Physica D 51 (1991) 52–98
North-Holland

State space reconstruction in the presence of noise

Martin Casdagli, Stephen Eubank, J. Doyne Farmer and John Gibson

Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA and Santa Fe Institute, 1120 Canyon Rd., Santa Fe, NM 87501, USA

Takens' theorem demonstrates that in the absence of noise a multidimensional state space can be reconstructed from a scalar time series. This theorem gives little guidance, however, about practical considerations for reconstructing a good state space. We extend Takens' treatment, applying statistical methods to incorporate the effects of observational noise and estimation error. We define the *distortion matrix*, which is proportional to the conditional covariance of a state, given a series of noisy measurements, and the *noise amplification*, which is proportional to root-mean-square time series prediction errors with an ideal model. We derive explicit formulae for these quantities, and we prove that in the low noise limit minimizing the distortion is equivalent to minimizing the noise amplification.

We identify several different scaling regimes for distortion and noise amplification, and derive asymptotic scaling laws. When the dimension and Lyapunov exponents are sufficiently large these scaling laws show that, no matter how the state space is reconstructed, there is an explosion in the noise amplification – from a practical point of view determinism is lost, and the time series is effectively a random process.

In the low noise, large data limit we show that the technique of local singular value decomposition is an optimal coordinate transformation, in the sense that it achieves the minimum distortion in a state space of the lowest possible dimension. However, in numerical experiments we find that estimation error complicates this issue. For local approximation methods, we analyze the effect of reconstruction on estimation error, derive a scaling law, and suggest an algorithm for reducing estimation errors.

Contents

0167-2789/91/$03.50

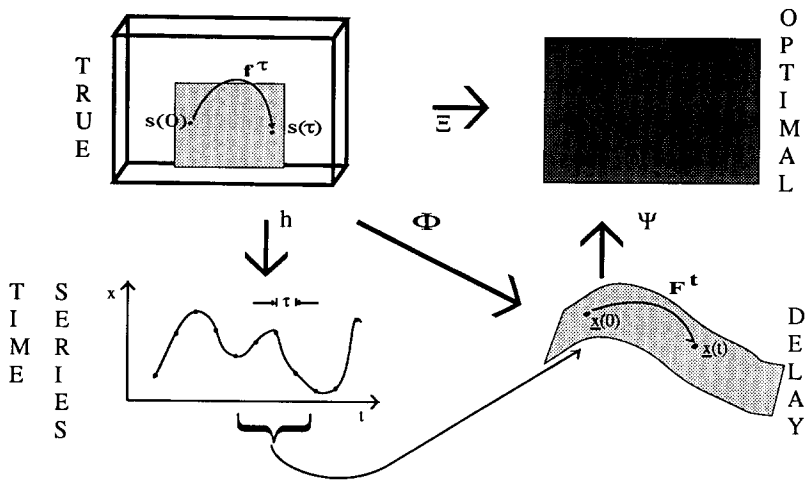

Fig. 1. The reconstruction problem. The true dynamical system f, its states s, and the measurement function h are unobservables, locked in a black box. Values of the time series x separated by intervals of the lag time τ form a delay vector $\underline{x}$ of dimension m. The *delay reconstruction map* Φ maps the original d-dimensional state s into the delay vector $\underline{x}$. The coordinate transformation Ψ further maps the delay vector $\underline{x}$ into a new state y, of dimension $d' \leq m$.

1. Introduction

1.1. Background

There are many situations in which a *time series* $\{x(t_i)\}$, $i = 1, \ldots, N$ is believed to be at least approximately described by a smooth dynamical system[#1] f on a d-dimensional manifold M:

$$s(t) = f^t(s(0)); \tag{1}$$

$s(t)$ is the state at time t. In the absence of noise, the time series is related to the dynamical system by

$$x(t) = h(s(t)). \tag{2}$$

We call h the *measurement function*. The time series $x(t)$ is D-dimensional, so that h: $\mathrm{M} \to \mathbb{R}^D$. We are most interested in dimension-reducing measurement functions, where $D < d$; we often implicitly assume $D = 1$. The state space reconstruction problem is that of recreating states when the only information available is contained in a time series. A schematic statement of the problem is given in fig. 1.

State space reconstruction is necessarily the first step that must be taken to analyze a time series in terms of dynamical systems theory. Typically f and h are both unknown, so that we cannot hope to reconstruct states in their original form. However, we may be able to construct a state space that is in some sense equivalent to the original. This state space can be used for qualitative analysis, such as phase portraits, or for quantitative statistical characterizations. We are particularly interested in state space reconstruction as it relates to the problem of nonlinear time series prediction, a subject that has received considerable attention in the last few years [8, 10, 11, 14, 15, 23, 28, 29, 32, 34, 42].

State space reconstruction was introduced into dynamical systems theory independently by Packard et al. [33], Ruelle[#2], and Takens [41]. In fact, in time series analysis this idea is quite old, going back at least as far as the work of Yule [44]. The important new contribution made in dynamical systems theory was the demonstration that it

[#1] This is one of several possible ways of representing a dynamical system. The map f^t takes an initial state $s(0)$ to a state $s(t)$. The time variable t can be either continuous or discrete. f^t is sometimes called the *time-t map* of the dynamical system. For simplicity, we will often implicitly assume that $\mathrm{M} = \mathbb{R}^d$.

[#2] Private communication.

is possible to preserve geometrical invariants, such as the eigenvalues of a fixed point, the fractal dimension of an attractor, or the Lyapunov exponents of a trajectory. This was demonstrated numerically by Packard et al. and was proven by Takens.

The basic idea behind state space reconstruction is that the past and future of a time series contain information about unobserved state variables that can be used to define a state at the present time. The past and future information contained in the time series can be encapsulated in the *delay vector* defined by eq. (3), where for convenience we assume that the sampling time is uniform,

$$\underline{x}(t) = \big(x(t+\tau m_{\mathrm{f}}), \ldots, x(t), \ldots, x(t-\tau m_{\mathrm{p}})\big)^{\dagger}. \tag{3}$$

Here † denotes the transpose, and we adopt the convention that states are represented by column vectors. The dimension of the delay vector is $m = 1 + m_{\mathrm{p}} + m_{\mathrm{f}}$. The number of samples taken from the past is m_{p}, and the number from the future is m_{f}. If $m_{\mathrm{f}} = 0$ then the reconstruction is *predictive*; otherwise it is *mixed*. The time separation between coordinates, τ, is the *lag time*.

Takens studied the *delay reconstruction map* Φ, which maps the states of a d-dimensional dynamical system into m-dimensional delay vectors:

$$\Phi(s) = (h(f^{\tau m_{\mathrm{f}}}(s)), \ldots, h(s), \ldots, h(f^{-\tau m_{\mathrm{p}}}(s)))^{\dagger}. \tag{4}$$

He showed that generically Φ is an embedding when $m \geq 2d + 1$. An *embedding* is a smooth, one-to-one coordinate transformation with a smooth inverse. If Φ is an embedding then a smooth dynamics F is induced on the space of reconstructed vectors:

$$F^{t}(\underline{x}) = \Phi \circ f^{t} \circ \Phi^{-1}(\underline{x}). \tag{5}$$

The reconstructed states can be used to estimate F, and since F is equivalent to the original dynamics f, we can use it for any purpose that we could use the original dynamics, such as prediction, computation of dimension, fixed points, etc.

1.2. Complications of the real world

Takens' proof is important because it gives a rigorous justification for state space reconstruction. However, it gives little guidance on reconstructing state spaces from real-world, noisy data. For example, the measurements $x(t)$ in the proof are arbitrarily precise, resulting in arbitrarily precise states. This makes the specific value of the lag time τ arbitrary, so that any reconstruction is as good as any other[#3]. However in practice, the presence of noise in the data blurs states and makes picking a good lag time critical. In this paper, we build on Takens' proof, by examining how states are affected when the assumption of arbitrary precision is relaxed.

There are several factors which complicate the reconstruction problem for real-world data:

– *Observational noise.* The measuring instruments are noisy; what we actually observe is $x(t) = \tilde{x}(t) + \xi(t)$, where $\tilde{x}(t)$ is the true value and $\xi(t)$ is noise.

– *Dynamic noise.* External influences perturb s, so that from the point of view of the system under study the evolution of s is not deterministic. f is thus a *stochastic* dynamical system.

– *Estimation error.* f and h are both unknown. We can estimate the dynamics in the reconstructed state space, but with a finite amount of data the approximation is never perfect.

1.3. Information flow and noise amplification

In real problems noise is always present. When we project a d-dimensional state onto a D-dimensional measurement with $D < d$, we

[#3] Provided it meets the conditions for genericity. For example, for a limit cycle, τ cannot be rationally related to the period.

throw away information. We can reconstruct some of this missing information from the past and future measurements. However, if the uncertainty of the reconstructed state is much higher than that of the individual measurements, then we have amplified the noise; the system appears less deterministic than it would if we could observe more information.

State space reconstruction relies on a flow of information from the unobserved variables to the observed variables. This can be qualitatively illustrated with the familiar Lorenz equations,

$$\dot{x} = 10(y - x),$$
$$\dot{y} = -xz + 28x - y,$$
$$\dot{z} = xy - \tfrac{10}{3}z. \qquad (6)$$

Assume that we observe x. Since $\dot{x}$ does not depend on z directly, information about z depends on the flow of information through y; when z changes it causes $\dot{y}$ to change, which causes y and hence $\dot{x}$ to change. When $x \approx 0$, since the only coupling to z is through the xz term, a large change in z causes only a small change in x. Equivalently, a small change in x corresponds to a large change in z. Thus the noise in the determination of z from noisy measurements of x is acutely amplified when $x \approx 0$. We refer to this phenomenon as *noise amplification*.

The formalism that we develop in this paper makes the notion of noise amplification precise, so that the qualitative analysis of the Lorenz equations in the previous paragraph becomes quantitative. It also provides guidance into the practical problem of reconstructing coordinates so that they minimize noise amplification.

Noise amplification depends on the following factors:

– *The measurement function*. Observation of one quantity may give more information than another.

– *The method of reconstruction*. A poor state space reconstruction amplifies noise more than a good state space reconstruction; noise amplification depends on factors such as m and τ.

– *The dynamical system*. Noise amplification depends on the flow of information between the individual degrees of freedom, which depends on properties of the dynamical system such as the dimension and Lyapunov exponents.

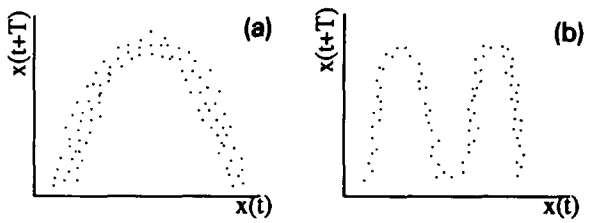

Fig. 2. Two hypothetical scenarios for prediction in a one dimensional state space. The horizontal axis is the state at time t, and the vertical axis is the state at time $t + T$. (a) shows a coordinate system with high noise amplification, while (b) shows a coordinate system with low noise amplification. This is evident from the thickness of the distribution of points at any given $x(t)$. However, since the functional form of (b) is more complicated, with a limited amount of data (b) might result in larger estimation error than (a).

1.4. Noise amplification versus estimation error

The difference between noise amplification and estimation error from the point of view of prediction is illustrated in fig. 2. The noise amplification is related to the "thickness" of the distribution of points. In fig. 2a the noise amplification is large, and in fig. 2b the noise amplification is small. However, the estimation error in (b) might be larger than that of (a).

Both noise amplification and estimation error cause prediction errors, and both of them depend on the reconstruction. The estimation error, however, also depends on the method of approximation. For most good approximation schemes, the estimation error goes to zero in the limit of a large number of data points. The prediction errors in this limit are entirely due to noise. The noise amplification thus tells us the prediction errors that remain even with a perfect model, setting a limit to predictability that is independent of the modeling procedure. As we shall

show, when the dimension and Lyapunov exponents are sufficiently large there can be a complete breakdown of predictability. The time series is unpredictable over times much shorter than the Lyapunov time, even with a perfect model (except for predictability through short-term linear correlation). In this limit the time series becomes a true random process.

1.5. Data compression and coordinate transformations

Any approach to state space reconstruction uses the information in delay coordinates as a starting point. For some purposes, such as reducing the dimension of a reconstruction, it may be desirable to make a further coordinate transformation to a new coordinate system y,

$$y = \Psi(\underline{x}). \tag{7}$$

As described in section 2, examples of such transformations Ψ are differentiation and principal value decomposition. By splitting the reconstruction process into Φ and Ψ, we have conveniently labeled the two parts of the problem. The choice of Φ determines the form of the delay coordinates, which are the raw information we have to work with, while Ψ determines how we use that information. The total reconstruction map $\Xi = \Psi \circ \Phi$ takes the original coordinates s to the reconstructed coordinates y. See fig. 1.

We will show that it is impossible to reduce the noise amplification by transforming delay coordinates by Ψ. The minimum possible noise amplification over all Ψ is obtained when $\Psi = \mathbb{1}$ and $y = \underline{x}$. However, as the noise level tends to zero, it *is* in general possible to compress all the information in $\underline{x}$ into a coordinate y with a lower dimension while keeping the noise amplification the same. The local principal value decomposition technique discussed in sections 6 and 8 accomplishes this in the minimum possible dimension. However, this technique is subject to estimation problems which sometimes outweigh the benefits of dimension reduction.

1.6. Approach and simplifying assumptions

The main goal of this paper is to develop a theory which gives insight into practical problems of state space reconstruction in the typical case in which a time series is the only available information. In order to get insight into the problem and develop a theory for its solution, we begin by assuming that we know both f and h. In sections 3 through 6, we develop an understanding of the effect that f and h have on the problem of determining s from noisy data. In section 7, we take a different viewpoint and investigate how the reconstruction affects the estimation of f and h. In section 8, we investigate the implications of these theoretical results for algorithms when only the time series is known.

Throughout this paper we assume that the noise is entirely observational. Treating dynamic noise is obviously important, but it is outside the scope of this paper. We also assume that the observational noise is independent and identically distributed (IID). In practice, noise tends to become correlated as sampling time goes to zero, so we will assume that the lag time τ is significantly greater than the correlation time. A similar problem arises if the measuring instrument records discrete, symbolic information rather than a continuous variable, but this will not be important if measurement errors are dominated by noise rather than quantization. We believe that the framework we have established here can be extended to treat dynamical, correlated and quantization noise as well.

1.7. Overview

In section 2, we review what is currently known about state space reconstruction. We begin by discussing methods currently available for state space reconstruction, such as delay coordinates, derivative coordinates, and principal value decomposition. We then review Takens' theorem, and present an intuitive discussion of why it is true.

In section 3, we derive formulae for the probabilistic treatment of this problem. We use several examples to develop intuition and to illustrate qualitatively what factors are essential for a good state space reconstruction.

From a practical point of view, it is important to have a simple criterion for selecting a reconstruction. A complete description of a reconstruction is contained in a probability density function, but this is too complicated; we need a number, or a set of a few numbers. In section 4, we examine several candidates and argue that for this problem, criteria based on the variance are more appropriate than other possibilities, such as mutual information. We define two quantities based on the variance: the distortion, which is related to errors in the state space, and noise amplification, which is related to errors in time series prediction. We derive explicit formulae for these quantities and investigate numerical examples.

In section 5, we study the dependence of distortion and noise amplification on the dynamical system and the methods of reconstruction. We demonstrate that for a given τ, distortion is a decreasing function of m. In the low noise limit, we derive scaling behaviors of the distortion as a function of m, τ, d, and the Lyapunov exponents. We show that for predictive coordinates an explosion in the noise amplification occurs when the Lyapunov exponents and dimension are sufficiently large. This causes a transition from behavior that is approximately deterministic for short times to behavior that is effectively random over almost any time scale. We use two examples to illustrate several aspects of the behavior of the distortion and noise amplification.

In section 6 we study the effect of making coordinate transformations from delay coordinates to more general coordinates. We demonstrate that in the low noise, large data limit, local singular value decomposition (SVD) is an optimal coordinate transformation in the sense that it minimizes the distortion with a coordinate system of the smallest possible dimension. In the low noise limit we prove that minimizing the distortion is equivalent to minimizing the noise amplification.

In section 7, we examine the effect of the reconstruction on estimation errors in prediction. We derive scaling laws for estimation error for local approximation methods. We show that noise amplification and estimation error are counteractive effects, and that the optimal state space for prediction balances between them. We discuss the possibility of defining quantities analogous to distortion for estimation error and dynamic noise.

Finally, in section 8, we discuss algorithms for constructing coordinates when only the time series is known. We show that local SVD can be estimated from a time series, through a technique we call local principal value decomposition (PVD). We perform numerical experiments comparing local PVD to other methods, such as delay coordinates and global PVD. Finally, we suggest an algorithm for reducing estimation errors.

1.8. Summary of notation

The notation we use in this paper is summarized in table 1.

2. Review of previous work

2.1. Current methods of state space reconstruction

The currently used possibilities for state space reconstruction include delay coordinates, derivative coordinates, and global principal value decomposition. Each of these is sometimes done in conjunction with filtering. As a matter of experience it is quite clear that the method of reconstruction can make a big difference in the quality of the resulting coordinates, but in general it is not clear which method is the best.

Delay coordinates are currently the most widely used choice. They have the nice property that the signal to noise ratio on each component is the same. They have the unpleasant property that in order to use them it is necessary to choose the

Table 1
Notation used in this paper.

Symbol	Description
M	d-dimensional manifold representing the state space
$s(t)$	d-dimensional state at time t
f^t	time-t map of dynamical system; $s(t) = f^t(s(0))$
$x(t)$	noisy D-dimensional value of time series at time t (we often assume $D = 1$)
$\xi(t)$	noise fluctuation, usually assumed to be Gaussian IID
h	measurement function; $x(t) = h(s(t)) + \xi(t)$
$S(t)$	$d - D$ dimensional measurement surface $S(t) = \{s: x(t) = h(s)\}$
τ	sampling time $t_{i+1} - t_i$
$A^\dagger$	transpose of a matrix or vector A
Tr A	trace of a matrix A
$\underline{x}$	m-dimensional delay vector $(x(t + \tau m_f), \ldots, x(t), \ldots, x(t - \tau m_p))^\dagger$
y	reconstructed d'-dimensional coordinate based on $\underline{x}$
Φ	delay reconstruction map $\underline{x} = \Phi(s)$
Ψ	coordinate transformation map $y = \Psi(\underline{x})$
Ξ	total reconstruction map $\Xi = \Psi \circ \Phi$
w_i	ith singular value of DΦ
$\underline{\xi}$	m-dimensional vector of noise fluctuations $(\xi(t + \tau m_f), \ldots, \xi(t), \ldots, \xi(t - \tau m_p))^\dagger$
$\tilde{\underline{x}}, \tilde{s}$	true values of $\underline{x}, s$ in absence of noise
$\hat{\underline{x}}, \hat{s}, \hat{f}$	best estimate for $\tilde{\underline{x}}, \tilde{s}, f$
p	probability density function (identified by its arguments)
$p(x\|y)$	conditional probability density for x given y
Σ	distortion matrix
δ	distortion $\delta = \sqrt{\text{Tr}\,\Sigma}$
$\sigma(T)$	noise amplification for extrapolation time T
w	window width $= (m - 1)\tau$
λ	largest Lyapunov exponent
$\langle\ \rangle_t$	time average
t_R	redundance time
t_I	irrelevance time
$\mathcal{O}(\epsilon)$	"of order ϵ"
$\sim$	"asymptotically scales as"

delay parameter τ. If τ is too small each coordinate is almost the same, and the trajectories of the reconstructed space are squeezed along the identity line; this phenomenon is known as *redundance*. If τ is too large, in the presence of chaos and noise, the dynamics at one time become effectively causally disconnected from the dynamics at a later time, so that even simple geometric objects look extremely complicated; this phenomenon is known as *irrelevance*. Most of the research on the state space reconstruction problem has centered on the problems of choosing τ and m for delay coordinates. The proposals for doing this include information-theoretic quantities [1, 17, 19], and others [9, 30, 31].

Another method in common use is *principal value decomposition*, also called *principal component analysis*, *factor analysis*, or *Karhunen–Loeve decomposition*. Broomhead and King originally proposed this for reconstructing a state space for chaotic dynamical systems [7]. The simplest way to implement this procedure is to compute the $m \times m$ covariance matrix $C_{ij} = \langle x(t)\, x(t + (i-j)\tau)\rangle_t$ and then compute its eigenvalues. The eigenvectors of C_{ij} define a new coordinate system, which is a rotation of the original delay coordinate system. The eigenvalues are the average root-mean-square projection of the m-dimensional delay coordinate time series onto the eigenvectors. Ordering them according to size, the first eigenvector has the maximum possible projection, the second has the largest possible projection for any fixed vector orthogonal to the first, and so on. Typically, one reduces dimension by using only eigenvectors whose eigenvalues are large.

Another method for reconstructing a state space is the *method of derivatives*, numerically investigated by Packard et al. [33]. The coordinates are derivatives of successively higher order,

$$y(t) = \left(x(t), \hat{x}'(t), \ldots, \hat{x}^{(m-1)}(t)\right)^\dagger, \tag{8}$$

where $\hat{x}^{(j)}(t)$ is a numerical approximation to the jth derivative of $x(t)$. As Takens proved, as long as m is sufficiently large, derivatives generically define an embedding. There are many different algorithms for the numerical computation of derivatives, so in this sense the method of derivatives actually defines a family of different methods, depending on the algorithm.

All of these methods can be used in conjunction with linear filtering. For example the quality

of derivative coordinates in the presence of noise can be considerably improved by low pass filtering the time series. Note that, since linear filtering can increase the dimension of the time series, it must be done with care [3]. We have recently shown that global principal value decomposition coordinates are closely related to low-pass filtered derivative coordinates [22].

At this point there is no clear statement as to which of these methods is superior. Fraser has presented evidence for situations in which delay coordinates are superior to global principal value decomposition [18]. However, we have observed examples where the opposite is true. The situation at this point is inconclusive, and it is not clear what causes one coordinate system to be better than another. One of our central motives for defining noise amplification is to compare different methods of state space reconstruction. This gives guidance for optimizing the parameters of a particular method, or for comparing two different methods.

