Published paper
Wei, H.L. and Billings, S.A. (2008) *Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information.* International Journal of Modelling Identification and Control, 3 (4). pp. 341-356.
http://dx.doi.org/10.1504/IJMIC.2008.020543

A prepublication draft of the paper published in
International Journal of Modelling, Identification and Control, Vol. 3, No. 4, pp.341-356, 2008.

Model Structure Selection Using an Integrated Forward Orthogonal Search Algorithm Assisted by Squared Correlation and Mutual Information

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@shef.ac.uk, w.hualiang@shef.ac.uk

Abstract: Model structure selection plays a key role in nonlinear system identification. The first step in nonlinear system identification is to determine which model terms should be included in the model. Once significant model terms have been determined, a model selection criterion can then be applied to select a suitable model subset. The well known orthogonal least squares type algorithms are one of the most efficient and commonly used techniques for model structure selection. However, it has been observed that the orthogonal least squares type algorithms may occasionally select incorrect model terms or yield a redundant model subset in the presence of particular noise structures or input signals. A very efficient integrated forward orthogonal search (IFOS) algorithm, which is assisted by the squared correlation and mutual information, and which incorporates a generalised cross-validation (GCV) criterion and hypothesis tests, is introduced to overcome these limitations in model structure selection.

Keywords: correlation, hypothesis tests, identification, model selection, mutual information, NARX/NARMAX model.

Model Structure Selection Using an Integrated Forward Orthogonal Search Algorithm Assisted by Squared Correlation and Mutual Information

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

Abstract: Model structure selection plays a key role in nonlinear system identification. The first step in nonlinear system identification is to determine which model terms should be included in the model. Once significant model terms have been determined, a model selection criterion can then be applied to select a suitable model subset. The well known orthogonal least squares type algorithms are one of the most efficient and commonly used techniques for model structure selection. However, it has been observed that the orthogonal least squares type algorithms may occasionally select incorrect model terms or yield a redundant model subset in the presence of particular noise structures or input signals. A very efficient integrated forward orthogonal search (IFOS) algorithm, which is assisted by the squared correlation and mutual information, and which incorporates a generalised cross-validation (GCV) criterion and hypothesis tests, is introduced to overcome these limitations in model structure selection.
Keywords: correlation, hypothesis tests, identification, model selection, mutual information, NARX/NARMAX model.

1. Introduction

Model structure selection is the central task in nonlinear system identification. This topic, which accompanies the development of system identification techniques, has been extensively studied in the literature. In a broader sense, model structure selection is closely related to many practical themes including data fitting, time series prediction, feature selection in classification, and complexity reduction in neural networks. The conventional Akaike information criterion (AIC) (Akaike 1974), the Bayesian information criterion (BIC) (Schwarz 1978), the minimum description length (MDL) (Rissanen 1978), the generalized cross-validation (GCV) (Golub et al. 1979), and many variants (Stoica et al. 1986, Miller 1990, Haber and Unbehauen 1990, Stoica and Selen 2004) have been proposed to determine the number of variables or regressors in the model, and this is often termed as model selection or model order determination. Both parametric and nonparametric techniques have been developed for variable selection (Hocking 1976, 1983, Breiman and Freedman 1983, Tjostheim and Auestad 1994, Breiman 1995, Vieu 1995, Rech et al. 2001, Huang and Yang 2004). Statistical methods, for example, conditional probability analysis (Savit and Green 1991) and hypothesis tests (Montgomery et al. 2001, Stark and Fitzgerald 1995, Anders and Korn 1999, Lind and Ljung 2005) have been studied for variable or regressor selection for some specific model structures. In network modeling, mutual information (Battiti 1994, Zheng and Billings 1996), genetic algorithms (Mao and Billings 1997, Madar et al. 2005), and robust regression and optimization methods (Hong and Harris

2002, 2002, Chen et al. 2003, Hong and Chen 2005), have been introduced for network training. In order to increase the robustness of a selected model for effectively handling ill-imposed problems or to avoid overfitting, regularisation methods have been introduced to complete the model structure detection procedure (Sjoberg and Ljung 1995, Orr 1995, 1998, Chen et al. 1996, Billings and Chen 1998). Some quantitative validation methods have also been proposed to measure model performance and dynamic signatures (Aguirre and Billings 1995c, Haynes and Billings 1994, Zheng and Billings 199, 2002).

In nonlinear system identification and function (signal) approximation, model structure selection can involve a large number of candidate model terms or basis functions. The first key step is to determine which terms or bases are significant and should be included in the model. There exist some situations where the static nonlinearity of the system (and the eigenvalue function) can indicate which term clusters are required. Of course, if the static nonlinearity is not known a priori, then the terms should be chosen in a purely black-box fashion, as proposed in this paper. Because the main justification of this paper is to introduce a new integrated method that improves on the error reduction ratio (ERR) based algorithm, when the latter fails (for example, when there is “missing information” due to “poor input signals” or excessive noise) it would be interesting to investigate the potential use of alternative sources of information, whenever available. A general discussion on some possible alternative approaches can be found in Aguirre et al. (2000, 2002, 2004).

It is known that inclusion of insignificant or redundant model terms might result in a much more complex model, involving a large number of parameters, and as a consequence the model may become oversensitive to training data and is likely to exhibit poor generalisation properties. For example, a redundant or overfitted model may lack a satisfactory long term predictive capability. One of the main tasks in nonlinear system identification therefore is to select a parsimonious model structure. Ideally, this requires that the resulting model structure is optimal or at least suboptimal with regard to specified modelling goals. Several approaches have been proposed to address this problem (Korenberg et al. 1988, Billings et al. 1988, Haber and Unbehauen 1990, Miller 1990, Mallat and Zhang 1993, David et al. 1994). One of the most efficient and popular model structure detection techniques are the class of orthogonal least squares (OLS) type algorithms (Korenberg et al. 1988, Billings et al. 1989, Chen et al. 1989), which have been widely applied in nonlinear system identification. The OLS type algorithms have a desirable advantage: the contributions of candidate model terms can be decoupled and decomposed, and as a consequence the significance of each candidate model term can be measured using the associated error reduction ratio (ERR). Significant model terms can thus be ranked according to the order that model terms are selected one at a time. The rank of selected model terms is independent of the positions in which the candidate model terms appear in the regression equation since at each step a significant model term is determined when the significance of all remaining candidate model terms have been evaluated (Wei et al. 2004). The incorporation of the OLS-ERR type algorithms with other modelling techniques has greatly raised the capability of improving the generalisation properties of the resulting models, see for example, Aguirre and Billings (1994, 1995a, 1995b), Billings, Chen and Backhouse (1989), Zhu and Billings (1996), Chen et al. (2003, 2005), Billings and Wei (2005a, 2005b), and Zhu et al. (2007).

It has been observed that the OLS-ERR type algorithms may occasionally select incorrect model terms or yield a redundant model subset when either the training data are contaminated by certain noise sequences (Mao and Billings 1997), or the input is poorly designed, for example a second order low frequency autoregressive process (Piroddi and Spinelli 2003). These are generic problems in nonlinear system identification and any algorithm may fail to produce correct models in these worse case scenarios. As will be seen later, however, the problems related to these cases can be avoided or alleviated by inspecting and comparing the performance of a few models produced from some trial-and-error tests. Piroddi and Spinelli (2003) proposed a promising approach to solve the model structure selection problem by minimizing the simulation error, which is defined as the discrepancy between the model predicted outputs and the measurements. However, the method of Piroddi and Spinelli requires calculating model predicted outputs for all candidate model term combinations and is thus very time demanding. Mao and Billings (1997) proposed a solution to the combined problem of model structure selection and parameter estimation by introducing a genetic searching algorithm, combined with the standard orthogonal least squares routine. Although this requires much less calculations compared with an optimal exhaustive search, the necessary computations are still quite large. In the present study, a much simpler but efficient approach, which is easier to implement and quicker to compute, for general nonlinear model structure selection, is proposed to solve the problem addressed in Piroddi and Spinelli (2003) and in Mao and Billings (1997).

This study focuses on the model structure selection problem in nonlinear dynamical system identification including model term detection and model subset selection. The main contributions of the work include: i) a new criterion for measuring the significance of model terms is introduced based on mutual information; the mutual information criterion can be used as a complementary approach or as an alternative to the ERR criterion; ii) a simple hypothesis test, based on the *t*-test, is incorporated into the new orthogonal forward search algorithm; for linear-in-the-parameters models, this kind of *t*-test provides an index to indicate which model terms are significant; iii) a new approach is proposed for selecting an accurate model subset for a given identification problem. The squared correlation and mutual information criteria, along with the *t*-tests and a general cross-validation (GCV) criterion, are all incorporated into the new forward orthogonal search algorithm. For convenience, the new *i*ntegrated *f*orward *o*rthogonal *s*earch algorithm assisted by squared correlation and mutual information will be referred to as the IFOS algorithm. The *t*-test has implicitly been applied in the OLS-ERR algorithm to aid the selection of significant mode terms in NARMAX modelling (Mendes and Billings 1993). In the present study, the *t*-test is incorporated into the IFOS algorithm to explicitly demonstrate how this test works for NARX modelling.

The remainder of the paper is organised as follows. In section 2 the orthogonal forward regression (OLS) algorithm is briefly reviewed and the performance of this algorithm is discussed and analysed. In section 3, the new integrated forward orthogonal search (IFOS) algorithm assisted by mutual information is proposed. Four examples are described in section 4 to demonstrate the effectiveness and applicability of the new IFOS algorithm. Some suggestions and discussions are included in section 5, and finally the work is concluded in section 6.

2. The OLS-ERR algorithm

In the following the discussion is restricted to models that can be expressed in a linear-in-the-parameters form. This is an important class of representations for nonlinear system identification and signal processing. Compared to nonlinear-in-the-parameters models, linear-in-the-parameters models are simpler to analyse mathematically and quicker to compute numerically. The polynomial NARX model will be used as an example to demonstrate the OLS-ERR algorithm. For the sake of convenience in the descriptions, the two terms 'system' and 'model' will not be strictly distinguished but the meanings of the two terms should be self-evident from the context.

2.1 The NARX model

The general form of the NARMAX (N*onlinear* A*uto*R*egressive* M*oving* A*verage* with *e*X*ogenous* inputs) model (Leontaritis and Billings 1985, Billings and Chen 1998, Pearson 1999, Piroddi and Spinelli 2003) takes the form of the following nonlinear recursive difference equation:

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

where f is some unknown nonlinear mapping, $u(t)$, $y(t)$ and $e(t)$ are the input, output, and the prediction error, n_u, n_y and n_e are the associated maximum lags. If the function f is specified as a polynomial function, model (1) can then be decomposed into a process related part and a noise related part as

$$y(t) = f^p(\varphi^p(t)) + f^n(\varphi^n(t)) + e(t) \tag{2}$$

where $\varphi^p(t) = [y(t-1),\cdots,\ y(t-n_y),\ u(t-1),\cdots,\ u(t-n_u)]^T$ is the process regressor vector, and $\varphi^n(t) = [y(t-1),\cdots,\ y(t-n_y),\ u(t-1),\cdots,\ u(t-n_u),e(t-1),\cdots,e(t-n_e)]^T$ is the extended regressor vector. The polynomial NARX (N*onlinear* A*uto*R*egressive* with *e*X*ogenous* inputs) model is a special case of the polynomial NARMAX model, where the noise related model f^n reduces to a single noise term $e(t)$ that can often be treated as an independent identical distributed (iid) zero mean noise sequence providing that the function f^p gives a sufficient description of the data set.

