

Chapter 1

Time series embedding and reconstruction

Nonlinear time series analysis is the study of time series data with computational techniques sensitive to nonlinearity in the data. While there is a long history of *linear* time series analysis, *nonlinear* methods have only just begun to reach maturity. When analysing time series data with linear methods, there are certain standard procedures one can follow, moreover the behaviour may be completely described by a relatively small set of parameters. For nonlinear time series analysis, this is not the case. While black box algorithms exist for the analysis of time series data with nonlinear methods, the application of these algorithms requires considerable knowledge and skill on the part of the operator: you cannot simply close your eyes and press the button.

Moreover, the growing artillery of nonlinear time series methods still has fairly few tangible, practical, applications: “applications” that satisfy the mathematicians or physicists definitely exist, but fairly few of these applications merit the attention of a practising engineer or physician.

This book is intended to be a guide book for the study of experimental time series data with nonlinear methods. We aim to present several applications, that, by the authors’ biased assessment, sit somewhere between a mathematician’s “application” and an engineer’s. To achieve such a lofty aim with such a slim volume, we need to restrict our attention somewhat: the subject of this book is the study of a single scalar time series measured from a deterministic dynamical system in the presence of observational and dynamic noise. A substantial portion of this volume concerns methods to obtain statistical evidence of nonlinear determinism rather than actual “applications”. The reason for this is simple: one should first establish that linear time series methods are insufficient before stepping outside the bounds of this venerable field.

In terms of understanding the underlying system, the focus of this book is two-fold. We are interested in statistically characterising the dynamics observed in a time series, and we are interested in developing methods to produce new time series exhibiting the same dynamics. The main tools which we will employ to achieve this are *the method of surrogate data*, *the estimation of dynamic invariants*, and *nonlinear modelling*.

In the remainder of this opening chapter, we provide the basic definitions and framework with which we will analyse data. We describe the various embedding theorems and appeal to these as the motivation of delay reconstruction and the subsequent methods. Throughout this opening salvo, we make one very significant assumption. We assume that the data are the deterministic output, measured with arbitrary (but finite) precision, of a deterministic dynamical system.

1.1 Stochasticity and determinism: Why should we bother?

So, let us begin with some mathematical necessities. The main purpose of delving into this detail so early in a book on *applied* methods, is to define precisely what it is we are interested in studying, and what we expect to find.

A deterministic dynamical system is one for which there is a rule, and, given sufficient knowledge of the current state of that system one can use that rule to predict future states. For notational convenience we will consider only discrete dynamical systems (this after all, is the only sort that modern digital computers can cope with). The extension to continuous systems is obvious and is dealt with at considerable depth elsewhere. Often, although our notation relates to a discrete signal it is in fact a continuous system sampled at a (hopefully) sufficiently high sampling rate. Let z_n be the current state of the system. We suppose that $z_n \in \mathcal{M} \subseteq \mathbf{R}^k$ is an k -dimensional state vector, \mathcal{M} is the attractor on which the dynamics evolve, and one has some *evolution operator* $\Phi : \mathcal{M} \times \mathbf{Z} \mapsto \mathcal{M}$ such that $\Phi(z_n, t) = z_{n+t}$. This situation is depicted in Fig. 1.1. Note that, in general one does not have to restrict $\mathcal{M} \subseteq \mathbf{R}^k$, but to do so does not really restrict our discussion either.

The dynamical system described by (\mathcal{M}, Φ) is said to be *deterministic* if the evolution operator Φ is deterministic. In other words, if one can write down some mathematical rule by which the future state z_{n+t} can be determined precisely from the current state z_n at any instance n for

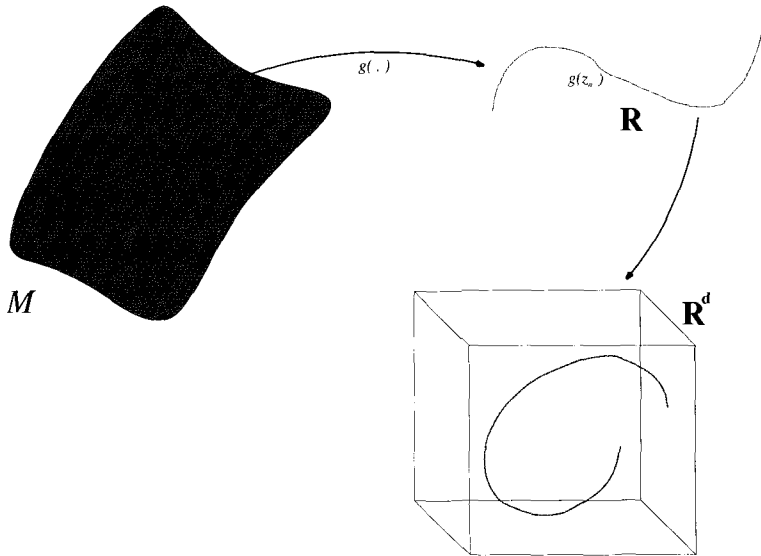


Fig. 1.1 **Evolution on a manifold.** The system state z_n exists within some manifold \mathcal{M} and evolves according to Φ . This dynamic is observed in \mathbf{R} by some C^2 observation function g , and, hence delay construction is possible. The delay reconstruction in \mathbf{R}^d is, in some sense, equivalent to the original trajectory.

some value of $t > 0$, then that rule, Φ , is said to be deterministic, and the dynamical system defined by that rule is also deterministic. We are currently only interested in discrete dynamical systems, and can therefore limit our discussion to $n, t \in \mathbf{Z}$.

Moreover, our definition insists that the evolution operator does not change with time (if $x_n = x_m$ then $\Phi(x_n, \cdot) = \Phi(x_m, \cdot)$ even if $m \neq n$); in this case, the dynamical system is *stationary*. Dynamical systems that are not stationary are exceedingly difficult to model from time series,¹ therefore we generally restrict our attention to the stationary case.

Notice that this definition of stationarity is not the same as for linear systems: a linear system is said to be stationary if all its moments² remain unchanged with time. However, restricting our attention to stationary systems can be easily justified by observing that the extent of the system is not

¹Unless one has *a priori* knowledge of the structure of the underlying system, the number of parameters will greatly exceed the number of available data.

²That is, the mean, standard deviation, kurtosis, skewness and all similar statistics concerning the distribution of the data.

necessarily constrained. One usually chooses the system $\mathcal{M} \subseteq \mathbf{R}^k$ to be the smallest system (lowest k) such that the corresponding $\Phi : \mathbf{Z} \times \mathcal{M} \mapsto \mathcal{M}$ is stationary. A non-stationary system is one which is subject to temporal dependence *based on some outside influence*. If we extend our definition of the system to include all outside influences, the system is stationary. In other words, an evolution operator Φ is not stationary, a new evolution operator $\hat{\Phi}$ can be constructed which is stationary simply by increasing k appropriately.

For example, an observer standing on the beach³ observing the rise and fall of the tide could justifiably say that the state of the system representing the level of the tide is nonstationary: it is subject to external, and for a suitably ignorant observer, not entirely predictable effects. However, if the same observer included the relative positions and orientation of the earth, moon and sun in their system, then, they would conclude that taken together they represent an (approximately) stationary system.

Now, let us suppose that we have a dynamical system which is both stationary and deterministic. We must now examine this system. Suppose that we can measure some single scalar quantity at any time. That is, we have an *observation function* $g : \mathcal{M} \mapsto \mathbf{R}$. This observation function provides us with a way to measure the current state of the system $g(z_n)$. Since $g(\cdot)$ gives us only a scalar value, it cannot offer a complete description of the system. But, observing $x_n = g(z_n)$ at many successive times will.

According to the celebrated Takens' embedding theorem [144], if d_e is sufficiently large, the evolution of $(x_n, x_{n-1}, x_{n-2}, \dots, x_{n-d_e})$ will be the same as z_n .

The argument for this theorem is topological, but it is easy to understand on an intuitive level. Suppose I can only measure one variable of a system. Then, except for one-dimensional systems, this is insufficient to describe the underlying dynamics. However, suppose that I can also measure the derivative of that variable, and further higher order derivatives up to some finite level d . Then, if the dimension of the system is less than d , I have enough information to completely describe the system: a system of d differential (or difference) equations.

But, for sufficiently high sampling rate, measuring d derivatives is equivalent to measuring the system at d different time intervals. Moreover, the embedding

$$x_n \longrightarrow (x_n, x_{n-1}, x_{n-2}, \dots, x_{n-d_e}) \quad (1.1)$$

³As Sir Isaac Newton did (when he wasn't standing on giants).

contains the same information as the original system: provided d_e is large enough, the measurement function g is twice differentiable,⁴ there is a sufficiently long data record available (for the comparison to the original system to be meaningful) and the data are sampled sufficiently often [144].

In practice, these conditions are very difficult to achieve. At the very least, digitisation (both sampling and quantisation) of the data represents a breach of the differentiability of g . Moreover, our informal argument concerning the differentiation of the system state by taking successive observations is substantially weakened in the presence of noise. Higher order derivatives are notoriously difficult to obtain numerically [95]. However, practitioners will presume that the conditions hold approximately; and the data embedded by Eq. (1.1) approximates the topology of the underlying attractor \mathcal{M} ; and the sequence of embedded points behaves under a deterministic rule approximately equivalent to the evolution operator Φ .

All of this approximation may not sit well with the theorist, but it is best to bear in mind that the motivation of nonlinear time series analysis is based firmly on these slightly faulty assumptions. One cannot necessarily achieve a perfect embedding (quantisation of data and finite sampling time violate the sufficient conditions of Takens' theorem), but we can still hope for a good one. But even our concept of a good embedding is limited by the purpose we have in mind [14]. To achieve a "good" embedding, the first parameter we need to estimate is the embedding dimension d_e .

1.2 Embedding dimension

Takens' embedding theorem [90; 144] and more recently work of Grebogi [26]⁵ and others [18; 6; 12; 83] give sufficient conditions on or suggest criteria for d_e . Unfortunately, the conditions require a prior knowledge of the fractal dimension of the object under study. In practice, one could guess a suitable value for d_e by successively embedding in higher dimensions and looking for consistency of results; this is the method that is generally employed.

In general, the aim of selecting an embedding dimension is to make sufficiently many observations of the system state so that the deterministic state of the system can be resolved unambiguously.⁶ Most methods to esti-

⁴That is, noise, or even ordinary digital quantisation are technically not allowed.

⁵Grebogi gives a sufficient condition on the value of d_e necessary to estimate the correlation dimension of an attractor, not to avoid all possible self intersections.

⁶It is best to remember that in the presence of observational noise and finite quantisa-

mate the embedding dimension aim to achieve unambiguity of the system state. The archetype of many of these methods is the so-called false nearest neighbour technique [32; 147].

1.2.1 False Nearest Neighbours

Suitable bounds on d_e can be deduced by using false nearest neighbour analysis [64]. The rationale of false nearest neighbour techniques is the following. One embeds a scalar time series y_t in increasingly higher dimensions, at each stage comparing the number of pairs of vectors v_t and v_t^{NN} (the nearest neighbour of v_t) which are close when embedded in \mathbf{R}^n but not close in \mathbf{R}^{n+1} . Each point

$$v_t = (y_{t-\tau}, y_{t-2\tau}, \dots, y_{t-n\tau})$$

has a nearest neighbour

$$v_t^{NN} = (y_{t'-\tau}, y_{t'-2\tau}, \dots, y_{t'-n\tau}).$$

When one has a large amount of data, the distance (Euclidean norm will do) between v_t and v_t^{NN} should be small. If these two points are genuine neighbours, they became close due to the system dynamics and should separate (relatively) slowly. However, these two points may have become close because the embedding in \mathbf{R}^n has produced trajectories that cross (or become close) due to the embedding rather than the system dynamics.⁷ For each pair of neighbours v_t and v_t^{NN} in \mathbf{R}^n , one can increase the embedding dimension by one so that

$$\hat{v}_t = (y_{t-\tau}, y_{t-2\tau}, \dots, y_{t-n\tau}, y_{t-(n+1)\tau})$$

and

$$\hat{v}_t^{NN} = (y_{t'-\tau}, y_{t'-2\tau}, \dots, y_{t'-n\tau}, y_{t'-(n+1)\tau})$$

tion this is not possible. Moreover, it has been shown that even with perfect observations over an arbitrary finite time interval, a “correct” embedding will still yield a set of states indistinguishable from the true state [58].