Principal value, derivative, and delay coordinates are related to each other by linear transformations. However, the transformation from delay coordinates to the original coordinates is typically *non*linear. As Fraser has demonstrated [18], nonlinear coordinate transformations can be greatly superior[#4]. The method of local principal value decomposition, discussed in sections 6 and 8, implements a nonlinear coordinate transformation, which gives it the potential for better performance.

2.2. *Takens' theorem revisited*

In order to understand when delay vectors form an embedding, Takens investigated the equation $\underline{x} = \Phi(s)$, assuming $\underline{x}$ is noise free. For a univariate time series ($D = 1$) this can be regarded as a set of m simultaneous nonlinear equations in d variables. The transformation Φ maps the d-dimensional state space M into an m-dimensional space. If the surface $\Phi(\mathrm{M})$ contains no self-intersections, then given any fixed $\underline{x} \in \Phi(\mathrm{M})$, there is a unique solution for s in terms of $\underline{x}$. If this solution also depends smoothly on $\underline{x}$, then Φ is an embedding[#5]. The case when $d = 2$ and $m = 3$, for example, is illustrated in fig. 3; in this case there are self-intersections along one-dimensional curves. When $m = d + 1$, the set of self-intersections is generically of dimension at most $d - 1$, and Φ is an embedding almost everywhere. As m increases by one, the dimension of the set of self-intersections generically decreases by one, until finally when $m > 2d$ there are no self-intersections at all. Thus generically, $m \geq 2d + 1$ *guarantees* that Φ is an embedding. It is *possible* that Φ will be an embedding with m as small as $m = d$, for example if Φ is sufficiently close to a nondegenerate linear map. See ref. [36]

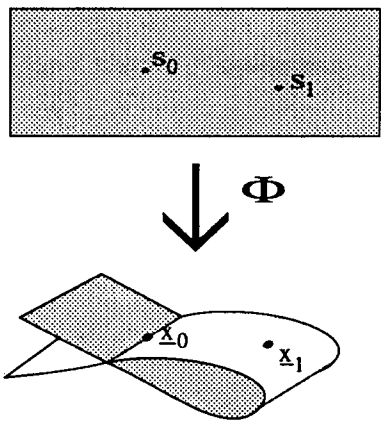

Fig. 3. Solutions of the equation $\underline{x} = \Phi(s)$ when $d = 2$ and $m = 3$. If M is the original two-dimensional state space shown above, the surface shown below is $\Phi(\mathrm{M})$. In this case there are self-intersections. The state s_0 is mapped onto a self-intersection, while s_1 is not. Except for special values of s like s_0, Φ defines an embedding.

[#4]Larimore has also considered nonlinear generalizations of canonical variate analysis for nonlinear modeling purposes [29].

[#5]By the implicit function theorem, the smoothness condition is satisfied if $D\Phi$ is of full rank everywhere. Since the set of points where $D\Phi$ fails to be of full rank is generically of lower dimension than the set of self-intersections [36], we will ignore smoothness problems in the discussion of this paragraph.

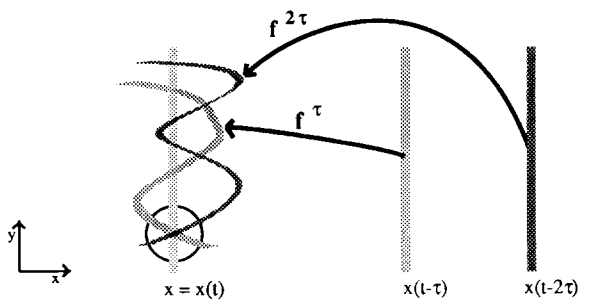

Fig. 4. A dynamical view of reconstruction in terms of the evolution of measurement surfaces, with $d = 2$ and $m = 3$. Suppose that the measurement function h corresponds to projection onto the horizontal axis, so that $h(s) = x$. A measurement at time t implies that s lies somewhere along the light gray vertical line defined by $x = x(t)$. Similarly, a measurement at time $t - \tau$ implies that it was on the darker line $x = x(t - \tau)$, and a measurement at time $t - 2\tau$ implies that it was on the darkest line $x = x(t - 2\tau)$. To see what this implies when they are taken together, each measurement surface can be mapped forward by f to the same time t. The state at time t lies on the intersection of these curves.

for a more complete discussion of Takens' theorem and its generalizations in the noise-free case.

The reconstruction process can also be considered in terms of the constraint that each measurement causes in the original state space, as illustrated in fig. 4. This gives a more dynamical point of view, which turns out to be useful for visualization in higher dimensions, and particularly in the presence of noise. Let the *measurement surface* $\mathrm{S}(t)$ be the set of possible states that are consistent with a given measurement $x(t)$, i.e. $\mathrm{S}(t) = \{s(t): x(t) = h(s(t))\}$. When h is smooth, $\mathrm{S}(t)$ is generically a surface of dimension $d - D$. For example, when $d = 2$ and h is projection onto the horizontal axis, the measurement surfaces consist of vertical lines. The effect of a series of measurements can be understood by transporting them to a common point in time. The state at that time must lie in their intersection $\mathrm{I}(t)$,

$$s(t) \in \mathrm{I}(t) = f^{-\tau m_{\mathrm{f}}}\mathrm{S}(t + \tau m_{\mathrm{f}}) \cap \ldots \cap \mathrm{S}(t) \cap \ldots \cap f^{\tau m_{\mathrm{p}}}\mathrm{S}(t - \tau m_{\mathrm{p}}). \qquad (9)$$

The intersection $\mathrm{I}(t)$ is never empty, since there must be at least one state consistent with all the measurements. If $\mathrm{I}(t)$ does not consist of a single point, Φ is not an embedding. If $\mathrm{I}(t)$ does consist of a single point, and if the intersection is transverse at this point, then Φ is *locally* an embedding in the neighborhood of $s(t)$. If Φ is locally an embedding everywhere, then it is a (global) embedding. The extent to which the intersection is transverse can be quantified by the singular values of the matrix $\mathrm{D}\Phi$ evaluated at $s(t)$, and will play an important role in section 4.

3. Geometry of reconstruction with noise

In the presence of noise there are many states that are consistent with a given series of measurements. The probability that a given state occurred can be characterized by a conditional probability density function $p(s|\underline{x})$. This illustrates how the presence of noise complicates the reconstruction problem: without noise a point is sufficient to characterize what is learned from a measurement, but with noise this requires a function giving the probability of all possible states. For chaotic dynamics the properties of $p(s|\underline{x})$ can be very complicated, as has been demonstrated by Geweke [21].

In this section we derive several formulae for $p(s|\underline{x})$ when h and f are known. We compute $p(s|\underline{x})$ for several examples, to illustrate qualitatively how it depends on $\underline{x}$, the noise level, and the reconstruction.

3.1. The likelihood function and the posterior

We can derive $p(s|\underline{x})$ from Bayes' theorem, making use of the fact that $p(\underline{x}|s)$ is easier to compute. According to the laws relating conditional and joint probability

$$p(s|\underline{x})\, p(\underline{x}) = p(\underline{x}|s)\, p(s). \qquad (10)$$

This can be rearranged as

$$p(s|\underline{x}) \propto p(s)\, p(\underline{x}|s). \tag{11}$$

The factor $p(\underline{x}|s)$ on the right is often called the *likelihood function*, since it represents the likelihood that the series of observations $\underline{x}$ is due to the underlying state s. Normally $p(\underline{x}|s)$ would be interpreted as a family of functions of $\underline{x}$, parameterized by the condition s; in eq. (11), however, we can regard $\underline{x}$ as given and interpret $p(\underline{x}|s)$ as a function of s. The *prior* $p(s)$ encapsulates any information that we had before these observations occurred. If we are studying a chaotic attractor, for example, and we know its natural measure, then we can take this as our prior. If we have no prior knowledge, however, then this term can be taken to be constant. The *posterior* $p(s|\underline{x})$ represents what we know about s after taking the observations $\underline{x}$ into account.

When f and h are known we can derive a formula for the likelihood function as follows. By definition we have $p(\underline{x}|s) = p(\underline{\xi})$, where $\underline{\xi} = \underline{x} - \underline{\tilde{x}} = \underline{x} - \Phi(s)$. If we assume that the noise is IID, from eq. (4) we obtain

$$\begin{aligned} p(\underline{x}|s) &= p(\underline{x} - \Phi(s)) \\ &= \prod_{i=-m_p}^{i=m_f} p\big(x(t+i\tau) - h(f^{i\tau}(s))\big). \end{aligned} \tag{12}$$

3.2. Gaussian noise

If we assume that $p(\xi)$ is a Gaussian of variance ϵ^2, eq. (12) becomes

$$p(\underline{x}|s) = \prod_{i=-m_p}^{i=m_f} \frac{1}{\sqrt{2\pi}\,\epsilon} \times \exp\left(- \frac{\big[x(t+i\tau) - h(f^{i\tau}(s))\big]^2}{2\epsilon^2} \right). \tag{13}$$

Letting $\|\cdot\|$ denote the Euclidean norm, then from the definition of Φ, eq. (13) can be rewritten as

$$p(\underline{x}|s) = A \exp\left(- \frac{1}{2\epsilon^2} \|\underline{x} - \Phi(s)\|^2 \right), \tag{14}$$

where A is a normalization constant.

Thus, $\rho(\underline{x}|s)$, interpreted as a function of $\underline{x}$, is quite simple: it is an isotropic Gaussian centered on the true delay vector $\underline{\tilde{x}} = \Phi(s)$. However, $\rho(x|s)$ interpreted as a function of s is not a Gaussian, because of the nonlinear function Φ. The probability for s given $\underline{x}$ is obtained using Bayes' theorem (eq. (11)), which gives

$$p(s|\underline{x}) = A' p(s) \exp\left(- \frac{1}{2\epsilon^2} \|\underline{x} - \Phi(s)\|^2 \right), \tag{15}$$

where A' is another normalization constant.

Eq. (15) describes how the behavior of $\Phi(s)$ determines the properties of a reconstruction. When the surface $\Phi(\mathrm{M})$ of fig. 3 is well-behaved, $p(s|\underline{x})$ is well-localized, as shown in fig. 5 for the case of a constant prior $p(s)$. However, self-intersections or regions where $\Phi(\mathrm{M})$ is tightly folded may complicate the structure of the conditional probability density $p(s|\underline{x})$. The properties of the reconstruction also depend on the stretching action of the map Φ on M.

The behavior of eq. (14) is illustrated in fig. 6, where we plot the likelihood function $p(\underline{x}|s)$ of the Ikeda map[#6] as a function of s for a fixed $\underline{x}$.

[#6] The Ikeda map is

$$\begin{aligned} &(x_{n+1}, y_{n+1}) \\ &\quad = \big(1 + \mu(x_n \cos t_n - y_n \sin t_n), \mu(x_n \sin t_n + y_n \cos t_n)\big), \end{aligned} \tag{16}$$

where $t_n = 0.4 - 6.0/(1 + x_n^2 + y_n^2)$. We take $\mu = 0.7$. The Ikeda map has an explicit inverse, and we use it in our numerical calculation of Φ. A single true state $\tilde{s}$ is randomly chosen and mapped by Φ into a noiseless delay vector $\underline{\tilde{x}}$, then perturbed by noise to obtain $\underline{x}$. For each point s on a grid, we calculate the likelihood function $p(\underline{x}|s)$ by eq. (14).

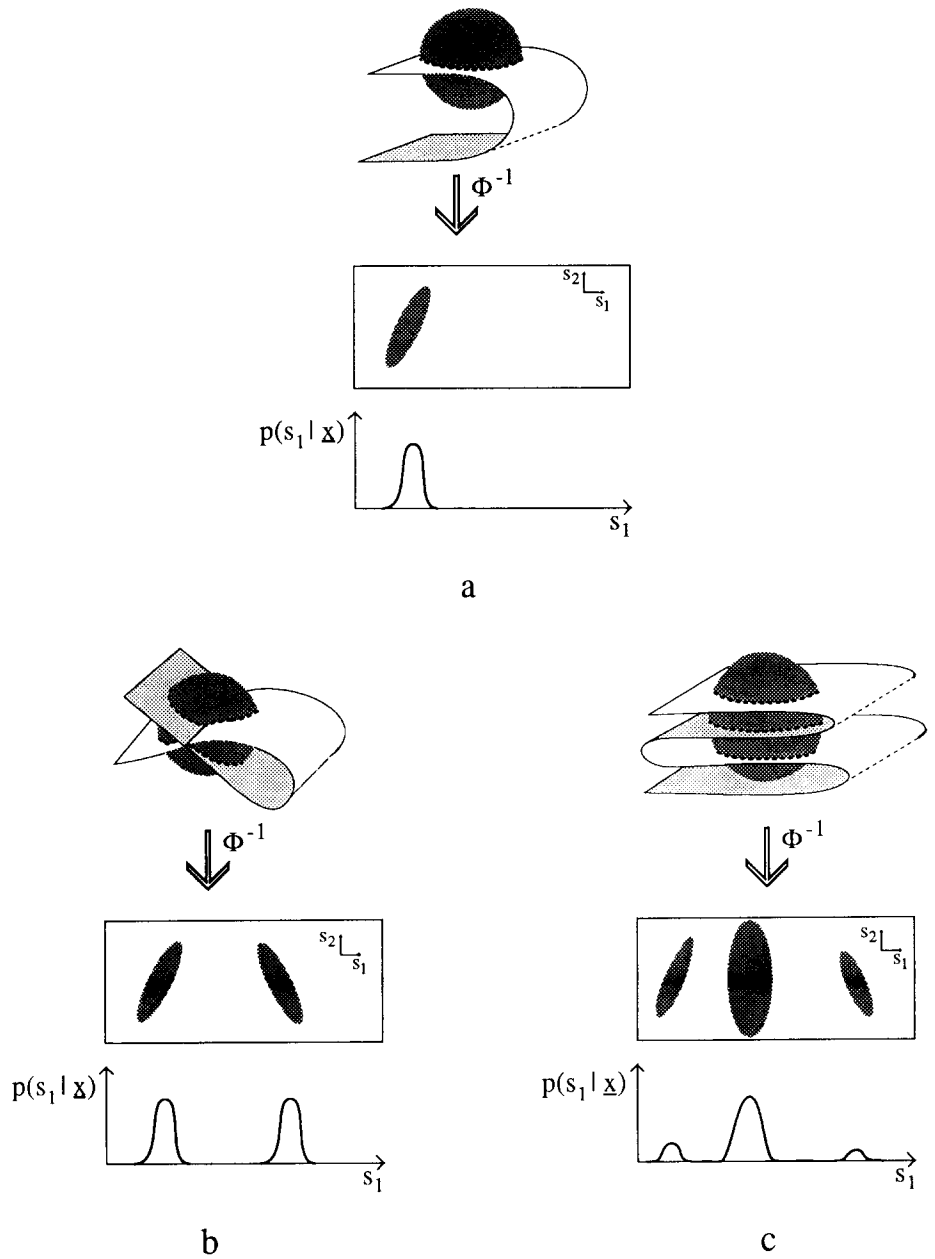

Fig. 5. Good and bad reconstructions. The quality of a reconstruction depends on the shape of the surface $\Phi(M)$. In (a) the surface $\Phi(M)$ is well-behaved within a "noise ball" of radius ϵ about the true state $\bar{s}$ and the resulting conditional probability density $p(s|\underline{x})$ is well-localized. In (b), $\bar{s}$ is near a self-intersection and $p(s|\underline{x})$ is bimodal. Even when Φ is a global embedding, problems can occur if $\Phi(M)$ is tightly folded, as illustrated in (c).

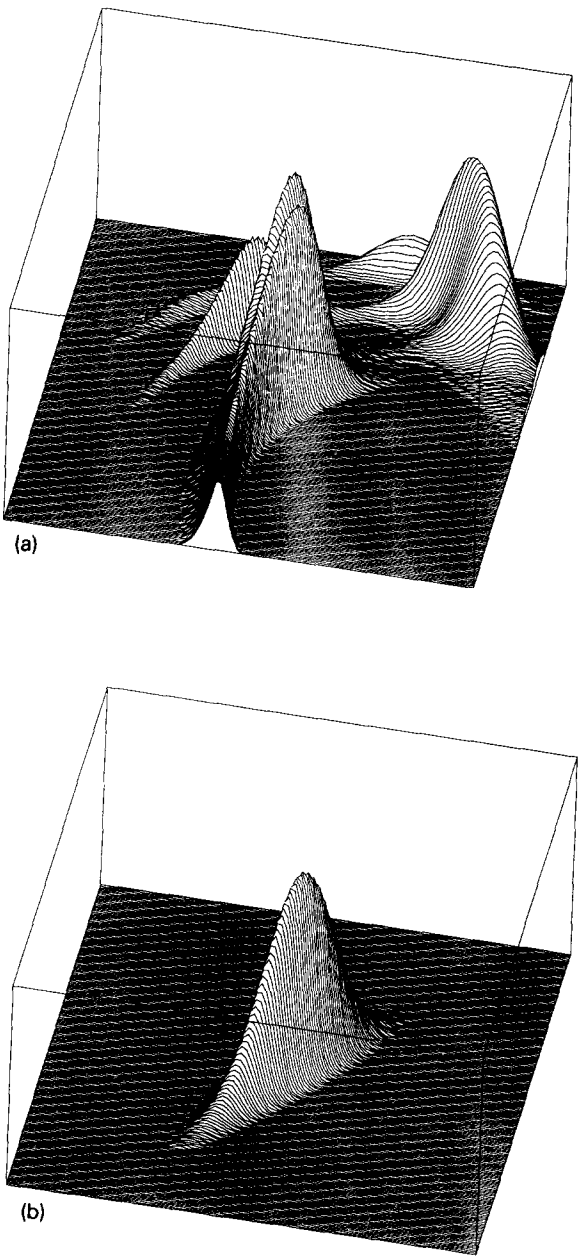

Fig. 6. Two likelihood functions for the Ikeda map, with the measurement function $h(x, y) = x$. The delay vector $\underline{x}$ is fixed, with $m_f = 2$, $m_p = 2$, and $\tau = 1$. The likelihood function $p(\underline{x}|s)$ is computed using eq. (14). The value of p is plotted vertically and $s = (x, y)$ horizontally. We assume Gaussian measurement errors with $\epsilon = 0.2$ in (a), and $\epsilon = 0.02$ in (b); the horizontal axes in (b) are blown up by a factor of 10 relative to (a). Note that in (a) p is complicated, but when the noise level is decreased in (b) it approaches a Gaussian.

Fig. 6 illustrates the case of Gaussian noise of two different variances ϵ^2, with $m_f = 2$ and $m_p = 2$. In fig. 6a we show the likelihood function for the case $\epsilon = 0.2$. With a high noise level, the likelihood function can be highly complex. In this case there are many local minima, so that it is a nontrivial task to find the maximum likelihood estimate $\hat{s}$ corresponding to the peak. In fig. 6b we show the likelihood function for the case $\epsilon = 0.02$. Here the likelihood function is approximately Gaussian.

