Dynamical effects of overparametrization in nonlinear models

Luis Antonio Aguirre\textsuperscript{a,1}, S.A. Billings\textsuperscript{b}

\textsuperscript{a} Centro de Pesquisa e Desenvolvimento em Engenharia Elétrica, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, MG, Brazil

\textsuperscript{b} Department of Automatic Control and Systems Engineering, University of Sheffield, P.O. Box 600, Mappin Street, Sheffield S1 4DU, UK

Received 17 May 1993; revised 14 March 1994; accepted 8 July 1994
Communicated by A.V. Holden

Abstract

This paper is concerned with dynamical reconstruction for nonlinear systems. The effects of the driving function and of the complexity of a given representation on the bifurcation pattern are investigated. It is shown that the use of different driving functions to excite the system may yield models with different bifurcation patterns. The complexity of the reconstructions considered is quantified by the embedding dimension and the number of estimated parameters. In this respect it appears that models which reproduce the original bifurcation behaviour are of limited complexity and that excessively complex models tend to induce ghost bifurcations and spurious dynamical regimes. Moreover, some results suggest that the effects of overparametrization on the global dynamical behaviour of a nonlinear model may be more deleterious than the presence of moderate noise levels. In order to precisely quantify the complexity of the reconstructions, global polynomials are used although the results are believed to apply to a much wider class of representations including neural networks.

1. Introduction

A major difficulty of some classical models used in the reconstruction of nonlinear systems is the huge number of parameters required to model relatively simple nonlinearities. For instance, the identification via Wiener's method of a simple system containing a second-order nonlinearity would require the evaluation of, typically, $10^{10}$ coefficients [1].

Although more recent representations like polynomials, rational functions, radial basis functions (RBF's) and neural networks permit modelling nonlinear systems with a rather restricted number of parameters, the total number of parameters can still run into the thousands. Moreover, such models are typically very flexible thereby enabling a very good fit to the data but such models can also become very easily too complex. Hence in order to make such representations of practical value it is important to restrict the structure of the final model. So far, the major inconvenience of considering model structures with a large number of parameters seems to have been related to the computation time and numerical problems. However, in view of an ever increasing computational power, there is the danger that the popularity of large and complex model structures, which hitherto were regarded as impractical, will also increase. Another danger is that the important problem of structure selection becomes overlooked.

Currently, there seems to be a tendency to use nonlinear models which are more complex than required.

\textsuperscript{1} e-mail: aguirre@cpdee.ufmg.br
Possible reasons for this are (i) inability to adequately choose a parsimonious representation, (ii) some structure selection algorithms may be too time-consuming, (iii) the variance of the residuals usually decreases as the complexity of a model is increased, (iv) large models are more flexible and consequently tend to fit the estimation data better. However, because flexible models which fit the data better do not necessarily capture the underlying dynamics appropriately, items (iii) and (iv) are misleading and may well be dangerous.

Although the need for parsimony is just beginning to be appreciated in some circles [2–4], the effects of unnecessary complexity in nonlinear representations on the global dynamics of such models have not been investigated in depth. It is the aim of the present paper to contribute some ideas and examples in this direction.

The paper is organised as follows. The problem of reconstructing dynamics from a single time series is briefly reviewed in Section 2. In Section 3 it is pointed out how models can be validated in order to assess the influence of the structure on the overall dynamics. In Section 4 it is shown that for non-autonomous systems, different inputs may lead to estimated models with different bifurcation patterns. Several aspects of the effects of overparametrization of nonlinear representations are discussed in Section 5. In Section 6 the model structure space is introduced in order to illustrate some points of Section 5 and also to provide a basic nomenclature for discussing the ideas concerning model structure selection. The main points of the paper are summarised in Section 7.

2. Reconstruction of nonlinear dynamics

Suppose that a measured time series \( y(1), y(2), \ldots, y(N) \) lies on a \( D \)-dimensional deterministic system. An important question in modelling is how to learn the dynamics of the \( D \)-dimensional attractor from the finite set of measurements \( y(1), y(2), \ldots, y(N) \).

The starting point for most current methods is to obtain an embedding, that is, a reconstructed space which is diffeomorphic to the original state space. A convenient, though not unique, representation is achieved by using delay coordinates [5,6]. A delay vector has the following form

\[
y(k) = [y(k) \ y(k-\tau) \ldots y(k-(d_e-1)\tau)]^T,
\]

where \( d_e \) is the embedding dimension and \( \tau \) is the delay time. Takens has shown that embeddings with \( d_e > 2n \) will be faithful generically so that there is a smooth map \( f^T : \mathbb{R}^{d_e} \to \mathbb{R} \) such that [6]

\[
y(k+T) = f^T(y(k))
\]

for all integers \( k \), and where the forecasting time \( T \) and \( \tau \) are also assumed to be integers. Recently, it has been shown that it is typically sufficient to take \( d_e > 2D \) for Eq. (2) to hold [7].