The polynomial NARX model can be expressed using a linear-in-the-parameters form

$$y(t) = \sum_{m=1}^{M} \theta_m \phi_m(t) + e(t) \tag{3}$$

where $\phi_m(t)=\phi_m(\varphi(t))$ are model terms generated in some way from the regressor vector $\varphi(t)$ $=[y(t-1),\cdots, y(t-n_y), u(t-1),\cdots, u(t-n_u)]^T$, θ_m are unknown parameters, and M is the total number of potential model terms. Clearly, the candidate model terms $\phi_m(t)$ are of the form $x_1^{i_1}(t)\cdots x_\ell^{i_\ell}(t)$, where $x_j^{i_j}(t)\in\{y(t-1),\cdots, y(t-n_y), u(t-1),\cdots, u(t-n_u)\}$ for j=1, 2, …, ℓ, with $0\le i_j\le \ell$ and $0\le i_1+\cdots+i_\ell\le\ell$. The maximum lag 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. Several algorithms are available for the determination of the maximum lags for both the input and the output (Bomberger and Seborg 1998, Feil et al. 2004, Wei et al. 2004).

2.2 The OLS-ERR algorithm

Consider the term selection problem for the linear-in-the-parameters model (3). Let $\mathbf{y}=[y(1),\cdots,y(N)]^T$ be a vector of measured outputs at N time instants, and $\boldsymbol{\varphi}_m=[\phi_m(1),\cdots,\phi_m(N)]^T$ be a vector formed by the mth candidate model term, where m=1,2, …, M. Let $\mathfrak{D}=\{\boldsymbol{\varphi}_1,\cdots,\boldsymbol{\varphi}_M\}$ be a dictionary composed of the M candidate bases. From the viewpoint of practical modelling and identification, the finite dimensional set $\mathfrak{D}$ is often redundant. The model term selection problem is equivalent to finding a full dimensional subset $\mathfrak{D}_n=\{\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_n\}=\{\boldsymbol{\varphi}_{i_1},\cdots,\boldsymbol{\varphi}_{i_n}\}$ of n ($n\le M$) bases, from the library $\mathfrak{D}$, where $\boldsymbol{\alpha}_k=\boldsymbol{\varphi}_{i_k}$, $i_k\in\{1,2,\cdots,M\}$ and k=1,2, …, n, so that $\mathbf{y}$ can be satisfactorily approximated using a linear combination of $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n$ as below

$$\mathbf{y}=\theta_1\boldsymbol{\alpha}_1+\cdots+\theta_n\boldsymbol{\alpha}_n+\mathbf{e} \tag{4}$$

or in a compact matrix form

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

where the matrix $\mathbf{A}=[\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_n]$ is assumed to be of full column rank, $\boldsymbol{\theta}=[\theta_1,\cdots,\theta_n]^T$ is a parameter vector, and $\mathbf{e}$ is the approximation error.

The model structure selection procedure starts from equation (3), with $\mathfrak{D}=\{\boldsymbol{\varphi}_1,\cdots,\boldsymbol{\varphi}_M\}$. For j=1,2,…, M, define

$$\mathrm{ERR}^{(1)}[j]=\frac{(\mathbf{y}^T\boldsymbol{\varphi}_j)^2}{(\mathbf{y}^T\mathbf{y})(\boldsymbol{\varphi}_j^T\boldsymbol{\varphi}_j)} \tag{6}$$

$$\ell_1=\arg\max_{1\le j\le M}\{\mathrm{ERR}^{(1)}[j]\} \tag{7}$$

The first significant basis can then be selected as $\boldsymbol{\alpha}_1=\boldsymbol{\varphi}_{\ell_1}$, and the first associated orthogonal variable can be chosen as $\mathbf{q}_1=\boldsymbol{\varphi}_{\ell_1}$.

Assume that a subset $\mathfrak{D}_{m-1}$, consisting of (m-1) significant bases, $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_{m-1}$, has been determined at step (m-1), and the (m-1) selected bases have been transformed into a new group of orthogonalized bases $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{m-1}$ via some orthogonal transformation. To select the mth significant basis $\boldsymbol{\alpha}_m$, let

$$\mathbf{q}_j^{(m)} = \boldsymbol{\varphi}_j - \sum_{k=1}^{m-1} \frac{\boldsymbol{\varphi}_j^T \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{q}_k} \mathbf{q}_k \tag{8}$$

$$\mathrm{ERR}^{(m)}[j] = \frac{(\mathbf{y}^T \mathbf{q}_j^{(m)})^2}{(\mathbf{y}^T \mathbf{y})[(\mathbf{q}_j^{(m)})^T \mathbf{q}_j^{(m)}]} \tag{9}$$

where $\boldsymbol{\varphi}_j \in \mathfrak{D} - \mathfrak{D}_{m-1}$. The mth significant basis can then be chosen as $\boldsymbol{\alpha}_m = \boldsymbol{\varphi}_{\ell_m}$ and the mth associated orthogonal basis can be chosen as $\mathbf{q}_m = \mathbf{q}_{\ell_m}^{(m)}$, where $\ell_m = \arg \max_{1 \le j \le M} \{\mathrm{ERR}^{(m)}[j]\}$. Subsequent significant bases can be selected in the same way step by step. At each step, the 'best' basis with the strongest capability to represent the output $\mathbf{y}$ is selected. The selection procedure can be terminated when some specified termination conditions are met.

The indices $\mathrm{ERR}^{(m)}[j]$ are referred to as the *error reduction ratios* (ERR), and provide a simple but effective means of selecting a subset of significant regressors. A more detailed explanation of ERR can be found in Billings et al. (1989) and Chen et al. (1989).

Note that in many cases the noise signal $e(t)$ in Eq. (3) may be a correlated or a coloured noise sequence. This is likely to be the case for most real data sets. The NARX model (3) will then become the NARMAX model. For the NARMAX model, the structure selection procedure starts from identifying the process NARX model, and the noise model can then be identified in the same way as selecting the NARX model structure (Billings and Chen 1998). The inclusion of noise terms is mainly used to reduce the bias in the parameters of the process NARX model.

2.3 The performance of the OLS-ERR algorithm

The OLS-ERR algorithm has been widely applied in model structure selection for nonlinear system identification (Billings and Chen 1998) and has already become a standard algorithm for nonlinear function approximation and neural network training (Haykin 1999, Nelles 2001, Harris et al. 2002). It has been observed, however, that this algorithm has some deficiencies when it is applied in some worse case situations, where there are some uncertainties in the data or the input signal is not very persistently exciting (Mao and Billings 1997, Piroddi and Spinelli 2003).

It has been observed that for some specific input signals, the model term $y(t\text{-}1)$ is nearly always selected as the first term with a very high ERR value, and as a consequence the contributions of other model terms, measured by the associated ERR values, become small and are sensitive to the effect of noise (Piroddi and Spinelli 2003). This problem arises largely because of the characteristics of the input: a low order, low frequency autoregressive (AR) process, though persistently exciting (of any finite order), by the standard definition for linear system identification (Ljung 1987, Söderström and Stoica 1989), may not be sufficient for nonlinear model identification. In fact, as noted in Piroddi and

Spinelli (2003), such a low frequency AR process often yields a slowly varying output signal. Assuming that the output signal, denoted by $y(t)$, is sampled at a fast sampling rate (oversampled), the signal $y(t)$ and the first few linear terms, $y(t\text{-}1)$, $y(t\text{-}2)$, …, will then become strongly correlated and thus indistinguishable, implying that $y(t) \approx y(t-1)$. This results in $\mathrm{ERR}(\mathbf{y}, \mathbf{y}_1) \approx 1$, where $\mathbf{y}$ and $\mathbf{y}_1$ are vectors formed by the output variable $y(t)$ and the term $y(t\text{-}1)$. Consequently, the term $y(t\text{-}1)$ is nearly always selected as the first term, regardless of whether the term $y(t\text{-}1)$ exists in the true model. The implication is that no matter what identification algorithms are employed, only the information on the actual system that is contained in the data can be extracted. Some algorithms may be capable of extracting more information from the data, in some situations, compared with other algorithms. For small sampling times the terms of a same term-cluster become indistinguishable, and hence for a practical identification problems the sampling time should not be chosen to be too small (Billings and Aguirre 1995). This is likely to be true for all identification models and algorithms.

Noise may also affect the model structure selection even when the training data are sampled with an appropriate sampling rate. While all correct model terms ('correct term' here means that the term exists in the original real model) can often be detected and included in the identified model, some 'unnecessary' (incorrect) model terms that do not exist in the original model may occasionally enter into the selected model subset before some correct model terms. In most cases, nonlinear identification is a structure-unknown problem. Almost all existing model structure selection algorithms are thus data-oriented, that is, any algorithm will try to find a model structure that reflects as closely as possible the information carried by the observed noisy data (it is assumed that the data cannot be cleaned by filtering), without any knowledge of the true model structure. Since realistically models must be learned from noise-contaminated data, spurious terms (incorrect terms) may also be included in the identified model subset. However, a good model structure selection algorithm should be able to provide a good model structure that minimizes the effects of incorrect (spurious) model terms to a negligible level, such that the main underlying dynamics embodied in the data can be revealed or captured by the identified model. Model validation can also be used to assist in model construction and is therefore an important part of all narmax modelling procedures. Validation provides an independent assessment of the final identified model.

It is an inherent problem that data uncertainty, sampling rates and the types of input signals will affect the selection of model structure in a nonlinear system identification task. The development of methods that can mitigate these effects will be highly desirable.

2.4 Two examples

Two simple examples will be used to illustrate some of the problems that arise if the training data are contaminated by noise, or if the input is not sufficiently exciting. The two artificial examples are given below:

Model I: $\mathrm{y(t)}=-1.7y(t\text{-}1)-0.8y(t\text{-}2)+u(t\text{-}1)+0.8u(t\text{-}2)+e(t)$ (10)

Model II: $y(t) = 0.7y(t-1) - 0.1y(t-2) + u(t-1)$ (11)

The input $u(t)$ in Model I is uniformly distributed on [−2,2], with the noise $e(t) \sim N(0, 0.1^2)$. The input $u(t)$ in Model II is a low frequency AR(2) process of the form: $u(t)$=1.6$u(t$-1)− 0.6375$u(t$-2)+$\xi(t)$, with $\xi(t) \sim N(0,1)$. Note that although the AR(2) process is persistently exciting of almost any finite order, it is a narrow band process behaving like a lowpass filter with minimum attenuation of low frequencies near ω=0, with sharply increasing attenuation as ω increases toward $\omega = \pi$. This kind of AR processes may not be sufficiently exciting for ARX and NARX model structure selection (Leontaritis and Billings 1987).

