

Exploiting Chaos to Predict the Future and Reduce Noise

J. Doyne Farmer and John J. Sidorowich¹

*Theoretical Division and Center for Nonlinear Studies
Los Alamos National Laboratory
Los Alamos, NM 87545.*

Abstract

We discuss new approaches to forecasting, noise reduction, and the analysis of experimental data. The basic idea is to embed the data in a state space and then use straightforward numerical techniques to build a nonlinear dynamical model. We pick an *ad hoc* nonlinear representation, and fit it to the data. For higher dimensional problems we find that breaking the domain into neighborhoods using local approximation is usually better than using an arbitrary global representation. When random behavior is caused by low dimensional chaos our short term forecasts can be several orders of magnitude better than those of standard linear methods. We derive error estimates for the accuracy of approximation in terms of attractor dimension and Lyapunov exponents, the number of data points, and the extrapolation time. We demonstrate that for a given extrapolation time T iterating a short-term estimate is superior to computing an estimate for T directly.

Once we have a nonlinear dynamical model that accurately represents a data set, all the tools that were previously available only in computer experiments are extended to physical experiments. Our error estimates suggest that the use of higher order approximation techniques can give significant improvements in computing quantities such as fractal dimension or Lyapunov exponents. Furthermore, forecasting provides strong self-consistency requirements on the identification of chaotic dynamics.

We propose a nonlinear averaging scheme for separating noise from deterministic dynamics. For chaotic time series the noise reduction possible depends exponentially on the length of the time series, whereas for non-chaotic behavior it is proportional to the square root. When the equations of motion are known exactly, we can achieve noise reductions of more than ten orders of magnitude. When the equations are not known the limitation comes from prediction error, but for low dimensional systems noise reductions of several orders of magnitude are still possible.

The basic principles underlying our methods are similar to those of neural nets, but are more straightforward. For forecasting we get equivalent or better results with vastly less computer time. We suggest that these ideas can be applied to a much larger class of problems.

¹Permanent address: Physics Department, UC Santa Cruz 95064

Contents

1	Introduction	279
1.1	Chaos and randomness	279
2	Model Building	283
2.1	State space reconstruction	283
2.2	Learning nonlinear transformations	285
2.2.1	Representations	286
2.2.2	Local approximation	288
2.2.3	Trajectory segmenting	292
2.2.4	Nonstationarity	292
2.2.5	Discontinuities	293
2.2.6	Implementing local approximation on computers	293
2.2.7	An historical note	295
2.3	Comparison to statistically motivated methods	295
3	Scaling of Error Estimates	296
3.1	Dependence on number of data points	297
3.2	Dependence on extrapolation time	297
3.2.1	Higher order Lyapunov exponents	300
3.2.2	Direct forecasting	302
3.2.3	Iterative forecasting	303
3.2.4	Temporal scaling with noise	307
3.3	Continuous time	307
3.4	Numerical results	308
3.5	Is there an optimal approach?	312
4	Experimental Data Analysis	314
4.1	Computing fractal dimension: A review	314
4.2	More accurate data analysis with higher order approximation	315
4.3	Forecasting as a measure of self-consistency	317
5	Noise Reduction	317
6	Adaptive Dynamics	322
7	Conclusions	324

1 Introduction

The great promise of chaos lies in the hope that randomness might become predictable. Although chaotic dynamics puts limits on long term prediction, it implies predictability over the short term. Applications of modern nonlinear data analysis techniques indicate that chaotic dynamics is quite common, and that in many cases random behavior is due to low dimensional chaos rather than complicated dynamics involving many irreducible degrees of freedom. Until recently, however, there has been no way to exploit the presence of low dimensional chaos to actually make predictions.

In this paper we investigate straightforward but powerful approaches to this problem. We embed the data in a state space, and use simple numerical techniques to construct a nonlinear model for the dynamics. For low dimensional chaos such models can be quite accurate, allowing good forecasts and significant levels of noise reduction. They can also be used to reduce the data requirements to achieve a given level of accuracy in the computation of fractal dimension, Lyapunov exponents, or metric entropy. A good numerical model of the dynamics that can generate a data set effectively extends all the techniques that are available in a numerical experiment to physical experiments.