⁷The standard example is the embedding of motion around a figure 8 in two dimension. At the crossing point in the centre of the figure, trajectories cross. However, one can imagine if this was embedded in three dimensions, then these trajectories may not intersect.

may or may not still be close. The increase in the distance between these two points is given only by the difference between the last components

$$\|\widehat{v}_t - \widehat{v}_t^{NN}\|^2 - \|v_t - v_t^{NN}\|^2 = (y_{t-(n+1)\tau} - y_{t'-(n+1)\tau})^2.$$

One will typically calculate the normalised increase to the distance between these two points and determine that two points are false nearest neighbours if

$$\frac{|y_{t-(n+1)\tau} - y_{t'-(n+1)\tau}|}{\|v_t - v_t^{NN}\|} \geq R_T. \quad (1.2)$$

A suitable value of R_T depends on the spatial distribution of the embedded data v_t . If R_T is too small, true near neighbours will be counted as false, if R_T is too large, some false near neighbours will not be included. Typically $10 \leq R_T \leq 30$, for convenience we find that $R_T = 15$ is a good starting point. One must ensure that the chosen value of R_T is suitable for the spatial distribution of the data under consideration — this may be done by trialling a variety of values of R_T . By determining if the closest neighbour to each point is false, one can then calculate the proportion of false nearest neighbours for a given embedding dimension n .

We can then choose as the embedding dimension d_e the minimum value of n for which the proportion of points which satisfy the above condition is below some small threshold. Typically one could expect the proportion of points satisfying this to gradually decrease; as the embedded data is “unfolded” in an increasing embedding dimension; and, eventually that proportion plateaus at a relatively low level.

1.2.2 False strands and so on

The idea in the previous section can easily be extended to consider trajectories, rather than simply successive points. Whereas, the method of false nearest neighbour relies on measuring the divergence of nearby *points* after a short time, the method of false strands [64] measures whether *trajectories* which cross continue to stay relatively close over their entire length. The difference is a subtle but important one.

When computing false nearest neighbours, one runs into a problem as the embedding dimension increases. The threshold, by which we determine whether a neighbour is false becomes insufficient, and one can readily observe that even white noise exhibits (according to the false nearest neighbour technique) a relatively low embedding dimension [63]. To overcome

this, the method of false nearest neighbour replaces Eq. (1.2) with the condition

$$\frac{\|\widehat{v}_t - \widehat{v}_t^{NN}\|}{R_A} \geq R_T, \quad (1.3)$$

where

$$R_A = \left(\frac{1}{N} \sum_{i=1}^N [y_i - \bar{y}]^2 \right)^{\frac{1}{2}}$$

and \bar{y} denotes the mean value of the data [63].

1.2.3 *Embed, embed and then embed*

One further approach that is worth mentioning, at least because it is commonly used in practice, we refer to as “the doctrine of ever increasing embedding dimension”. The idea goes something like this: embed the data in increasing dimensions until one observes consistency. Usually, one will choose increasing embedding dimensions, and in each case measure the correlation integral.⁸ When we no longer observe changes in the behaviour of the correlation integral with increasing dimension, we have found a sufficiently large embedding dimension.

Although, consistency will not necessarily imply correctness, the rationale of this approach is that the correlation integral (or any other dynamic invariant) should be independent of the embedding dimension provided that it is sufficiently large. Therefore, when one observes no dependence on the embedding dimension, then this must be a sufficiently large dimension. While this approach may work in practice the fallacy of this logic is that if the sufficient conditions of Takens’ theorem are not satisfied (either the data set is too short, too noisy, has been digitised, or is not even finite dimensional), no embedding dimension may necessarily be acceptable. However, for a finite length data set, with some noise, one will almost always observe consistent behaviour provided that the dimension is sufficiently large. This consistency is simply a result of embedding in an ever increasing dimension. Eventually the embedding dimension will become sufficiently large so that increasing it by one does not appreciable change the properties of the data (i.e. as $n \rightarrow \infty$, $n + 1 \approx n$).

⁸Correlation dimension and other dynamic invariants are discussed in more detail in Chapter 2.

A further trap one may often observe from this approach is that the dynamic invariants being estimated may not behave exactly as expected. As the embedding dimension increases, the distribution of points in space changes. For large embedding dimensions, one observes the counter-intuitive fact that the majority of a volume is located in a thin shell on the exterior (it has a virtually empty interior). Many correlation dimension estimation algorithms do not correct this and one can observe increasing estimates of correlation dimension even after the minimal suitable choice of embedding dimension has been exceeded.

1.2.4 *Embed and model, and then embed again*

The approaches for selecting the embedding dimension described in the previous section aim to achieve an unambiguous reconstruction of the state of the dynamical system. An alternative, less heavily exploited approach is to select the embedding dimension which achieves the best model of the underlying dynamics. In Sec. 1.6.1, we will return to this idea and describe an efficient implementation that utilises information theory to evaluate the quality of the model. In [7] a similar, although somewhat different approach is described.

If one supposes that the object of embedding (and selecting the embedding dimension) is to reconstruct the underlying dynamics, one should choose the embedding dimension that corresponds to the best model of the underlying system. Unfortunately, this requires one to choose a model class, select a model from within that class and then to determine which model is best.

Ataei [7] and co-workers achieve this by restricting themselves to the class of polynomial models and assessing model quality based on model prediction errors. However, it is known that the class of polynomial models is often a poor choice for nonlinear systems [17; 127]. Of course, selecting wider model classes involves proportionately more work.

We will describe a solution to this problem in the coming sections, and also a more robust way of assessing the quality of a model, using an information theoretic criterion.

1.3 Embedding lag

In the previous sections, we described several distinct methods for choosing the embedding dimension, and we assumed that the selection of embedding lag was straightforward (or at they very least, was not a problem). In theory, any value of τ is acceptable, but the shape of the embedded time series will depend critically on the choice of τ and it is wise to select a value of τ which separates the data as much as possible. Typically, one is concerned with the evolution of the dynamics in phase space. By ensuring that the data are maximally spread in phase space, the vector field will be maximally smooth. Spreading the data out minimises possibly sharp changes in direction amongst the data. From a topological view-point, spreading data maximally, makes fine features of phase space (and the underlying attractor) more easily discernible.

In his thorough book, Abarbanel [1] suggests the first minimum of the mutual information criterion [96; 103]; and, the first zero of the autocorrelation function [97] or one of several other criteria to choose τ . Our experience and numerical experiments suggest that selecting a lag approximately equal to one quarter of the approximate period of the time series produces comparable results to the autocorrelation function but is more expedient. Note that the first zero of the autocorrelation function will be approximately the same as one quarter of the approximate period if the data are very strongly periodic.⁹ In the following subsections, we review each of these popular methods.

1.3.1 Autocorrelation

Define the sample autocorrelation of a scalar time series y_t of N measurements to be

$$\rho(T) = \frac{\sum_{n=1}^N (y_{n+T} - \bar{y})(y_n - \bar{y})}{\sum_{n=1}^N (y_n - \bar{y})^2}$$

where $\bar{y} = \frac{1}{N} \sum_{n=1}^N y_n$ is the sample mean. The smallest positive value of T for which $\rho(T) \leq 0$ is often used as embedding lag. Data which exhibits a strong periodic component suggests a value for which the successive coordinates of the embedded data will be virtually uncorrelated whilst still being close (temporally). We stress that a choice of $\tau = T$ such that the

⁹Moreover, in some specific cases, such as the Lorenz system, the autocorrelation is strictly positive, but the approximate period is still well defined, and easy to determine.

sample autocorrelation is zero is purely prescriptive. Some simple systems produce poor results for this choice, and, as an alternative, some authors recommend choosing the lag such that autocorrelation has first dropped below $\frac{1}{e}$ (the so-called *decorrelation time*). Sample autocorrelation is only an estimate of the autocorrelation of the underlying process, but is sufficient for estimating time lag.

1.3.2 Mutual information

A competing criterion relies on the information theoretic concept of mutual information, the mutual information criterion (MIC). In the context of a scalar time series, the information $I(T)$ can be defined by

$$I(T) = \sum_{n=1}^N P(y_n, y_{n+T}) \log_2 \frac{P(y_n, y_{n+T})}{P(y_n)P(y_{n+T})},$$

where $P(y_n, y_{n+T})$ is the probability of observing y_n and y_{n+T} , and $P(y_n)$ is the probability of observing y_n . $I(T)$ is the amount of information we have about y_n by observing y_{n+T} , and so one sets τ to be the first local minima of $I(T)$. The primary difficulty with estimating mutual information is that one must first estimate a probability distribution on the systems states (and, moreover, do this on vector states). Inappropriate selection of the histogram binning can easily lead to poor results: the issue of density estimation is a well-established field in its own right [118]. However, provided one can obtain a good estimate of the underlying density, the mutual information will usually provide a good guide to appropriate choice of embedding lag [60].

1.3.3 Approximate period

The rationale of these previous two methods is to choose the lag so that the coordinate components of v_t are reasonably uncorrelated while still being “close” to one another. When the data exhibits strong periodicity, a value of τ that is one quarter of the length of the average breath generally gives a good embedding. This lag is approximately the same as the time of the first zero of the autocorrelation function. Coordinates produced by this method are within a few cycles of each other (even in relatively high dimensional embeddings) whilst being spread out as much as possible over a single period. Moreover, for embedding in three or four dimensions

(as is commonly used in certain widely studied systems¹⁰ the data are spread out over one half to three quarters of a period. This means that the coordinates of a single point in the three or four dimensional vector time series v_t represents most of the information for an entire cycle. This choice of lag is extremely easy to calculate and for the data sets that we consider it also seems to give much more reliable results than the mutual information criterion.

1.3.4 Generalised embedding lags

A generalisation of the ideas presented in the previous sections has been described by Luo [79]. In [79] the authors describe the *redundance and irrelevance tradeoff exponent (RITE)* which is a general technique for choosing embedding lag. In general, one considers two competing criteria for the selection of embedding lag: (i) the lag must be large enough so that the various co-ordinates contain as much new information as possible, and (ii) the lag must be small enough so that the various co-ordinates are not entirely independent. We can see this principle in each of the previous ideas. With AMI, we choose the first minimum of mutual information, as the first minimum gives the least redundant (i.e. mutual) information as possible without being too large. Similarly, with the first zero of autocorrelation, the co-ordinates are linearly uncorrelated, and yet not too far apart.

Let us generalise this idea as follows. Let $R(\tau)$ measure the *redundance* between a time series $\{x_n\}$ and the time series $\{x_{n+\tau}\}$, and let $I(\tau)$ measure the *irrelevance*. Then, given some weight k , we define the RITE as

$$kR(\tau) + (1 - k)I(\tau).$$

For example, we may consider (as Luo does [79]), the case where redundance and irrelevance are measured by second order autocorrelation $\rho(\tau)$ and the weights assigned to the two measures are $\langle x_n^2 \rangle$ and $\langle x_n \rangle^2$. In other

¹⁰From examining the literature, it may seem that most chaotic systems which are actually studied, can be embedded in no more than three or four dimensions. We speculate that, while three or four dimensions is necessary to observe chaotic flows [75], higher dimensional systems are often simply too difficult to study in practice.

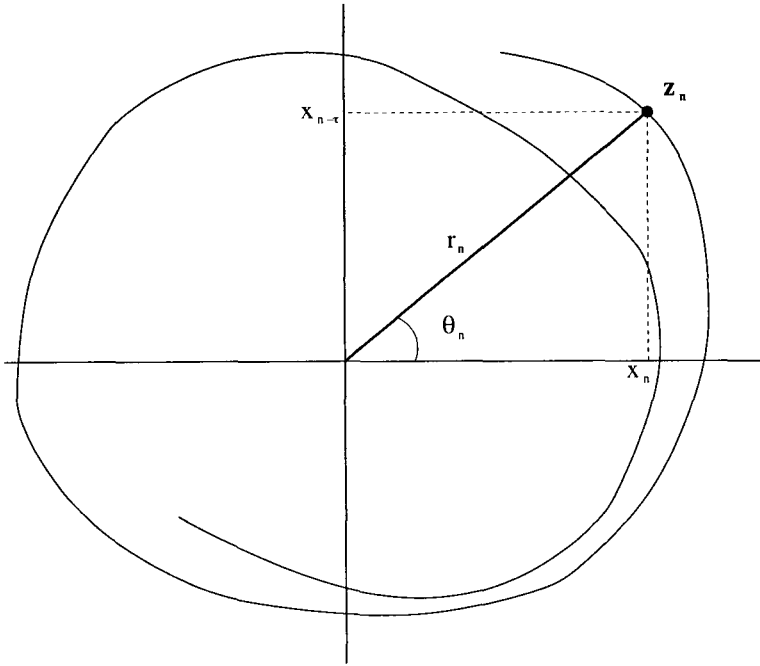


Fig. 1.2 **Embedding in two dimensions.** The vector point z_n can be decomposed into its polar form consisting of the radius r_n and the angle θ_n . The distance of the point z_n from the diagonal line is proportional to $|x_{n+\tau} - x_n|$ and the projection of z_n onto that line is proportional to $|x_{n+\tau} + x_n|$.

words,

$$\begin{aligned}
 R(\tau) &= \rho(\tau) \\
 I(\tau) &= 1 - \rho(\tau) \\
 k &= \frac{\langle x_n^2 \rangle}{\langle x_n^2 \rangle + \langle x_n \rangle^2} \\
 1 - k &= \frac{\langle x_n \rangle^2}{\langle x_n^2 \rangle + \langle x_n \rangle^2} \\
 kR(\tau) + (1 - k)I(\tau) &= \frac{\rho(\tau) \langle x_n^2 \rangle + (1 - \rho(\tau)) \langle x_n \rangle^2}{\langle x_n^2 \rangle + \langle x_n \rangle^2}.
 \end{aligned}$$

In fact, in this special case, the RITE reduces exactly to the standard second order autocorrelation. A more interesting case is that, by considering a two dimensional embedding, we may examine the distance from the diagonal

$d_n \propto |x_{n+\tau} - x_n|$, the projection onto that identity line $p_n \propto |x_{n+\tau} + x_n|$, and the angle subtended from that line $\theta_n \propto \tan^{-1} \left| \frac{x_{n+\tau} - x_n}{x_{n+\tau} + x_n} \right|$. Fig. 1.2 depicts the general situation. Now, each of these measures can be employed in Eq. (1.4) in the place of x_n , and one now obtains a nontrivial, unique, nonlinear measure of redundancy and irrelevance. In computational studies [79], it can be seen that these measures perform as well as the competing criteria, and often provide more robust estimates of appropriate embedding lags.

