ELSEVIER

Physica D 120 (1998) 273–286

PHYSICA D

Embedding as a modeling problem

Kevin Judd *, Alistair Mees

Centre for Applied Dynamics and Optimization, The University of Western Australia, Nedlands, WA 6907, Australia

Received 7 May 1997; received in revised form 22 January 1998; accepted 23 February 1998
Communicated by A.M. Albano

Abstract

Standard approaches to time-delay embedding will often fail to provide an embedding that is useful for many common applications. This happens in particular when there are multiple timescales in the dynamics. We present a modified procedure, *non-uniform embedding*, which overcomes such problems in many cases. For more complex nonlinear dynamics we introduce *variable embedding*, where, in a suitable sense, the embedding changes with the state of the system. We show how to implement these procedures by combining embedding and modeling into a single procedure with a single optimization goal. © 1998 Elsevier Science B.V.

1. Embedding and modeling

This paper concerns the analysis of time series, in particular, those that are recordings from nonlinear dynamical systems. In recent years, many new time-series analysis and modeling techniques have been derived from nonlinear dynamical systems theory. A common procedural feature of these techniques is *embedding*, which is performed before, and is quite distinct from, the modeling process. The practice of embedding rests upon a celebrated theorem of Takens [20] and a number of significant algorithms related to it [2,4,7]. It is well known that there are difficulties with the embedding step that precedes modeling but until recently no better alternative had suggested itself. Here we present an alternative that appears to be powerful and widely applicable.

Instead of treating embedding and modeling as two distinct procedures, we combine embedding and modeling into a single procedure with a single optimization criterion. In a sense, the requirement of a carefully constructed global embedding of the dynamical system is dispensed with, since for the purposes of time series prediction such an embedding is not needed, as we shall see. Our alternative procedures build either non-uniform global embeddings, or non-uniform local embeddings that vary with system state, the embeddings in either case being chosen so as to optimize a modeling criterion.

In order to understand the development of our modeling procedures, one has to be familiar with the minimum description length (MDL) principle for modeling time series. The main part of this paper begins (Section 2) with a brief description of the MDL principle and an algorithm for implementing it.

Section 3 is a critique of commonly accepted embedding procedures from a modeling point of view,

* Corresponding author.

0167-2789/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PII: S0167-2789(98)00089-X

in particular from the standpoint of time series prediction. Section 4 describes how to avoid some of the difficulties we identify, using *non-uniform* embeddings. As a special case we describe the building of linear models of time series; this results in a linear modeling technique, *reduced autoregressive* modeling, which has advantages over standard linear autoregressive modeling. The ideas on linear modeling are not central to the thesis we are developing; they are a special case that merits description in its own right. Section 4 includes applications of non-uniform embedding to both linear and nonlinear modeling of the sunspot time series, and to recordings of infant breathing during natural sleep.

Section 5 describes models using *variable embeddings*, which can be thought of as a set of local embeddings that vary with the system state. When variable embedding is applied to radial-basis models, which are a particular nonlinear modeling technique, one arrives at a natural generalization we call *cylindrical basis* models. We describe an application to the modeling of a spoken vowel; in effect, we are producing a phenomenological model of the dynamics of the vocal tract as used in producing that particular sound. We also consider the improved ability of nonlinear models with variable embeddings to capture the dynamics of a system from a time series by comparing radial and cylindrical basis models of the sunspot time series.

The principal motivation of this work is the desire to obtain models that accurately reflect the dynamics of the system being modeled. That is, a model should not only fit data and predict it well, but should also have dynamical behavior like that of the measured system. This is a stringent criterion that is satisfied by very few modeling methods; the work described here provides only a partial solution.

2. The minimum description length principle

The work described in this paper relies on two theoretical and procedural tools: the *minimum description length* principle, and an algorithm for building models in a general class of nonlinear models called *pseudo-linear* models. We have described these in other publications and readers are referred to them for greater detail than is provided here [10,11].

The minimum description length principle is an application of Occam's razor in a modeling context: it defines the best model for a time series to be the one that achieves the most concise description of the data. To understand how the principle works, suppose you (the "sender") have collected an experimental time series $x(t)$, $t = 1, \dots, n$, measured to an accuracy of (say) 12 bits, and you wish to communicate this data to a colleague (the "recipient"). You could send the raw data. Alternatively, you could construct a dynamical model from the data that enables the recipient to predict a value of $x(t)$ from earlier values. If you and your colleague have previously agreed on a class of models, then you could communicate the data by sending the parameters of a model, enough initial data to start predicting future values of the time series, and the errors between the true time series and the values predicted by the model. [1] Given this information, the recipient can reconstruct the experimental data to its full measured accuracy. An important point is that the parameters and errors need only be specified to finite accuracy. Furthermore, if the model is good, then the total number of bits required to transmit parameters, initial values and errors will be less than the number of bits of raw data.

In practice the minimum description length principle requires calculating an approximation to the *description length* of the time series and model, which is effectively the number of bits required to transmit the model plus the number of bits required to transmit the errors. (The initial conditions are included in the parameter count, though their effect only matters when we are comparing different embedding dimensions.) Under fairly general assumptions one can write

(Description length)

$\approx$ (number of data)

[1] Rissanen [16] presents another way to find models with minimal description length, using so-called "honest" prediction errors. This introduces difficulties in building nonlinear models, and we leave it for future work.

$$\times \log(\text{Mean square prediction error})$$
$$+ (\text{Penalty for number and}$$
$$\text{accuracy of parameters}).$$

As the number of parameters in a model increases the prediction errors decrease, but eventually, the penalty for introducing another parameter outweighs the benefit it has in reducing prediction errors. The model that attains the minimum description length is defined to be the optimal model within the class of models considered. We do not have space here to discuss in detail why this is successful; extensive discussions are to be found elsewhere [10,16,21].