A number of papers have been devoted to the question of how to obtain an approximation of the map \( f^T \).

Most methods can be separated into two major groups, namely local and global approximation techniques.

The local approaches usually begin by partitioning the embedding space into neighbourhoods \( \{ \mathcal{U}_i \}_{i=1}^{N_n} \) within which the dynamics can be appropriately described by a linear map \( g^T : \mathbb{R}^{d_e} \to \mathbb{R} \) such that

\[
y(k+T) \approx g^T_i(y(k)) \quad \text{for} \quad y(k) \in \mathcal{U}_i, \quad i = 1, \ldots, N_n.
\]

Several choices for \( g^T \) have been suggested in the literature such as linear polynomials [8,9] which can be interpolated to obtain an approximation of the map \( f^T \) [10]. Simpler choices include zero-order approximations, also known as local constant predictors [8,11] and a weighted predictor [12]. Other approaches use weighted local maps [13].

One way of avoiding the need for constructing neighbourhoods is to fit global models to the data. In other words, all the data pertain to a single neighbourhood namely the entire embedding space. This considerable advantage has prompted some authors to investigate the estimation and use of global models such as polynomials [14,15,3,16], RBF's [17,18] and neural networks [19,20].
The use of global models, however, has two major difficulties namely (i) the choice of a representation for the model which should be sufficiently complex to approximate the dynamics of \( f^T \), and (ii) the selection of the correct structure or basis within the chosen representation. In the case of global models the representation has to be nonlinear. Consequently the number and sizes of all possible model bases within a given representation become very large as the dimension and the degree of nonlinearity increase and therefore overparametrization becomes an important consideration.

Nonlinear autoregressive moving average models with exogenous inputs (NARMAX) expanded as polynomials polynomials, which are briefly reviewed in the next section, were used to investigate the effects of overparametrization on the dynamics of nonlinear representations in general. This choice seems justified by a number of reasons including (i) such a representation is global and therefore better suited for analysing global dynamics, (ii) the complexity of the models is easily quantified by the degree of nonlinearity, the number of terms and the maximum lag used (the embedding dimension), and (iii) moderate amounts of noise can be conveniently handled thus making it possible to focus on the effects of the model structure. However, the deleterious effects of excessive complexity in the structure of nonlinear representations is believed to apply to a much wider class of models which include RBF's and neural networks. Indeed, some authors have recently reported (though not quantified nor characterised) the harmful effects of overparametrization in neural networks [21,22]. On the other hand, other researchers have suggested ways of reducing the complexity in dynamical models such as RBF's [4] and polynomials [3].

2.1. The NARMAX representation

A NARMAX model can be represented as follows [23,24]

\[
y(k) = F^T[y(k - 1), \ldots, y(k - n_y), u(k - d), \ldots, u(k - d - n_u + 1), e(k), \ldots e(k - n_e)],
\]

where \( n_y, n_u \) and \( n_e \) are the maximum lags considered for the output, input and noise terms, respectively and \( d \) is the delay measured in sampling intervals, \( T_s \). Moreover, \( u(k) \) and \( y(k) \) are respectively input and output time series obtained by sampling the continuous data \( u(t) \) and \( y(t) \) at \( T_s \). The sequence \( e(k) \) accounts for uncertainties, possible noise, unmodelled dynamics, etc. In this paper, the map \( F^T[\cdot] \) is taken to be a polynomial of degree \( l \).

There are some similarities between the NARMAX representation and the more general setting discussed in Section 2. In particular, the NARMAX approach does not involve finding neighbourhoods, thus \( N_1 = 1 \) and all the data belong to a unique “neighbourhood”, that is, \( y(k) \in U_k \), \( k = d_e, \ldots, N \). The delay time is taken to be equal to the sampling period, thus \( \tau = T_s \) and the maximum output lag, \( n_y \), is analogous to the embedding dimension \( d_e \).

There are, however, a number of important differences. Firstly, a NARMAX model includes input terms. This enables fitting data from non-autonomous systems and therefore estimating input/output maps. An immediate consequence of this is that for input/output systems, it is not required that the output be on any particular attractor. In fact a number of input waveforms can be used to enhance estimation accuracy and this will be briefly illustrated in Section 4.