1.4 Which comes first?

Clearly estimating embedding dimension requires one to first estimate the embedding lag. But the value of embedding lag selected is, at least implicitly, dependent on the embedding dimension. Although most of the techniques for selecting the embedding lag appear to be independent of embedding dimension, this is not the case.

By selecting one-quarter of the pseudo-period as the embedding lag (or, almost equivalently, the first minimum of mutual information or the first zero of autocorrelation), one implicitly supposes that the embedding dimension is approximately 4. The heuristic motivation for selecting embedding lags according to these criteria (for data with a strong periodic component) is to have each single embedded point representative of the dynamics over an entire period.

Although it is common practice to estimate the embedding lag first, then choose embedding dimension, we prefer an alternative approach. Provided the data is “clean enough”, starting with an embedding lag of 1 and estimating the embedding dimension will give a good guide to the number of degrees of freedom. From this, one can choose an embedding lag. For most low-dimensional (i.e. 3–5 degrees of freedom) systems, an embedding lag estimated with one of the methods described in the previous section will suffice. If the embedding dimension does not fall neatly into the range 3–5, an alternative value of embedding lag may be better. Having selected embedding lag, one can go back and check that the embedding dimension, using this chosen value of embedding lag, is appropriate.

Alternatively, one may start by estimating the embedding window (Sec. 1.7), then determine the most appropriate irregular embedding (Sec. 1.6). Before describing this alternative, we believe, superior approach, we present some illustrations of selection of embedding lag and embedding dimension for various sets of data.

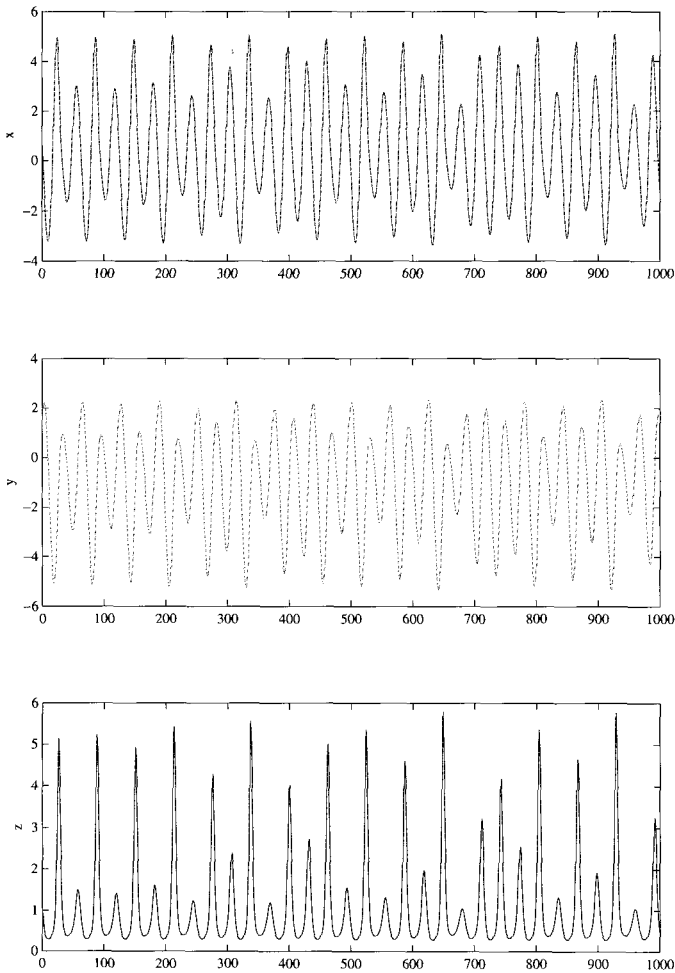


Fig. 1.3 Typical time series for the Rössler dynamical system. x , y and z components (from top to bottom) of the three dimensional Rössler system (1.4) integrated (and sampled) with a time step of 0.2.

1.5 An embedding zoo

Let us pause now, and present several archetypal dynamical systems and some of our favourite data sets.

Two examples of *continuous* dynamical systems that are stationary, and

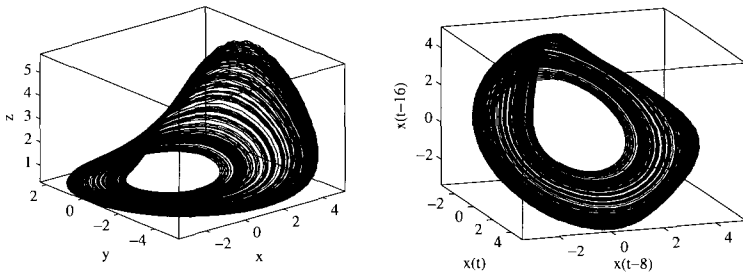


Fig. 1.4 **Rössler attractor and reconstruction of the attractor.** The left panel is a single 5000 point trajectory plotted in x , y and z co-ordinates, the right hand panel is a delay embedding in three dimension (embedding lag of 8) of the x co-ordinate. One can see from both original and embedded co-ordinates that chaos in this system is generated by a gradual stretching apart of trajectories over most of the attractors, combined with rapid folding and compressing at one point.

in all other respects satisfy the above conditions. The Rössler equations [107] are given by

$$\begin{aligned}\dot{x} &= -y - z, \\ \dot{y} &= x + ay, \\ \dot{z} &= b + z(x - c)\end{aligned}\tag{1.4}$$

where, for $a = 0.398$, $b = 2$ and $c = 4$ the system exhibits “single-band” chaos [152]. Figure 1.3 depicts a discrete sampling with a sample step size¹¹ of 0.2 of a typical time series of the x , y and z co-ordinates. In Fig. 1.4, we have produced the attractor of this system (left panel) by plotting the x , y , and z components against one another to illustrate a single trajectory. The right hand panel shows a reconstruction of this system from the x component alone.

The Lorenz system [77] is defined by

$$\begin{aligned}\dot{x} &= s(y - x), \\ \dot{y} &= rx - y - xz, \\ \dot{z} &= xy - bz\end{aligned}\tag{1.5}$$

where ($s = 10$, $r = 28$, $b = 8/3$) yields chaotic dynamics [152]. In Figs. 1.5 and 1.6 we depict the dynamics of a discrete sampling of the Lorenz system

¹¹That is, the equations are numerically integrated with an integration time step of 0.2.

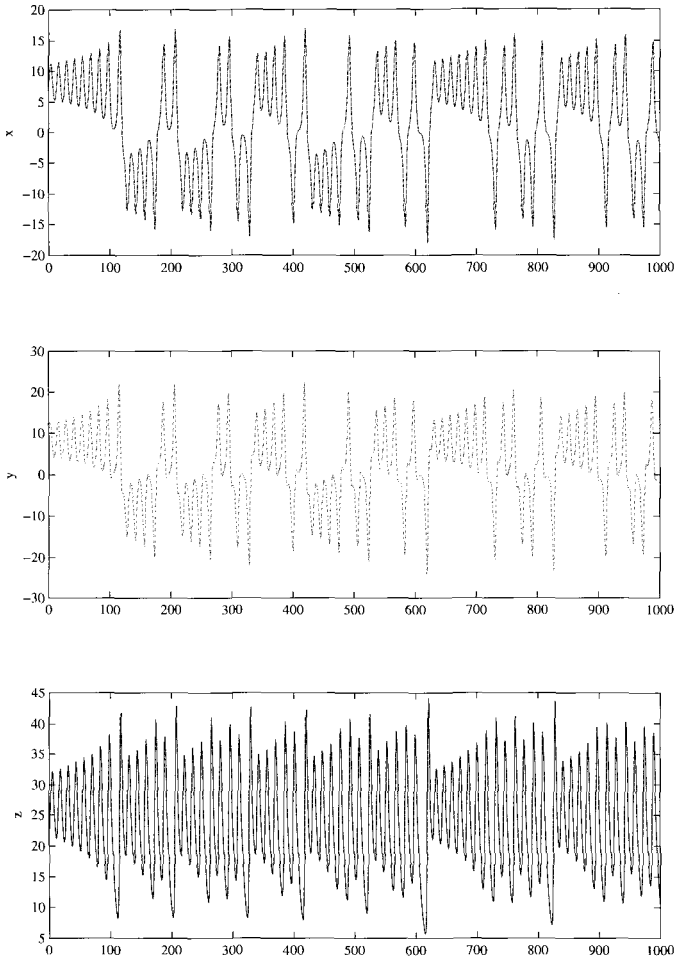


Fig. 1.5 **Typical time series for the Lorenz dynamical system.** x , y and z components (from top to bottom) of the three dimensional Lorenz system (1.5) integrated (and sampled) with a time step of 0.05.

and typical reconstructions. Throughout the remainder of this book we will frequently refer to these two systems.

Furthermore, let us introduce two discrete dynamical systems (maps) that exhibit chaos and are also widely studied. Perhaps the simplest of all

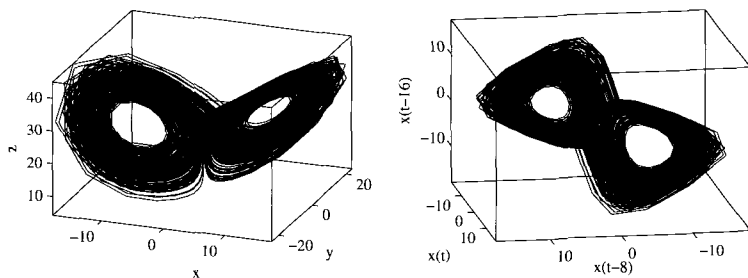


Fig. 1.6 **Lorenz attractor and reconstruction of the attractor.** The left panel is a single 5000 point trajectory plotted in x , y and z co-ordinates, the right hand panel is a delay embedding in three dimensions (embedding lag of 3) of the x co-ordinate. Notice that the original attractor (and to a lesser extent, when viewed from this angle, the reconstructed one) exhibits two flat (two dimensional wings) and a more complex central region. The dynamics on the wings is relatively simple, only at the central separatrix do the complex crossings and splittings that generate chaos in this system occur.

is the Logistic map [61] given by

$$x_{n+1} = \mu x_n(1 - x_n) \quad (1.6)$$

where, for $\mu > 4$ the system is chaotic. In Fig. 1.7, we see a short trajectory of the Logistic map together with its attractor in x_n -vs- x_{n+1} co-ordinates (note that because this system is one dimensional, delay reconstruction is both unnecessary and trivial). However, the representation of successive values of the logistic map is widely used to provide a complete description of the underlying dynamics (see for example [152]).

A less trivial map (and one with a more attractive attractor) is the Ikeda map defined by

$$\begin{aligned} x_{n+1} &= 1 + 0.7(x_n \cos \theta_n - y_n \sin \theta_n), \\ y_{n+1} &= 0.7(x_n \sin \theta_n + y_n \cos \theta_n), \\ \theta_n &= 0.4 - \frac{6}{1 + x_n^2 + y_n^2}, \end{aligned} \quad (1.7)$$

where, for the parameter values given, the dynamics are chaotic. In Fig. 1.8, we show a short segment of the time series together with its attractor and delay reconstruction (from the x_n component). Notice that the delay reconstruction here is non-trivial, and, even for a “good”¹² choice of em-

¹²And, obvious.