3.3. Uniform bounded noise

Another case that is easily treated is that of uniform bounded noise of variance ϵ^2,

$$p(\xi) = 1/2\sqrt{3}\epsilon \quad \text{if } |\xi| \leq \sqrt{3}\epsilon,$$
$$= 0 \qquad \text{if } |\xi| > \sqrt{3}\epsilon. \tag{17}$$

The effect of a given measurement can be visualized geometrically in terms of the *measurement strip* $S_\epsilon(t) = \{s: |x(t) - h(s)| < \sqrt{3}\epsilon\}$. The measurement strip is the support of p, and is similar to the measurement surface $S(t)$ discussed earlier, except that it is "thickened" by ϵ. Following eq. (12), the likelihood function can be computed in a manner analogous to eq. (9). The state s must lie inside the intersection of the measurement strips,

$$s(t) \in I_\epsilon(t) = f^{-\tau m_f} S_\epsilon(t + \tau m_f) \cap \ldots \cap S_\epsilon(t) \cap \ldots \cap f^{\tau m_p} S_\epsilon(t - \tau m_p). \tag{18}$$

The likelihood function is uniform over the domain defined by $I_\epsilon(t)$, and zero outside this domain. For an invertible dynamical system, a simple method for determining whether a given point s lies within $I_\epsilon(t)$ is to test whether it

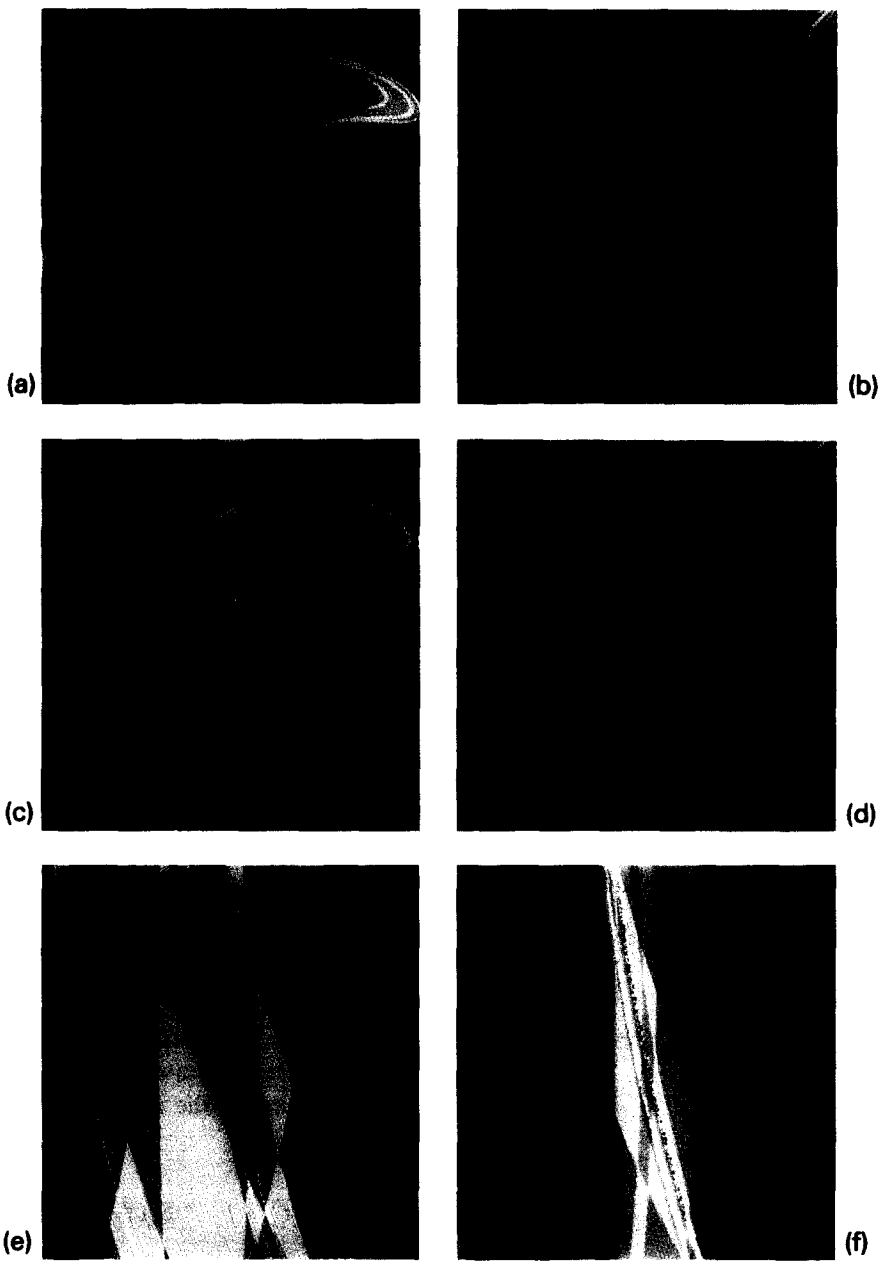

Fig. 7. State space reconstruction based on measurements of the x-coordinate of the Ikeda map with uniform noise of standard deviation 0.02. (a) and (c) are similar to fig. 4; each evolved measurement strip $f^i S_\epsilon(-i)$ is assigned a different color, with the bluest corresponding to the past (largest i), and the reddest corresponding to the future (smallest i). In (b) and (d)–(f), s is colored according to how many evolved measurement strips it lies within; blue corresponds to lying in one measurement strip, red corresponds to lying in the intersection of all the measurement strips. The red point are therfore possible states, consistent with the entire sequence of measurements. For reference, sample points on the attractor are colored white. (a), (b) have $m_f = 0$, $m_p = 2$. Figures (c), (d) have the same state $\bar{s}$ as (a), (b), but $m_f = 2$, $m_p = 2$. The scale of the first two figures on the right (b), (d) is expanded relative to (a), (c) on the left. In (e), the state $\bar{s}$ is near a homoclinic tangency, with $m_f = 2$, $m_p = 2$. (f) is the same as (e), but $m_f = 4$, $m_p = 4$.

satisfies the condition

$$f^{\tau m_f}(s) \in S_\epsilon(t + \tau m_f) \wedge \ldots \wedge s \in S_\epsilon(t) \wedge \ldots$$
$$\wedge f^{-\tau m_p}(s) \in S_\epsilon(t - \tau m_p) \qquad (19)$$

where "$\wedge$" denotes the logical "and" function.

To gain geometric insight into how the likelihood function $p(\underline{x}|s)$ is influenced by the state space reconstruction and by the properties of the dynamical system, in fig. 7 we have applied eq. (19) to the Ikeda map (eq. (16)) in a variety of different situations[#7]. As expected, in each figure there is a unique connected region of points that are in the intersection of all the evolved measurement strips. The true state lies inside this region. Figs. 7a, 7b correspond to a predictive reconstruction with $m = 3$. The likelihood function is well-localized along the stable manifold, but not along the unstable manifold. However, by using a nonpredictive reconstruction with $m_f = 2$ and $m_p = 2$, it is possible to make the likelihood function well-localized along both unstable and the stable manifolds, as shown in figs. 7c, 7d.

In fig. 7e, the state $\tilde{s}$ is near a homoclinic tangency. The likelihood function is spread out along the attractor. This is because the images of the appropriate measurement strips $S_\epsilon(i)$ intersect almost tangentially. In fig. 7f, more measurements are taken, and the likelihood function becomes more well-localized.

The geometric interplay between properties of the dynamics and properties of the reconstruction are investigated in more detail in section 5.3. However, before we can make this discussion more quantitative, we must introduce criteria for judging the localization of $p(s|x)$, and hence the quality of an embedding. This is discussed in the next section.

[#7] Figs. 7a–7f were made in the following manner: A single state $\tilde{s}$ was chosen on the attractor at random. A single noisy delay vector $\underline{x}$ was obtained from $\tilde{s}$ by iterating and applying the measurement function and then perturbing with a random number generator to generate $\underline{x}$. Then points $s \in \mathbb{R}^2$ on a 400×400 grid were tested to see how many of the individual conditions $f^i(s) \in S_\epsilon(i)$ of eq. (19) were satisfied, and colored according to the description in the caption.

4. Criteria for optimality of coordinates

As we showed in the previous section, the properties of a reconstructed coordinate system in the presence of noise depend on a conditional probability density function. To compare two functions quantitatively, we must adopt a criterion which assigns a scalar to each possible function p. In this section we discuss various criteria, and investigate the properties of the criterion that we choose.

4.1. Evaluating predictability

For convenience, we assume the current state corresponds to $t = 0$, and that predictions are desired at $t = T$. We couch the discussion in terms of a general set of coordinates $y = \Psi(\underline{x})$; for the special case of delay coordinates, Ψ is the identity.

In the previous section we discussed the reconstruction problem in terms of $p(s|y)$, the probability of the original state s given a series of measurements. This is useful for theoretical analysis, but since s is unobservable, it is inadequate for many practical purposes. For time series prediction, the probability density function that is directly relevant is $p(x(T)|y)$, the probability of a given value of the time series at a future time T. In the discussion that follows, the function p can be either $p(x(T)|y)$ or $p(s|y)$. In section 4.4 we derive a relationship relating one to the other.

4.1.1. Possible criteria

Some criteria commonly used to assess predictability are:

– *Maximum expectation*. The function p is ranked according to its maximum value. This is a criterion one might choose in a gambling prob-

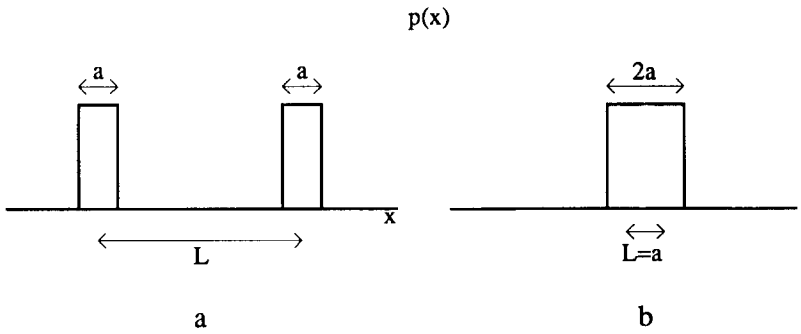

Fig. 8. Hypothetical conditional probability density functions for prediction errors. (a) is not localized, corresponding to the behavior one might expect from a reconstruction that is not an embedding. (b) is localized. The conditional variance of (a) is much higher than that of (b), but their entropies are the same. To determine whether or not a reconstruction is an embedding, conditional variance is a more sensitive test than mutual information.

lem, to maximize the expected return for a bet placed on the predicted value.

– *Mutual information*. Let H represent the entropy[#8]

$$H(x) = -\int p(x) \log p(x)\, dx. \tag{20}$$

The mutual information between the variables x and y is $I(x, y) = H(x) - H(x|y)$, where $H(x|y)$ is the entropy associated with the conditional probability density $p(x|y)$ averaged over y.

– *Mean-square error (conditional variance)* is defined as

$$\mathrm{Var}(x|y) = \int x^2 p(x|y)\, dx - \left(\int x p(x|y)\, dx\right)^2. \tag{21}$$

$\mathrm{Var}(x|y)$ measures the mean-square errors in x given y, and depends on the value taken on by y (a quantity analogous to mutual information could be defined by integrating over y). Since the expectation $\hat{x} = \int x p(x|y)\, dx$ minimizes mean-square prediction errors [35], $\mathrm{Var}(x|y)$ is a lower bound on the mean-square prediction error. If x is vector-valued, then eq. (21) is modified so that $\mathrm{Var}(x|y)$ is a covariance matrix.

[#8]Note that the entropy is actually a functional of $p(x)$ rather than a function of x.

– *Mean-absolute error*. The arithmetic mean-absolute error or geometric mean-absolute error are other common measures of predictability.

4.1.2. Comparison of criteria

Intuitively, for prediction of a continuous variable, the conditional probability p should be as well-localized as possible. Criteria such as mean-square error or mean-absolute error enforce this. In contrast, maximum expectation and mutual information do not enforce localization. Because of this they are more appropriate for discrete variables[#9]. For example, consider the probability density function

$$\begin{aligned} p(x) &= 1/2a \quad |x - \tfrac{1}{2}L| < \tfrac{1}{2}a \text{ or } |x + \tfrac{1}{2}L| < \tfrac{1}{2}a, \\ &= 0 \qquad \text{otherwise}, \end{aligned} \tag{22}$$

shown in fig. 8 for two values of L. The entropy for this density is $H = \log(2a)$ and its variance is $\frac{1}{4}(L^2 + \frac{1}{3}a^2)$. Any of the criteria based on mean errors will assign a low value to fig. 8b, and a high

[#9]At any finite level of resolution, x and y may be thought of as "messages", with a given number of bits [38, 39]. The mutual information gives the average uncertainty for predicting message x from message y. It weights the low order bits equally with the high order bits. In predicting a continuous variable, however, the consequences of an error in the highest order bit are usually worse than one in the lowest order bit. The fact that mutual information does not make this distinction makes it a poor predictability criterion for continuous variables.

value to fig. 8a. This is in accord with the fact that 8b is well-localized and 8a is not. However, the mutual information for figs. 8a and 8b is the same, and so is the maximum expectation. Criteria based on mean errors are better at evaluating localization, and hence are better for detecting whether or not a reconstruction is an embedding.

The requirement of locality leads us to choose mean errors as our criterion for predictability. Mean-square error as compared to mean-absolute error has the disadvantage that it over-emphasizes outliers. However, it has the important advantage that, when used in conjunction with Gaussian noise, many computations can be performed in closed form, a property of which we make much use in the next sections. Thus, locality and computational tractability are our primary reasons for using mean-square error to select reconstructed coordinates.

4.1.3. Previous work

Conditional variance[#10] was originally suggested as a criterion for reconstruction by Packard et al. [33]. This was developed by Čenys and Pyragas [9], who used a more efficient method of estimating it, and considered scaling with the estimator resolution and τ. Variations which amount to different estimators of conditional variance or related quantities, have also been suggested by Guckenheimer [24], Liebert et al. [30], Aleksić [2], and Savit and Green [37].

Shaw [39] originally suggested that the best coordinates should be those that maximize the mutual information between past and future states. This was pursued by Fraser and Swinney [19]. However, they did not compute it for the full reconstructed state space. Instead, they computed $I(x(\tau), x(0))$. This amounts to the mutual information between past and future in a one-dimensional projection of the dynamics. They then proposed that the value of τ corresponding to the first minimum of $I(x(\tau), x(0))$ should be a good choice for delay coordinates. They justified this procedure on the grounds that a small value of $I(x(\tau), x(0))$ implies that $x(0)$ is statistically independent of $x(\tau)$, minimizing the redundance of the coordinates. There are several problems, though: There is no obvious reason to prefer the first minimum of $I(x(\tau), x(0))$ over others, and $I(x(\tau), x(0))$ may not even have any minima at finite τ. Fraser [17] later proposed another heuristic quantity, which was designed to provide a compromise between redundance and relevance and to be applicable to higher dimensional systems. However, the connection with Shaw's original criteria of *maximizing* the mutual information between the past and the future is unclear. Finally, there are the problems with using mutual information for continuous variables mentioned in the previous section.

Another heuristic which is sometimes used is to choose τ at the first minimum of the autocorrelation function, or alternatively, to choose a value of τ that makes the autocorrelation function "small". This has some justification from the point of view of minimizing linear redundance. However, in general a statistic such as the correlation function that measures only linear dependence is simply inadequate, as discussed in section 5.2.

4.2. Noise amplification

As we argued in section 4.1.2, a natural criterion for assessing predictability is the variance of the conditional probability density function $p(x(T)|y)$. This quantity can be interpreted as measuring the thickness of the points in fig. 2 in the vertical direction. The conditional variance depends on the noise level ϵ. When the reconstruction is an embedding, for small ϵ the conditional variance is asymptotically proportional to ϵ^2. The constant of proportionality quantifies the predictive value of the reconstructed coordinate y at a given noise level. When the constant of proportionality is large, then the reconstructed coordinates amplify noise.

[#10] Estimators of conditional variance can be used to measure the total prediction error, which is a combination of effects due to estimation error and noise. See section 7.

This motivates us to define the *noise amplification at a given noise level* ϵ as

$$\sigma_\epsilon(T) = \frac{1}{\epsilon}\sqrt{\mathrm{Var}(x(T)|y)}\,, \tag{23}$$

where for convenience we have suppressed the dependence of $\sigma_\epsilon(T)$ on y. We define the *noise amplification* σ by taking the limit $\epsilon \to 0$,

$$\sigma(T) = \lim_{\epsilon \to 0} \sigma_\epsilon(T). \tag{24}$$

The noise amplification $\sigma(T)$ characterizes the predictive value of a reconstructed coordinate y. In contrast to the conditional variance, it is *independent of the noise level* ϵ. It depends on purely geometric factors, such as the dynamical system, the measurement function, and the reconstruction. Taking the limit as the noise goes to zero is quite different from simply *setting* the noise to zero, as was effectively done by Takens [41]. When the noise is set to zero, all reconstructions that are embeddings are equivalent. In the *limit* as the noise goes to zero, however, two embeddings may have quite different noise amplifications.

The limit involved in defining $\sigma(T)$ may not always exist; for example, it does not exist when the reconstruction is not an embedding. There are other situations where it does not exist because σ_ϵ oscillates in the limit as $\epsilon \to 0$. This is true for highly regular fractals, for example, a simple Cantor set. In these cases, $\sigma(T)$ can be made well-defined by replacing the simple limit with a limit of the supremum.

If we are interested in a geometric object with an ergodic measure, such as a chaotic attractor, we can also eliminate the dependence on the state y by taking an average over the values of y with respect to this measure. We will call this the *average nose amplification*:

$$\langle \sigma \rangle^2 = \langle \sigma^2(y) \rangle_y. \tag{25}$$

For some purposes, such as noise reduction, we wish to predict the *true* value $\tilde{x}(T)$, i.e. the value of $x(T)$ in the absence of noise. In this case we can define a quantity $\tilde{\sigma}$ in terms of $\mathrm{Var}(\tilde{x}(T)|y)$, by analogy with eqs. (23) and (24). Since $x(t) = \tilde{x}(t) + \xi(t)$, it follows that

$$\tilde{\sigma}^2 = \sigma^2 - 1. \tag{26}$$

4.3. Distortion

For many purposes it is useful to consider how the uncertainties in a reconstructed state y are manifested in the original state s. Although the probability density of the noise is isotropic in delay coordinates, in the original state space it is typically anisotropic. This was illustrated in fig. 6b. For example, for Gaussian noise the surface on which the probability density function $p(\underline{x}|s)$ is a constant is an m-dimensional sphere. If Φ is an embedding, in the low noise limit the intersection of this sphere and $\Phi(\mathrm{M})$ will map into a d-dimensional ellipsoid in the original state space M, as was illustrated in fig. 5a. The noise distribution is thus "distorted" when transformed to the original state space.

We define the *distortion matrix at noise level* ϵ as

$$\Sigma_\epsilon = \frac{1}{\epsilon^2}\mathrm{Var}(s|y). \tag{27}$$

The dependence on ϵ can be removed by taking the limit as $\epsilon \to 0$,

$$\Sigma = \lim_{\epsilon \to 0} \Sigma_\epsilon. \tag{28}$$

The *distortion matrix* Σ is a $d \times d$ symmetric real matrix, whose eigenvalues are proportional to the squares of the lengths of the principal axes of the distorted ellipsoid in the original space.

The distortion matrix describes the noise amplification in each direction in d dimensions. For

an overall summary, it is often more convenient to consider

$$\delta_\epsilon = \sqrt{\mathrm{Tr}\,\Sigma_\epsilon} = \frac{1}{\epsilon}\sqrt{\mathrm{Var}(\|s\|\|y)}\,. \tag{29}$$

We have taken the square root to make it easier to compare with noise amplification. As before, we can eliminate the dependence on ϵ by taking the limit as $\epsilon \to 0$. We call δ the *distortion*[#11]

$$\delta = \lim_{\epsilon \to 0} \delta_\epsilon. \tag{30}$$

Compared with noise amplification, the distortion has the advantage that it does not depend on the extrapolation time T. However, it has two disadvantages: First, it depends on the coordinates used to describe the dynamical system[#12]; for example, rescaling s changes the distortion. Second, it is not observable, and cannot be computed from a time series alone. Nonetheless, the distortion matrix is a valuable tool because of its relation to noise amplification, as shown in section 4.4.

In addition, the distortion is of interest in its own right. In some engineering problems the form of f and h is known, and it is desirable to estimate the "hidden variables" s, or to estimate the unknown parameters of f and h, from a noisy time series. For example, in section 1, we considered how accurately z could be inferred from x for the Lorenz equations. This is a problem sometimes faced in extended Kalman filtering, and has also been considered by Breeden et al. [4].

4.4. Relation between noise amplification and distortion

In the low noise limit, there is a simple relation between noise amplification and distortion. Let a variation of $x(T)$ about its true value $\tilde{x}(T)$ be $\Delta x = x(T) - \tilde{x}(T)$, and similarly let $\Delta s = s - \tilde{s}$. When Δs is small, $\Delta x \approx \mathrm{D}h\,\mathrm{D}f^T \Delta s + \xi(T)$. The noise amplification at resolution ϵ is

$$\begin{aligned}\sigma_\epsilon^2(T) &= \frac{1}{\epsilon^2}\langle \Delta x\, \Delta x^\dagger \rangle \\ &\approx \frac{1}{\epsilon^2}\Big\langle \big[\mathrm{D}h\,\mathrm{D}f^T \Delta s + \xi(T)\big] \\ &\quad \times \big[\mathrm{D}h\,\mathrm{D}f^T \Delta s + \xi(T)\big]^\dagger \Big\rangle. \end{aligned} \tag{31}$$

By definition $\Sigma_\epsilon = (1/\epsilon^2)\langle \Delta s\, \Delta s^\dagger \rangle$, and $\langle \xi^2 \rangle = \varepsilon^2$. Since Δs and $\xi(T)$ are independent this implies, on taking the limit $\epsilon \to 0$,

$$\sigma^2(T) = 1 + \mathrm{D}h\,\mathrm{D}f^T\, \Sigma \big(\mathrm{D}f^T\big)^\dagger\, \mathrm{D}h^\dagger. \tag{32}$$

Intuitively this makes sense; the uncertainty in the initial state is first altered by the derivative of the dynamics, then projected down onto the time series. The first term is the result of convolution with noise.

4.5. Low noise limit

When Φ is an embedding, the likelihood function $p(\underline{x}|s)$ has a simple form in the low noise limit. This was illustrated for the Ikeda map in fig. 6b. In this section, we derive analytical formulae for the distortion matrix in the case of Gaussian noise with a uniform prior.

With the assumption of a constant prior $p(s)$, eq. (15) can be rewritten as

$$p(s|\underline{x}) = A\, \mathrm{e}^{-Q(s)/2\epsilon^2}, \tag{33}$$

where A is a normalization constant and $Q(s) = \|\underline{x} - \Phi(s)\|^2$. If f and h are smooth then Q is also smooth. When Φ is an embedding and ϵ is small enough, $p(s|\underline{x})$ has a unique maximum $\hat{s}$, called

[#11] The term "distortion" was originally used for another related quantity defined by Fraser [18].

[#12] The noise amplification depends on the coordinates of $x(t)$, but, as long as these are fixed, it does not depend on the coordinates of s.

the *maximum likelihood estimate*. In this case it is possible to get a good approximation for $p(s|\underline{x})$ by expanding Q in a Taylor series about $\hat{s}$, making use of the fact that $DQ(\hat{s}) = 0$:

$$Q(s) = Q(\hat{s}) + \tfrac{1}{2}(s-\hat{s})^{\dagger} D^2Q(\hat{s})(s-\hat{s}) + \ldots . \tag{34}$$

To differentiate Q, we take advantage of the fact that it is of the form $Q = v^{\dagger}v$, where $v = \underline{x} - \Phi(s)$. Differentiating gives $DQ = Dv^{\dagger}v + v^{\dagger}Dv = 2Dv^{\dagger}v$, and $D^2Q = 2[(D^2v^{\dagger})v + Dv^{\dagger}Dv]$. Since v is of order ϵ, while $Dv = D\Phi$ is typically of order one, $D^2Q(\hat{s}) \approx 2D\Phi^{\dagger}D\Phi$. To leading order in $s - \hat{s}$, this gives

$$p(s|\underline{x}) \approx A' \exp\left(-\frac{1}{2\epsilon^2}(s-\hat{s})^{\dagger} D\Phi^{\dagger} D\Phi (s-\hat{s})\right), \tag{35}$$

where A' is a normalization constant, which in the limit $\epsilon \to 0$ becomes equal to A in eq. (33). The variance is $\mathrm{Var}(s|\underline{x}) = \epsilon^2(D\Phi^{\dagger}D\Phi)^{-1}$. By definition (eqs. (27) and (28)) the distortion matrix is

$$\Sigma = (D\Phi^{\dagger}D\Phi)^{-1}. \tag{36}$$

The derivative $D\Phi$ is evaluated at $s = \hat{s}$, which depends on the particular realization $\underline{\xi}$ of the noise that gave rise to $\underline{x}$. However, $\hat{s} - \tilde{s}$ is almost always of order ϵ. Since $D\Phi(\hat{s}) = D\Phi(\tilde{s}) + D^2\Phi(\tilde{s})(\hat{s} - \tilde{s}) + \ldots$, from the definition of the distortion matrix it follows that, to leading order, $D\Phi^{\dagger}(\hat{s})D\Phi(\hat{s}) \approx D\Phi^{\dagger}(\tilde{s})D\Phi(\tilde{s})$. Thus, taking the limit as $\epsilon \to 0$, the distortion does not depend on the realization. We make use of this fact in numerical experiments, in which we compute the distortion matrix by evaluating the derivative $D\Phi$ at $s = \tilde{s}$.

