LA-UR 90-4342

CONF-9007195

Received by OSTI

JAN 0 8 1991

LA-UR--90-4342

DE91 005851

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: A THEORY OF STATE SPACE RECONSTRUCTION IN THE PRESENCE OF NOISE

AUTHOR(S): Martin Casdagli, Theoretical Division, LANL and Santa Fe Institute
Stephen Eubank, J. Doyne Farmer, and John Gibson
Theoretical Division and Center for Nonlinear Studies
Los Alamos National Laboratory, Los Alamos, NM

SUBMITTED TO:

## DISCLAIMER



Los Alamos National Laboratory
Los Alamos, New Mexico 87545



FORM NO. 836 R4
ST NO 2629 5/81

# A Theory of 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.*
and

*Santa Fe Institute*
*1120 Canyon Rd.*
*Santa Fe, NM 87501*

**Abstract**

Takens' theorem demonstrates that in the absence of noise a multidimensional state space can be reconstructed from a single time series. This theorem does not treat the effect of noise, however, and so gives no guidance about practical considerations for reconstructing a good state space. We study the problem of reconstructing a state space with observational noise, examining the likelihood for a particular state given a series of noisy observations. We define a quantity called the *distortion*, which is proportional to the covariance of the likelihood function in a reconstructed state space. This is related to the *noise amplification*, which corresponds to the root-mean-square errors for time series prediction with an ideal model. We prove that in the low noise limit minimizing the distortion is equivalent to minimizing the noise amplification.

We derive several asymptotic scaling laws for distortion and noise amplification. They depend on properties of the state space reconstruction, such as the sampling time and the reconstruction dimension, and properties of the dynamical system, such as the dimension and Lyapunov exponents. 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 all determinism is lost, even for short times, so that the time series is effectively a random process.

In the low noise, large data limit we show that the technique of local principal value decomposition (PVD) is an optimal method of state space reconstruction, in the sense that it achieves the minimum distortion in a state space of the lowest possible dimension.

# Contents

# 1 Introduction

There are many situations in which we observe a *time series* $\{x(t_i)\}, i = 1, \ldots, N$ which we believe to be at least approximately described by a $d$-dimensional dynamical system[1] $f$.

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

The time series is related to the original dynamical system by

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

We call $h$ the *measurement function.* The observed time series $x(t)$ is $D$-dimensional, so that $h : \Re^d \rightarrow \Re^D$. We are most interested in dimension-reducing measurement functions, where $D < d$, and we will 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 of reconstructing a state space is given in Figure (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, for example to construct a phase plot or one dimensional map, or for quantitative statistical characterizations, such as fractal dimension, Lyapunov exponents, or the eigenvalues of fixed points. We are particularly interested in state space reconstruction as it relates to the problem of nonlinear time series forecasting, a subject that has received considerable attention in the last few years [4, 6, 7, 8, 9, 15, 18, 19, 27, 21].

State space reconstruction was introduced into dynamical systems theory independently by Packard et al. [20], Ruelle[2] and Takens [26]. In fact, in time series analysis this idea is quite old, going back at least as far as the work of Yule in 1927 [28]. The important new contribution made in dynamical systems theory was the demonstration that it 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 a *delay vector*[3]

$$\underline{x}(t) = (x(t + \tau m_+), \ldots, x(t), \ldots, x(t - \tau m_-)) \tag{3}$$

[1] This is one of several possible ways of representing a dynamical system. $f^t$ is the map that 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.

[2] Private communication.

[3] For convenience we assume that the sampling time is uniform.

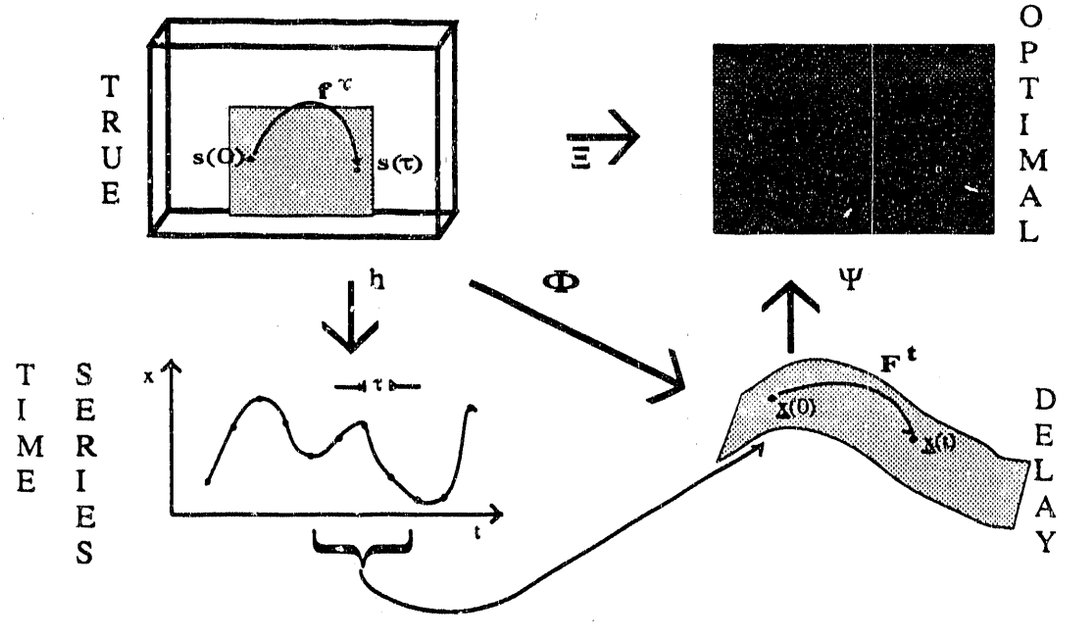


Figure 1: The true dynamical system $f$, its states $s$, and the measurement function $h$ are unobservables, locked in a black box. Values of the time series $x$ separated by intervals of the lag time $\tau$ form a delay vector $\underline{x}$ of dimension $m = 1 + m_+ + m_-$, where $m_+$ is the number of values from the future and $m_-$ is the number of values from the past. $\Phi$ maps the original $d$-dimensional state $s$ into the delay vector $\underline{x}$. $\Psi$ further maps the delay vector $\underline{x}$ into a new state $y$, of dimension $d' \leq m$.

The dimension of the delay vector is $m = 1 + m_- + m_+$. The number of samples taken from the past is $m_-$, and the number from the future is $m_+$. If $m_+ = 0$ then the reconstruction is *predictive*; otherwise it is *mixed.* The time separation between coordinates, $\tau$, is the *lag time.*

Takens studied the reconstruction map $\Phi$, which maps the states of a $d$-dimensional dynamical system into $m$-dimensional delay vectors.

$$\underline{x} = \Phi(s) = h(f^{\tau m_+}(s)), \ldots, h(s), \ldots, h(f^{-\tau m_-}(s)) \tag{4}$$

He showed that generically $\Phi$ is an embedding when $m \geq 2d + 1$. An *embedding* is a smooth, one-to-one coordinate transformation with a smooth inverse. When $\Phi$ is an embedding, $\Phi(\Re^d)$ is diffeomorphic to $\Re^d$.

If $\Phi$ is an embedding then a smooth dynamics $F$ is induced on the space of reconstructed vectors.

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

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

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 $\tau$ arbitrary[4], and any reconstruction is as good as any other. However in practice, the presence of noise in the data blurs states and makes picking a good lag time critical. Our work "fleshes out" Takens' proof, by examining how states are affected when conditions such as arbitrary precision are relaxed.

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

1. *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.

2. *Dynamical 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.

3. *Estimation error.* $f$ and $h$ are both unknown. They can be estimated, but with a finite amount of data some uncertainty remains.

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

[4] Provided it meets the conditions for genericity.

the individual measurements, then we have amplified the noise; the system is not as deterministic as it would be if we could observe more information.

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

$$\begin{aligned} \frac{dx}{dt} &= 10(y-x) \\ \frac{dy}{dt} &= -xz + 28x - y \\ \frac{dz}{dt} &= xy - \frac{10}{3}z \end{aligned} \tag{6}$$

Assume that we observe $x$. Since $\frac{dx}{dt}$ does not depend on $z$ directly, information about $z$ depends on the flow of information through $y$; when $z$ changes it causes $\frac{dy}{dt}$ to change, which causes $y$ and hence $\frac{dx}{dt}$ to change. When $x \approx 0$, since the only coupling 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 noise in the determination of $z$ from measurements of $x$ is acutely amplified when $x \approx 0$.

When noise is present, state space reconstruction becomes a problem in statistical estimation. The formalism that we develop in this paper makes the notion of noise amplification more 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 three factors:

- *The measurement function.* One observation 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 $\tau$.
- *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.

In assessing predictability it is important to distinguish between estimation error and noise amplification. Figure (2) shows a hypothetical comparison of two prediction problems in the idealized case of a one dimensional state space. The noise amplification is related to the "thickness" of the distribution of points. In Figure (2a) the noise amplification is large, and in Figure (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. The estimation error depends on the procedure used to approximate the dynamics. Noise

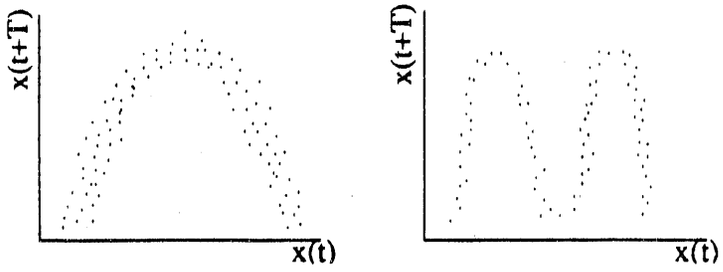


Figure 2: Two hypothetical scenarios for prediction in a one dimensional state space. The horizontal axis is the state at time $t$, and the vertical axis is the state a time $t+1$ in the future. (a) shows a coordinate system with high noise amplification, while (b) shows a coordinate system with low noise amplification. This is evident from the "thickness" of the distribution of points at any given $y$. 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).

amplification, however, does not. It sets a limit to predictability that is independent of the modeling procedure. In the limit of a large number of data points for most good approximation schemes the estimation error goes to zero. The prediction errors in this limit are given purely by the noise amplification. The noise amplification thus tells us the prediction errors that remain even with a perfect model.

As we shall show, when the dimension and Lyapunov exponents are sufficiently large there can be a complete breakdown of predictability, so that even with a perfect model the time series is unpredictable, even for short times. This is the limit in which a time series becomes a true random process.

Any approach to state space reconstruction uses the information in delay coordinates as a starting point. For some purposes, such as reducing the dimension, 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 $\Psi$ are differentiation or the singular value decomposition in PVD. By splitting the reconstruction process into $\Phi$ and $\Psi$, we have conveniently labeled the two parts of the problem. The choice of $\Phi$ determines the form of the delay coordinates, which are the raw information we have to work with, while $\Psi$ determines how we use that information. The total reconstruction map $\Xi = \Psi \circ \Phi$ takes the original coordinates $s$ to the reconstructed coordinates $y$.

We show here that it is impossible to reduce the noise amplification by transforming delay coordinates by $\Psi$. The minimum possible noise amplification over all $\Psi$ is obtained when $\Psi = \mathbf{1}$ and $y = \underline{x}$. However, 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. This can be desirable for reducing statistical estima-

tion errors, which typically increase with dimension (often exponentially). The local principal value decomposition technique discussed in Section 6 accomplishes this in the minimum possible dimension.

## 1.1 Approach and overview

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 times 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. We are currently investigating the implications of these theoretical results for the case when $f$ and $h$ are unknown, and will report the results in a future publication.

Throughout this paper we assume that the noise is entirely observational. Treating dynamical noise is obviously important, but it is outside the scope of this paper. However, we suspect that many aspects of the framework we have established here can be used to analyze dynamical noise as well.

