Finding Chaos in Noisy Systems

by

Douglas Nychka†[1], Stephen Ellner[2], Daniel McCaffrey[3] and A. R. Gallant[1]

Institute of Statistics Mimeograph Series No. 1996

July 30, 1991.

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

Finding Chaos in Noisy Systems

by

Douglas Nychka†[1], Stephen Ellner[2], Daniel McCaffrey[3] and A. R. Gallant[1]

July 30, 1991.

Abstract

In the past twenty years there has been much interest in the physical and biological sciences in nonlinear dynamical systems that appear to have random, unpredictable behavior. One important parameter of a dynamic system is the dominant Lyapunov exponent (LE). When the behavior of the system is compared for two similar initial conditions, this exponent is related to the rate at which the subsequent trajectories diverge. A bounded system with a positive LE is one operational definition of chaotic behavior. Most methods for determining the LE have assumed thousands of observations generated from carefully controlled physical experiments. Less attention has been given to estimating the LE for biological and economic systems that are subjected to random perturbations and observed over a limited amount of time. Using nonparametric regression techniques (Neural Networks and Thin Plate Splines) it is possible to consistently estimate the LE. The properties of these methods have been studied using simulated data and are applied to a biological time series: marten fur returns for the Hudson Bay Company (1820-1900). Based on a nonparametric analysis there is little evidence for low-dimensional chaos in these data. Although these methods appear to work well for systems perturbed by small amounts of noise, finding chaos in a system with a significant stochastic component may be difficult.

Keywords: Lyapunov Exponent, Nonparametric Regression, Thin Plate Splines, Cross-validation, Neural Networks, Dynamical Systems, Marten Fur Returns

[1] Department of Statistics, North Carolina State University

[2] Biomathematics Graduate Program, North Carolina State University

[3] The Rand Corporation

† Address for correspondence: Department of Statistics Box 8203, North Carolina State University,
Raleigh, NC, 27695-8203

1. Introduction

In the past twenty years much interest has been generated in the physical and biological sciences by deterministic mathematical systems that appear to have random, unpredictable behavior. These kind of phenomena, spanning a diverse range of fields, have been collected under the common heading of chaos. Models for chaotic behavior are important because they suggest a parsimonious representation for systems following seemingly complex behavior. Also, the chaotic nature of a system puts limits on the predictability of the future behavior from past history. These limits are present even in the absence of any random components. This article discusses the statistical analysis of dynamic systems based on estimating the dominant Lyapunov exponent, λ. When the behavior of the system is compared for two similar initial conditions, λ is related to the rate at which the subsequent trajectories diverge. A bounded system with $\lambda > 0$ is one operational definition of chaotic behavior. Data analytical methods developed over the last decade in theoretical physics (Schuster 1988, Mayer-Kress 1986) have concentrated on very large data sets generated from carefully controlled physical experiments. Less attention has been given to estimating λ for systems subjected to random perturbations and observed over a limited amount of time. These constraints are relevant for many biological and economic systems and thus we are interested in the feasibility of statistical methods when the dynamical information is limited by sample size and masked by noise.

Traditionally chaos has referred only to purely deterministic systems and has been considered a distinct alternative to stochastic modeling (Farmer and Sidorwich 1988). Ruelle (1989), however, defines a system to be chaotic if it exhibits sensitive dependence on initial conditions for all initial conditions. Such sensitivity distinguishes chaotic systems from non-chaotic ones. We have found this general definition useful because particularly in ecological or epidemiologic systems there is no *a priori* evidence to suggest a strictly deterministic model. By focusing on the dominant Lyapunov exponent we are able to fit dynamic models to time series and to estimate the degree to which f is chaotic, without the presupposition that the system is deterministic.

Given a times series we propose to estimate the Lyapunov exponent using nonparametric regression. We assume that the data $\{x_t\}$ are generated by a nonlinear autoregressive model

(1.1a) $$x_t = f(x_{t-1}, x_{t-2}, \cdots x_{t-d}) + e_t, \qquad 1 \leq t \leq N,$$

or more generally

(1.1b) $$x_t = f(x_{t-L}, x_{t-2L}, \cdots x_{t-dL}) + e_t, \qquad 1 \leq t \leq N.$$

Here $x_t \in \mathbb{R}$, f is a smooth, unknown function and $\{e_t\}$ are a sequence of independent random variables with $E(e_t) = 0$ and $Var(e_t)=\sigma^2$.

An autoregressive model for chaotic data may be motivated by Takens's Theorem from dynamical systems theory: a deterministic chaotic system $U(t) = (u_1(t), u_2(t), \cdots u_m(t))$ on an attractor with dimension $D < \infty$ generically satisfies an equation of the form

$$x(t) = f(x(t\text{-}L), x(t\text{-}2L), x(t\text{-}dL))$$

for any $d > 2D+1$ and $L > 0$, where x is any one of the variables $u_1, u_2, \cdots, u_m$ (see Eckmann and Ruelle 1985 for a precise statement of the theorem). This result is important because it suggests that time lags of a single variable can serve as surrogates for the unobserved variables of the system. Data analyses based on this result include the widely-used method of "attractor reconstruction in time-delay co-ordinates" (Schuster 1988). Thus our basic model (1.1) is a generalization of attractor reconstruction to allow for random perturbations. Under the broader definition of chaos cited above, systems like (1.1) may be chaotic.

In Section 2 we review the properties of Lyapunov exponents and compare this measure of a dynamic system to the dimension of the attracting set. Section 3 describes two nonparametric regression estimates of f in (1.1) and these estimates are used to derive estimates of λ. Section 4 evaluates the performance of these methods for simulated data and Section 5 compares these methods on a short biological time series: marten fur returns for Northern Canada from 1820-1900. The last section discusses these results from the simulations and the data analysis.

2. Quantifying Dynamical Properties of a System.

In order to follow a system in time it is useful to think in terms of a map acting on a state vector. Let $X_t^T = (x_t, \cdot\cdot\cdot, x_{t-d+1})$, $E_t^T = (e_t, 0, \cdot\cdot\cdot, 0)$ and define the map function F: $\mathbb{R}^d \rightarrow \mathbb{R}^d$ such that

$$X_t = F(X_{t-1}) + E_t \tag{2.1}$$

is an equivalent model to (1.1a). This form makes it clearer how the system evolves over time and the sequence of state vectors, $\{X_t\}$ $1 \leq t \leq N$, will be referred to as the system's trajectory. One basic feature of a chaotic system is that for arbitrarily close state vectors the resulting trajectories will diverge at an exponential rate. In order to illustrate this phenomenon it is first necessary to discuss the set traced out by a trajectory.

2.1 Attracting Sets.

For a deterministic dynamic system given by (2.1) with $E_t \equiv 0$, let $\mathcal{A}$ denote a set with the following properties:

1) If $X \in \mathcal{A}$ then $F(X) \in \mathcal{A}$

2) If X_0 is sufficiently close to $\mathcal{A}$ then distance$(X_t, \mathcal{A}) \rightarrow 0$ as $t \rightarrow \infty$.

Because of the second property, $\mathcal{A}$ is called an attractor: trajectories starting near the attractor converge onto it and their subsequent motion is coi.fined to the attractor. (However, it is difficult to give a formal mathematical definition of attractors that can apply in all cases, and less restrictive definitions are often used; see Guckenheimer and Holmes 1983, Chapter 5). For deterministic chaotic systems, the attractor will often be a complicated set with a fractional dimension. Due to this correspondence, the identification of chaos in observational data may focus on dimension estimates based on the set of observed state vectors. One problem with this approach is that dimension estimates are sensitive to the amount of noise in the system; this difficulty will be illustrated by considering a simple system without and with noise.

Figure 1 is an example of the attractor for a simple deterministic system that will be referred to as the cosine map (a relative of the Hénon map):

$$x_t = \cos(2.8\, x_{t-1}) + .3x_{t-2}$$

Note that in the vector formulation the state vectors are only two dimensional and thus it is easy to depict the trajectories of this system. We see that the attractor is a bounded set with complicated structure; using a series of 2500 values the correlation dimension was estimated to be approximately 1.2. Figure 2 is a similar plot of the attractor when a random component is added to the system as in equation (1.1). In this case $\{e_t\}$ are independent $N\left(0,\, (.2)^2\right)$ random variables. These two figures indicate that the attractor changes dramatically when noise is added. The deterministic system yields an unusual set with fractional dimension and Lebesgue measure zero. Due to the blurring by the random component, in the mixed system the attractor is a two dimensional set. Consequently, using dimension estimates to identify a chaotic element may be ambiguous. Correlation dimension estimates for moderate distance scales may be fractional but will tend toward 2 as the distance scale decreases. The reader is referred to Smith (1991) for a discussion of correlation dimension estimates and modifications to adjust for noise.

2.2 Sensitive Dependence on Initial Conditions.

The attractor is a static object and the dimension of this set does not directly quantify the dynamic behavior of the trajectories. An alternative to studying the attractor is to consider the evolution of trajectories. Figure 3 and Figure 4 are panels of plots that follow 500 points with similar initial conditions through 40 iterations of the cosine map. The second set (Figure 4) include a random component, with the same realization, $\{e_t\}$, of random shocks being used for all the trajectories. In either case, although the initial state vectors are distributed in a small circle, subsequent iterations of the map rapidly distribute these points uniformly over the attracting set. By 20 iterations the positions of these state vectors carry little information concerning their common origin and are essentially independent. Note that the deterministic system tends to map these points into the characteristic pattern of the attractor (c.f. Figure 1). The trajectories associated with the noisy system do not appear to converge to a fixed set because a new random shock is added with each iteration. However, the pattern of these points is suggestive of the stable distribution depicted by the attracting set for the noisy system in Figure 2.

Both cases exhibit a sensitivity to initial conditions, with trajectories initially diverging at an exponential rate. Let Ω_0 denote the cohort of initial state vectors in these examples and let Ω_t denote the set of state vectors after t iterations of the system. The size of these sets can be quantified by the mean pairwise distance (MPD) between points. Figure 5 is a plot of the log MPD of Ω_t as a function of t for the deterministic system and 5 realizations (5 different sequences of random shocks) for the cosine map. In the deterministic case there is an initial decrease in the size of Ω_t as trajectories are absorbed onto the attractor, but on this set there is "sensitive dependence on initial conditions" and the size of Ω_t then grows exponentially. The expansion stops abruptly as the size of Ω_t approaches the size of the attractor and at this point the action of the map is to fold Ω_t back onto the attractor. The size of Ω_t for the noisy system exhibits the same dependence on t: an exponential increase that plateaus when the set expands to the same size as the attractor. Because these trajectories depend on a random component, there is some variation between them in the growth of Ω_t. However, once trajectories settle onto the attractor there is relatively little variation in the slope during the exponential growth phase.

2.3 Lyapunov Exponents for Stochastic Systems.

The example given above suggests that sensitive dependence on initial conditions is a common feature of both deterministic and noisy chaotic systems. More specifically the slope (or average slope) for the curves in Figure 5 may be a useful measure for the degree of chaos. The Lyapunov exponent is essentially the limiting slope as the initial state vectors are confined to an infinitesimally small neighborhood.

Let X_0^A and $X_0^B \in \mathbb{R}^d$ denote two initial state vectors such that $X_0^A - X_0^B = U\delta$ where U is a fixed vector of unit length. After one iteration of (2.1) *with the same random shock* we have

$$\left\| X_1^A - X_1^B \right\| = \left\| F(X_0^A) - F(X_0^B) \right\|$$

$$= \left\| DF(X_0^A)\ (X_0^A - X_0^B) \right\| + o(\delta)$$

where DF is the d x d Jacobian matrix of partial derivatives for the map F. Now assume that D^2F is uniformly bounded for all X in the attractor, set $J_t = DF(X_t^A)$ and $T_M = J_M \bullet J_{M-1} \cdots \bullet J_1$.. By application of the chain rule for differentiation : $(d/d\delta)\ (X_M^A - X_M^B) = T_M U$ it is possible to show that

$$\left\| X_M^A - X_M^B \right\| = \left\| T_M U \right\| \delta + o(\delta)$$

under the circumstances that $M \to \infty$, $\delta \to 0$ such that $K^M \delta \to 0$ where K is an upper bound based on the the first and second order derivatives of F. Under such limiting conditions, T_M gives a linear approximation to the action of iterating F, M steps. Indeed, the Lyapunov exponent is related to the largest singular value of T_M.

Let $\nu_1(M)$ denote the largest eigenvalue of $T_M^T T_M$. The formal definition of the Lyapunov exponent is

$$\lambda \overset{\text{def}}{=} \lim_{M \to \infty} \frac{1}{2M} \log \left| \nu_1(M) \right|. \tag{2.2}$$

At this point it is far from clear whether such a limit exists, especially for a noisy system. This definition can be made rigorous, however, and some necessary conditions on the system are given at the end of this section. It should be noted, that if λ exists and is independent of X_0^A,

$$\left\| X_M^A - X_M^B \right\| \leq e^{\lambda M} \delta + o(\delta)$$

provided $\delta \to 0$ and $M \to \infty$ and $\delta K^M \to 0$. Also, if we let $\mathcal{T} = \lim_{M \to \infty} (T_M^T T_M)^{1/2M}$, then λ corresponds to the largest eigenvalue of $\mathcal{T}$. The inequality given above will actually be attained if $X_0^A - X_0^B$ is not orthogonal to the eigenvector of $\mathcal{T}$ corresponding to λ. Although λ will be a constant, for a system with a random component the limiting matrix $\mathcal{T}$ will be random and will depend on the particular sequence of errors $\{e_t\}$. Thus the eigenvectors of $\mathcal{T}$ will vary from one realization of the system to another even though the eigenvalues are constant (Eckmann and Ruelle 1985).

One can actually define a vector of Lyapunov exponents associated with the log eigenvalues of $\mathcal{T}$. Positive exponents correspond to directions where the action of F causes trajectories to diverge while negative exponents identify directions of contraction. Moreover the positive exponents may be used to estimate the dimension of the attractor *via* the Kaplan-York conjecture (see Abarbanel, 1991). We have chosen to concentrate on the dominant exponent because it is much easier to estimate and by itself provides evidence for chaotic dynamics. There remains some controversy as to the feasibility of estimating exponents other than λ even in the case of deterministic systems.

2.4 Existence of λ.

The existence of the limit in (2.2) is based on the system (1.1) generating a time series that is ergodic. The conditions leading to a rigorous justification of (2.2) will now be developed.

0) $X_t = F(X_{t-1}) + E_t$ where $\{E_t\}$ are iid random vectors.

1 a) There is an invariant set, $\mathcal{A}$, and a unique Borel probability measure, μ, on $(\mathbb{R}^d, \mathfrak{B})$ such that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} I_B(X_t) = \mu(B)$$

for all $B \in \mathfrak{B}$ and $X_0 \in \mathcal{A}$, where $I_B(x)$ is the indicator function for the set B ($I_B(x) = 1$ if $x \in B$, 0 otherwise).

1 b) X_0 is randomly sampled from the distribution μ.