One thousand input-output data points were generated from Model I. The candidate model terms were set to be y(t-k) and u(t-k) where k=1,2,3,4,5. By applying the OLS-ERR algorithm to the given 10 candidate model terms, a model of 8 terms was produced as shown in Table 1, where the model terms are ranked according to the order in which they were selected. It can be seen from Table 1 that even though all the correct model terms were selected, the resulting model structure is not the minimum or correct structure. The structure is a redundant model structure due to the inclusion of some incorrect model terms. As will be seen later, all the incorrectly selected model terms can however easily be eliminated by introducing a simple t-statistic.

Table 1 Model selection results for Model I using the OLS-ERR algorithm

Term	Parameter		ERR(%)
	True	Estimate	
y(t-1)	-1.7	-1.704552	67.4213
u(t-1)	1.0	1.000453	28.0911
y(t-4)	0	-0.007688	2.9753
u(t-4)	0	0.008823	0.5170
y(t-3)	0	-0.020076	0.4823
u(t-3)	0	0.011086	0.1250
u(t-2)	0.8	0.801407	0.1524
y(t-2)	-0.8	-0.815569	0.0342

Table 2 Model selection results for Model II using the OLS-ERR algorithm

Selected model structure	Number of times selected out of 100
y(t)= 0.39y(t-2)-0.07y(t-3)+u(t-1)+ 0.7u(t-2)	35
y(t)= 0.557143y(t-1)-0.014286y(t-3)+u(t-1)+0.142857u(t-2)	31
y(t)= 0.5205y(t-1)-0.00256y(t-3)+u(t-1)+0.1795u(t-2)+0.02564u(t-3)	18
y(t)= 0.7y(t-1)-0.1y(t-2)+u(t-1)	16

Model II was simulated 100 times and at each time 1000 input-output data points were recorded. By setting the candidate model terms to be the same as in Model I, the OLS-ERR algorithm, coupled with the GCV criterion, was applied over the 100 data sets respectively, and the model selection results are illustrated in Table 2. From Table 2, it can be seen that the true model structure was only correctly selected 16 times out of a 100 when the input signal was chosen to be a low frequency AR(2) process, even though noise free data were used. These results suggest that the low frequency AR(2) input process is so slowly varying that it is not sufficiently exciting for ARX or NARX model structure identification. An interesting phenomenon is that, although the 4 models given in Table 2 have different structures, they all produce the same (in fact indistinguishable) model predicted or long term outputs for any given input. Thus, in this regard, the four models are equivalent in output behaviour, even thought they are different in structure. It was also noticed that if the input signal was set to a high frequency AR(2) process, say $u(t)$=0.6$u(t$-1)− 0.0875$u(t$-2)+$\xi(t)$ with $\xi(t) \sim N(0,1)$, then the true model structure will be correctly identified.

As noted earlier, many factors can affect model structure selection including the presence of noise, the sampling rate and the richness of the input signal. Some subjective factors such as the selected maximum lags in the input and output terms, and the nonlinear degree specified for nonlinear candidate model terms will also affect the model structure selection. It has been verified by numerous simulation examples that if the maximum lags or key variables of the system can be appropriately chosen, then most of the irrelative model terms can be excluded and confidence of correctly selecting a minimum model structure or nearly minimum model structure can be significantly increased. Thus determining suitable values for the maximum lags and selecting significant variables as a first stage in model structure selection is likely to be highly beneficial. Several algorithms are available for the determination of the maximum lags for both the input and the output (Bomberger and Seborg 1998, Feil et al. 2004, Wei et al. 2004). In many cases, however, suitable maximum lags and significant variables may be difficult to determine, and some alternatives are thus worthy of investigation.

3. The new IFOS algorithm

The above discussion suggests that there is a need to improve the OLS-ERR algorithm to try and ensure that the correct model structure can be determined even when the data sets are not ideal. This motivates the development of the new integrated forward orthogonal search (IFOS) algorithm assisted by both the squared correlation and mutual information criteria. Before describing the IFOS algorithm, some preliminaries will be described first.

3.1 Some definitions

Definition 1: Primary variables and derivative variables

A primary variable is a dependent variable that originally exists in the model which characterises a given system. A derivative variable is derived from the primary variables. Generally, a primary variable is explicit in the model, but a derivative variable is implicit.

Consider the model below

$$y(t) = f(y(t-1), y(t-2), u(t-1)) \tag{12}$$

The variables $y(t-1), y(t-2), u(t-1)$ here are the primary dependent variables. Iterating (12) by one step with respect to the primary variable $y(t$-1), yields

$$\begin{aligned} y(t) &= f(y(t-1), y(t-2), u(t-1)) \\ &= f(f(y(t-2), y(t-3), u(t-2)), y(t-2), u(t-1)) \end{aligned} \tag{13}$$

The induced model (13) now involves 4 variables $y(t-2)$, $y(t-3)$, $u(t-1)$ and $u(t-2)$, where $y(t\text{-}3)$ and $u(t\text{-}2)$ are derived variables. Inspection of the results in Table 1 for Model 1 in section 2.4 shows that some of the derived variables may have been induced by the presence of noise if the candidate maximum lags are set to be too high. Therefore, if the primary variables of the system can be determined initially from the observational data, the accuracy of the model structure selection can then be significantly improved. Notice that the non-uniqueness which produces the result that the models in Eqs. (12) and (13) are equivalent is in general a direct result of the discrete model form. This non-uniqueness is common in most discrete forms and is often independent of the structure selection algorithms employed.

Definition 2: Model term dictionary

A model term dictionary $\mathscr{D}$ is a set whose elements are some specified (candidate) model terms (also called atoms or bases in signal processing). A dictionary $\mathscr{D}$ is said to be *over-complete* if all the true model terms are included in $\mathscr{D}$. A dictionary $\mathscr{D}$ is said to be *under-complete* (or *incomplete*) if at least one true model term is not included in $\mathscr{D}$. A dictionary $\mathscr{D}$ is said to be *exactly-complete* if all the true model terms are included in $\mathscr{D}$, but $\mathscr{D}$ contains no other candidate model terms. Clearly, for an *exactly-complete* dictionary the identification problem reduces to a structure-known estimation problem.

Assume that a system is described by the model: $y(t) = 0.7y(t-1) - 0.1y(t-2) + u(t-1)$, then $\mathscr{D}_1 = \{y(t-1), y(t-2), u(t-1), u(t-2)\}$ is over-complete; $\mathscr{D}_2 = \{y(t-1),\ y(t-1)u(t-1),\ u(t-2)\}$ is under-complete; and $\mathscr{D}_3 = \{y(t-1), y(t-2), u(t-1)\}$ is exactly-complete.

For a NARX model with a nonlinear degree ℓ and maximum lags n_y (for output) and n_u (for input), the candidate model term dictionary, including the constant term, is

$$\mathscr{D}_{n_y, n_u, \ell} = \{x_1^{i_1}(t) \cdots x_\ell^{i_\ell}(t) : x_j^{i_j} \in V_{n_y, n_u},\ 1 \le j \le \ell,\ 0 \le i_j \le \ell,\ 0 \le i_1 + \cdots + i_1 \le \ell\} \tag{14}$$

where $V_{n_y, n_u} = \{y(t-1), \cdots,\ y(t-n_y),\ \ u(t-1), \cdots,\ u(t-n_u)\}$. The number of elements in the dictionary $\mathscr{D}_{n_y, n_u, \ell}$ is $C_\ell^{n_y+n_u+\ell} = [(n_y + n_u + \ell)!]/[(n_y + n_u)!\ell!]$.

Definition 3: Model library

A model library $\mathscr{L}$ is a set whose elements are some specified models. A model selection criterion is always performed over a given model library.

Given a model library $\mathscr{L}$, the objective of model selection is to find the 'best' model from the library. All model selection criteria are relative, and there exists no absolute criterion that is able to measure all model fits under all conditions. A criterion will select the 'best' model structure over all the others even when the model library is inadequate ('inadequate' here means that no models in the library are exactly correct but only approximately correct). What the 'best' model is, depends on the specific situation. For example, the first three models given in Table 2 are structure incorrect compared with the true model. However, all the four models are equivalent if the model predicted outputs are of the most concern. The 'correctness' of a model structure is thus always relative and the ultimate objective is always to find the model structure that best fits, in either the structure and the output behaviour, the real system under study.

Definition 4: Model behaviour equivalence

Two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are said to be equivalent with each other in behaviour, if the (model predicted) outputs of the two models, driven by the same input, are the same.

In practice, it may be impossible to get exactly the same output behaviour for two different models. Thus, two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are often considered approximately equivalent when their outputs are sufficiently close when justified using a given criteria.

Assume that an identified model, $\mathcal{M}$, is given by

$$y(t) = f(y(t-1),\cdots, y(t-n_y), u(t-1),\cdots, y(t-n_u)) + e(t) \tag{15}$$

At a given time instance t_0, setting $\hat{y}^{mpo}(t_0-k) = y(t_0-k)$ for k=1,2, …, n_y, model predicted outputs at time instances $t \geq t_0$ are defined as

$$\hat{y}^{mpo}(t) = \hat{y}^{mpo}(t,\mathcal{M}) = f(\hat{y}^{mpo}(t-1),\cdots,\hat{y}^{mpo}(t-n_y),u(t-1),\cdots,u(t-n_u)) \tag{16}$$

While one-step-ahead predictions are often used to validate an identified model, previous experience shows that even a poor (e.g., insufficient, biased, unstable, etc.) model can provide good one-step-ahead predictions. Model predicted outputs can reveal severe model deficiencies which would otherwise go undetected by one-step-ahead predictions. However, in some cases, model predicted outputs may be unstable or may decay to zero, implying that model predicted outputs become invalid. In this case, a trade-off between one-step-ahead predictions and model predicted outputs is to use multi-step-ahead predictions.

3.2 Model term selection aided by mutual information

In the OLS-ERR algorithm described in section 2.2, a non-centralised squared correlation coefficient was used to measure the dependency between two vectors. This coefficient, between two vectors **x** and **y**, is defined as

$$C(\mathbf{x},\mathbf{y}) = \frac{(\mathbf{x}^T\mathbf{y})^2}{(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})} = \frac{(\sum_{i=1}^{N} x_i y_i)^2}{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2} \tag{17}$$

In the following, the well-known mutual information function will be employed as an alternative to the non-centralised squared correlation function, to aid the selection of significant model regressors.

3.2.1 Mutual information

Mutual information has now been extensively studied in the literature and applied to various areas. Following Cover and Thomas (1991), mutual information is defined as follows.

Consider two random discrete variables **x** and **y** with alphabet $\mathscr{X}$ and $\mathscr{Y}$, respectively, and with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(\mathbf{x},\mathbf{y})$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$, given as

$$I(\mathbf{x},\mathbf{y}) = \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{18}$$

The mutual information $I(\mathbf{x},\mathbf{y})$ is the reduction in the uncertainty of **y** due to the knowledge of **x**, and vice versa. Mutual information provides a measure of the amount of information that one variable shares with another one. If **y** is chosen to be the system output (the response), and **x** is one regressor in a linear model, $I(\mathbf{x},\mathbf{y})$ can be used to measure the coherence of **x** with **y** in the model. Several algorithms have been proposed to estimate mutual information from observed data, see for example Moddemeijer (1989, 1999), Darbellay and Vajda (1999), and Paninski (2003) and the references therein. In this study, the adaptive partitioning histogram method proposed in Darbellay and Vajda (1999) was employed to estimate relevant mutual information.

