and the selected terms and the corresponding ERRs are shown in Table 7. The one-step-ahead predictions and the model predicted outputs along with the original output data are shown in Figure 11, where only the data between 300 and 500 are shown in order to illustrate the fit more clearly.

Table 7 The selected variables, estimated parameters and the corresponding ERR values for the system described by (38) in Example 5

Search Steps	Terms	Estimates of Parameters	ERRs
1	constant	0.19076E+01	0.56336E+00
2	y(t-2)	-0.13536E+00	0.18738E+00
3	y(t-3)	-0.96857E+00	0.77672E-01
4	y(t-3)u(t-2)	0.42188E+00	0.55734E-01
5	y(t-1)y(t-1)y(t-1)	-0.90980E-01	0.22329E-01
6	y(t-2)y(t-2)	-0.22479E+00	0.16928E-01
7	y(t-1)u(t-1)	0.20278E+00	0.16749E-01
8	y(t-3)y(t-3)y(t-3)	0.91578E-01	0.85737E-02
9	y(t-1)y(t-3)y(t-3)	0.13027E+00	0.93642E-02

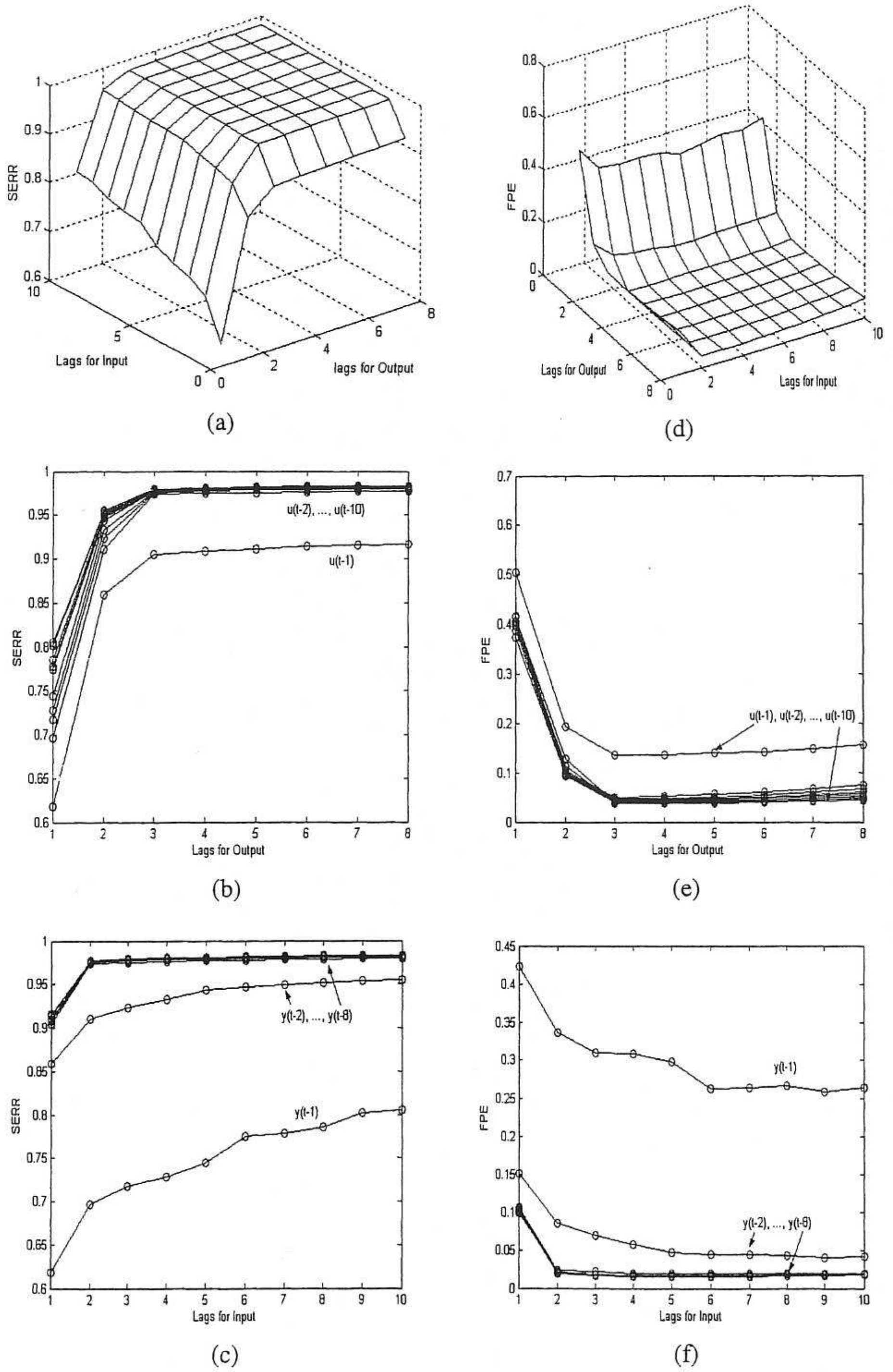


(a) (d)

(b) (e)

(c) (f)

Figure 9 The criteria SERR and FPE calculated from a set of linear models for the system described by Eq. (38) in Example 5. (a) SERR; (b) SERR vs n_y; (c) SERR vs n_u; (d) FPE; (e) FPE vs n_y; (f) FPE vs n_u.