2) $$\int_{\mathcal{A}} \max\left(\log \| DF(X) \|, 0\right) \, d\mu(X) < +\infty$$

Under these conditions it follows that (2.2) will exist almost surely and be a constant. Condition 0 implies that the trajectory constitutes a Markov process. This structure is important because it allows us to characterize the joint distribution of the times series by the stationary distribution of the process. Condition 1 is the requirement that X_t be a stationary, ergodic process. In this manner a single realization of the process is rich enough to reproduce integration with respect to the stationary measure. Although the limiting properties at (2.2) depend directly on the matrix process $\{J_t\} = DF(X_t)$, this process will also be ergodic and stationary due to Condition 1. The independence of λ to the initial state vector can be tied to the requirements of ergodicity. It is possible to use the work of Kifer (1986) to relax Condition 1b at the expense of more stringent conditions on F. Since stationarity is often a basic requirement of a times series model we consider Condition 1b a reasonable assumption.

Although the Lyapunov exponent is almost surely a constant, the convergence to this limit may be slow. The practical implication is that the eigenvalues of $(T_M^T \, T_M)^{1/2M}$ may be quite variable for small M and are best considered as random quantities. Some examples of this variability are given in Section 4 (e.g. Figure 8). Also, because this mathematical structure rests heavily on ergodic properties, the stationarity of $\{X_t\}$ is crucial.

3. Estimates of the Map and the Lyapunov Exponent

The approaches for estimating λ can be classified into two groups: Direct methods and Jacobian methods . Direct methods seek to find similar pairs of state vectors within the observed series and estimate how the subsequent trajectories diverge (Guckenheimer 1982, Wolf et al. 1985). Besides requiring long data series, the procedure is sensitive to noise. For noisy systems, the trajectories being compared will not have the same sequence of random shocks. Thus the divergence between them might simply be due to the random component.

Jacobian methods generate estimates of λ through the intermediate step of estimating the individual Jacobian matrices. Let $\widehat{J}_t$ denote the estimate of J_t and $\widehat{T}_M = \widehat{J}_M \cdots \widehat{J}_1$. The estimate of the Lyapunov exponent is then $(1/2M) \log|\widehat{\nu}_1(M)|$, where $\widehat{\nu}_1(M)$ is the largest eigenvalue of $(\widehat{T}_M^T \, \widehat{T}_M)^{1/2M}$. (Note however that for $L > 1$ (1.1b) is a system on $\mathbb{R}^{dL}$ having dL Lyapunov exponents consisting of d groups of L identical exponents; thus we estimate λ by averaging the d largest exponents.) This estimate depends intrinsically on the limiting relationship (2.2) for consistency. Early work on this problem used a local linear regression procedure to estimate the Jacobian matrices (Eckmann et al. 1986). We have improved on this approach by introducing more sophisticated function estimation techniques (M^{c}Caffrey 1991, M^{c}Caffrey et al. 1991, Ellner et al. 1991) and data based methods for smoothing parameter selection.

A basic theoretical question is the relationship between the consistency of the Jacobian estimates and the estimated Lyapunov exponent. M^{c}Caffrey et al. (1991) give a consistency proof and conjecture on the rate of convergence. Abstracting the main conjecture of this work,

$$\textit{If} \quad \sup\left|\widehat{J}_t - J_t\right| = \dot{\mathcal{O}}_p(\beta_N) \ \textit{for some } \beta_N \rightarrow 0 \ \textit{and} \ \beta_N M \rightarrow 0 \ \textit{then} \ (\widehat{\lambda} - \lambda) = \mathcal{O}_p(\beta_N) \ \textit{as} \ N, M \rightarrow \infty.$$

Two important aspects are highlighted by this conjecture. The convergence rate for $\widehat{\lambda}$ is directly related to the convergence rate on estimates of derivatives of the map. Also, the number of data points may need to be larger than the number of terms in the matrix product.

When a random component is present in (1.1) and a nonparametric regression estimate is used to estimate the Jacobian matrices, $\beta_N \sim N^{-\delta}$ with $\delta < \frac{1}{2}$. Moreover, $\delta \sim \frac{1}{d}$ as d, the number of lags in the model (1.1), increases. Thus, for a given sample size increasing the number of lags may have a drastic effect on reducing the accuracy of $\widehat{\lambda}$. A similar phenomenon can be observed in dimension estimates as the embedding dimension is increased. This problem, described as "the curse of dimensionality", is not limited to the analysis of dynamic systems but is a general feature of multivariate data. One way to avoid this problem is to formulate map estimates that represent the full multivariate function using lower dimensional surfaces. Of course this strategy will only work provided that the true map is well approximated by a set of lower dimensional functions.

Because of the typical convergence rate expected for β_N, the condition $M\beta_N \rightarrow 0$ can only be satisfied if M is asymptotically negligible relative to N. This suggests that in practice one might estimate the partial derivatives of f based on the full data but only consider products of the matrices over blocks of size $M < N$. The estimates for the individual blocks (there will be roughly N/M of them) could then be averaged to yield an overall estimate. While theory suggests $M < N$ our experience with simulation experiments has suggested that M be taken as large as possible (i.e. $M = N$).

Moving away from the general discussion of estimates for λ the remaining part of this section will focus on two nonparametric regression estimates of f. It should be kept in mind that these function estimates are considered with the intent of differentiating the estimated map and constructing estimates of the Jacobian matrices.

3.2 Thin Plate Splines

A fundamental definition of a spline is as the solution to a minimization problem. Although one-dimensional splines are most widely known as piecewise polynomial curves, this characterization does not extend easily to two or more dimensions. Consider data of the form

$$y_t = f(X_t) + e_t \tag{3.1}$$

where $X_t \in \mathbb{R}^d$, $\{e_t\}$ follow the assumptions of (1.1) and f has square integrable partial derivatives up to degree m. (For estimating f in (1.1), note that $y_t \equiv x_t$ and $X_t \equiv (x_{t\text{-}L}, x_{t\text{-}2L}, \cdots, x_{t\text{-}dL})$.) Let

$$\mathcal{L}(h) = \frac{1}{N} \sum_{t=1}^{N} \left(y_t - h(X_t)\right)^2 + \rho \mathfrak{J}_{m,\, d}(h)$$

where $\rho > 0$ and

$$\mathfrak{J}_{m,\, d}(h) = \sum_{\alpha_1 + \ldots + \alpha_d = m} \binom{m}{\alpha_1 \ldots \alpha_d} \int_{\mathbb{R}^d} \left[\frac{\partial^m}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} h(X) \right]^2 dX.$$

The m^{th} order thin plate spline estimate of f is the function that minimizes $\mathcal{L}(h)$ over all h such that $\mathfrak{J}_{m,\, d}(h) < \infty$ (see Wahba 1990). $\mathfrak{J}_{m,\, d}(h)$ is a general (rotation invariant) measure of roughness in the function h and by varying the value of ρ one can control the resulting smoothness of the estimated map.

Let $\hat{f}_\rho$ denote the thin plate spline estimate of f to emphasize its dependence on this smoothing parameter. Large values of ρ will constrain the estimate to be very smooth and gradual at the expense of not fitting the observed data closely. This choice for ρ would be advantageous when the errors have a large variance or f has simple structure. In the limit as $\rho \rightarrow \infty$, $\hat{f}_\rho$ will converge to a polynomial of degree $m - 1$ where the coefficients are determined by ordinary least squares regression. Small values of ρ would typically arise from fitting data with little or no noise. As $\rho \rightarrow 0$ the resulting estimate will interpolate the observed data, but will remain a smooth function (relative to the criterion $\mathfrak{J}_{d,\, m}$) in between data points. Some asymptotic properties of thin plate splines are given by Cox (1984).

Although spline functions are defined abstractly as the solution to a minimization problem, they are readily computable up to a sample size of several hundred (Bates et al. 1987). The solution will be a linear combination of the $\binom{d+m-1}{d}$ monomials up to degree $m - 1$ and N radial basis functions. Moreover, the coefficients in this linear combination are linear functions of $Y = (y_1, \ldots, y_N)^T$. For this reason there exists an $N \times N$ matrix $A(\rho)$ such $\left(\hat{f}_\rho(X_1), \cdots \cdot \hat{f}_\rho(X_N)\right)^T = A(\rho)Y$. Here $A(\rho)$ depends on ρ, m, d and $\{X_t\}$ but not on Y. If A was a projection matrix then the trace of $A(\rho)$, would give the number of parameters in the representation of the function. For a smoothing spline, $A^2 \leq A$, however, it is s'ill reasonable to use *trace* $A(\rho)$ as a measure of the effective number of parameters in the estimate. This also suggests identifying $N - \mathit{trace}\ A(\rho)$ with the degrees of freedom for the residual vector and an estimate of σ^2 based on this correspondence is

$$\hat{\sigma}^2 = \left\|[I - A(\rho)]Y\right\|^2 / \left(N - \mathit{trace}\ A(\rho)\right). \tag{3.2}$$

Some asymptotic properties of $\hat{\sigma}^2$ are given in Nychka (1990).

From the discussion above it should be clear that the accuracy of $\hat{f}_\rho$ may depend greatly on the choice of ρ. Although the smoothing parameter is often chosen subjectively, it is also useful to have a data based procedure for determining ρ. Perhaps the most common procedure is generalized cross-validation (GCV). Here $\hat{\rho}$ is taken to be the value that minimizes

$$V(\rho) = \frac{1}{N} \left\|\left(I - A(\rho)\right)Y\right\|^2 / \left(1 - \frac{\mathit{trace}\ A(\rho)}{N}\right)^2$$

One motivation for using GCV is that $\hat{\rho}$ will tend to minimize the expected average squared error (EASE): $\mathrm{EASE}(\rho) = (1/N) \sum_{t=1}^{N} E\left(\hat{f}_\rho(X_t) - f(X_t)\right)^2$. Although GCV tends to give good estimates for ρ on the average, in a small fraction of cases GCV may drastically under smooth ($\rho \approx 0$) noisy data (Nychka 1991). In these cases $\hat{f}_{\hat{\rho}}$ will yield poor estimates of the Lyapunov exponent. To avoid this problem, a modification of GCV was considered that can give added weight to larger values of ρ.

$$V_C(\rho) = \frac{1}{N} \left\|\left(I - A(\rho)\right)Y\right\|^2 / \left(1 - \frac{C\ \mathit{trace}\ A(\rho)}{N}\right)^2$$

This criterion is the same as usual cross-validation except for the addition of the constant C in the denominator. Setting $C = 2$ creates a pole at *trace* $A(\rho) = N/2$ and thus constrains the effective number of parameters to always be less than half the number of observations. Although this modified form does not provide estimates that minimize EASE, we conjecture that $\hat{\rho}$ will be related to minimizing a weighted sum of the bias and variance components. For $C > 1$ more weight will be given to the average variance than to the average squared bias.

3.3 Nonparametric Regression with Neural Network Models

"Neural networks" are a class of nonlinear models inspired by the neural architecture of the brain. The study of neural networks (or equivalently neural nets) began with McCulloch and Pitts' (1943) analysis of the logical computations that could be performed by appropriately configured networks of simple input-output devices modeling individual neurons. The growth of interest in neural nets stems from their recently discovered ability to perform some computational tasks that are more difficult to handle with standard algorithmic approaches, such as pattern recognition and classification based on imperfect data (see e.g. Caudill and Butler 1987, Lapedes and Farber 1987, IEEE 1988, IJCNN 1989). In addition, neural nets are capable of approximating arbitrary continuous maps on finite dimensional spaces, which allows their use in nonlinear regression.

Here we consider only one type of network -- feedforward single hidden layer networks with a single output (Rumelhart, Hinton and Williams, 1986) -- which has been the predominant model in statistical research on neural nets. The network structure is illustrated in Figure 6. The input values x_1 and x_2 are received by the two *input units*, which simply pass the input forward to the *hidden units* u_i. Each connection (indicated by an arrow) performs a linear transformation determined by the *connection strength* γ_{ij}, so the total input for hidden unit u_i is $\sum_{j=1}^{d} \gamma_{ij} x_j$. Each unit performs a nonlinear transformation on its total input, producing output

$$o_i = \psi\Big(\sum_{j=1}^{d} \gamma_{ij} x_j + \gamma_{i0}\Big).$$

The *activation function* ψ is the same for all units, but each unit may have its own bias γ_{i0} representing an external input or the neuron's intrinsic activity level. Typically ψ is a sigmoid function with limiting values 0 and 1 as $x \longrightarrow -\infty$ and $+\infty$ respectively; in our work we used $\psi(x) = x(1+|x/2|)/(2+|x|+x^2/2)$. The hidden layer outputs o_i are passed along to the single *output unit*, which performs an affine transformation on its total input. The network output O can therefore be represented as

$$O = \beta_0 + \sum_{i=1}^{q} \beta_i \psi(X'\gamma_i + \gamma_{i0}) \tag{3.3}$$

for d inputs and q units in the hidden layer, where X is the vector of inputs and $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \cdots \gamma_{id})$.

Like projection pursuit regression (Friedman and Stuetzle, 1981), neural network models estimate a multi-dimensional map by a sum of univariate functions of projections of the multi-dimensional vectors. However unlike projection pursuit, the univariate functions are not adaptive like kernels or splines, but rather are parametric functions selected *a priori*. Thus the neural net model resembles multi-dimensional series expansions such as Fourier series expansions. In fact, Gallant and White (1988) showed that for any unknown function on a bounded domain, an appropriately configured network with a piecewise trigonometric activation function yields a multivariate Fourier series approximation of the function. Hence networks with sufficiently many hidden units inherit the ability of Fourier series to approximate square-integrable functions to any specified degree of accuracy.

The utility of neural net models as empirical tools is not ensured by these approximation results. For regression applications, given a data set of inputs and their associated outputs it must be possible to choose parameters $\theta = (\beta_i, \gamma_{ij})$ that give an accurate estimate of the unknown function. Suppose we are given a sample of size N generated by $y_t = f(X_t) + \epsilon_t$, $1 \leq t \leq N$, where the ϵ_t are *iid* random errors. Let f_N be estimated by least-squares based on the model at (3.3) and subject to the constraints $\sum_{i=1}^{q} | \beta_i | < \Delta$ and $\sum_{i=1}^{q} \sum_{j=0}^{d} | \gamma_{ij} | < \Delta q$. If Δ and q are appropriately chosen functions of N which prohibit network complexity from growing too rapidly, then $f_N \longrightarrow f$ in probability and one can also estimate consistently functionals of f, including its partial derivatives if f is sufficiently smooth (White 1989, Gallant and White 1989). McCaffrey (1991) and Barron (1991) extend these results by giving rates of convergence.

In previous work on estimating Lyapunov exponents (McCaffrey et al. 1991, McCaffrey 1991), neural nets emerged as the regression method of choice for large time series (2000 or more values) from chaotic systems with low levels of noise. The main advantage of the neural nets was their robustness against incorrect choice of the model's dimension d (i.e., the number of lags in equation (1.1)). In practice the number of lags is not known , so a common strategy is to increase d with a fixed time-delay L until estimates of the quantity of interest stabilize (Mayer-Kress 1986). The other methods we studied -- local thin-plate splines, radial basis functions, and projection pursuit regression -- became unreliable when the model included lags which were not present in the equations generating the data.