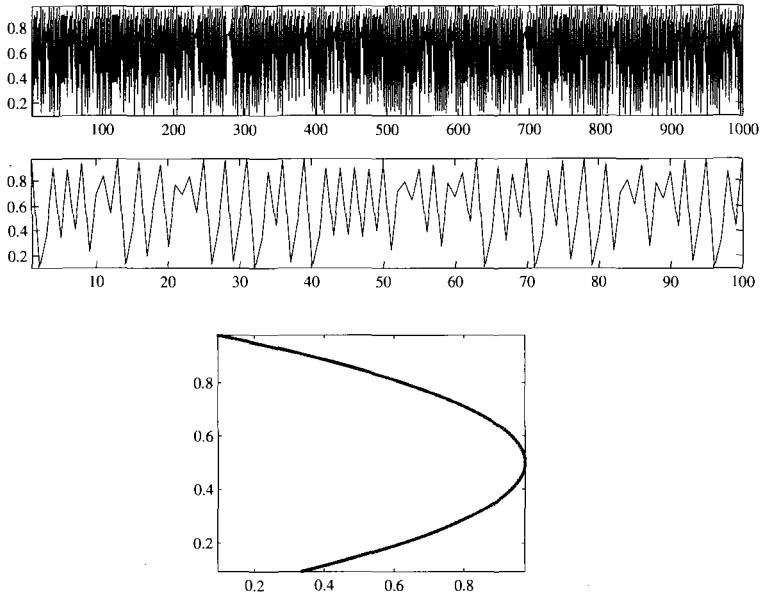


Fig. 1.7 **Typical time series and attractor for the Logistic map system.** The top two panels show the time evolution of the logistic map system (1.6) on different time scales. The bottom panel is the first return map (aka the time delay embedding with lag 1 in two dimensions).

bedding parameters, there is significant convolution in the reconstructed attractor.

Several experimental data sets which we constantly find both useful and interesting are illustrated in Fig. 1.9. In Fig. 1.10, we show the effect of embedding the first data set with various embedding lags in two dimensions. In Fig. 1.11, we can see that no low dimensional embedding provides a satisfactory reconstruction of the ECG data. Conversely, in Fig. 1.12, we show the laser data successfully reconstructed in three dimensions.

1.6 Irregular embeddings

The methods to estimate d_e and τ described in the previous sections assume that a single embedding lag is sufficient and having chosen d_e and τ , the embedding defined by

$$x_n \longrightarrow (x_n, x_{n-\tau}, x_{n-2\tau}, \dots, x_{n-(d_e-1)\tau}) \quad (1.8)$$

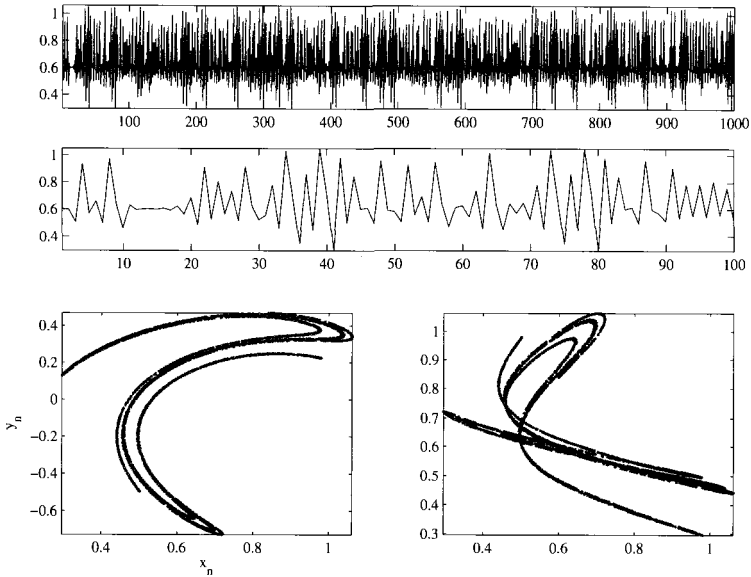


Fig. 1.8 Typical time series and attractor for the Ikeda map system. The top two panels show the time evolution of the Ikeda map system (1.7), the lower two panels are the plot of x_n against y_n (left panel) and a delay one embedding of the first co-ordinate (i.e. x_n against x_{n-1} , right panel).

is adequate. For estimating dynamic invariants, this may well be the case,¹³ however, if one is concerned with recreating the underlying dynamics and estimating the evolution operator (i.e. nonlinear modelling), there is no reason to suppose that this would be the case. In the first part of this book, we are concerned primarily with the estimation of dynamic invariants, and therefore irregular embeddings are not necessary. However, later we will be highly interested in estimating the underlying evolution operator (i.e. modelling); in this context an irregular embedding can be invaluable.

In [55], Judd and Mees describe the embedding (1.8) as a *uniform embedding* and propose a *nonuniform embedding* according to

$$x_n \longrightarrow (x_{n-l_1}, x_{n-l_2}, x_{n-l_3}, \dots, x_{n-l_{d_e}}) \quad (1.9)$$

where the parameters $0 \leq l_1 \leq l_i < l_{i+1} \leq l_{d_e}$ are the *embedding lags*. Collectively, the problem is now finding the correct set of lags $(l_1, l_2, \dots, l_{d_e})$.

¹³I.e. to the best of my knowledge there is no evidence to the contrary.

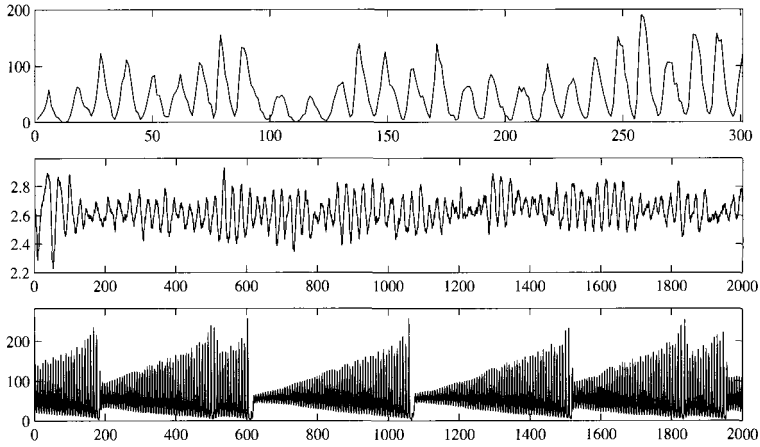


Fig. 1.9 **Experimental time series data.** The two experimental time series examined in this application are depicted (from top to bottom): annual sunspot numbers for the period 1700 to 2000, a recording of ECG (electrocardiogram) activity during Ventricular Fibrillation, and the chaotic “Santa-Fe” laser times series. For the bottom panel only the first 2000 points are utilised for time series modelling.

However, as Judd and Mees point out [55], even this may not be sufficient. Often, one encounters dynamical systems where the important variables are different in different parts of phase space. Another way of describing this is to say that the embedding is not constant. For the Lorenz attractor (see Fig. 1.6), the “wings” are two dimensional, and a two dimensional embedding is sufficient, however, at the central separatrix three dimensions are required. Such non-constant and potentially non-uniform embeddings are called *variable embeddings*.

1.6.1 Finding irregular embeddings

The full problem of finding the correct (i.e. “best”, or even just “good”) embedding for a particular time series will depend entirely on the model and the model selection scheme one chooses to employ. We will consider these problems in more detail much later in this book. For now we would just like to demonstrate how one may practically obtain variable embeddings (1.8) that behave quantitatively better than uniform embeddings.

The problem of estimating the optimal embedding can be stated as: find the parameters $\{\ell_i | i = 1, \dots, k\}$ and the embedding window d_w (to be

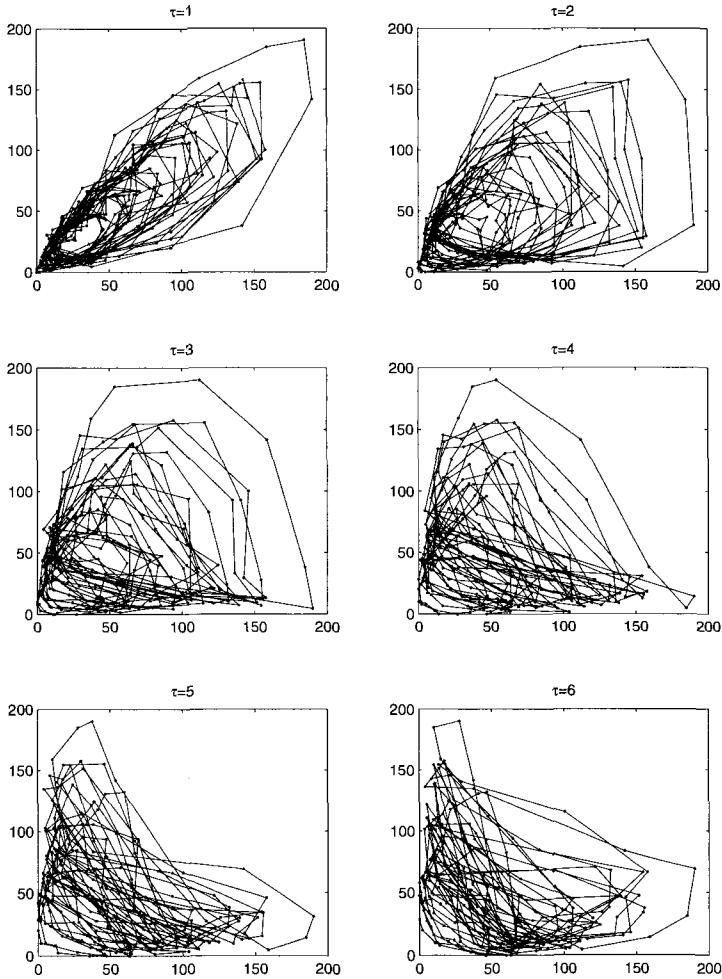


Fig. 1.10 **Embedding of the sunspot data.** Six embeddings in two dimensions, with embedding lag from 1 to 6 are shown for the sunspot time series depicted in Fig. 1.9. The upper panels show that a small embedding lag (i.e. 1) allows a delay reconstruction with a central hole (for pseudo-periodic dynamics such as this, it is reasonable to presume that the central region will exhibit an unstable focus). However, larger embedding lags (i.e. 2 and 3) show a better reconstruction of the amplitude of oscillations. These two main features (oscillations about a central point, and amplitude modulation) are evident, but in different embeddings. We will revisit this problem later when we see that for a third data set a *non-uniform* embedding actually performs much better.

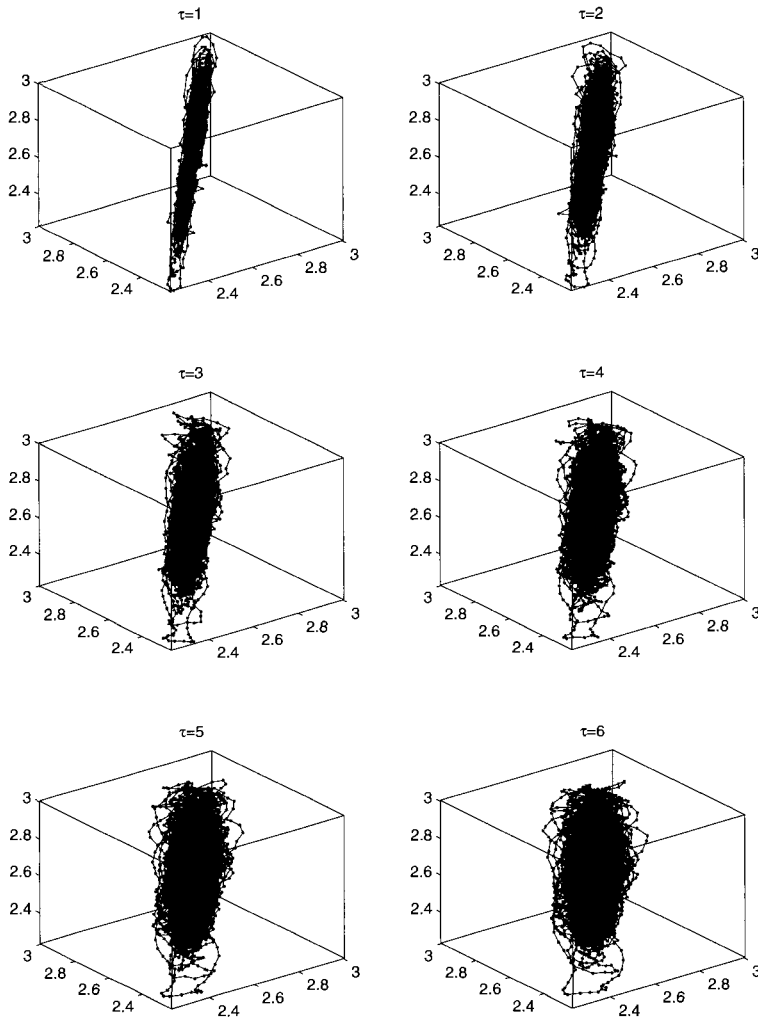


Fig. 1.11 **Embedding of the ECG VF data.** Six embeddings in two dimensions, with embedding lag from 1 to 6 are shown for the ECG data depicted in Fig. 1.9. The embeddings we achieve of this data are clearly very poor. The main feature of these embeddings is a complex (high dimensional or non-stationary) structure which is not adequately unraveled in three dimensions. Worse, for small embedding lags the data are tightly distributed along the identity line (too highly correlated).