Recently, Takens’ results have been extended for input-output systems [25].

Another important difference is the presence of noise terms, that is, the moving average part of the model. It should be noted that Eq. (2) will only hold in the unlikely case when noise is absent. Any noise in the data or any imperfection in the estimation of the map \( f^T \) will result in an extra term in the right hand side of Eq. (2). Such a term would be responsible for modelling the mismatch introduced by the noise and unmodelled dynamics. It is a well known result in the theory of system identification that if such a
term is omitted from the model structure, the estimate of the map \( f^T \) will become biased during parameter estimation. In order to avoid bias, in the NARMAX approach a stochastic component is appended to the model during parameter estimation. Once the parameters have been estimated, only the deterministic part of the model which is an approximation to the dynamics is retained. This procedure can handle moderate amounts of white and correlated noise.

The problem of how to compose the basis of a global polynomial has been solved using the error reduction ration ERR [26]. A similar approach has recently been used to select the structure of RBF's [4]. Using ERR it is possible to include in a model only the terms which are relevant and thereby to exclude the vast majority of a usually large set of candidate terms.

3. Model validation

A number of dynamical invariants can be used to assess the quality of nonlinear reconstructions. For autonomous systems, adequate validation tools include Lyapunov exponents, the correlation dimension and two-dimensional plots of the reconstructed trajectories [16,19].

For nonautonomous systems however, it has been pointed out that bifurcation diagrams are better suited for validation purposes because such diagrams are far more sensitive to the effects of structure and parameter changes than local invariants such as Poincaré sections, Lyapunov exponents and dimensions [27].

In order to facilitate the comparison of similar bifurcation diagrams, the following quality index will be used

\[
J_0 = \sum_{i}^{N_b} (A_i - a_i)^2 w_i, \tag{5}
\]

where \( A_i \in \mathbb{R} \) are the values of the bifurcation parameter \( A \). The \( a_i \)'s are defined likewise for the model being validated and \( w = [w_1 \ w_2 \ \ldots \ w_{N_b}] \) is a vector of weights which can be chosen to reflect the particular needs of the problem. In this paper \( \{w_i\}_{i=1}^{N_b} = 1 \) and the summation will be taken over all the \( (N_b) \) bifurcation points of interest.

4. Input design

For nonlinear systems, as opposed to linear systems, the frequency content of the output and input signals may be totally different. In some nonlinear systems, a single frequency input may produce an output with a wide frequency content. This is clearly the case for systems presenting chaotic motions. In practice, however, this will not be necessarily helpful because it might not be known a priori if a certain input will drive the system into chaos, nor if the system is at all chaotic. Therefore there is a need to design inputs which will excite the dynamics appropriately.

To illustrate the fact that different inputs may yield models with different bifurcation patterns, three inputs have been used to excite the Duffing-Ueda oscillator [28,29]

\[
\ddot{y} + k\dot{y} + y^3 = u(t), \tag{6}
\]

where \( u(t) = A \cos(\omega t) \). This system provides a bench test for the study of nonlinear dynamics in electrical science [30]. The respective bifurcation diagram is shown in Fig. 1.

**Sinusoidal inputs.** For some chaotic nonautonomous systems a single input frequency may suffice to excite the system properly providing the input drives the system into a sufficiently chaotic regime. This has been verified for the Duffing-Holmes and modified van der Pol oscillators.

![Bifurcation diagram of the Duffing-Ueda oscillator with \( u(t) = A \cos(t) \).](image)
Exciting the Duffing-Ueda system with a single frequency, however, did not yield the best results. Inputs in the chaotic window (see Fig. 1) for $5.55 \leq A \leq 5.82$ do not provide the level of excitation required and numerical problems occur. On the other hand, input in the chaotic window for $9.94 \leq A \leq 11.64$ does provide sufficient excitation but, although the bifurcation diagram was quite accurate, the Poincaré section of the attractor at $A = 11$ was rather distorted. In passing, it should be said that this was the only instance in which the bifurcation diagram suggested that the estimated model was very accurate as opposed to the Poincaré section which revealed some inaccuracies.

**Square wave plus Gaussian sequence inputs.** A square wave of amplitude $\pm 10$ and frequency equal to 1 rad/s was superimposed on a gaussian-distributed random variable of zero mean and variance $\sigma^2 = 9.1$. It has been shown that this type of input is more adequate than a pure Gaussian sequence [31].