One practical difficulty in regression with neural net models is selecting among the many possible combinations of d, L and q. Based on Gallant and Tauchen's (1990) experience in fitting seminonparametric GARCH nonlinear time series models, we propose model selection based on the "Bayesian Information Criterion" (BIC) (Schwarz 1978, Pötscher 1989). Assuming Gaussian errors in our case the criterion is computed as

$$\mathrm{BIC} = \frac{1}{2}\{1+\ln(2\pi)+ 2\ln(\mathrm{RMS})+\mathrm{P}\ln(\mathrm{n})/\mathrm{n}\}$$

where n is the number of data points, P is the number of parameters in the model, and RMS is the root-mean-square one step ahead prediction error. For a fixed number of parameters, minimizing BIC is equivalent to least squares.

3.4 Fitting a neural network to data using least squares

Choosing the network parameters (β_i,γ_{ij}) to optimize some performance criterion is a high-dimensional nonlinear minimization problem. The parallelism of nets results in many different parameter combinations achieving identical or nearly identical input-output maps, hence the objective function has many local minima and the fitting procedure must get past these to a global minimum. We experimented with a variety of optimization methods (including Levenberg-Marquardt as implemented in MINPACK; the Nelder-Mead simplex method, Powell's method, and conjugate gradient methods in Press et al. (1986); Brent's PRAXIS method from NETLIB; and genetic algorithms). The most efficient and reliable results were obtained with two implementations of the BFGS gradient method (Gill et el. 1981) with exact derivatives. BFGS is based on secant updates of an approximate Hessian or inverse Hessian, which is used to determine a search direction.

Our experience with neural nets matches some general conclusions of Gill et al. (1981) for "large residual" nonlinear least squares problems: a general-purpose gradient method performed better than one expoiting the structure of least squares, and exact calculation of the Hessian (rather than approximate updating) was not advantageous. Our first implementation used standard BFGS updates on an approximate inverse Hessian, the algorithm of Berndt et al. (1974) for choosing the steplength and identifying unacceptable search directions, and the termination criterion suggested by Gill et al. (1981) for unconstrained minimization. Numerical differentiation of the gradient was used to initialize the Hessian, and re-initialize it whenever a search direction proved unacceptable; more often than not re-initialization failed because the true Hessian was not positive definite, and the approximate Hessian was then re-initialized to the identity matrix. The second implementation used the NPSOL constrained minimization package (Gill et al. 1986). NPSOL uses a modified BFGS update of an approximate Hessian which preserves positive definiteness. The first approach (coded in GAUSS) was the most efficient we found, but the second (coded in FORTRAN) was faster due to the difference in computing platforms.

4. Simulation Studies of Thin Plate Splines and Neural Net Estimates of λ.

4.1 Mackey-Glass System

The methods described in Section 4 were evaluated using simulated data from a simple dynamic model that has some biological justification. The map considered is

$$f(x_1, \cdots, x_d) = ax_1 + b\phi(x_d) \tag{4.1}$$

where $\phi(u) = u/(1+u^k)$ and (a, b, k, d) are parameters of this system. For k= 10 this is a discretized version of the Mackey-Glass delay differential equation, originally developed to model the production and loss of white blood cells (Glass and Mackey 1988). It can also be interpreted as a model for population dynamics. If $0 < a < 1$ and $b > 0$ and if x_t denotes the number of adults, then a is the survival rate of adults and d is the time delay between birth and maturation . The term $b\phi(x_d)$ accounts for the recruitment of new adults due to births d years in the past, which is nonlinear due to decreased fecundity at higher population levels. For such biological systems multiplicative errors are more reasonable and so series were generated according to the model

$$x_{t+1} = f(x_t, \cdots, x_{t-d+1})\omega_t$$

where ω_t is distributed log $N(0, \sigma^2)$. In order to satisfy the additive model assumption in (1.1), the methods for estimating λ were applied to the log transformed series. (It should be noted that the Lyapunov exponent is unaffected by this transformation of the state vectors.)

The parameters values (a, b, k, d) = (.2, 2, 6, 2) were used in this study. The values of σ were taken to be (.01, .05,.1) and have corresponding Lyapunov exponents (.14, .14, .12). Figure 7 compares realizations of this system based on deterministic evolution and with noise ($\sigma = .1$). The addition of noise to this system has the effect of occasionally introducing sharp positive and negative spikes in the time series. The behavior of the deterministic system exhibits amplitudes that are more regular in size.

4.2 Thin Plate Spline Estimates

The first part of this study was designed to understand the statistical variability associated with estimates of λ derived from a thin plate spline estimate of the map. In this case the correct number of lags (d=2, L=1) was assumed to be known. Three levels of σ (.01, .05, .1) and 5 estimates of $\widehat{\lambda}$ were considered. It is useful to list these estimates in order of their expected accuracy.

J known: $\widehat{\lambda}$ is calculated from (2.1) using the true Jacobian matrix.

f known: Jacobian matrices are computed from a thin plate spline fit to the exact values of the map. ($y_k = f(X_k)$ in (3.1))

Rho fixed: The map is estimated from the time series using a fixed value of rho. This value was chosen to minimize the variance of the resulting estimates of $\widehat{\lambda}$.

GCV2: Map estimated from the time series where ρ is chosen by modifed cross validation with C=2

GCV1: Same as the GCV2 estimate except that ρ is chosen based on the usual cross-validation function (C=1).

Note that only the last two estimates are completely data-based and do not require some knowledge about f.

For each level of σ, 200 series with length N=80 were simulated, and an estimate of λ based on the five alternatives listed above was computed for each series. Boxplots in Figure 8 summarize the resulting distributions. Not surprisingly the variability of the estimates based on estimating f increases with σ. The choice of smoothing parameter also appears to have a significant effect on the distribution of $\hat{\lambda}$. Estimates that used an "optimal" choice of ρ have a variability comparable to that if the map was known. Estimates based on cross-validation are substantially less accurate but the skewness of the distribution depends on C. For ordinary cross-validation (C=1) the estimates tend to produce more spurious high values. When C=2 the estimates tend to underestimate λ. Figure 9 illustrates the dependence of $\hat{\lambda}$ to poor estimates of ρ. For the case σ=.1, the Lyapunov exponent estimates have been plotted verses the estimate of σ based on the GCV spline estimate, C=2. When the spline oversmooths the data (ρ too large), it was found that $\hat{\sigma} \gg .1$. For large values of $\hat{\sigma}$ we see more variable estimates of λ and also a substantial bias. Similar patterns were found for the other levels and estimators in this study.

4.3 Neural Network Estimates

Our simulations focussed on the stability of numerical least squares estimates of the network parameters and the accuracy of the corresponding estimates of λ. For each model specification (# of lags d, time-delay L and number of units q) we randomly generated 100 to 200 parameter values and used the best of these (lowest sum of squares) as initial values for a minimization with a lax termination criterion, repeating this process 150 times and saving each final set of parameter values. The 10 best of these 150 were then used as initial values for minimization with a stringent termination criterion. The estimate of λ for a given combination of d, L and q corresponded to the map estimate that minimized BIC over the 10 replicates.

The results of fitting a series from the Mackey-Glass system (N=125, σ=.1) are summarized in Figure 10. As with longer time series (McCaffrey et al 1991), plotting the resulting estimate of λ estimate vs. d produces a curve with a plateau very near the correct value, for L = 1 or 2. Since the qualitative presence or absence of chaos hinges simply on whether λ is positive or negative, the slight variation of the estimates is unimportant. The BIC criterion chose models with the correct time-delay (L = 1), and for a given d and L models with the BIC-preferred number of hidden units generally gave the most accurate estimate of λ (Figure 10b). In addition, with the values of d, L and q chosen by BIC, the variability among the estimates of λ from the 10 best-fitting parameter sets was quite low, indicating numerically stable estimates of the map f. Very similar results were obtained with other systems, including the cosine map described above with $N(0,0.2^2)$ additive dynamical noise, and the Hénon map with $N(0, 0.05^2)$ additive measurement errors.

In the Mackey-Glass example the results for L=3 are much less accurate and stable than those for L=1 or 2. This is not surprising, given that the true map depends only on x_{t-1} and x_{t-2} , so any model with L=3 gives a much poorer approximation to the true dynamics. This is reflected in BIC values (Figure 10b) which clearly favor models with L=1 or 2. However we rarely see such clear gaps in performance when fitting empirical data on population dynamics. One possible reason for the difference is that real-world populations are running in continuous time, and so they rarely have such non-smooth dynamics with sharp dependence on specific past values. Simulation results for a continuous-time chaotic system (the Rössler equations) with random perturbations are consistent with this interpretation (Figure 11). BIC favors shorter time-lags, but there is no clear gap between successive values of L, and estimates of λ based on a series of length 200 are sufficiently accurate and stable for L=1 to 5 to identify the system as chaotic. For comparison, using the standard Wolf et al. (1985) method on the same system without noise Vastano and Kostelich (1985) report: "results are poor with only 1024 points, but the Wolf algorithm gives reasonable estimates with 4096 points".

However these positive findings are critically dependent on finding accurate nonlinear least squares parameter estimates, rather than local minima that are far from optimal. With any of the minimization methods we examined, this required a large number of trials with different initial parameter values for each model specification. Attempting to reduce the number of trials by using an "upward search" strategy over model specifications (as in Gallant and Tauchen 1990) was ineffective. With upward search, the BIC criterion often chose models with too few parameters and consequently inaccurate estimates of λ, and minimal-BIC estimates of λ were less robust against incorrect choice of the time-delay L.

5. Analysis of Marten Fur Returns.

5.1 The Data

Based on records kept by the Hudson Bay Company, Jones (1914) tabulates by species the total number of pelts brought to market each year. Figure 12a graphs these observations for marten over the period of 1820 to 1900. As is typical with animal abundance series, the analysis will use the log of the original counts. Also, the log abundance is standardized to have a sample variance of one (Figure 12b). Note that in this case the log transformation has minimal effect on the marginal distribution for these data.

5.2 Spline Estimates of λ.

The analysis based on thin plate splines consisted of estimating maps for all pairs of lags between 2 and 15. The first lag was excluded from these subsets due to high lag one autocorrelation (.6) in the observed series. Although the embedding dimension is varied over a large range, each map estimate is just a two dimensional surface. This strategy is a compromise between estimating a high dimensional function and just focusing on the first few lags of the series. For the roughness penalty, $\mathfrak{J}_{d,m}$, m = 3. Thus $\hat{f}_{\infty}$ will be a second degree polynomial and *trace* $A(\infty)= 6$. The $\binom{14}{2} = 91$ pairs of lags each suggested estimates of λ and σ. In order to choose among these possibilities the BIC criterion (see Section 3.4) was calculated for each fit. (In the formula for BIC *trace* $A(\hat{\rho})$ was substituted for the number of parameters, P.) Figures 13a and 13b are scatterplots of the estimated Lyapunov exponents verses BIC for two different choices of C (C=1,2 respectively) in the cross-validation criterion for estimating ρ. When $C = 1$ the estimates of ρ tended to be slightly smaller except for a small fraction of lag pairs where the data was interpolated. The large positive Lyapunov exponent estimates in Figure 13a) correspond to these interpolatory map estimates. These interpolating splines were considered spurious, however, due to their rough appearance and thus these estimates of λ were discounted. For the estimates with C=2, the 5 lag pairs with the smallest BIC are reported in Table 1. One surprising feature when C=2 is that $\hat{\rho} = \infty$ for these five pairs of lags. Moreover, the estimates of ρ for the five cases reported in Table 1 did not change when C = 1 was used in the GCV function. It should be noted that except for the interpolatory cases all the thin plate spline estimates yielded negative values for $\hat{\lambda}$.

5.3 Neural Network Estimates of λ .

Figure 14 summarizes the results from fitting neural network models to the log-transformed marten data. All models with time-lags L = 2 to 6 were examined for d = 1 to 6 lags. The minimal-BIC model has 3 hidden units, 6 lags with time-delay 2. It is strongly chaotic ($\widehat{\lambda} \doteq 0.32$). The next cluster of models all have longer time-delays (L $\geq$ 4), and are less strongly chaotic ($\widehat{\lambda} \doteq 0.1$ to 0.15). Although these results differ from the spline estimates, these estimates are derived from high-dimensional models for the map (d $\geq$ 4). Restricting attention to lower-dimensional models (d $\leq$ 3, Table 2), the minimal-BIC neural net models are all nonchaotic and there is general agreement with the spline estimates regarding the level of noise and the complexity of the map (i.e., the number of parameters in the model).

6 Discussion and Conclusions

In this article we have presented a statistical framework for the analysis of chaotic systems. One basic question is whether it is ever possible to identify chaotic dynamics in short, noisy systems. Simulation results based on thin plate spline estimates suggest that under some circumstances it is possible. However, the accuracy of the estimate degrades as the variance of the random component increases and is sensitive to the choice of smoothing parameter.

This work also suggests that map estimates based on neural networks are feasible for noisy systems with a small number of observations, but the advantages of this model can be retained only if care is taken to find accurate least squares parameter values. Consequently the neural net model has a heavy computational cost compared with linear estimators such as splines (2 to 3 orders of magnitude for sample sizes of roughly 100). The cost is bearable for analyzing a few short data sets, but it is an obstacle to extensive simulation studies or to inference based on bootstrapping.

The basic strategy in formulating map estimates is to avoid estimating the full d-dimensional multivariate function and thus avoid the "curse of dimensionality". In the analysis of the marten series we have taken two approaches. One is to estimate the best two-dimensional surface from searching over a wide range of possible pairs of lags. Since a thin plate spline has good approximation properties it is assumed that a flexible surface based on two lags may serve as a surrogate for a higher dimensional map with a larger embedding dimension d. The neural net estimate represents the map *via* univariate functions taking linear combinations of the lagged values as arguments. This strategy will be successful if the true map has a simple structure with respect to several projections of the lagged values onto 1-dimensional subspaces.

For the marten fur returns we do not find any evidence for "low dimensional" chaos ($d \leq 3$). Both the spline and neural net estimates consistently yield weakly negative estimates of λ for low-dimensional models, except in several spurious cases where the spline interpolated the data. Because these two function estimates are very different in their form, our confidence in the analysis of the marten data is based on the rough agreement between the estimates of λ for splines and neural nets. It is also encouraging that both methods identified models which base predictions on values roughly 10 to 15 years in the past. In higher dimensions the neural net estimates favor chaotic models, but this should be interpreted with caution given that all our simulation studies involved low-dimensional systems.

Most work on the effects of noise on chaotic data analysis has aimed at estimating properties of the system with the noise deleted. This is appropriate if the noise is primarily measurement error, which should be "filtered" out to get a better picture of the true system (e.g., R. Smith 1991). In contrast, we do not require the assumption of perfectly deterministic dynamics, and the Lyapunov exponent as defined here is a joint property of the intrinsic nonlinear map and the extrinsic random shocks to the system. In the absence of random shocks there would be a different attractor, hence a different distribution of Jacobians entering into the definition of λ. The cosine map provides a good example: with additive $N(0,0.2^2)$ noise $\lambda \doteq 0.35$, but the noise-free system has $\lambda \doteq 0.5$. The difference between these is not an error due to noise, but a genuine effect of noise on the system's qualitative dynamics. The noisy system spends relatively more time in regions of state space where the short-term sensitivity to initial conditions is smaller, so the long-term sensitivity measured by the Lyapunov exponent is smaller. Both λ's are legitimate objects of study, but we emphasize the former because it reflects the system as it is-- e.g., marten in the wild and subject to fluctuations of climate, food supply, predator abundance, etc.