In special model classes, explicit approximations to the description length can be calculated. A particularly useful class of parametrized nonlinear autoregressive model consists of those we call *pseudo-linear* models, which have the form

$$x(t+1) = \sum_{i=1}^{m} \lambda_i\, f_i(v(t)) + \varepsilon_t, \tag{1}$$

$$v(t) = (x(t), x(t-1), \ldots, x(t-d)) \tag{2}$$

for some selection of nonlinear functions f_i, unknown parameters λ_i and unknown i.i.d. random variates ε_t. (Observe in passing that choosing $v(t)$ amounts to using a particular embedding.) Define

$$V_i = (f_i(v(1)), \ldots, f_i(v(n)))^{\mathrm{T}}, \quad i = 1, \ldots, m, \tag{3}$$

$$y = (x(1), \ldots, x(n))^{\mathrm{T}}, \tag{4}$$

$$\lambda = (\lambda_1, \ldots, \lambda_m)^{\mathrm{T}}, \tag{5}$$

and let V be the matrix whose columns are V_i, $i = 1, \ldots, m$. If the ε_t are assumed to be Gaussian and λ has been chosen to minimize the sum of squares of the prediction errors $e = y - V\lambda$, then the description length [10] is bounded by

$$\left(\frac{n}{2} - 1\right) \ln \frac{e^{\mathrm{T}} e}{n} + (k+1)\left(\frac{1}{2} + \ln \gamma\right) - \sum_{j=1}^{k} \ln \delta_j, \tag{6}$$

where k is the number of non-zero components of λ, γ is related to the scale of the data and δ solves $[Q\delta]_j = 1/\delta_j$ where

$$Q = \hat{V}^{\mathrm{T}} \hat{V} / e^{\mathrm{T}} e,$$

and $\hat{V}$ is composed of just those columns of V that correspond to non-zero elements of λ. The variables δ can be interpreted as the relative precisions to which the parameters λ are specified.

The attraction of pseudo-linear models is that the parameters λ are easily calculated, since the sum of squares of the prediction errors $e = y - V\lambda$ can be minimized efficiently using singular value decomposition or any of its many equivalents. What makes general pseudo-linear models different from, and more powerful than, special cases such as linear or global polynomial models, is that the basis functions f_i can be chosen in many ways.

The critical problem is therefore how to select the basis functions f_i, which will, in general, be nonlinear functions depending on various parameters over which they are optimized. Unfortunately, this optimization is nonlinear and so is in general difficult, slow and prone to capture by local minima. (This problem is well-known in modeling via single-layer neural nets, a particular pseudo-linear approach.) Instead of optimizing the parameters of a few basis functions, we can instead generate a large number of fixed basis functions, not only at the start but also adaptively as the model-building progresses, and select a subset of them that optimizes the description length.

This alternative scheme requires an efficient combinatorial optimization method to select an optimal subset of the basis functions. It would appear that we have made the problem worse, since combinatorial optimization is notoriously hard, but in fact the following subset selection algorithm, described in detail elsewhere [10] is very successful in most of the applications we have considered. The algorithm selects subsets that are near-optimal according to the description length criterion, and hence produces good pseudo-linear models. It operates by adding and removing candidate functions from a given basis set according to a local optimality criterion, and accepting a set of given size as optimal if the same candidate is removed as was just added. The size of the basis set is increased until the description length criterion says it has become too large, and then the best set found so far is selected as the overall optimum.

In the algorithm, B represents any set of $k < m$ indices in $\{1, \dots, m\}$. We write V_B for the $n \times k$ matrix formed from the columns of V with indices in B, λ_B for the least squares solution to $y = V_B\lambda$, and $e_B = y - V_B\lambda_B$.

Algorithm 1.

(1) Normalize the columns of V to have unit length.

(2) Let $S_0 = (\frac{1}{2}n - 1)\ln(y^{\mathrm{T}}y/n) + \frac{1}{2} + \ln\gamma$.

(3) Let $B = \{j\}$ where V_j is the column of V such that $|V_j^{\mathrm{T}}y|$ is maximum. (This selects as the first basis function the one that most closely matches the data y. Note that $\lambda_B = V_j^{\mathrm{T}}y/V_j^{\mathrm{T}}V_j$ in this case.)

(4) Let $\mu = V^{\mathrm{T}}e_B$ and i be the index of the component of μ with maximum absolute value. Let $B' = B \cup \{i\}$. (The components of the vector μ measure how closely each of the basis functions not currently in use will match the error of the current model. Extend the current model with basis function that best matches the current error.)

(5) Calculate $\lambda_{B'}$. Let o be the index in B' corresponding to the component of $\lambda_{B'}$ with smallest absolute value. (Here o is index of the basis function that makes the smallest contribution to the current extended model.)

(6) If $i \neq o$, then put $B = B' \setminus \{o\}$ and go to step 4. (Throw *out* the "worst" basis function o if it is not i, the last one we brought *in*; then go back and try again.)

(7) Define $B_k = B$, where $k = |B|$. Find δ such that $(V_B^{\mathrm{T}}V_B\delta)_j = 1/\delta_j$ for each $j = \{1, \dots, k\}$ and calculate $S_k = (\frac{1}{2}n - 1)\ln(\hat{e}^{\mathrm{T}}\hat{e}/n) + (k+1)(\frac{1}{2} + \ln\gamma) - \sum_{j=1}^{k}\ln\hat{\delta}_j$. (At this stage we have found the best model of size k that can be built from the best model of size $k-1$ by "bringing in the best and throwing out the worst.")

(8) If $S_k < S_{k-1}$, then go to step 4. (Continue until the description length stops decreasing.)

(9) Take the basis B_k such that S_k is minimum as the optimal model.

3. Embedding

The modern practice of nonlinear time series analysis consists of two distinct steps: embedding and modeling. This section is a critique of the embedding step.

Embedding is a procedure where a scalar time series $x(t) \in \mathbb{R}$ is converted to a vector time series $v(t) \in \mathbb{R}^d$ for some integer d; the parameter d is referred to as the *embedding dimension*. The scalar time series is assumed to be a nonlinear function of the unknown system state Ξ, so that

$$x(t) = c(\Xi(t)).$$

The system state is assumed to evolve in discrete time according to

$$\Xi(t) = \Phi(\Xi(t-1))$$

and we call Φ the map governing the dynamics. There are several embedding procedures, but all are based on Takens' embedding theorem [20] which states that for almost any measurement $x(t) \in \mathbb{R}$ of a finite-dimensional dynamical system and for sufficiently large d,

$$v_t = (x(t), x(t-1), \dots, x(t-d+1)) \in \mathbb{R}^d$$

is an embedding of the dynamics of the original system in $\mathbb{R}^d$. Oversimplifying only slightly, the embedding induces a map F on the embedded states v_t such that

$$v_{t+1} = F(v_t),$$

and F is diffeomorphically conjugate to the dynamical map Φ that governs the original system; hence the systems are equivalent in all important respects. The embedding dimension d depends on both the dimension and the topology of the system's phase space, or at least its attractor, and is generally unknown. However, there are effective methods, such as the method of false nearest neighbors [2], for determining a suitable value of d.