Note that if Φ is an embedding then $D\Phi$ is of full rank and Σ is well-defined. At low noise levels the uncertainty in the estimate of s is approximately an anisotropic Gaussian of covariance matrix $\epsilon^2\Sigma$, centered on the maximum likelihood estimate $\hat{s}$. This was illustrated in fig. 6b. Small eigenvalues of Σ imply that the Gaussian is sharply peaked.

4.6. The observability matrix

Since Φ is the vector function whose components are $\Phi_i = h(f^{i\tau})$, according to the chain rule the components of the derivative are $D\Phi_i = Dh\,Df^{i\tau}$. When the measurement function h is one-dimensional, $D\Phi$ is the $m \times d$ matrix

$$D\Phi = \begin{pmatrix} Dh\,Df^{\tau m_1} \\ \vdots \\ Dh \\ \vdots \\ Dh\,Df^{-\tau m_p} \end{pmatrix}. \tag{37}$$

As long as Φ is an embedding, $D\Phi$ has d nonzero singular values. The inverse squares of these singular values are equal to the eigenvalues of Σ. We often use this fact to compute the distortion δ directly from the singular values of $D\Phi$.

The matrix $D\Phi$ has a simple interpretation. In control theory it is called the *observability matrix*. For a system to be observable, in the sense that inferences about the state s can be made from the time series, the observability matrix must have full rank. This is one of the conditions for Φ to be an embedding. Whether $D\Phi$ has full rank depends on detailed properties of the coupling between variables in f, and on the measurement function h. For example, if the dynamical system f can be split into two noninteracting subsystems, and h measures only one of them, the other subsystem is unobservable. All the columns of the observability matrix corresponding to this subsystem are zero, and $D\Phi$ is not of full rank. On the other hand, if the measurement function depends on both subsystems, or if they are coupled, then from Takens' theorem $D\Phi$ is generically of full rank.

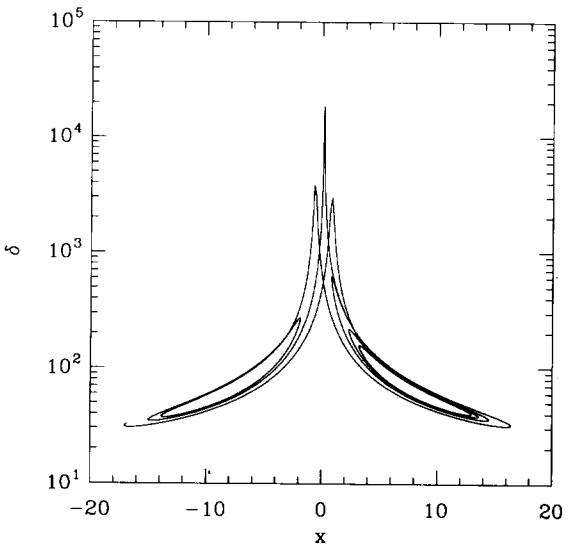

Fig. 9. The distortion computed along a typical trajectory of the Lorenz equations, using five dimensional delay coordinates with $m_f = 0$, $m_p = 4$ and $\tau = 0.01$.

4.7. State dependence of distortion

When f and h are known, the distortion matrix can easily be computed using eqs. (36), (37). This provides a useful quantitative tool for understanding the properties of a reconstruction. For example, we can now make the discussion of information flow in the Lorenz equations from section 1 more precise by simply computing the distortion δ: Let h be projection onto the x axis. The dynamics f^T can be computed by numerically integrating the Lorenz equations[#13]. The distortion δ along a typical trajectory is shown in fig. 9. The graph is multi-valued, since δ depends on y and z as well as x. The blowup of the distortion at $x = 0$ is a result of the poor information flow from z to x when $x \approx 0$. Note that when τ is small, all the coordinates in the delay vector are sometimes near zero simultaneously; when τ is large the blowup is less severe.

4.8. Comparison of finite noise and the zero noise limit

At small noise levels, σ, which is computed from purely deterministic quantities, can be used to estimate the noise amplification σ_ϵ at finite noise levels. In this section, for the Lorenz equations we numerically investigate the accuracy of this approximation. Since this numerical experiment involves a long time integration of the Lorenz equations, it is natural to take the prior $p(s)$ to be the natural measure on the Lorenz attractor.

To compute the distortion at finite noise levels we make use of eq. (15), which gives an exact formula for $p(s|\underline{x})$ in terms of Φ and $p(s)$. Φ is known from the dynamics, and $p(s)$ can be estimated numerically by computing a time average[#14]. In order to compute the conditional variance as defined in eq. (21), we compute time averages of $\phi_1(s) = \|s\|^2 p(\underline{x}|s)$ and $\phi_2(s) = sp(\underline{x}|s)$. For fixed $\underline{x}$, the likelihood function $p(\underline{x}|s)$ is proportional to $\omega_i \equiv \exp[-\|\underline{x} - \Phi(s(t_i))\|^2/2\epsilon^2]$, where $s(t_i) = f^{i\tau}(s_0)$. Putting these statements together gives

$$\epsilon^2 \delta_\epsilon^2 = \lim_{N\to\infty} \frac{\sum_{i=1}^N \|f^{i\tau}(s_0)\|^2 \omega_i}{\sum_{i=1}^N \omega} - \left\| \frac{\sum_{i=1}^N f^{i\tau}(s_0)\omega_i}{\sum_{i=1}^N \omega_i} \right\|^2 . \tag{39}$$

The terms in the denominators make sure that this is properly normalized. For a numerical approximation, N is taken large enough for convergence. Note that the smaller ϵ is, the larger N must be for convergence.

[#13]The derivative matrix $Df^{-i\tau}$ of the map associated with the Lorenz equations is found by integrating the equations for the differentials, as is done in computing Lyapunov exponents. For numerical stability, we are often forced to integrate forwards. We then use singular value decomposition to invert the resulting matrices. Finally, we compute the distortion from the singular value decomposition of the matrix $D\Phi$.

[#14]Since the system is ergodic, we can compute an ensemble average of any function ϕ by a time average

$$\int \phi(s)\, p(s)\, ds = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \phi\big(f^{i\tau}(s_0)\big). \tag{38}$$

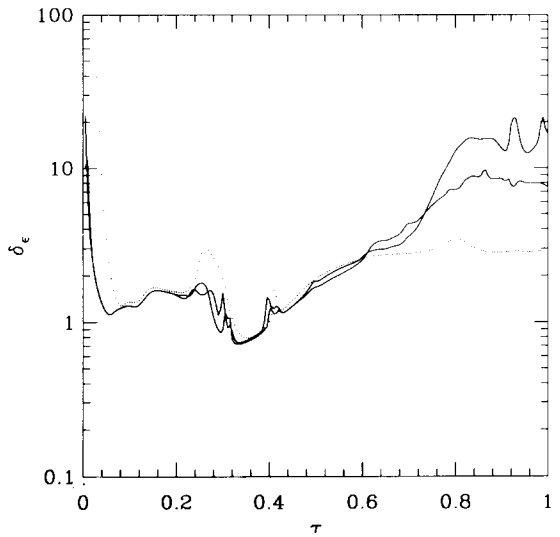

Fig. 10. δ_ϵ at finite resolution ϵ as a function of τ for the Lorenz equation. The solid lines are for $\epsilon = 0.5$ and $\epsilon = 0.25$. The dotted line is for the limit $\epsilon \to 0$. All of these are for a predictive embedding with $m = 5$, and a fixed state $(-1.8867, -5.1366, 24.7979)$.

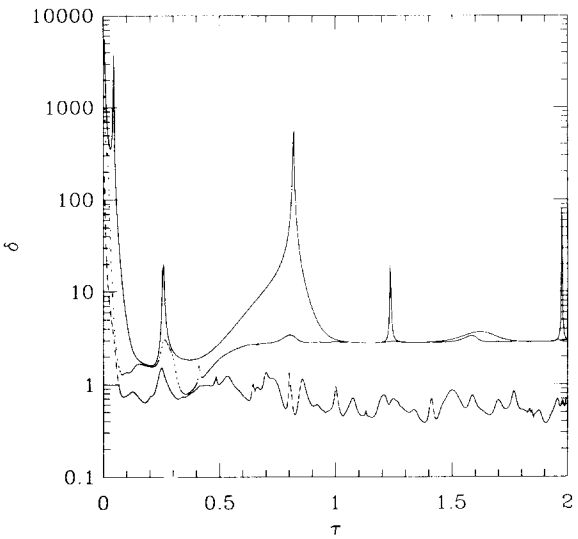

Fig. 11. The distortion of the Lorenz equations as a function of the lag time τ. We arbitrarily fix the true state as in fig. 10. The upper curve corresponds to a reconstruction with $m_f = 0$ and $m_p = 2$; the singularities occur because the embedding dimension $m = d = 3$ is too low. The middle curve is for $m_f = 0$ and $m_p = 4$, and the lower curve is a mixed reconstruction with $m_f = 5$ and $m_p = 4$. The third reconstruction incorporates both past and future information, and yields a lower distortion.

Fig. 10 shows the distortion δ_ϵ as a function of τ at finite noise levels corresponding to signal to noise ratios of about 20 and 40. This is compared to the low noise limit distortion δ as computed from eq. (36). Note that for roughly $0 < \tau < 0.5$, δ_ϵ has converged quite well. Through this range δ provides a good upper bound for δ_ϵ. However, δ does not always provide a good *approximation* to δ_ϵ, because a uniform prior was assumed in the analysis of section 4.5. The low noise limit approximation breaks down for $\tau > 0.5$. We believe this is due to the phenomenon of multimodality, illustrated in fig. 5c, which cannot be approximated using the local analysis of section 4.5.

4.9. *Effect of singularities*

When the embedding dimension $m < 2d$ there may be points where $D\Phi$ is not of full rank. These cause singularities in the distortion. For example, in fig. 11 we compute the distortion as a function of τ for several different embeddings. There are three reconstructions shown: for the first $m = 3$, which is too low, and δ is singular for several values of τ. For the second $m = 5$, and the singularities disappear. When $m \leq 2d$, a state space average of Σ is not well defined unless the singularities of Σ are integrable. We believe that the singularities are generically integrable as long as $m \geq d + 1$.

5. Parameter dependence and limits to predictability

The noise amplification depends on properties of the reconstruction, such as m_f, m_p, and τ, as well as properties of the problem, such as the measurement function and dynamical system. Understanding the dependence on the reconstruction provides guidance for constructing the best possible coordinates. The properties of the dynamical system, such as the dimension and Lyapunov exponents, along with the measurement function, determine the limits to

predictability. In appropriate limits these dependencies can be characterized by scaling laws.

One of the interesting results that emerges from our analysis is that in some situations the noise amplification is so large that determinism is completely lost. This result is important because it shows how the projection of a chaotic dynamical system onto a low dimensional time series can generate an irreducible random process which is unpredictable except for very short times, much shorter than the Lyapunov time, $\log(1/\epsilon)/\lambda$.

For convenience we state most of our results in terms of distortion rather than the noise amplification, since distortion does not depend on the extrapolation time T. Distortion and noise amplification are simply related by eq. (32), and we discuss effects relating to the extrapolation time T in section 5.5. Also, in this section we study only delay coordinates. As already mentioned, delay coordinates determine the information set on which the reconstruction is based. As we demonstrate in section 6, the choice of the information set provides a lower bound on the distortion, so delay coordinates alone are sufficient to give us an understanding of the limitations to general state space reconstruction in the presence of noise.

5.1. More information implies less distortion

We define an ordering on distortion matrices as follows: $\Sigma_1 \leq \Sigma_2$ if $\Sigma_2 - \Sigma_1$ is positive semi-definite[#15]. One fact that is immediately apparent is that *gathering more information can only decrease the distortion matrix*. Suppose we are given two delay vectors $\underline{x}^{(1)}$ and $\underline{x}^{(2)}$ for which $\underline{x}^{(1)} \subset \underline{x}^{(2)}$, i.e. $\underline{x}^{(2)}$ is of higher dimension than $\underline{x}^{(1)}$, and contains $\underline{x}^{(1)}$ as a subset. Then, letting $\Sigma^{(1)}$ be the distortion matrix associated with $x^{(1)}$, and similarly for $x^{(2)}$, we have

$$\Sigma^{(2)} \leq \Sigma^{(1)}. \tag{40}$$

[#15]By definition a $d \times d$ matrix M is positive semi-definite if $v^{\dagger} M v \geq 0$ for all vectors $v \in \mathbb{R}^d$.

This follows from an elementary property of the conditional probability density function $p(s|\underline{x}^{(i)})$. The more conditions that are imposed, the more sharply localized is the state s. Thus, the distortion is a monotonic nonincreasing function of the dimensions m_f and m_p. The distortion can typically be reduced by increasing the dimension of the reconstructed space.

It should be kept in mind that, with finite data, prediction error depends on the estimation error as well as distortion. While distortion decreases with m, estimation error *increases*. To make the best possible predictions requires an optimal compromise between distortion and estimation errors. In this section we focus our attention on the behavior of the error due to distortion, and address the problem of estimation error in section 7.

5.2. Redundance and irrelevance

The distortion is strongly influenced by two effects that we call redundance and irrelevance. For a smooth time series, measurements with τ very small are *redundant*. Geometrically this means that measurement surfaces corresponding to successive measurements are roughly parallel near the true state, as illustrated in fig. 12b. Because these surfaces intersect at a small angle, the intersection of the corresponding noisy measurement strips is delocalized along one or more directions, even for small noise levels ϵ. We call the characteristic time for this to occur the *redundance time* τ_R. It depends on ϵ, as will be made precise in section 5.3. If the *window width* $w = (m-1)\tau < \tau_R$, then the distortion is very large.

At the other extreme, for a chaotic system with predictive coordinates, measurements made in the distant past are irrelevant. When transported to the present, the associated measurement strips collapse onto the unstable manifold in the vicinity of the true state. This is illustrated in fig. 12a, and was also illustrated earlier in fig. 7b. While measurements from the distant past may determine the state arbitrarily accurately along the stable

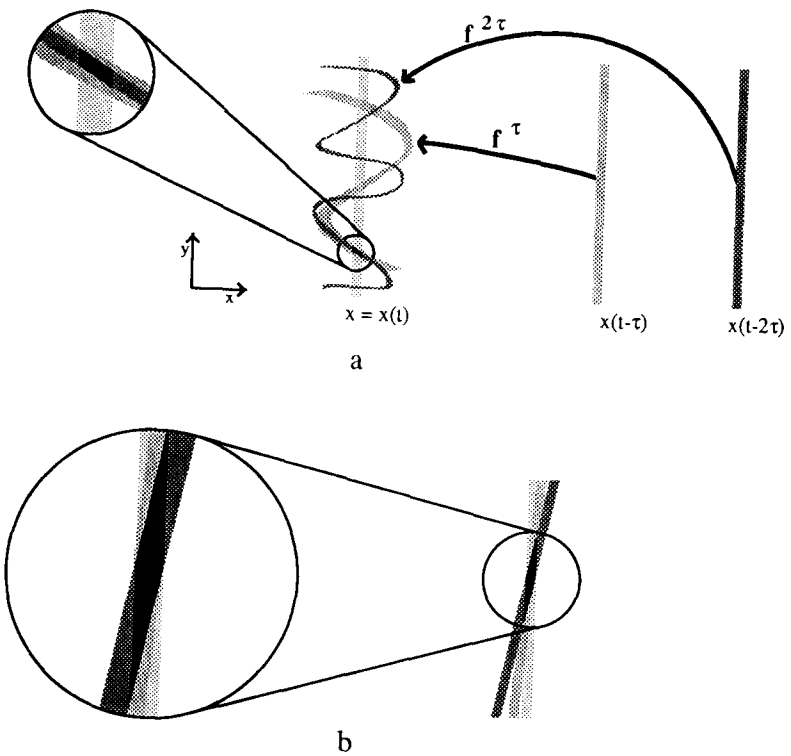

Fig. 12. Redundance and irrelevance. Images of measurement strips $S_\epsilon(t - i\tau)$, transported to the same time t. (See fig. 4.) (a) illustrates irrelevance; τ is large, and f^τ is highly nonlinear. The measurement strips are complicated. Strips from the distant past, with large $i\tau$, are roughly parallel along the unstable manifold near the true state $\tilde{s}$. Increasing $i\tau$ better determines the state along the stable manifold, but gives no new information about the unstable manifold. Thus at a finite level of coarse-graining, measurements from the distant past are irrelevant, since the limiting factor is determination along the unstable manifold. (b) illustrates redundance: When τ is small f^τ is close to the identity, and is approximately linear, so that the images of the measurement strips are nearly parallel at time t. Their intersection is delocalized, making the conditional variance large.

manifold, the eigenvalues of the distortion matrix associated with the unstable manifold reach a limiting value. As we prove later, for large times the eigenvectors of the distortion matrix are related to those associated with the Lyapunov exponents. We call the *irrelevance time* τ_I the time when measurement strips become effectively tangent relative to the noise level ϵ, so that making $w > \tau_I$ gives no significant decrease in the leading eigenvalue of the distortion matrix[#16].

[#16]The irrelevance time is related to the uncertainty time for prediction, $-\log \epsilon/\lambda$. However, the irrelevance time depends on other geometric factors, such as rotation rates onto the unstable manifold, and is more complicated.

5.3. Scaling laws

5.3.1. Overview

In certain limits the distortion behaves according to well-defined scaling laws. There are several distinct scaling regimes, which are organized schematically in fig. 13. As shown in the diagram, the scaling regime depends on the window width, the redundance time, whether the dynamics are chaotic, whether the coordinates are predictive, and whether $\tau_R > \tau_I$. An example that illustrates several distinct scaling regimes is shown in fig. 14. We will describe this example and consider it in some detail in section 5.4.

For an overview of the scaling behavior see figs. 13 and 14. The scaling laws quoted are

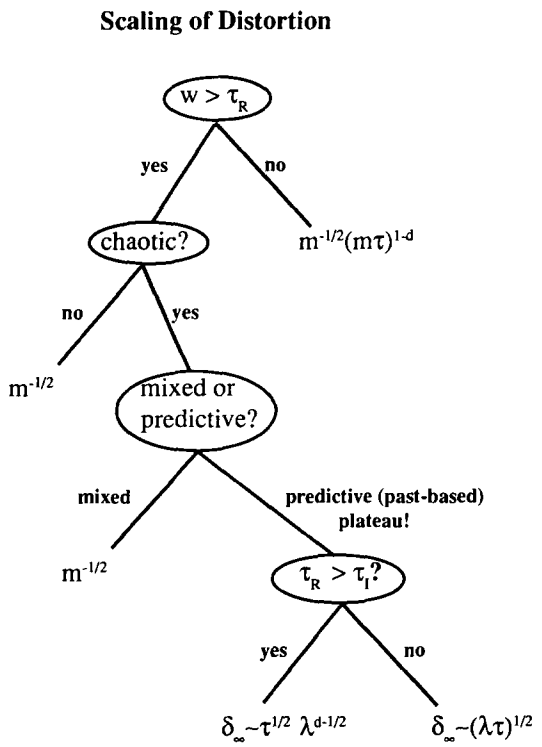

Fig. 13. The scaling regimes of the distortion are defined according to the conditions shown. w is the window width, τ is the lag time, m is the delay coordinate dimension, τ_R is the redundance time, τ_I is the irrelevance time, and δ_∞ is the distortion in the limit as $m \to \infty$.

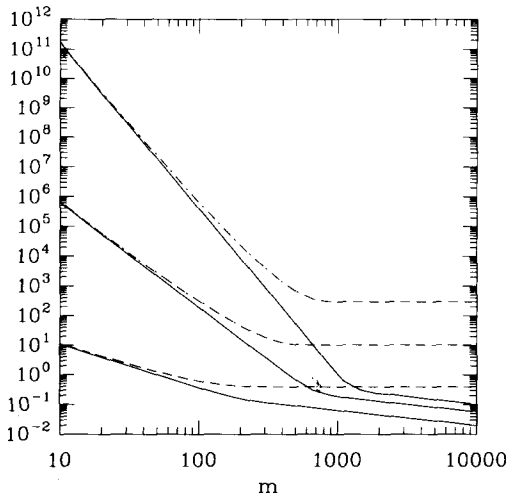

Fig. 14. The distortion δ as a function of the delay coordinate dimension m for the system defined by eqs. (60) and (61). The reconstruction uses predictive coordinates with a fixed delay time $\tau = 0.01$. For the dashed curves the Lyapunov exponents $\lambda_i = \lambda = 1$ and the system is chaotic, while for the solid curves $\lambda = 0$ and the system is not chaotic. Three different dimensions are shown, $d = 2$, 4, and 6; the curves with larger distortion have higher dimension. For small $m, w < \tau_R$, and the behavior is dominated by the effect of redundance; for large m, when $\lambda > 0, w > \tau_I$, and the behavior is dominated by irrelevance.

derived later in this section. When w is small the behavior is dominated by redundance. This is seen for small m in fig. 14. In the limit as $w \to 0$ and $m \to \infty, \delta \sim m^{-1/2}(m\tau)^{1-d}$, independent of any other conditions. This gives a quantitative explanation for the well-known observation that making τ too small results in poor coordinates. The exponent that determines the rate at which the distortion blows up in this limit is proportional to the dimension, so this effect is much worse in higher dimensional systems. This is apparent in fig. 14.

When w is large and the dynamics are not chaotic the distortion goes to zero as $m \to \infty$. The ability to isolate a state is determined by the central limit theorem, and the distortion goes to zero as $\delta \sim m^{-1/2}$. This behavior is seen for large m in fig. 14. Note that the distortion also increases with d.