(a) (d)

(b) (e)

(c) (f)

Figure 10 The criteria SERR and FPE calculated from a set of cross-bilinear models for the system described by Eq. (38) in Example 5.
(a) SERR; (b) SERR vs n_y; (c) SERR vs n_u; (d) FPE; (e) FPE vs n_y; (f) FPE vs n_u.

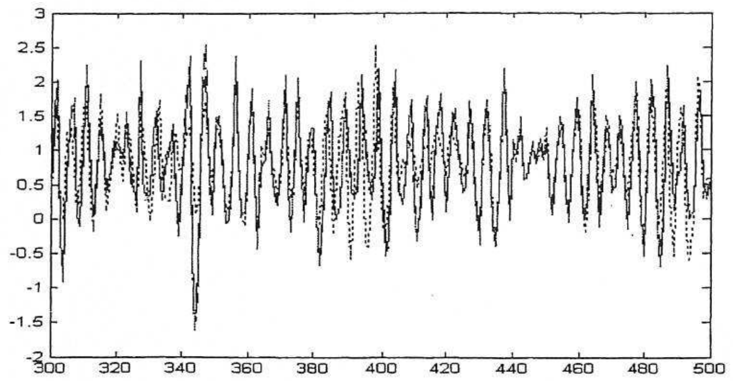


Figure 11 One-step-ahead predicted outputs, model predicted outputs and the original data for Example 5. (Solid--original data; Dashed--model predicted outputs; Dotted--one-step-ahead predictions)

(C) Example 6—a terrestrial magnetosphere system

Figure 12 shows 1000 data points of the measurement of the solar wind parameter VB_s (input) and D_{st} index (output) with a sample period *T*=1hour. Based on this data set, a set of linear models with the form of (25) for different values of p and q were initially fitted and the two criteria SERR and FPE were calculated and are shown in Figure 13. The values of SERR and FPE clearly indicate that the model order is n_y=7 and n_u=2. Note that at least 96.53% of the system output variance can be explained by the ten variables {*const*, *y*(t-1), *y*(t-2), *y*(t-3), *y*(t-4), *y*(t-5), *y*(t-6), *y*(t-7), *u*(t-1), *u*(t-2)}.

The ten variables can be used to form a linear or nonlinear model for the data set. For example, by setting n_y=7 and n_u=2, a wavelet-based additive submodel decomposition of the NARX model can be expressed as below

$$\begin{aligned} y(t) &= f(y(t-1), y(t-2), \cdots, y(t-7), u(t-1), u(t-2)) + e(t) \\ &= \sum_{i=1}^{7} f_i(y(t-i)) + \sum_{i=1}^{2} f_{i+7}(u(t-i)) + \sum_{i=1}^{6}\sum_{j=i+1}^{7} f_{ij}(y(t-i), y(t-j)) \\ &+ \sum_{i=1}^{7}\sum_{j=1}^{2} f_{i(j+7)}(y(t-i), u(t-j)) + f_{89}(u(t-1), u(t-2)) + e(t) \end{aligned} \tag{39}$$

where f_i and f_{ij} are unknown univariate and bivariate functions which can be approximated by one- and two-dimensional wavelet decompositions. In this example, both the input and output data points were initially normalized and the modelling procedure was performed on the standard hypercube $[0,1]^n$, where $n=9$. The first 500 input-output data points were used for model identification and the remaining 500 data points were used for model testing. By expanding each f_i and f_{ij} using the wavelet series decompositions (the 4th-order B-sline scaling functions were used in each decomposition), model (39) can be converted into a linear-in-the-parameters problem and can be estimated using the forward OLS algorithm in Section 3. The final identified model, which involves 16 B-spline wavelet regressors selected from 891 candidate terms, is given as

$$y(t) = \sum_{i=1}^{16} \theta_i B_i(t) \qquad (40)$$

where $B_i(t)$ are wavelet regressors formed by the 4th-order B-spline scaling functions, and θ_i are estimated parameters. The terms, parameters and corresponding ERR values are listed in Table 8. Notice again that each variable in the model (39) and (40) was initially normalized to $[0,1]$, and the model outputs were recovered to the original system operating domain by taking inverse transforms. The six-step-ahead predictions and the model predicted outputs are shown in Figure 14.