In practice one generally uses a simple time-delay embedding

$$z(t) = (x(t), x(t-l), \dots, x(t-(d-1)l)) \in \mathbb{R}^d,$$

where l is called the *lag*; in anticipation of our coming discussion we will refer to this embedding as a *uniform* embedding. The lag is introduced to improve the observability of, for example, noisy time series. The lag is chosen to optimize the spread of the embedded time series without confusing the dynamics. There are two principal methods of choosing the lag: the first zero of the autocorrelation function [4] and the minimum of the mutual information [7].

Uniform embeddings for modeling purposes are at their most effective when embedding time series with a single, dominant periodicity or recurrence time. Both of the above-mentioned methods for calculating lags give similar lags for such time series and the lag is approximately one quarter of the dominant period. For this lag the embedded time series is ring-shaped; shorter and longer lags result in elliptical rings. The optimal lag in this case keeps states that correspond to similar phases close together and anti-phase states as far apart as possible. Uniform embeddings are quite suitable for classic chaotic systems such as the Rössler and Lorenz systems, which have a single dominant periodicity or recurrence time.

Uniform embeddings can fail when there are multiple strong periodicities with greatly differing timescales. For example, consider a quasi-periodic time series with greatly differing frequency components, or with very close frequencies that lead to a "carrier" frequency and a "modulation" of greatly differing periods. Fig. 1 shows three time series, from quite different systems, that all possess short and long period recurrences. Uniform embedding fails for such time series because a short lag is optimal for the high frequency component and a long lag is optimal for the low frequency components and modulation, while a compromise lag is inadequate for both timescales.

The problem is that embedded points corresponding to different phase states can lie close together in the embedding; when the time series are noisy there can be close proximity, and even overlap, of parts of an attractor that are distinct and should be kept widely separated in the embedding. The widely accepted criterion for an embedding to be good "geometrically" is that the embedded attractor should avoid self-intersection, and close proximity of distinct parts.

One method of overcoming the problem of multiple timescales is to use a *non-uniform* embedding: choose a vector of positive integers $(l_1, l_2, \ldots, l_d)$, called a *lag vector*, and define the embedding as

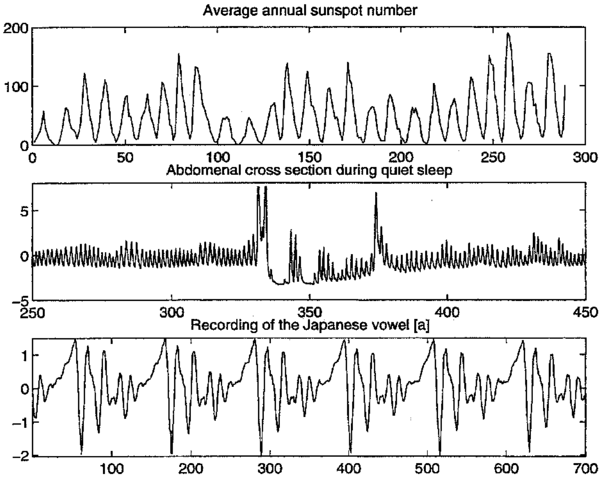


Fig. 1. Three time series that have multiple, strong periodicities: (a) the average annual sunspot number; (b) a sampling at 50 Hz of a measurement that is proportional to the cross sectional area of the abdomen of a child during quiet sleep: the phenomenon observed here is called "periodic breathing"; (c) a 12 kHz recording of the Japanese vowel [a].

$$v_t = (x(t-l_1), x(t-l_2), \ldots, x(t-l_d)) \in \mathbb{R}^d. \qquad (7)$$

Non-uniform embeddings can deal with multiple timescales by having greatly differing lags; short lags deal with short timescale recurrence and long lags with long timescale recurrence. Selection of the lag vector presents a problem, since there is a combinatorial explosion of possible embeddings as the embedding dimension increases. The next section presents a solution to the problem of finding an optimal lag vector.

4. Optimal lag vectors

In this section we present a method for finding the optimal lag vector for a non-uniform embedding strategy. Our method is based on the minimum description length principle and incidentally provides a linear autoregressive modeling technique that is useful in its own right.

The theme of this paper is that the question of what is a good embedding cannot be divorced from the question of what one is going to do with the embedded data; thus the correct criterion for optimizing the embedding in a modeling context is whatever one uses to measure the quality of a model. Whatever criterion is used should automatically ensure that the embedding is good geometrically as described in the previous section. Since the criterion we use is description length, we need to explain why minimizing description length should avoid close proximity and overlap of parts of an attractor. In the case of close proximity, the map fitted to the embedded data will have to vary rapidly with position, and to model rapid variations one generally needs more functions, or more complicated functions, and hence a larger description length. On the other hand, overlap will result in poor prediction, which means larger errors and so also a larger description length. An embedding that helps to minimize the description length, which implies minimizing both prediction error and complexity, is therefore expected to avoid the geometrical embedding problems of close proximity and overlap.

To find an optimal non-uniform embedding we first select a class of models such as pseudo-linear models with non-uniform embedding, that is,

$$x(t+1) = \sum_{i=1}^{m} \lambda_i \; f_i(v(t)) + \varepsilon_t, \tag{8}$$

$$v(t) = (x(t-l_1), x(t-l_2), \ldots, x(t-l_d)). \tag{9}$$

Compare these equations with Eqs. (1) and (2). For a given non-uniform embedding, specified by a lag vector $(l_1, l_2, \ldots, l_d)$, there is a model that is optimal with respect to description length, which can be found (possibly approximately) using Algorithm 1 described previously. Hence, for each lag vector there corresponds a description length, which is that of the optimal model given that embedding. The optimal embedding is found by a combinatorial optimization of the description length over all lag vectors.

The combinatorial optimization will generally be impractical but in some cases Algorithm 1 can be used successfully. The simplest case is the class of linear models. Contrary to what one might expect, the outcome of this application is not trivial. We call the models that result *reduced autoregressive models.* Although the idea does appear in the literature [9], it does not appear to have received much attention. One way that we shall use the results for the linear case is to obtain a global non-uniform embedding from the reduced autoregressive model, and then use this embedding as the basis for a nonlinear model.

A standard autoregressive linear model has the form

$$\begin{aligned} x(t) = a_0 &+ a_1 x(t-1) + \cdots \\ &+ a_w x(t-w) + \varepsilon_t, \end{aligned} \tag{10}$$

which is a linear model for a uniform embedding of dimension w and unit lag. The order d of the optimal model is usually taken to be the value of w where one attains the minimum of the Akaike Information Criterion (AIC) [3]

$$\mathrm{AIC}(k) = n \ln\left(\sum_{i=1}^{n} e_t^2 \Big/ n \right) + 2w, \tag{11}$$

or the minimum of the Schwarz Information Criterion (SIC) [17]

$$\mathrm{SIC}(k) = n \ln\left(\sum_{i=1}^{n} e_t^2 \Big/ n \right) + w \ln(n), \tag{12}$$

where e_t is the residual error of the model fit. Since it can be shown that the Schwarz Information Criterion is an asymptotic approximation to the description length of a linear model [16], it is preferable to use the description length formula of Eq. (6). Whichever criterion we use, we are optimizing over a pseudo-linear model with basis functions f_j which are the coordinate functions (that is, projections onto the coordinate axes) and a constant function.