For a chaotic system with mixed coordinates, different eigenvalues of the distortion matrix exhibit different scaling behavior. Past data can be used to localize the state along the stable manifold; the distortion in this direction decreases exponentially according to $\lambda_i m_p \tau$, where λ_i is the associated Lyapunov exponent. Similarly, future data can be used to localize the state along the unstable manifold; the distortion in this direction decreases exponentially according to $-\lambda_i m_f \tau$. For a dynamical system that has any Lyapunov exponents equal to zero (a continuous-time system always has at least one), the eigenvalues of the distortion matrix associated with the neutral manifold go to zero as m^{-1}, following the central limit theorem (recall that δ involves a square root, whereas Σ does not). Since this decrease

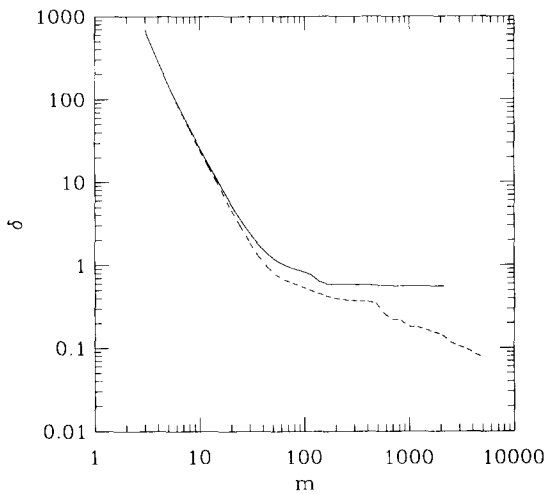

Fig. 15. The distortion of the Lorenz equations as a function of the delay dimension m, with $\tau = 0.005$, and fixed $\bar{s}$ (different from that of fig. 11). The solid curve is for a predictive embedding with $m_f = 0$, and the dashed curve is for a mixed embedding with $m_f = \frac{1}{2}m$ and $m_p = \frac{1}{2}m - 1$.

follows a power law rather than an exponential, when there are zero exponents the neutral manifold is the limiting factor on the distortion.

For chaotic systems with predictive coordinates the behavior is dominated by the fact that successive measurements are irrelevant to the position on the unstable manifold. The distortion approaches a constant $\delta_\infty > 0$, as seen for large m in fig. 14. Thus, no matter how many measurements are included, there is a limit to the localization of the state along the unstable manifold, as was illustrated in fig. 12a.

Some of the above relationships are apparent in fig. 15 where we compute δ for the Lorenz equations. Fig. 15 illustrates the distortion as a function of m. For small m, the scaling goes as $m^{-5/2}$, as predicted by eq. (41). For larger m, for the case of mixed coordinates, the distortion decreases as $m^{-1/2}$. In the case of purely predictive coordinates, a plateau is reached at $m \approx 250$, illustrating the fundamental limitation to predictability.

Note that δ_∞ depends on τ; by decreasing τ it is possible to decrease δ_∞. However, in any physical system there is inevitably a limiting sampling time, below which measurements become correlated and contain no new information. This implies a lower bound to δ_∞ – it is never possible to reduce it to zero. Thus, for chaotic systems with predictive coordinates there is a limit to predictability that does not exist in any of the other scaling regimes shown in fig. 13.

The limiting distortion δ_∞ is interesting, since it provides a quantitative measure of how prediction is limited by noise. If δ_∞ is reasonably low, then it is possible to determine well-localized states using measurements of reasonable precision, and approximate the time series as a deterministic dynamical system. Whenever $\tau_R \ll \tau_I$ it is possible to make δ reasonably small.

When $\tau_R > \tau_I$, however, there is a complete breakdown of predictability. With measurements of reasonable accuracy states cannot be well-localized, and the dynamics are not deterministic, even as an approximation. There is no predictability, except over a *very* short time scale, which we conjecture is related to the autocorrelation function rather than the leading Lyapunov exponent. The time series is effectively a random process.

The origin of unpredictability comes from a lack of relevant data. The redundancy time τ_R represents the minimum window width needed to resolve independent coordinates. If $\tau_R > \tau_I$, then it is not possible to make enough relevant measurements to resolve all the degrees of freedom during the time interval where the measurements are relevant to the present state. As a result the state is not localized, and the system is not deterministic in any sense.

Note that, since τ_R and τ_I both depend on ϵ, it is possible in principle to move from the deterministic to the random regime by varying ϵ. However, as we show in an example in section 5.4, the limiting distortion can increase exponentially as $\lambda^{d-1/2}$. When λ and d are even moderately large, δ_∞ quickly reaches astronomical proportions, making embedding impossible with any realistically obtainable precision.

5.3.2. Precise statement and derivation of scaling laws

Small window width limit, $w \to 0$. When $m\tau \to 0, m \to \infty$, the scaling law is

$$\delta = \mathcal{O}\left(m^{-1/2}(m\tau)^{1-d}\right), \qquad m\tau \to 0, \quad m \to \infty. \tag{41}$$

"$\mathcal{O}(\ )$" denotes "the order of". For $d > 1$ the distortion blows up in the limit as $\tau \to 0$, with an exponent that increases with dimension. Note that for almost all states the condition $m \to \infty$ can be relaxed to $m > d$, and the τ part of the scaling law still holds.

Derivation. Let m be sufficiently large so that Φ is locally an embedding. Consider the expansion of $D\Phi$ in a Taylor series in time around $t = 0$. For convenience assume a predictive embedding, with the first row of $D\Phi$ simply Dh; this does not effect the result. The rows of $D\Phi$ are of the form

$$D\Phi_{i+1} = a^{(0)} + a^{(1)}(i\tau) + a^{(2)}(i\tau)^2 + \ldots, \tag{42}$$

where $i = 0, \ldots, m-1$, labels the row, and the $a^{(j)}$ are fixed d-dimensional row vectors. If we truncate the Taylor series at order $d-2$ the matrix cannot be of full rank, since it is constructed from linear combinations of only $d-1$ independent vectors $a^{(j)}$. Consequently the dth singular value is zero to order $(m\tau)^{d-2}$. But if we truncate the Taylor series at order $d-1$ the matrix will generically be of full rank at almost all states s, since the d vectors $a^{(j)}$ of dimension d involved in the expansion are typically independent. Therefore the dth singular value is typically of order $(m\tau)^{d-1}$. The dominant eigenvalue of Σ is the square of the inverse of the dth largest singular value of $D\Phi$, which implies the $m\tau$ scaling in eq. (41).

The m scaling comes from the law of large numbers. If we fix the window width w at a small value and increase m, then the variance decreases as m^{-1} because of the assumed independence of the measurement errors. These two arguments taken together give the scaling law of eq. (41). □

Eq. (41) can be used to make the definition of the redundance time more precise. The reconstructed dynamical system ceases to be effectively deterministic[#17] when $\delta \geq 1/\epsilon$. Substituting this for δ and setting $m\tau = \tau_R$ in eq. (41) gives

$$\tau_R \approx \left(\epsilon^2 m^{-1}\right)^{1/2(d-1)}. \tag{43}$$

Nonchaotic systems, $w \to \infty, m \to \infty$. When w is large the measurement surfaces are no longer nearly parallel. In this case

$$\delta = \mathcal{O}(m^{-1/2}), \qquad w \to \infty, \quad m \to \infty. \tag{44}$$

This is reasonable from the law of large numbers. This result can be obtained more rigorously as a special case of the scaling laws derived below for chaotic systems.

We also hypothesize that δ typically increases with d, since when d is large the information in the time series is spread over more coordinates. In the system studied in section 5.4, for example, we show that $\delta = (1/\sqrt{2})\, d^{3/2} m^{-1/2}$. However, the precise nature of this scaling may depend on factors such as the measurement function, and it is not clear that a general scaling law exists.

Chaotic systems with mixed coordinates. When both past and future information are used for reconstruction, $\delta \to 0$ in the limit as $m \to \infty$ and $w \to \infty$. Transforming to coordinates where Σ is diagonal, the eigenvalues scale as

Unstable manifold ($\lambda_i > 0$):

$$\Sigma_{ii} = \mathcal{O}(e^{-2m_f \lambda_i \tau}), \tag{45}$$

[#17] For a discussion of what we mean by "effectively deterministic", see section 5.5.

Neutral manifold ($\lambda_i = 0$):

$$\Sigma_{ii} = \mathcal{O}(m^{-1}), \tag{46}$$

Stable manifold ($\lambda_i < 0$):

$$\Sigma_{ii} = \mathcal{O}(e^{2m_p\lambda_i\tau}). \tag{47}$$

For a flow the distortion is dominated by the scaling of the neutral manifold; for a discrete time map it decreases exponentially.

Chaotic systems with predictive coordinates. This case is equivalent to that of mixed coordinates above, except for the unstable manifold. Eq. (45) is replaced by

Unstable manifold ($\lambda_i > 0$):

$$\Sigma_{ii} = \mathcal{O}(1). \tag{48}$$

The eigenvalues of distortion matrix for the unstable manifold approach a constant, and provide the dominant term for noise amplification.

Derivation of eqs. (44)–(48). Let $\underline{x}$ be a delay vector of dimension m perturbed by IID Gaussian noise of variance ϵ^2, and let d' be the minimum dimension for which delay vectors define a predictive embedding $\Phi_{d'}$. Let $z = \Phi_{d'}(s)$ be a noise free d'-dimensional delay vector. Assume $m \gg d'$. In this derivation we investigate the scaling behavior of $p(\underline{x}|s)$ (and hence Σ) indirectly, by considering the likelihood function $p(\underline{x}|z)$ in d'-dimensional delay space.

We first consider the case where $\underline{x}$ is a predictive delay vector, i.e. $m_f = 0$. For convenience we take m so that it is an integer multiple of d'. We now use the strategy of splitting the noisy delay vector $\underline{x}$ into d'-dimensional pieces: Let

$$\underline{x}_j^{(d')} = (x(-jd'\tau), \ldots, x(-(jd' + d' - 1)\tau))^{\dagger} \tag{49}$$

be a d'-dimensional delay vector rooted at time $-jd'\tau$, and let

$$\underline{\xi}_j^{(d')} = (\xi(-jd'\tau), \ldots, \xi(-(jd' + d' - 1)\tau))^{\dagger} \tag{50}$$

be a vector of d'-dimensional measurement errors. Let F be the d'-dimensional dynamics induced by f, and let $j = 0, \ldots, m/d' - 1$. Transporting z back to time $-jd'\tau$, the fact that the noisy vector is the sum of the true vector and the measurements errors implies

$$\underline{x}_j^{(d')} = F^{-jd'\tau}(z) + \underline{\xi}_j^{(d')}. \tag{51}$$

As in section 3.1, the likelihood function $p(\underline{x}|z)$ is given by

$$\begin{aligned} p(\underline{x}|z) &= \prod_{j=0}^{m/d'-1} p\left(\underline{x}_j^{(d')} - F^{-jd'\tau}(z)\right) \\ &= A \exp\left(-\frac{1}{2\epsilon^2} \sum_{j=0}^{m/d'-1} \left\| \underline{x}_j^{(d')} - F^{-jd'\tau}(z) \right\|^2 \right). \end{aligned} \tag{52}$$

A is a normalization constant. Since ϵ is assumed to be small, eq. (52) can be simplified by using a Taylor series expansion, as in section 4.5, to obtain

$$p(\underline{x}|z) \approx C \exp\left(-\frac{1}{2\epsilon^2} (z - \hat{z})^{\dagger} (\Sigma')^{-1} (z - \hat{z}) \right), \tag{53}$$

where C is a normalization constant, $\hat{z}$ is the maximum likelihood estimate for z (based on $\underline{x}$), and the matrix Σ' is given by eq. 54, where the derivatives $DF^{-jd'\tau}$ are evaluated at $\hat{z}$,

$$\Sigma' = \left(\sum_{j=0}^{m/d'-1} (DF^{-jd'\tau})^{\dagger} (DF^{-jd'\tau}) \right)^{-1}. \tag{54}$$

It follows from the definition of the Lyapunov exponents that for sufficiently large $jd'\tau$, $(DF^{-jd'\tau})^{\dagger}(DF^{-jd'\tau})$ tends to a matrix with eigenvalues $e^{-2j\lambda_1 d'\tau}, \ldots, e^{-2j\lambda_{d'} d'\tau}$. Furthermore, for

large $jd'\tau$ the eigenvectors approach limiting values, independent of j. Since we are interested in the scaling behavior as $m \to \infty$, we assume that this happens rapidly enough that we can neglect the small j terms. Evaluating eq. (54) in the basis of eigenvectors yields

$$\Sigma'_{ii} = \left(\sum_{j=0}^{m/d'-1} e^{-2j\lambda_i d'\tau} \right)^{-1}. \tag{55}$$

When $\lambda \neq 0$ the summation reduces to

$$\Sigma'_{ii} = \frac{1 - e^{-2\lambda_i d'\tau}}{1 - e^{-2m\lambda_i \tau}}. \tag{56}$$

To compute the distortion matrix Σ for $p(\underline{x}|s)$ we transform from the d'-dimensional delay space to the d-dimensional state space. Since $z = \Phi_{d'}(s)$, we obtain

$$\Sigma' = D\Phi_{d'} \Sigma D\Phi_{d'}^{\dagger}. \tag{57}$$

Since d' is independent of m, it follows that the distortion matrix Σ has the same scaling laws in m as the matrix Σ'. Therefore, the scaling relationships are evident by considering eqs. (55) and (56).

Although for simplicity we assumed predictive coordinates in the above, the calculation for mixed coordinates is essentially the same except that all the sums and products must be taken from $-m_f/d'$ to m_p/d'. When $\lambda_i < 0$ the second term in the denominator of eq. (56) dominates and we obtain eq. (47). When $\lambda_i = 0$, from eq. (55) the sum is of order m, and we obtain eq. (46), which implies eq. (44). When $\lambda_i > 0$, for predictive coordinates the denominator of eq. (56) approaches 1 as $m \to \infty$, and we obtain eq. (48). For mixed coordinates the behavior of the unstable manifold mimics that of the stable manifold, with m_f replacing m_p, and we obtain eq. (45). □

$\tau_R > \tau_I$: *Irreducible random process.* We can get a rough estimate of the scaling in this regime by assuming that $\tau_I \approx 1/\lambda$, and substituting $w = 1/\lambda$ in eq. (41). This gives

$$\delta = \mathcal{O}(\tau^{1/2}\lambda^{d-1/2}). \tag{58}$$

This derivation may be unreliable, since it is not clear that $\tau_I \approx 1/\lambda$ is generally true, and there may be prefactors in eq. (41) that depend on λ. Nonetheless, it is encouraging that in the example of section 5.4 this at least approximately agrees with the observed behavior.

In the regime where λ is smaller so that $\tau_R < \tau_I$, we can repeat the argument above using eq. (44) rather than eq. (41). This gives the result that

$$\delta = \mathcal{O}(\sqrt{\lambda\tau}). \tag{59}$$

This result can also be derived using eq. (56) in the limit where $m \to \infty$ and $\lambda d\tau \ll 1$.

5.4. *A solvable example*

In this section we investigate the distortion for an example that is sufficiently simple that the observability matrix can be calculated explicitly. Consider a system of $d/2$ negatively damped harmonic oscillators:

$$\frac{d}{dt}\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} \lambda_i & -\omega_i \\ \omega_i & \lambda_i \end{pmatrix}\begin{pmatrix} u_i \\ v_i \end{pmatrix}, \quad i = 1, \ldots, d/2. \tag{60}$$

The state space dimension d is even. u_i and v_i are both taken modulo 1, corresponding to (piecewise smooth) motion on a torus. $\lambda_i > 0$ are the Lyapunov exponents. For convenience we will take $\lambda_i = \lambda$ = constant. We take the measurement function to be

$$h = \frac{2}{d} \sum_{i=1}^{d/2} u_i. \tag{61}$$

We assume a predictive reconstruction with $m_f = 0$.

This example is admittedly rather contrived. The oscillators are independent, so measurements only give information about the whole sys-

tem because the measurement function involves a combination of all the degrees of freedom. In a more typical example the flow of information depends on the coupling of the unobserved degrees of freedom to the observed degrees of freedom[#18]. Nonetheless, as we shall see, even this very simple example exhibits nontrivial behavior. Furthermore, the behavior agrees with the general scaling laws derived in the previous section.

This system has the following analytic solution:

$$u_j(t) = c_{1j}\, e^{\lambda_j t} \cos(\omega_j t + c_{2j}) \mod 1,$$
$$v_j(t) = c_{1j}\, e^{\lambda_j t} \sin(\omega_j t + c_{2j}) \mod 1, \qquad (62)$$

where c_{1j} and c_{2j} are arbitrary constants. Applying the definition of Φ and differentiating, the observability matrix can be calculated explicitly everywhere except at the discontinuities:

$$D\Phi_{i,2j-1} = \frac{2}{d} e^{-(i-1)\lambda_j \tau} \cos(i-1)\omega_j \tau,$$
$$D\Phi_{i,2j} = -\frac{2}{d} e^{-(i-1)\lambda_j \tau} \sin(i-1)\omega_j \tau, \qquad (63)$$

where i ranges from 1 to m and j ranges from 1 to $d/2$. Note that $D\Phi$ is constant throughout the state space.

To compute the distortion we must first evaluate $D\Phi^\dagger D\Phi$. The odd rows and columns are

$$(D\Phi^\dagger D\Phi)_{2i-1,2j-1} = \frac{4}{d^2} \sum_{k=0}^{m-1} e^{-k(\lambda_i+\lambda_j)\tau} \cos k\omega_i \tau \cos k\omega_j \tau. \qquad (64)$$

There are similar expressions for the other terms, with sin cos and sin sin instead of cos cos. The distortion can be obtained from the singular value decomposition of $D\Phi^\dagger D\Phi$ using eq. (36).

[#18]This example can be thought of as one in which we observe equally all the eigenmodes of a system of oscillators. In a more realistic example we might observe a single variable which is not an eigenmode. When rotated into the eigenspace, for a typical system this will correspond to unequal observations of each eigenmode. However, providing this inequality is not too extreme, it should not affect the scaling behavior.

In fig. 14 we plot δ as a function of m for several different values of the dimension and Lyapunov exponents. This illustrates several of the features derived in section 5.3.

– Small w: For small values of m the window width w is also small. The chaotic and nonchaotic cases are approximately the same. As m decreases the distortion increases as a power law with the predicted exponent $\frac{1}{2} - d$.

– Large w: For the chaotic case the distortion approaches a constant $\delta_\infty > 0$, while for the nonchaotic case the distortion decreases to zero according to $m^{-1/2}$. For the nonchaotic case the asymptotic behavior can be derived by considering eq. (64) in the limit $\tau \to 0, m\tau \to \infty$. Approximating the sum by an integral, to leading order in m the diagonal terms are $2m/d^2$, and the off-diagonal terms are of order 1. Thus this matrix is trivial to invert to leading order in m, and by taking the trace we obtain $\delta^2 = d \cdot d^2/2m$, which implies

$$\delta \sim \frac{1}{\sqrt{2}} d^{3/2} m^{-1/2}, \quad m\tau \to \infty, \quad \tau \to 0. \qquad (65)$$

For chaotic systems, the behavior of δ_∞ can be investigated by taking the limit $\tau \to 0$ in eq. (64) and approximating by an integral. This gives

$$(D\Phi^\dagger D\Phi)_{2i-1,2j-1} \approx \frac{2(\lambda_i + \lambda_j)}{\tau d^2} \times \left\{ \left[(\lambda_i + \lambda_j)^2 + (\omega_i + \omega_j)^2 \right]^{-1} + \left[(\lambda_i + \lambda_j)^2 + (\omega_i - \omega_j)^2 \right]^{-1} \right\}. \qquad (66)$$

The behavior of δ_∞ under changes in parameter values is investigated by using eq. (66), and the associated sin cos and sin sin expressions, with $\lambda_i = \lambda = \text{const.}$, and the frequencies uniformly spaced so that $\omega_i = 2/d, 4/d, \ldots, d/d$. The result is shown in fig. 16.

There are two scaling regimes, one for low λ, and one for high λ. In the small λ regime, the scaling can be derived by considering eq. (66).

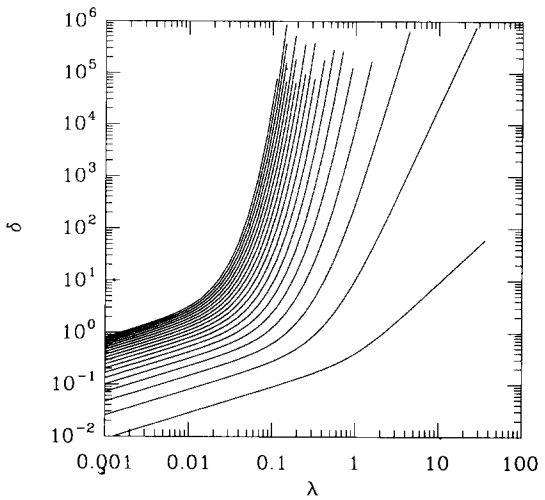

Fig. 16. The limiting distortion δ at $m = \infty$ plotted as a function of the Lyapunov exponent λ for dimensions $d = 2, 4, \ldots, 40$, for the example of eqs. (60), (61). The curves with the lowest distortion have the lowest dimension d. Notice that there are two scaling regimes, one for low λ and another for higher λ. In the high λ regime the enormous distortion means that even for small ϵ the system effectively behaves as a random process.

For sufficiently small λ the diagonal terms dominate, and to leading order in λ are $1/\lambda\tau d^2$. Thus this matrix is trivial to invert to leading order in λ. Taking the trace gives $\delta^2 = d \cdot d^2 \lambda\tau$, which implies

$$\delta_\infty \sim d^{3/2}\sqrt{\lambda\tau}\,, \quad \lambda \to 0, \quad \tau \to 0. \tag{67}$$

Thus, in the low λ regime the motion is effectively predictable and $\delta = \mathcal{O}(\lambda^{1/2})$, as predicted by eq. (59).