3.2.3 Inclusion of mutual information in model structure selection

Mutual information can easily be incorporated into the orthogonalization procedure in the same way as the squared correlation coefficient. Let $\mathscr{D} = \{\boldsymbol{\varphi}_j : j = 1 \le j \le M\}$ be a given model term dictionary. Let $\mathbf{r}_0 = \mathbf{y}$, and

$$\ell_1 = \arg\max_{1\le j\le M}\{I(\mathbf{r}_0, \boldsymbol{\varphi}_j)\} \tag{19}$$

where $I(\cdot,\cdot)$ is the mutual information function given by (18). The first significant basis can thus be selected as $\boldsymbol{\alpha}_1 = \boldsymbol{\varphi}_{\ell_1}$, and the first associated orthogonal basis can be chosen as $\mathbf{q}_1 = \boldsymbol{\varphi}_{\ell_1}$. Set

$$\mathbf{r}_1 = \mathbf{r}_0 - \frac{\mathbf{r}_0^T\mathbf{q}_1}{\mathbf{q}_1^T\mathbf{q}_1}\mathbf{q}_1 \tag{20}$$

In general, the *m*th significant model term can be chosen as follows. Assume that at the (*m*-1)th step, a subset $\mathscr{D}_{m-1}$, consisting of (*m*-1) significant bases, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_{m-1}$, has been determined, and the (*m*-1) selected bases have been transformed into a new group of orthogonal bases $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{m-1}$ via some orthogonal transformation. Let

$$\mathbf{r}_{m-1} = \mathbf{r}_{m-2} - \frac{\mathbf{r}_{m-2}^T\mathbf{q}_{m-1}}{\mathbf{q}_{m-1}^T\mathbf{q}_{m-1}}\mathbf{q}_{m-1} \tag{21}$$

$$\mathbf{q}_j^{(m)} = \boldsymbol{\varphi}_j - \sum_{k=1}^{m-1} \frac{\boldsymbol{\varphi}_j^T \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{q}_k} \mathbf{q}_k \tag{22}$$

$$\ell_m = \arg \max_{j \neq \ell_k, 1 \leq k \leq m-1} \{ I(\mathbf{r}_{m-1}, \mathbf{q}_j^{(m)}) \} \tag{23}$$

where $\boldsymbol{\varphi}_j \in \mathfrak{D} - \mathfrak{D}_{m-1}$. The mth significant basis can then be chosen as $\boldsymbol{\alpha}_m = \boldsymbol{\varphi}_{\ell_m}$ and the mth associated orthogonal basis can be chosen as $\mathbf{q}_m = \mathbf{q}_{\ell_m}^{(m)}$. Subsequent significant bases can be selected in the same way step by step.

From (21), the vectors $\mathbf{r}_{m-1}$ and $\mathbf{q}_{m-1}$ are orthogonal, thus

$$\| \mathbf{r}_{m-1} \|^2 = \| \mathbf{r}_{m-2} \|^2 - \frac{(\mathbf{r}_{m-2}^T \mathbf{q}_{m-1})^2}{\mathbf{q}_{m-1}^T \mathbf{q}_{m-1}} \tag{24}$$

By respectively summing (21) and (24) for m from 2 to n+1, yields

$$\mathbf{y} = \sum_{m=1}^{n} \frac{\mathbf{r}_{m-1}^T \mathbf{q}_m}{\mathbf{q}_m^T \mathbf{q}_m} \mathbf{q}_m + \mathbf{r}_n \tag{25}$$

$$\| \mathbf{r}_n \|^2 = \| \mathbf{y} \|^2 - \sum_{m=1}^{n} \frac{(\mathbf{r}_{m-1}^T \mathbf{q}_m)^2}{\mathbf{q}_m^T \mathbf{q}_m} \tag{26}$$

The residual sum of squares, also called the sum-squared-error, $\| \mathbf{r}_n \|^2$, or its variants including the mean-square-error (MSE) or the error-to-signal ratio defined as $\| \mathbf{r}_n \|^2 / \| \mathbf{y} \|^2$, can be used to form criteria for model selection. The model term selection procedure can be terminated when some specified termination conditions are met.

The motivation for introducing the mutual information assisted criterion here is not to totally replace the commonly used ERR criterion, but rather to provide an alternative and complementary approach to the ERR criterion. Further details will be given in Section 4.

3.2.3 Parameter estimation

It is easy to verify that the relationship between the selected original bases $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_m$, and the associated orthogonal bases $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_m$, is given by

$$\mathbf{A}_m = \mathbf{Q}_m \mathbf{R}_m \tag{27}$$

where $\mathbf{A}_m = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_m]$, $\mathbf{Q}_m$ is an $N \times m$ matrix with orthogonal columns $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_m$, and $\mathbf{R}_m$ is an $m \times m$ unit upper triangular matrix whose entries $u_{ij} (1 \leq i \leq j \leq m)$ are calculated during the orthogonalization procedure. The unknown parameter vector, denoted by $\boldsymbol{\theta}_m = [\theta_1, \theta_2, \cdots, \theta_m]^T$, for the model with respect the original bases (similar to (4)), can be calculated from the triangular equation $\mathbf{R}_m \boldsymbol{\theta}_m = \mathbf{g}_m$ with $\mathbf{g}_m = [g_1, g_2, \cdots, g_m]^T$, where $g_k = (\mathbf{r}_{k-1}^T \mathbf{q}_k)/(\mathbf{q}_k^T \mathbf{q}_k)$ or $g_k = (\mathbf{y}^T \mathbf{q}_k)/(\mathbf{q}_k^T \mathbf{q}_k)$.

Note that some tricks can be used to avoid selecting strongly correlated model terms. Assume that at the (m-1)th step, a subset $\mathfrak{D}_{m-1}$, consisting of (m-1) significant bases, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_{m-1}$, has been determined. Also assume that $\boldsymbol{\varphi}_j \in \mathfrak{D} - \mathfrak{D}_{m-1}$ is strongly correlated with some bases in $\mathfrak{D}_{m-1}$, that is, $\boldsymbol{\varphi}_j$

is a linear combination of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_{m-1}$. Thus there exist ($m$-1) real numbers $k_1, k_2, \cdots, k_{m-1}$, at least one of which is different from zero, such that

$$\boldsymbol{\varphi}_j = k_1\boldsymbol{\alpha}_1 + k_2\boldsymbol{\alpha}_2 + \cdots k_{m-1}\boldsymbol{\alpha}_{m-1} \tag{28}$$

From (27), there exists another set of real numbers, $\mu_1, \mu_2, \cdots, \mu_{m-1}$, such that

$$\boldsymbol{\varphi}_j = \mu_1\mathbf{q}_1 + \mu_2\mathbf{q}_2 + \cdots \mu_{m-1}\mathbf{q}_{m-1} \tag{29}$$

For the candidate basis given by (29), equation (22) becomes

$$\mathbf{q}_j^{(m)} = \boldsymbol{\varphi}_j - \sum_{k=1}^{m-1} \frac{\boldsymbol{\varphi}_j^T\mathbf{q}_k}{\mathbf{q}_k^T\mathbf{q}_k}\mathbf{q}_k = \mathbf{0} \tag{30}$$

Therefore, $(\mathbf{q}_j^{(m)})^T\mathbf{q}_j^{(m)} = 0$. In the IOFS algorithm, the candidate basis $\boldsymbol{\varphi}_j \in \mathfrak{D} - \mathfrak{D}_m$ will be automatically discarded if $(\mathbf{q}_j^{(m)})^T\mathbf{q}_j^{(m)} < 10^{-\tau}\{1, \boldsymbol{\varphi}_j^T\boldsymbol{\varphi}_j\}$, where τ is a positive number that is large enough. In this way, any severe multicolinearity or ill-conditioning can be avoided.

3.3 Model length determination

The role of model length determination in dynamical system identification has been widely recognised and various model selection criteria have been well established, see for example the recent review paper by Stoica and Selen (2004). Model selection criteria are often established on the basis of estimates of prediction errors, by inspecting how the identified model performs on future (never used) data sets. One general routine for model selection, which tries to avoid or reduce any possible bias introduced by relying on any particular test data sets, is cross validation (Stone 1974). Cross-validation has a number of variations, two commonly used variants of which are the leave-one-out (LOO), also called predicted sum of squares (PRESS) (Allen 1974), and generalised cross-validation (GCV) (Golub et al. 1979). Generalised cross-validation, due to its convenience of use and effectiveness for avoiding overfitting, has been widely accepted.

Following Golub et al. (1979) and Orr (1995), the generalised cross-validation criterion below will be used for model size determination

$$\mathrm{GCV}(n) = \left(\frac{N}{N-n}\right)^2 \mathrm{MSE}(n) \tag{31}$$

where N is the length of the training data set, n is the effective number of model terms (Moody 1992) where GCV is a minimum value, and $\mathrm{MSE}(n) = \|\mathbf{r}_n\|^2 / N$ is the mean-square-error that is associated to the model of n model terms. It should be pointed out that the GCV criterion tends to produce overfitted models (Friedman and Silverman 1989, Barron and Xiao 1991). The evaluation of the performance of the GCV criterion and some relative improvements have been reported in Billings and Wei (2007a).

3.4 Hypothesis testing on individual regression coefficients

Statistical methods play a unique role in the diagnosis and analysis of linear models. One aspect of the application of statistical methods for linear model analysis is hypothesis testing on regression coefficients (Hocking 1976, 1983, Montgomery et al. 2001). Consider the linear regression model with k regressors below

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e} \tag{32}$$

where $\mathbf{y}$ is $N \times 1$, $\mathbf{X}$ is $N \times n$ (if a constant term is included in (32), then all the elements of the first column of $\mathbf{X}$ are assumed to be unit), $\boldsymbol{\theta}$ is $n \times 1$, $\mathbf{e}$ is $n \times 1$, and $n=k+1$. Equation (32) is equivalent to (5), where $\mathbf{y}$ and $\mathbf{X}$ in (32) can be viewed as the output vector and the associated regression matrix $\mathbf{A}$ in (5), respectively. A frequently asked question is: do all the k regressors contribute significantly to the regression model?

The simplest but useful hypothesis for testing the significance of any individual regression coefficient, for instance θ_j in the model (32), is

$$H_0 : \theta_j = 0 \tag{33a}$$

$$H_1 : \theta_j \neq 0 \tag{33b}$$

If there is no sufficient reason to reject the null hypothesis $H_0 : \theta_j = 0$, then the corresponding regressor $\mathbf{x}_j$ can be removed from the model. The test statistic for this hypothesis is

$$t_0 = \frac{|\hat{\theta}_j|}{\mathrm{se}(\hat{\theta}_j)} \tag{34}$$

where $\mathrm{se}(\hat{\theta}_j) = \sqrt{\hat{\sigma}^2 c_{jj}^*}$ is the standard error of the regression coefficient θ_j, c_{jj}^* is the diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ corresponding to $\hat{\theta}_j$, and $\hat{\sigma}^2 = MS_{\mathrm{Res}} = \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}/(N-n)$ is the unbiased estimator of variance.