Beyond reliable estimates of the Lyapunov exponent it is important to be able to quantify the variability of the estimate. For example it would be useful to be able to construct a confidence interval for λ. Because the distribution of $\hat{\lambda}$ is a complicated, nonlinear function of the map estimate it is difficult to derive an asymptotic expression for the standard error. One alternative is to use bootstrapping. Bootstrapping will only be useful if the map estimate is accurate enough to yield distributional information about the variance and bias of $\hat{\lambda}$. In situations where the noise is large it is not clear that the map estimates are this accurate. Part of the problem is in smoothing parameter selection. If the map estimate over-smooths the data and gives a large estimate of σ then bootstrap simulations from this estimated model may be misleading. In situations where data is limited it would be helpful to have an independent estimate of the size of the random component. This suggest a role for short-term experiments on the system to estimate σ.

Acknowledgements: This research was supported in part by National Science Foundation grants DMS-8715756 and SES-8808015 and the North Carolina Experiment Station Project NCO-6134.

References

Abarbanel, H, D. (1991), "Lyapunov Exponents in Chaotic Systems: Their Importance and their Evaluation using Observed Data," to appear *Modern Physics Letters* B.

Barron , A. R. (1991), "Approximation and Estimation for Artificial Neural Networks," *University of Illinois at Urbann-Champaign, Department of Statistics, Technical Report* #58.

Bates, D., M. Lindstrom, G. Wahba, and B. Yandell (1987), "GCVPACK-Routines for generalized cross-validation," *Communications in Statistical Simulation and Computing* 16 263-297.

Berndt, E.K., B.H. Hall, R.E. Hall, and J.A. Hausman (1974), "Estimation and inference in nonlinear structural models", Ann. Econ. & Social Measurement 3/4, 653-665.

Caudill, M. and C. Butler (eds). 1987. *IEEE First International Conference on Neural Networks: Sheraton Harbor Island East, San Diego, California, June 21-24, 1987.* San Diego: SOS Printing.

Cox, D. D. (1984), "Multivariate smoothing splines," *SIAM Journal of Numerical Analysis*, 21, 789-813.

Ellner, S., A.R. Gallant, D. McCaffrey, and D. Nychka (1991), "Convergence rates and data requirements for Jacobian-based estimates of Lyapunov exponents from data". *Physics Letters A* 153, 357-363.

Eckmann, J.-P. and D. Ruelle (1985), "Ergodic theory of chaos and strange attractors," *Reviews of Modern Physics*, 57(3), 617-656.

Eckmann, J.-P., S. O. Kamphorst, D. Ruelle, and S. Ciliberto (1986), "Liapunov exponents from time series," *Physical Review A*, 34, 4971-4979.

Farmer, J.D. and J.J. Sidorowich (1988), "Exploiting chaos to predict the future and reduce noise," in Y.C. Lee (ed.) *Evolution, Learning, and Cognition*, World Scientific Press, Singapore, p. 277.

Friedman, J, and W. Stuetzle (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association* 76, 817-823.

Gallant, A.R. & G. Tauchen. (1990) "A nonparametric approach to nonlinear time series analysis: estimation and simulation". IMA Preprint series #705, Institute for Mathematics and its Applications, University of Minnesota, Minneapolis.

Gallant, A. R., and H. White (1988), "There Exists a Neural Network that Does Not Make Avoidable Mistakes," *IEEE International Conference on Neural Networks: Sheraton Harbor Island, San Diego, California, July 24-27, 1988.* San Diego: SOS Printing, I.657-I.664.

Gallant, A. R., and H. White (1989), "On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks," North Carolina State University, Institute of Statistics Mimeograph Series No. 1964.

Gill, P.E., W. Murray, and M.H. Wright (1981), *Practical Optimization* . Academic Press, London and New York.

Gill, P.E., W. Murray, M.A. Saunders, and M.H. Wright (1986), "Users Guide for NPSOL (Version 4.0): A Fortran Package for Nonlinear Programming", Technical Report SOL 86-2, Systems

Optimization Laboratory, Department of Operations Research, Stanford University, Stanford CA 94305.

Glass, L. and Mackey, M.C. (1988), *From Clocks to Chaos: the Rhythms of Life.* Princeton Univ. Press, Princeton.

Guckenhiemer, J. (1982) "Noise in chaotic systems", *Nature* 298, 358-361.

Guckenheimer, J. and P. Holmes 1983. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* Springer-Verlag, New York.

IEEE, 1988. *IEEE International Conference on Neural Networks: Sheraton Harbor Island, San Diego, California, July 24-27, 1988* (1988). San Diego: SOS Printing.

IJCNN, 1989. *IJCNN, International Joint Conference on Neural Networks: Sheraton Washington Hotel* (1989). San Diego: SOS Printing.

Jones, J.W. 1914. *Fur-Farming in Canada.* Second edition. The Mortimer Company, Ltd., Ottawa.

Kifer, Y. (1986). *Ergodic Theory of Random Transformations*, Birkhäuser, Basel.

Lapedes, A., and R. Farber (1987), "Nonlinear Signal Processing Using Neural Networks: Prediction and System Modelling," Los Alamos National Laboratory, Technical Report LA-UR-87-2662.

Mayer-Kress, G. (ed.) (1986), *Dimensions and Entropies in Chaotic Systems*, Springer- Verlag, Berlin.

M^{c}Caffrey, D. (1991) "Estimating Lyapunov exponents with nonparametric regression and convergence rates for feedforward single hidden layer networks" Ph. D. Thesis, North Carolina State University, Raleigh, NC.

M^{c}Caffrey, D.F., S.Ellner, A.R. Gallant, and D.W. Nychka (1991). "Estimating the Lyapunov exponent of a chaotic system with nonparametric regression". *Journal of the American Statistical Association* (to appear).

McCulloch, W.S. and W. Pitts (1943),"A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics* 5, 115-143.

Nychka, D.W. (1990), "The Average Posterior Variance of a Smoothing Spline and a Consistent Estimate of the Average Squared Error," *Annals of Statistics* 18, 415-428.

Nychka, D.W. (1991), "Choosing a range for the amount of smoothing in nonparametric regression," *Journal of the American Statistical Association* 86 (in press)

Pötscher, B.M. 1989. "Model selection under nonstationarity: autoregressive models and stochastic linear regression models," *Annals of Statistics* 17, 1257-1274.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge UK.

Ruelle, D. (1989), *Chaotic Evolution and Strange Attractors*, Cambridge University Press, Cambridge UK.

Rumelhart, D.E., G.E. Hinton and R.J. Williams (1986), "Learning Internal Representions by Error

Propagation," in Rumelhart, D.E. and J.L. McClelland (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol 1. Cambridge MA: MIT Press, 318-362.

Schuster, H.G. (1988), *Deterministic Chaos: An Introduction,* (2nd revised edition), VCH Verlagsgesellschaft, Weinheim FRG.

Schwartz, G. (1978) "Estimating the dimension of a model", *Annals of Statistics* 6,461-464.

Smith, R.L. (1991), "Estimating Dimension in noisy chaotic systems," manuscript.

Vastano, J.A. and E.J. Kostelich (1985), "Comparison of algorithms for determining Lyapunov exponents from experimental data", pp. 100-107 in Mayer-Kress, G. (ed), *Dimensions and Entropies in Chaotic Systems*, Springer-Verlag, Berlin.

Wahba, G. (1990), *Spline Models for Observational Data,* SIAM, Philadelphia.

Wolf, A., J.B. Swift, H.L. Swinney, and J.A. Vastano (1985), "Determining Lyapunov exponents from a time series," *Physica* 16D, 285-315.

Figure Legends

Figure 1. Attractor for the deterministic cosine map system. Plotted are the state vectors, (x_{t-1}, x_t), for 1000 iterations for the cosine map defined in Section 2.1. The starting point for this sequence was obtained by first iterating the map from an arbitrary starting value several thousand times. The dominant Lyapunov exponent is approximately 0.5 and the correlation dimension is approximately 1.2.

Figure 2. Attractor for the cosine map with noise. Plotted are the 1000 state vectors for the cosine map with additive $N(0,(.2)^2)$ perturbations (see (1.1)). In this case the dominant Lyapunov exponent has been estimated to be approximately .35.

Figure 3. Action of the deterministic cosine map on the cohort of state vectors, Ω_t. This panel traces the action of the cosine map on 500 state vectors initially clustered at (0,0). The cohort of state vectors are plotted at t=(0,5,10,15,20,40).

Figure 4. Action of the cosine map with noise on a cohort of state vectors. The initial conditions for the state vectors are the same as that in Figure 3. The subsequent states are generated according to (1.1) where e_t, the random component is a $N(0,(.2)^2)$ random variable. The *same* value of e_t is used across the cohort and these state vectors are plotted at t=(5,10,15,20,45,50).

Figure 5. Exponential divergence of trajectories from similar initial conditions. For the deterministic and noisy cosine systems depicted in FIgures 3 and 4, the divergence of the individual trajectories is followed over time. Plotted are the log average distance among state vectors in the cohort for the first twenty time points. The solid line is the result for the deterministic map and the dashed lines indicate the behavior for five realizations of the random system. These latter trajectories differ in the sequence of errors $\{e_t\}$. Note that for a particular sequence of random errors, the *same* value of e_t is used to generate all points in the cohort for the next time period.

Figure 6. Network architecture for a single-hidden-layer feedforward network with a single output, and three units in the hidden layer. Inputs x_1 and x_2 are received by the two input units and passed on to each of the hidden units u_1,u_2,u_3. The output unit performs an affine transformation on the summed hidden layer outputs to produce the network's output. This network computes the function $F(X) = \beta_0 + \sum_{j=1}^{3} \beta_j \psi(\gamma_{j1} x_1 + \gamma_{j2} x_2 + \gamma_{j0})$ where ψ is the activation function of the hidden units and the parameters β_j, γ_{ji}, are described in Section 3.3.

Figure 7. Realizations from the Mackey Glass system (4.1) with and without noise. The two realizations consist of times series of 400 points where a) is the time series for the deterministic system and b) the same for a noisy system ($\sigma = .1$).

Figure 8. Distribution of Lyapunov exponent estimates based on thin plate splines. Time series of N=80 values were simulated from a Mackey-Glass system at three levels of noise variance (σ=.01, .05, .10). For each level of σ, 200 realizations were generated and λ was estimated using the five different methods described in Section 4.2. The boxplots summarize the resulting distribution of the estimates. It should be noted that "J known" refers to an estimator based on the true partial derivatives of the map. Only the last two estimators (GCV 2 and GCV 1) are actually based entirely on the observed data and are constructed without knowledge of the true map.

Figure 9. Dependence of $\hat{\lambda}$ on $\hat{\sigma}$. For the level σ=.1 in Figure 8, the 200 estimates of the dominant Lyapunov exponent λ based on modified GCV with C=2 are plotted against an estimate of σ.

Figure 10. Neural net estimates of λ for a single realization of the discrete Mackey-Glass system with $\sigma = 0.1$. a) The first series of plots indicate the stability of the estimates for different starting values and across different parametrizations. For each combination of L (1,2,3) and d (1,2,3,4,5,6) the estimate with the smallest BIC is identified among the range of hidden units, q, (1,2,3,4,5,6). Let $\hat{q}$(L,d) denote the number of hidden units associated with this estimate. Besides the estimate with lowest BIC for a given choice of L and d there are also 9 other replicate estimates for the model (L,d,$\hat{q}$) based on different starting values. Accordingly, boxplots are plotted for these 10 estimates as a function of d and L. The scatter in these estimates illustrate the numerical stability of recovering a similar estimate when only the starting values are varied. b) Estimates of λ are plotted verses the corresponding value of BIC over all combinations of (L,d,q). The plotted numeral indicates the value of L for each case and circled points are those where q= $\hat{q}$(L,d).

Figure 11. Neural net estimates of λ for a realization of the Rössler system, a standard example of chaos in simple differential equations (e.g., Schuster 1988), with random perturbations. The Rössler equations are the three-variable system: $dx/dt = -z - y$, $dy/dt = x+ay$, $dz/dt = b+z(x-c)$. Parameter values a=0.15, b = 0.2, c = 10.0, were used, which in the absence of noise give a chaotic system with $\lambda \doteq 0.09$ (Mayer-Kress 1986). The equations were integrated numerically using fourth order Runge-Kutta, but at times t = 0.25, 0.5, 0.75, $\cdots$, the integration was halted and the values of x(t), y(t), and z(t) were multiplied by independent lognormal $(0, 0.05^2)$ random perturbations. The estimates of λ shown here are based on the univariate time series of length 200 from x(t) sampled at

t = 25.5, 30.0, 30.5, $\cdots$. The layout of this figure is the same as Figure 10. Note, however that only the estimates for L =(1,3,5) are reported in part a).

Figure 12. Hudson Bay Company marten fur records. Figure 12a) Annual marten fur records from Jones (1914). The standardized series in Figure 12b) is obtained by taking the log of the raw data and dividing by the marginal standard deviation.

Figure 13. Lyapunov exponent estimates for the standardized marten series based on thin plate splines. All subsets of two lags in the range (2-15) are considered and for each map estimate the exponent is plotted against the value of BIC. The numerals indicate the maximum lag in each subset. a) ρ estimated by GCV with C=1 and b) ρ estimated by GCV with C=2.

Figure 14. Lyapunov exponent estimates for the standardized marten series based on neural nets. These plots are similar to those in Figure 10. Note that the time delays plotted in part a) are L= (2,4,6).

Table 1. Thin plate spline estimates (GCV C = 2) with the five smallest BIC values for the transformed Marten Fur Series (log transformed and standardized to unit variance).

Lags	BIC	$\hat{\sigma}$	trace $A(\hat{\rho})$	$\hat{\lambda}$
10, 15	1.193	.69	6.0	-.025
10, 14	1.210	.70	6.0	-.006
9, 15	1.261	.74	6.0	-.020
2, 14	1.261	.74	6.0	-.029
11, 15	1.274	.75	6.0	-.016

Table 2. Neural net estimates for 3 or fewer lags with the five smallest BIC values for the transformed Marten Fur Series. P is the number of parameters in the model, P = 1+q(d+2), where q is the number of hidden units and d is the number of lags. $\hat{\sigma}$ is the root-mean-square prediction error of the model.