Several NARMAX polynomial models with an increasing number of terms were estimated with cubic nonlinearities and embedding dimension equal to three. Bifurcation diagrams were computed for each model. The bifurcation points are shown in Fig. 2. In this figure, the horizontal axis indicates the number of terms in the model, $n_p$, and the vertical axis denotes the input amplitude, $A$. The dashed lines represent the bifurcation points of the original system and have been labelled B1 to B9 for convenience. Thus B1 $\rightarrow$ B2 $\rightarrow$ B3 is the first period-doubling cascade, and so forth. In this diagram the observed bifurcation points of the identified models are indicated by symbols which also appear on the right side of the figure for reference. Hence bifurcation points corresponding to B1 are marked “o”, those corresponding to B2 are marked “+” and so on. If a model does not exhibit a particular bifurcation for the range of amplitudes considered, such a bifurcation point is omitted in the figure. On the other hand, if a model has spurious bifurcations which are not present in the original bifurcation diagram, an asterisk is placed at the bottom of the figure.

This figure shows that none of the estimated models exhibit all the nine bifurcations and, in addition, most models present spurious dynamical regimes. Moreover, the bifurcation points which are reproduced by the estimated models are severely misplaced.

**Square wave with increasing amplitude plus Gaussian sequence input.** This type of input is similar to the previous one, the main difference is that the amplitude of the square wave in this case was gradually increased. In this example, the square wave amplitude was varied in the range $\pm 0.5$ to $\pm 13$ and the Gaussian component had variance $\sigma^2 = 2.0$. The amplitude limits of the square wave were chosen to span the range of values for which the control parameter was varied in the bifurcation diagrams. The bifurcation structure...
of the estimated models are summarised in Fig. 3.

Inspection of Fig. 3 shows that this input has made possible the identification of a family of models which are far better than those represented in Fig. 2.

4.1. Discussion

It is noted that the frequency spectra of the last two inputs are virtually identical. Nonetheless the latter input produces results which are clearly better. This illustrates that in designing an input for the identification of nonlinear systems amplitude considerations should be taken into account. This contrasts with the approach usually followed for linear systems [32].

It appears that for nonlinear systems therefore the requirements on the frequency content of the input may be somewhat relaxed due to the ability of such systems to transfer energy between different frequencies and actually create new frequencies at the output. On the other hand, greater attention must be paid to the amplitude profile of the input as illustrated in this section. This requirement on the input amplitude seems equivalent to imposing that the input should cause the system to explore all the regions of interest in the state space [33].

5. Model overparametrization

This section investigates the consequences of overparametrization on the dynamics of nonlinear representations. NARMAX polynomials are used as a tool to illustrate some of the effects.

5.1. Minimum and maximum number of terms

The minimum number of terms which should be included in a model can be roughly thought of as the minimum number of terms required such that the estimated model passes some validity test. In other words the minimum number of terms is directly associated with the size of the simplest model valid according to some specific criterion.

In practice, this number can be found by gradually allowing extra terms in a model and verifying if the final model reproduces the desired dynamics or not. When the estimated model is nonlinear, it is usually difficult to determine the best size for the model and more terms than strictly necessary are often included.

Consider the family of models estimated from data obtained from the Duffing-Ueda system excited by a square wave with increasing amplitude plus Gaussian sequence input sampled at $T_s = \pi/60$. Fixing the maximum lags allowed as $n_x = n_u = 3$, a number of models with different numbers of terms were estimated from the data. The bifurcation points of these models are shown in Fig. 4.

Comparing Figs. 3 and 4 reveals that $T_s = \pi/60$ is a better choice for the sampling period in this example. A clear difference between these figures is that all of the models represented in Fig. 4 exhibit all the bifurcation points.

It is worth pointing out that the models with six, seven, twelve, thirteen and fourteen terms have spurious bifurcation points, as indicated by the asterisks in the bottom line of Fig. 4. The evident displacement of the bifurcation points of the models with six and seven terms would certainly prevent such models from being considered as acceptable. By contrast, the models with twelve, thirteen and fourteen terms exhibit very accurate bifurcation behaviour but such models also present spurious or ghost bifurcations which indicate dynamical regimes which are qualitatively different from the original system.
Fig. 5. Bifurcation quality, $J_q$, for models with different number of terms and order.

Based on Fig. 4 it is clear that for the chosen values of $T_s$ and the maximum lag $n_y = n_u$, the minimum and maximum number of terms are eight and eleven, respectively. This figure illustrates that a model which is unnecessarily complex may induce spurious dynamical regimes and this has important implications for model structure selection and modelling in general.