For a given α, if $t_0 > t_{\alpha/2, N-n}$, the null hypothesis $H_0 : \theta_j = 0$ can then be rejected. Note that this is really a partial or marginal test (Montgomery et al. 2001) because the regression coefficient $\hat{\theta}_j$ depends on all of the other regressors that are in the model. Thus it is a test of the contribution of $\mathbf{x}_j$ given the other regressors in the model.

For practical identification problems, where $N-n>120$, $t_{\alpha/2,N-n} \approx 1.96$ if α is set to 0.05, an equivalent test to (34) is

$$t_0 = \frac{|\hat{\theta}_j|}{1.96\,\mathrm{se}(\hat{\theta}_j)} \tag{35}$$

If $t_0 > 1$, the null hypothesis $H_0 : \theta_j = 0$ can then be rejected at the 95% confidence interval. Detailed information for the explanation of the rationale and the theoretical derivations for the above statistic can be found in (Montgomery et al. 2001).

The ERR (and mutual information) criterion and the t-test criterion are used in two separate procedures in the IFOS algorithm. The ERR criterion is used in the forward orthogonal search procedure, where significant model terms are selected step by step, one at a time. This procedure, coupled with the GCV criterion, will produce a model (or a set of models) formed by the selected significant model terms. But it has been noted that some spurious model terms may still be in the resultant model when the input is poorly designed. To refine the resultant model, the t-test is then used to detect and then remove the most probable spurious model terms.

In the next section, illustrations will be presented to show how the proposed IFOS algorithm works. A general procedure and some recommendation on how to apply the IFOS algorithm will then be given in Section 5. In the following, the IFOS algorithm, aided by the squared correlation, will be referred to as IFOS-SC. Similarly, the algorithm aided by mutual information criterion, will be referred to as IFOS-MI. As will be seen, for the same identification problem, IFOS-SC and IFOS-MI may or may not produce exactly the same model structure. By evaluating the performance of the resulting models, in accordance with some specified measures, a more accurate model structure can often be obtained. The relationship between ERR and MI criteria have not yet been found (Billings and Wei 2007b). A further study is still needed to reveal conditions under which one algorithm outperforms the other.

4. Case studies

In this section, several examples are provided to illustrate how to select an accurate model structure using the new IFOS algorithm. It will be shown that the IFOS algorithm can detect spurious model terms even when the data are contaminated with noise. A spurious model term here means that the model term is not in the true model but is selected with an ERR value that is not small. For cases where the input may not be sufficiently exciting, a trial-and-error approach can be used to avoid selecting the terms $y(t\text{-}1)$, $y(t\text{-}2)$, etc., since these terms are most likely to be selected even if they are not in the true model. In cases where the true model is not known, the performance of model predicted outputs will be examined to find the best model from a model set containing several candidate models generated from different libraries.

Notice that in the given examples, both artificial models and real data sets, where it is believed to be difficult to find the correct model structure, have been deliberately chosen to illustrate the effectiveness of the new IFOS algorithm. The examples are therefore far more demanding than typical model structure selection problems.

4.1 Example 1—the input is white

The following model was taken from Mao and Billings (1997)

$$\begin{aligned} y(t) = &-0.5y(t-2) + 0.7y(t-1)u(t-1) + 0.6u^2(t-2) \\ &+ 0.2y^3(t-1) - 0.7y(t-2)u^2(t-2) + e(t) \end{aligned} \tag{36}$$

Table 3 Identified model structure for system (36) using the IFOS-SC algorithm

Term	Parameter		ERR(%)	t-test	GCV
	True	Estimate			
$y(t-1)u^2(t-2)$	0	0.014704	34.9921	0.7382	0.074273
$y(t-1)u(t-1)$	0.7	0.706678	21.9095	69.9612	0.049441
$u^2(t-2)$	0.6	0.601460	12.3828	99.9614	0.035379
$y(t-2)$	-0.5	-0.491838	23.6688	59.4477	0.008150
$y^3(t-1)$	0.2	0.204638	4.5382	33.6203	0.002915
$y(t-2)u^2(t-2)$	-0.7	-0.708220	2.1595	27.4588	0.000412
$y(t-1)u(t-4)$	0	-0.026297	0.0045	1.1833	0.000403
$y^2(t-2)u(t-3)$	0	-0.012915	0.0044	1.1315	0.000400
$y(t-4)u(t-2)$	0	0.017330	0.0043	1.6214	**0.000396**
$y(t-3)y(t-4)u(t-2)$	0	-0.025846	0.0032	1.1110	0.000397
Run time: 0.906s					

Table 4 Identified model structure for system (36) using the IFOS-MI algorithm

Term	Parameter		Mutual Info	t-test	GCV
	True	Estimate			
$y(t-2)u^2(t-2)$	-0.7	-0.690247	0.251193	42.2583	0.118617
$u^2(t-2)$	0.6	0.599793	0.320914	149.1860	0.048510
$y(t-1)u(t-1)$	0.7	0.705487	0.188335	99.3864	0.026045
$y(t-2)$	-0.5	-0.501902	0.227581	66.5005	0.014168
$y^3(t-1)$	0.2	0.201394	0.214758	65.9884	**0.000393**
$u^2(t-1)u(t-4)$	0	-0.002367	0.012226	0.3664	0.000394
$u^2(t-2)u(t-3)$	0	-0.001729	0.008698	0.2627	0.000396
$y(t-4)u(t-2)u(t-4)$	0	-0.010032	0.008073	0.8780	0.000396
Run time: 2.126s					

where the input $u(t)$ was uniformly distributed on [−1, 1], with the noise $e(t) \sim N(0, 0.02^2)$. Following Mao and Billings (1997), the maximum lags of both the input and the output were assumed to be 4 and the nonlinear degree to be 3. Five hundred input-output data were generated and were used for model structure selection. The new IFOS algorithm, which incorporates the *t*-test given by (35), was applied to the data set, and the results are shown in Tables 3 and 4.

From Table 3, the ERR values show that the first 6 model terms are significant and should be included in the model. The first selected term, $y(t-1)u^2(t-2)$, with the highest ERR value is spurious. The t-tests show that among all the 10 model terms selected with the ERR criterion, only 5 are significant and the t-tests of the 5 terms are significantly different from unity. This means that the 5 terms with the highest t-tests dominate the regression model. This can easily be confirmed by inspecting the model predicted outputs based on the model with regard to the 5 model terms. The GCV values show that the appropriate number of model terms is 9, but clearly a model of 9 terms is overfitted.

Compared with Table 3, results given in Table 4 are quite optimistic. The t-tests show that only 5 model terms are significant, and the five model terms are exactly consistent with the 5 true model terms. In addition, GCV provides a correct indication of the structure, suggesting that 5 model terms are appropriate. Thus, from the results given by Table 3 and 4, all model terms can be correctly determined.

4.2 Example 2—the input is non-white

Consider the following two systems

$$\mathbb{S}_1: \quad w(t) = 0.5w(t-1) + 0.8u(t-2) + u^2(t-1) - 0.05w^2(t-2) + 0.5 \tag{37a}$$

$$y(t) = w(t) + \frac{1}{1-0.5q^{-1}}\xi(t), \ \xi(t) \sim N(0, 0.05^2) \tag{37b}$$

$$\mathbb{S}_2: \quad w(t) = u(t-1) + 0.5u(t-2) + 25u(t-1)u(t-2) - 0.3u^3(t-1) \tag{38a}$$

$$y(t) = w(t) + \frac{1}{1-0.8q^{-1}}\xi(t), \ \xi(t) \sim N(0, 0.02^2) \tag{38b}$$

Following Piroddi and Spinelli (2003), the input $u(t)$ to the two systems were chosen as a low frequency AR(2) process of the form: $u(t)$=1.6$u(t$-1)-0.6375$u(t$-2)+ $0.16\zeta(t)$, with $\zeta(t) \sim N(0,1)$. Two data sets of 500 input-output samples were generated from each system and the two data sets were used for model structure selection.

4.2.1 Experiments for system $\mathbb{S}_1$

Following Piroddi and Spinelli (2003), the maximum lags of both the input and the output were assumed to be 2 and the degree of nonlinearity to be 2. Model structure selection results for system $\mathbb{S}_1$ are reported in Tables 5 and 6. Following the analysis in Example 1, it is clear that the significant model terms should be selected as y(t-1), u(t-2), u^2(t-1), y^2(t-2), and the *const* term, which are exactly the same as the true model. Note that once the 5 model terms have been determined, the parameters need to be re-estimated based on just these selected model terms.

4.2.2 Experiments for system $\mathbb{S}_2$

Following Piroddi and Spinelli (2003), the maximum lags of both the input and the output were assumed to be 2 and the degree of nonlinearity to be 3. To ensure selection of the correct model subset, the IFOS-SC algorithm was applied over the following 5 different candidate model term dictionaries:

$\mathfrak{D}^u = \mathfrak{D}_{0,2,3}$, $\mathfrak{D}^0 = \mathfrak{D}_{2,2,3}$,

$\mathfrak{D}^1 = \mathfrak{D}^0 - \{y(t-1)\}$,

$\mathfrak{D}^2 = \mathfrak{D}^0 - \{y(t-2)\}$,

$\mathfrak{D}^3 = \mathfrak{D}^0 - \{y(t-1), y(t-2)\}$,

where the model term dictionary $\mathfrak{D}_{n_y, n_u, \ell}$ was defined by (14). The reason that the 5 different candidate dictionaries were considered here was two fold: one goal is to illustrate that the choice of initial dictionaries will affect the model selection results, and another goal is to show that spurious model terms can be detected using the *t*-test. Five different models, corresponding to the 5 dictionaries, were selected and the identified models are shown in Table 7. Similar results were also obtained using the IFOS-MI algorithm, but are not shown to save space.

While it is not quite apparent which model terms should be included in the model from the results with respect to $\mathfrak{D}^0$ and $\mathfrak{D}^2$, it is quite clear from the results with regard to $\mathfrak{D}^u$, $\mathfrak{D}^1$ and $\mathfrak{D}^3$ that the significant model terms included in the model should be $u(t\text{-}1)$, $u(t\text{-}2)$, $u(t\text{-}1)u(t\text{-}2)$, and $u^3(t\text{-}1)$, which are exactly the same as required by the system. Note that the search time to select the model terms is quite short, and it is less than 0.1s for each of the 5 cases.