Lags	BIC	$\hat{\sigma}$	P	$\hat{\lambda}$
5, 10, 15	1.29	.70	6	-0.02
5, 10, 15	1.33	.60	11	-0.03
3, 6, 9	1.34	.73	6	-0.07
2, 4	1.34	.65	9	-0.36
5, 10	1.35	.77	5	-0.48

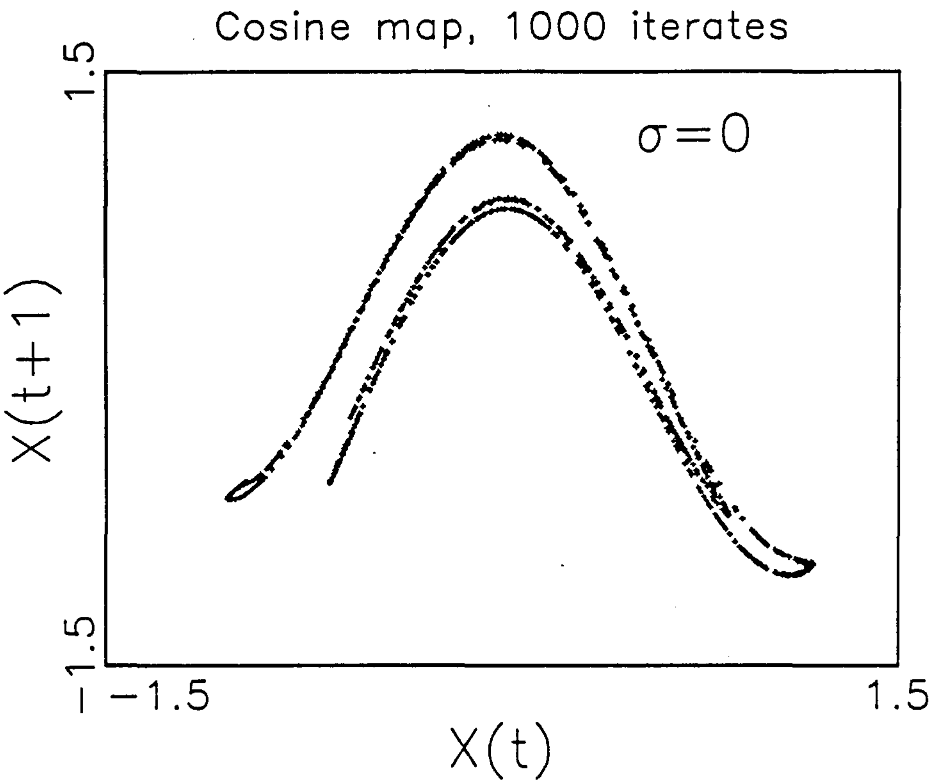


Figure 1

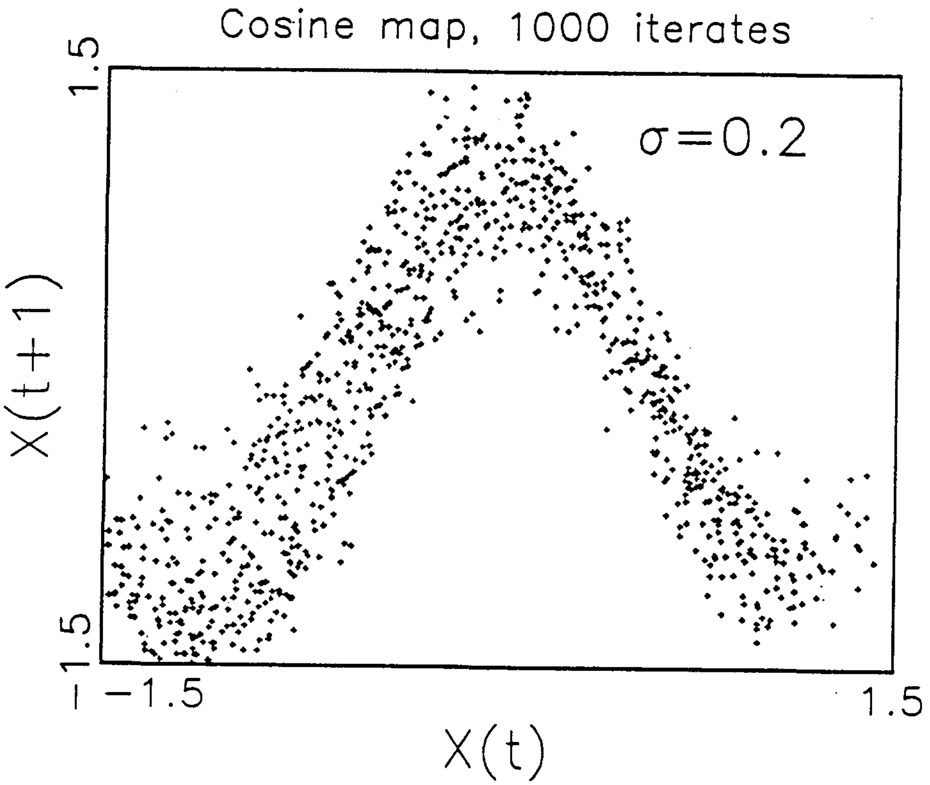
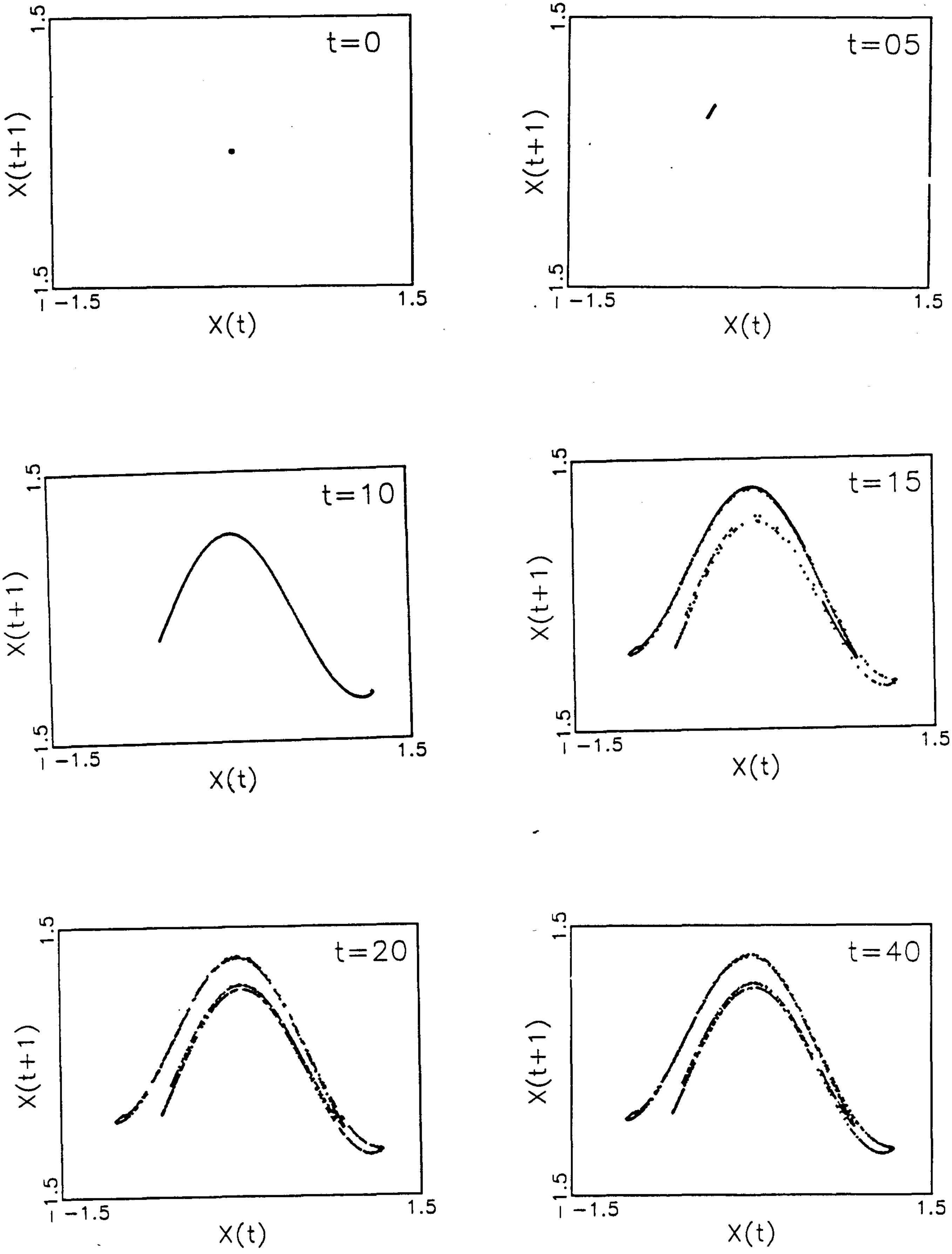
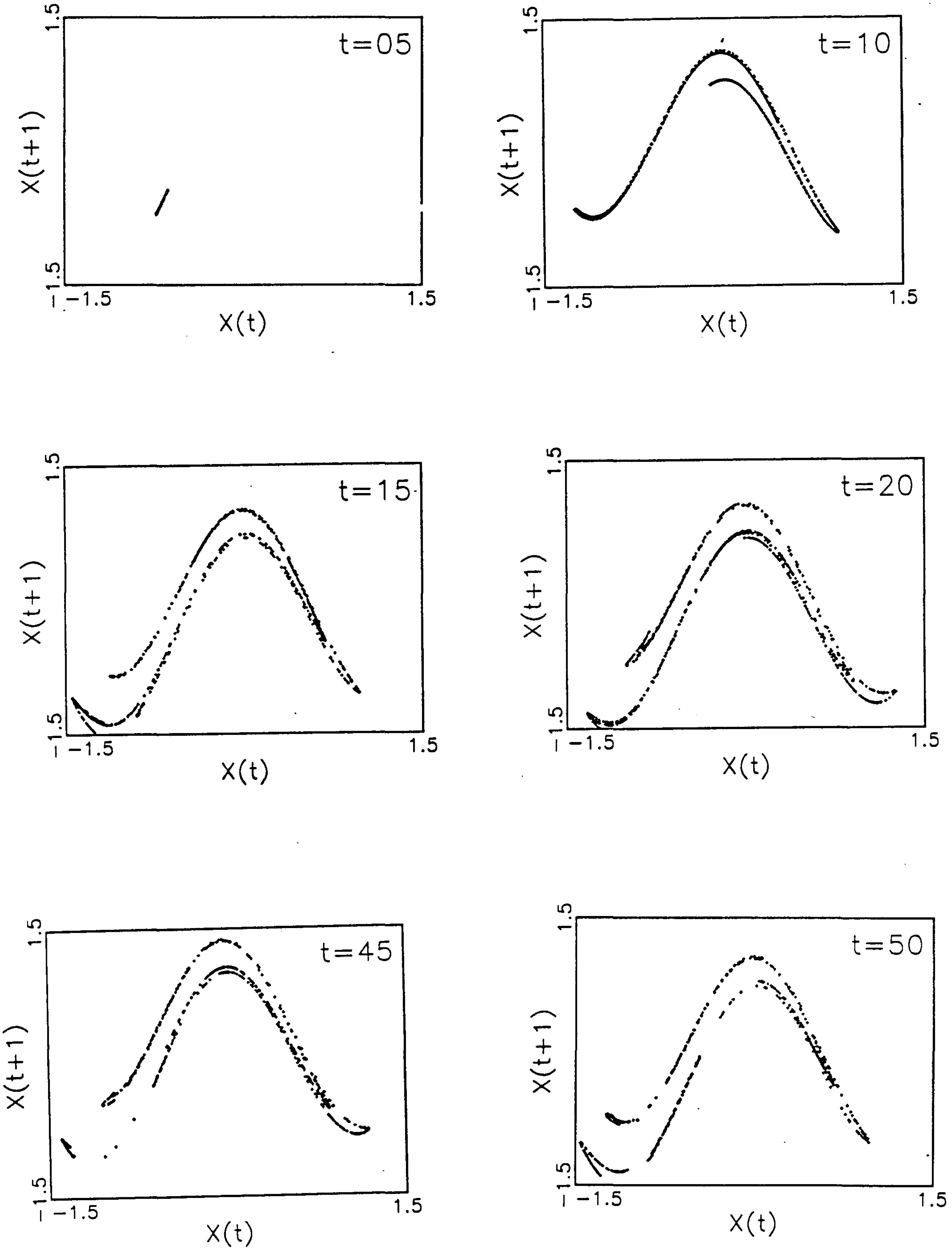


Figure 2


t=0
X(t+1)
1.5
-1.5
-1.5
1.5
X(t)
t=05
X(t+1)
1.5
-1.5
-1.5
1.5
X(t)
t=10
X(t+1)
1.5
-1.5
-1.5
1.5
X(t)
t=15
X(t+1)
1.5
-1.5
-1.5
1.5
X(t)
t=20
X(t+1)
1.5
-1.5
-1.5
1.5
X(t)
t=40
X(t+1)
1.5
-1.5
-1.5
1.5
X(t)


t=05
1.5
X(t+1)
-1.5
-1.5
1.5
X(t)
t=10
1.5
X(t+1)
-1.5
-1.5
1.5
X(t)
t=15
1.5
X(t+1)
-1.5
-1.5
1.5
X(t)
t=20
1.5
X(t+1)
-1.5
-1.5
1.5
X(t)
t=45
1.5
X(t+1)
-1.5
-1.5
1.5
X(t)
t=50
1.5
X(t+1)
-1.5
-1.5
1.5
X(t)