The corresponding linear model for a non-uniform embedding with a lag vector $(l_1, \ldots, l_d)$ is

$$\begin{aligned} x(t) = a_0 &+ a_{l_1} x(t-l_1) + \cdots \\ &+ a_{l_d} x(t-l_d) + \varepsilon_t. \end{aligned} \tag{13}$$

Here $l_d \leq w$ is the maximum lag, and w is conveniently thought of as the width of a window that is slid along the time series. When the lag vector is chosen optimally, we call Eq. (13) a reduced autoregressive model. Reduced autoregressive models remove terms from the standard AR model that do not contribute significantly to the model, as assessed by the description length criterion.

To find the optimal reduced autoregressive model we choose the matrix V of Section 2 (Eq. (3)) so that its tth row is

$$(1, x(t-1), \ldots, x(t-w))$$

and so

$$\lambda = (a_0, a_1, \ldots, a_w).$$

Now we apply Algorithm 1 for sufficiently large w to find the optimal non-uniform embedding for a linear model; if it has $d \leq w$ non-zero coefficients other than a_0, this corresponds to a non-uniform d-dimensional embedding which is optimal (or nearly so) among all embeddings with maximal lag at most w.

Since in practice we do not know a good value of w until after Algorithm 1 has run, it is necessary to apply Algorithm 1 repeatedly to an increasing sequence of w values and watch for the value of d, and the chosen lags, to stabilize. To avoid spurious results due to the need to hold back more and more data at the beginning of the time series to supply initial conditions, it is best to specify a $w_{\max}$ in advance, and then run the increasing w sequence but always discarding $x(t)$ for $t = 1, \ldots, w_{\max}$. If it turns out that we are using lags as large as $w_{\max}$, we must discard everything, increase $w_{\max}$, and start again.

Fig. 2 shows the variation of the lag vector for an example data set. Observe how the lag vector more or less stabilizes as w increases. For sufficiently large w, Algorithm 1 typically selects some number $d < w$ lags for the lag vector.

4.1. Examples of non-uniform embeddings

We illustrate with two applications of non-uniform embeddings. The first example is the sunspot time series, which is well known to be difficult to model. We show that reduced autoregressive models give substantial improvement over standard autoregressive models. We also show that a particular nonlinear model is improved when a non-uniform lag vector is used. A second example is the analysis of infant breathing time series. This example illustrates that the optimal lag vector provides useful information about recurrences of a time series that are sometimes not obvious to the eye.

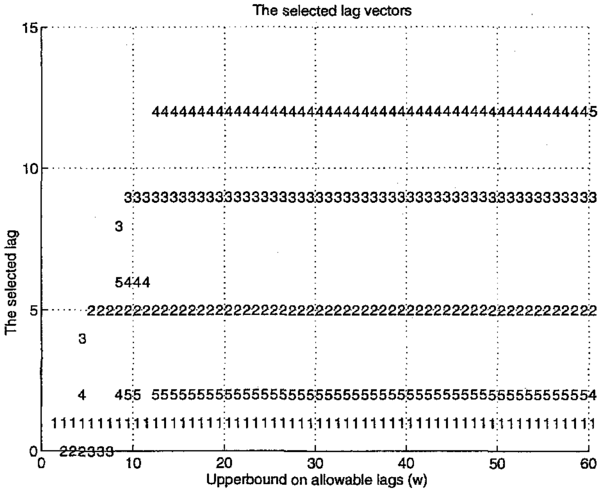


Fig. 2. For the data described in Section 4.1.2, this figure simultaneously shows the lags selected for a reduced autoregressive model (Eq. (13)) using Algorithm 1 for given upper bound w on the lag and shows the order in which the lags were selected. Vertically above each value of w there appear the numbers $1, 2, \ldots$; the height of label 1 is the value of the first selected lag, the height of label 2 the second selected lag and so on. For example, when $w = 20$ the lags in order of selection are 1, 5, 9, 12 and 2. As the upper bound on the lag w is increased the selected lag vector stabilizes to $(1, 2, 5, 9, 12)$ when $w > 12$. (Note: the reduced autoregressive model (Eq. (13)) may sometimes involve a constant term, when it happens it is indicated by a lag of zero; this only occurred when $2 \leq w \leq 8$.)

4.1.1. Sunspots

The authors have previously discussed the modeling of the average sunspot numbers (Fig. 1(a)) using reduced autoregressive models [10]. It was found that the optimal lag vector $(0, 2, 8)$ is good for autoregressive modeling and prediction of sunspot numbers one year in advance. (This embedding should be compared with the optimal uniform embedding with $d = 3$ and $l = 2$, equivalent to a lag vector $(0, 2, 4)$, that is recommended by the mutual-information and false-nearest-neighbor criteria.) A lag of 0 implies next year's average sunspot number will be close to this year's. The lag of 2 enables the model to estimate the rate of change of the average sunspot number (i.e. the first derivative) and is consistent with the uniform embedding. Although the three lags of a uniform embedding enable the model to estimate the rate of change of the rate of change (i.e. the second derivative or curvature), it is seen that the optimal lag vector uses a much longer third lag of 8.

Although a uniform embedding (0, 2, 4) provides enough information to predict a periodic variation in the sunspot numbers, it makes a poor prediction of the period. When free run, the uniform embedding models quickly lose synchronization with the 11 year sunspot cycle. The nonuniform embedding (0, 2, 8) performs much better, the reason being that the longer third lag gives better information about the period of the 11 year cycle. One way to look at this is to say that a lag of 8 in a period of near 11 is similar to a lag of -3, so we are able to put a parabola through the time series points x_{t-2}, x_t and x_{t+3} in order to estimate x_{t+1}. (Recall what was said about curvature in the previous paragraph.) The fact that this works well is an indication that the sunspot period is relatively stable: the period of the last cycle does not usually vary much from the period of the present one.

One can obtain slightly better models of the sunspot time series than those we reported previously [10], by using a nonlinear radial basis model with the above non-uniform embedding (0, 2, 8). Even better results are available from an application of the method of variable embedding, and this is described in the Section 5.2.