Table 5 Identified model structure for the system (37) using the IFOS-SC algorithm

Term	Parameter		ERR(%)	t-test	GCV
	True	Estimate			
y(t-1)	0.5	0.500106	91.1027	71.4985	1.511037
y^2(t-2)	-0.05	-0.049757	3.5098	128.3416	0.922388
u^2(t-1)	1	1.000401	2.0742	132.8120	0.571884
u(t-2)	0.8	0.806721	2.8537	125.5270	0.079973
const	0.5	0.493459	0.4406	43.4106	**0.003336**
y^2(t-1)	0	-0.000419	0.0001	0.8359	0.003343
u^2(t-2)	0	0.006367	0.0001	0.6223	0.003360
Run time: 0.032s					

Table 6 Identified model structure for the system (37) using the IFOS-MI algorithm

Term	Parameter		Mutual Info	t-test	GCV
	True	Estimate			
u(t-1)	0	0.006148	1.313614	0.3120	15.160800
u^2(t-1)	1	0.994118	1.203510	61.4893	1.587509
y(t-1)	0.5	0.496906	0.244386	84.2243	1.077226
y^2(t-2)	-0.05	-0.049833	0.818507	135.5297	0.102098
u(t-1)u(t-2)	0	0.011942	0.332722	0.5739	0.091160
const	0.5	0.499216	0.218877	51.2285	0.039561
u(t-2)	0.8	0.800587	1.156804	36.8467	**0.003281**
y(t-1)u(t-1)	0	0.000024	0.000976	0.0210	0.003294
Run time: 0.141s					

Table 7 Identified model structures for the system (38) using the IFOS-SC algorithm

Term		Parameter		ERR(%)	t-test	GCV
		True	Estimate			
$\mathfrak{D}^u$	u(t-2)	0.5	0.496879	66.5315	31.3303	0.344189
	u^2(t-1)u(t-2)	0	0.000176	16.4164	0.0154	0.176546
	u(t-1)u(t-2)	0.25	0.253131	14.2253	113.7397	0.029466
	u(t-1)	1	1.002408	2.2567	61.4645	0.005983
	u^3(t-1)	-0.3	-0.299978	0.4670	26.4503	**0.001090**
	const	0	-0.002844	0.0005	0.8391	0.001092
$\mathfrak{D}^0$	y(t-1)	0	0.117996	90.4984	3.2882	0.121247
	y(t-2)	0	-0.012730	3.8298	1.2854	0.072865
	u^2(t-1)	0	0.040058	0.1612	2.4779	0.071273
	u(t-1)u(t-2)	0.25	0.184041	1.1284	18.3499	0.057063
	u(t-1)	1	1.026177	0.3607	52.7857	0.008343
	u^3(t-1)	-0.3	-0.296222	3.3894	85.2908	0.008343
	u(t-2)	0.5	0.318613	0.5477	15.5183	0.001121
	u^3(t-2)	0	0.027746	0.0044	2.8930	**0.001070**
$\mathfrak{D}^1$	y(t-2)	0	0.005719	81.2615	0.8224	0.195498
	u(t-1)	1	1.005003	5.5294	72.3156	0.138739
	u^3(t-1)	-0.3	-0.297251	5.5040	121.4937	0.081477
	u(t-1)u(t-2)	0.25	0.251067	6.9853	91.0853	0.007663
	u(t-2)	0.5	0.490089	0.6127	29.7224	**0.001148**
	const	0	0.003600	0.0007	0.9898	0.001148
$\mathfrak{D}^2$	y(t-1)	0	0.097761	94.6515	4.0993	0.072308
	u(t-1)	1	1.021391	0.3734	60.3493	0.067714
	u^3(t-1)	-0.3	-0.307184	1.4250	55.8901	0.048646
	u^2(t-1)	0	-0.029880	3.0680	1.9263	0.006651
	u(t-2)	0.5	0.336549	0.2329	7.6580	0.003461
	u(t-1)u(t-2)	0.25	0.265645	0.1777	19.8444	0.001000
	u(t-1)u^2(t-2)	0	0.034981	0.0036	3.1287	0.000955
	y^2(t-1)	0	-0.006900	0.0022	2.2771	**0.000930**
$\mathfrak{D}^3$	y(t-1)u^2(t-1)	0	0.000027	71.7306	0.0242	0.981663
	y^2(t-1)u(t-1)	0	-0.000045	12.3847	0.1902	0.555321
	u(t-2)	0.5	0.496203	4.9718	43.4379	0.384091
	u^3(t-1)	-0.3	-0.298608	6.4838	228.1314	0.156944
	u(t-1)u(t-2)	0.25	0.251097	3.1894	141.8768	0.044227
	u(t-1)	1	1.000408	1.2084	77.1917	**0.001123**
Run time: $\mathfrak{D}^u$ (0.031s), $\mathfrak{D}^0$ (0.059s), $\mathfrak{D}^1$ (0.079s), $\mathfrak{D}^2$ (0.094s), $\mathfrak{D}^3$ (0.047s)						

4.3 Example 3—forecasting annual sunspot numbers

The data set used in this example contains 301 observations of the annual sunspot numbers from 1700 to 2000. The first 280 samples for years 1700 to 1979 were used for model identification and the remaining 22 data were used for model performance testing. The candidate model term dictionaries were chosen as $\mathfrak{D}^0 = \mathfrak{D}_{12,0,1} = \{y(t-1),\cdots, y(t-12)\}$, and $\mathfrak{D}^1 = \mathfrak{D}^0 - \{y(t\text{-}1), y(t\text{-}2)\}$. The reason that the maximum lag was chosen to be 12 is due to the fact that the annual sunspot time series has a cycle that is about 11years. Although a nonlinear model for the sunspot time series may be more appropriate, the objective in this example is to illustrate the efficiency of the new IFOS algorithm for model structure selection, and a linear model was thus adopted.

The selected model structures from the dictionary $\mathfrak{D}^0$ using both IFOS-SC and IFOS-MI are shown in Table 8. Both algorithms suggested that the best model subset be chosen as $\{y(t\text{-}1), y(t\text{-}2), y(t\text{-}9), const\}$. The selected model structures from the dictionary $\mathfrak{D}^1$ by both IFOS-SC and IFOS-MI required 5 model terms: $y(t\text{-}3)$, $y(t\text{-}4)$, $y(t\text{-}9)$, $y(t\text{-}11)$, and *const*. It easily be shown that the performance of the model generated from $\mathfrak{D}^1$ is much inferior compared with the model generated from $\mathfrak{D}^0$.

The fact that the two different criteria (squared correlation and mutual information) yield the same results indicates that the linear regression model is dominated by the three significant variables $y(t\text{-}1)$, $y(t\text{-}2)$ and $y(t\text{-}9)$. This result enhances the previous conclusion (Wei et al. 2004) that $y(t\text{-}1)$, $y(t\text{-}2)$ and $y(t\text{-}9)$ are the three most important variables for describing the sunspot time series over the period from 1700 to 1979. By re-estimating the parameters in a linear model, the final identified model was given by y(t)=6.0223+1.2352y(t-1)−0.5404y(t-2)+0.1917y(t-9). One-step-ahead predictions and model predicted outputs produced by the identified model over the test data set are shown in Figure 1.

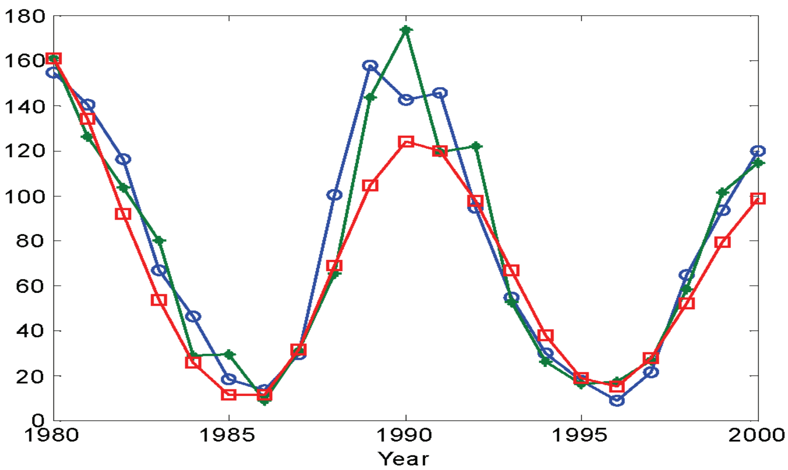

Fig. 1. One-step-ahead predictions and model predicted outputs produced from the identified model (with 4 model terms) for the sunspot time series. Solid line with circles indicates the measurements; dashed line with stars, one-step-ahead predictions; and dotted line with squares, model predicted outputs.

Table 8 Identified model structures for the sunspot time series

Term		Parameter	ERR(%) or Mutual info	t-test	GCV
SC	y(t-1)	1.202332	86.0183	10.1523	551.750797
	y(t-9)	0.187390	5.2192	3.3646	348.392854
	y(t-2)	-0.428369	2.7622	2.2895	240.374414
	const	6.275233	0.1884	1.2828	**234.594548**
	y(t-3)	-0.134668	0.0262	0.7185	235.314457
	y(t-4)	0.054645	0.0193	0.4780	236.322559
MI	y(t-1)	1.215845	0.442097	10.3688	551.750797
	y(t-2)	-0.532471	0.239983	4.2013	358.789312
	y(t-9)	0.161627	0.171117	1.6646	240.374414
	const	6.469004	0.036343	1.3200	**234.594548**
	y(t-10)	0.038577	0.045810	0.3668	235.862834
	y(t-4)	-0.005922	0.030401	0.0835	237.642482
Run time: IFOS-SC (0.078s), IFOS-MI (0.094s)					

4.4 Example 4—Drosphila or fruit fly modelling

This data set came from experiments and observations on a fruit fly, called Drosophila. The system input was the response of the photoreceptors, and the output was the response of the large monopolar cells. Recordings of 1000 points, sampled at a rate of 1kHz, on wild-type flies were collected.

The relationship between the input and the output in the fruit fly experiment is complex, because in addition to the response from the photoreceptors, several other factors may also affect the output response of the large monopolar cells. Identification of models relating these responses is therefore quite challenging. The objective of this example is to find a model that reflects, as closely as possible, the relationship between the response of the photoreceptors (the input) and the response of the large monopolar cells (the output), to facilitate the analysis and understanding of the associate behaviour of this kind of insect.

For the fruit fly modeling, the 1000 points in the data set were partitioned into two parts: the first 600 points were used for model identification, and the remaining 400 points were used for model testing. The input and the output over the test data set are shown in Figure 2.

The maximum lag for the input and the output were chosen to be 5 and 3, respectively, and the degree of nonlinearity to be 3. Similar to previous examples, the following 6 candidate model term dictionaries will be considered:

$\mathfrak{D}^u = \mathfrak{D}_{0,3,5}$, $\mathfrak{D}^0 = \mathfrak{D}_{3,5,3}$, $\mathfrak{D}^1 = \mathfrak{D}^0 - \{y(t-1)\}$, $\mathfrak{D}^2 = \mathfrak{D}^0 - \{y(t-2)\}$,

$\mathfrak{D}^3 = \mathfrak{D}^0 - \{y(t-1), y(t-2)\}$, $\mathfrak{D}^4 = \mathfrak{D}^0 - \{y(t-1), y(t-2), y(t-3)\}$,

where the set V_{n_y,n_u} was defined as defined as V_{n_y,n_u} ={y(t-1), …, y(t- n_y), u(t), u(t-1), …, u(t- n_u)}. The reason that the 6 different candidate dictionaries were considered here was as follows. Experience has shown that the terms y(t-1), y(t-2), etc. are most likely to be selected even if they are not in the true model. Based on this observation, the 6 different initial dictionaries were considered and these led to 6 different models. The model that produces the best output performance was chosen to be the final model. The average time used by the IFOS-SC algorithm for model structure selection, over different model term dictionaries, was 2.425s, and 4.688s for the IFOS-MI algorithm running on a standard PC.