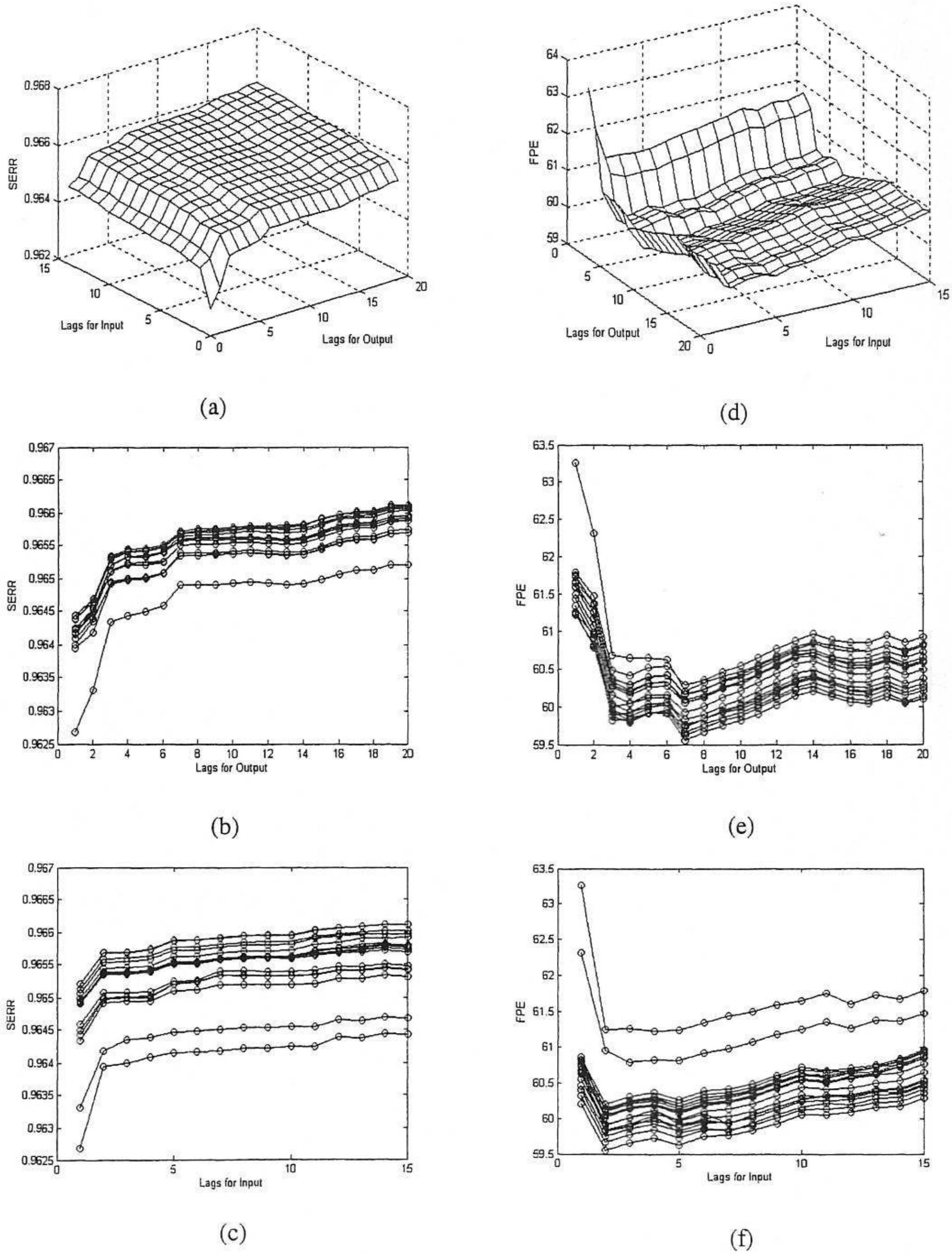


Figure13 The criteria SERR and FPE calculated from a set of linear models for the system in Example 6.
(a) SERR; (b) SERR vs n_y; (c) SERR vs n_u; (d) FPE; (e) FPE vs n_y; (f) FPE vs n_u.

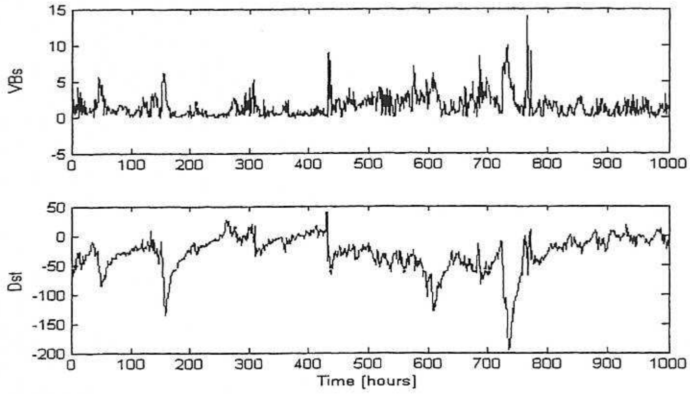


Figure 12 The input and output data of the terrestrial magnetospheric dynamic system in Example 6