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 extensions to Takens' theorem, and present an intuitive discussion of why it is true.

In Section 3 we derive formulas for the likelihood function and compute it for several examples. We use color graphics to develop intuition and to qualitatively illustrate what factors are necessary in order to obtain a good state space reconstruction.

From a practical point of view it is important to have simple criteria for selecting a reconstruction. The likelihood function gives a complete description of a reconstruction, but it 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 mean-square errors in the state space, and noise amplification, which is related to errors in time series prediction. We discuss the relationship between distortion and noise amplification, showing that minimizing one is equivalent to minimizing the other.

In Section 5 we discuss the problem of deciding how much noise amplification to expect in a given situation. We demonstrate that for a given $\tau$ noise amplification is a monotonically non-increasing function of $m$. We derive the behavior of the noise amplification as a function of $m$, $\tau$, $d$, and the Lyapunov exponents, testing our conclusions on a simple example. We show that for predictive coordinates an explosion in the noise amplification occurs when the Lyapunov exponents and dimension are sufficiently large, which causes a transition from behavior that is approximately de-

| symbol | description |
|---|---|
| $s(t)$ | $d$-dimensional state at time $t$ |
| $f^t$ | time-$t$ map of dynamical system; $s(t) = f^t(s(0))$ |
| $x(t)$ | $D$-dimensional value of time series at time $t$. (We often assume $D = 1$.) |
| $h$ | measurement function $x(t) = h(s(t))$ |
| $S(t)$ | $d - D$ dimensional measurement surface $S(t) = \{s : x = h(s)\}$ |
| $\tau$ | sampling time $t_{i+1} - t_i$ |
| $\underline{x}$ | $m$-dimensional delay vector $(x(t + \tau m_+), \ldots, x(t), \ldots, x(t - \tau m_-))$ |
| $y$ | reconstructed $d'$-dimensional coordinate based on $\underline{x}$ |
| $\Phi$ | delay reconstruction map $\underline{x} = \Phi(s)$ |
| $\Psi$ | map taking delay vector to new coordinate $y = \Psi(\underline{x})$ |
| $\Xi$ | total reconstruction map $\Xi = \Psi \circ \Phi$ |
| $\xi(t)$ | noise fluctuation, usually assumed to be Gaussian IID |
| $\underline{\xi}$ | $m$-dimensional vector of noise fluctuations $(\xi(t + \tau m_+), \ldots, \xi(t), \ldots, \xi(t - \tau m_-))$ |
| $\check{\underline{x}}$, $\check{s}$ | true values of $\underline{x}$, $s$ in absence of noise |
| $\hat{s}$, $\hat{x}$, $\hat{f}$ | best estimate for $s$, $x$, $f$ |
| $p$ | probability density function (identified by its arguments) |
| $p(x\|y)$ | probability density for $x$ given $y$ |
| $\Sigma$ | distortion matrix |
| $\delta$ | Trace of $\Sigma$ |

Table 1: Notation used in this paper.

terministic for short times to behavior that is effectively random over any time scale. Finally, we use two examples to illustrate several aspects of the behavior of the noise amplification.

In Section 6 we study the effect of making further coordinate transformations to delay coordinates. We demonstrate that in the low noise, large data limit, local principal value decomposition is an optimal state space reconstruction method in the sense that it minimizes the noise amplification with a coordinate system of the smallest possible dimension.

### 1.2 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 PCA. 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 is not clear which method is the best.

Delay coordinates are currently the most widely used choice. They have the nice property that the statistical properties of each dimension are the same. They have the unpleasant property that in order to use them it is necessary to choose the delay parameter $\tau$. If $\tau$ is too small each coordinate is almost the same, and the trajectories of the reconstructed space are squeezed along the identity line; if $\tau$ 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. Most of the research on the state space reconstruction problem has centered on the problems of choosing $\tau$ and $m$ for delay coordinates. The proposals for doing this include information-theoretic quantities [13, 11, 1], and others [5].

Another method for reconstructing a state space is the *method of derivatives*, originally investigated by Packard et al. [20]. The coordinates are derivatives of successively higher order.

$$\begin{aligned} y_1(t) &= x(t), \\ y_2(t) &= \hat{x}'(t), \\ &\vdots \\ y_m(t) &= \hat{x}^{(m-1)}(t). \end{aligned} \tag{8}$$

$\hat{x}^{(j)}(t)$ is a numerical approximation to the $j^{th}$ derivative of $x(t)$. As Takens proved, as long as $m$ is sufficiently large derivatives define an embedding.

There are many different methods 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 used. Straightforward methods of numerical differentiation act as a high pass filter, with a response function that is proportional to frequency. The quality of derivative coordinates in the presence of noise can be considerably improved by using a numerical algorithm that uses low pass filtering to balance the response function.

The other 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 [3]. The simplest way to implement their procedure is to compute the covariance matrix $C_{ij} = \langle x_i x_j \rangle_t$, and then compute its eigenvalues $\alpha_i$. ($x_i$ represents the $i$th coordinate of the delay vector $\underline{x}$; $\langle\rangle_t$ denotes a time average.) 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.

We have recently shown that PVD coordinates are very closely related to (appropriately low pass filtered) derivative coordinates [14].

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 PVD [12]. 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 [12], nonlinear coordinate transformations can be greatly superior. The method of local PVD, discussed in Section 6 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 $\underline{x} = \Phi(s)$. For a univariate time series ($D = 1$) this can be regarded as a set of $m$ simultaneous nonlinear equations in $d$ variables. The transformation $\Phi$ maps a $d$-dimensional surface into an $m$-dimensional space. If the surface $\Phi(s)$ contains no self intersections then there is a unique solution for $s$ given any $x$, and $\Phi$ is an embedding. If $\Phi$ is sufficiently close to a linear transformation then this may be possible with $m = d$. In general, however, for a unique solution we must have $m > d$. Generically, when $2d \geq m$ a $d$-dimensional surface in $m$ dimensions has self intersections on sets of dimension $2d - m$; when $2d < m$ generically it has no self intersections at all. The case when $d = 2$ and $m = 3$, for example, is shown in Figure (3); in this case there are typically self intersections along one dimensional curves. When $m = d + 1$ the set of self intersections is typically of dimension $d - 1$, and $\Phi$ is an embedding almost everywhere. As $m$ increases the dimension of the set of self intersections decreases, until finally when $m > 2d$ there are no self intersections at all. Thus, $m \geq 2d + 1$ *guarantees* that $\Phi$ is an embedding, it is *possible* that it will be an embedding with $m$ as small as $m = d$. See reference [23] for a more complete discussion.

The reconstruction process can also be considered in terms of the constraint that each measurement causes in the original state space. 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* $S(t)$ be the set of possible states that are consistent with a given measurement $x(t)$, i.e., $S(t) = \{s(t) : x(t) = h(s(t))\}$. When $h$ is smooth $S(t)$ is a surface of dimension $d - D$. For example, when $d = 2$ and $h$ is projection onto the horizontal axis, the measurement surfaces consist of horizontal lines. The effect

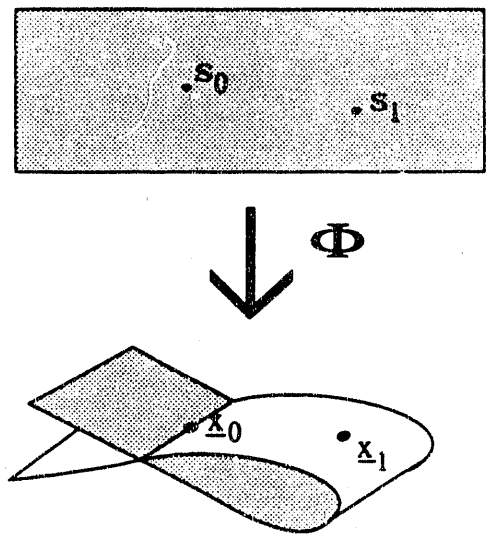


Figure 3: Solutions of the equation $\underline{x} = \Phi(s)$ when $d = 2$ and $m = 3$. There are typically self intersections along curves. For example, the state $s_0$ is mapped onto a self intersection, while $s_1$ is not. Except for special values of $s$ like $s_0$, $\Phi$ defines an embedding.

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 $I(t)$.

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

The intersection $I(t)$ is never empty, since there must be at least one state consistent with all the measurements. If $I(t)$ does not consist of a single point, $\Phi$ is not an embedding. An example for the case when $d = 2$ and $m = 3$ is shown in Figure (4).

In most real situations $f$, $h$, and consequently $\Phi$ are unknown. Nonetheless, as long as there is a smooth one-to-one correspondence between the delay coordinate and the original state we know that there is an embedding, so that the delay coordinate $\underline{x}$ can be used in place of the original coordinate $s$.

# 3 Geometry of reconstruction with noise

The goal of reconstruction is to assign a state based on a series of measurements. With noise this task is considerably more difficult because the measurements are uncertain, and 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[5] $p(s|\underline{x})$. This illustrates how the presence of noise com-

[5] We use probability density functions rather than measures only because we want to keep the discussion accessible to the widest audience possible. All of the statements given here can be recast in more rigorous terms using measures.

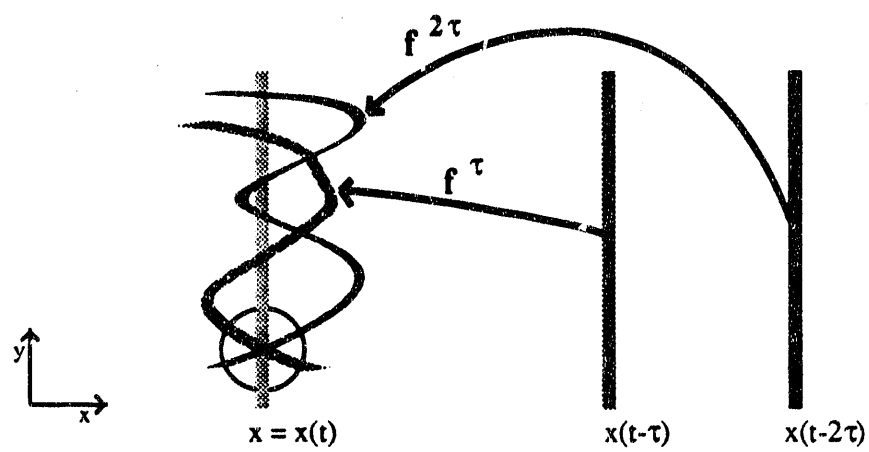


Figure 4: *A dynamical view of reconstruction in terms of the evolution of measurement surfaces.* Suppose that the measurement function $h$ corresponds to projection onto the horizontal axis, so that $h(s) = x$. A measurement at time $t$ implies that $s$ lies somewhere along the light gray vertical line 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 curve can be mapped forward by $f$ to the same time $t$. If their intersection is a not single point, then the reconstruction is not an embedding.

plicates 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 a very complicated, as has been demonstrated by Geweke.

In this section we derive several formulas 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 $\underline{x}$ and on the properties of the reconstruction problem.

## 3.1 The likelihood function

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

$$p(s|\underline{x})p(\underline{x}) = p(\underline{x}|s)p(s) \tag{10}$$

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 of a given series of observations. 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 term on the left represents what we know about $s$ after taking the observations $\underline{x}$ into account, and is called the *posterior*.