To derive the behavior in the large λ limit we use eq. (66) and related expressions. The only case we have been able to solve in closed form is $d = 2$, which yields

$$\mathrm{D}\Phi^\dagger \mathrm{D}\Phi = \frac{1}{4\tau(\lambda^2+\omega^2)\lambda}\begin{pmatrix} 2\lambda^2+\omega^2 & -\omega\lambda \\ -\omega\lambda & \omega^2 \end{pmatrix}. \tag{68}$$

Using the fact that for 2×2 matrices

$$\mathrm{Tr}(\mathrm{D}\Phi^\dagger \mathrm{D}\Phi)^{-1} = \frac{\mathrm{Tr}\,\mathrm{D}\Phi^\dagger \mathrm{D}\Phi}{\det \mathrm{D}\Phi^\dagger \mathrm{D}\Phi}, \tag{69}$$

it follows that

$$\delta_\infty \sim \sqrt{8\tau\lambda(1+\lambda^2/\omega^2)}\,, \quad \tau \to 0. \tag{70}$$

Thus in the case $d = 2$, in the high λ limit the distortion diverges at a rate which is consistent with the scaling law $\delta = \mathcal{O}(\lambda^{d-1/2})$, as predicted by eq. (58). For larger values of d, this scaling law is observed in fig. 16. However, we have observed somewhat different behavior in other examples.

5.5. *When chaotic dynamics becomes a random process*

A very large value of the distortion can cause a chaotic dynamical system of sufficiently high dimension to produce a time series for any practical purpose must be regarded as a random process. Before elaborating on this, we should make the distinction between a deterministic dynamical system and a "random process" more precise.

In physical systems perfect determinism never exists. Operationally, when we say that a system is "deterministic", we mean that measurements with precision ϵ result in predictions that are accurate to roughly ϵ. For a nonchaotic dynamical system this is true for any extrapolation time T; for a chaotic system, it is roughly true for times $T < 1/\lambda$.

However, as we have shown above, when d and λ are sufficiently large, projection onto a low dimensional time series can make it impossible to reconstruct a well-localized state. Because d is large, many measurements are needed to recover the state; because λ is large, measurements sufficiently far in the past are irrelevant to the present. As a result $\tau_R > \tau_I$, and the distortion diverges, at a rate that depends exponentially on

the dimension. In this case reconstructing a well-defined state becomes impossible for measurements of any reasonable precision. For example, when $\delta_\infty = 10^6$, as observed for $\lambda > 0.1$ and $d > 20$ in fig. 16, it would be necessary to make $\epsilon \approx 10^{-8}$ in order to reconstruct a state that was well-localized to within one part in a hundred[#19].

Even when the state is not well-localized, the time series is still predictable for a short time. This is because the most recent measurement causes the state to be confined within the associated measurement strip $S_\epsilon(0)$, and this strip takes a finite time to rotate so that its projection onto the time series is no longer well-localized. More precisely, this is because the measurement $x(0) = h(s(0))$ confines large eigendirections of Σ in a direction approximately orthogonal to Dh. However, if the distortion matrix Σ has at least one very large eigenvalue, then generically, the vector $Dh\, Df^T$ will have a significant component in the corresponding eigendirection after a short time T, causing an explosion in the noise amplification $\sigma(T)$.

Fig. 17 compares the behavior of the noise amplification for dynamical systems of low versus high dimension, using the example of section 5.4. The noise amplification $\sigma(T)$ at $m = \infty$ is plotted against T, using the integral approximation of eq. (66) to compute $\sigma(T)$. We take the limit $m_p \to \infty$, $m_f = 0$, so that the delay vector $\underline{x}$ represents the *entire past* of the time series. For the case $d = 2$ the noise amplification, which sets the limit to predictability with an ideal model, grows at a rate governed by the largest Lyapunov exponent. When $d = 40$, however, the noise amplification grows at a much faster rate, which is governed by the rotation rate associated with the linear dynamics, and which we conjecture is related to the correlation time of the time series. If $\epsilon > 10^{-9}$, the time series is predictable only over much shorter times than the Lyapunov time, and thus behaves like a random process.

[#19] We will assume in such calculations that the variance of states s and the time series $x(t)$ are on the order of 1.

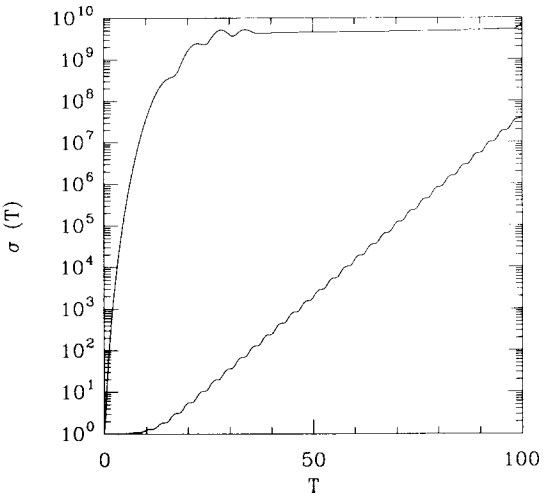

Fig. 17. The noise amplification $\sigma(T)$ for the example of section 5.4, using predictive coordinates with $m = \infty$ and $\tau = 0.01$. The largest Lyapunov exponent $\lambda = 0.2$. The curve on the bottom corresponds to dimension $d = 2$. In this case the state is well-localized, so the noise amplification starts out low and grows exponentially with time, at a rate governed by λ. For $d = 40$, in contrast, the initial state is not well-localized, and $\sigma(T)$ grows at a much faster rate, determined by the rotation of the measurement surface S(0).

From a practical point of view, when data is limited, estimation error may cause a time series to appear to be a random process when the dimension of the dynamics is sufficiently high. This is true even for nonchaotic dynamics. The mechanism that we discuss here, however, creates a type of random process which is in a certain sense more fundamental than that caused by limited data, in that it produces unpredictability *even when the optimal model is known*.

6. Coordinate transformations

Any reconstruction can be broken into two parts, $\Xi = \Psi \circ \Phi$. The transformation $\underline{x} = \Phi(s)$ specifies the delay coordinates, which determine the "information set". Thus far we have focused our attention on this part of the problem. In this

section we study the possibilities of transforming to new coordinates $y = \Psi(\underline{x})$. Commonly used examples of linear transformations Ψ include derivatives and principal components. In part motivated by the work of Fraser [18], we are particularly interested in the case where Ψ is *nonlinear*. In section 6.2 we show a general method for constructing a nonlinear transformation Ψ^* that (in a certain sense) provides optimal coordinates.

6.1. *Effect on noise amplification*

What is the effect of a coordinate transformation Ψ on the noise amplification? Two facts are immediately apparent:

– *Invertible coordinate transformations do not change the noise amplification.* This is clear from the fact that the conditional probability density $p(x(T)|\Psi(\underline{x}))$ is a function of $x(T)$ alone; $\Psi(\underline{x})$ is not an argument of p, but rather a label that identifies this as a particular member of a family of different functions. As long as the function Ψ is one-to-one, it leaves $p(x)$ and hence σ unchanged.

– *Noninvertible coordinate transformations can increase the noise amplification, but they cannot decrease it.* If more than one state $\underline{x}$ is mapped into the same state $\Psi(\underline{x})$, this generally has the effect of broadening p. This is evident since

$$p(x(T)|y) = \sum_{\{\underline{x}: y = \Psi(\underline{x})\}} p(x(T)|\underline{x}) p(\underline{x}). \quad (71)$$

Summing probability densities either increases the variance or leaves it unchanged, so

$$\mathrm{Var}(x(T)|y) \geq \big\langle \mathrm{Var}(x(T)|\underline{x}) \big\rangle_{\{\underline{x}: y = \Psi(\underline{x})\}}. \quad (72)$$

While coordinate transformations are not useful to reduce noise amplification, they can be useful for information compression. In some cases it is possible to reduce the dimension and leave the noise amplification unchanged, packing the same information into fewer coordinates[#20]. We will show in section 6.3 that it suffices to study the effect of changes of coordinates on the distortion matrix. The following results are a generalization of sections 3.1 and 4.5 to include (possibly noninvertible) changes of coordinates.

In the low noise limit, to first order in ϵ the transformation Ψ can be approximated locally by its derivative $\mathrm{D}\Psi$ (we leave out the constant term since it is an invertible translation). An expression for $p(s|\mathrm{D}\Psi(\underline{x}))$ can be derived using a generalization of the argument of section 3.1, as follows: Assuming a uniform prior gives $p(s|\mathrm{D}\Psi(\underline{x})) \propto p(\mathrm{D}\Psi(\underline{x})|s)$. But $p(\mathrm{D}\Psi(\underline{x})|s) = p(\mathrm{D}\Psi\,\underline{\xi})$, where $\underline{\xi} = \underline{x} - \Phi(s)$. We obtain eq. (73) by transforming the isotropic Gaussian distribution of the noise $\underline{\xi}$ through the linear map $\mathrm{D}\Psi$,

$$p(s|\mathrm{D}\Psi(\underline{x})) \approx A \exp\Big(\frac{-1}{2\epsilon^2} [\mathrm{D}\Psi\,\underline{x} - \mathrm{D}\Psi\,\Phi(s)]^\dagger \times (\mathrm{D}\Psi\,\mathrm{D}\Psi^\dagger)^{-1} [\mathrm{D}\Psi\,\underline{x} - \mathrm{D}\Psi\,\Phi(x)] \Big). \quad (73)$$

As before, in the limit that ϵ is small we can expand this in a Taylor series. The arguments parallel those leading to eq. (35), and the result is that

$$p(s|y) \approx C \exp\Big(-\frac{1}{2\epsilon^2} (s - \hat{s})^\dagger \Sigma^{-1} (s - \hat{s}) \Big), \quad (74)$$

where the distortion matrix Σ depends on Ψ according to eq. (75),

$$\Sigma = \big[\mathrm{D}\Phi^\dagger\,\mathrm{D}\Psi^\dagger (\mathrm{D}\Psi\,\mathrm{D}\Psi^\dagger)^{-1}\,\mathrm{D}\Psi\,\mathrm{D}\Phi \big]^{-1}. \quad (75)$$

As expected, a locally invertible coordinate trans-

[#20] This is trivial to do if nonsmooth coordinate transformations are allowed, as all the information in a delay vector can be compressed into one dimension by coding the decimal expansions of the components of the delay vector into one real number. However, in this section, we restrict attention to smooth coordinate transformations Ψ, since we are ultimately interested in modeling smooth dynamics.

formation Ψ does not alter the distortion matrix, since when $D\Psi$ is invertible $(D\Psi\, D\Psi^{\dagger})^{-1} = (D\Psi^{\dagger})^{-1} D\Psi^{-1}$, and eq. (75) reduces to eq. (36).

6.2. Optimal coordinate transformation

In this section we show that in the low noise limit it is possible to compress the information contained in a delay vector $\underline{x}$ into a smaller number of dimensions, while retaining all the available relevant information. We also show how to compute a transformation Ψ^* that does this from a singular value decomposition of the matrix $D\Phi$. In section 8 we discuss the estimation of this transformation from a time series.

$D\Phi$ is an $m \times d$ matrix that maps variations in the d-dimensional state, $\delta\tilde{s}$, into variations in the delay vector, $\delta\underline{\tilde{x}}$. The technique of singular value decomposition expresses $D\Phi$ as the product of three linear transformations, U, W, and $V^{\dagger}$:

$$D\Phi = UWV^{\dagger}. \tag{76}$$

The first of these, $V^{\dagger}$, is represented by an orthonormal $d \times d$ matrix that performs a rotation about $\tilde{s}$ onto the principal axes in the tangent space $\mathbb{R}^d$ to the state space M. The second transformation W is represented by a diagonal $d \times d$ matrix that stretches or contracts the principal axes; its diagonal elements w_i are called the *singular values* of $D\Phi$. The third transformation U is represented by a column orthonormal[#21] $m \times d$ matrix that maps the d-dimensional tangent space of the state space M into the tangent space to $\Phi(\mathrm{M})$ at $\Phi(\tilde{s})$.

The distortion matrix for delay coordinates can be decomposed in these terms by inserting eq. (76) into eq. (36), which gives

$$\Sigma = (D\Phi^{\dagger} D\Phi)^{-1} = (VWU^{\dagger}UWV^{\dagger})^{-1} = VW^{-2}V^{\dagger}. \tag{77}$$

[#21] U is column orthonormal if $U^{\dagger}U = \mathbb{1}_d$.

The eigenvalues of the distortion matrix are the inverse squares of the singular values, since V can be viewed as a similarity transformation which diagonalizes the distortion matrix:

$$V^{\dagger}\Sigma V = W^{-2}. \tag{78}$$

The singular values w_i describe how well the observations determine the original state s along each of the principal axes of Σ. If w_i is small then the observations are highly uncertain along the corresponding axis. To reduce distortion, the best coordinates are obviously those that make w_i as large as possible for all i.

Using eq. (79), we define a nonlinear transformation Ψ^*: $\mathbb{R}^m \to \mathbb{R}^d$, called *local singular value decomposition*,

$$\Psi^*(\underline{x}) = U^{\dagger}\underline{x}, \tag{79}$$

where $U^{\dagger}$ is from the singular value decomposition of $D\Phi$ at $\tilde{s}$. Geometrically, the transformation Ψ^* projects noisy delay vectors in a direction orthogonal to the appropriate tangent space to the embedded state space $\Phi(\mathrm{M})$, collapsing the m-dimensional delay space onto a d-dimensional subspace. All information in directions orthogonal to the tangent space is consequently lost. This is desirable, since these are the directions dominated by noise. The transformation Ψ^* is nonlinear, since a singular value decomposition at each point in the state space produces a different linear map $U^{\dagger}$.

The local singular value decomposition Ψ^* results in the same distortion matrix as raw delay coordinates, i.e. it compresses all the relevant information in the m-dimensional delay coordinate into d dimensions. This can be seen by substituting $D\Psi^* = U^{\dagger}$ and $D\Phi = UWV^{\dagger}$ into eq. (75),

$$\Sigma_{\Psi^*} = (VWU^{\dagger}UU^{\dagger}UWV^{\dagger})^{-1} = VW^{-2}V^{\dagger}. \tag{80}$$

Comparing eq. (80) to eq. (77) shows that the distortion of Ψ^* is equal to that of pure m-

dimensional delay coordinates. Thus Ψ^* is optimal in the sense that it has the minimal distortion of any coordinate transformation from $\mathbb{R}^m \to \mathbb{R}^d$. This minimum is not unique, as Ψ^* can be composed with any invertible transformation of $\mathbb{R}^d$ and the distortion will be unchanged.

6.3. Simultaneous minimization of distortion and noise amplification

In general, when we observe a time series we cannot observe the original coordinates, and so it is impossible to compute the distortion from the time series. Fraser originally posed the question of whether or not it is possible to find a reconstruction which minimizes distortion by using only the information available in a time series [18]. We answer this question partially by showing that minimizing the distortion matrix Σ over coordinate transformations Ψ is equivalent to minimizing the noise amplification $\sigma(T)$, which can be estimated from a time series. This provides only a partial answer, because we optimize over Ψ, holding Φ fixed, whereas Fraser's question concerned the total reconstruction map $\Xi = \Psi\Phi$.

Consider the set of all coordinate transformations $y = \Psi(\underline{x})$, where Ψ: $\mathbb{R}^m \to \mathbb{R}^{d'}$ and m, τ and d' are fixed. If there exists a Ψ^* which satisfies $\Sigma_{\Psi^*} \le \Sigma_\Psi$ for all Ψ (according to the ordering defined in section 5.1), we claim Ψ^* will also satisfy $\sigma_{\Psi^*}(T) \le \sigma_\Psi(T)$ for all Ψ and T. The converse is also true under the condition that there exists a finite set of times $T_1, \ldots, T_p$ such that the p vectors $Dh\, Df^{T_i}$ span the tangent space $\mathbb{R}^d$ to the state space M. Note that by Takens' theorem this condition is a generic property of h and f for $p \ge 2d + 1$. Then the converse states that if there exists a Ψ' that satisfies $\sigma_{\Psi'}(T_i) \le \sigma_\Psi(T_i)$ for all Ψ and $i = 1, \ldots, p$, then $\Sigma_{\Psi'} \le \Sigma_\Psi$ for all Ψ. Since $\sigma(T)$ is an observable, in principle it can be minimized by finding a transformation that gives a simultaneous minimum for several different times. In section 6.2 we showed how to construct the coordinates $y^* = \Psi^*(\underline{x})$ by minimizing the distortion matrix, so the above optimal coordinate transformations do indeed exist. In section 8 we show how to estimate them directly from a time series.

Derivation. We can use eq. (32) to demonstrate that any coordinate transformation Ψ^* that minimizes the distortion Σ will also minimize the noise amplification $\sigma(T)$ for any time T as follows: Let $y^* = \Psi^*(\underline{x})$. Then $v^\dagger(\Sigma_\Psi - \Sigma_{\Psi^*})v \ge 0$ for any d-dimensional vector v and any coordinate transformation Ψ. By taking $v^\dagger = Dh\, Df^T$, we have $\sigma_\Psi(T) - \sigma_{\Psi^*}(T) \ge 0$ for all T.

To demonstrate the converse, let there be a Ψ' such that $\sigma_{\Psi'}(T_i) \le \sigma_\Psi(T_i)$ for all Ψ and $i = 1, \ldots, p$. We have shown in section 6.2 that there exists a transformation Ψ^* such that $\Sigma_{\Psi^*} \le \Sigma_\Psi$ for all Ψ. It suffices to show that $\Sigma_{\Psi'} = \Sigma_{\Psi^*}$. By definition of Ψ', we have that $\sigma_{\Psi'}(T_i) \le \sigma_{\Psi^*}(T_i)$ for $i = 1, \ldots, p$, and by the first part of the derivation we have that $\sigma_{\Psi'}(T) \ge \sigma_{\Psi^*}(T)$ for all T. It follows that $v_i^\dagger L v_i = 0$ for $i = 1, \ldots, p$, where $v_i^\dagger = Dh\, Df^{T_i}$, and $L = \Sigma_{\Psi'} - \Sigma_{\Psi^*}$ is necessarily a positive semi-definite matrix. To complete the demonstration we must show that $L = 0$. Now transform to coordinates so that $L = \mathrm{diag}(l_1, \ldots, l_d)$. We obtain a contradiction if one or more of the l_i are non-zero, as follows: Suppose (without loss of generality) that $l_1 > 0$. Then $v_i^\dagger L v_i \ge l_1 \|v_i^{(1)}\|^2 > 0$, where $v_i^{(1)}$ denotes the first component of v_i in the new coordinates. Note that there must exist an i such that $\|v_i^{(1)}\|^2 > 0$, by the condition that the v_i span $\mathbb{R}^d$. □

6.4. Linear versus nonlinear decomposition

In this section we study the local singular value decomposition numerically, using the Lorenz equations as an example. We compare this to a global (linear) decomposition. For convenience, we take advantage of a result from ref. [22], showing that for small window widths global principal value decomposition is well approximated by Legendre polynomials. We use the first three Legendre polynomials as basis functions for a reconstruction, and use eq. (75) to compute the

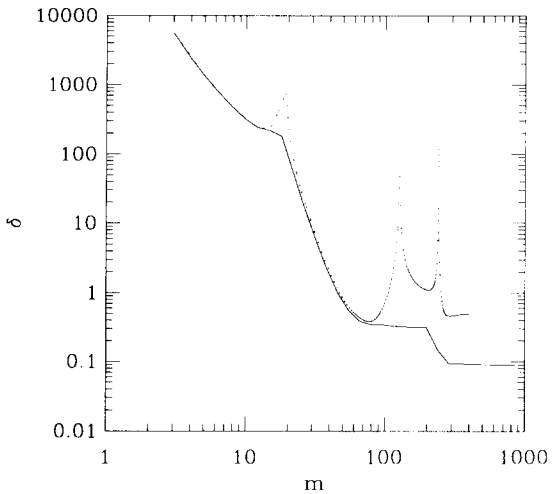

Fig. 18. A comparison of the distortion for local versus global decomposition, for the Lorenz equations. The solid curve is for local singular value decomposition, and the dashed curve is for Legendre polynomials. In both cases $d = 3$, $\tau = 0.005$, and $m_f = 0$; the state is the same as that used in fig. 11.

distortion. To compute the distortion for local singular value decomposition we make use of the fact we just derived, that the distortion of local SVD is optimal, and is equal to that of delay coordinates. The result of the comparison is shown in fig. 18.

For this example the linear transformation is close to optimal over a wide range of window widths. We believe this is because the transformation $U^\dagger$ generally converges onto Legendre polynomials in the limit of small window widths. At larger window widths this is no longer true, and the optimal nonlinear method gives distortions that are sometimes lower by as much as an order of magnitude.