Most forecasting is currently done with linear methods. Linear dynamics cannot produce chaos, and linear models cannot produce good forecasts for chaotic time series. While nonlinear forecasting is an active field of investigation with a long history [33,76,64,75,63], as far as we know, the word "chaos" is not mentioned anywhere in the current forecasting literature. In this paper we re-examine forecasting problems in terms of the new paradigm that chaos offers, extending the results of a previous letter [23], where we demonstrated that local approximation can be used to make good forecasts for dynamical systems such as the Mackey-Glass differential delay equation, or chaotic convection in a H^3 - H^4 mixture [43]. We have reproduced the convection results in Figure (1). For short times our forecasts are roughly 50 times as good as those of a standard linear forecasting model.

In this paper we extend our previous results to address topics such as higher order approximation, data analysis, and noise reduction, presenting derivations of some of the scaling properties we conjectured in our previous letter. A summary of some of these results will appear soon [27].

We apologize if some of our remarks are still speculative; given the current interest in these problems [35,16,18,12,49], we have decided to report work in progress along with completed work. We hope to complete a new version of this paper shortly, adding more extensive numerical results to address some of the unresolved points.

1.1 Chaos and randomness

Chaos [53,68,17,7] has caused a fundamental change in the way we think about randomness. This influence is felt strongly in physical models for random phenomena. A good example is the problem of "excess noise" in Josephson junctions. Popular models for this phenomenon [45] are now often formulated in terms of simple sys-

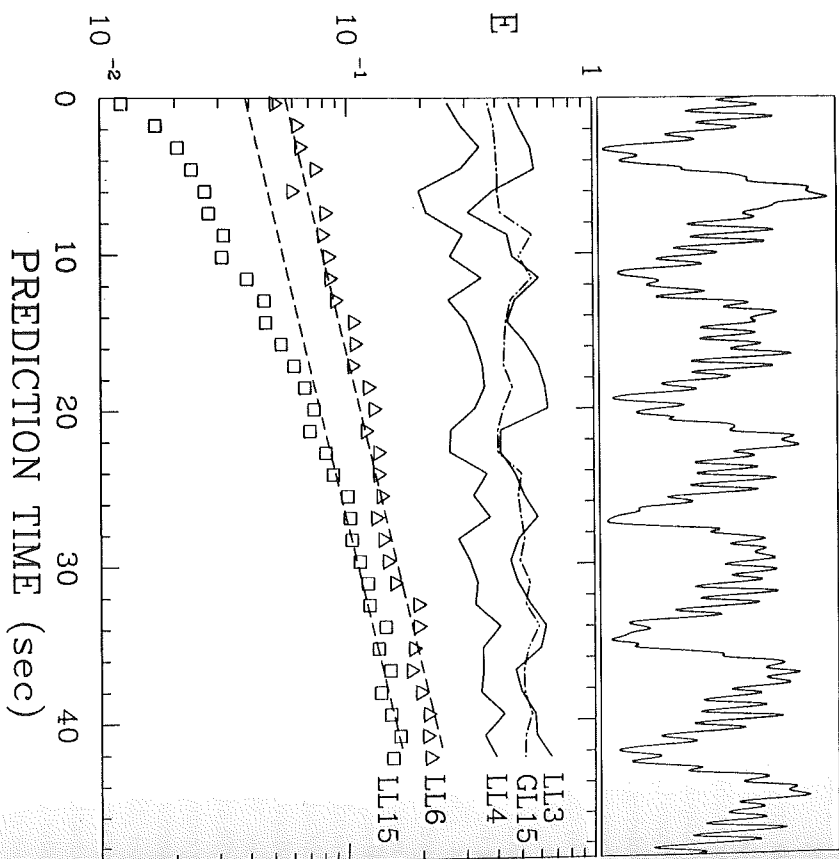


Figure 1: Top: An experimental time series obtained from Rayleigh-Bernard convection in an He^3 - He^4 mixture [43], with Rayleigh number $R/R_c = 12.24$ and dimension $D = 3.1$. Bottom: The normalized prediction error E (defined in Equation (15)) making forecasts with the Local Linear and Global Linear (linear autoregressive) methods. Numbers following the initial indicate the embedding dimension. The dashed lines are from Equation (34), using computed values of the metric entropy from reference [43]. Our predictions were based on $N = 30,000$ data points; with a sampling time $\Delta t = 0.07$ seconds, and a delay embedding time $\tau = 10\Delta t$. Based on the mean frequency in the power spectrum the characteristic time is roughly $t_c = 1.5$ seconds, so our database contains roughly 1400 characteristic times.

tems of deterministic differential equations, quite different from the statistical models that were the only option twenty years ago. Similarly, in this paper we show how thinking in terms of deterministic dynamical systems, and assuming that randomness arises out of chaos rather than complexity, leads to new approaches to forecasting and nonlinear modeling.