When $f$ and $h$ are known we can write down a formula for the likelihood function. Assume the noise $\xi$ is zero mean.

$$p(\underline{x}|s) = p(\underline{x} - \tilde{\underline{x}}) \tag{12}$$

where $\tilde{\underline{x}}$ is the "true" value of $\underline{x}$, in the absence of noise. It is related to the state $s$ by $\tilde{\underline{x}} = \Phi(s)$. If we furthermore assume that the noise is IID (each fluctuation is statistically Independent, and Identically Distributed with probability $p$),

$$p(\underline{x}|s) = \prod_{i=-m_-}^{i=m_+} p(x(t+i\tau) - h(f^{\tau i}(s))) \tag{13}$$

### 3.1.1 Gaussian noise

If we assume that $p(\xi)$ is a Gaussian of variance $\epsilon^2$, this becomes

$$p(\underline{x}|s) = \prod_{i=-m_-}^{i=m_+} \frac{1}{\sqrt{2\pi}\epsilon} \exp -\frac{(x(t+i\tau) - \tilde{x}(t+i\tau))^2}{2\epsilon^2} \tag{14}$$

If we assume the Euclidean norm, using the definition of $\Phi$ this can be rewritten as

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

where $A$ is a normalization constant.

Thus, the probability for $\underline{x}$ given the true value of $s$ is quite simple: it is an isotropic Gaussian centered on the true delay vector $\tilde{\underline{x}} = \Phi(s)$. The probability for $s$ given $\underline{x}$, in contrast, is much more complicated; using Bayes theorem (Equation (11)) gives

$$p(s|\underline{x}) = A'p(s)\exp -\frac{1}{2\epsilon^2}\|\underline{x} - \Phi(s)\|^2 \tag{16}$$

where $A'$ is another normalization constant. Although this looks quite similar to Equation (15), it is actually quite different, as it is interpreted as a function of $s$ rather than $\underline{x}$. Because of the nonlinear function $\Phi$, it is not a Gaussian.

### 3.1.2 Uniform bounded noise

Another case that is easily treated is that of uniform bounded noise,

$$p(\xi) = \begin{cases} \frac{1}{2\epsilon}, & \text{if } |\xi| \leq \epsilon \\ 0, & \text{if } |\xi| > \epsilon. \end{cases} \tag{17}$$

The effect of a given measurement can be visualized geometrically in terms of the *measurement strip* $S_\epsilon(t) = \{s : |x(t) - h(s)| < \epsilon\}$. The measurement strip is the support of $p$, and is similar to the measurement surface discussed earlier, except that it is "thickened" by $\epsilon$. Following Equation (13) the likelihood function can be computed in a manner analogous to Equation (9). The state must lie inside the intersection of the measurement strips.

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

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 tool for determining whether or not a given point $s \in I_\epsilon(t)$ is to test whether it satisfies the condition

$$f^{\tau m_+}(s) \in S_\epsilon(t + \tau m_+) \wedge \ldots \wedge s \in S_\epsilon(t) \wedge \ldots \wedge f^{-\tau m_-}(s) \in \wedge S_\epsilon(t - \tau m_-) \tag{19}$$

### 3.1.3 Chaotic geometry

We have performed several numerical experiments using the above formulae for likelihood functions, in particular for two dimensional dynamical systems such as the Ikeda map. This has allowed us to investigate the effects of varying the number of measurements, varying the noise level, and the effect of homoclinic tangencies. We have found the use of color graphics particularly helpful in exploring how these various effects interact with each other. We intend to illustrate these findings in a future publication.

# 4 Criteria for optimality of coordinates

To choose between different reconstructed coordinate systems we must first decide what we mean by "best". Since our primary interest is in prediction, in principle we could just use each coordinate system to estimate a dynamical system $\hat{f}$, and test to see which coordinate system gives the best predictions. In practice, however, this is cumbersome, since it is necessary to test many different statistical estimation procedures as well as many different coordinate systems. The errors would contain the effects of both estimation error and observational noise, and the result of such an experiment might give very little theoretical insight. Instead, we are going to begin by developing a formalism for understanding the effects of observational error alone, and we defer the problem of estimation error to a later publication.

The presence of noise forces us to think about the time series in probabilistic terms. The predictive value of a given reconstructed coordinate $y$ is given by the conditional probability density of a future value of the time series given $y$.

$$p(x(T)|y(0)) = \textit{probability } x(T) \textit{ given } y(0) \tag{20}$$

$p(x(T)|y(0))$ tells us what we can predict about $x(T)$ given $y(0)$, assuming a perfect estimation procedure. This applies to either iterative or direct estimation procedures [4, 9]. Any criterion for evaluating the effect of observational noise must be based on $p(x(T)|y(0))$.

## 4.1 Error measures

To choose between different coordinate systems, we must decide which property of $p(x(T)|y(0))$ to optimize. This amounts to deciding what we mean by the "best" predictions. There are several natural possibilities:

- *Maximum expectation.* We can view this as a gaming problem in which we must bet on the value of $x(T)$. We naturally want to maximize our expectation, which we can do by choosing the embedding which gives $p(x(T)|y(0))$ with the largest possible maximum value. However, since $x(T)$ is a continuous variable, we can never predict it exactly, and this is not well defined unless a coarse graining is specified. Also, this is only optimizes the rate of return with an infinite bank.

- *Mutual information.*

  Let $H$ represent the entropy function $H(x) = \int p(x) \log p(x) dx$. We can choose the coordinate system that maximizes the mutual information $I(x, y) = H(x) - H(x|y)$, where $H(x|y)$ is the entropy associated with the conditional probability density $p(x|y)$. Note that since $H(x)$ is fixed, this is equivalent to minimizing the conditional entropy $H(x|y)$. This criterion has the disadvantage that for continuous variables it may have little to do with prediction error. For example, a $p(x|y)$ with sharp peaks at different values of $x$ may have a very low entropy

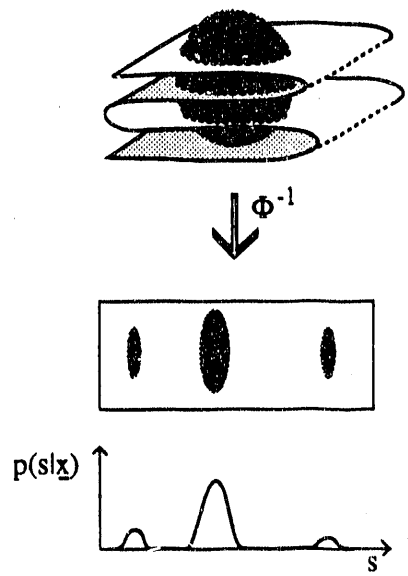


Figure 5: *The conditional probability of $s(t)$ given $\underline{x}(t)$.* Observational noise induces an uncertainty in the delay vector, here symbolized as a "noise ball". The noise ball, in turn, induces a conditional probability density for the true state given the delay vector. The variance of this density quantifies the quality of the embedding. Note that a sharp, multimodal density function can have a low entropy but a large variance.

but a high variance[6]. This makes the mutual information very insensitive to whether or not a reconstructed coordinate system forms an embedding. Several authors have looked at various criteria relating to mutual information [25, 13, 11, 12].

- *Variance.* The variance

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

is a lower bound on the mean-square prediction error $E(x - \hat{x})^2$ as follows (see [22] for details). It can be shown that the mean square prediction error is minimized by taking the predictor $\hat{x} = E(x|y)$. The prediction error of this ideal predictor is then $Var(x|y)$, which is a lower bound for any other predictor. For continuous variables this is a very natural way to evaluate the quality of predictions. Furthermore, it is easy to calculate analytically and estimate numerically. We shall use this measure.

[6] At any finite level of resolution, $x(T)$ may be thought of as a "message", with a corresponding number of bits, as originally proposed by Shaw [24, 25]. The entropy tells us our uncertainty in predicting this message. However, it weights the low order bits equally with the high order bits. In predicting a continuous variable, however, an error in the highest order bit is usually much worse than one in the lowest order bit. The inability of the entropy to make this distinction makes it a poor measure of the quality of predictions.

- *Other measures*, such as mean-absolute error or the geometric mean error have the advantage that, compared to mean-square error, they do not emphasize outliers. Our choice of variance as compared to these other measures is primarily one of convenience.

## 4.2 Noise amplification

The observational errors have a variance of $\epsilon^2$. We define the *noise amplification at a given noise level* $\epsilon$, $\sigma_\epsilon$, as the ratio of the variance in the future value $x(T)$ given the present reconstructed coordinate $y(0)$, to the variance of the observational errors.

$$\sigma_\epsilon(T) = \frac{1}{\epsilon}\sqrt{\mathrm{Var}(x(T)|y(0))} \tag{22}$$

The quantity $\sigma_\epsilon$ has the advantage that it can be estimated directly from a time series [7], and so can be used as an operational test for the quality of a given set of coordinates.

The quantity $\sigma_\epsilon$ depends on both $\epsilon$ and $p(\xi)$. We can remove this dependence by assuming the noise is Gaussian and taking the limit as $\epsilon \to 0$. We will call this simply the *noise amplification* $\sigma$.

$$\sigma(T) = \lim_{\epsilon \to 0} \sigma_\epsilon(T) \tag{23}$$

This limit may not always exist; in particular, when the state space reconstruction is not an embedding it will tend to infinity. However we will show in Section 4.4, that for state space reconstructions that are embeddings the limit exists, and in the case of Gaussian noise, it depends solely on geometric factors, specifically, the dynamical system $f$, the measurement function $h$, the embedding dimensions $m_+$ and $m_-$, the state $y(0)$, and the prediction time $T$. In Section 4.6 we will illustrate how in some situations the limit depends on properties of an underlying attractor, and the realization of the noise perturbations. At small noise levels, $\sigma$ can be used to provide an estimate of the "true" noise amplification $\sigma_\epsilon$. Geometrically, $\sigma_\epsilon$ measures the "thickness" of Figure (2) in the vertical direction at the state $x(t) = y(0)$.

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 [26]. 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.

If we are interested in a geometric object such as a chaotic attractor that has an ergodic measure, we can also eliminate the dependence on the state $y(0)$ by taking an averaging over the values of $y(0)$ with respect to this measure. We will call this the *average noise amplification*

$$\langle\sigma\rangle = \langle\sigma\rangle_{y(0)} \tag{24}$$

[7] Disregarding the factor of $1/\epsilon$.

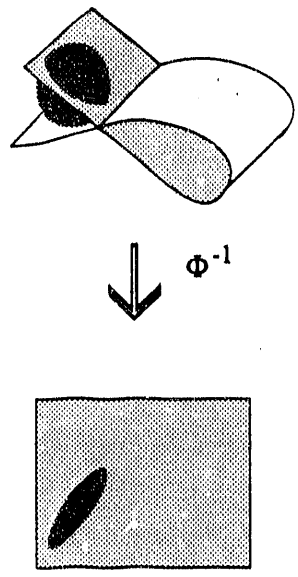


Figure 6: Since the errors in the time series are identically distributed, the probability density function for noisy delay coordinates is isotropic (top). This induces a non-isotropic distribution in the original space $S$ (bottom).

## 4.3 Distortion

The noise amplification has the disadvantage that it depends on the time $T$. This problem can be overcome by defining a quantity $\Sigma_\epsilon$ that is related to noise amplification, called the *distortion matrix at a given noise level* $\epsilon$ as follows

$$\Sigma_\epsilon = \frac{1}{\epsilon^2}\mathrm{Var}(s|y) \tag{25}$$

We define a related scalar quantity $\delta_\epsilon$ called the *distortion at a given noise level* $\epsilon$ [8]. by Equation (26).

$$\delta_\epsilon = \frac{1}{\epsilon}\sqrt{\mathrm{Var}(\|s\| \mid y)} = \sqrt{\mathrm{trace}\Sigma_\epsilon} \tag{26}$$

Finally, by analogy with Equation (23), we define the $\epsilon$-independent quantities $\Sigma = \lim_{\epsilon\to 0}\ \Sigma_\epsilon$ and $\delta = \lim_{\epsilon\to 0}\ \delta_\epsilon$. The motivation for defining the distortion in this way comes from considering the geometrical effects of reconstruction. In delay coordinates the probability distribution corresponding to the noise is isotropic. For example, for Gaussian noise a surface on which the probability density function $p(x)$ is a constant is an $m$-dimensional sphere, as shown in Figure (6). Assuming that $\Phi$ is an embedding in a neighborhood of this sphere, in the low noise limit this sphere will map into a $d$-dimensional ellipsoid in the original state space $S$. The distortion $\Sigma$ is a $d \times d$ symmetric real matrix, whose eigenvalues are proportional to the squares of the principal axes of this ellipsoid.