In fig. 19 we illustrate the structure of the basis vectors produced by local singular value decomposition, at two different values of m. For $m = 10$ the basis vectors are quite similar to Legendre polynomials, as expected. For $m = 100$, however, this is no longer true. For a given window width these basis vectors may be thought of as optimal predictive filters – noise can be removed from the time series by convolving it with each basis vector to reconstruct states of dimension 3. Unfortunately, when f and h are unknown these filters

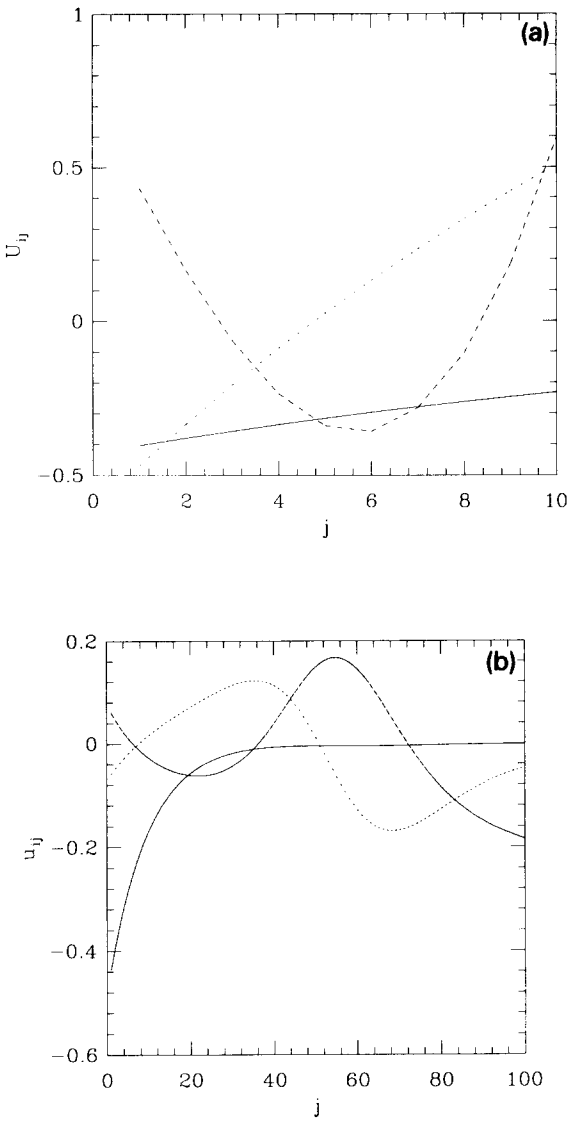

Fig. 19. First three basis vectors for local singular value decomposition of the Lorenz equations. $U_{ij}^\dagger$ from eq. (76) is plotted against j for $i = 1, 2, 3$. The solid curve is for the case $i = 1$, which corresponds to the largest singular value. In (a) $m = 10$, and the basis vectors are quite similar to Legendre polynomials; in (b) $m = 100$, and they reflect the more nonlinear behavior associated with larger window widths.

must be *estimated* from a time series $x(t)$. This introduces problems, particularly for large window widths. This is not the case for Legendre polynomials. At least for this example, they are close to optimal over a fairly wide range of window widths, and so are difficult to improve upon.

7. Estimation error

In this section, we compare noise amplification to prediction errors, and show how the framework we have developed can be used to understand the relationship between state space reconstruction and estimation errors. We show how scaling laws for estimation errors, together with those for noise amplification, explain the behavior of prediction errors under changes of coordinates. For delay coordinates this gives insight into the selection of parameters such as m and τ for optimizing predictions.

7.1. Analysis of estimation error

In this section, we will initially assume the states s are observable and the dynamics f and measurement function h are unknown. Once we have discussed estimation error in these terms, we will address the problem of estimation error for delay coordinates.

In constructing a nonlinear model of a dynamical system from a time series, with a finite number of data points there is inevitably a discrepancy between the true dynamics f and the approximation $\hat{f}$. Furthermore, there is a discrepancy between the true measurement function h and the approximation $\hat{h}$. This leads to *estimation error* $E(s) = \hat{h}\hat{f}(s) - hf(s)$. The total prediction error is governed by the sum of estimation error and the error due to noise amplification.

Estimation errors depend on the method of approximation. There are many possible methods of approximation, and studying them all is beyond the scope of this paper. We shall focus on one class of methods, local approximation, which have been shown to be effective in many prediction problems in nonlinear dynamics [8, 14, 15]. For clarity we shall begin by being even more specific, and studying first order local approximation. A simple example is nearest neighbor approximation: For the current state $s(t)$, find the nearest neighbor $s'(t')$ in the historical record. The prediction $\hat{x}$ is given by $\hat{x}(t, T) = x(t' + T)$. There are many variations on this method, for example using weighted averages over several nearby states. Such methods come under the general heading of *kernel density estimation* [40].

For first order local approximation, for small ϵ the error in predicting the time series is

$$E \approx \mathrm{D}h\, \mathrm{D}f^{T}\boldsymbol{\epsilon}, \tag{81}$$

where $\boldsymbol{\epsilon} = s'(t) - s(t)$ is the difference vector between the current state and the kth nearest neighbor (although we also use ϵ to represent noise level, the meaning should be clear from the context). Eq. (81) gives an error estimate that is accurate to first order in $\epsilon = \|\boldsymbol{\epsilon}\|$.

Providing ϵ is chosen small enough, we expect to find k points out of N inside a ball of radius ϵ if

$$\frac{k}{N} \approx c\epsilon^{d_I}\langle p(s)\rangle, \tag{82}$$

where $\langle p(s)\rangle$ is the average probability density in the ball, d_I is the information dimension, and c is a constant which depends on the dimension of the ball.

To understand the estimation error, our overall strategy is to use the original state space as a fixed reference frame. First we transform the probability density function $p(s)$ in the original space into delay space; this will allow us to compute the radius of a ball containing the k nearest neighbors in delay space. We then transform this ball back into the original space, and use eq. (81) to compute the estimation error. We will assume throughout that Φ is an embedding.

Providing the probability density is smooth, under a coordinate transformation $\underline{x} = \Phi(s)$ the probability density transforms as

$$p(\underline{x}) = |\mathrm{D}\Phi^{\dagger}\, \mathrm{D}\Phi|^{-1/2} p(s), \tag{83}$$

where $|\ |$ denotes the determinant.

In the case that $p(s)$ is fractal[#22], eq. (83) must be modified. For example, consider a chaotic attractor which is locally the Cartesian product of a Euclidean manifold and a Cantor set. For convenience, assume that the first $d-1$ eigendirections of the distortion matrix $\Sigma = (\mathrm{D}\Phi^{\dagger}\,\mathrm{D}\Phi)^{-1}$ are tangent to the Euclidean manifold, and the last is transverse. For a chaotic attractor, from the derivations of section 5.3, this becomes true in the large window width limit. Letting the singular values of $\mathrm{D}\Phi$ be w_i, the transformation rule is

$$p(\underline{x}) = p(s)\, w_d^{-(d_1-[d_1])} \prod_{j=1}^{d-1} w_j^{-1}, \tag{84}$$

where $[d_1]$ is the integer part of the information dimension. Note that eq. (84) approaches eq. (83) in the limit as $d_1 - [d_1] \to 1$.

Using eqs. (82) and (84), the radius of a ball containing k points in reconstructed coordinates is

$$\epsilon_{\underline{x}} \approx \epsilon\, w_d^{(d_1-[d_1])/d_1} \prod_{j=1}^{d-1} w_j^{1/d_1}. \tag{85}$$

When a ball in delay coordinates is transformed back to original coordinates, to first order in ϵ, it becomes an ellipsoid whose principal axes are of length

$$\epsilon_i' \approx \epsilon\, w_i^{-1} w_d^{(d_1-[d_1])/d_1} \prod_{j=1}^{d-1} w_j^{1/d_1}. \tag{86}$$

The principal axes are the eigendirections of Σ. Substituting $v_i\epsilon_i'$, where v_i is the ith singular vector of $\mathrm{D}\Phi$, for $\boldsymbol{\epsilon}$ in eq. (81) gives the estimation error along axis i,

$$E_i \approx \mathrm{D}h\, \mathrm{D}f^T(s)\, v_i \epsilon\, w_i^{-1} w_d^{(d_1-[d_1])/d_1} \prod_{j=1}^{d-1} w_j^{1/d_1}. \tag{87}$$

Since $\mathrm{D}h$, $\mathrm{D}f^T$, and ϵ are in the original space, under changes of coordinates these terms may regarded as constants. We expect changes in v_i to be second order. Changes in estimation error then occur because of changes in the singular values w_i.

The underlying reason for the change in estimation error is that changing coordinates changes neighborhood relationships. This is illustrated in figs. 20 and 21. Consider a ball in M containing k points which has radius ϵ. $\mathrm{D}\phi$ maps this ball into an ellipsoid centered on $\underline{x}$, with principal axes of length $w_i\epsilon$. In contrast, a ball in Φ(M) generally contains a different set of points. When mapped back to the original space these points are stretched along the principal axes of the distortion matrix, according to w_i^{-1}. Thus, if the singular values of $\mathrm{D}\Phi$ vary across a wide range, the stretching is severe, and the nearest neighbors in reconstructed coordinates correspond to an unnatural set of "neighbors" from the point of view of the original space.

Scaling laws for the estimation error can be obtained by substituting the scaling laws for the distortion matrix, as derived in section 5.3, into eq. (86). For chaotic dynamical systems with predictive coordinates, in the limit as the window width $w \to \infty$, according to eqs. (46)–(48) the singular values w_i corresponding to positive Lyapunov exponents approach a constant, while those corresponding to negative Lyapunov exponents diverge as $w_i \sim \mathrm{e}^{-\lambda_i w}$. Thus the singular values are very different, and we expect unnatural neighbors. To simplify the scaling law, we assume that the Lyapunov dimension can be substituted for the information dimension, according to the

[#22] For a fractal the measure is singular, so strictly speaking $p(s)$ must be viewed as a functional. This alters the transformation.

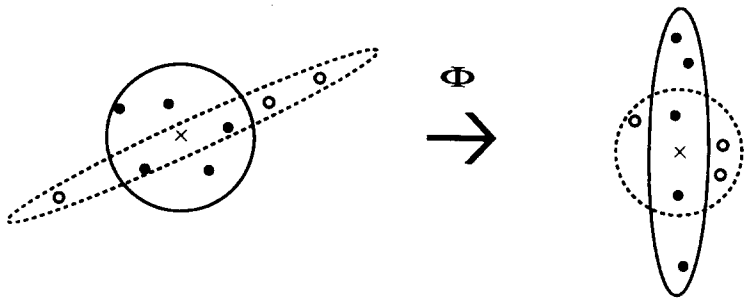

Fig. 20. Geometrical cause of change in estimation error due to coordinate reconstruction. The nearest neighbors in the original space may differ from the nearest neighbors in a reconstructed space. In the original state space (a), the solid points are the five nearest neighbors to the point marked by "×". These points are bounded by the solid ball. In a reconstructed state space (b) this ball is distorted; the nearest neighbors in this space include the open circles shown and are bounded by the dashed ball. The ball enclosing the nearest neighbors in the reconstructed space is distorted when it is transformed back to the original space. Neighbors found in the reconstructed space may form an unnatural set of "neighbors" for local approximation, causing a change in estimation error.

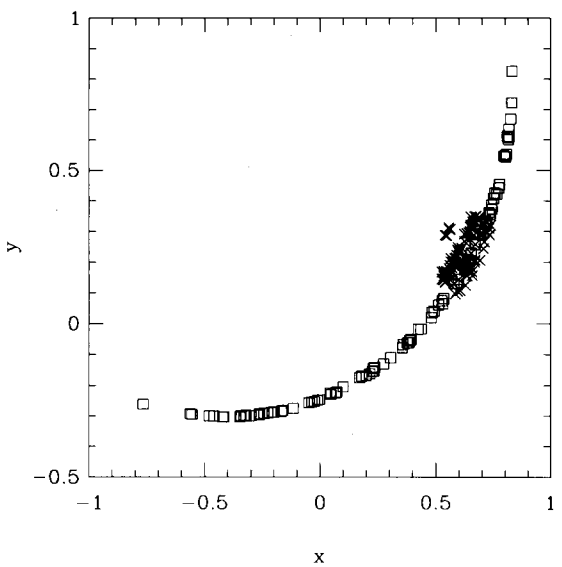

Fig. 21. Neighborhoods for the original state space compared to those in delay coordinates, for the Ikeda map with $\mu = 0.9$. The points marked with "×" are neighbors of a given point in the original state space. Points marked by a square are neighbors of the same point in a delay delay space with $\tau = 1$ and $m = 6$, but are no longer neighbors in the original space.

Kaplan–Yorke conjecture[#23] [27, 13]

$$d_{\mathrm{I}} = n + \frac{\sum_{i=1}^{n} \lambda_i}{\lambda_{n+1}}, \tag{88}$$

where n is the largest integer such that $\sum_{i=1}^{n} \lambda_i \geq 0$. Substituting into eq. (87), taking logarithms, and averaging implies that in the limit as $\epsilon \to 0$ and $w \to \infty$,

$$\langle \log|E| \rangle \sim C + \frac{h_\mu w}{d_{\mathrm{I}}}, \tag{89}$$

where h_μ is the metric entropy, which is equal to the sum of positive Lyapunov exponents, and C is a constant that depends on properties of the dynamical system in the original coordinates, as well as the number of data points and extrapolation times[#24]. This scaling behavior is illustrated in figs. 22 and 23.

[#23] The Kaplan–Yorke conjecture essentially requires that the attractor is locally the Cartesian product of a Euclidean manifold and a Cantor set, just as we have have assumed in eq. (84).

[#24] The scaling properties of C under variations in N and T were derived in ref. [15].

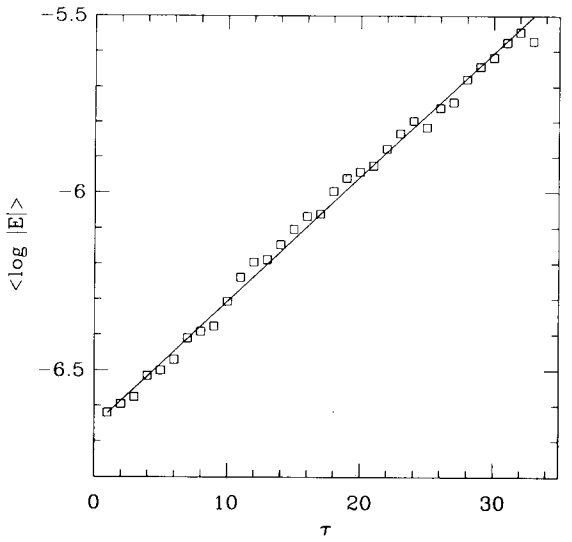

Fig. 22. Estimation error as a function of τ. Prediction errors for nearest neighbor prediction are plotted with boxes, and the expected scaling based on eq. (89) (with the prefactor fit by inspection) is plotted with a solid line. The noiseless 10000 point time series is from the x coordinate of the Ikeda map (eq. (16)) with $\mu = 0.6465$. The embedding dimension is $m = 4$. In this and subsequent figures, the prediction error is evaluated with respect to the noiseless time series, $E = \hat{x}(T) - \tilde{x}(T)$. Here $T = 1$.

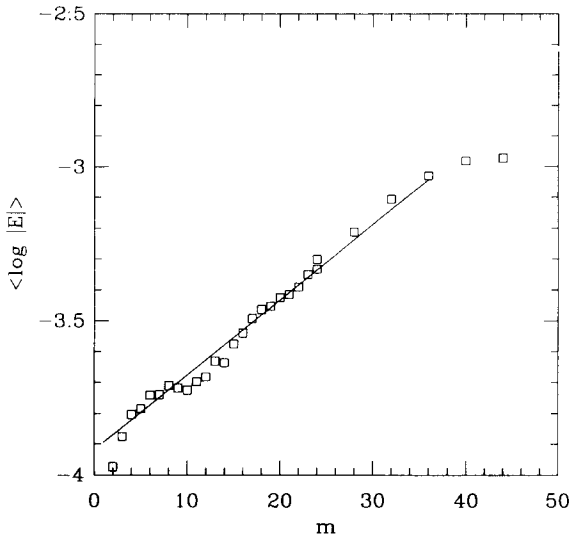

Fig. 23. Estimation error as a function of embedding dimension m. Similar to fig. 22, except that the lag time is fixed at $\tau = 17$ and m is varied. The time series is from the Mackey–Glass equation [12] at a delay parameter value of 17. The extrapolation time is $T = 1$.

The derivation of these scaling laws depends on the assumption that the error is small (or equivalently, that the data set is sufficiently large), so that eqs. (81)–(87) are all valid[#25]. Furthermore, the scaling law of eq. (89) assumes that the window width is sufficiently large to reach the asymptotic behavior derived in section 5.3. We have tested eq. (89) for a variety of different examples, including the Hénon and Ikeda maps and the Lorenz and Mackey–Glass equations, and we find that it is in good agreement with numerical experiments, providing these conditions are met.

This analysis makes it clear that the geometric causes of distortion and estimation error are similar. In both cases a ball in reconstructed coordinates is distorted when it is mapped back into the original space. There is an important difference, however: For distortion the size of this ball is fixed by the noise level, and is independent of the reconstruction. For estimation error, in contrast, the size of the ball varies. This causes the estimation error to vary differently from noise amplification, as illustrated in fig. 24.

Fig. 24 demonstrates how prediction errors are governed by the sum of noise amplification error and estimation error. For low values of m the noise amplification provides the dominant source of error; it is a decreasing function of m which reaches a plateau as m becomes large. The estimation error, in contrast, increases exponentially with m. The best value of m occurs when these two effects are roughly comparable. At least for simple examples, the prefactors of the scaling law for estimation error may be worked out. This makes it possible to sum the estimation error and the error due to noise amplification, and to find the optimal m and τ by minimization. The opti-

[#25]For systems whose fractal dimension is less than two, there may be significant predictive information in a single coordinate. In this case, even as the window width $w \to \infty$ some predictability is retained. In a plot of error as a function of τ, for example, this will cause $\langle \log|E| \rangle$ to prematurely reach a plateau.

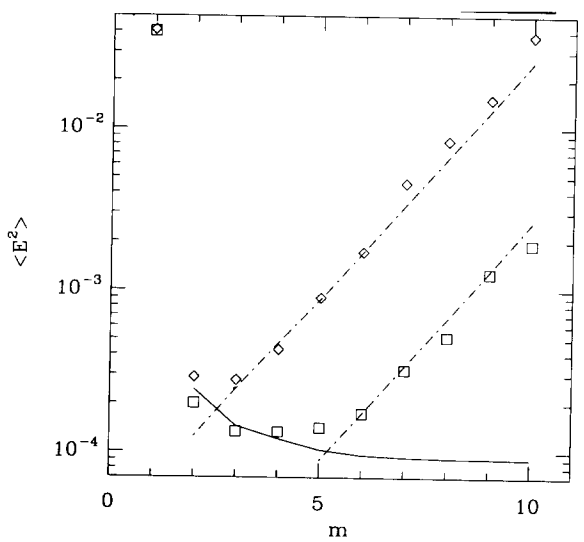

Fig. 24. Prediction error as a function of embedding dimension m for the Hénon map. The time series is $x(t)$ with additive Gaussian noise of variance 5.2×10^{-5} ($\epsilon \approx 1\%$). A state space is reconstructed using delay coordinates with $\tau = 1$. Predictions were made using the average value of the five nearest neighbors. The points marked with diamonds are prediction errors for a training set of $N = 1000$ points; those marked by squares are for $N = 10000$. The dot-dashed lines show the predictions of eq. (87) for the estimation error. The solid line is an estimate of errors due to noise amplification, $\epsilon^2\tilde{\sigma}^2$, based on eqs. (26), (32), and (36). Some predictions dip slightly below the minimum indicated by noise amplification; we attribute this to the effects of a fractal prior and a finite ϵ.

mal values depend on factors that might not have been obvious in advance, such as the number of data points.

The analysis given here can be extended to higher order local approximation. For qth order approximation in one dimension, the error is roughly $E \approx f^{(q)}(x)\epsilon^q$, where $f^{(q)}$ is the qth order derivative. Higher dimensions require multiple separation vectors, which complicates the situation. We conjecture that the result of generalizing eq. (87) to higher orders gives a similar answer, but involving powers of q and higher derivatives. For direct approximation (see ref. [15]), the scaling law of eq. (89) is the same, except that the last term is multiplied by a factor of q. This roughly agrees with the behavior we have observed in numerical experiments.

7.2. *Extensions of noise amplification to estimation error and dynamic noise*

When we defined the conditional variance of section 4.1, we assumed that the uncertainties in the time series came from observational noise. However, we conjecture that the ideas of that section can be extended to include estimation error and dynamic noise.

A data set can be viewed as a particular realization of an ensemble of possible data sets. For a given realization, function approximation produces a prediction $\hat{x}(T)$ for the time series value $x(T)$. Using the same estimation procedure, another realization of the data set would yield a different value for $\hat{x}(T)$. There is thus an ensemble of different estimates, characterized by a probability density function $p(\hat{x}(T)|\underline{x})$. This density function in turn defines a conditional variance associated with the estimation error. The square of the estimation error computed in eq. (81) is an estimate of the variance; this estimate can be improved by averaging over several neighbors, several nearby values of $\underline{x}$, or several realizations.

A conditional variance can also be associated with dynamic noise. It is important to realize that the properties of dynamic noise are significantly different from those of observational noise. One of the essential differences is that observational noise acts on the time series, whereas dynamic noise acts in the original state space. Thus, as a first approximation the effect of dynamic noise is closer to that of estimation error, and follows the geometrical behavior illustrated in fig. 20. As a result, we conjecture that the scaling properties of dynamic noise should be similar to those of estimation error. This obviously requires further investigation.

When either observational noise, dynamic noise, or estimation error dominates, we can assign a corresponding notion of noise amplifica-

tion. This is done by dividing the square root of the conditional variance by ϵ, as in eq. (23). For estimation error ϵ can be defined as the mean distance to a nearest neighbor, and for dynamic noise as the dynamic noise level. Just as for observational noise, normalizing by ϵ has the advantage that when Φ is an embedding, the result becomes independent of ϵ in the limit as $\epsilon \to 0$ (except for possible oscillation problems mentioned earlier). This can be very useful for theoretical analysis, derivation of scaling laws, etc.

In real time series the effects of observational noise, dynamic noise, and estimation error are combined, and may be difficult to separate. For some practical purposes, such as optimization of parameters, it may be more useful to work directly in terms of estimators of the conditional variance. This has been the point of view in earlier work which has not focused on the distinction between different sources of error [2, 9, 24, 30, 33, 37].

8. Practical implications for time series analysis

In this section we discuss the practical implications of our theory for the problems of numerical state space reconstruction and prediction from a time series. In section 8.1, we examine a procedure for estimating the optimal nonlinear coordinates discussed in section 6.2, when the dynamics f and the measurement function h are unknown. We perform numerical experiments comparing prediction errors using estimated nonlinear coordinates to prediction errors using delay coordinates. We also introduce a new algorithm for reducing estimation errors by "distorting" coordinates.

8.1. Numerical local principal value decomposition

In section 6.2, for the case when f and h are known, we showed how to construct nonlinear coordinates through local singular value decomposition. The resulting coordinates are optimal, in the sense that for a given information set they have minimal distortion in only d dimensions. In this section we discuss a method of estimating these coordinates directly from the time series.