Until recently it was usually assumed that randomness was caused by extreme complication, i.e. the presence of many irreducible degrees of freedom. This naturally led to Kolmogorov's theory of random processes, which he defined in terms of the joint probability distribution \mathcal{P} [46]. For a time series $\{x_i\}$, the d^{th} order distribution is

$$\mathcal{P}(\xi_1, \dots, \xi_d) = \text{Probability}\{x_1 < \xi_1, \dots, x_d < \xi_d\}. \quad (1)$$

$\{x_i\}$ can represent events at discrete times, or samples of a continuous function. Random processes can also be discussed in terms of the probability density function P , defined as

$$\int_0^{\xi_1} \dots \int_0^{\xi_d} P(x_1, \dots, x_d) dx_1 \dots dx_d = \mathcal{P}(\xi_1, \dots, \xi_d). \quad (2)$$

The process is *deterministic* if there is some value of d such that the probability density approaches a delta function in the limit of perfect measurements of $\{x_i\}$.

Many people speak of random processes as though they were a fundamental *source* of randomness. This is misleading. The theory of random processes is an empirical technique for coping with inadequate information, and makes no statements about *causes* of randomness. As far as we know, the only truly fundamental source of randomness is the uncertainty principle of quantum mechanics; everything else is deterministic, at least in principle. Nonetheless, we call many phenomena such as fluid turbulence or economics random, even though they have no obvious connection to quantum mechanics. It has traditionally been assumed that the apparent randomness of these phenomena derives solely from their complication.

We will take the practical viewpoint that randomness occurs to the extent that something cannot be predicted, which usually depends on the available information. With more data or more accurate observations, a phenomenon that previously seemed random might become more predictable, and hence less random. Randomness is in the eye of the beholder.² Furthermore, randomness is a matter of degree – some systems are more predictable than others.

As originally pointed out by Poincaré [60], many of the classic examples of randomness are not complicated. The dynamics of a flipping coin or a roulette ball, for example, involve only a few degrees of freedom. Their randomness comes from *sensitive dependence on initial conditions* – a small perturbation causes a much larger effect at a later time, making prediction difficult. When sensitive dependence on

² Another perhaps more fundamental notion of randomness is due to Kolmogorov and Chaitin [14]. Whether or not chaotic systems are random in this sense is controversial [78,28].

initial conditions occurs in a sustained way it is called *chaos*³. Since chaos is defined in the context of deterministic dynamics, in some very strict sense it might be incorrect to say that chaos is random – ultimately uncertainty originates from something external to the dynamics, such as measurement error or external “noise”. But sensitive dependence exaggerates uncertainty, so that small uncertainties turn into large ones. Since chaos amplifies noise exponentially any uncertainty at all is amplified to macroscopic proportions in finite time, and short-term determinism becomes long-term randomness.

Chaos creates randomness by strongly amplifying what we don't know. Even with only a few degrees of freedom and a very small source of uncertainty, points in a chaotic time series that are far apart in time do not appear deterministic, unless the source of uncertainty is reduced to unreasonable proportions. Chaotic systems pass many classic “tests” of randomness; for example, some simple chaotic maps produce uncorrelated time series, with $\langle x_i x_{i+j} \rangle = 0$ unless $j = 0$. Furthermore, chaotic trajectories look random.

In contrast, if the dynamics are not chaotic errors grow slowly and the main requirement for determinism is that d be large enough. As long as this condition is satisfied forecasts can be made far into the future.

Dissipative dynamical systems often have the property that undisturbed trajectories approach a subset of the state space, called an *attractor*. This can cause a drastic reduction in the number of degrees of freedom. Fluid flows, for example, have an effectively infinite dimensional state space, but can have low dimensional chaotic attractors [56,9,43].