Table 8 The selected model terms, estimated parameters and the corresponding ERR values for the model (40) in Example 6

Number	$B_i(t)$	θ_i	$ERR_i \times 100\%$
1	$\phi_{0,-1}(y(t-1))\phi_{0,-1}(y(t-2))$	1.01366e+000	99.74059
2	$\phi_{0,-1}(y(t-1))\phi_{0,-2}(u(t-1))$	9.65055e-001	0.08546
3	$\phi_{0,0}(y(t-1))\phi_{0,0}(y(t-4))$	5.31751e+000	0.05724
4	$\phi_{5,12}(u(t-2))$	-1.49386e-002	0.00578
5	$\phi_{0,0}(y(t-2))\phi_{0,-2}(u(t-2))$	-2.40523e+000	0.00434
6	$\phi_{5,27}(y(t-4))$	-8.47144e-003	0.00392
7	$\phi_{5,26}(y(t-2))$	3.69205e-003	0.00264
8	$\phi_{5,12}(u(t-1))$	-1.18424e-002	0.00167
9	$\phi_{0,0}(y(t-1))\phi_{0,0}(y(t-3))$	5.10037e+000	0.00146
10	$\phi_{0,-3}(u(t-1))\phi_{0,-1}(u(t-2))$	7.95453e-001	0.00138
11	$\phi_{0,0}(y(t-1))\phi_{0,0}(y(t-7))$	7.25787e-001	0.00131
12	$\phi_{5,10}(u(t-1))$	1.21153e-002	0.00127
13	$\phi_{5,23}(y(t-2))$	-3.92299e-003	0.00096
14	$\phi_{5,23}(y(t-5))$	3.42082e-003	0.00158
15	$\phi_{0,-3}(y(t-2))\phi_{0,0}(y(t-3))$	-4.03856e+001	0.00087
16	$\phi_{5,26}(y(t-6))$	3.27063e-003	0.00073
Note:	$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ — the 4th-order B-spline scaling functions		

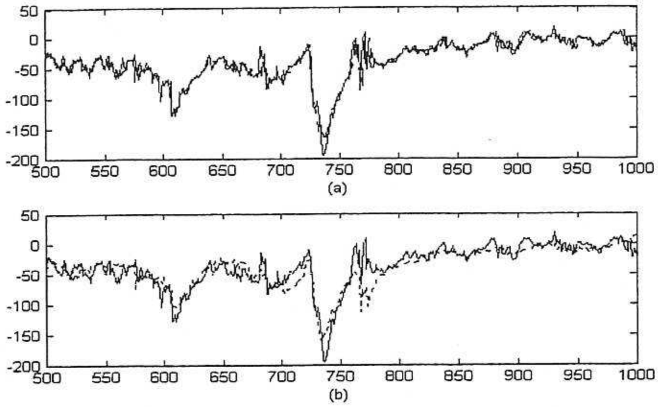


Figure 14 Comparisons of the 6-step-ahead predictions, model predicted outputs and the measurement for the solar wind Dst index in Example 6. (a) Six-step-ahead predictions; (b) Model predicted outputs. (Solid—measurements; Dashed—6-step-ahead predicted outputs; Dotted—model predicted outputs)

(D) Example 7—a multi-input and multi-output system

Consider a mechanical system with viscous damping and constant Coulomb friction, which was taken from the work of Chen and Tomlinson et al. (1996) and Ghanem and Romeo (2001). The motion equations of the system are given by

$$m_1\ddot{x}_1 + c_1\dot{x}_1 + \delta_1 sign(\dot{x}_1) + k_{11}x_1 + k_{12}x_2 = u_1(t) \tag{41a}$$

$$m_2\ddot{x}_2 + c_2\dot{x}_2 + \delta_2 sign(\dot{x}_2) + k_{21}x_1 + k_{22}x_2 = u_2(t) \tag{41b}$$

$$y_1(t) = x_1(t) + e_1(t) \tag{41c}$$

$$y_2(t) = x_2(t) + e_2(t) \tag{41d}$$

where $u_1(t) = A_1\sin(\omega_1 t)$ and $u_2(t) = A_2\sin(\omega_2 t)$ are the forces applied on the first and second mass, respectively; $y_1(t)$ and $y_2(t)$ are the outputs of the two subsystems, and $e_1(t)$ and $e_2(t)$ are additive noise uniformly distributed on $[-\varepsilon_0, \varepsilon_0]$, where $\varepsilon_0 = 10^{-4}$. The parameters were set to be $m_1 = m_2$ =1.0, $c_1 = c_2 = 10$, $\delta_1 = \delta_2$ =1.0, $k_{11} = k_{22}$ =20000, $k_{12} = k_{21}$=-10000.