[8]The term "distortion" was originally used for another closely related quantity defined by Fraser [12].

The distortion has two disadvantages when compared with the noise amplification. First, it depends on the coordinates used to describe the dynamical system; for example, rescaling $s$ changes the distortion. Second, it is not observable, and cannot be computed from a time series alone. Nonetheless, the distortion is a valuable tool for understanding the behavior of different coordinate reconstructions. This is because of some theoretical relationships between distortion and noise amplification which we will describe in Section 4.5. These results tell us that, in order to understand the effect of noise on the embedding process, it should be helpful to examine how the uncertainty in measurements of the time series translates into uncertainty in the original state space.

From a practical point of view the distortion may seem less relevant than the noise amplification. However there are some cases where we may know parametric forms for $f$ and $h$, for example from scientific laws, and want to estimate the "hidden variables" $s$, or the unknown parameters of $f$ and $h$ from a noisy time series. Note that unknown parameters are just like dummy hidden variables. This is the problem in extended Kalman filtering, and has also been considered by Breiden et. al. [2]. Then the distortion matrix is of direct interest in quantifying the uncertainties in the estimates for the hidden variables. For example, in the Lorenz system, we considered how accurately it was possible to infer the value of $z$ given values for $x$.

## 4.4 Low noise limit

In the low noise limit, when $\Phi$ is an embedding the probability density $p(s|\underline{x})$ becomes much simpler. For example, take Equation (15), which assumes Gaussian noise and a uniform prior. We can rewrite it as

$$p(\underline{x}|s) = Ae^{-\frac{Q(s)}{2\epsilon^2}} \tag{27}$$

where $A$ is a normalization constant and $Q(s) = \|\underline{x} - \Phi(s)\|^2$. The most likely value of $s$ (the maximum likelihood solution $\hat{s}$) occurs at the maximum value of $p(s|\underline{x})$, where $DQ(\hat{s}) = 0$. If $\Phi$ is an embedding, then there exists a unique maximum for sufficiently small noise levels. When $\epsilon$ is small, $p(s|\underline{x})$ is concentrated near its maximum, and it is possible to get a good approximation for $p(s|\underline{x})$ by expanding $Q$ in a Taylor series about $\hat{s}$. To leading order in $s - \hat{s}$ this is[9]

$$p(s|\underline{x}) \approx C \exp -\frac{1}{2\epsilon^2}(s-\hat{s})^\dagger D\Phi^\dagger D\Phi(s-\hat{s}), \tag{28}$$

[9]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 $D^2Q = 2[D^2v^\dagger v + Dv^\dagger Dv]$. But $v$ is of order $\epsilon$, so the dominant term is $D^2Q(\hat{s}) = 2D\Phi^\dagger D\Phi$.

where the derivatives $D\Phi^\dagger$ and $D\Phi$ are evaluated at $s = \hat{s}$, and $C$ is a normalization constant. By inspection, from the definition of the distortion matrix we see that

$$\Sigma = (D\Phi^\dagger D\Phi)^{-1}. \tag{29}$$

Note that if $\Phi$ is an embedding then $D\Phi$ is of full rank, and $\Sigma$ is well defined. The uncertainty in the estimate of $s$ is thus an anisotropic Gaussian centered on the maximum likelihood estimate $\hat{s}$. The distortion of this Gaussian is given by the eigenvalues of $\Sigma$. The larger the eigenvalues of $\Sigma$, the more well defined the initial state.

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

$$D\Phi = \begin{pmatrix} DhDf^{\tau m+} \\ \vdots \\ Dh \\ \vdots \\ DhDf^{\tau m-} \end{pmatrix}. \tag{30}$$

Since $s$ is $d$-dimensional, as long as $\Phi$ is an embedding $D\Phi$ has $d$ nonzero singular values. The squares of these singular values are equal to the eigenvalues of $\Sigma$.

In control theory $D\Phi$ is called the *observability matrix.* A system is observable if the observability matrix has full rank, which is one of the conditions for $\Phi$ to be an embedding. Whether $D\Phi$ has full rank evidently depends on detailed properties of the coupling between variables in $f$, and on the measurement function $h$. For example if the dynamical system $f$ has a representation such that it splits into two non-interacting subsystems, and the measurement function is a constant (for example zero) on one of the subsystems, then intuitively one would expect that this subsystem is unobservable. Indeed in such an example all the columns of the observability matrix corresponding to this subsystem are zero, and it is not of full rank. On the other hand if the measurement function depends on both subsystems, then, by Takens' theorem, generically full rank will be attained. We will consider such an example in Section 5.

Finally, we can use Equations (28) and (29) above to derive Equation (31) for the noise amplification by transforming variables to $\check{x}(T) = h(f^T(s(0)))$. In the low noise limit we can take $h$ and $f$ to be approximately linear, so that a small variation of $\check{x}(T)$ about its mean value $\hat{x}(T)$ is $x(T) - \hat{x}(T) \approx DhDf^T(s(0) - \hat{s}(0))$. Then, since $s(0)$ has a Gaussian density with covariance matrix $\Sigma$, and covariance matrices transform under linear transformations $L$ according to $\Sigma \rightarrow L\Sigma L^\dagger$, it follows that $x(T)$ (the *noisy* future observation) has a Gaussian density with variance

$$\sigma(T) = 1 + DhDf^T\Sigma(Df^T)^\dagger Dh^\dagger. \tag{31}$$

Intuitively this makes sense; the uncertainty in the initial state is first altered by the derivative of the dynamics, and then projected down onto the time series and

finally convolved with noise. Also it is straightforward to verify that Equation 31 is invariant with respect to the representation of the underlying state space dynamics and measurement function.

## 4.5 Relation between noise amplification and distortion

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 minimize the distortion of a reconstruction by using only the information available in a time series [12]. We demonstrate that this is indeed possible in the low noise limit, by demonstrating a relation between the distortion matrix $\Sigma$ and the noise amplification $\sigma(T)$, which *is* computable from a time series.

Define an ordering on distortion matrices by $\Sigma_1 \leq \Sigma_2$ if $\Sigma_2 - \Sigma_1$ is positive semi-definite[10]. Consider the set of all reconstructions $y = \Psi(\underline{x})$, where $\Psi : \Re^m \rightarrow \Re^{d'}$ and $m$ and $d'$ are fixed. Then our result firstly states that if there exists a $y^*$ such that $\Sigma(y^*) \leq \Sigma(y)$ for all $y$, then $y^*$ will also satisfy $\sigma_T(y^*) \leq \sigma_T(y)$ for all $y$ and $T$. [11] The converse is also true generically: any reconstruction $y'$ that minimizes $\sigma(T)$ over $\Psi$ for all $T$ will also minimize $\Sigma$. Thus, 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. As we will see in Section 6, this is not as difficult as it might seem.

**Derivation.**

We can use Equation (31) to demonstrate that any reconstruction $y^*$ that minimizes the distortion $\Sigma$ will also minimize the noise amplification $\sigma(T)$ for any time $T$ as follows. Let $y^* = \Psi^*(\underline{x})$. Then $v(\Sigma(y) - \Sigma(y^*))v^\dagger \geq 0$ for any $d$-dimensional vector $v$ and any reconstruction $y$. By taking $v^\dagger = DhDf^T$, we have $\sigma(T, y) - \sigma(T, y^*) \geq 0$ for all $T$.

To demonstrate the converse, we proceed as follows. Let $y' = \Psi'\underline{x}$. As will be shown in Section 6, there exists a transformation $y^*$ such that $\Sigma(y^*) \leq \Sigma(y)$ for all $y$. It suffices to show that $\Sigma(y') = \Sigma(y^*)$. But by definition of $y'$ we know $\sigma(y', T) \leq \sigma(y^*, T)$ for all $T$, and by the the first part of the demonstration we know that $\sigma(y', T) \geq \sigma(y^*, T)$ for all $T$. It follows that $v_T^\dagger M v_T = 0$ for all $T$, where $v_T^\dagger = DhDf^T$, and $M = \Sigma(y') - \Sigma(y^*)$ is necessarily a positive semi-definite matrix. To complete the demonstration we must show that $M = 0$. Now transform to coordinates so that $M = diag(m_1, .., m_d)$. We obtain a contradiction if one or more of the $m_i$ are non-zero, because suppose (without loss of generality) that $m_1 > 0$. Then $v_T M v_T^\dagger \geq m_1 \|v_T^{(1)}\|^2 > 0$, where $v_T^{(1)}$ denotes the first component of $v_T$ in the new coordinates. Note that there must generically exist a $T$ such that $\|v_T^{(1)}\|^2 > 0$,

[10] By definition a $d \times d$ matrix $M$ is positive semi-definite if $v^\dagger M v \geq 0$ for all $d$-dimensional vectors $v$.

[11] In Section 6 we will show that such a $y^*$ exists, is generically unique up to invertible coordinate transformations, and show how to compute it straightforwardly from $f$ and $h$.

because a finite subset of the vectors $v_T$ make up an observability matrix of the form $D\Phi$, which by Takens' theorem is generically of full rank, so that the $v_T$ span $\Re^d$.

## 4.6 Numerical example: The Lorenz equations

In this subsection we illustrate the above ideas using the Lorenz equations as an example.

### 4.6.1 Low noise limit distortion

In Section 1 we gave an intuitive sketch of the flow of information between variables in the Lorenz system. We argued that when $x \approx 0$, the observations of $x$ tell us little about $z$. The distortion makes this notion precise. To illustrate how the flow becomes restricted as $x$ nears zero we numerically compute the distortion along a typical trajectory of the Lorenz attractor, using five dimensional delay coordinates with $m_+ = 0$ and $m_- = 4$, and $\tau = 0.01$ (by keeping the $\tau$ small, we guarantee that all the coordinates in the delay vector may be near zero simultaneously). Since the measurement function is projection onto the $x$ axis, $Dh$ is the row vector $(1,0,0)$. The derivative matrix $Df^{-i\tau}$ of the map associated with the Lorenz equations can be found by integrating the equations for the differentials, i.e. as is done in computing Lyapunov exponents for an ODE. For numerical stability, we are often forced to integrate forwards along an orbit segment, and we then use singular value decomposition to invert the resulting matrices. In order to visualize the distortion's $x$-dependence, we plot $\delta$ against the $x$ coordinate, as shown in Figure (7). The graph is multi-valued, since $\delta$ depends on $y$ and $z$ as well as $x$.

To illustrate the dependence of the distortion on the time lag $\tau$ used, we arbitrarily fix a state $s = (-1.8867, -5.1366, 24.7979)$, and plot $\delta$ against $\tau$. See Figure (8). We chose three different embedding dimensions as follows. The upper curve is for $m_+ = 0, m_- = 2$, and because of the low embedding dimension, there are singularities. The middle curve is for $m_+ = 0, m_- = 4$, and the singularities have vanished. Note also that as $\tau$ increases, there is very little advantage in using a higher embedding dimension. Intuitively, this is because the motion on the Lorenz attractor is chaotic, and measurements in the far past fail to give new information in the unstable direction. We will return to this topic in Section 5. Finally, the lower curve is for a non-predictive embedding with $m_+ = 5, m_- = 4$. Significant noise reduction has been achieved since future coordinates do give information in the unstable direction.