4.1.2. Breathing periodicities

This section considers reduced autoregressive models of the breathing patterns of infants during quiet sleep; Fig. 1(b) showed a typical recording of a measurement that is proportional to the abdomen cross-section. This recording shows a brief transition to what is called "periodic breathing" by clinicians, although a dynamical systems theorist would call it quasi-periodic. Periodic breathing usually only occurs in infants under four months of age and is of interest to those studying lung development. Later in this recording the child entered a long phase of periodic breathing: see Fig. 3. The periodic breathing in Fig. 1(b) occurs after a sigh, but significantly there appear to be modulations of breath amplitude prior to the sigh, with a similar period to the periodic breathing. One of the authors (KJ) has been using reduced autoregressive models to identify and measure the period of the modulation [18,19].

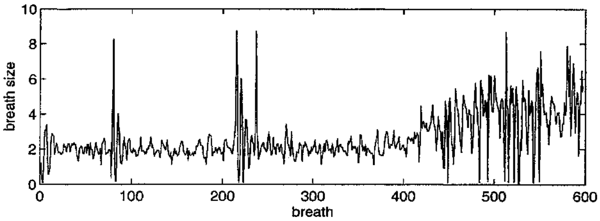


Fig. 3. A long recording of tidal volume of an infant during quiet sleep. Tidal volume is approximated as the difference between a peak and trough heights of abdomen cross section shown in Fig. 1(b). Strong periodic breathing begins after 450th recorded breath and Fig. 1(b) corresponds to breaths 150–250. The time series shows clear evidence of a periodicity which corresponds to a modulation of the breath amplitude of Fig. 1(b).

The time series analyzed is an approximation of the tidal volume series, estimated as the difference between peak and trough heights of Fig. 1(b). The breathing was recorded at 50 Hz so that peaks and troughs are well-enough resolved given the measurement noise. A tidal volume time series is shown in Fig. 3. Fig. 4 shows the power spectrum of the tidal volume series and Fig. 5 the autocorrelation function. As can be seen from Fig. 2, the optimal nonuniform lag selected is (1, 2, 5, 9, 12).

All three methods indicate modulations, although they are not in exact agreement. The lags 1 and 2 of the selected lag vector indicate that a breath is strongly correlated with the last breath and its previous breath; these provide a linear extrapolation from the current state of the system. The lag of 5 provides the tonic oscillation. The longer lags of 9 and 12 indicate a longer recurrence period, responsible for the modulations. The power spectrum and autocorrelation function show signs of longer term periodicities, because of the peak in the spectrum with side bands (that is, the modulation appears as "beating") and the periodicity superimposed on the decay of the autocorrelation function.

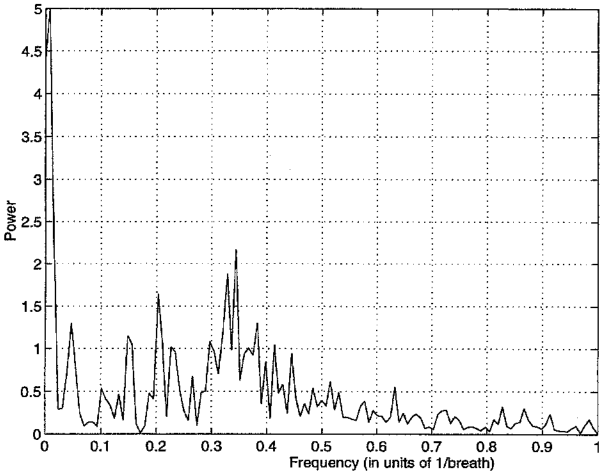


Fig. 4. The power spectrum of the tidal volume time series shown Fig. 3. The spectrum is consistent with a modulation of about 3–5 breaths, and possibly higher modulations.

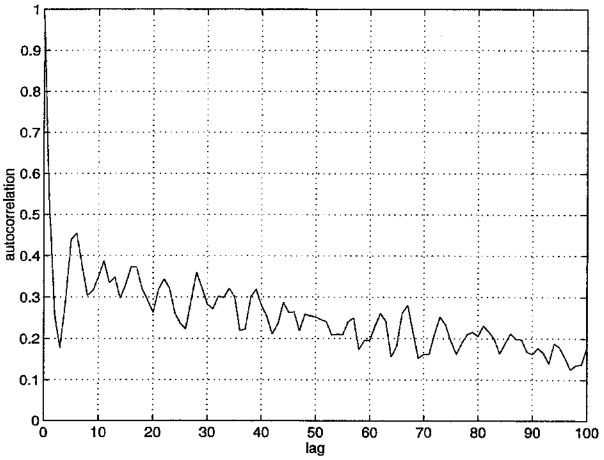


Fig. 5. The autocorrelation function of the tidal volume time series shown in Fig. 3. The autocorrelation function is consistent with a modulation of about 6 breaths.

5. Variable embedding

Despite the success of non-uniform embedding in overcoming multiple periodicities and recurrences, there are more subtle structural complexities of a time series that even non-uniform embedding cannot adequately resolve within the limitations of finite data sets and noisy systems. In this section we propose a technique to overcome these additional complexities. The price we pay is that we have to abandon the requirement of an optimized, global embedding.

Observe that specifying a fixed lag vector, whether uniform or non-uniform, implies a global embedding of the time series. We now argue that a fixed global embedding is not necessary and may, in fact, be counterproductive because it neglects information available in the time series. Instead, we look for locally optimal embeddings: that is, we imagine an embedding space of large dimension w as in Section 4, but now instead of finding a single projection we use *different* projections in different parts of the space. This idea is prefigured in the work of Broomhead and King on local singular value decomposition [5] and in the work of Abarbanel and his students on local false-nearest-neighbors [1]. Our contribution is to recognize that local embeddings can be found and optimized as part of the modeling procedure, rather than from local linear (or low-order polynomial) models.

To visualize why it is useful to do this, consider modeling the Lorenz system with its butterfly-shaped attractor. When the system state is out on the “wings” of the attractor, a two-dimensional embedding is sufficient to model and predict motions – one would require a lot of very high quality data to discern the thickness of these wings. However, near the origin where the cross over of the wings occur, a three-dimensional embedding is essential. One could imagine constructing a perfectly adequate model that does not use a global embedding, but rather uses appropriate local embeddings as the system state varies.

To distinguish local, state dependent embedding from other strategies we coin the name *variable embedding*. Finding variable embeddings is not necessarily easy, but we can go a long way with variants of the procedures we have already proposed. Recall that the guiding principle we have proposed for selecting a lag vector in a non-uniform embedding is that the embedding provide the best predictive model of the time series as measured by an information criterion. When one uses a variable embedding the processes of embedding and modeling are merged into one process with a single optimization goal. In Section 5.2 we describe one implementation of variable embedding to pseudo-linear models.