The aim here was to identify a multi-input and multi-output ARMAX or NARMAX model for the system based on the inputs and outputs observed with additive noise. A general form of the time discrete model for a MIMO system with r inputs and m outputs is

$$\begin{aligned} y_i(t) = f_i[&y_1(t-1),\cdots,y_1(t-n_{y_1}^{(i)}), y_2(t-1),\cdots,y_2(t-n_{y_2}^{(i)}),\cdots\cdots, y_m(t-1),\cdots,y_m(t-n_{y_m}^{(i)}), \\ &u_1(t-1),\cdots,u_1(t-n_{u_1}^{(i)}), u_2(t-1),\cdots,u_2(t-n_{u_2}^{(i)}),\cdots\cdots,u_r(t-1),\cdots,u_r(t-n_{u_r}^{(i)})] \\ &i = 1,2,\cdots,m \end{aligned} \tag{42}$$

The system in Eq. (41) was simulated by setting A_1=5, A_2=16, $\omega_1 = 20\pi$, $\omega_2 = 16\pi$, and 1000 equi-spaced samples were obtained with a sampling interval of T=0.005 time units. Based on the first half (t=1,2,$\cdots$,

500) of the data set, the model order for the system (41) was initially determined using the approach in Section 4 and the model order was selected to be $n_{y_1}^{(1)}=2$, $n_{y_2}^{(1)}=2$, $n_{y_1}^{(2)}=2$, $n_{y_2}^{(2)}=2$, $n_{u_1}^{(1)}=2$, $n_{u_2}^{(1)}=0$, $n_{u_1}^{(2)}=0$, $n_{u_2}^{(2)}=2$. Then an ARX model was identified using the first 500 input-output data. The terms (variables) and the corresponding parameters are listed in Table 9. Based on this ARX model, the model predicted outputs were compared and are shown in Figure 15, which clearly indicates that the model predicts very well.

Table 9 The terms, parameters and the corresponding ERR values of the ARX model for the 2-input and 2-output system in Example 7

Subsystem 1			Subsystem 2		
Terms	Parameters	ERRs	Terms	Parameters	ERRs
y1(t-1)	0.15483E+01	0.92550E+00	u2(t-1)	0.77178E-05	0.93989E+00
y1(t-2)	-0.94950E+00	0.73012E-01	y1(t-1)	0.17761E+00	0.41358E-02
y2(t-1)	0.12665E+00	0.47887E-03	y1(t-2)	0.14436E-01	0.33789E-02
u1(t-1)	0.65726E-05	0.63781E-03	y2(t-1)	0.14947E+01	0.76123E-03
u1(t-2)	0.13099E-04	0.72448E-04	y2(t-2)	-0.91632E+00	0.26097E-02
y2(t-2)	0.59403E-01	0.22247E-04	u2(t-2)	0.13743E-04	0.55511E-04

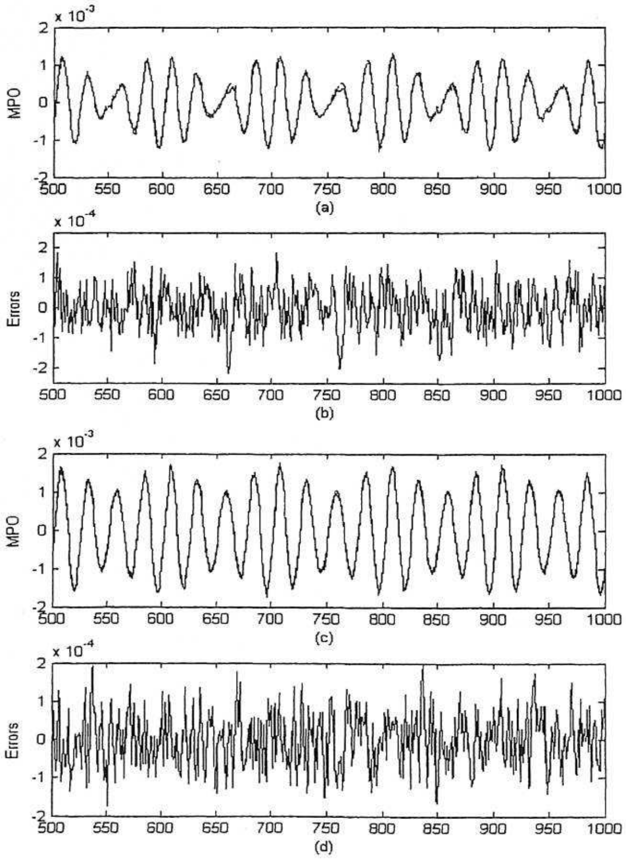


Figure 15 Comparisons of the model predicted outputs and the measurement for the 2-input and 2-output system in Example 6.
(a),(b) Subsystem 1; (c) ,(d) Subsystem 2
(Solid--measurement; Dashed--model-predicted outputs) (Input-output data with 1000 sampling points were used.)