Fig. 5 shows the bifurcation quality index, $J_q$, of the models identified using the same data as Fig. 4 for different values of maximum lag $n_y = n_u$ and total number of process model terms $n_p$. It is noted that only models with complete and non-spurious bifurcation diagrams were considered. This figure reveals that (i) fixing the order of the system, $n_y = n_u$, there is an optimum number of terms beyond which the bifurcation diagram begins to deteriorate, and (ii) there seems to be an optimum value for $n_y = n_u$ beyond which the bifurcation diagram also deteriorates.

Figs. 4 and 5 suggest that there is a minimum number of terms which should be included in the model in order to adequately capture the system dynamics, that there is a maximum number of terms beyond which spurious dynamical regimes are induced in the model and that in between these limits lies an optimum \(^2\) number of terms.

\(^2\) This is not a mathematical assertion but rather an observation inferred from Figs. 4 and 5 where the optimum is in terms of the quality index, $J_q$, defined in Eq. (5).

Fig. 6. Bifurcation diagrams of the (a) model estimated with $T_s = \pi/30$, $n_y = n_u = 4$, $n_p = 14$ and (b) $T_s = \pi/60$, $n_y = n_u = 3$, $n_p = 14$. These diagrams illustrate typical ghost bifurcations and spurious dynamical regimes induced as a consequence of overparametrization.

Figs. 6a–b show the bifurcation diagrams of the Duffing-Ueda oscillator and of the models estimated with $T_s = \pi/30$, $n_y = n_u = 4$, $n_p = 14$ and $T_s = \pi/60$, $n_y = n_u = 3$, $n_p = 14$, respectively. In the bifurcation diagram of Fig. 6a, spurious supercritical and subcritical flip bifurcations occur at $A \approx 7.2$ and $A \approx 7.6$, respectively. Furthermore, a spurious chaotic window is observed at approximately $8.65 \leq A \leq 8.95$. It should be noted that this chaotic regime has been observed for all of the estimated models with $T_s = \pi/30$, $n_y = n_u = 4$ and $14 \leq n_p \leq 17$. The spurious flip bifurcations have been observed for models with $n_p = 12, 14$ and 15.

A similar situation has also been verified for mod-
els estimated from data sampled at $T_s = \pi/60$. The respective bifurcation diagram is presented in Figs. 6b.

It is interesting to note that the Poincaré section of the attractors at $A = 11$ for the aforementioned model compare very well to the original attractor revealing that such sections are rather insensitive to overparametrization. Thus although the extra terms included in such models have induced spurious bifurcations to the original local dynamical regimes remain very much the same. In other words, the global behaviour of the identified model can be quite different to the behaviour of the underlying system. This highlights the need for a thorough and systematic validation procedure. In spite of this, validation is sometimes based solely on how well a model predicts over a different set of data, see for example [22] and most papers which use neural networks to model systems. But predictive model performance does not give any indication about how adequate the structure of the identified model is. Bifurcation diagrams give detailed information concerning the dynamical behaviour over a wide range of parameters and therefore seem more adequate for validation [27].

Similar results were observed for other systems including the modified van der Pol and Duffing-Holmes oscillators and Chua’s circuit. A few results concerning the latter system will be discussed in Section 6 below. In what follows, however, an example is provided which illustrates that excessive complexity in a nonlinear model may be more deleterious to the overall dynamics than noise on the data.

To investigate this point, the modified van der Pol oscillator [34]

$$\ddot{y} + \mu (y^2 - 1) \dot{y} + y^3 = A \cos(\omega t)$$ (7)

was used to generate data with $\mu = 0.2$, $A = 17$ and $\omega = 4$ rad/s. A second set of data was obtained by adding white Gaussian distributed noise to the clean data. The signal to noise ratio of the noisy data was $20 \log(\sigma_y^2/\sigma_n^2) = 70.6$ dB. Models were fitted to this data set. The bifurcation diagrams of the original system and of such models are shown in Fig. 7a–c.

The model estimated from the clean data had twelve terms and an embedding dimension equal to three. On the other hand, the model estimated from noise-corrupted data had sixteen terms and an embedding dimension equal to six. The clear increase in complexity was prompted by the need to handle the noise in the data. The only way to estimate valid NARMAX polynomials without such an increase in the complexity seems to be by smoothing the data before parameter estimation [35].