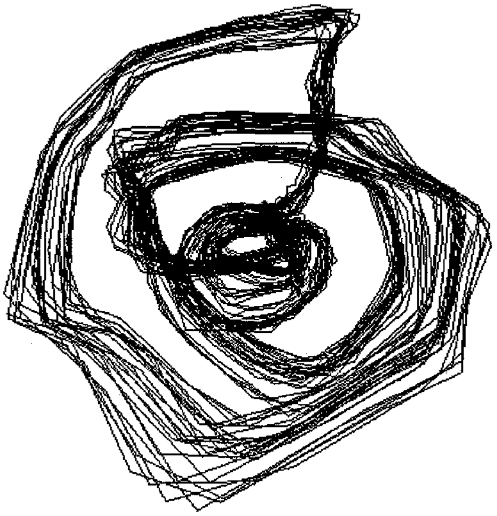


Fig. 6. A uniform embedding of the vowel time series shown in Fig. 1(c).

5.1. Advantages of variable embedding

In this section we describe some of the additional structural complexities which cannot be resolved by an optimized global embedding (uniform or non-uniform) but which can be resolved by variable embedding.

Consider the time series of Fig. 1(c), which is a recording of the Japanese vowel [a]. This time series has a complex structure with many periodicities. Fig. 6 shows a uniform embedding of the time series which reveals more of the complexity of the recurrences; with the aid of an interactive 3D viewer[2] the complexity, and some of the following discussion, becomes more obvious.

The time series possesses two features that make it difficult to model with a global embedding, even a non-uniform one. One feature is that with any reasonable low-dimensional embedding the tube-like embedded time series passes very close to, and even intersects, itself as it bends and spirals. A second feature, not so obvious from the embedding shown, is that the "speed" of a trajectory around the "attractor" varies greatly. For example, the system appears to have a Shil'nikov type mechanism [8,14] where it spirals in toward a fixed point along its stable manifold, then leaves along its unstable manifold. The timescale of the spiraling on the stable manifold is very much different from that of the departure along the unstable manifold. Accurate prediction of the moment of departure requires some long lags in the embedding, in this case lags of around 20. However, far from the fixed point the trajectories moves rapidly, and the optimal embedding has short lags of around 2 or 3. In both regions the local embedding dimension is 3 or 4.

It is clearly impossible to construct even a non-uniform global embedding that is suitable in both regions unless one goes to very high dimension, which invariably increases the difficulty of model construction. For example, in a region where a low-dimensional embedding, with short lags, is required, additional long lags would introduce irrelevant information that would generally contribute only noise and make predictions worse. On the other hand, near the unstable fixed point eliminating the long lags results in poor predictions of the sudden escape from that region.

5.2. Cylindrical basis models

The standard radial basis model [6,12,13] is pseudo-linear, with each of the nonlinear functions depending only on the radial distance from a certain point called a center. That is, in the standard pseudo-linear model of Eq. (1), the functions are here $f_i(v) = \phi(|z - c_i|/r_i)$ for suitably chosen centers c_i, radii r_i and radial basis function ϕ. If the function ϕ is decreasing, we can think of each f_i as each acting on a ball. If we ignore some coordinates (as a result of a local non-uniform embedding) then the functions act on cylinders instead, so a reasonable name is *cylindrical basis models*. To construct such models we must define the various cylinders.

Thus a cylinder is defined by a center c_i, a radius r_i and a lag vector $(l_1, l_2, \dots, l_k)$; the lag vector corresponds to a projection P such that

$$P(v(t)) = (x(t - l_1), x(t - l_2), \dots, x(t - l_k)).$$

The basis functions

$$f_i(z) = \phi(|P_i(z - c_i)|/r_i), \tag{14}$$

[2] For example, a VRML document is available at `http://cado.maths.uwa.edu.au/kevin/vowel`.

with decreasing ϕ, have the effect of localizing the embedding in the neighborhood of c_i. To simplify the notation we have introduced a redundancy in (14), because the center c_i is unique only up to the projection P_i.

To construct a cylindrical basis model using the selection methods we have described, one would generate a large set of basis functions (14), with decreasing function ϕ, having different centers c_i, radii r_i and projections P_i (which define the lag vectors). There is an obvious combinatorial explosion here because for each potential center there are $2^d - 1$ possible lag vectors. This could perhaps be tackled using genetic algorithms or simulated annealing, but we have found that a relatively simple adaptation of our earlier algorithms appears to avoid the worst of this explosion.

Algorithm 2.

(1) Let S represent an initial set of candidate basis functions. These functions are likely to be generated randomly but possibly using some additional information to choose likely candidates for selection. One good way to start is to generate centers and radii as usual, and then apply the reduced autoregressive method *locally* in the region around each center to get a local lag vector. This lag vector is only an initial guess, and will typically be thinned later in the process.

(2) Apply Algorithm 1 to determine the best model using the basis functions of S. Let S^* be the selected basis functions.

(3) If desired, locally optimize S^* by tuning its parameters using some standard method such as the Levenberg–Marquardt algorithm [15].

(4) Generate a new set of candidate basis functions S that includes S^*. The new candidate functions might be chosen with additional knowledge gained from the selection of S^*. (For example, we might try simplifying cylinders by applying addition projections, randomly perturbing selected cylinders, or putting new cylinders near where the model makes its worst predictions.)

(5) Return to step 2, and continue to do so while S^* is changing (significantly).

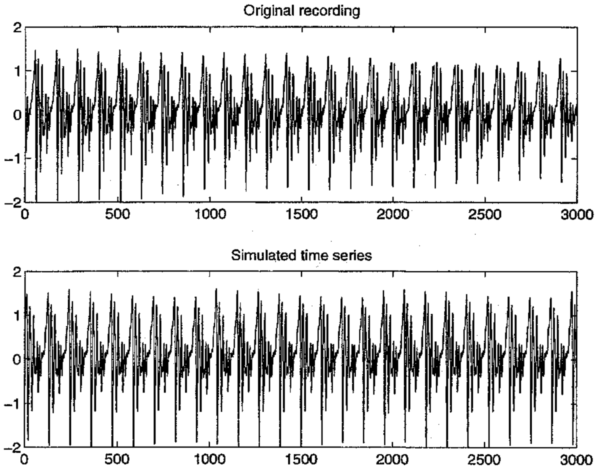


Fig. 7. This figure shows the complete original recording of the vowel sound and a simulation of a variable embedding (cylindrical basis) model. Observe that the period of the simulation appears to be exactly that of the original; this is remarkable since the model makes a one step prediction and the period of the signal is about 140 steps. Fig. 8 shows that the model also accurately reproduces many details. The only obvious differences in the two time series are very long term changes in amplitude over the whole time series. The simulation has an almost constant amplitude, although there is even some modulation of this that is similar to that the original signal.