Note that in all three cases, the distortion blows up at $\tau = 0$. This is to be expected, since in this limit, measurements become redundant. In fact a general result of Section 5 implies that for this example $\delta \to \tau^{-2}$ as $\tau \to 0$. On the other hand, for this chaotic example, intuitively we should expect to see the distortion increasing as $\tau$ increases, due to irrelevancy. However, this is clearly not reflected in the numerics. In fact, the low noise limit approximation must ultimately break down as $\tau$ increases, even for small noise levels. This situation may be visualized in Figure (6). As $\tau$ increases, there will be more and more folds, and the induced

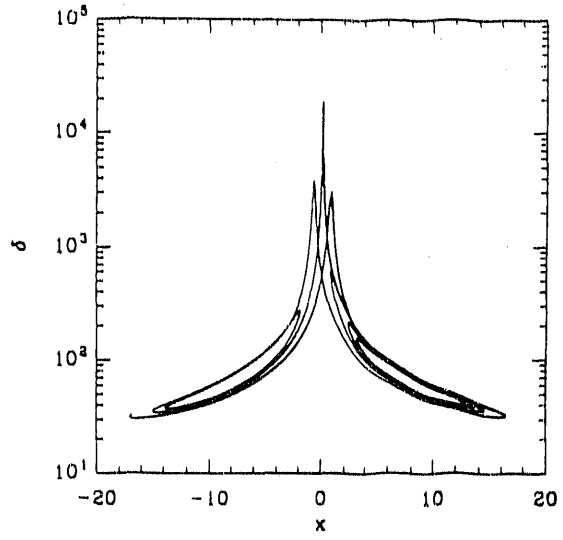


Figure 7: *Local variation of the distortion for the Lorenz equations* for a typical trajectory on the Lorenz attractor. The blowup of the distortion along $x = 0$ is a result of the poor information flow from $z$ to $x$ when $x = 0$.

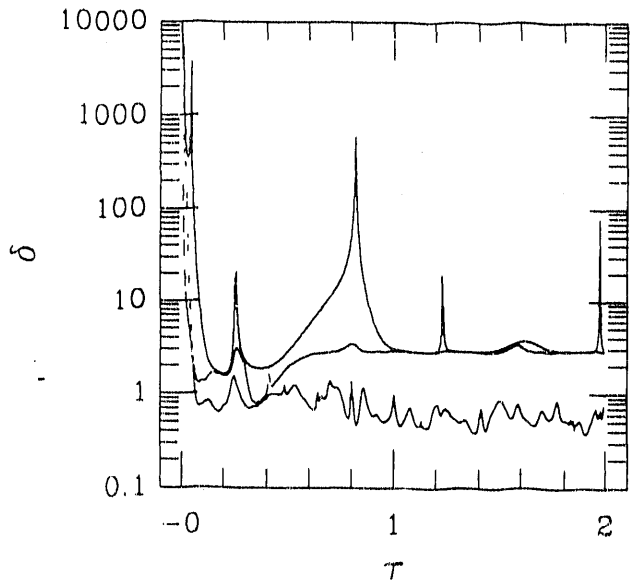


Figure 8: *The distortion as a function of* $\tau$ for three different embedding dimensions.

distribution for $p(s|\underline{x})$ will become multimodal, causing an explosion in the distortion at finite resolution. This effect cannot be obtained by the purely local analysis of Section 4.4. We return to this problem in Section 4.7. However, the above results show that it is a very complicated problem to settle on an optimal value for the time delay $\tau$.

In Section 5 we will consider the dependence of the distortion on the embedding dimension $m$ in more detail.

### 4.6.2 Finite noise distortion

In this subsection, we investigate numerically the accuracy of the low noise limit formulae above for approximating the distortion at finite resolution $\delta_\epsilon$. A similar investigation could be done for the noise amplification. Recall that the noise amplification at finite resolution measures the "thickness" of Figure 2. One could attempt to measure this thickness directly by a lengthy numerical simulation, and compare the result to the formula for noise amplification. We will now describe an algorithm for this idea in the case of distortion which we refer to as a Monte Carlo Simulation.

We use the exact likelihood function $p(\underline{x}|s)$ of Equation (15), and take the prior $p(s)$ to be the natural measure $p_\nu(s)$ on the attractor to obtain $p(s|\underline{x})$. We assume that the dynamics has sufficiently nice mixing properties so that Equation (32) holds for almost all initial conditions $s_0$ in the basin of attraction, where $\phi$ is any smooth function and $\Delta t$ is held fixed at some small value.

$$\int \phi(s) p_\nu(s) ds = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \phi(f^{i\Delta t}(s_0)) \tag{32}$$

Then taking $\phi_1(s) = \|s\|^2 p(\underline{x}|s)$ and $\phi_2(s) = s\, p(\underline{x}|s)$, we obtain Equation(33), where $w_i \equiv \exp(-\|\underline{x} - \Phi(f^{i\Delta t}(s_0))\|^2/2\epsilon^2)$.

$$\epsilon^2 \delta_\epsilon^2 = (\sum_{i=1}^{\infty} w_i \|f^{i\Delta t}(s_0)\|^2 / \sum_{i=1}^{\infty} w_i) - \|\sum_{i=1}^{\infty} w_i f^{i\Delta t}(s_0) / \sum_{i=1}^{\infty} w_i\|^2 \tag{33}$$

This is turned into a numerical approximation by truncating after $N$ terms, where $N$ is varied until satisfactory convergence has been achieved. Note that the smaller $\epsilon$ is taken, the larger $N$ must be taken for convergence. This approximation is clearly much more CPU intensive than the analytical formula of Section 4.4 for the low noise limit $\delta$.

Figure (9) illustrates the results of such a computation for the case of the Lorenz equations. The two solid lines are plots of $\delta_\epsilon$ against $\tau$ for noise levels $\epsilon = 0.5$ and $\epsilon = 0.25$, which represents a signal to noise ratio of about 20 and 40. A predictive embedding was chosen with $m_+ = 0$ and $m_- = 4$. The noisy delay vector $\underline{x}$ was generated from a state corresponding to Figure (8). The dotted line is the corresponding plot of $\delta$ against $\tau$ taken from Figure (8).

We make the following observations about this figure.

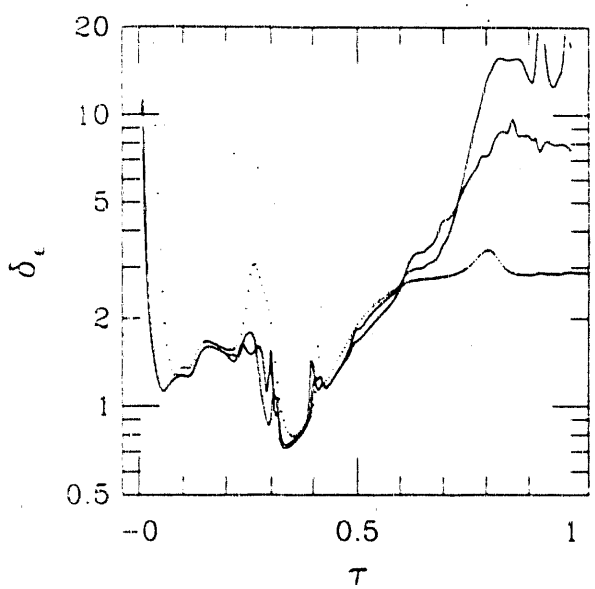


Figure 9: *The distortion $\delta_\epsilon$ at finite resolution $\epsilon$ as a function of $\tau$* for $\epsilon = 0.5$ and $\epsilon = 0.25$ are plotted in solid lines from a Monte Carlo simulation. The dotted line is the corresponding plot for the low noise limit of the distortion using the analytical formula of Section 4.4.

1. The distortion at finite resolution appears to have converged well for a value of $\epsilon$ as high as 0.5, for a wide range of values of $\tau$.

2. For the range of $\tau$ over which convergence at finite resolution has been achieved, $\delta$ provides an upper bound for $\delta_\epsilon$. This upper bound is also very sharp over a wide of values of $\tau$. We believe that the failure of $\delta$ to bound $\delta_\epsilon$ for large values of $\tau$ is due to the phenomenon of irrelevancy and bimodality mentioned in Section 4.6. We also believe that the failure of $\delta$ to bound $\delta_\epsilon$ sharply for all small values of $\tau$ is due to the that the Monte Carlo simulation for $\delta_\epsilon$ is carried out on a trajectory, effectively including the prior information of being on an attractor. In our calculation of $\Sigma$, we use a uniform prior, so that we should only expect $\delta = \sqrt{\text{trace}\Sigma}$ to provide an upper bound on $\delta_\epsilon$.

   We have performed other Monte Carlo simulations and the results indicate that the situation can be more complicated in other examples. Firstly, we performed simulations with the Ikeda map, and Gaussian noise. We observed that the limit of $\delta_\epsilon$ as $\epsilon \to 0$ sometimes fails to exist. We believe this is due to the highly fractal structure of the underlying attractor. This is not a problem at realistic noise levels for the Lorenz example, because in that case, the fractal structure is only apparent at an extremely small resolution. Secondly, we performed simulations for the Lorenz example, but using uniform noise. We observed that the limit of $\delta_\epsilon$ exists, but is dependent on the realization of the noise used to produce the delay vector $\underline{x}$. To obtain a well defined limit requires taking an ensemble average over many realizations of the noise. Fortunately, as demonstrated in

Section 4.4, this problem does not arise for Gaussian noise.

# 5 Limits to predictability

For a given noise level the noise amplification tells us how much loss of predictability occurs purely because of the reconstruction process; it therefore sets a limit to prediction that is independent of the number of data points or the modeling technique. The distortion[12] (or equivalently, the noise amplification) depends on the state space reconstruction, for example, on the parameters $m_+$, $m_-$, and $\tau$. It also depends on the properties of the underlying dynamical system such as the dimension and Lyapunov exponents, and on the measurement function. In this section we show that there are some general scaling laws that make it possible to estimate the way the distortion will change as these parameters are varied. These scaling laws set upper bounds to predictability.

To study the dependence on the reconstruction it is sufficient to consider delay coordinates. As we prove in the Section 6, this is because delay coordinates provide a lower bound on distortion, in the sense that a coordinate transformation of delay coordinates cannot reduce the distortion. One fact that is immediately apparent is that *gathering more information can only decrease the distortion.* This follows from an elementary property of conditional probabilities. 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

$$\Sigma(s|\underline{x}^{(2)}) \leq \Sigma(s|\underline{x}^{(1)}), \tag{34}$$

in the sense of Section 4. Thus, to reduce the distortion the dimension of the reconstructed space should be as high as possible.

As a practical matter, however, finite data resources usually impose a limit on the state space dimension. It is therefore important to know *which* information is most useful. For uniform lag times this translates into choosing the best values for $\tau$, $m_+$, and $m_-$. The scaling laws derived in Sections 5.1 and 5.2 provide insight into this question.

Another fact that is intuitively obvious is that when $\tau$ is sufficiently small successive measurements become almost *redundant*, in the sense that in the absence of noise they approach the same value; the difference in their value is mainly due to measurement noise. In this case images of the measurement surfaces are roughly parallel in the neighborhood of the true state. Let $t_r$ denote the *redundancy time*, above which measurement surfaces intersect at a significant angle. Then we expect that if the *window width* $w = m\tau$ is much

[12] In this section we study the distortion rather than the noise amplification because distortion does not depend on the prediction time. However, from the results of the previous section, the results will apply to either quantity.

less than $t_r$, then the distortion will be very large. To avoid this we should choose $w > t_r$. On the other hand, for a chaotic system, let $t_i$ denote the *irrelevancy time*, of order $\log \epsilon/\lambda$, where $\lambda$ is the largest Lyapunov exponent. Intuitively, it should be expected that measurements made outside a window width $w$ much greater than $t_i$ will be irrelevant, in the sense that the images of the measurement surfaces will line up along the unstable direction, and so give no information in that direction. To avoid this we should choose $w < t_i$. In the case that $t_i < t_r$, one of the above conditions on $w$ must be violated, and one would expect a very large distortion. In the this section we will investigate the extent to which the above intuition is born out quantitatively, by deriving general scaling laws for the distortion, and working out some examples.