Following the same procedures as described in previous examples, the IFOS-MI identified model, selected over the dictionary $\mathfrak{D}^2$, was found to be the best model, because the performance of the long-term predictions produced by this model were superior to the other identified models. The final IFOS-MI identified model contained 10 model terms. A comparison between the model predicted outputs and the measurements over the validation data set is shown in Figure 3. Clearly, the identified model fitted the experimental data extremely well.

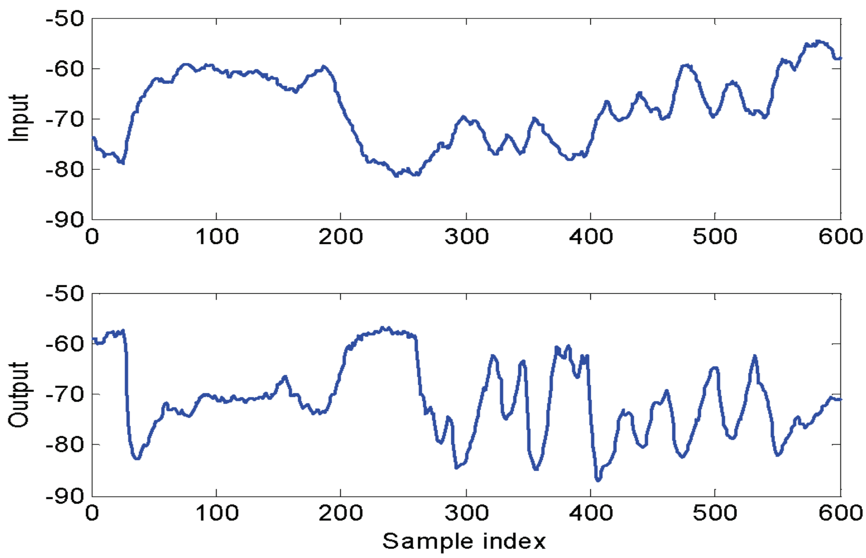

Fig. 2. The input and output data used for model estimation for the fruit fly modeling problem.

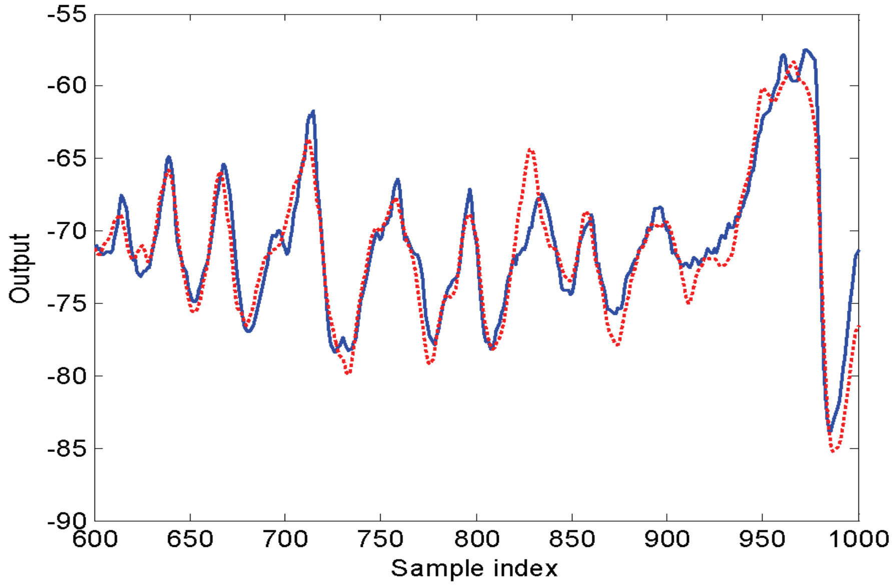

Fig. 3. A comparison between model predicted outputs and the measurements over the validation data set the fruit fly modeling problem. Solid line indicates the measurements and the dashed line indicates the model predicted outputs from the identified model for the fruit fly data set.

5. Discussions and recommendations

Model structure selection is a central issue in any nonlinear system identification problem. In addition to the input signal and sampling interval, many other factors, including the initial choice of the maximum lags for both the input and the output, the determination of the primary variables, the choice of initial candidate model term dictionaries, and the presence of noise (uncertainty in the data), all affect model structure selection. All these are generic problems in nonlinear system identification.

It is known that if the maximum lags or key (primary) variables for the system can be appropriately determined in advance, then irrelevant model terms can be precluded. Thus determining suitable maximum lags and selecting significant variables is a key step that could greatly improve the accuracy of all model structure selection procedures.

Results from numerous examples and applications in the literature have shown that the OLS-ERR algorithm can select accurate model structures for general nonlinear system identification problems. The algorithm may however occasionally produce redundant or incorrect model subsets in the presence of noise or if the input signal is not very exciting over the system bandwidth. To solve this problem, Piroddi and Spinelli (2003) suggested a simulation error based approach, which was implemented by minimizing the simulation error. This method, however, has two main drawbacks. First, it requires the calculation of model predicted or simulated outputs for all candidate model terms and can thus be extremely time consuming. Secondly, when the model predicted output diverges (the model is unstable) this is a clear indication that the model should be rejected. The examples provided in Piroddi and Spinelli (2003) and other related works show that the models rejected for this reason have an incorrect structure. For example, assume that a system is totally determined by a model subset of n model terms. An often encountered scenario is that, models formed by any subset of up to r ($< n$) terms may be unstable (infinitely divergent) or over attenuated (converge to zero), the model predicted output may thus be either infinite or zero. Clearly, the simulation error based approach will not work well for these cases and will not select any correct model subsets.

This study suggests the following four-stage trial-and-error experiments:

- Stage 1—*Select candidate model term dictionaries*.

  Let $\mathfrak{D}^u = \mathfrak{D}_{0,n_u,\ell}$, $\mathfrak{D}^0 = \mathfrak{D}_{n_y,n_u,\ell}$, $\mathfrak{D}^1 = \mathfrak{D}^0 - \{y(t-1)\}$, $\mathfrak{D}^2 = \mathfrak{D}^0 - \{y(t-2)\}$, and $\mathfrak{D}^3 = \mathfrak{D}^0 - \{y(t-1), y(t-2)\}$, where the model term dictionary $\mathfrak{D}_{n_y,n_u,\ell}$ is defined by (14).
- Stage 2—*Model structure selection*.

  Perform the forward orthogonal search algorithm over the 5 candidate dictionaries, respectively. This will lead to different model structures. Note that some spurious model terms may still be in the resultant models in cases where the input is poorly designed.
- Stage 3—*Model comparison*.

  Compare the performance of the identified models selected over the different model term dictionaries $\mathfrak{D}^u$, $\mathfrak{D}^0$, $\mathfrak{D}^1$, $\mathfrak{D}^2$ and $\mathfrak{D}^3$. Select the best model according to a specified criterion, for example the performance of model predicted outputs or multi-step-ahead predictions.
- Stage 4—*Model refinement*.

  Check the values of the t-test statistic, which indicates which model terms might be removed from the resultant model structure. Re-estimate model parameters if a couple of model terms need to be removed from or added into the selected model in Stage 3.

Note that the time spent on model structure selection using the orthogonal least squares type algorithms, for instance the IFOS algorithm here, is very short even for general cases. The above 4-stage trial-and-error experiments are thus not time demanding and can often be completed in a very short time. From the experience of numerous experiments including the four examples described in the present study, this 4-stage approach will usually provide accurate model structures.

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. In this case the NARX model (3) may fail to give a sufficient description due to the bias in the parameter estimates. As a consequence, the identified NARX model may not be sufficiently accurate if the model is used for other types of input signals. 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) and Billings and Chen (1998) for details about the NARMAX modelling methodology.

6. Conclusions

A new integrated forward orthogonal search (IFOS) algorithm, which is assisted by both the squared correlation and mutual information, and which incorporates a t-test and a general cross-validation (GCV) procedure, has been proposed for nonlinear system identification. The incorporation of the t-tests into the new IFOS algorithm has greatly enhanced the capability of detecting and hence removing any incorrect (spurious) model terms. The incorporation of a GCV into the new algorithm provides an important index for choosing an appropriate number of model terms.

It has been observed that for some input signals with a specific structure, the model term $y(t\text{-}1)$ is nearly always selected as the first term with a very high ERR value, and as a consequence the contributions of the other model terms, measured by the associated ERR values, can become small and sensitive to the effects of noise. This problem, however, has been effectively solved by introducing the four stage model selection procedure.

The new mutual information criterion can be used as a complementary approach or alternative to the squared correlation criterion. For a given identification problem, the two criteria may or may not produce exactly the same model structure. By inspecting and comparing the performance of the resulting models, in accordance with some specified measures, for example model predicted outputs, or multi-step-ahead predictions, a more accurate model structure can often be obtained. In this way, the accuracy of the identified model structure will be significantly improved compared with results based on any one single criterion.

The application of IFOS algorithm is not limited to the polynomial NARMAX model. The key idea in the IFOS algorithm can be applied to any linear-in-the-parameters model identification including the configuration and training of radial basis function (RBF) networks and wavelet modelling.

It should be noted that a comprehensive comparison between the ERR and MI criteria still needs to be considered, because cconditions under which one algorithm outperforms the other or vice versa, have not yet been found. This is why we suggest using the two algorithms in parallel, to lead to a model that is better than, or at least as good as, those produced by any one single criterion.

Acknowledgements

The authors gratefully acknowledge that this work was supported by EPSRC(UK). They are grateful to Dr M. Juusola, University of Sheffield, for providing the fruit fly data.

References

L. A. Aguirre and S. A. Billings, "Validating identified nonlinear models with chaotic dynamics", *International Journal of Bifurcation and Chaos*, 4, pp. 109-125, 1994.

L. A. Aguirre and S. A. Billings, "Dynamical effects of overparametrization in nonlinear models", *Physica D*, 80, pp. 26-40, 1995a.

L. A. Aguirre and S. A. Billings, "Improved structure selection for nonlinear models based on term clustering", *International Journal of Control*, 62, pp. 569-587, 1995b.

L. A. Aguirre and S. A. Billings, "Retrieving dynamical invariants from chaotic data using NARMAX models", *International Journal of Bifurcation and Chaos*, 5, pp. 449-474, 1995c.

L. A. Aguirre, P. F. Donoso-Garcia, and R. Santos, "Use of a priori information in the identification of global nonlinear models - A case study using a buck converter", *IEEE Transactions on Circuits and Systems I-Fundamental Theory and Applications*, 47(7), pp.1081-1085, 2000.

L. A. Aguirre, M. V. Correa, and C.C.S. Cassini, "Nonlinearities in NARX polynomial models: representation and estimation", *IEE Proceedings-Control Theory and Applications*, 149(4), pp.343-348, 2002.