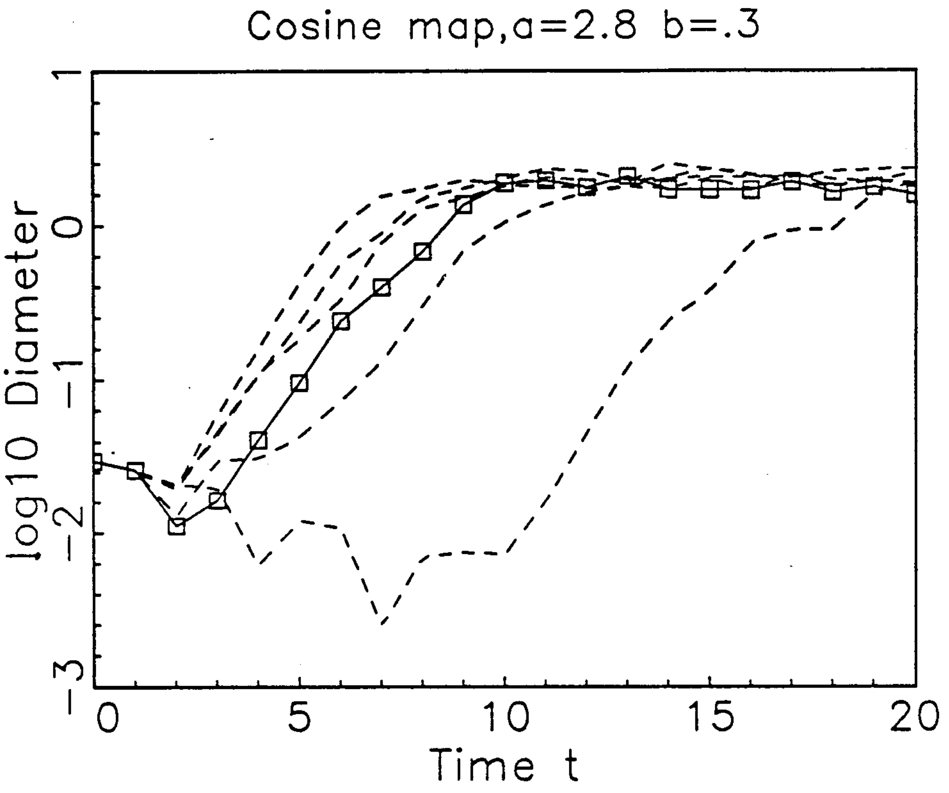


Figure 5

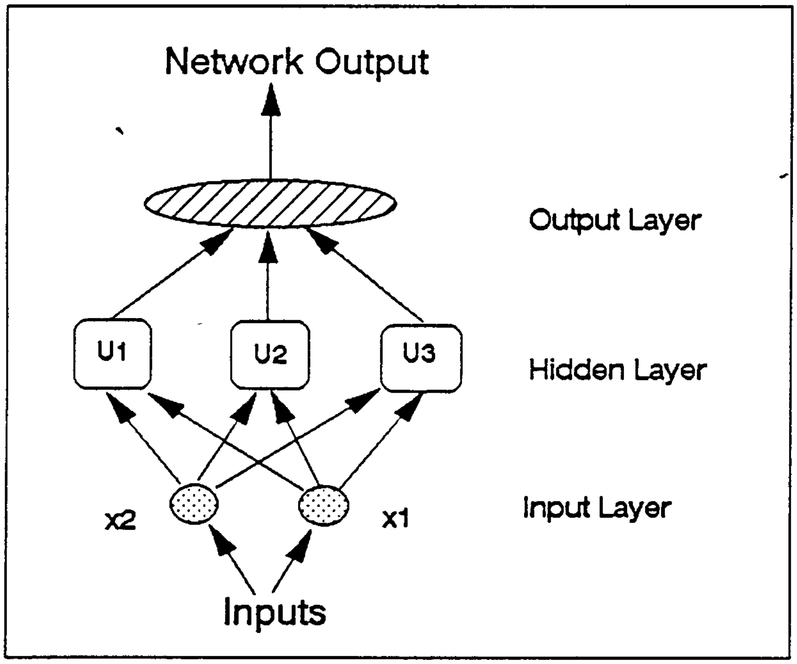


Figure 6

Figure 7a) Deterministic system

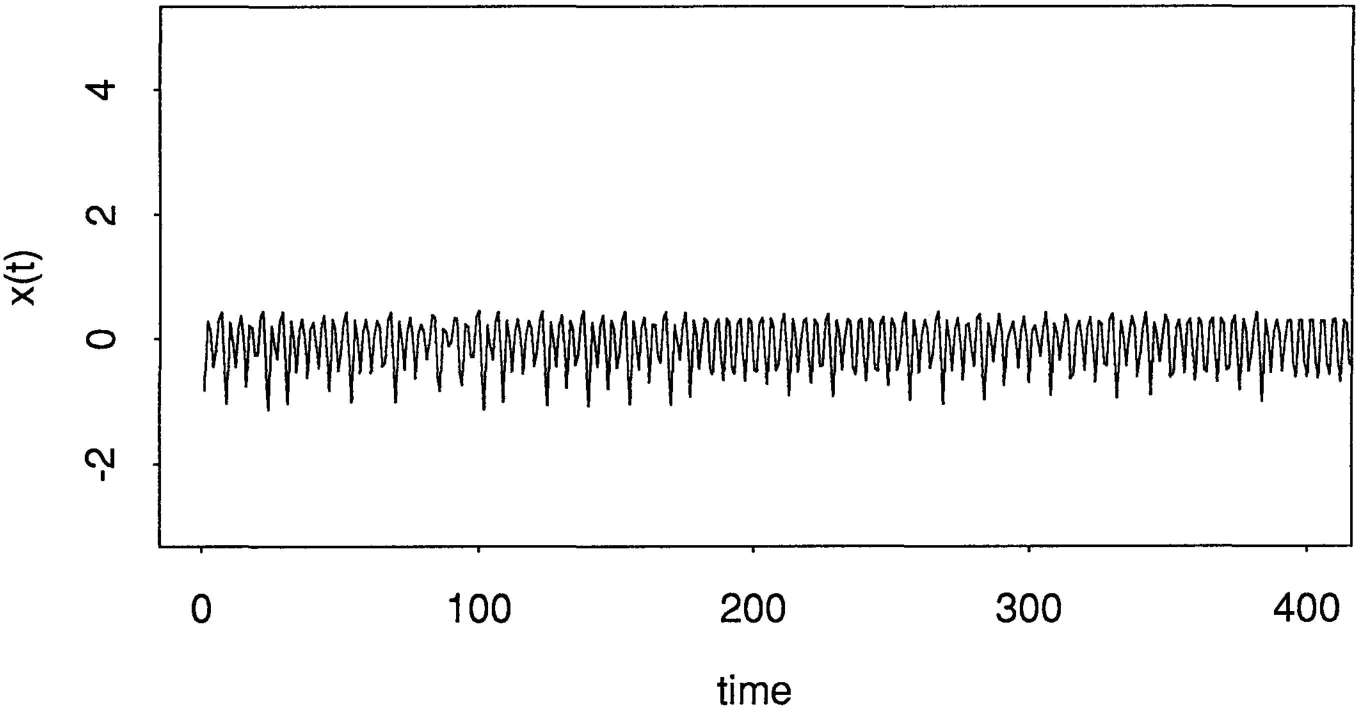


Figure 7b) Noisy system sigma =.1

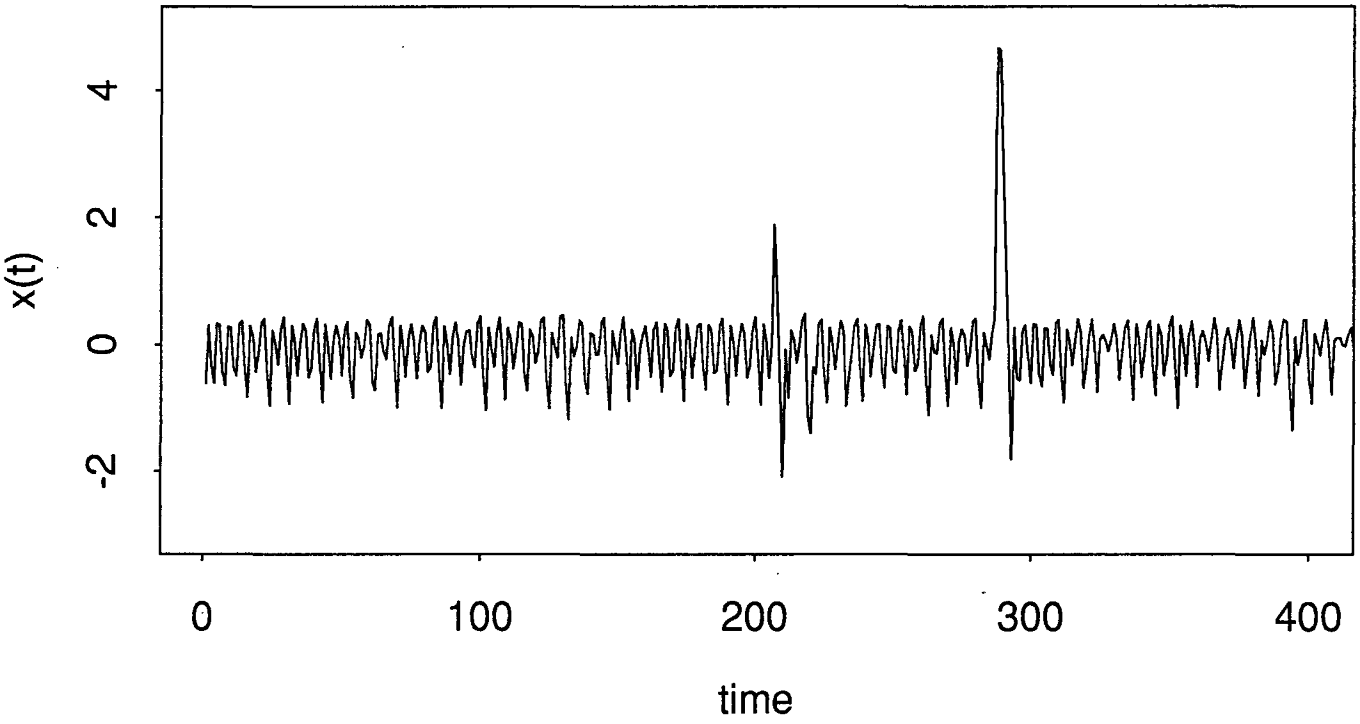
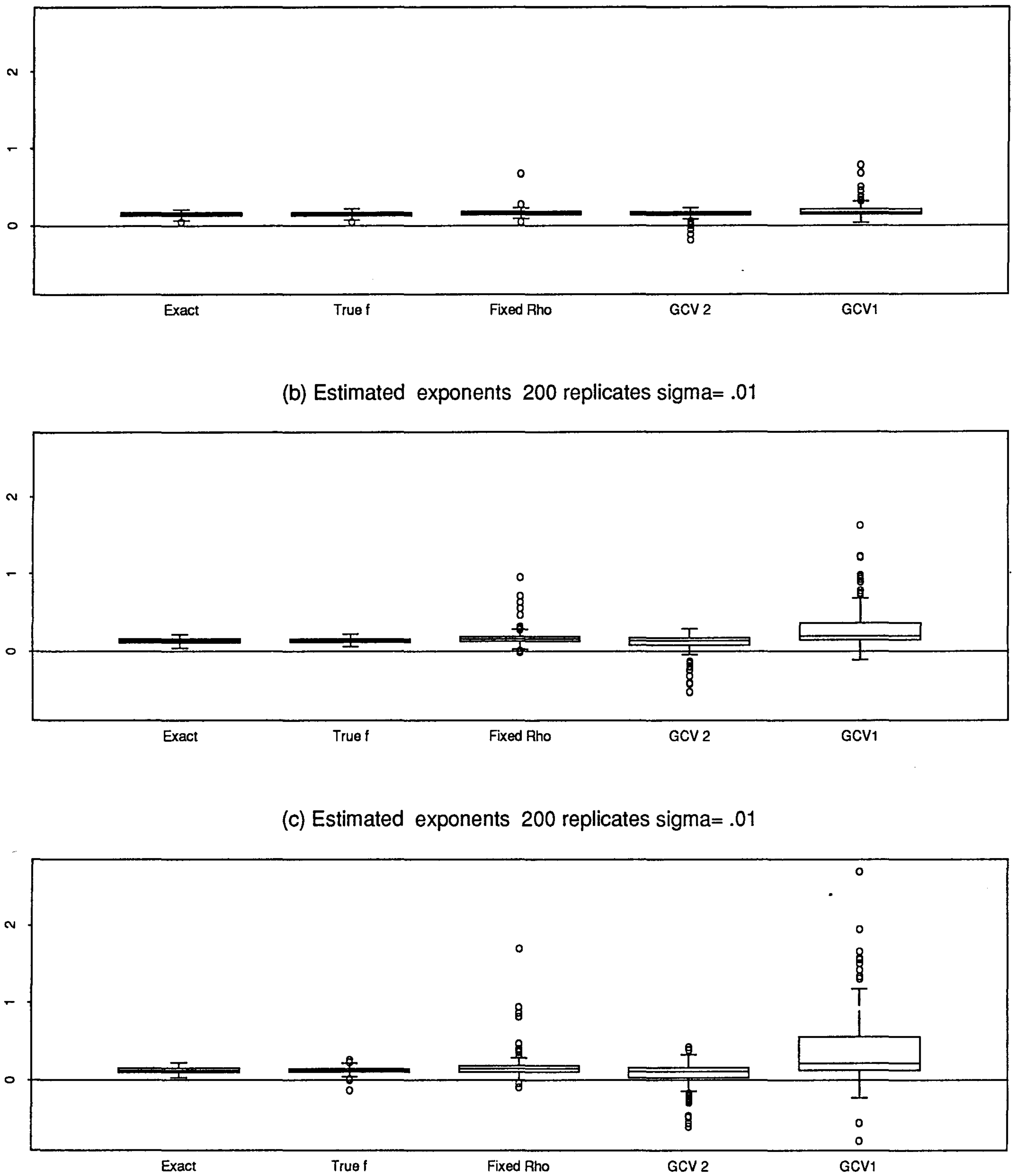


Figure 8: Mackey Glass System Lyapunov Exponent Estimates



(a) Estimated exponents 200 replicates sigma= .01

(b) Estimated exponents 200 replicates sigma= .01

(c) Estimated exponents 200 replicates sigma= .01

Figure 9: Dependence of Estimates on Residual Variance sigma= .1

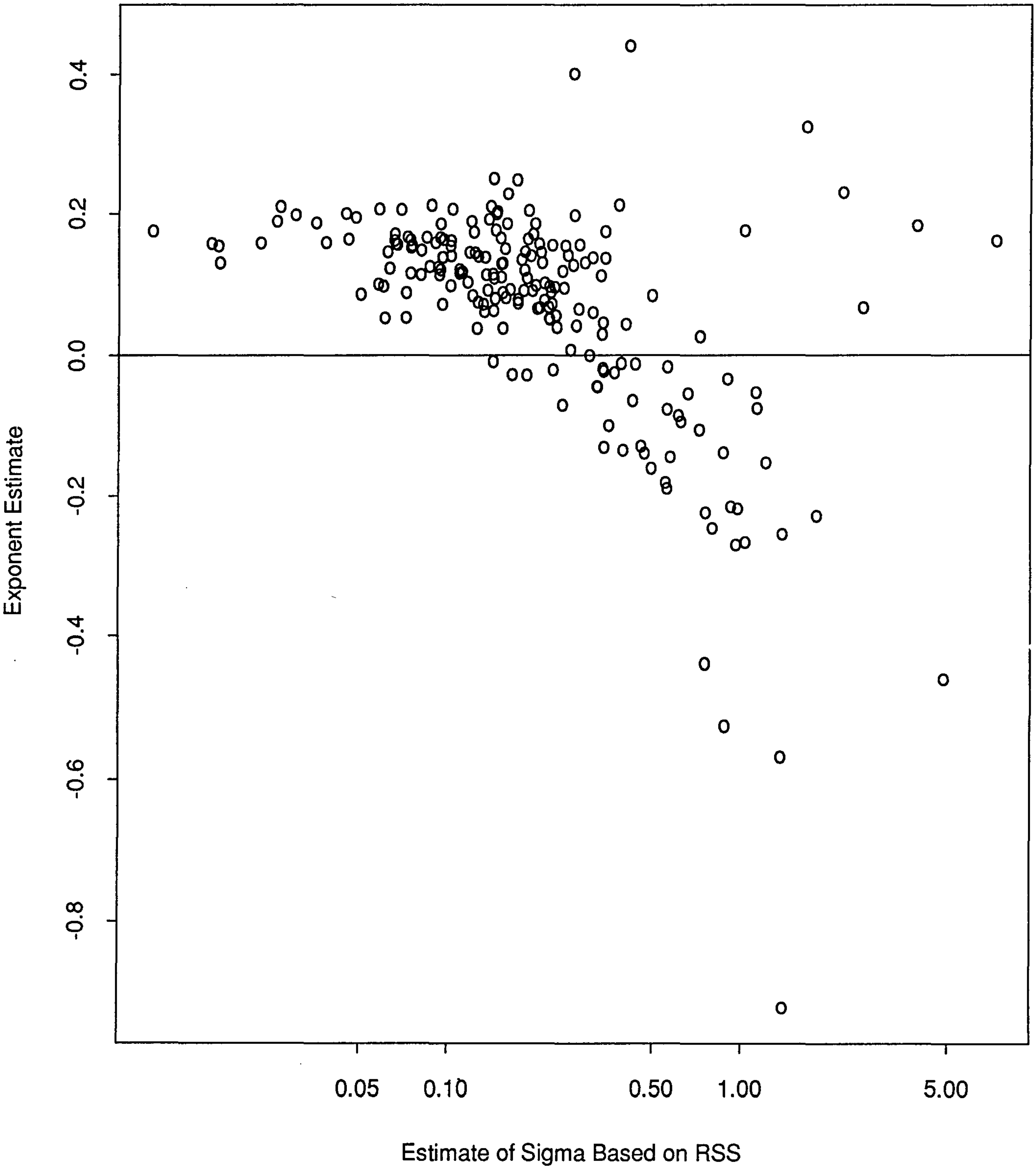


Figure 10a) Neural Net Estimates for Mackey Glass System as a Function of Embedding Dimension