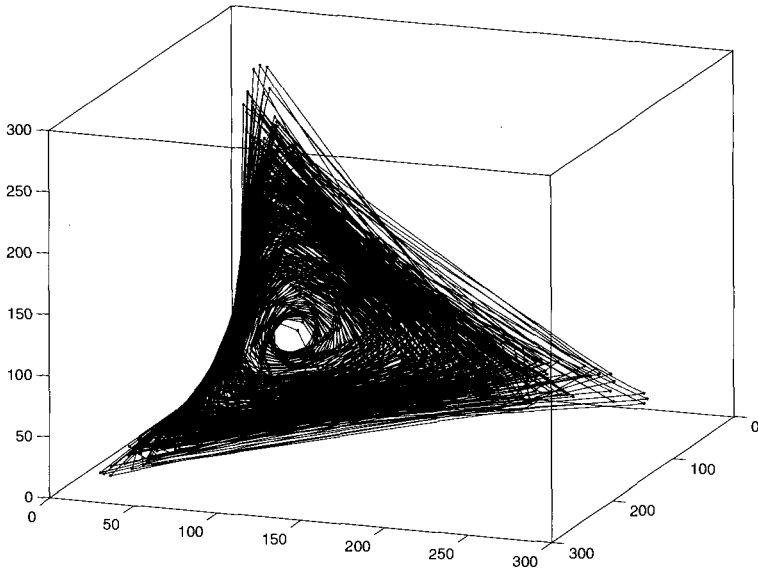


Fig. 1.12 **Embedding of the laser data.** An embedding in 3 dimensions with an embedding lag of 2 (the first zero of autocorrelation) for the laser data depicted in Fig. 1.9. The smooth geometric structure of this system is evident from this simple low-dimensional embedding. Moreover, this indicates that the dynamics of this system can be described and modelled within only three dimensions.

discussed in more detail in Sec. 1.7), where $1 \leq \ell_1 \leq \ell_i < \ell_{i+1} \leq \ell_k \leq d_w$ and the time delay embedding

$$x_t \rightarrow (x_{t-\ell_1}, x_{t-\ell_2}, x_{t-\ell_3}, \dots, x_{t-\ell_k}) \quad (1.10)$$

is somehow the “best”.

Unfortunately, application of Eq. (1.10) makes the problem of selecting embedding parameters considerably more complicated. Here, we describe one suitable criterion for quantitatively comparing embedding strategies and an efficient scheme for the computation of $\{\ell_i | i = 1, \dots, k\}$ and k . The, so-called *optimal embedding* strategy achieves results superior to the standard techniques (1.1) and (1.8). Employing this optimal embedding strategy may allow one to reconstruct the complex nonlinear dynamics of the underlying system more accurately than would otherwise be possible.

Often, the purpose of time delay embedding is to estimate correlation dimension [167] or other dynamic invariants [2] (to be described in Chapter

2). In such situations, embeddings such as (1.8) are usually adequate. But, what if one is interested in the more complex problem of estimating the underlying evolution operator of the dynamical system. Hence, we are interested in obtaining the most accurate prediction of the observed data values. By doing so, we hope to capture the long term dynamics of the underlying system. To achieve this we adopt the information theoretic measure *description length* [103] and seek to choose the embedding which provides the minimum description length.

We will revisit the topic of description length later in this book. Roughly speaking, the description length of a time series is the compression of the finite precision data afforded by the model of that data [53]. If a model is poor, it will be more economical to simply describe the model prediction errors. Conversely, if a model fits the data well, the description of that model and the (presumably small) model prediction errors will be more compact. However, if a model over-fits the data [130], the description of the model itself will be too large. In [132] (see Sec. 1.7) we showed that the description length $DL(\cdot)$ of a time series $\{x_t\}$ is approximated by

$$\begin{aligned} DL(\{x_t\}) \approx & \frac{N}{2} (1 + \ln 2\pi) + \frac{d}{2} \ln \left[\frac{1}{d} \sum_{i=1}^d (x_i - \bar{x})^2 \right] \\ & + \frac{N-d}{2} \ln \left[\frac{1}{N-d} \sum_{i=d+1}^N e_i^2 \right] \\ & + d + DL(d) + DL(\bar{x}) + DL(\mathcal{P}). \end{aligned} \quad (1.11)$$

where $d = \max_i \{\ell_i\} = \ell_{d_c}$, \bar{x} is the mean of the data, e_i is the model prediction error, and $DL(\mathcal{P})$ is the description length of the model parameters. The description length of an integer d can be shown to be $DL(d) = \lceil \log d \rceil + \lceil \log \lceil \log d \rceil \rceil + \dots$ where each term on the right is an integer and the last term in the series is 0 [103]. Furthermore, $\frac{N}{2} (1 + \ln 2\pi) + DL(\bar{x})$ is independent of the embedding strategy. Hence, the optimal embedding strategy is that which minimises

$$\begin{aligned} & \frac{d}{2} \ln \left[\frac{1}{d} \sum_{i=1}^d (x_i - \bar{x})^2 \right] + d + DL(d) + \\ & \frac{N-d}{2} \ln \left[\frac{1}{N-d} \sum_{i=d+1}^N e_i^2 \right] + DL(\mathcal{P}). \end{aligned} \quad (1.12)$$

The first three terms in (1.12) may be computed directly. However, the last

Table 1.1 **Embedding parameters for various data sets.** Embedding parameters for the various data sets described so far: data length (N), embedding dimension computed with the method of false nearest neighbours (d_e), embedding lag estimated by the first zero of autocorrelation (τ), optimal embedding window computed with the method described in Sec. 1.7 (d_w), and the optimal set of embedding lags (such that $(x_{t-\ell_1}, \dots, x_{t-\ell_k})$ is used to predict x_t). With the exception of the Lorenz system, τ is approximately one-quarter of the pseudo-period of the time series.

data	N	d_e	τ	d_w	d	ℓ_1, \dots, ℓ_k
Sunspots	301	6	3	7	10	1,2,5
Fibrillation	6000	6	7	2	10	1,5,6
Laser	4000	8	2	32	30	1, 2, 6, 7, 11, 14, 15, 21, 24, 25, 30
Rössler	1000	3	3	6	10	1, 5, 7
Rössler+noise	1000	5	3	9	10	1, 2, 4, 5, 6, 7, 9
Lorenz	1000	5	43	4	10	1, 3
Lorenz+noise	1000	5	42	8	10	1, 2, 3, 5, 6, 7, 9

two terms require one to estimate the optimal model.

As in [132], for the purposes of computational expediency, we restrict ourselves to the class of local constant models. In the current context this is not unreasonable as we hope to obtain an embedding which spreads the data in phase space based on the deterministic dynamic evolution. Under this assumption, $DL(\mathcal{P}) = 0$ and the model prediction error

$$\frac{1}{N-d} \sum_{i=d+1}^N e_i^2$$

may be computed via “drop-one-out” interpolation. That is, $e_{i+1} = y_{i+1} - y_{j+1}$ where $j \in \{1, 2, \dots, N\} \setminus \{i\}$ is such that $\|y_i - y_j\|$ is minimal. Note that, in the limit as $N \rightarrow \infty$ (i.e. $N \gg d$), optimising (1.12) is equivalent to finding the embedding which provides the best prediction (the last two terms of (1.12) dominate).

To minimise Eq. (1.12) we assume that the maximum number of inputs, d , has already been calculated. We choose $d = d_w$, the embedding window computed using the method described in [132]. Alternatively, one may either assign an arbitrary value for d or use $d = d_e \tau$ where both d_e and τ are estimated by one of the many standard techniques.

An exhaustive search on the 2^d possible embedding strategies is only feasible for small d . For large d (i.e. $d > 10$) we utilise a genetic algorithm [31] to determine the optimum embedding strategies. Furthermore, to reduce the computational effort in estimating the model prediction error for large N ($N > 1000$) we minimise the prediction error only on a randomly selected subset of the data. Our calculations show that neither of these

approximations adversely affects our results. The results of the genetic algorithm are robust and accurate, and the final solution is independent of the data subset selected.¹⁴

In Table 1.1, we demonstrate the application of this algorithm with data from three experimental systems (the famous annual sunspot time series, a chaotic laser [158], and a recording of human electroencephalogram during ventricular fibrillation [132; 134], these data are depicted in Fig. 1.9); and two computational simulations (Rössler and Lorenz equations, see Figs. 1.3 and 1.5) both with and without the addition of Gaussian observational noise with a standard deviation of 5% that of the data. For each data set, we estimated the embedding window d_w [132], the embedding dimension d_e (via false nearest neighbours) and the embedding lag τ (using the first zero of autocorrelation).

For each of these systems, we estimated the optimal embedding strategy using a genetic algorithm and (where necessary) the sub-sample selection scheme 30 times. All the data sets except the longest (the ECG recording and the laser system) produced identical results on repeated execution. For the two longest data sets, the most often observed embedding strategy was also the best (indicating that the sub-sample selection scheme is expedient but perhaps not always accurate). Table 1.1 also illustrates that, in most cases, the optimal embedding covered a smaller range of embedding lags than the standard method (i.e. $\ell_k < d_e \tau$) and is often of lower dimension ($k < d_e$). Perhaps intuitively, noisier time series required larger k . Furthermore, we note that in none of the cases was the optimal embedding strategy uniform.

Although time delay embedding is a fundamental technique for the reconstruction of nonlinear dynamical systems from time series, we can see here that it is not optimal, and that in general one should apply a non-uniform embedding such as (1.10). However, the problem with adopting this approach is that selection of the optimal embedding strategy becomes computationally intractable for even moderate d and N . The solution we propose here is a simple estimate of the nonlinear prediction error, and a combination of genetic algorithm and sub-sample selection. Undoubtedly more sophisticated and efficient methods for solving this NP-hard problem exist.

However, for a wide variety of experimental and simulated time series, the method we employ provides alternative embedding strategies which are

¹⁴Provided that the subset is selected *with* replacement and that it is moderately large.

Table 1.2 Comparison of correlation dimension estimates for the data and local constant model simulations using either the standard or optimal variable embedding strategy. We computed 30 simulations with either embedding strategy for each data set and report here the median, mean and standard deviation of the correlation dimension estimates (computed with the values d_e and τ listed earlier). The value of correlation dimension estimated from the time series data is also provided for reference.

data	correlation dimension				"true"
	standard		optimal		
	median	mean	median	mean	
Sunspots	2.095	3.043±4.957	2.245	2.191±0.4137	1.889
Fibrillation	1.467	1.464±0.04860	1.560	1.559±0.0387	1.619
Laser	2.212	2.205±0.116	2.478	2.448±0.190	2.129
Rössler	1.306	1.323±0.1281	1.337	1.338±0.0969	1.588
Rössler+noise	1.961	1.936±0.135	1.886	1.861±0.141	1.819
Lorenz	1.981	1.983±0.0789	1.861	1.860±0.0968	1.966
Lorenz+noise	1.642	1.644±0.0743	1.636	1.628±0.0576	1.612

often smaller ($k < d_e$ and $\ell_k < d_e\tau$) than standard methods, and perform at least as well. We have applied correlation dimension as a quantitative measure of the accuracy of dynamic reconstruction and find that the optimal embedding strategy described here produces models which behave more like the true data. Hence, the application of this embedding methodology will allow more accurate modelling of the underlying dynamics.

1.7 Embedding window

As we have seen, the celebrated theorem of Takens [144] guarantees that, for a sufficiently long time series of scalar observations of an n -dimensional dynamical system with a twice differentiable¹⁵ measurement function, one may recreate the underlying dynamics (up to homeomorphism) with a time delay embedding. Unfortunately the theorem is silent on exactly how to proceed when the data is limited or contaminated by noise. In practice, time delay embedding is routinely employed as a first step in the analysis of experimentally observed nonlinear dynamical systems (see [2; 60]). Typically, one identifies some characteristic embedding lag τ (usually related to the sampling rate and time scale of the time series under consideration) and utilises d_e lagged version of the scalar observable for sufficiently large d_e . In general, τ is determined by identifying linear or nonlinear temporal correlations in the data and one will progressively increase d_e until

¹⁵Technically, C^2 .

the results obtained are self consistent.

In the preceding section we described one possible (albeit computationally expensive) solution to the problem of finding the best irregular embedding for a particular time series. We now describe an alternative, simpler, approach to the same problem: the problem of reconstructing the underlying dynamics from a finite scalar time series in the presence of noise. We recognise that in general the quality of the reconstruction will depend on the length of the time series and the amount of noise present in the system. Employing the minimum description length model selection criteria, we show that the optimal model of the dynamics does not depend on the choice of the embedding lag, only on the maximum lag ($d_e\tau$ in Eq. 1.8). We call that maximum embedding lag $d_w := d_e\tau$, the embedding window, and show that for long noise-free time series, the optimal d_w minimises the one-step model prediction error. For short or noisy data, the optimal value of d_w is data dependent. To estimate the one-step model prediction error and d_w , we apply a generic local constant modelling scheme to several computational examples. We will return to this extremely useful modelling scheme in several different guises throughout this book.