## 5.1 Scaling laws

When $m$ is sufficiently large or $\tau$ is sufficiently small the distortion behaves according to well defined scaling laws. There are two regimes. One of these occurs when the window width $w = m\tau$ is small, and the other occurs when the window width is large.

### 5.1.1 Small window width limit

The scaling is the same whether or not the dynamics is chaotic. The scaling law is

$$m\tau \to 0 \qquad \delta = O(m^{-1/2}(m\tau)^{(1-d)}), \tag{35}$$

(where "$O()$" denotes "the order of"). Note that for $d > 1$ the distortion blows up in the limit as $\tau \to 0$, with an exponent that increases with dimension.

**Example: The Lorenz equations.** In Figure (10) we plot the distortion $\delta$ as a function of the embedding dimension $m$, with $\tau$ fixed at 0.005, and $s$ fixed at the same value as for Figure (8). A predictive and a non predictive embedding are shown. Observe that for small $m$, in both cases the scaling goes as $m^{-3/2}$, as predicted by Equation (35). At larger $m$, a different behavior is apparent, as will be discussed in Section 5.1.2. For another example, see Figure (11) of Section 5.2.

**Derivation.** Expand $D\Phi$ in a Taylor's series in time around $t = 0$. For convenience assume a predictive embedding, with the first row simply $Dh$. Then the rows of $D\Phi$ are of the form

$$D\Phi_{i+1} = a^{(0)} + a^{(1)}(i\tau) + a^{(2)}(i\tau)^2 + \ldots \tag{36}$$

where $i = 0, \ldots, m-1$, labels the row, and the $a^{(j)}$ are fixed $d$-dimensional row vectors. For sufficiently small values of $\tau$ the embedding surfaces are approximately linear, and there is a unique crossing when $m \geq d$. If we truncate the

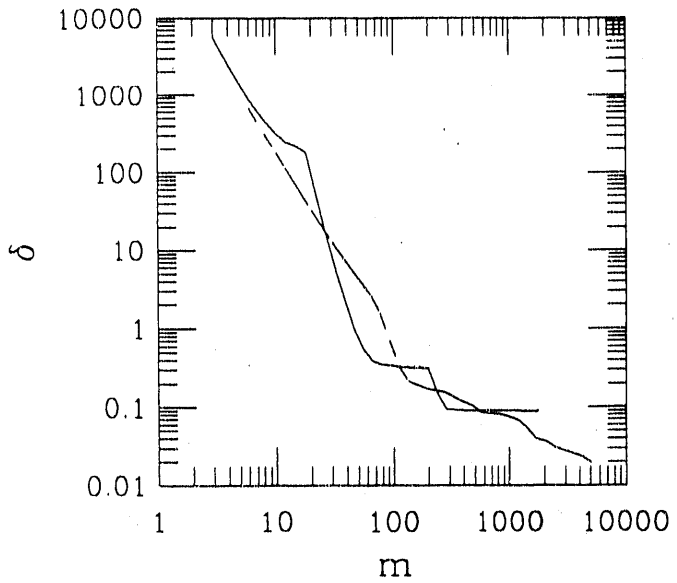


Figure 10: *The distortion as a function of m.* The solid curve is for a predictive embedding with $m_+ = 0$, and the dashed curve is for a nonpredictive embedding with $m_+ = m/2$ and $m_- = m/2 - 1$.

Taylor series at order $d-2$ the matrix cannot be of full rank, since there are only $d-1$ independent vectors $a^{(j)}$. Consequently the $d^{th}$ singular value is zero to order $\tau^{d-2}$. But if we truncate the Taylor series at order $d-1$ the matrix will generically be of full rank at almost all states $s$ because the $d$ $d$-dimensional vectors $a^{(j)}$ involved in the expansion are typically independent. Therefore the $d^{th}$ singular value is typically of order $(m\tau)^{(d-1)}$. The dominant eigenvalue of $\Sigma$ is the square of the inverse of the $d$ largest singular values of $D\Phi$, which implies the $\tau$ scaling in Equation (35).

The $m$ scaling comes from the law of large numbers. If we fix the window width (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 Equation (35).

**Remark.** In the small window limit, it can also be demonstrated that the singular vectors of $D\Phi$ converge onto Legendre polynomials for almost all states $s$. It was previously shown by us [14] that in the limit of low noise and small window width, global Broomhead and King coordinates converge onto Legendre polynomials, with the singular spectrum satisfying similar scaling laws. The ideas behind the proofs are very similar, with the observability matrix replacing the covariance matrix $C_{ij}$ defined in Section 2.1.

### 5.1.2 Large window width limit.

In the limit as $m \to \infty$ when the window width is large, there are different scaling behaviors, depending on whether or not the dynamics is chaotic.

**Nonchaotic systems.** When $\tau$ is large the measurement surfaces are no longer nearly parallel. Each measurement can be treated as independent, and according to the law of large numbers the scaling is

$$m \to \infty, \qquad \delta = O(m^{-1/2}) \tag{37}$$

In general it is intuitively clear that this also increases with $d$, since when $d$ is large the information in the time series is spread over more coordinates. In the example in Section 5.2, it is shown that $\delta = \sqrt{2}d^{3/2}m^{-1/2}$.

**Chaotic systems with predictive coordinates.** As described at the beginning of this section, for a chaotic system measurements in the distant past provide no information about the position along the unstable direction. While information is provided in the stable direction, this information is at a fine scale of resolution that is typically below instrumental error. Since the uncertainty of the position along the unstable manifold is the limiting factor, this information is irrelevant for prediction.

In the limit of sufficiently large $\tau$ and small $\lambda$ it is possible to derive a scaling law for the distortion matrix. This is possible because in this case the eigenvectors of the distortion matrix line up with the stable and unstable manifolds. The eigenvalues of the distortion matrix have three different behaviors,

$$\textbf{Unstable manifold} \quad (\lambda_i > 0) \tag{38}$$
$$m \to \infty \quad \Sigma_{ii} = O(1 - e^{-2\lambda_i d'\tau})$$
$$\textbf{Neutral manifold} \quad (\lambda_i = 0) \tag{39}$$
$$m \to \infty \quad \Sigma_{ii} = O(m^{-1})$$
$$\textbf{Stable manifold} \quad (\lambda_i < 0) \tag{40}$$
$$m \to \infty \quad \Sigma_{ii} = O(e^{2m\lambda_i\tau})$$

In the above equations we have transformed the distortion matrix $\Sigma$ to the appropriate coordinates. The distortion in the unstable manifold approaches a constant, while in the neutral manifold it goes to zero as a polynomial, and in stable manifold it goes to zero exponentially with $m$. The mean-square error in a prediction is related to the trace of the distortion, which is dominated by the largest eigenvalues. As we demonstrate in the derivation below, this approaches a constant.

**Chaotic systems with mixed coordinates.** With mixed coordinates the situation is quite different from that of predictive coordinates, since future information makes it possible to pinpoint the position along the unstable manifold

precisely, and all the eigenvalues go to zero as $m \to \infty$. The calculation of the distortion follows closely that for predictive coordinates, except that all the sums and products must be taken from $\frac{-m_+}{d'}$ to $\frac{m_-}{d'}$. The results for the stable and neutral manifolds remain essentially the same, but the unstable manifold is now dominated by the contributions from the future terms. We get instead

$$\begin{aligned} &\textbf{Unstable manifold} \quad (\lambda_i > 0) \\ m \to \infty \quad \Sigma_{ii} \quad &= \quad O(e^{-2m_+\lambda_i\tau}) \\ &\textbf{Neutral manifold} \quad (\lambda_i = 0) \\ m \to \infty \quad \Sigma_{ii} \quad &= \quad O(m^{-1}) \\ &\textbf{Stable manifold} \quad (\lambda_i < 0) \\ m \to \infty \quad \Sigma_{ii} \quad &= \quad O(e^{2m_-\lambda_i\tau}) \end{aligned} \tag{41}$$

The main difference is that all the eigenvalues $\Sigma_{ii} \to 0$ as $m \to \infty$; the eigenvalues for the stable and unstable manifold go to zero exponentially. The neutral manifold thus provides the leading order contribution to the distortion.

The above relationships are apparent in Figure (10). The relationships are also not valid when $d'\tau \gg \frac{\log\epsilon}{\lambda}$; when this assumption is violated the behavior is entirely different, as we discuss in Section 5.2.

The following derivation of Equations (38)-(40) is admittedly rather loose; to turn these it into a more rigorous statement may involve placing restrictions on quantities such as the measurement function and the nature of the dynamical system. It should probably be ommitted at a first reading.

**Derivation of Equations (38)-(40) .**

The $m$-dimensional delay vector $\underline{x}^{(m)}$ can be broken into a series of lower dimensional delay vectors rooted at different times. To derive the scaling we transport all of them to the same time and examine their joint likelihood function. Let $d'$ be the minimum dimension for which delay vectors define a global embedding, and for convenience pick $m$ so that it is an integer multiple of $d'$. Let $\underline{x}_j^{(d')}$ be the $d'$-dimensional delay vector rooted at time $-jd'\tau$, $\underline{x}_j^{(d')} = (x(-jd'\tau), \ldots, x(-(jd'+d'-1)\tau))$. Assume the measurement errors are Gaussian with variance $\epsilon^2$, and let $\underline{\xi}_j^{(d')} = ((\xi(-jd'\tau), \ldots, \xi(-(jd'+d'-1)\tau))$ be the vector of $d'$-dimensional measurement errors rooted at time $-j(d')\tau$. Let $F$ be the induced $d'$-dimensional dynamics in delay space. In the limit as $\epsilon \to 0$ the vector of measurement errors rooted at time $-jd'\tau$ transported to time 0 is $\underline{\xi}_j^{(d')} = DF^{j(d')\tau}(\underline{\xi}_j^{(d')})$. The noise p.d.f. $p(\underline{\xi}_j^{(d')})$ is an isotropic Gaussian of variance $\epsilon^2$; following a calculation similar to that of Section 4.4, to leading order in $\epsilon$

$$p(\underline{\xi}_j^{(d')}) = A \exp \frac{1}{2\epsilon^2} (\underline{\xi}_j^{(d')})^\dagger \Theta_j^{-1} \underline{\xi}_j^{(d')}) \tag{42}$$

where $A$ is a normalization constant, and $\Theta_j$ is a $d' \times d'$ dimensional matrix $\Theta_j = (DF^{jd'\tau}(\underline{x}_j^{(d')}))^\dagger DF^{jd'\tau}(\underline{x}_j^{(d')})$.