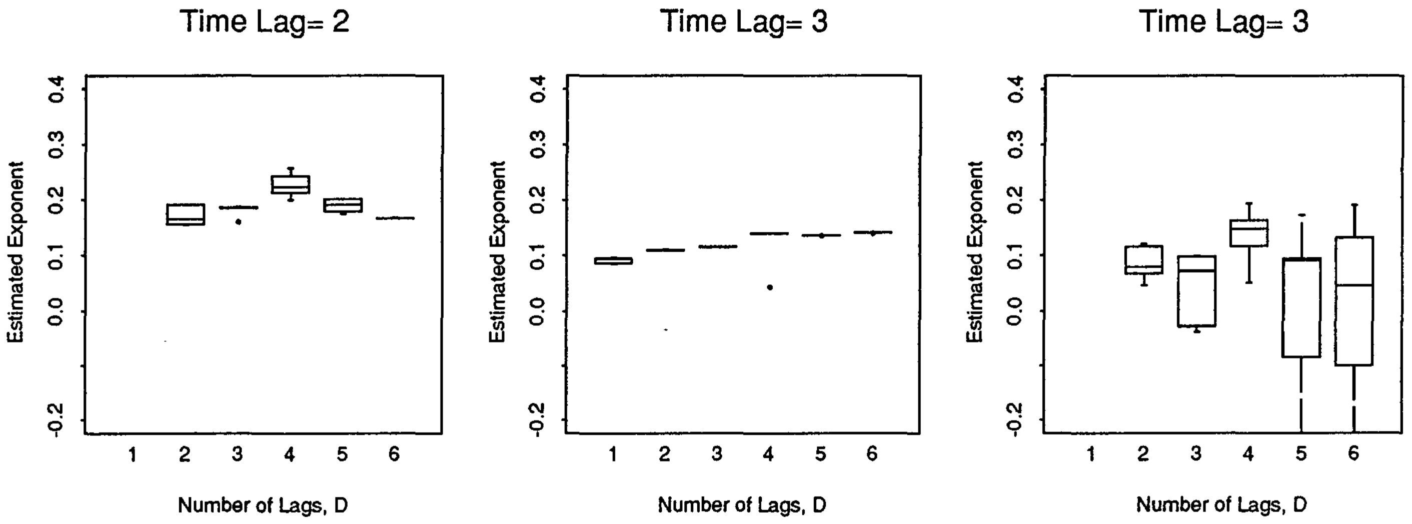


Figure 10b) Relationship of Neural Net Estimates with BIC

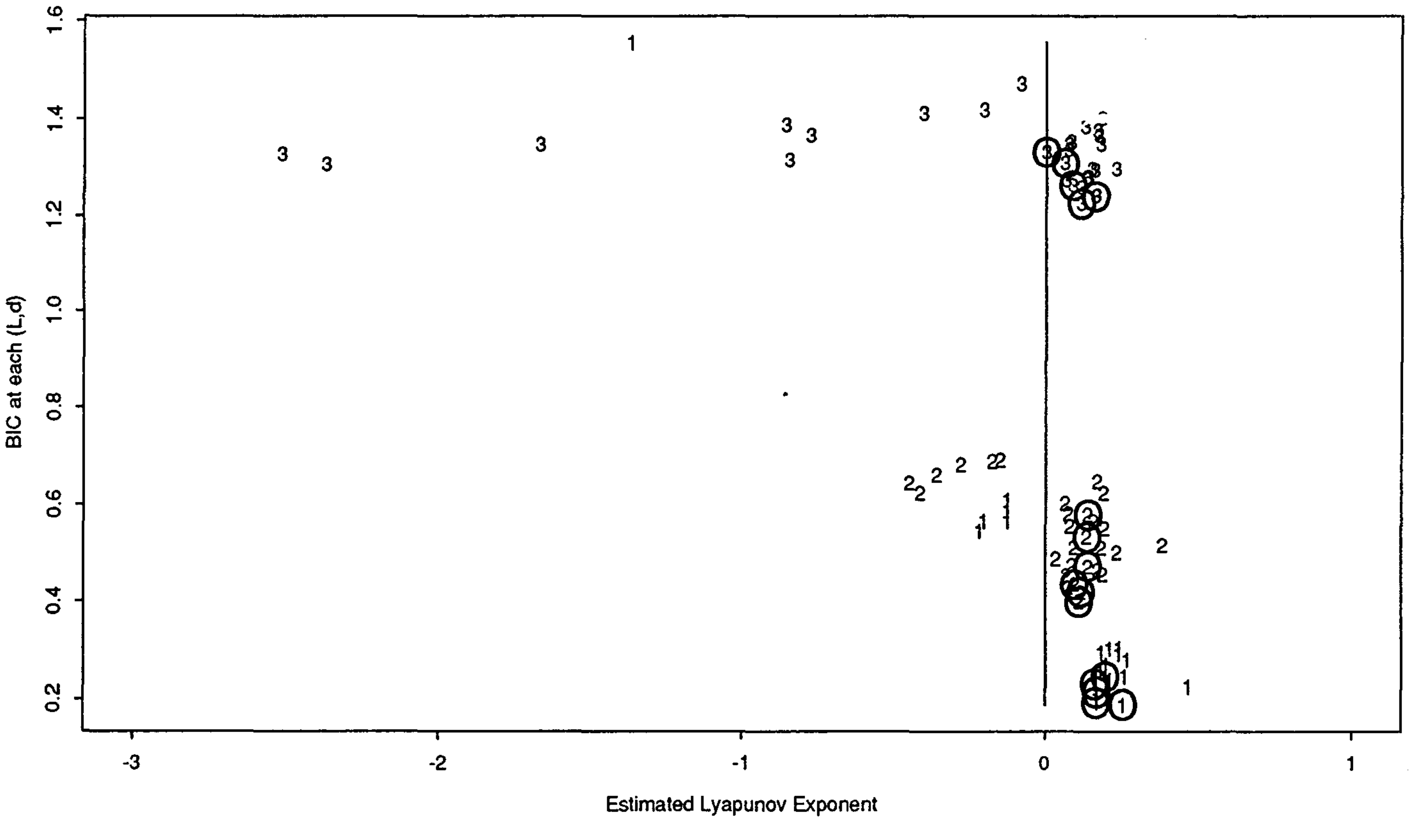


Figure 11a) Neural Net Estimates for the Rossler System as a Function of Embedding Dimension