Thus, we should not distinguish chaos and randomness, but rather we should distinguish systems with low dimensional attractors from those with high dimensional attractors. If a time series is produced by motion on a very high dimensional attractor, then from a practical point of view it is impossible to gather enough information to exploit the underlying determinism. If we model the dynamics in a state space whose dimension is lower than that of the attractor, we only see a projection of the dynamics, and determinism is invisible – the dynamics look random. Even if the dimension of the model is large enough, the amount of data needed to make a good model for a high dimensional attractor may be prohibitive. This problem gets exponentially worse as the dimension increases [31,39], as is apparent in the error estimates presented in Section 3.

With many degrees of freedom, the statistical approach is probably as good as any – linear models may even be optimal. But if random behavior comes from low dimensional chaos, we can make forecasts that are much better than those of linear models. Furthermore, the resulting models can give useful diagnostic information about the nature of the underlying dynamics, aiding the search for a description in terms of first principles.

³ A trajectory is *chaotic* if it has positive Lyapunov exponents, i.e., if on average it is locally unstable. The motion of a roulette ball is sensitive to initial conditions, but strictly speaking not chaotic, since it comes to rest, it does not have any positive Lyapunov exponents. Chaos is a special case of sensitive dependence to initial conditions, where there is also sustained motion.

2 Model Building

2.1 State space reconstruction

Consider a time series $\{v(t_i)\}$, $i = 0, 1, \dots, N$. Assume that $\{v(t_i)\}$ is stationary. This is automatic if it comes from an attractor.⁴ We will assume for the moment that v is a scalar, although the extension to the case that it is a vector is straightforward. Typically $\{v(t_i)\}$ is a projection of dynamics in a higher dimensional state space, so that in order to make use of any determinism in $\{v(t_i)\}$ we must *reconstruct* a state space.

A typical example occurs in fluid flow experiments, which in principle can be modeled accurately by deterministic partial differential equations. However, in practice this may not be useful unless the data is in the correct form. For example, suppose a single probe measures a given component of the velocity at a fixed point in space. This data is simply inadequate to provide initial conditions for the Navier-Stokes equations. To build a model from the data at hand we are forced to reconstruct a state space from a single time series.

A method for doing this⁵ was introduced by Packard *et al.* [59] and put on a firm mathematical foundation by Takens [71]. Suppose we create a state vector $x(t)$ by assigning coordinates

$$\begin{aligned} x_1(t) &= v(t), \\ x_2(t) &= v(t - \tau), \\ &\vdots \\ x_d(t) &= v(t - (d-1)\tau), \end{aligned} \quad (3)$$

where τ is a delay time. If the dynamics takes place on an attractor of dimension D , then a necessary condition for determinism is $d \geq D$. If τ is the dimension of a manifold containing the attractor, Takens showed that $d = 2\tau + 1$ is sufficient, at least in principle.

In principle, τ is arbitrary as long as it is not rationally related to $x(t)$. In practice, if τ is too small the coordinates become singular, so that $x_j \approx x_{j+1}$. If τ is too big, chaos makes x_1 and x_d causally disconnected. Taken together, these two considerations imply an effective upper bound on the embedding dimension d . In practice d is often chosen by trial and error, starting with a low value and increasing it, searching for optimal results. A more systematic procedure based on mutual information has been explored by Fraser and Swinney [30].

The use of delay coordinates to reconstruct a state space is not original to dynamical systems theory. It goes at least as far back as Yule, who in 1927 made a model for sunspot activity based on a linear combination of past values [79]. This

⁴ We will sometimes use subscripts to indicate coordinates, but at other times we will use them to indicate time. We hope that the context makes this clear.

⁵ This was also suggested by David Ruelle

idea is also implicit in Kolmogorov's definition of a random process. The important contribution from dynamical systems theory is the demonstration that reconstruction preserves geometrical *invariants* of the dynamics, such as attractor dimension, metric entropy, and the positive Lyapunov exponents.

Delays are not the only way to embed data in a high dimensional space. Another example is derivatives, $x_1(t) = x(t)$, $x_2(t) = x'(t)$, ..., $x_d = x^{(d-1)}(t)$, which for clean data usually produce nicer embeddings than delays. But since differentiation amplifies noise, high dimensional embeddings are impractical [59,31].