Finally, an extra term was added to the twelve-term model and parameters were subsequently reestimated. The embedding dimension of this model was also three. The bifurcation diagram of the resulting model is shown in Fig. 7d. The effects of overparametrization are clear. It should be noted that good models were estimated from the clean data with the number of terms varying between nine and twelve. Models with a larger number of terms presented increasingly inaccurate bifurcation diagrams of which Fig. 7d is the first one.

Comparing Figs. 7a–d, it becomes clear that overparametrization might in some cases have a stronger effect on the overall dynamics of the final model than noise on the data.

5.2. True identification versus data fitting

The parameters in a NARMAX model are estimated by minimizing the variance of the residuals which are defined as the measured output minus the one-step-ahead predicted output. Intuitively, as model complexity is increased, the more “flexible” such a model will become and a better fit to the estimation data will usually be possible. This is particularly true in nonlinear applications because the nonlinearities in the data will require a greater flexibility from the models to adjust to the records. Implicit in this process is the hope that a model which fits the data well will also capture the underlying dynamics of the true system in some appropriate manner. This is verified for chaotic systems in general as long as the noise present in the data is not excessive.

To illustrate this point further, an estimated model which reproduces the bifurcation pattern of the original Duffing-Ueda system [27] was used to predict over the estimation data. The resulting time series is shown in Fig. 8a. This is a typical scenario for a chaotic
system, namely the short term predictions are very accurate but gradually the model output drifts away from the measurements as a consequence of the sensitive dependence on initial conditions.

Allowing four extra terms and reestimating the model parameters gives a thirteen-term model which predicts the estimation data extremely well, see Fig. 8b. However this model, as noted in Fig. 4, includes spurious bifurcations. This clearly illustrates that a model which fits the estimation data better is not necessarily the best model. This is a very important observation. In a lot of the literature on estimation and especially neural networks, the claim is often made that one model produces superior predictions compared with another and must therefore be better. The prediction over a chaotic data set, see Fig. 8a, provides a good illustration of the folly of such a conjecture. The true model of the system is expected to exhibit poor long-term predictions. This is an inherent and fundamental property of systems which display extreme sensitivity to initial conditions. Any model therefore which produces apparently excellent predictions, as in Fig. 8b, cannot possibly be representative of the underlying system, it is just a curve fit to one piece of data. This is also one of the reasons why the validation should always be performed using...
5.3. Information criteria

It is clear that the number of terms in a model should be trade-off against the goodness of fit attained with such a model. Some information criteria are composed of two terms the first of which measures how well the model fits the estimation data and the second term penalises models with a large number of terms. These types of criteria were investigated.

The final prediction error (FPE) [36], Akaike’s information criterion (AIC) [36], Khundrin’s law of iterated logarithm criterion (LILC) [37] and the Bayesian information criterion (BIC) [38] were investigated. Other related criteria include the model entropy [15] and the Schwarz information criterion [4].

For the Duffing-Ueda system sampled at $T_s = \pi/30$ and $n_y = n_u = 4$, AIC(4), LILC and BIC indicated that the correct structure was that of the model with twelve terms. Although the bifurcation diagram of this model was similar to the diagram of the original system, spurious flip (supercritical and subcritical) bifurcations were observed. However, a thirteen-term model reproduced the bifurcation pattern fairly accurately. Similarly, for the double scroll Chua’s attractor [39,40], the four criteria pointed to a fifteen-term model. The embedded attractor reconstructed using such a model was clearly distorted whereas a sixteen-term model faithfully reproduced the dynamical invariants of the original attractor [16]. Finally, for the modified van der Pol oscillator the aforementioned criteria presented local minima for models with eight, eleven, thirteen and fifteen terms. For this system, models with a number of terms varying between nine and twelve were capable of reproducing the overall dynamics rather faithfully.

In view of such results, it appears that whilst such information criteria do not usually select the model with the best dynamical properties, the number of terms indicated as a statistical optimum is in general close to the dynamical optimum. Finally, it should be pointed out that a difficulty in applying such criteria to nonlinear problems is the appearance of local minima. Which of these minima to choose is not usually obvious and hence some search in the space of model structures seems inevitable. Such a space is the focus of the next section.

6. The model structure space (MSS)

Consider an $n_y = n_u$th-order NARX model with $n_p$ process terms which has been estimated from data sampled at $T_s$. The structure of this model can be char-
characterised, to some extent, by the numbers $T_s, n_p$ and $n_x = n_y$ and, hence, can be represented as a point in an $\mathbb{R} \times \mathbb{N}^2$ space which will be called the *model structure space* (MSS).