This method of estimating d_w proves to be consistent and, even better, robust. Moreover, the results that we obtain capture the salient features of the underlying dynamics. Finally, we also find that in general there is no single characteristic time lag τ . Generically, the optimal reconstruction may be obtained by considering the lag vector

$$(\tau_1, \tau_2, \dots, \tau_k) \tag{1.13}$$

where $0 < \tau_i < \tau_{i+1} \leq d_w$.¹⁶

The textbooks [2; 60] and even the preceding sections of this volume, contain copious detail on the estimation of d_e and τ . We briefly revisit some developments relevant to the estimation of embedding window here.

Often, the primary aim of time delay embedding is to estimate dynamic invariants. In these instances, one may estimate τ with a variety of heuristic techniques: usually autocorrelation, pseudo-period or mutual information. One then computes the dynamic invariant for increasing values of d_e until some sort of plateau onset occurs (see [65] and the references therein). For estimation of correlation dimension, d_c , it has been shown that $d_e > d_c$ is sufficient [26]. However, for reconstruction of the underlying dynamics this is not the case. Alternatively, the method of false nearest neighbours [64]

¹⁶This is the so called “variable embedding” described in [55] and elsewhere.

and its various extensions apply a topological reasoning: one increases d_e until the geometry of the time series does not change.

We note that several authors have speculated on whether the individual parameters d_e and τ , or only their product $d_e\tau$, is significant. For example, Lai and Lerner [72] provide an overview of selection of embedding parameters to estimate dynamic invariants (in their case, correlation dimension). They impose some fairly generous constraints on the correlation integral and use these to estimate the optimal value of d_e and τ . Their numerical results from long clean data imply that correct selection of τ is crucial, selection of d_w (and therefore d_e) is not. Conversely, utilising the BDS statistic [11], Kim and co-workers [65] concluded that the crucial parameter for estimating correlation dimension is d_w .

Unlike these previous methods, the question we consider now is: “What is the optimal choice of embedding parameters to reconstruct the underlying dynamic evolution from a time series?” In answering this question we conclude that only the embedding window d_w is significant, selection of optimal embedding lags is, essentially, a modelling problem [55]. Clearly, successful reconstruction of the underlying dynamics will depend on one’s ability to identify any underlying periodicity (and therefore τ). These results show that it is possible to estimate the optimal value of d_w , and subsequently use this optimal value to derive a suitable embedding lag τ . However, as previous authors have observed in many examples, estimation of τ for nonlinear systems is model dependent [55] (and may even be *state dependent*) [55]).

In the following sub-section we introduce the rationale for the calculations that follow. Section 1.7.2 demonstrates the application of this method to several test systems, and Application 1.8 studies the problem of modelling several experimental time series.

1.7.1 A modelling paradigm

As before, let $\phi : \mathcal{M} \rightarrow \mathcal{M}$ be the evolution operator of a dynamical system, and $h : \mathcal{M} \rightarrow \mathbf{R}$ a C^2 differentiable observation function. Through some experiment we obtain the time series $\{h(X_1), h(X_2), \dots, h(X_N)\}$. Denote $x_i \equiv h(X_i)$. Takens’ theorem [144] states that for some $m > 0$ the mapping g

$$x_i \xrightarrow{g} (x_i, x_{i-1}, x_{i-2}, \dots, x_{i-m-1}) \quad (1.14)$$

is such that the evolution of $g(x_i) = (x_i, x_{i-1}, \dots, x_{i-m-1})$ ¹⁷ is homeomorphic to ϕ

We will generalise the embedding map (1.14) and consider \hat{g} as

$$x_i \xrightarrow{\hat{g}} (a_1 x_i, a_2 x_{i-1}, a_3 x_{i-2}, \dots, a_d x_{i-d-1}). \quad (1.15)$$

The objective of a successful embedding is to find $a = [a_1, a_2, \dots, a_d]$ where $a_i \in \{0, 1\}$. Note that $\hat{g}(x_i)$ is simply the subspace projection of $g(x_i)$ onto a ,

$$\hat{g}(x_i) = \text{Proj}_a g(x_i).$$

Using this notational convenience, the embedding is completely defined by $a \in \{0, 1\}^d$ and we wish to make the best choice of a and d , which we write (a, d) . Note that, in general one could consider $a \in \mathbf{R}^{d \times \tau}$. We restrict ourselves to $\{0, 1\}^d$ as the more general case is concerned with the optimal model of the dynamics rather than the necessary information. For a uniform embedding with embedding parameters d_e and τ , we have that $a \in \{0, 1\}^d$ and $(a)_i \neq 0$ if and only if τ divides i .

Let $z_i = \hat{g}(x_i) \in \mathbf{R}^d$ and let

$$f(z) = \sum_{i=1}^m \lambda_i \theta(z; w_i) \quad (1.16)$$

where θ is some basis and $\lambda_i \in \mathbf{R}$ and $w_i \in \mathbf{R}^k$ are linear and nonlinear model parameters. The selection of this particular model architecture is arbitrary, but does not alter the results. We assume that there exists some algorithm to select $\mathcal{P} = (m, \lambda_1, \lambda_2, \dots, \lambda_m, w_1, w_2, \dots, w_k)$ such that $e_i = f(z_{i-1}) - z_i \sim N(0, \sigma)$ (or at the very least, $\sum (f(z_{i-1}) - z_i)^2 = \sigma^2$ is minimised). We do not consider the model selection problem here, rather we seek to find out what is the best choice of (a, d) . Our own model selection work is summarised in [130].

The most obvious approach to this problem is to look for the maximum likelihood solution:

$$\max_{(a,d)} \max_{\mathcal{P}} P(x|x_0, a, d, \mathcal{P})$$

where x is the vector of all the time series observations and $x_0 \in \mathbf{R}^d$ is a vector of model initial conditions. Unfortunately this leads to the

¹⁷In writing $g(x_i) = (x_i, x_{i-1}, \dots, x_{i-m-1})$ (i.e. in using equality for assignment) we take a slight liberty with the notation, but the meaning remains clear.

redundant solution $d = N$. To solve this problem one could either resort to Bayesian regularisation [80] or the minimum description length model selection criteria [103]. We choose the later approach.¹⁸

The description length of a time series is the length of the shortest (most compact) description of that time series. The description length of a time series with respect to a given model is the length of the description of that model, the initial conditions of that model and the model prediction error. We intend to optimise the description length of the observed time series $\{x_i\}_{i=1}^N = x$ with respect to (a, d) . At this point we make the fairly cavalier assumption that for a given (a, d) one can obtain the optimal model \mathcal{P} . We will address this assumption in more detail later in this section.

The description length of the data $DL(x)$ is given by

$$DL(x) = DL(x|x_0, a, d, \mathcal{P}) + DL(x_0) + DL(a, d) + DL(\mathcal{P}) \quad (1.17)$$

where $x_0 = (x_1, x_2, \dots, x_d)$ are the model initial conditions. Notice that the description length of the model prediction errors $DL(x|x_0, a, d, \mathcal{P})$, is equal to the negative log likelihood of the errors under the assumed distribution. Similarly x_0 is a sequence of d real numbers which for small d we approximate by d realisations of a random variable. Therefore $DL(x_0)$ can also be computed as a negative log-likelihood of some probability distribution. If we assume that x and x_0 are approximated by Gaussian random variables with variance σ^2 and σ_D^2 respectively, (1.17) becomes

$$DL(x) \approx -\ln P(x|N(0, \sigma^2)) - \ln P(x_0|N(0, \sigma_X^2)) + d + DL(d) + DL(\mathcal{P}). \quad (1.18)$$

Since a is a sequence of d independent zeros or ones, $DL(a) = d$, furthermore the description length of an integer d is given by $DL(d) = \lceil \log(d) \rceil + \lceil \log \lceil \log(d) \rceil \rceil + \dots$ where the last term in this expansion is 0 [103]. Compared to the term d , $DL(d)$ is very slowly varying and has little effect on the results. The final term $DL(\mathcal{P})$ is the description length of the optimal model for the given (a, d) .

Substituting for the probability distributions $P(x|N(0, \sigma^2))$ and $P(x_0|N(0, \sigma_X^2))$ and estimating σ^2 and σ_X^2 directly from the data, one fi-

¹⁸The application of Bayesian regularisation to this problem is left as an ‘‘exercise for the reader’’.

nally obtains

$$DL(x) \approx \frac{d}{2} (1 + \ln 2\pi\sigma_X^2) + \frac{N-d}{2} (1 + \ln 2\pi\sigma^2) + DL(\bar{x}) + d + DL(d) + DL(\mathcal{P}) \quad (1.19)$$

$$\approx \frac{d}{2} \ln \left[\frac{1}{d} \sum_{i=1}^d (x_i - \bar{x})^2 \right] + \frac{N-d}{2} \ln \left[\frac{1}{N-d} \sum_{i=d+1}^N e_i^2 \right] + \frac{N}{2} (1 + \ln 2\pi) + d + DL(d) + DL(\bar{x}) + DL(\mathcal{P}). \quad (1.20)$$

In this form, Eq. (1.20) provides the first suggestion of what the optimal embedding strategy should be. We see that a does not feature in this calculation. Hence, if we adopt the modelling paradigm suggested here, the embedding lag (or more generally the embedding strategy) is not crucial: one should only be concerned with the maximum embedding dimension d . Of course, this does not mean that to reconstruct the dynamics, the embedding lag is unimportant. When one applies numerical modelling to reconstruct the dynamics, embedding strategies are of very great significance; however, selecting the optimal embedding co-ordinates (or rather those that are most significant in predicting the dynamics) is inherently part of the modelling process [55]. Furthermore, the modelling algorithm should be allowed to choose from all possible embedding lags within the embedding window. Indeed, one often finds that the “optimal” embedding strategy is not fixed within a single model [55]. This result shows that it is preferable to identify the embedding window d_w and let the model building process determine which of the d_w co-ordinates are most useful.

The description length of the mean of the data $DL(\bar{x})$ is a fixed constant and we drop it from the calculation. Optimising (1.20) over all (a, d) requires selection of the optimal model for a given (a, d) and computation of the model prediction error of that model. For a given model, $DL(\mathcal{P})$ can be calculated precisely [53]. However, selection of the optimal model is a more difficult problem.

Instead, we restrict our attention to a particular *class* of model, and choose the optimal model from that class. To simplify the computation of (1.20) we restrict our attention to the class of local constant models on the attractor. We have two good reasons for choosing this particular class. Firstly, because the models are simple, estimates of the error as a function of (a, d) are relatively well behaved. Secondly, these models rely

on no additional parameters and therefore $DL(\mathcal{P}) = 0$, simplifying our calculation considerably.¹⁹

In trials [132], we tested many alternative model classes. We found radial basis functions [53] and neural networks [130] to be excessively non-linear and difficult to optimise for the purpose of determining embedding windows. Complex local modelling regimes, such as triangulation and tessellation [84] or parameter dependent local linear schemes [46], we found to be overly sensitive to small changes in the data. In comparison the local constant scheme we employ here appears remarkably robust.

As local constant models have no explicit parameters (other than the embedding strategy (a, d)), $DL(\mathcal{P}) = 0$. Therefore, for a given (a, d) , computation of (1.20) only requires estimation of $\sum e_t^2$. We employ an in-sample local constant prediction strategy. Let z_s be the nearest neighbour to z_t (excluding z_t), then

$$x_{t+1} = x_{s+1} \tag{1.21}$$

and therefore $e_{t+1} = x_{t+1} - x_{s+1}$. In other words, for each point in the time series we determine the prediction error based on the difference between the successor to that point and the successor to its nearest neighbour.²⁰ Since this is a form of interpolation rather than extrapolation, this strategy does not provide a *predictive* model, likewise (as with all local techniques) it does not describe the underlying dynamics. However, the strength of this particular approach is that it is simple and it provides a realistic estimate of the size of the optimal model's prediction error as a function of (a, d) .

The proposed algorithm may be summarised as follows. We seek to minimise (1.20) over d . To achieve this we need to estimate the model prediction error as a function of d . Hence, for increasing values of d we employ the local constant "modelling" scheme suggested by (1.21) to compute the model prediction error and substitute this into (1.20). The optimal embedding window d_w is the value of d that minimises (1.20).

1.7.2 Examples

We now demonstrate the application of the above method to several numerical time series. First, we examine the performance of the algorithm and importance of the choice of modelling algorithm (1.21).

¹⁹Alternatively, one could argue that the data are the parameters, in either case the description length of the model is constant.

²⁰This is a technique sometimes referred to as "drop-one-out" interpolation.