6. Conclusions

The model term and variable selection problem has been considered for nonlinear system identification, and the application of the forward orthogonal least squares (OLS) algorithm to term detection has been described. A new variable selection algorithm based on calculating the sum of the error reduction ratios (SERR) and the final prediction errors (FPE) for local linear and cross-bilinear models has been proposed. It has been proved that bias which would normally affect approximate models does not affect the relative values of the ERR and this provides a robust variable selection procedure based on the new SERR criterion.

The new term detection procedure can be used to detect significant terms for all linear-in-the-parameter model structures including NARMAX models, neural networks and fuzzy logic based models and so has wide applicability.

Unlike operating region dependent local linear model based approaches, the variable selection procedure proposed here can find the significant variables which are not operating region dependent. These significant variables can then be used directly to construct nonlinear parametric and nonparametric model structures. The applicability and effectiveness of the new algorithm has been demonstrated using several numerical examples.

Appendix

The Effects of Noise on ERR Values

Assume that the actual output $y_a(t)$ of a given noise free system can be described by a deterministic linear regression model

$$y_a(t) = \sum_{i=1}^{M} \theta_i p_i(t) \, , \; t = 1,2,\cdots,N \qquad \text{(a1)}$$

where $y_a(t)$ is the dependent variable, $p_i(t)$ are regressors (predictors), θ_i are the unknown parameters to be estimated , N is the data length and M is the number of all the candidate regressors. For input-output systems, $p_i(t)$ (i=1,2,...,M) are usually formed from the lagged output variables $y_a(t-j)$ (j=1,2,..., n_y) and lagged input variables $u(t-k)$ (k=1,2,..., n_u) of the system under study, that is, $p_i(t) = p_i(x(t))$ with $x(t) = [y_a(t-1),\cdots, y_a(t-n_y), u(t-1),\cdots,u(t-n_u)]^T \in R^n$, and $n = n_y + n_u$. A compact matrix form corresponding to (a1) is

$$Y_a = P\Theta \qquad \text{(a2)}$$

where $Y_a = [y_a(1), y_a(2),\cdots, y_a(N)]^T$ is the actual observation vector, $P = [p_1, p_2,\cdots, p_M]$, $p_i = [p_i(1), p_i(2),\cdots, p_i(N)]^T$, $\Theta = [\theta_1, \theta_2,\cdots,\theta_M]^T$.

Some assumptions for the regression equation (a2) are now made as below:

(A1) *The regression matrix P is full rank in columns to guarantee that $P^T P$ is positive definite.*

(A2) *The system output $y_a(t)$ is corrupted by a noise sequence $\xi(t)$ which might be an intrinsic state noise or an additive output noise and which is uncorrelated with the regressors $p_i(t)$, (i=1,2,..., M).*

(A3) *Among all the M regressors* $\{p_1, p_2, \cdots, p_M\}$, *only K (≤M) regressors, say* $\{p_1, p_2, \cdots, p_K\}$ *are related to the lagged output variables* $y_a(t-1)$, $y_a(t-2)$,..., $y_a(t-n_y)$ *and are contaminated by the noise* $\xi(t)$.

(A4) *There exists some modelling error* $\zeta(t)$ *when using the linear regression equation (a1).*

Denote $y(t)$ as the noisy dependent variable corresponding to $y_a(t)$ in the regression equation (a1), and $Y = [y(1), y(2), \cdots, y(N)]^T$ the noisy measurement vector. Based on assumptions (*A*2)-(*A*4), the regression equation for the noisy dependent variable $y(t)$ can be described as

$$\begin{aligned} y(t) &= \sum_{i=1}^{K} \theta_i p_i(x(t);\xi(t)) + \sum_{i=K+1}^{M} \theta_i p_i(x(t)) + \zeta(t) \\ &= \sum_{i=1}^{K} \theta_i [p_i(t) + \xi_i(t)] + \sum_{i=K+1}^{M} \theta_i p_i(t) + \zeta(t) \\ &= \sum_{i=1}^{M} \theta_i p_i(t) + \sum_{i=1}^{K} \theta_i \xi_i(t) + \zeta(t) \\ &= \sum_{i=1}^{M} \theta_i p_i(t) + \eta(t), \quad t = 1,2,\cdots,N \end{aligned} \tag{a3}$$

where $\xi_i(t)$ are the noise effects on the regressors $p_i(t)$ (*i*=1,2,...,*K*) introduced by the noise $\xi(t)$, $\zeta(t)$ is some modelling error which is uncorrelated with $p_i(t)$ (*i*=1,2,...,*M*) , and $\eta(t) = \sum_{i=1}^{K} \theta_i \xi_i(t) + \zeta(t)$ is a coloured noise which includes the effects of measurement noise, unmeasured disturbances and modelling errors, such as unmodelled nonlinear effects.