Broomhead and King [11] have suggested an alternative approach. They apply the Karhunen and Loeve principal value decomposition to the delay coordinate representation, and produce embeddings that seem to have the nice properties of derivatives, but without the numerical problems. The simplest way to implement their procedure is to compute the covariance matrix $\langle x_i(t)x_j(t) \rangle_t$ and compute its eigenvectors and eigenvalues α_i . The eigenvalues α_i are the average root-mean-square projection of the d -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. The numerical calculations of Broomhead and King demonstrate that under good circumstances α_i falls off exponentially with i , until it reaches a floor determined by the noise level. The fall off steepens as the sampling time Δt decreases.

A nice feature of the Broomhead and King procedure is that a new global embedding dimension \tilde{d} is computed automatically, simply by counting the number of eigenvalues above the noise floor. As long as the original embedding dimension d is sufficiently large and τ is sufficiently small, \tilde{d} only depends on the *lag window* $\tau_L = \tau(d-1)$. This eliminates much of the trial and error procedure that is usually necessary to determine the dimension of the state space. If τ_L is too small, the embedding makes incomplete use of the available information, but if τ_L is too large, the determinism is overwhelmed by the amplification of noise. Our suggestion is to weight the delay coordinates before performing the principal value decomposition, according to

$$\begin{aligned} x_1(t) &= v(t), \\ x_2(t) &= e^{-h\tau} v(t-\tau), \\ &\vdots \\ x_d(t) &= e^{-h(d-1)\tau} v(t-(d-1)\tau), \end{aligned} \quad (4)$$

where h is the metric entropy. h can be calculated by a variety of different algorithms [57]. The resulting coordinates are linearly optimal from an information theoretic point of view. This technique can also be extended for assessing the relevance of spatial samples or multivariate time series by maximizing information, as will be discussed in more detail later [25]. As pointed out by Fraser [29], while the Broomhead and King procedure has some nice features, maximizing the mutual information is a much better criterion for a good embedding than diagonalizing the covariance matrix.

Up until now we have assumed that the state $x(t)$ is constructed directly from the time series. This is usually referred to as autoregressive (AR) modeling. For a deterministic dynamical system it is the most natural approach. An alternative is to make predictions in terms of the residuals, $\delta(t) = x(t) - \hat{x}(t)$, where $\hat{x}(t)$ is a prediction of $x(t)$. This approach, originally due to Slutsky [70], is called the moving average (MA) model. An alternative, called the ARMA model [63], extends the state space to include both. AR, MA, and ARMA models are formally equivalent, but they are not necessarily equivalent in practice. The ARMA approach may offer some advantages, even for modeling fully deterministic dynamics. We intend to investigate this further.

2.2 Learning nonlinear transformations

Once we have found a state space representation, the next task is to fit a model to the data. There are several approaches. The simplest is to make time discrete, and assume that the dynamics can be written as a map in the form

$$x(t+T) = f_T(x(t)) \quad (5)$$

where the current state is $x(t)$, and $x(t+T)$ is a future state. f and x are both d -dimensional vectors. The problem is to estimate $x(t+T)$. We will call this estimate $\hat{x}(t, T)$, and approximate the dynamics by a map \hat{f} of the form

$$\hat{x}(t, T) = \hat{f}_T(x(t)). \quad (6)$$

We often iterate this equation, writing

$$\hat{x}(t, T) = \hat{f}_T(\hat{x}(t-T, T)). \quad (7)$$

An alternative approach for continuous time systems is to approximate the differential,

$$\frac{d\hat{x}}{dt} = \hat{f}(\hat{x}(t)) \quad (8)$$

and integrate. In this paper we mainly discuss the discrete-time approach, since it is faster, and we have studied it more. Differential models may offer improvements in accuracy, though, and we intend to investigate this in a future paper.

In general inaccurate measurements, external disturbances, or round off errors introduce high dimensional uncertainties that are more conveniently viewed as non-deterministic. At any given time the purely deterministic state $y(t)$ may be perturbed by $n(t)$, so that the observed state is

$$x(t) = y(t) + n(t).$$

We call $n(t)$ *noise*, and generally assume that it is small, and do our best to ignore it. For observational noise $y(t)$ is an iterate of a deterministic equation, $y(t) = f^t(y_0)$, but for external disturbances this is not possible unless the trajectory is shadowed by a purely deterministic trajectory [1,8,42]. As we demonstrate in Section 5, with a good model noise can be reduced considerably.