A number of examples involving different nonlinear systems suggest that there are lower and upper bounds on the parameters which determine an appropriate model structure. Thus, for an adequate polynomial model, $T_{s, \text{min}} \leq T_s \leq T_{s, \text{max}}, n_{p, \text{min}} \leq n_p \leq n_{p, \text{max}}$ and $n_{x, \text{min}} \leq n_x \leq n_{x, \text{max}}$. It should be noted that the upper bound on the number of terms in polynomials is often overlooked and consequently $n_p$ is sometimes equal to all possible terms up to a certain degree of nonlinearity.

These bounds exist for different reasons such as (i) theoretical, $T_{s, \text{max}}$ is limited by the sampling theorem, (ii) numerical, if $T_s \ll T_{s, \text{min}}$ the problem may become numerically ill-conditioned, (iii) dynamical, if $n_p < n_{p, \text{min}}$ the model may not capture the dynamics of the original system and if $n_p > n_{p, \text{max}}$ spurious dynamical regimes may be induced, and so on.

For some examples using the Duffing-Ueda oscillator, the following practical values were found: $T_{s, \text{min}} \approx \pi/100$ s, $T_{s, \text{max}} \approx \pi/30$ s, $n_{x, \text{min}} = 2, n_{x, \text{max}} = 5$, $n_{p, \text{min}} = 4$ and $n_{p, \text{max}} = 13$. These limits define a polyhedron in the MSS, see Fig. 9, where the best models estimated for $T_s = \pi/30, \pi/60$ and $\pi/100$ were represented by black dots. The oval dots are the projections of the model structures on the planes $T_s \times n_x, T_s \times n_p$ and $n_p \times n_y$.

The subregion $Q$ in Fig. 9 contains the structure of the best estimated models. Moreover, the best models presented in Fig. 5 are also in $Q$ and the projections fall on the shaded regions on the projection planes.

Fig. 9 might give the impression that the subspace $Q$ has a rather constant transversal section for $n_y \geq 4$. This is because $Q$ is only defined for integer values of $n_x$ and $n_p$ and, in the scales used, it would be difficult to represent the narrowing of $Q$ along the $n_y$ direction. This can be better appreciated in Fig. 5 by noticing that the bifurcation characteristics of the models deteriorate for $n_y \geq 4$.

It is noted that Fig. 9 is just an empirical representation of some of the results described in the paper. Other aspects of the shape of $Q$ relate to the sampling period $T_s$ and will be discussed elsewhere. No attempt
has been made to define or quantify the shape and dimensions of the subspace \( Q \). Nevertheless, this Fig. helps to illustrate that such a subspace is limited. This is in accordance with experimental results which show that valid models tend to be found in a bounded region of the MSS, and that increasing \( n_y \) and/or \( n_p \) beyond certain limits deteriorates the quality of the identified models.

A projection of \( Q \) onto the \( n_y \times n_p \) plane for the double scroll Chua’s attractor is shown in Fig. 10. Because the nonlinearity in this system is piecewise linear, neither the number of terms nor the embedding dimension are as sensitive to \( T_s \) as for other systems such as the Duffing-Ueda oscillator [16]. Thus the projection onto the \( n_y \times n_p \) plane seems a natural choice. Despite major differences between these two systems the main points observed for the Duffing-Ueda oscillator were also verified for the double scroll, namely, (i) there is a minimum number of terms necessary to provide the complexity needed to reproduce the chaotic attractor, (ii) there is a maximum number of terms beyond which the models become unstable, (iii) such bounds on the number of terms depend on the embedding dimension, and (iv) \( Q \) is a limited subspace of the MSS.

6.1. Rough estimates of the bounds of \( Q \)

Because the best models are confined to \( Q \), it would be helpful to have rough estimates for the limit values of \( T_s \), \( n_y \), and \( n_p \). This is discussed in the following.

\( T_{s,\text{max}} \) This is limited by the sampling theorem as \( T_{s,\text{max}} \leq 1/2f_m \), where \( f_m \) is the highest frequency of interest in the data.

\( T_{s,\text{min}} \) If \( T_s \) is taken too small, consecutive samples in the data will be highly correlated. The choice of \( T_{s,\text{min}} \) can be based on the first zero-crossing of the autocorrelation function of the data or by using results from information theory [41]. \( T_{s,\text{min}} \) can, of course, also be determined based on \( f_m \) as \( T_{s,\text{min}} \geq 1/20f_m \).