In section 6 we showed that an optimal coordinate transformation Ψ can be constructed by decomposing $D\Phi = UWV^\dagger$, and choosing $\Psi(\underline{x}) = U^\dagger \underline{x}$. But since $U^\dagger$ is a transformation which maps noisy delay vectors onto the tangent space to $\Phi(M)$ its domain is observable, and its range can be estimated from observables. This allows us to estimate Ψ directly from a time series. Although Ψ is in general a nonlinear transformation, the estimation can be done locally, by finding a set of nearest neighbors of the given delay vector $\underline{x}$ and performing a principal value decomposition[#26].

The principal value decomposition yields a $d \times m$ matrix $\hat{U}^\dagger$ which spans the d principal directions of the data. Let N be the number of points in the data set. In the limit as $N \to \infty$ and the noise level $\epsilon \to 0$, providing the data spans the region of state space of interest, the space spanned by the data is the same as the tangent space of the embedded state space. The dimension d is equal to the number of nonzero singular values in the local decomposition. Except for rotations within the tangent space, as $N \to \infty$, $\hat{U}^\dagger$ converges to $U^\dagger$. For finite N we call the matrix $\hat{U}^\dagger$ the *local principal value decomposition*[#27] (local PVD), to distinguish it from the exact transformation $U^\dagger$.

Local PVD is fundamentally different from global PVD. For local PVD, flatness of neighborhoods in the limit as the neighborhood size goes to zero implies that there are only d nonzero

[#26] In general, the point $\underline{x}$ and its neighbors are not centered on the origin, so we perform the PVD in translated coordinates whose origin is at the mean of the local neighborhood.

[#27] This procedure was originally suggested as a means of computing dimension by Broomhead et al. [6], which improved upon a related method suggested by Froehling et al. [20]. It has also been used for computing local coordinates for spatiotemporal chaos [5].

singular values. For global PVD, in contrast, due to curvature of the embedded state space there are an infinite number of nonzero singular values [22]. Of course, for finite neighborhood sizes there are always deviations from flatness, which causes problems. The neighborhood must be chosen large enough to contain sufficient data, and it must also be large enough to ensure that the extension of the data in the tangent space is significantly greater than the magnitude of the noise. To the extent that there is significant curvature within the neighborhood, the advantages of local over global PVD are lost.

In the absence of noise, the degree to which the tangent space is a bad approximation to $\Phi(M)$ is measured by the combined magnitudes of the last $m-d$ singular values[#28]. Fig. 25 illustrates typical local singular spectra for the Ikeda map, with varying m. Note that the last $m-d$ singular values increase as the dimension of the delay reconstruction increases, indicating greater curvature within neighborhoods for larger m.

Curvature within neighborhoods tends to cause estimation problems in prediction with local PVD coordinates. Fig. 26 illustrates this for the Ikeda map. For low embedding dimensions m, local PVD coordinates are roughly as good as delays. However, as m increases, curvature within neighborhoods causes local PVD coordinates to give predictions several orders of magnitude less accurate than delays.

In contrast, when the prediction errors are dominated by noise rather than estimation error, local PVD is often superior to delay coordinates, particularly when the time series is finely sampled from a continuous time system. In fig. 27, we show the results of a prediction experiment for $x(t)$ of the Lorenz attractor. We increase m and decrease τ in order to hold the window width constant. As m increases the noise amplification

[#28] Whether noise or curvature dominates a singular value can be determined by examining its scaling with increasing neighborhood size. See ref. [6].

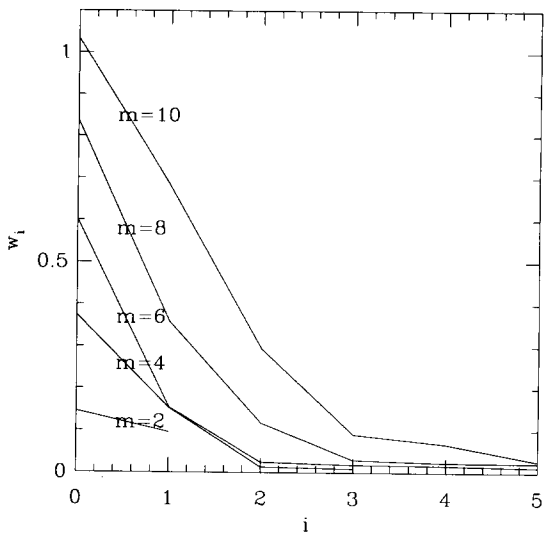

Fig. 25. Singular spectra for local PVD, based on a 1000 point time series of the x coordinate of the Ikeda map with 1% noise. The decomposition is based on predictive delay coordinates with $\tau = 1$ and $m = 2$ through 10. Local PVD is performed on the 20 nearest neighbors of an arbitrary point. Note that as m increases, the distinction between the first two singular values and the rest becomes less significant, due to increasing curvature.

decreases; however, with constant window width the curvature stays the same. For large m and small τ, local PVD coordinates are better than delays. However, in this regime we have found that appropriate linear filtering of delay coordinates produces results that are just as good as those of local PVD. Linear filtering of delay coordinates has the advantage of being less computationally intensive, but the disadvantage that it must be done carefully, or it can increase the dimension of the time series [3]. We do not understand why global PVD for $m \gg d$ behaves so poorly in this example.

In conclusion, while local PVD is optimal in the limit of low noise and a large number of data points, because of estimation problems due to curvature this advantage may be difficult to realize in practice. However, with certain data sets, Townshend [43] and Hunter [26] have found local PVD to be significantly advantageous.

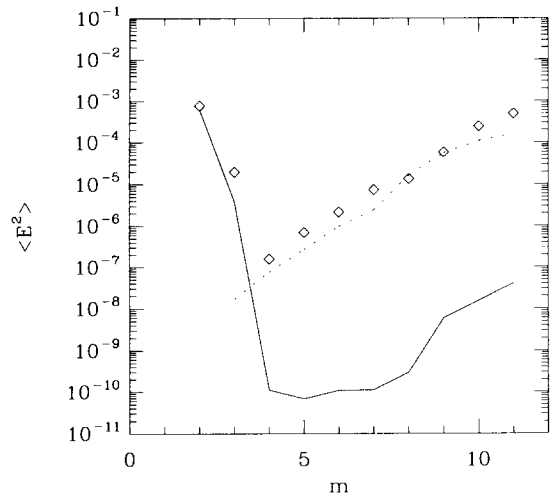

Fig. 26. Average prediction error for local PVD compared to delay coordinates, for the x-coordinate of the Ikeda map in the absence of noise. Both coordinate systems are based on predictive delay coordinates with $\tau = 1$ and $m = 2$ through 10. The vertical axis show the average error for 1000 predictions, made with local linear approximation using neighborhoods of 20 points from a 10000 point data set. The solid lines indicate prediction errors for delay coordinates. Diamonds indicate prediction errors for local PVD coordinates, where $d = 2$. For small m, delays and local PVD are roughly equivalent. At larger values of m, errors due to curvature dominate local PVD, resulting in less accurate predictions than delay coordinates. The dotted lines indicate errors due to curvature within local neighborhoods, estimated by the first (i.e. future-most) component of the difference between delay vectors and their projections onto the estimated tangent spaces.

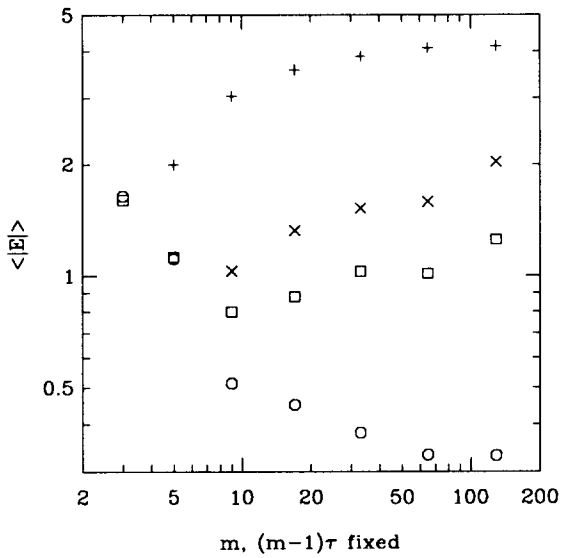

Fig. 27. Average prediction error for different coordinates from the Lorenz $x(t)$ time series versus m. In this plot, as m increases, τ decreases to keep $w = (m-1)\tau$ constant. Prediction errors for delay coordinates are plotted with squares, local PVD ($d' = 3$) with octagons, global PVD ($d' = 4$) with plus signs, and global PVD ($d' = 7$) with crosses. Local PVD coordinates are superior to delay coordinates for large m, because of a smaller number of parameters needed to estimate maps. All prediction was done with local linear approximation from a 10000 point time series, sampled at $\Delta t = 0.01$ in the time units of the Lorenz equations. The window was fixed at $w = 1.28$; predictions were made for extrapolation time $T = 0.10$.

8.2. *Improving estimation by warping of coordinates*

The analysis of section 7.1 shows that the estimation errors in one coordinate system may differ from those in another coordinate system. This immediately suggests that estimation errors can be reduced by an appropriate warping of coordinates. In this section we propose an algorithm for reducing estimation errors. Although this algorithm was motivated by the problem of state space reconstruction, we believe that it can be used to improve performance of local function approximation methods in a general context.

From eq. (81), for delay coordinates, the estimation error is

$$E \approx \pi \, DF^T \epsilon_{\underline{x}}, \tag{90}$$

where $F^T = \Phi f^T \Phi^{-1}$ is the dynamical system in delay coordinates, which can be estimated directly from the time series, π is a row vector which projects a delay vector onto the first coordinate axis, and $\epsilon_{\underline{x}}$ is the distance vector to the nearest neighbor in delay coordinates. The orientation of $\epsilon_{\underline{x}}$ that gives the largest error in an approximation of πF^T for a given $\|\epsilon_{\underline{x}}\|$ is given by

the unit vector

$$v = \frac{(\mathrm{D}F^T)^\dagger \pi^\dagger}{\left\|(\mathrm{D}F^T)^\dagger \pi^\dagger\right\|}. \tag{91}$$

The shape of the local neighborhood can be changed to reduce the errors by simply transforming to coordinates x' that are dilated by a factor $\alpha > 1$ along the v axis. A ball in the primed coordinates corresponds to an ellipsoid which is squeezed along the v-axis in the original coordinates.

In order to guarantee that the neighborhood contains K points, the neighborhood must be expanded in other directions to compensate for its contraction in the ν direction. However, in the limit as $\epsilon \to 0$, directions orthogonal to v do not effect the estimation error *in the direction corresponding to the projection operator* π. As a result, the estimation error in this direction is reduced by a factor of α^{-1}. Of course, for finite $\epsilon_{\underline{x}}$ there is an upper bound on α above which nonlinearities will dominate, and beyond which it is impossible to reduce the estimation error without carrying this procedure to higher order.

This suggests the following algorithm for adaptively improving the approximation of $\pi F^T(\underline{x})$ near the point $\underline{\tilde{x}}$.

(1) Approximate $\pi \hat{F}^T \approx \pi F^T$ near $\underline{\tilde{x}}$ using balls in the original coordinate system.

(2) Compute the vector v for the projection $\pi \hat{F}^T$ according to eq. (91).

(3) Transform to new coordinates $\underline{x}'$ by dilating the coordinates of each nearby point $\underline{x}$ by a factor α along the v-axis.

(4) Use the distorted neighborhood to approximate $\pi \hat{F}^T$ locally.

(5) Compute the estimation error $\langle E^2 \rangle = \langle (\pi \, \mathrm{D}\hat{F}^T \epsilon_{\underline{x}'})^2 \rangle$ for the points in the neighborhood of $\underline{x}'$.

(6) Try another value of the scalar parameter α, and repeat steps (3)–(5), searching for the value that minimizes $\langle E^2 \rangle$.

This procedure is straightforward to generalize to higher orders, for example distorting the neighborhood by a quadratic rather than a linear transformation. This may improve the effectiveness of the algorithm if the original approximation $\pi \hat{F}^T$ is accurate enough to support this.

This algorithm could be applied to the local approximation of more general mappings F and π.

9. Conclusions

In dynamical systems theory it is customary to emphasize invariance under changes of coordinates. In practical problems, however, the choice of coordinates can be very important. State space reconstruction is a case in point. Although a naive interpretation of Takens' theorem might suggest that any coordinate system that forms an embedding is equivalent to any other, in practice the choice of coordinates dramatically affects the ability to make predictions. A poor reconstruction amplifies noise and increases estimation error.

The theoretical treatment that we have presented here is designed to help understand these problems, and give insight into the construction of the best possible coordinates. Errors due to state space reconstruction involve a tradeoff between two effects: For small window widths, observational noise amplification is typically the dominant source of errors; it can be minimized by making the embedding dimension and window width as large as possible. For large window widths, estimation error is typically the dominant effect; it can be minimized by making the window width as small as possible. An optimal choice of coordinates balances these two effects.

For chaotic dynamical systems, mixed reconstructions, which use both past and future information, are preferable to those that are exclusively based on either the past or future. By making the past and future window widths sufficiently large, the noise amplification of a mixed

reconstruction can be made arbitrarily small. For problems such as computation of fractal dimension, mixed coordinates, in combination with a noise reduction algorithm, may give better results [15, 16, 25]. However, for prediction problems past-based coordinates are unavoidable.

The limits to state space reconstruction depend on the properties of the dynamical system. The behavior of the noise amplification is complicated, with several different scaling regimes (as schematically illustrated in fig. 13). However, if all other factors are held constant, the noise amplification generally increases according to the largest Lyapunov exponent and the dimension of the attractor. Similarly, for large window widths the logarithm of the estimation error increases according to the ratio of the metric entropy over the dimension (eq. (89)). We have suggested an algorithm for minimizing the effect of estimation errors, which we hope someone else will explore in more detail in a future paper.

One of the problems we have investigated is that of reconstructing smooth coordinates which compress all the available information into the smallest possible dimensional space. In general, to do this it is necessary that the coordinates be nonlinear. We have shown that in the low noise limit, coordinates which are optimal in this sense can be constructed by the method of local singular value decomposition. When the dynamical system and measurement function are unknown, these coordinates can be estimated directly from a time series. However, this involves estimation errors. If predictability is limited by noise such nonlinear coordinates may be useful, but if predictability is limited by lack of data, delay coordinates are probably a better choice.

Perhaps our most significant result in this paper concerns the limits of predictability. It is now a well known fact that chaos limits long-term predictability. We have shown that *when projected into lower dimensions, chaos may also impose limits to short term predictability*. For a dynamical system whose dimension and leading Lyapunov exponent are sufficiently large, projection onto a low dimensional time series causes an explosion in the noise amplification. As a result, it is impossible to reconstruct localized states from measurements of any reasonable precision. The time series is unpredictable for times much less than the Lyapunov time and it becomes indistinguishable from one generated by a random process. This is true even when the dynamical system is known. Note that this is *not* true for nonchaotic systems – as long as the dimension is finite, it is always possible to localize states by taking a sufficient number of measurements. The projection of chaos onto lower dimensions may explain the origin of many random processes.

The results we have presented here suggest many avenues for future work. One obvious problem is to extend the framework we have developed to include dynamic noise; although we have argued that the effects of dynamic noise are similar to those of local estimation error, this is only true as a first approximation, and more work needs to be done. Another interesting problem is to extend the treatment of estimation error to other function approximation methods, in particular those involving global function representations. Finally, with a limited number of data points and finite noise levels the problem of optimal coordinate reconstruction still remains to be solved.

Acknowledgements

We would like to thank David Broomhead, Peter Grassberger, and Wallace Larimore for useful conversations, and James Theiler for a critical reading of the manuscript. We are grateful for support from the National Institute for Mental Health under grant 1-R01-MH47184-01.

References

[1] H.D.I. Abarbanel, R. Brown and J.B. Kadtke, Prediction and system identification in chaotic nonlinear systems: Time series with broadband spectra, Phys. Lett. A 18 (1989) 401–408.

[2] Z. Aleksić, Estimating the embedding dimension, Physica D, to appear.

[3] R. Badii, G. Broggi, B. Derighetti, M. Ravani, S. Ciliberto, A. Politi and M.A. Rubio, Dimension increase in filtered chaotic signals, Phys. Rev. Lett. 60 (1988) 979.

[4] J.L. Breeden and A. Hübler, Reconstructing equations of motion from experimental data with unobserved variables, Phys Rev. A 42 (1990) 5817–5826.

[5] D.S. Broomhead, R. Indik, A.C. Newell and D.A. Rand, Local adaptive Galerkin bases for large dimensional dynamical systems, Nonlinearity (1991), to appear.

[6] D.S. Broomhead, R. Jones and G.P. King, Topological dimension and local coordinates from time series data, J. Phys. A 20 (1987) L563–L569.

[7] D.S. Broomhead and G.P. King, Extracting qualitative dynamics from experimental data, Physica D 20 (1987) 217.

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

[9] A. Čenys and K. Pyragas, Estimation of the number of degrees of freedom from chaotic time series, Phys. Lett. A 129 (1988) 227.

[10] J. Cremers and A. Hübler, Construction of differential equations from experimental data, Z. Naturforsch., 42a (1987) 797–802.

[11] J.P. Crutchfield and B.S. McNamara, Equations of motion from a data series, Complex Systems 1 (1987) 417–452.

[12] J.D. Farmer, Chaotic attractors of an infinite-dimensional dynamical system, Physica D 4 (1982) 366–393.

[13] J.D. Farmer, E. Ott and J.A. Yorke, The dimension of chaotic attractors, Physica D 7 (1983) 153–180.

[14] J.D. Farmer and J.J. Sidorowich, Predicting chaotic time series, Phys. Rev. Lett. 59 (1987) 845–848.

[15] J.D. Farmer and J.J. Sidorowich, Exploiting chaos to predict the future and reduce noise, in: Evolution, Learing and Cognition, ed. Y.C. Lee (World Scientific, Singapore, 1988).

[16] J.D. Farmer and J.J. Sidorowich, Optimal shadowing and noise reduction, Physica D 47 (1991) 373.

[17] A.M. Fraser, Information and entropy in strange attractors, IEEE Transactions on Information Theory, IT-35 (1989).

[18] A.M. Fraser, Reconstructing attractors from scalar time series: A comparison of singular system and redundancy criteria, Physica D 34 (1989) 391–404.

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

[20] H. Froehling, J.P. Crutchfield, J.P. Farmer, N.H. Packard and R.S. Shaw, On determining the dimension of chaotic flows, Physica D 3 (1981) 605.

[21] J. Geweke, Inference and forecasting for chaotic nonlinear time series, in preparation.

[22] J.F. Gibson, M. Casdagli, S. Eubank and J.D. Farmer. Principal component analysis and derivatives of time series, Technical report LA-UR-90-2117, Los Alamos National Lab (1990).

[23] P. Grassberger, Information content and predictability of lumped and distributed dynamical systems, Technical Report WU-B-87-8, University of Wuppertal (1987).

[24] J. Guckenheimer, Noise in chaotic systems, Nature 298 (1982) 358–361.

[25] S.M. Hammel, A noise-reduction method for chaotic systems, Phys. Lett. A 148 (1990) 421–428;
E.J. Kostelich and J.A. Yorke, Noise reduction in dynamical systems, Phys. Rev. A 38(3) (1988).

[26] N.F. Hunter, Pleistocene climate as a dynamic system, in: Nonlinear Prediction and Modeling, eds. M. Casdagli and S. Eubank (Addison-Wesley, Reading, MA, 1991).

[27] J.L. Kaplan and J.A. Yorke, Chaotic behavior of multidimensional difference equations, in: Functional Differential Equations and Approximations of Fixed Points, eds. H.-O. Peitgen and H.-O. Walther, Springer Lecture Notes in Mathematics, Vol. 730 (Springer, Berlin, 1979) p. 204.

[28] A.S. Lapedes and R. Farber, Nonlinear signal processing using neural networks: Prediction and system modeling, Technical Report LA-UR-87-2662, Los Alamos National Laboratory (1987).

[29] W. Larimore, System identification, reduced order filtering, and modelling via canonical variate analysis, in: Proc. 1983 American Control Conf. (1983).

[30] W. Liebert, K. Pawelzik and H.G. Schuster, Optimal embedding of chaotic attractors from topological considerations (1989).

[31] W. Liebert and H.G. Schuster, Proper choice of the time delay for the analysis of chaotic time series, Phys. Lett. A 142 (1988) 107–111.

[32] A.I. Mees, Modelling complex systems, in: Dynamics of Complex Interconnected Biological Systems, eds. T. Vincent, L.S. Jennings and A.I. Mees, (Birkhauser, Boston, 1990) pp. 104–124.

[33] N.H. Packard, J.P. Crutchfield, J.D. Farmer and R.S. Shaw, Geometry from a time series, Phys. Rev. Lett. 45 (1980) 712–716.

[34] M.B. Priestley, State dependent models: A general approach to nonlinear time series analysis, J. Time Series Anal. 1 (1980) 47–71.

[35] M.B. Priestley, Spectral Analysis of Time Series (Academic Press, New York, 1981).

[36] T. Sauer, J. Yorke, M. Casdagli and E. Kostelich, Embedology, Technical report, University of Maryland (1990).

[37] R. Savit and M. Green, Time series and independent variables, Physica D 50 (1991) 95–116.

[38] R.S. Shaw, Strange attractors, chaotic behavior, and information flow, Z. Naturforsch. 36a (1981) 80–112.

[39] R.S. Shaw, The Dripping Faucet as a Model Dynamical System (Aerial Press, Santa Cruz, 1984).

[40] B.W. Silverman, Kernel Density Estimation Techniques for Statistics and Data Analysis (Chapman Hall, London, 1986).

[41] F. Takens, Detecting strange attractors in fluid turbulence, in: Dynamical Systems and Turbulence, eds. D. Rand and L.-S. Young (Springer, Berlin, 1981).

[42] H. Tong and K.S. Lim, Threshold autoregression, limit cycles and cyclical data, J. R. Stat. Soc. B 42(3) (1980) 245–292.

[43] B. Townshend, Nonlinear prediction of speech signals, in: Nonlinear Prediction and Modeling, eds. M. Casdagli and S. Eubank (Addison-Wesley, Reading, MA, 1991), to appear.

[44] G.U. Yule, Phil. Trans. R. Soc. London A 226 (1927) 267.