Before analysing the noise effects on the ERR values, another assumption is added and stated as below:

(A5) *The derivative noise effects on the regressors* $p_i(t)$ (*i*=1,2,...,*K*) *introduced by* $\xi_i(t)$ *are uncorrelated with the regression matrix P.*

A matrix form for (a3) is

$$Y = P\Theta + \Pi \tag{a4}$$

with $\Pi = [\eta(1), \eta(2), \cdots, \eta(N)]^T$.

Guaranteed by the assumption (*A*1), the regression matrix can be orthogonally decomposed as

$$P = WA \tag{a5}$$

where A is an $M \times M$ unit upper triangular matrix and W is an $N \times M$ matrix with orthogonal columns $w_1, w_2, \cdots, w_M$ in the sense that $W^T W = D = diag[d_1, d_2, \cdots, d_M]$ with $d_i = <w_i, w_i> = \sum_{t=1}^{N} w_i(t) w_i(t)$, and the symbol $<\cdot,\cdot>$ denotes the inner product of two vectors. The space spanned by the orthogonal basis $w_1, w_2, \cdots, w_M$ is the same as that spanned by the basis set $p_1, p_2, \cdots, p_M$, and (a5) can be expressed as

$$Y = (PA^{-1})(A\Theta) + \Pi = WG + \Pi \tag{a6}$$

where $G = [g_1, g_2, \cdots, g_M]^T$ is an auxiliary parameter vector. From the assumptions (*A*2) and (*A*5), the noisy output variance can be expressed as

$$\begin{aligned} Y^T Y &= G^T DG + 2(WG)^T \Pi + \Pi^T \Pi \\ &= G^T DG + \Pi^T \Pi + 2(P\Theta)^T \Pi \\ &= G^T DG + \Pi^T \Pi + 2\Theta^T (P^T \Pi) \\ &= G^T DG + \Pi^T \Pi \\ &= Y_a^T Y_a^T + \Pi^T \Pi \end{aligned} \tag{a7}$$

or

$$\sigma_y^2 + \mu_y^2 = \sigma_{y_a}^2 + \sigma_\eta^2 + \mu_{y_a}^2 + \mu_\eta^2 \tag{a8}$$

where μ_y, μ_{y_a} and μ_η are the means of the output measurements $y(t)$, the actual system output $y_a(t)$ and the coloured noise $\eta(t)$; $\sigma_y^2, \sigma_{y_a}^2$ and σ_η^2 are the standard variations of $y(t)$, $y_a(t)$ and $\eta(t)$, respectively. Therefore, the error reduction ratio effected by a noise $\eta(t)$, $ERR(i)$, introduced by w_i, can be calculated by

$$\begin{aligned} ERR(i) &= \frac{g_i^2 < w_i, w_i >}{< Y, Y >} = \frac{< Y, w_i >^2}{< Y, Y >< w_i, w_i >} \\ &= \frac{< Y_a + \Pi, w_i >^2}{< Y_a + \Pi, Y_a + \Pi >< w_i, w_i >} \\ &= \frac{< Y_a, w_i >^2}{[< Y_a, Y_a > + < \Pi, \Pi >] < w_i, w_i >} \\ &= \frac{< Y_a, Y_a >}{< Y_a, Y_a > + < \Pi, \Pi >} \frac{< Y_a, w_i >^2}{< Y_a, Y_a >< w_i, w_i >} \\ &= \frac{\sigma_{y_a}^2 + \mu_{y_a}^2}{\sigma_{y_a}^2 + \sigma_\eta^2 + \mu_{y_a}^2 + \mu_\eta^2} \frac{< Y_a, w_i >^2}{< Y_a, Y_a >< w_i, w_i >} \\ &= \lambda \, ERR_a(i) \\ &\le ERR_a(i) , \quad i = 1, 2, \cdots, M , \end{aligned} \tag{a9}$$

where

$$\lambda = \frac{\sigma_{y_a}^2 + \mu_{y_a}^2}{\sigma_{y_a}^2 + \sigma_\eta^2 + \mu_{y_a}^2 + \mu_\eta^2} = \frac{\sigma_{y_a}^2 + \mu_{y_a}^2}{\sigma_y^2 + \mu_y^2} \le 1 \tag{a10}$$

Eq. (a9) clearly indicates that the coloured noise $\eta(t)$ makes the values of ERR become smaller compared with the actual values ERR_a. This implies that the contribution made by the ith regressor to the system measurement Y may appear to be slightly less significant than would be the case if Y_a were used and the data were noise free. Notice, however, that the correct terms (regressors) can still be selected despite the coloured noise $\eta(t)$, and just as important the order of the selected terms will be the same as in the noise free case. The only difference will be that terms will be selected with slightly smaller ERR values because of the effects of the noise. Clearly, if $\sigma_\eta^2 + \mu_\eta^2 \approx 0$ or $\sigma_\eta^2 + \mu_\eta^2 << \sigma_{y_a}^2 + \mu_{y_a}^2$, then $\lambda \approx 1$, this means that the values of the ERR will be affected to a neglected degree by the noise $\eta(t)$.