L. A. Aguirre, M.F.S. Barroso, R.R. Saldanha, and E.M.A.M. Mendes, "Imposing steady-state performance on identified nonlinear polynomial models by means of constrained parameter estimation", *IEE Proceedings-Control Theory and Applications*, 151(2), pp174-179, 2004.

H. Akaike, "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, 19, pp. 716-723, 1974.

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

U. Anders and O. Korn, "Model selection in neural networks", *Neural Networks*, 12, pp. 309-323, 1999.

A. R. Barron and X. Y. Xiao, "Discussion: multivariate adaptive regression splines", *Ann. Statistics*, 19(1), pp. 67-82, 1991.

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

S. A. Billings and L. A. Aguirre, "Effects of the sampling time on the dynamics and identification of nonlinear models", *International Journal of Bifurcation and Chaos*, 5, pp. 1541-1556, 1995.

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

S. A. Billings, S. Chen, and R. J. Backhouse, "The identification of linear and non-linear models of a turbo charged diesel-engine", *Mechanical Systems and Signal Processing*, 3(2), pp. 123-142, 1989.

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

S. A. Billings, M. Korenberg, and S. Chen, "Identification of nonlinear output-affine systems using an orthogonal least-squares algorithm", *International Journal of Systems Science*, 19, pp.1559-1568, 1988.

S. A. Billings and H. L. Wei, "The wavelet-NARMAX representation: a hybrid model structure combining the polynomial models and multiresolution wavelet decompositions", *International Journal of Systems Science*, 36, pp. 137-152, 2005a.

S. A. Billings and H. L. Wei, "A new class of wavelet networks for nonlinear system identification", *IEEE Transactions on Neural Networks*, 16, pp. 862-874, 2005b.

S. A. Billings and H. L. Wei, "An adaptive orthogonal search algorithm for model subset selection and nonlinear system identification", *International Journal of Control*, 2007a (in press).

S. A. Billings and H. L. Wei, "Sparse model identification using a forward orthogonal regression algorithm aided by mutual information", *IEEE Transactions on Neural Networks*, 18(1), pp. 306-310, 2007b.

J. D. Bomberger and D. E. Seborg, "Determination of model order for NARX models directly from input-output data", *Journal of Process Control*, 8, pp.459-468, 1998.

L. Breiman and D. Freedman, "How many variables should be entered in a regression equation", *Journal of the American Statistical Association*, 78, pp. 131-136, 1983.

L. Breiman, "Better subset regression using the nonnegative garrote", *Technometrics*, 37, pp. 373-384, 1995.

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

S. Chen, E. S. Chng, and K. Alkadhimi, "Regularized orthogonal least squares algorithm for constructing radial basis function networks", *International Journal of Control*, 64, pp. 829-837, 1996.

S. Chen, X. Hong, and C. J. Harris, "Sparse kernel regression modeling using combined locally regularized orthogonal least squares and D-optimality experimental design", *IEEE Transactions on Automatic Control*, 48, pp. 1029-1036, 2003.

S. Chen, X. X. Wang, and D. J. Brown, "Sparse incremental regression modeling using correlation criterion with boosting search", *IEEE Signal Processing Letters*, 12, pp. 198-201, 2005.

T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

G. Davis, S. Mallat, Z. Zhang, "Adaptive time-frequency decompositions", *Optical Engineering*, 33, pp. 2183–2191, 1994.

G.A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space", *IEEE Transactions on Information Theory*, 45(4), pp.1315-1321, 1999.

Feil B, Abonyi J, Szeifert F, "Model order selection of nonlinear input-output models - a clustering based approach", *Journal of Process Control*, 14(6), pp.593-602, 2004.

J. H. Friedman and B. W. Silverman, "Flexible parsimonious smoothing and additive modeling", *Technometrics*, 31(1), pp. 3-21, 1989.

G. H. Golub, M. Heath, and G. Wahha, "Generalized cross-validation as a method for choosing a good ridge parameter", *Technometrics*, 21, pp. 215-223, 1979.

R. Haber and H. Unbehauen, "Structure identification of nonlinear dynamic-systems - a survey on input output approaches", *Automatica*, 26, pp. 651-677, 1990.

C.J. Harris, X. Hong, and Q. Gan, *Adaptive Modelling, Estimation and Fusion from Data : A Neurofuzzy Approach*. Berlin ; London : Springer-Verlag, 2002.

S. Haykin, *Neural networks: A Comprehensive Foundation* (2nd Ed). New York: Macmillan; Oxford: Maxwell Macmillan International, 1999.

B.R. Haynes and S.A. Billings, "Global analysis and model validation in nonlinear-system identification", *Nonlinear Dynamics*, 5(1), pp.93-130, 1994.

R. R. Hocking, "Analysis and selection of variables in linear-regression", *Biometrics*, 32, pp. 1-49, 1976.

R. R. Hocking, “Developments in linear-regression methodology:1959-1982”, *Technometrics*, 25, pp. 219-230, 1983.

X. Hong and C. J. Harris, “Nonlinear model structure design and construction using orthogonal least squares and D-optimality design”, *IEEE Transactions on Neural Networks*, 13, pp. 1245-1250, 2002.

X. Hong and S. Chen, “M-estimator and D-optimality model construction using orthogonal forward least regression”, *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 30, pp. 155-162, 2005.

J. H. Z. Huang and L. J. Yang, “Identification of non-linear additive autoregressive models”, *Journal of Royal Statistical Society Series B-Statistical Methodology*, Part 2, 66, pp. 463-477, 2004.

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

I. J. Leontaritis and S. A. Billings, “Experimental-design and identifiability for nonlinear-systems”, *International Journal of Systems Science*, 18, pp. 189-202, 1987.

I. J. Leontaritis and S. A. Billings, “Input-output parametric models for non-linear systems, part I: deterministic non-linear systems”, *International Journal of Control*, 41, pp. 303-344, 1985.

I. Lind and L. Ljung, “Regressor selection with the analysis of variance method”, *Automatica*, 41, pp. 693-700, 2005.

L. Ljung, *System Identification : Theory for the User*. Englewood Cliffs : Prentice-Hall, 1987.

J. Madar, J. Abonyi, and F. Szeifert, “Genetic programming for the identification of nonlinear input-output models”, *Industrial & Engineering Chemistry Research*, 44(9),pp. 3178-3186,2005.

S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries”, *IEEE Transactions on Signal Processing*, 41, pp. 3397-3415, 1993.

K. Z. Mao and S. A. Billings, “Algorithms for minimal model structure detection in nonlinear dynamic system identification”, *International Journal of Control*, 68, pp. 311-330, 1997.

E. Mendes and S. A. Billings, “Orthogonal estimator with intelligent structure detection: users guide”, Dept ACSE, The University of Sheffield, U.K., 1993.

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

R. Moddemeijer, “A statistic to estimate the variance of the histogram-based mutual information estimator based on dependent pairs of observations”, *Signal Processing*, 75(1), pp 51-63, 1999.

R. Moddemeijer, “On estimation of entropy and mutual information of continuous distributions”, *Signal Processing*, 16(3), pp. 233-246, 1989.

D. C. Montgomery, E. A. Peck, and G. G. Vining. Introduction to linear regression analysis (3rd Ed). New York: John Wiley & Sons, 2001.

J. E. Moody, “The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems”, in *Advances in Neural Information Processing Systems* (NIPS), J.E. Moody, S.J. Hanson, and R.P. Lippmann, Ed., pp. 847-854. Morgan Kaufmann, San Mateo CA, 1992.

O. Nelles, *Nonlinear System Identification: From Classical Approach to Neural Networks and Fuzzy Models*. Berlin: Springer-Verlag, 2001.

M. J. L. Orr, “Regularization in the selection of radial basis function centers”, *Neural Computation*, 7, pp. 606-623, 1995.

M. J. L. Orr, “Optimising the widths of radial basis functions”, *The Fifth Brazilian Symposium on Neural Networks*, pp. 26-29, Belo Horizonte, Brazil, Dec. 9-11, 1998.

L. Paninski, “Estimation of entropy and mutual information”, *Neural Computation*, 15, pp.1191-1253, 2003.

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

L. Piroddi and W. Spinelli, “An identification algorithm for polynomial NARX models based on simulation error minimization”, *International Journal of Control*, 76, pp. 1767-1781, 2003.

G. Rech, T. Terasvirta, and R. Tschernig, “A simple variable selection technique for nonlinear models”, *Communications in Statistics-Theory and Methods*, 30, pp. 1227-1241, 2001.

J. Rissanen, “Modelling by shortest data description”, *Automatica*, 14, pp. 465-471, 1978.

R. Savit and M. Green, “Time series and dependent variables”, *Physica D* 50, pp. 95-116, 1991.

J. Shao, “Linear-model selection by cross-validation”, *Journal of the American Statistical Association*, 88, pp. 486-494, 1993.

G. Schwarz, “Estimating the dimension of a model”, *The Annals of Statistics*, 6, pp. 461-464, 1978.

J. Sjoberg and L. Ljung, “Overtraining, regularization and searching for a minimum, with application to neural networks”, *International Journal of Control*, 62, pp. 1391-1407, 1995.

T. Söderström and P. Stoica, System Identification. New York : Prentice Hall, 1989

J. A. Stark and W. J. Fitzgerald, “Parameter-based hypothesis tests for model selection”, *Signal Processing*, 46, pp. 169-178, 1995.

P. Stoica, P. Eykhoff, P. Janssen, and T. Soderstrom, “Model-structure selection by cross-validation”, *International Journal of Control*, 43, pp. 1841-1878, 1986.

P. Stoica and Y. Selen, “Model-order selection: a review of information criterion rules”, *IEEE Signal Proc. Magazine*, 21, , pp. 36-47, 2004.

M. Stone, “Cross-validatory choice and assessment of statistical predictions”, *Journal of Royal Statistical Society Series B- Methodology*, 36(2), pp.111-147, 1974.

D. Tjostheim and B. H. Auestad, “ Nonparametric identification of nonlinear time series: projections”, *Journal of the American Statistical Association*, 89, pp. 1398-1409, 1994.

P. Vieu, “Order choice in nonlinear autoregressive models,” *Statistics*, 26, pp. 307-328, 1995.

H. L. Wei, S. A. Billings, and J. Liu, “Term and variable selection for nonlinear system identification,” *International Journal of Control*, **77**, pp. 86-110, 2004.

G. L. Zheng and S. A. Billings, “Radial basis function networks configuration using mutual information and the orthogonal least squares algorithm”, *Neural Networks*, 9, pp.1619-1637, 1996.

G. L. Zheng and S. A. Billings, “Qualitative validation and generalization in non-linear system identification”, *International Journal of Control*, 72(17), pp.1592-1608, 1999.

G. L. Zheng and S. A. Billings, “Effects of overparametrization in non-linear system identification and neural networks”, *International Journal of Systems Science*, 33(5), pp.331-349, 2002.

D. Zhu, M. Balikhin and S.A. Billings, “Variable and term selection for the identification of MIMO nonlinear systems, *International Journal of Modelling, Identification and Control*, 2007 (in press)

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