The algorithm applies a classical purification technique from chemistry. The algorithm distills from S a good set of basis functions S^*, then dilutes and distills off S^* again and again until a (nearly) unchanging set of basis functions is obtained. The number of loops through Algorithm 2 depends on the size of S and the skill with which S was generated. Convergence is quickly achieved, and even though this may have more to do with the limitations of our selection schemes than with attainment of global optimality, the results are good as we shall see in Section 5.2.

There are a couple of final points that should be made about the cylindrical basis implementation of variable embedding. Firstly, the global embedding $v(t)$, on which the projections apply, need not be a uniform embedding of unit lag; the algorithm will run more quickly if a lag greater than one or a nonuniform embedding is used, and at least for over-sampled time series the exact lags chosen are unlikely to be critically important. The second point is the same

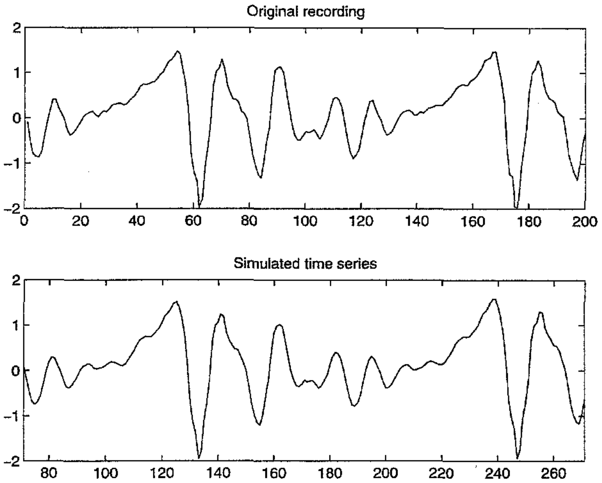


Fig. 8. This figure shows just a single period of the original recording of the vowel sound and the simulation of a variable embedding (cylindrical basis) model. Observe that the simulation reproduces a remarkable amount of detail as well as accurately reproducing the period.

one that arose in finding non-uniform embeddings: one needs to specify a window size $w_{\max}$ in advance, and increase it if local embeddings are being selected with lags near to $w_{\max}$.

5.2.1. Modeling a vowel sound

In this section we build a cylindrical basis model for the vowel sound recording of Fig. 1(c).

The data modeled is a half-second recording sampled at 12 kHz of a male pronouncing the Japanese vowel sound [a]. We constructed a one step prediction using a cylindrical basis model as follows. The global embedding was determined by increasing maximum lag method; specifically Algorithm 2 was used to build models on uniform global embeddings of lag 3 and dimension $d = 5$, 10, 15 and 20, and $d = 15$ provided a sufficient maximum lags. The set S of Algorithm 2 contained 200 randomly generated Gaussian cylinders as basis functions. The centers c_i of the cylinders were embedded time series points plus a Gaussian random variate with standard deviation 0.3 of the standard deviation σ of the time series. The radii r_i of the cylinders were uniformly distributed between 0.5 of the minimum inter-point distance of the embedded data and 2σ. The local embeddings were chosen with uniform probability from all possible local embeddings that could be projected from the global embedding and having dimension at least 2. During Algorithm 2 we also generated a variant for each member of S^* by perturbing the center by 0.3σ and keeping the same radius.

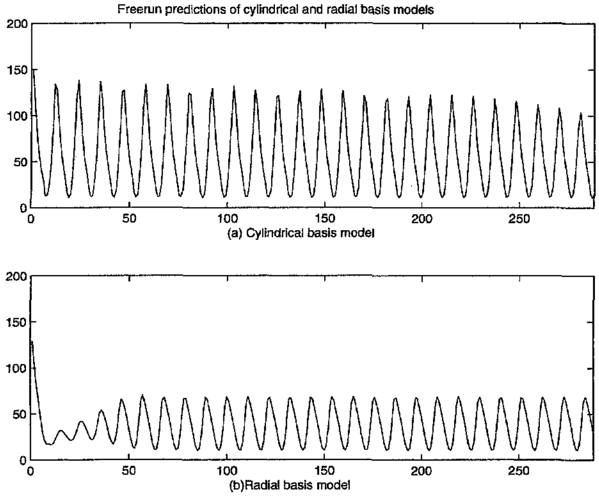


Fig. 9. This figure shows the typical free-run behavior (that is, zero noise simulations of Eq. (2)) of (a) cylindrical basis models and (b) radial basis models of the sunspot time series, when the free-run behavior was asymptotic to a limit cycle in both cases. The free-run starts from the final state of the embedded time series.

Algorithm 2 "converged" after 5 or 6 iterations in the sense that it was reselecting the same set of basis functions S^*, or minor variants of functions previously in S^*, with little change in the mean square error or description length. The quality of the final models is good. Fig. 7 shows the original time series and a simulation run of the dynamical model. (That is, the model is iterated many times without any corrections from the data.) The simulation accurately reproduces the waveform. Observe that the simulation has the same period as the original; this is remarkable since the model makes only a one step prediction and the period is approximately 140 steps. Fig. 8 shows just one period of the original signal and the simulation; the shape of the signal has been modeled accurately.

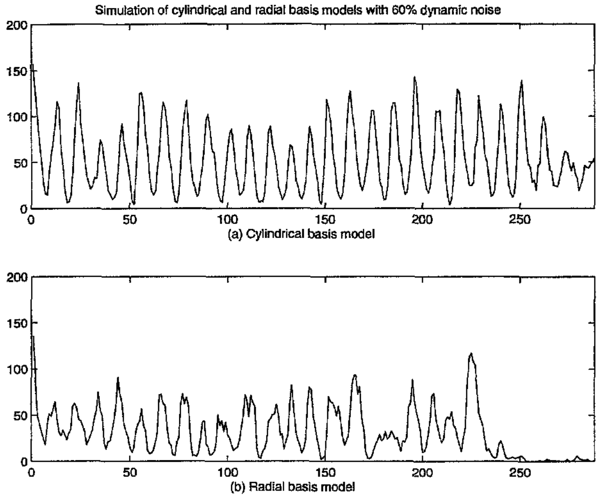
The model that generated the simulations used 70 cylinders plus a constant and 8 linear terms. The maximum local embedding dimension was 13 and the maximum lag 42. The minimum local embedding dimension was 3 and the average 6. The total description length is marginally less than taking a one period of the signal and laying this end to end to provide a prediction. The cylindrical basis model is certainly too big, because our calculation of the description length did not take into account the cost of sending position, radius and orientation of the cylinders, due to the fact that calculating the description length with these additional parameters is difficult. A crude approach to accounting for these additional parameters is to use the Schwarz criterion (12), where k counts the number of cylindrical basis functions and the radius and coordinates of the centers are treated as parameters. (Only the coordinates of the center that are significant after projection need be counted.) Although we obtain much smaller models with only marginally larger sum of squares of prediction error with this stronger criterion, we do not obtain the excellent free-run behavior described above. There are a number of possible reasons for this; most likely the selection algorithm needs to be improved to avoid problems of local minima in the optimization.