The above analysis shows that the relative values of the ERR are always preserved, which means that it should still be possible to select the significant model terms even with coloured noise. However, the existence of noise, especially coloured noise with a large variance, will affect the choice of the threshold value mentioned in *Remark* 3.

Acknowledgment

The authors gratefully acknowledge that part of this work was supported by EPSRC. We are grateful to Dr M. Balikhin for providing the magnetosphere data.

References

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

Battiti, R.(1994), Using mutual information for selecting features in supervised neural net learning, *IEEE Trans. Neural Networks*, **5(4)**,537-550.

Billings,S.A. and Voon, W.S.F.(1987), Piecewise linear identification of nonlinear systems, *International Journal of Control*, **46(1)**,215-235.

Billings, S.A., Korenberg, M. and Chen, S.(1988), Identification of nonlinear output-affine systems using an orthogonal least-squares algorithm, *Int. Journal of Systems Sci*, **19(8)**,1559-1568.

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

Billings,S.A., and Zhu, Q.M..(1994), A structure detection algorithm for nonlinear rational models, *International Journal of Control*, **59(6)**,1439-1463.

Chen, Q. and Tomlinson, G.R. (1996), Parametric identification of systems with dry friction and nonlinear stiffness a time series model, *Journal of Vibration and Acoustics-Transactions of the ASME*, **118 (2)**, 252-263.

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

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

Chng, E.S., Yang, H.H. and Bos, S.(1996), Orthogonal least-squares learning algorithm with local adaptation process for the radial basis function networks, *IEEE Signal Processing Letters*, **3 (8)**, 253-255

Friedman,J.H. and Stuetzle, W., 1981, Projection pursuit regression, *Journal of the American Statistical Association*, **76(376)**, 817-823.

George, E.I. and McCulloch, R.E. (1993), Variable selection via Gibbs Sampling, *Journal of the American Statistical Association*, **889(423)**, 881-889.

Ghanem, R. and Romeo, F. (2001), A wavelet-based approach for model and parameter identification of non-linear systems, *International Journal of Non-Linear Mechanics*, **36(5)**, 835-859.

Gomm, J.B. and Yu, D.L.(2000), Order and delay selection for neural network modelling by identification of linearized models, *International Journal of Systems Science*, **31(10)**, 1273-1283.

Hong, X. and Harris, C. J.(2001a), Variable selection algorithm for the construction of MIMO operating point dependent neurofuzzy networks, *IEEE Transactions On Fuzzy Systems*, **9(1)**, 88-101.

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

Korenberg, M., Billings, S.A., Liu, Y. P. and McIlroy P.J.(1988), Orthogonal parameter estimation algorithm for non-linear stochastic systems, *International Journal of Control*, **48(1)**,193-210.

Leontaritis,I.J., Billings,S.A.(1985), Input-output parametric models foe non-linear systems, part I: deterministic non-linear systems; part II: stochastic non-linear systems, *International Journal of Control*, **41(2)**,303-344.

Mao, K.Z. and Billings, S.A.(1999), Variable selection in nonlinear systems modelling, *Mechanical Systems and Signal Processing*, **13(2)**,351-366.

Miller, A.J.(1990), *Subset selection in regression*. London: Chapman and Hall.

Oja, E.(1992), Principle components, minor components and linear neural networks, *Neural Networks*, **5(6)**,927-935.

Roy, E., Stewart, R.W., and Durrani, T.S.(1996), High-order system identification with an adaptive recursive second-order polynomial filter, *IEEE Signal Processing Letters*, **3(10)**, 276-279.

Savit, R. and Green M. (1991), Time series and dependent variables, *Physica **D 50***, 95-116.

Sjoberg, J., Zhang, Q.H., Ljung,L., Benveniste, A., Delyon, B., Glorennec, R.Y., Hjalmarsson,H., and Juditsky, A.(1995), Nonlinear black-box modelling in system identification: a unified overview, *Automatica*, **31(12)**, 1691-1724.

Wang, L.X. and Mendel, J.M.(1992), Fuzzy basis functions, universal approximations, and orthogonal least squares learning, *IEEE Trans Neural Networks*, **3(5)**,807-814.

Zheng, G.L. and Billings, S.A.(1995), Radial basis function networks configuration using mutual information and the orthogonal least squares algorithm, *Neural Networks*, **9(9)**, 1619-1637.

Zhu, Q.M. and Billings, S.A.(1996), Fast orthogonal identification of nonlinear stochastic models and radial basis function neural networks, *International Journal of Control*, **64(5)**,871-886.