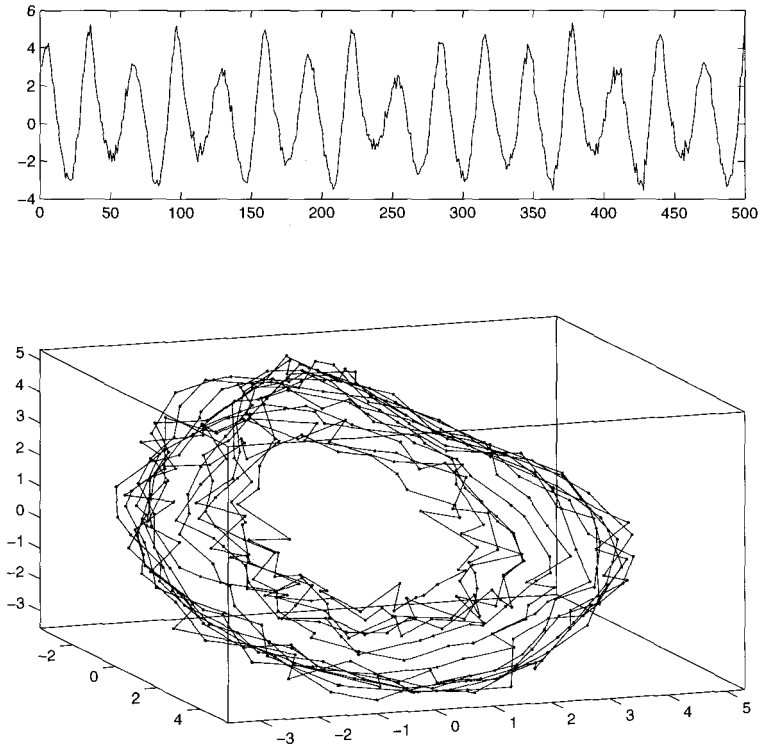


Fig. 1.13 Typical noisy time series for the Rössler dynamical system and the reconstructed attractor. Noisy x component (top panel) and a three dimensional reconstruction of the Rössler system (1.4). The system is integrated (and sampled) with a time step of 0.2, and subjected to additive Gaussian noise with a standard deviation of 10% of the standard deviation of the data.

The example we consider is 2000 points of the x component of a numerically integrated (sampling rate of 0.2) trajectory of the Rössler system, contaminated by additive Gaussian noise with a standard deviation of 5% of the standard deviation of the data. The effect of noise on this time series and its attractor are shown in Fig. 1.13.

Figure 1.14 demonstrates the computation of (1.20) as a function of embedding window. To estimate model prediction error we employ the rather simple interpolative scheme described in the previous section. For comparison, the performance of alternative (more complex) modelling schemes is also shown in Fig. 1.14. We find that alternative, more parametric, mod-

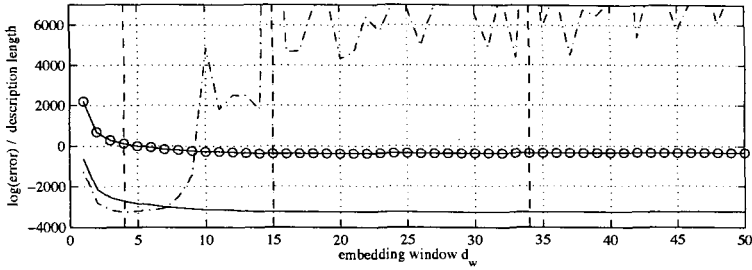


Fig. 1.14 **Computation of description length as a function of embedding window for Rössler time series data.** The solid line and dot-dashed line are proportional to the logarithm of the sum of the squares of the model prediction error using a local constant and local linear method respectively. The local constant model utilised is described in Sec. 1.7.1, the local linear scheme is described in the text. The second modelling scheme exhibits a clear minimum which occurs at 4. The local constant modelling scheme employs only lags that provide an improvement to model prediction error. Its error as a function of embedding window is therefore monotonic (plateau onset occurs at 34). For small values of embedding window the linear scheme performs best, but for large values, behaviour is poor and extremely erratic. Computation of description length utilising the local constant scheme (solid line with circles) yields an optimal embedding window of 15. For clarity, the values $d_w = 4, 15, 34$ are marked as vertical dashed lines.

elling methods produce results which are sensitively dependent on “correct” choice of modelling algorithm parameters.²¹

The first zero of the autocorrelation function occurs at a lag of 8 and the data exhibits a pseudo period of about 31 samples. With the embedding lag set at 8, false nearest neighbours indicates a minimum embedding dimension of 4. Standard methods, therefore, suggest an embedding window of roughly 32.

By coincidence,²² the minimum of the model prediction error for a constant model occurs at this value. Conversely, the minimum of the error of the local linear model occurs at a value of 4. This comparatively low value of embedding window is due to the relative complexity of the local linear modelling scheme [137]. Although this scheme performs best for small embedding windows, the additional information introduced with larger embedding windows is not recognised by this scheme. The main reason for this is that the parameters of the scheme (neighbourhood size, neighbourhood

²¹By modelling algorithm parameters we mean parameters associated with the model selection scheme itself rather than only the parameters optimised by that scheme.

²²In other examples, and for other amounts of noise or with other lengths of data this proved not to be the case.

weights and so on) are also dependent on the embedding dimension and embedding lag. For example, values of neighbourhood size which work well for a small dimension embedding may not work well for larger embedding dimension. Moreover, as embedding dimension becomes larger, it becomes difficult to find good values for these parameters.. This general behaviour is observed in every example we consider. Therefore, although the local linear scheme often provides a good estimate of the optimal embedding *dimension* (as would false nearest neighbours), the description length estimated from a local constant model provides a much better estimate of the optimal embedding *window*.

We have already mentioned that the local constant modelling scheme selects only lags that provide some improvement in model prediction error. Clearly, as d_w increases there is a combinatorial explosion. To address this combinatorial explosion is both difficult and beyond the requirements of this algorithm. We consider only whether the addition of *successive* lags offers an improvement. Suppose for a d_e dimensional embedding the chosen model includes the lags $\{\ell_1, \ell_2, \dots, \ell_k\}$ (where $0 \leq \ell_1 \leq \ell_i < \ell_{i+1} \leq \ell_k < d_e$). To determine the set of model lags for the $(d_e + 1)$ -dimensional embedding we consider the performance of the local constant model with lags $\{\ell_1, \ell_2, \dots, \ell_k, d_e\}$. If this model performs better than the model with lags $\{\ell_1, \ell_2, \dots, \ell_k\}$ then it is accepted, otherwise we retain only the lags $\{\ell_1, \ell_2, \dots, \ell_k\}$.

Therefore, the selected lags may be used as an estimate of the optimal lags for a generalised variable embedding (1.13). In the case of the Rössler system data analysed in Fig. 1.14, the optimal lags were 1 to 15 and 19, 20, 24, 26, 29, 32 and 34. Altogether, 22 different lags. Clearly, a 22 dimensional embedding is excessive, and some subset of these lags would probably prove sufficient. Moreover, the minimum description length optimal embedding window is 15, limiting the selection to the first 15 lags. It is reasonable to suppose that each of these large number of lags may contribute some significant novel information to the modelling scheme. However, the expression we hope to optimise (1.20) is independent of which lags are included (indeed, in this example, they are *all* included²³), and therefore we do not consider this issue more closely here. We defer the selection of optimal lags from this set for the modelling phase of dynamic reconstruction.

In Fig. 1.15 we examine the effect of various noise levels and different

²³This is not the case in general.

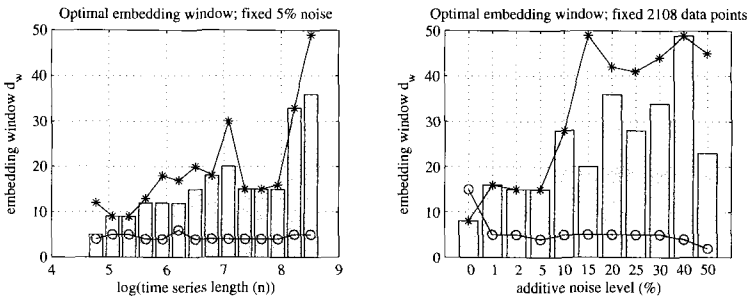


Fig. 1.15 **Optimal embedding dimension as a function of data length and noise level.** The solid bars depict the optimal model size utilising the methods described in this paper for a single realisation of Rössler time series data. The panel on the left is for a fixed noise level of 5% and time series length between 118 and 5000 data. The panel on the right is for fixed data length of 2108 data and various noise levels (expressed as percentage of the standard deviation of the data). For the cases where noise was added to the time series, the results depicted here are for a single realisation of that noise (not an average). This is the likely cause of the moderate variation in the results observed for larger noise levels. For comparison, the embedding window that yielded minimum error for the local constant (asterisks) and local linear (circles) models is also shown.

length time series on the selection of embedding window. We observe that for longer time series, the optimal embedding window is larger. This is consistent with what one might expect. For short time series the optimal model can only capture the short term dynamics and therefore only recent past history (a small embedding window) is required. For larger quantities of data one is able to characterise the more sensitive long term dynamics and a larger embedding window provides significant advantage. Initially, an embedding window of about 10 is sufficient, while for the longest time series an embedding window of 35 is optimal. Significantly, these two values correspond to approximately the first zero of the autocorrelation function (or one-quarter of the pseudo-period) and the pseudo-period of the observed time series.

We note in passing, that, the optimal embedding window for the local constant window is an upper bound on the minimum description length best window. This is as we would expect. The description length is the sum of a term proportional to the model prediction error and a function which increase monotonically with embedding dimension (the description length of the local constant model). Therefore the minimum of the model prediction error must be no less than the minimum of the description length.

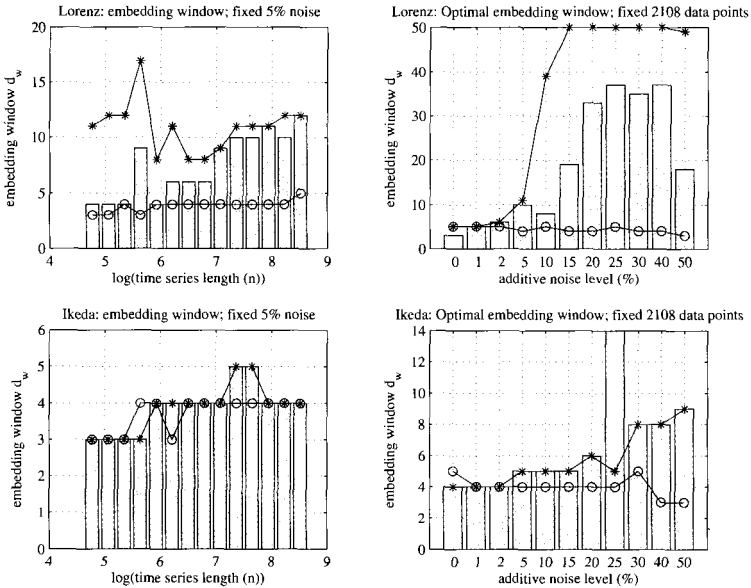


Fig. 1.16 **Optimal embedding windows for Lorenz and Ikeda time series.** The calculations depicted in Fig. 1.15 are repeated for time series of two standard systems. The top two panels are for a single realisation of the chaotic Lorenz system, the bottom two panels are for a single realisation of the chaotic Ikeda map. The solid bars depict the optimal model size utilising the methods described in this paper. The leftmost panels are for a fixed noise level of 5% and time series length between 118 and 5000 data. The panels on the right are for fixed data length of 2108 data and various noise levels (expressed as a percentage of the standard deviation of the data). This is the likely cause of the moderate variation in the results observed for larger noise levels. For comparison, the embedding window that yielded minimum error for the local constant (asterisks) and local linear (circles) models is also shown.

Conversely, we find that the optimal embedding window for the local linear method remains about 4 or 5 (roughly corresponding to the optimal embedding dimension).

Variation in the noise level for a fixed length time series demonstrates similar behaviour. For noisier time series, a larger embedding window is required, as increasing the noise on each observation decreases the useful information provided. As the information provided to the optimal model by each observation decreases, more observations (a larger embedding window) is required to provide all the available information. For noise levels of up to 30% this method provides consistent, repeatable, results. Noisier

time series tend to yield a larger variation in the optimal estimates of the embedding window. Note that in contrast, the local linear scheme performs progressively worse, utilising a diminishing window as the noise level is increased. We believe that this is due to the additional parametric complexity of this modelling method. As more noise is added to the data, the (relatively) complex rules used to determine near neighbours and derive a weighted linear prediction from these, become more prone to the system noise, and actually perform worse.

In Fig. 1.16 we repeat the above calculations for time series generated from the standard chaotic Lorenz system and the Ikeda map [61]. Variation of optimal embedding window as a function of noise and data length for the Lorenz data is very similar to the results depicted in Fig. 1.15 for the Rössler system. Increasing noise level or time series length yields a larger optimal model. Furthermore, optimal embedding window values tend to coincide with the pseudo-period of the time series, or one-quarter, or one-half of this value.

Results for the Ikeda map are substantially different. In this case the optimal embedding window estimated coincides with the value that minimises the error of the local constant and linear models. In general, an embedding dimension of 3 or 4 is suggested, and this is what one would expect for this system.²⁴

We now return to the main purpose of estimating the embedding window, namely the reconstruction of the dynamics. For the Rössler system analysed in Figs. 1.14 and 1.15 we build nonlinear models following the methods described in [55] with embedding suggested by either autocorrelation or false nearest neighbours (namely $d_e = 4$ and $\tau = 8$), hereafter referred to as a *Standard Embedding*, or with the embedding window (of 34), hereafter a *Windowed Embedding*. Table 1.3 compares the average model size (number of nonlinear basis functions in the optimal model) and model prediction error for 60 models of this time series (2000 observations and 5% noise) with each of these two embedding strategies. These models are built to minimise the description length of the data given the model, and therefore a comparison of the optimal model description length is also given. These qualitative measures show a consistent improvement in the model performance for the model built from the windowed embedding.