5.2.2. *Modeling sunspot dynamics*

In this section we compare radial basis models with cylindrical basis models for the sunspot time series shown in Fig. 1(a), and observe that cylindrical basis models more successfully capture the dynamics from the times series.

For the sunspot time series we built 50 of each of radial basis and cylindrical basis models using Algorithm 2 with a uniform global embedding of lag 1 and dimension 10 and basis function generation as described in the last section for the vowel sound model.

When these models were free run (that is, Eq. (2) is iterated with $\varepsilon_t = 0$) they asymptotically approached either a fixed point or a limit cycle; clearly a limit cycle is the more appropriate dynamical behavior. Only 10% of the radial basis models had desirable limit cycle behavior but only 6% of the cylindrical basis models had undesirable fixed point behavior. Fig. 9 shows typical limit cycle behavior of the radial and cylindrical basis models. It can be observed that even when the radial basis model has a limit cycle, the amplitude and transient are unconvincing, that is, the limit cycle appears to be an accident or uncertain; the cylindrical basis model is much more convincing.



(a) Cylindrical basis model

(b) Radial basis model

Fig. 10. This figure shows typical simulations of a (a) cylindrical basis model and (b) radial basis model, when the free-run behavior was asymptotic to a limit cycle in both cases. The simulation starts from the final state of the embedded time series and dynamic noise of 60% of the RMS error of the fitted error of the model is added.

We also studied noise driven simulations of the radial and cylindrical basis models that had limit cycle behavior in free run. What we hope to observe, for example, is the same statistical variation in the amplitude of the sunspot maxima as is seen in the origin time series. We simulated the time series starting from the final state of the original time series and injecting gaussian random variates (ε_t in Eq. (2)) with 60% of the RMS error of the fitted error of the model. (The value of 60% is a nominal value that allows some of the residual error to be due to measurement error and not dynamical noise.) We observed that cylindrical basis models had simulations that were much closer to the original time series; indeed the radial basis model simulations were quite poor. Fig. 10 shows a typical simulation of the cylindrical and radial basis model that have limit cycle behavior in free run. The cylindrical basis model simulations show good shape and convincing variation in the amplitude of maxima; the principal difference from the original time series is that the maxima and minima of the model are not as extreme as the original time series. This last observation is a common problem and the authors believe it is the result of not correctly accounting for the influence of dynamical noise in the model; this is the subject of current research.

Acknowledgements

This research was supported by a grant from the Australian Research Council. The authors thank Tohru Ikeguchi, Steven Stick and Michael Small for providing time series data. Michael Small did the calculations for Figs. 2, 4, 5, 9 and 10.

References

[1] H.D.I. Abarbanel, Analysis of Observed Chaotic Data, Springer, New York, 1996.

[2] H.D.I. Abarbanel, M.B. Kennel, Local false nearest neighbors and dynamical dimensions from observed chaotic data, Technical report, Department of Physics, University of California, San Diego, 1992.

[3] H. Akaike, A new look at the statistical identification model, IEEE Trans Automatic Control 19 (1974) 716–723.

[4] A.M. Albano, A.I. Mees, G.C. deGuzman, P.E. Rapp, Data requirements for reliable estimation of correlation dimensions, in: H. Degn, A.V. Holden, L.F. Olsen (Eds.), Chaos in biological systems, Plenum, New York, 1987, pp. 207–220

[5] D.S. Broomhead, R. Jones, G.P. King, Topological dimension and local coordinates from time series data, Technical report, Department of Mathematics, Imperial College, London, 1986.

[6] M. Casdagli, Nonlinear prediction of chaotic time series, Physica D 35 (1989) 335–356.

[7] A.M. Fraser, H.L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (1986) 1134–1140.

[8] P. Glendinning, C.T. Sparrow, Local and global behavior near homoclinic orbits, J. Stat. Phys. 35 (1983) 645–697.

[9] V. Haggan, O.B. Oyetunji, On the selection of subset autoregression time series models, J. Time Ser. Anal. 5 (1984) 103–113.

[10] K. Judd, A.I. Mees, On selecting models for nonlinear time series, Physica D 82 (1995) 426–444.

[11] K. Judd, A.I. Mees, Modeling chaotic motions of a string from experimental data, Physica D 92 (1996) 221–236.

[12] A.I. Mees, Parsimonious dynamical reconstruction, Int. J. Bifurcation and Chaos 3 (1993) 669–675.

[13] A.I. Mees, M.F. Jackson, L.O. Chua, Device modeling by radial basis functions, IEEE Trans CAS/FTA 39 (1992) 19–27.

[14] A.I. Mees, C.T. Sparrow, Some tools for analyzing chaos, Proceedings IEEE 75 (1987) 1058–1070.

[15] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, 1988.

[16] J. Rissanen, Stochastic Complexity in Statistical Inquiry, vol. 15, World Scientific, Singapore, 1989.

[17] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.

[18] M. Small, K. Judd, M. Lowe, S. Stick, Is breathing in infants chaotic? Dimension estimates for respiratory patterns during quiet sleep, Journal of Applied Physiology, submitted.

[19] M. Small, K. Judd, S. Stick, Linear modelling techniques detect periodic respiratory behaviour in infants during regular breathing in quiet sleep, Am. J. Respiratory Critical Care Medicine 153 (1996) 79.

[20] F. Takens, Detecting strange attractors in turbulence, in: D.A. Rand, L.S. Young (Eds.), Dynamical Systems and Turbulence, vol. 898, Springer, Berlin, 1981, pp. 365–381

[21] P. Vitanyi, M. Li, Ideal MDL and its relation to Bayesianism, Information, Statistics and Induction in Science, 1996.