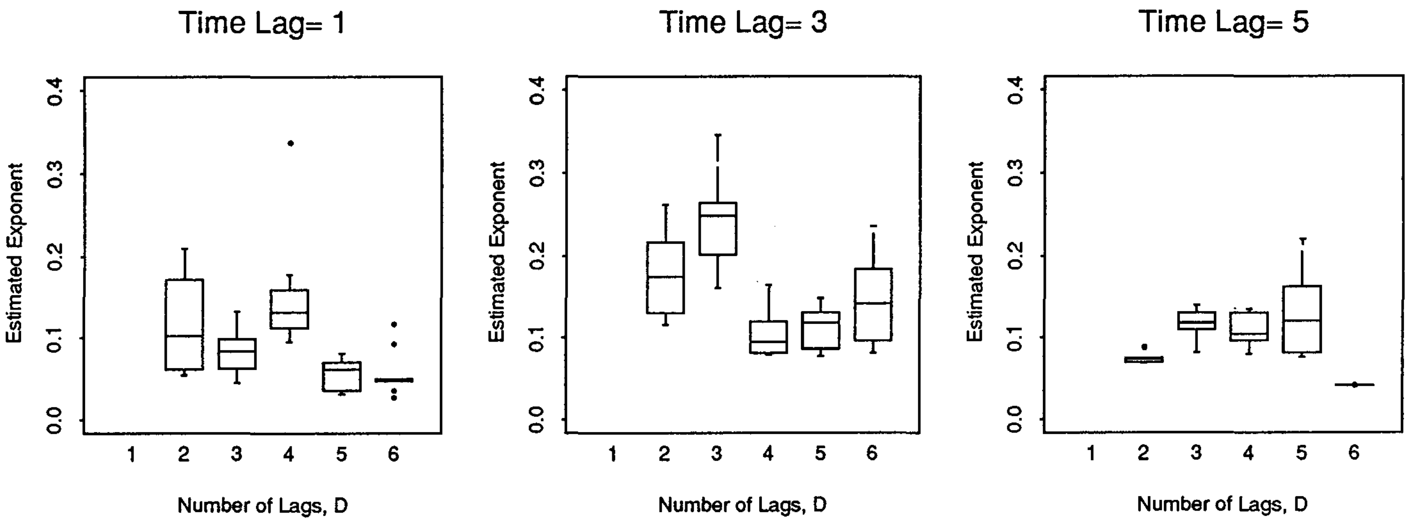


Figure 11b) Relationship of Estimates to BIC

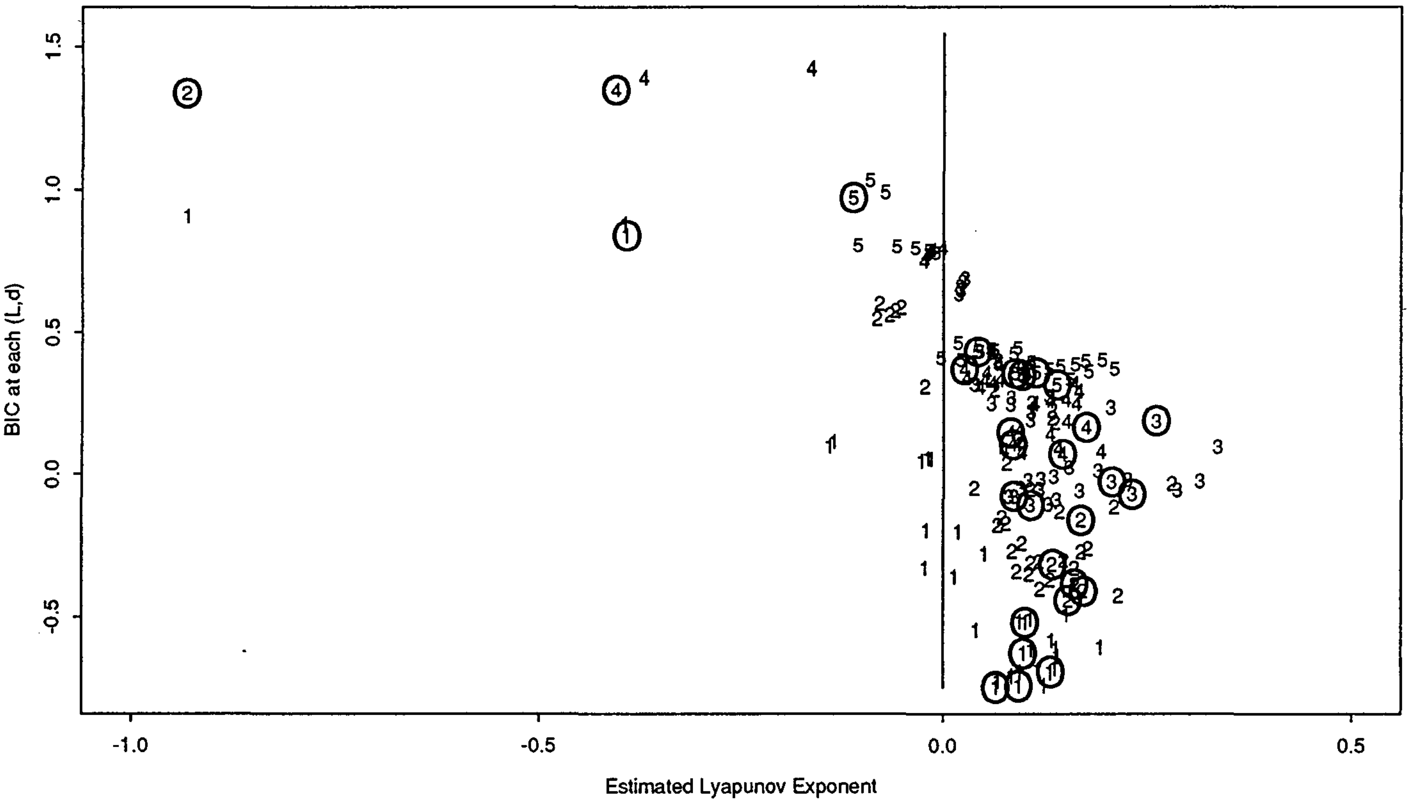


Figure 12a): Hudson Bay Company Marten Fur Records

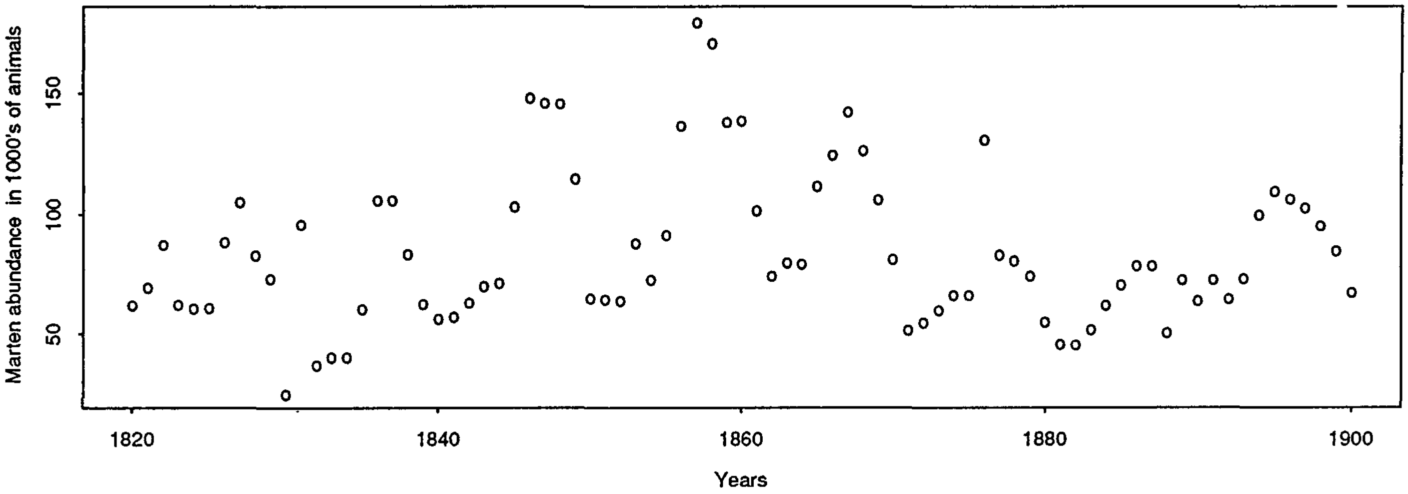


Figure 12b): Standardized Time Series

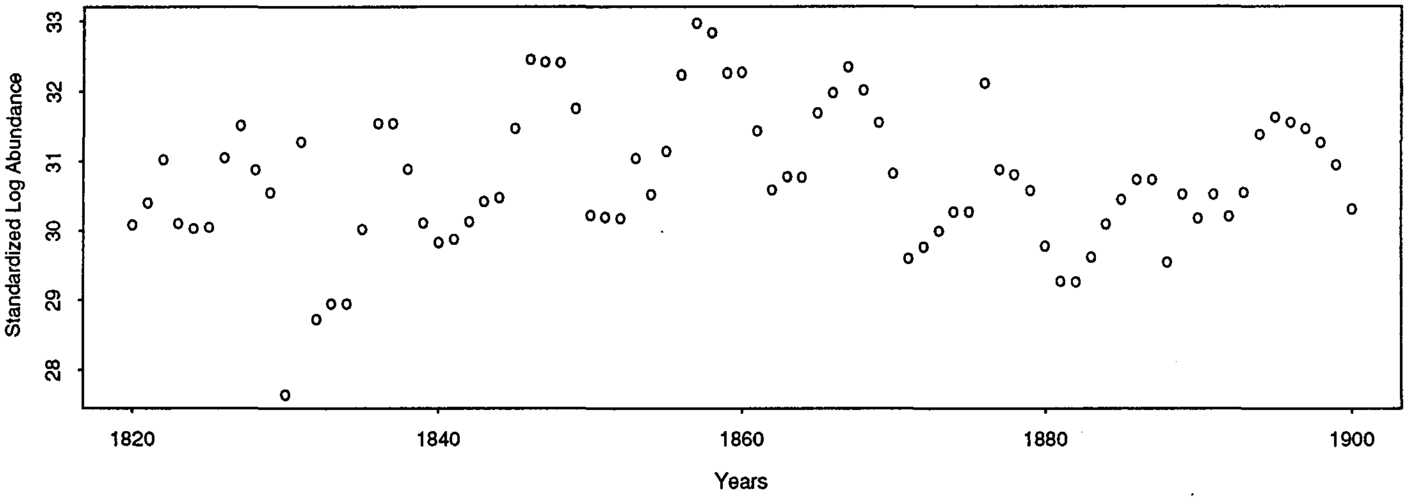


Figure 13a): Lyapunov Exponent estimates GCV C=1

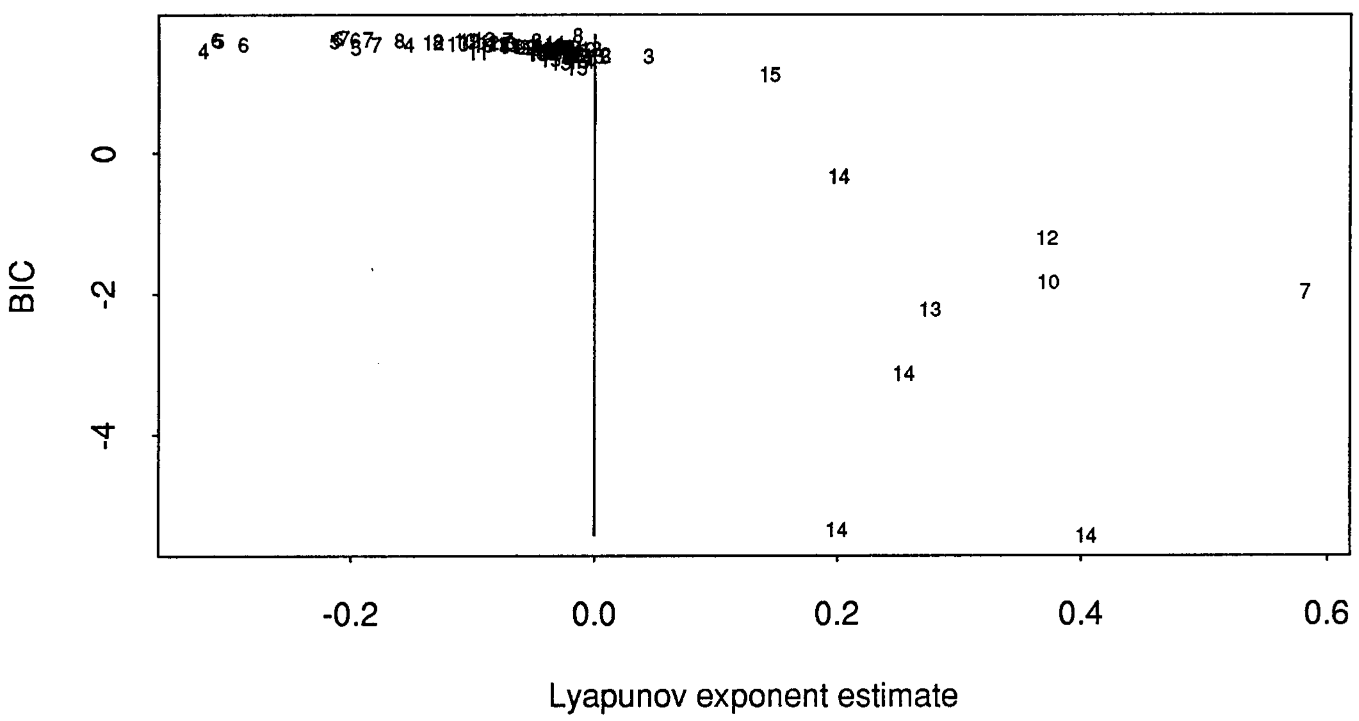


Figure 13b): Lyapunov Exponent estimates GCV C=2

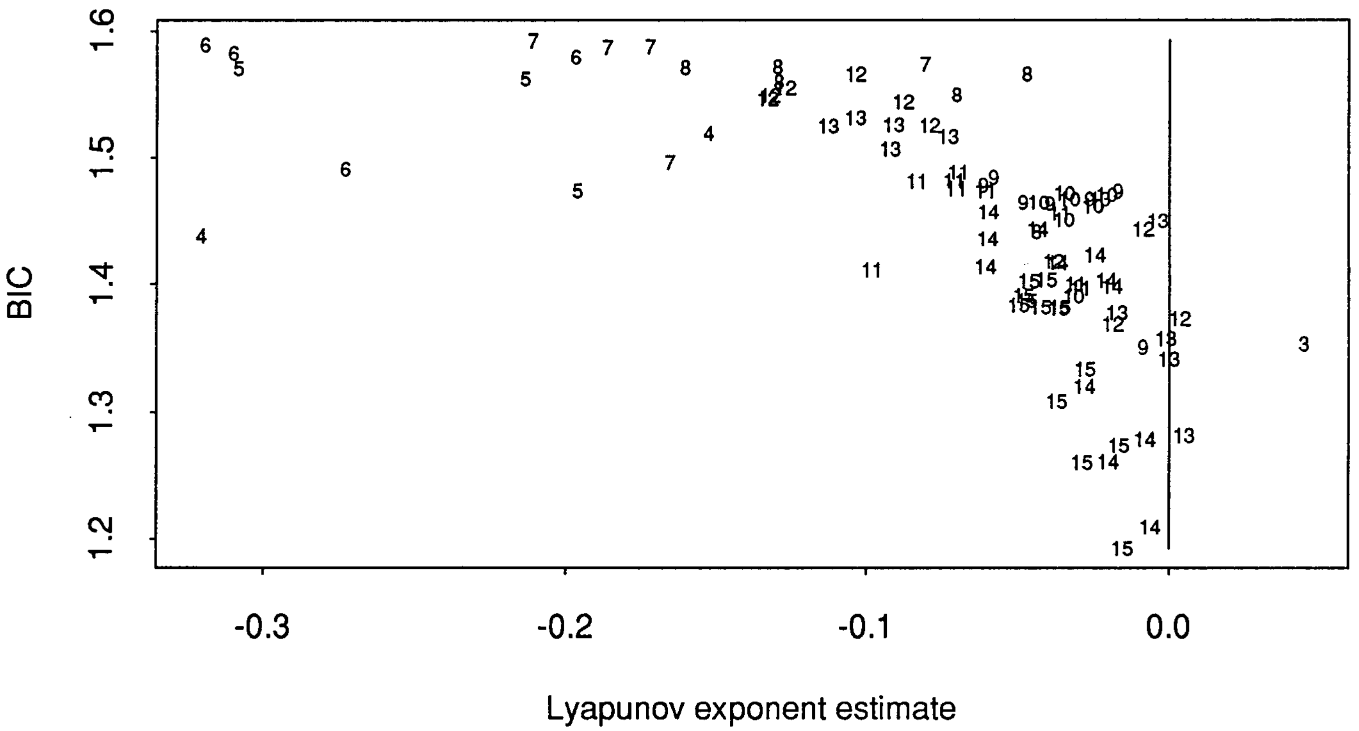


Figure 14a) Neural Net Estimates for Marten Fur Returns as a Function of Embedding Dimension

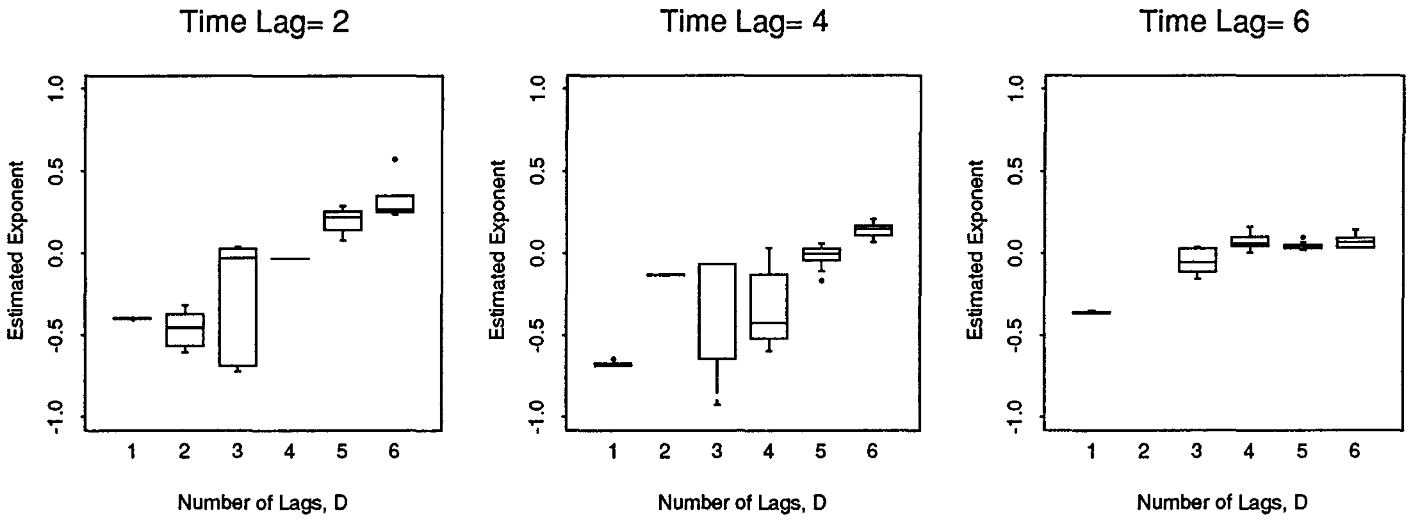
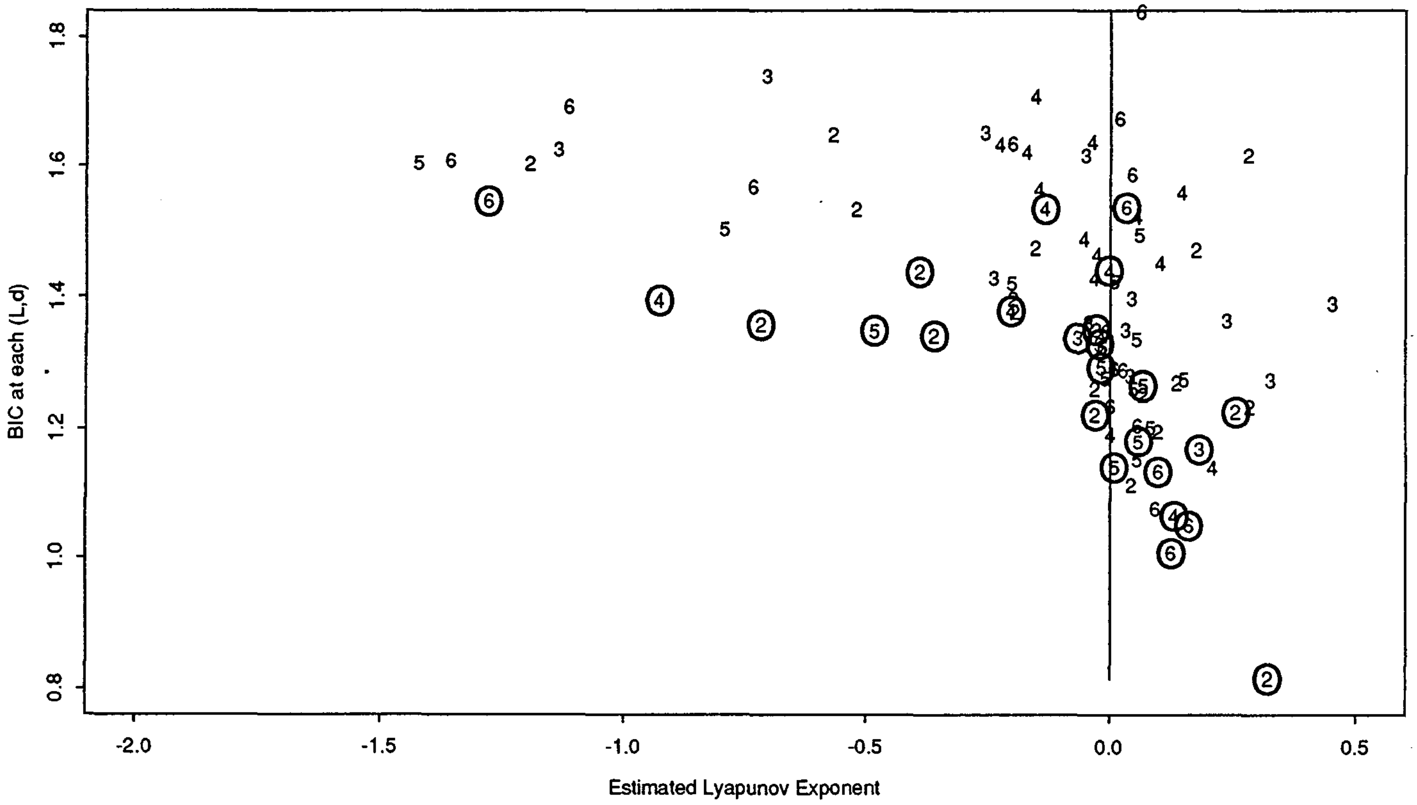


Figure 14b) Relationship of Estimate to BIC


BIC at each (L,d)
Estimated Lyapunov Exponent