Let $\dot{\underline{x}}_j^{(d')}$ be the noise free delay vector such that $\dot{\underline{x}}_j^{(d')} - \underline{x}^{(d')}(0) = \dot{\underline{\xi}}_j^{(d')}$. The set $\{\dot{\underline{x}}_j^{(d')}\}, j = 0, 1, \ldots, \frac{m}{d'}$ contains the same information as the $m$-dimensional delay vector $\underline{x}^{(m)}$. Furthermore, $\{\dot{\underline{\xi}}_j^{(d')}\}$, $j = 0, 1, \ldots, \frac{m}{d'}$ is a collection of independent random variables. Following similar reasoning to that of Section 3.1, the above statements plus Bayes' theorem (with a uniform prior) imply

$$\begin{aligned} p(\underline{x}^{(d')}(0)|\underline{x}^{(m)}) &= Ap(\{\dot{\underline{x}}_j^{(d')}\}|\underline{x}^{(d')}(0)) \qquad (43) \\ &= p(\{\dot{\underline{\xi}}_j^{(d')}\}) \\ &= A' \prod_{j=0}^{\frac{m}{d'}-1} \exp \frac{1}{2\epsilon^2} (\dot{\underline{\xi}}_j^{(d')})^\dagger \Theta_j^{-1} \dot{\underline{\xi}}_j^{(d')} \qquad (44) \\ &= A' \exp \frac{1}{2\epsilon^2} \sum_{j=0}^{\frac{m}{d'}-1} (\dot{\underline{x}}_j^{(d')} - \underline{x}^{(d')}(0))^\dagger \Theta_j^{-1} (\dot{\underline{x}}_j^{(d')} - \underline{x}^{(d')}(0)) \end{aligned}$$

where $A$ and $A'$ are normalization constants. The distortion can obtained by expanding in a Taylor series, as in Section 4.4. Hence we obtain

$$\Sigma = \left( \sum_{j=0}^{\frac{m}{d'}-1} \theta_j^{-1} \right)^{-1} \qquad (45)$$

It follows from the definition of the Lyapunov exponents that when $d'\tau$ is sufficiently large $\Theta_j$ approaches a matrix whose eigenvalues are $e^{2j\lambda_1 d'\tau}, \ldots, e^{2j\lambda_{d'} d'\tau}$. Furthermore, for large $d'\tau$ the eigenvectors approach limiting values, independent of $j$. In this case we can evaluate Equation (45) in the basis of eigenvectors.

$$\left( \sum_{j=0}^{\frac{m}{d'}-1} \Theta_j^{-1} \right)_{ii}^{-1} = \sum_{j=0}^{\frac{m}{d'}-1} e^{-2j\lambda_i d'\tau} = \frac{1 - e^{-2m\lambda_i \tau}}{1 - e^{-2\lambda_i d'\tau}} \qquad (46)$$

When $\lambda_i > 0$ the numerator approaches 1 as $m \to \infty$, and by inverting we obtain Equation (38). When $\lambda_i < 0$ the second term in the numerator dominates and we obtain Equation (40). When $\lambda_i = 0$ the summation in the previous equation is no longer valid; however, the sum is clearly of order $m$, and we obtain Equation (39).

Note that while $\underline{x}^{(d')}(0)$ is related to $s$ by a coordinate transformation, because $\Sigma$ is not invariant under coordinate transformations the distortion is not in general the same. Nonetheless, we expect their scalings properties to be the same.

Note also that in this derivation, by taking delay vectors of dimension $d'$ we are assuming that the predictability changes very little over times $d'\tau$, i.e., $d'\tau \ll \frac{\log \epsilon}{\lambda}$. When this assumption breaks down the scaling is radically different, as we demonstrate in the example of the next section.

## 5.2 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

$$\frac{d}{dt}\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} \lambda_i & -\omega_i \\ \omega_i & \lambda_i \end{pmatrix}\begin{pmatrix} u_i \\ v_i \end{pmatrix} \qquad i = 1, .., \frac{d}{2}. \tag{47}$$

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 sometimes take $\lambda_i = \lambda =$ constant. We take the measurement function to be

$$h = \frac{2}{d}\sum_{i=1}^{\frac{d}{2}} u_i \tag{48}$$

We will consider a predictive reconstruction with $m_+ = 0$.

This example is admittedly rather contrived. The oscillators are independent, so measurements only give information about the whole system because the measurement function involves a combination of all of 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. Nonetheless, as we shall see, even this very simple example exhibits nontrivial behavior.

This system has the following analytic solution.

$$\begin{aligned} u_j(t) &= u_j(0)e^{-\lambda_j t}\cos\omega_j t \\ v_j(t) &= v_j(0)e^{-\lambda_j t}\sin\omega_j t \end{aligned} \tag{49}$$

Applying the definition of $\Phi$ and differentiating, the observability matrix can be calculated explicitly.

$$\begin{aligned} D\Phi_{i,2j-1} &= \frac{2}{d}e^{-(i-1)\lambda_j\tau}\cos(i-1)\omega_j\tau \\ D\Phi_{i,2j} &= -\frac{2}{d}e^{-(i-1)\lambda_j\tau}\sin(i-1)\omega_j\tau, \end{aligned} \tag{50}$$

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.

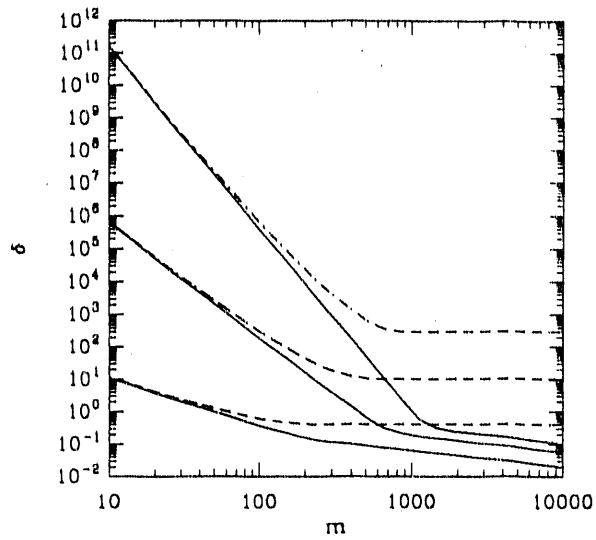


Figure 11: *The distortion $\delta$ plotted as a function of the delay coordinate dimension $m$* with a fixed delay time $\tau = 0.01$, for the system defined by Equations (47) and (48). For the solid curves $\lambda = 1$ and the system is chaotic, while for the dashed curves $\lambda = 0$ and the system is not. Three different dimensions are shown, $d = 2, 4$, and 6.

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

$$D\Phi^\dagger D\Phi|_{2i-1\ 2j-1} = \frac{4}{d^2} \sum_{k=0}^{m} e^{-k(\lambda_i+\lambda_j)\tau} \cos k\omega_i\tau \cos k\omega_j\tau. \tag{51}$$

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

In Figure (11) we plot $\delta$ as a function of $m$ for several different values of the dimension and Lypunov exponents.

This illustrates several of the features derived in Section 5.1.

- **Small w:** For small values of $m$ the window width $w$ is also small. The chaotic and nonchaotic cases behave approximately the same. As $m$ decreases the distortion increases as a power law with the predicted exponent $1/2 - d$.
- **Large w:** For the chaotic case the distortion approaches a constant while for the nonchaotic case the distortion decreases according to $m^{-1/2}$, independent of the dimension.

Note that, as it must, the distortion decreases monotonically with $m$.

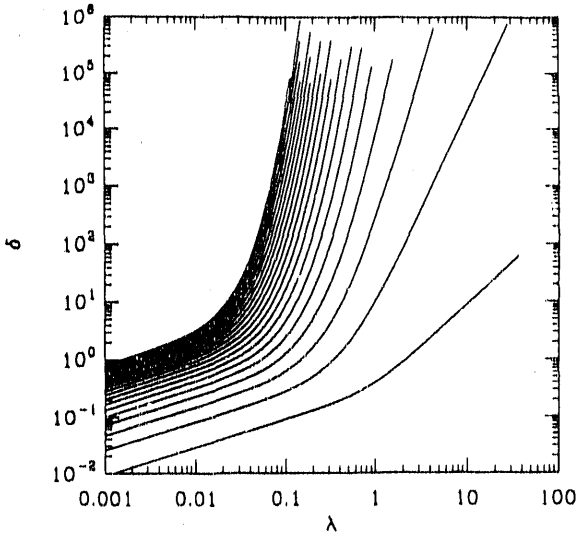


Figure 12: The distortion $\delta$ plotted as a function of the Lyapunov exponent $\lambda$ for dimensions $d = 2, 4, \ldots, 20$. The curves with the lowest distortion have the lowest dimension. Notice that there are two scaling regimes, one for low $\lambda$ and another for higher $\lambda$. In the high $\lambda$ regime the enormous noise amplification means that even a small noise level $\epsilon$ makes the system behave effectively as a random process. There is no predictability, *even for short times.*

The behavior of the plateau at $m = \infty$ for chaotic systems can be investigated by taking the limit $\tau \to 0$ in Equation (51) and approximating by an integral. This gives

$$D\Phi^{\dagger} D\Phi|_{2i-1\ 2j-1} \approx \frac{2(\lambda_i + \lambda_j)}{\tau d^2} \{[(\lambda_i+\lambda_j)^2+(\omega_i+\omega_j)^2]^{-1}+[(\lambda_i+\lambda_j)^2+(\omega_i-\omega_j)^2]^{-1}\} \tag{52}$$

The behavior of the plateau under changes in parameter values is investigated by using Equation (52) with $\lambda_i = \lambda = const$, and the frequencies uniformly spaced so that $\omega_i = \frac{2}{d}, \frac{4}{d}, \ldots, \frac{d}{d}$. The result is shown in Figure (12).

There are two scaling regimes, one for low $\lambda$, and one for high $\lambda$. In the low $\lambda$ regime the motion is effectively predictable and $\delta = O(\lambda^{1/2})$. In the high $\lambda$ limit, however, the distortion appears to diverge at a rate that increases with dimension, $\delta = O(\lambda^{d-1/2})$. Note that the crossover between the two scaling limits occurs at a lower value of $\lambda$ when the dimension increases. The scaling law for the high (respectively low) $\lambda$ limit can be obtained by substituting $m\tau = O(1/\lambda)$ in the small (respectively large) window width scaling law of Equation (35) (respectively (37)). However, there are some problems with this argument, as the prefactor of Equation (35) might depend on $\lambda$ and alter the $\lambda$ scaling behavior. Indeed we have investigated other high dimensional examples,

and the behavior can be more involved than that described above, though we often see a crossover effect.

### 5.3 When a time series becomes a random process

The divergence of the distortion in the high dimension limit seen in the previous example is particularly significant, because it demonstrates how noise amplification causes a system that is sufficiently chaotic and sufficiently high dimensional to become a random process. With a noise amplification of $10^6$, unless the noise level $\epsilon < 10^{-6}$ (a very rare occurrence), the dynamics is fundamentally unpredictable, to first order in $\epsilon$, even for short times. Note that in the previous example the distortion exceeds $10^6$ when $d > 20$ and $\lambda > 0.1$. In this case there is simply not enough information in the time series to make the motion deterministic, on any time scale. We add the caveat that when the distortion is extremely large, there may be important effects of second order in $\epsilon$ which are beyond the above local analysis.

## 6 Coordinate transformations

Up until now we have assumed that the reconstructed coordinates are simple delay coordinates, so that the reconstruction map $\Xi = \Phi$. Delay coordinates have the advantage of being simple and direct. However, the question arises of whether we can get better results by transforming to new coordinates. In general we may want to consider other coordinates $y = \Psi(\underline{x})$ where we further transform the delay coordinates so that the total reconstruction map $\Xi = \Psi \circ \Phi$.

### 6.1 Effect on noise amplification

There are two senses in which we might hope to make the coordinates "better": The first is that we might attempt to reduce the noise amplification by reducing noise, thereby locating the state more precisely. The second is that we might hope to reduce the dimension of the coordinate system, which reduces estimation error.

We will first address the question of changing the noise amplification. Two basic facts are apparent:

- *Invertible coordinate transformations cannot change the noise amplification.* This is evident from the fact that the conditional probability density $p(x(T)|\Psi(\underline{x}(0)))$ is a function of $x(T)$ alone; $\Psi(\underline{x}(0))$ 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 $\Psi$ is one-to-one, it leaves the corresponding function $p$ unchanged.

- *Non-invertible coordinate transformations cannot decrease the noise amplification.* 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

$$p(x(T)|y) = \sum_{\{\underline{x}:y=\Psi(\underline{x})\}} p(x(T)|\underline{x}) \tag{53}$$

Summing probability densities either increases the variance or leaves it unchanged. Hence, the noise amplification either increases or remains the same.