\( n_{y,\text{min}} \) The minimum number of degrees of freedom required to characterize the dynamics underlying the data can be obtained from estimates of the fractal dimension \( D \). If the system which produced the data has dynamical order \( n \), then \( D < n \) [30]. Methods of direct determination of the minimum embedding dimension are available in the literature [42-44].

\( n_{p,\text{max}} \) Apparently there is no way of determining an a priori value for this limit. Although Takens’ theorem provides a sufficient lower bound, in practice it has been found that this value is often larger than is required [42,45,44] and therefore could be used as an upper bound in some cases.

\( n_{p,\text{min}} \) This limit should mark the minimum number of terms for which some validation criterion is satisfied.

\( n_{p,\text{max}} \) There seems to be no way of determining an a priori value for this upper bound. Thus in practice information criteria may be computed as the model complexity is increased and the search may be halted not immediately but shortly after any of the criteria reaches a minimum.

It is believed that these considerations help to reduce the search space. This is very desirable in practice since term selection is usually very time consuming. Different model structures, such as neural networks, would be difficult to represent in the MSS. However it is believed that for neural networks the “subspace” in which good models are likely to be estimated is also limited.

\(^3\) It should be noted that \( D \) characterizes the attracting set (data) and not necessarily the system which produced the data.
7. Conclusions

This paper has been concerned with some aspects of the reconstruction of nonlinear dynamics. In particular, the influence of excessive complexity in nonlinear model representations has been investigated by using NARMAX polynomials as a means of quantifying and assessing overparametrization and studying the consequent effects on the bifurcation structure of such models.

The first aspect which has been investigated was input design. In linear systems the choice of the input signal is almost invariably specified based on frequency considerations in such a way as to guarantee that the signal is persistently exciting. For nonlinear systems not only the frequency content but also the amplitude profile of the input should be carefully chosen. Because of the energy transfer among different frequencies in nonlinear systems, the requirements on the frequency content of the input may be somewhat relaxed. On the other hand, because different dynamical regimes may be induced by simply altering the amplitude of a sinusoidal input, additional excitation can be attained by specifying a particular amplitude profile. This has been shown by way of an example.

Regarding model complexity, it has been argued that if a nonlinear representation is unnecessarily complex, such an approximation is prone to induce spurious dynamical regimes. This deleterious aspect of model overparametrization is in contrast to the sometimes misleading enhanced predictive ability that such models exhibit over the estimation data. Examples have been provided to illustrate that models which fit the estimation data better are not necessarily the models which capture the underlying dynamics adequately. This is highly relevant because as a consequence of the ever increasing computational power available, there seems to be a "natural" tendency to overparametrize nonlinear models.

The use of four information criteria (FPE, AIC(4), LILC and BIC) in selecting the best model from within a family has also been investigated. Although such criteria did not indicate the best model in most examples, the cut off was always close. However, because occasionally one extra term may induce a number of spurious bifurcations it seems that the best use of information criteria would be to indicate a suboptimal model size. The optimal model would then have to be searched for in the vicinity of the suboptimal result.

It has also been pointed out that a pitfall in the reconstruction of nonlinear dynamics is that most of the harmful effects of overparametrisation are not revealed by other classical tools used in signal processing such as correlation tests and nor by some techniques developed for the analysis of nonlinear dynamics such as Poincaré sections. Hence, bifurcation diagrams have been used to assess how the complexity of a model affects the overall dynamics.

Finally, mapping estimated models in what has been called the model structure space (MSS) has revealed that the subspace of the MSS where valid models are likely to be found is limited, thereby highlighting the dangers of overparametrisation. Suggestions on how to obtain rough estimates of the bounds of such a subspace have been proposed. This has been done for NARMAX polynomial models though similar spaces could be defined for different representations.

The field of nonlinear dynamics has experienced a great deal of excitement in the last few years as a consequence of the interest in chaos. The main reason for this was that very complicated dynamics, even unpredictable motions, could be produced and therefore modelled by simple deterministic equations. If on the one hand the chaos revolution was disclosing the fact that apparently random dynamical regimes could be characterised without resorting to stochastic theory, on the other hand it was reminding practitioners that in some cases simple models would suffice to model complicated dynamics. Thus it seems paradoxical that when approximating nonlinear (perhaps chaotic) systems large models and networks tend to be used almost as a default.

Acknowledgments

The authors are grateful to Professor Arun Holden for helpful comments on an early draft of this paper. LAA gratefully acknowledges financial support from
the Brazilian Council of Scientific and Technological Development - CNPq, under grant 200597/90-6. SAB gratefully acknowledges that this work was supported by SERC under grant GR/H 35286.

References