²⁴Although the fractal dimension of the Ikeda map is less than two, a delay reconstruction of this map is highly “twisted” and requires an embedding dimension of 3 or 4 to successfully remove all intersecting trajectories. This is evident in Fig. 1.8.

Table 1.3 **Comparison of model performance with standard constant lag embedding (a *Standard Embedding*) and embedding over the embedding window suggested in Fig. 1.15 (a *Windowed Embedding*)**. Figures quoted are the mean of 60 nonlinear models, fitted with a stochastic optimisation routine to the same data set, and standard deviations. Figures quoted here are for 2000 data points with 5% noise; other values of these parameters gave similar, consistent, results. The three indicators are minimum description length (MDL) of the optimal model, root-mean-square model prediction error (RMS) and the model size (number of nonlinear terms in the optimal model). For each indicator, the new embedding strategy shows clear improvement. MDL and RMS have decreased, indicating a more compact description of the data and a smaller prediction error, respectively. Conversely, the mean model size has increased indicating that more structure is extracted from the data. Several other measures were also considered: mean amplitude of oscillation, correlation dimension, entropy and estimated noise level (see chapter. 2). However, for each of these measures the variance between simulations of models built using the same embedding strategy was as large as that between the different embedding strategies. The results of these calculations are therefore omitted.

model	MDL	RMS	size
Standard ($d_e = 4, \tau = 8$)	-655 ± 23	0.158 ± 0.003	15.6 ± 2.9
Windowed ($d_w = 15$)	-716 ± 17	0.151 ± 0.004	21.1 ± 5.5

1.8 Application: Sunspots and chaotic laser dynamics: Improved modelling and superior dynamics

We now consider the application of this method to two experimental time series: the annual sunspots times series [153] and experimental laser intensity data [158; 108]. A third, unsuccessful example is described in [132]. The raw time series data are depicted in Fig. 1.9.

Since the main motivation for selection of the embedding window with the method described here is to improve modelling results, we concentrate exclusively on the comparison of the performance of nonlinear models of this data with standard embedding techniques and the windowed embedding suggested by the algorithm proposed here. By construction, the local constant modelling scheme performs best with the windowed embedding. Therefore, we consider a more complicated nonlinear radial basis modelling algorithm, first proposed in [53] and most recently described in [130]. Like the windowed embedding strategy, this modelling scheme is designed to optimise the description length of the time series [130].

We are interested in two types of measures of performance: short term behaviour (for example, mean square prediction error) and dynamic be-

Table 1.4 **Comparison of model performance with standard constant lag embedding and embedding over the embedding window suggested in Fig. 1.15.** Figures quoted are the mean of 60 nonlinear models, fitted with a stochastic optimisation routine to the same data set, and standard deviations. Figures quoted here are for 2000 data points, where more data is available; longer time series samples gave similar, consistent, results. The four indicators are minimum description length (MDL) of the optimal model, root-mean-square model prediction error (RMS), the model size (number of nonlinear terms in the optimal model), and the correlation dimension (CD) of the free run dynamics. For the laser time series, none of the models built using the standard embedding produced stable dynamics and it was therefore not possible to estimate correlation dimension. The correlation dimension estimated directly from these three data sets was 0.396, 1.090, 1.182 (note that the low value for the first data set is an artefact of the short time series).

data	MDL	RMS	size	CD
sunspots				
($d_e = 6, \tau = 3$)	1267.9 ± 12.1	13.16 ± 1.116	7.32 ± 1.818	0.938 ± 0.456
($d_w = 6$)	1230.1 ± 11.6	12.31 ± 0.6886	6.96 ± 1.50	0.7836 ± 0.4145
laser				
($d_e = 5, \tau = 2$)	5753.6 ± 153.9	2.405 ± 0.2954	100.8 ± 12.3	n/a
($d_w = 10$)	5239.8 ± 159.0	1.767 ± 0.1992	109.5 ± 12.3	0.8637 ± 0.7999

behaviour (invariant measures of the dynamical systems). Results equivalent to those depicted in table 1.3 have also been computed and are summarised in table 1.4.

Table 1.4 shows that for the sunspot time series and the experimental laser intensity recording, the windowed embedding improved model performance. That is, the description length was lower, the one-step model prediction error was less and the models were larger. However, with the exception of one step model prediction error the difference in these measures was not statistically significant. For the recording of human VF, the new method did not improve model performance and, in fact, the optimal embedding window was $d_w = 2$: substantially smaller than one would reasonably expect from such a complex biological system. It seems plausible, that in this case, the time series under consideration is too short, noisy or non-stationary (this conclusion is supported by Fig. 1.9). Finally, we note that the result for the sunspot time series is particularly encouraging because this improvement in short term predictability is achieved with a much smaller embedding ($d_w = 6$ compared to $d_e \tau = 18$).

However, as has been observed elsewhere [130], short term predictability is not the best criteria with which to compare models of nonlinear dynamical systems. Therefore, for each model, we estimated correlation dimension, noise level and entropy, using a method described in [167]. Furthermore,

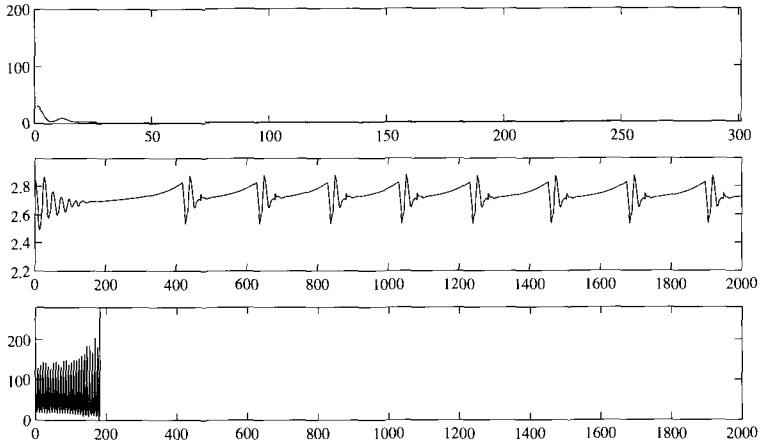


Fig. 1.17 **Typical model behaviour using the standard embedding strategy** (d_e, τ): Two simulated time series from models of the experimental data examined in this paper are depicted. The panels correspond to those of Fig. 1.9 and the horizontal and vertical axes in these figures are fixed to be the same values as the corresponding panels of Fig. 1.9.

under the premise that these models should exhibit pseudo-periodic dynamics we also computed mean limit cycle diameter (i.e. the amplitude of the limit cycle oscillations). In every case we found that the dynamics exhibited by models built from the traditional (i.e. uniform) embedding strategy was more likely to either be a stable fixed point or divergent.

Finally, Figs. 1.17 and 1.18 show typical noise free dynamics in models of each of these three systems. No effort was made to ensure that the models performed well and the models and simulations presented in these figures were selected at random. For both data sets, the new method clearly performs better. Typically, the original method produced laser dynamics and sunspots simulations that were divergent and a stable fixed point (respectively). These results are typical. In contrast, the windowed method yields models which exhibit bounded (almost) aperiodic dynamics.²⁵

²⁵Closer examination of the laser dynamics indicates that it eventually settles to a stable periodic orbit (this phenomenon can be observed towards the end of the time series depicted in Fig. 1.18).

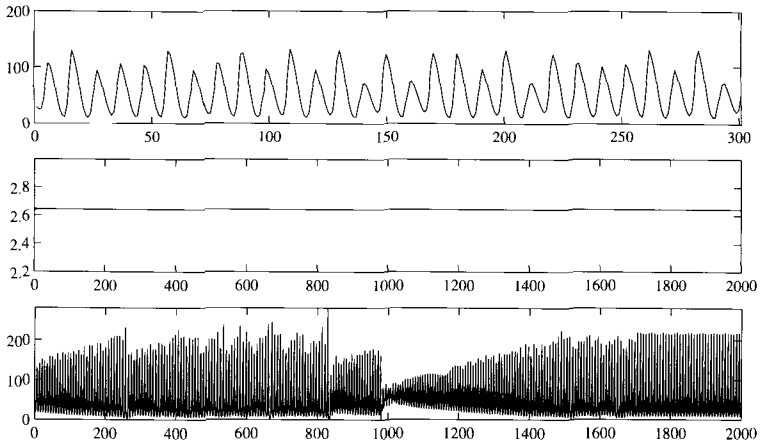


Fig. 1.18 Typical model behaviour using the windowed embedding strategy (d_w): Two simulated time series from the experimental data examined in this paper are depicted. The panels correspond to those of Fig. 1.9 and the horizontal and vertical axes in these figures are fixed to be the same values as the corresponding panels of Fig. 1.9.

1.9 Summary

We have approached the problem of optimal embedding from both the modelling perspective (Secs. 1.6 and 1.7) and from the perspective of estimating dynamic invariants (in the earlier sections). In contrast to previous reports (which focused on estimating dynamic invariants), our primary concern was selection of embedding parameters that provide the optimal reconstruction of the underlying dynamics for an observed time series. To achieve this, we assumed that the optimal model is that which minimises the description length of the data. From this foundation, we showed that the best embedding has a constant lag ($\tau = 1$) and a relatively large embedding window d_w . In general, the optimal d_w will be determined by the amount of noise and the length of the the time series. From an information theoretic perspective this is what one would expect: $\tau > 1$ implies some information is missing from the embedding. The optimal value of d_w reflects a balance between a small embedding with too little information to reconstruct the dynamics and a large embedding where the model ceases to describe the dynamics.

To compute the quantity d_w we introduced an extremely simple *non-*

predictive local constant model of the data and selected the value of d_w for which this model performs best. One can see that this offers a new and intuitive method for selection of embedding parameters. In essence, one could neglect description length and simply choose the embedding such that this model performs best. However, the addition of description length makes the optimal d_w dependent not only on the noise but also on the length of the time series. We see that for short time series, one shouldn't be confident of a large embedding window.

The similarity between this new method of embedding window selection and the well established false nearest neighbour technique [12] is more than superficial.²⁶ In Sec. 1.7.2 and 1.8 we provided an explicit comparison between our technique and the “standard” false nearest neighbour method. However, there are various improvements to this algorithm (such as [12]) which are worthy of further consideration. Nonetheless, there are several important distinctions between our method, and these false nearest neighbour techniques. As we have already emphasised, the aim of this method (to achieve the best model of the dynamics) differs from that of false nearest neighbours (topological unfolding). Furthermore, the incorporation of minimum description length means that our method explicitly penalises short or noisy time series.

At a functional level, the two algorithms are similar because both methods seek to avoid data points which are close, but which quickly diverge. Such points are (respectively) either false nearest neighbours or bad nonlinear predictors of one another. However, whereas false nearest neighbour methods seek only to avoid this situation (i.e. spreading out the data is sufficient), the windowed embedding method insists that the neighbours which are the best predictors be found.

Consider the situation where a system's dynamics either stochastic or extremely high dimensional. Using false nearest neighbour methods, one may simply embed the data in a high enough dimension so that the data are sufficiently sparse. However, doing so does not improve the nonlinear prediction error, consequently, the windowed embedding method would prefer a small embedding window.

Conversely, consider the situation at a separatrix. Points which are close do rapidly diverge from one another and so they will appear as false near neighbours for large embedding dimensions, until (at a time scale similar

²⁶The comparison of this method to that described in [12] is particularly apt. Cao introduces a modified false nearest neighbour approach which, like our method, avoids many of the subjective parameters of alternative techniques.

to that of the underlying system) the points are eventually, sufficiently spread. But from a nonlinear prediction view-point, these points are equally difficult to predict for all embedding dimension, and again the windowed embedding method will indicate a much smaller embedding dimension than that suggested by a strict application of false nearest neighbours.²⁷

Finally, we note that the examples of Sec. 1.7.2 showed that this method performed consistently and the applications in Sec. 1.8 showed that selecting embedding parameters in this way improved the model one-step prediction error. In effect, this is a demonstration that the method is working as expected. More significantly, we found that the dynamics produced by models built from windowed embedding also behaved more like the experimental dynamics than for models built from a standard embedding. This is a very positive results however, we are now faced with a more substantial problem: the problem of building the best nonlinear model for the data once the embedding window has been determined [130]. Information theory has shown us that the optimal embedding should fix $\tau = 1$, we now need to consider the practice of nonlinear modelling to determine which lags $\ell = 1, 2, 3, \dots, d_w$ are significant for practical reconstruction from specific experimental systems.

²⁷We acknowledge that this problem is actually related to the “plateau” observed in plots of the fraction of false nearest neighbours against embedding dimension. In many cases, prudent selection of “plateau-onset” can minimise the problem. However, this remains somewhat subjective.