Thus, we see that we cannot decrease the noise amplification by a change of coordinates. In order to decrease the noise amplification we must alter the original information set, by changing $\Phi$. For example, we can increase the dimension of the original delay space. However, from the point of view of noise amplification a coordinate transformation on the original delay coordinates is at best neutral.

Changing coordinates can be quite useful, however, for improving the estimation problem. This is particularly true for reducing the dimensionality. The estimation problem generally becomes exponentially worse as the dimension increases. Thus, we wish to find coordinates that make the dimension as small as possible while leaving the noise amplification unchanged.

## 6.2 Local analysis

In the low noise limit, to first order in $\epsilon$ the transformation $\Psi$ can be approximated locally by its derivative $D\Psi$ (the constant term plays no role in the following). An expression for $p(s|D\Psi(\underline{x}))$ can be derived using a generalization of the argument of Section 3.3 as follows. Assuming a uniform prior, we have $p(s|D\Psi(\underline{x})) \propto p(D\Psi(\underline{x})|s)$. But $p(D\Psi(\underline{x})|s) = p(D\Psi\vec{\xi})$, where $\xi = \underline{x} - \Phi(s)$. Hence we obtain Eq. 54 by transforming the isotropic Gaussian distribution of the noise $\vec{\xi}$ through the linear map $D\Psi$.

$$p(s|D\Psi(\underline{x})) \approx A \exp \frac{-1}{2\epsilon^2}(D\Psi\underline{x} - D\Psi\Phi(s))^\dagger (D\Psi D\Psi^\dagger)^{-1}(D\Psi\underline{x} - D\Psi\Phi(s)) \tag{54}$$

As before, in the limit that $\epsilon$ is small we can expand this in a Taylor's series. The arguments parallel those leading to Equation (28), except that $\Phi$ is replaced by $\Psi \circ \Phi$. The result is that

$$p(s|y) \approx C \exp -\frac{1}{2\epsilon^2}(s - \hat{s})^\dagger \Sigma^{-1}(s - \hat{s}), \tag{55}$$

where

$$\Sigma = (D\Phi^\dagger D\Psi^\dagger (D\Psi D\Psi^\dagger)^{-1} D\Psi D\Phi)^{-1}. \tag{56}$$

Note that, as expected, a locally invertible coordinate transformation will not alter the distortion, since $(D\Psi D\Psi^\dagger)^{-1} = (D\Psi^\dagger)^{-1} D\Psi^{-1}$. In the next section we show how to minimize $\Sigma$ with respect to $\Psi$.

## 6.3 Optimal reconstruction

We will now show that 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. Local principal value decomposition provides a way of achieving this.

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

$$D\Phi = UWV^\dagger \tag{57}$$

The first of these, $V^\dagger$, is represented by an orthogonal $d \times d$ matrix that performs a rotation onto the principal axes. The second transformation $W$ is represented by a diagonal $d \times d$ matrix that stretches or contracts the principle axes; its diagonal elements $w_i$ are called the *singular values* of $D\Phi$. The third transformation $U$ is represented by a column orthogonal $m \times d$ matrix that maps onto the $m$-dimensional delay space, so that $U^\dagger U = \mathbf{1}$, the identity in $d$ dimensions.

Inserting this into Equation (56) we get

$$\Sigma = (D\Phi^\dagger D\Phi)^{-1} = (VWU^\dagger UWV^\dagger)^{-1} = VW^{-2}V^\dagger \tag{58}$$

$$\delta \equiv \sqrt{Tr\Sigma} = \sqrt{Tr(VW^{-2}V^\dagger)} = \sqrt{TrW^{-2}} = \left\{ \sum_i w_i^{-2} \right\}^{1/2}. \tag{59}$$

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:

$$V^\dagger \Sigma V = W^{-2} \tag{60}$$

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

Note that the singular values depend only on the way in which we construct the original delay coordinates (and of course on the dynamical system and the measurement function). In order to reduce the uncertainty of our coordinates

we must gather more information, for example, by increasing the dimension of the delay space or by choosing a better value of $\tau$. The search for an optimal $\tau$ is nontrivial, and increasing the dimension of the delay space worsens estimation error. Ideally, we would like a reconstruction algorithm which incorporates all the information in a given window in the lowest possible dimension.

Local principal value decomposition coordinates are the best possible coordinates which can be derived from the information set represented by a given delay embedding, in the sense that they compress all the available information into the smallest possible dimension, $d$. We define local PVD coordinates by introducing a transformation $\Psi = U^\dagger$ from the delay space of dimension $m$ to a space of dimension $d < m$ which projects onto the $d$ principal axes determined by singular value decomposition. They are local in the sense that a principal value decomposition at each point in the delay space produces a different $U$. Geometrically, the transformation $U^\dagger$ maps noisy delay vectors back onto the tangent space to the embedded state space $\Phi(\Re^d)$, and is thus a natural candidate for an optimal reconstruction.

Although the optimality of PVD coordinates is almost intuitively clear, to make sure it is understood we give a proof: First, we show that any reconstruction in fewer than $d$ dimensions has infinite distortion, then we show that the PVD coordinates have the same distortion as the delay embedding. First, suppose that the total reconstruction map $\Xi$ maps points from $d$ dimensions to $d' < d$. Performing a singular value decomposition on its transpose,

$$D\Xi^\dagger = UWV^\dagger \tag{61}$$

yields for the distortion matrix:

$$\Sigma = (UW^{-2}U^\dagger)^{-1} \tag{62}$$

But this is a $d \times d$ dimensional matrix with at most $d'$ nonzero singular values, since $W$ is $d' \times d'$. The distortion, as above, is the square root of the sum of squares of the inverses of all $d$ singular values, hence it must diverge. Obviously, we cannot embed the data in fewer than $d$ dimensions with a finite distortion. Consider instead the map from $d$ to $d$ dimensions defined by $\Xi = \Psi \circ \Phi$, where $\Psi = U^\dagger$. The distortion matrix in this case is just:

$$\begin{aligned} \Sigma &= (VWU^\dagger\Psi^\dagger\Psi UWV^\dagger)^{-1} = (VWU^\dagger UU^\dagger UWV^\dagger)^{-1} && (63) \\ &= (VW^2V^\dagger)^{-1}, && (64) \end{aligned}$$

the same as for the delay coordinates themselves. Local PVD coordinates give us a reconstruction which takes advantage of all the information available in a high dimensional embedding yet minimizes estimation error by minimizing the embedding dimension. [13]

[13] Of course, any linear transformation which can be obtained from $U^\dagger$ by a rotation will have the same property.

# 7 Conclusion

## 7.1 Results

Takens' theorem establishes that delay coordinates from a dynamical system are diffeomorphic to the system's original coordinates if their dimension is sufficiently large. The particular values of $\tau$ and $m$ are not important when the data is infinitely precise, as long as $m$ is large enough. However, in the presence of noise, the quality of an embedding is highly dependent on these parameters. The distortion and noise amplification quantify the quality of an embedding, and we give analytical formulae for them in terms of the dynamics and the measurement function in the low noise limit. The scaling laws for these quantities give fundamental bounds on the predictability of a dynamical system and show how a deterministic system becomes a random process.

Reconstruction techniques such as PVD always start with delay vectors. But since the distortion of a delay reconstruction provides a lower bound on the distortion of any further transformation, the only advantage of such transformations is dimension reduction. Local PVD is an optimal method of state space reconstruction, since it retains the distortion of its delay vectors while projecting them into the lowest possible dimension.

## 7.2 Open Questions

Although we have developed a theory of state space reconstruction in the presence of noise, in this paper we have not addressed the practical issues which arise when constructing nonlinear predictive models from time series. We are currently conducting numerical experiments in order to find out if our local principal value reconstruction technique has advantages over existing techniques for modeling. However, these experiments are subject to a number of complications which prevent their inclusion at this date. For example, forecasting error includes both noise amplification and estimation errors due to finite data sets. We have found that estimation error often dominates forecast error for quite reasonable noise levels and data set sizes. Further, estimation error varies with the reconstruction, making it difficult to distinguish the contribution of either effect to the overall error. We intend to publish the results of these experiments and an analysis of the estimation problem in **Physica D.**

**We have also left several theoretical questions unanswered as follows. We expect that our theory will need to be modified in the case of dynamical or correlated noise. Also, the example we use to illustrate the breakdown of predictability as dimension and Lyapunov exponents increase is very special. We intend to explore this phenomenon for more realistic examples.**

# 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. U.C. San Diego, 1989.

[2] J. Breiden and A. Hubler. Reconstructing equations of motion from experimental data with hidden variables. Technical report, University of Illinois, 1990.

[3] D.S. Broomhead and G.P. King. Extracting qualitative dynamics from experimental data. *Physica*, 20D:217, 1987.

[4] M. Casdagli. Nonlinear prediction of chaotic time series. *Physica D*, 35, 1989.

[5] A. Cenys and K. Pyragas. Estimation of the number of degrees of freedom from chaotic time series. Technical report, Institute of Semiconductor Physics, Academy of Sciences of the Lithuanain SSR, Vilnius 232600, 1987.

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

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

[8] J. D. Farmer and J. J. Sidorowich. Predicting chaotic time series. *Physical Review Letters*, 59(8):845–848, 1987.

[9] J.D. Farmer and J.J. Sidorowich. Exploiting chaos to predict the future and reduce noise. In Y.C. Lee, editor, *Evolution, Learning and Cognition*. World Scientific, 1988.

[10] J.D. Farmer and J.J. Sidorowich. Optimal shadowing and noise reduction. *to appear in Physica D*, 1990.

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

[12] A.M. Fraser. Reconstructing attractors from scalar time series: A comparison of singular system and redundancy criteria. *Physica D*, 34, 1989.

[13] A.M. Fraser and H.L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review*, 33A:1134–1140, 1986.

[14] J. F. Gibson, M. Casdagli, S. Eubank, and J. D. Farmer. Principal component analysis and derivatives of time series. Technical report, Los Alamos National Lab, 1990.

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

[16] S.M. Hammel. Noise reduction for chaotic systems. Naval Surface Warfare Center, Silver Spring Maryland, 1989.

[17] E.J. Kostelich and J.A. Yorke. Noise reduction in dynamical systems. *Physical Review A*, 38(3), 1988.

[18] A. Lapedes and R. Farber. Technical Report LA-UR-87-2662, Los Alamos National Laboratory, 1987.

[19] A.I. Mees. Modelling complex systems. Math Department, University of Western Australia, 1989.

[20] N.H. Packard, J.P. Crutchfield, J.D. Farmer, and R.S. Shaw. Geometry from a time series. *Physical Review Letters*, 45:712–716, 1980.

[21] M.B. Priestley. State dependent models: A general approach to nonlinear time series analysis. *Journal of Time Series Analysis*, 1:47–71, 1980.

[22] M.B. Priestley. *Spectral Analysis of Time Series.* Academic Press, 1981.

[23] T. Sauer, J. Yorke, M. Casdagli, and E. Kostelich. Embedology, 1990.

[24] R.S. Shaw. Strange attractors, chaotic behavior, and information flow. *Z. Naturforschung*, 36a:80–112, 1981.

[25] R.S. Shaw. *The Dripping Faucet as a Model Dynamical System.* Aerial Press, Santa Cruz, CA 95060, 1984.

[26] F. Takens. Detecting strange attractors in fluid turbulence. In D. Rand and L.-S. Young, editors, *Dynamical Systems and Turbulence*, Berlin, 1981. Springer-Verlag.

[27] H. Tong and K.S. Lim. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society B*, 42(3):245–292, 1980.

[28] G. U. Yule. *Philos. Trans. Roy. Soc. London A*, 226:267, 1927.

# END

# DATE FILMED

02/11/91