

ESTIMATING LYAPUNOV EXPONENTS  
WITH NONPARAMETRIC REGRESSION

Douglas Nychka<sup>1</sup>, Daniel McCaffrey, Stephen Ellner, and A. Ronald Gallant

Institute of Statistics Mimeograph Series No. 1977

July, 1990

# Estimating Lyapunov Exponents with Nonparametric Regression

Daniel McCaffrey, Stephen Ellner, A. Ronald Gallant, and Douglas Nychka<sup>1</sup>

Department of Statistics

North Carolina State University

May 1990.

**Abstract.** We discuss procedures based on nonparametric regression for estimating the dominant Lyapunov exponent  $\lambda_1$  from time-series data generated by a system  $x_t = f(x_{t-1}, x_{t-2}, \dots, x_{t-d}) + \sigma e_t$ , where  $x_t \in \mathbb{R}$ , and  $\{e_t\}$  is an *iid* sequence of random variables. For systems with bounded fluctuations in  $x_t$ ,  $\lambda_1 > 0$  is the defining feature of chaos. We show that any consistent estimator of the partial  $\partial f / \partial x_j$  can be used to obtain a consistent estimator of  $\lambda_1$ . The rate of convergence we establish is quite slow. A better rate of convergence is derived heuristically, and supported by simulations. Simulation results from several implementations, one "local" (thin-plate splines) and three "global" (neural nets, radial basis functions, projection pursuit) are presented for two deterministic ( $\sigma = 0$ ) chaotic systems. Local splines and the neural nets yield accurate estimates of the Lyapunov exponent. However, the spline method is sensitive to the choice of the embedding dimension. Limited results for a noisy ( $\sigma > 0$ ) Henon system suggest that the thin-plate spline and neural net regression methods also provide reliable values of the Lyapunov exponent.

**Key Words and Phrases:** Thin plate smoothing splines, neural networks, projection pursuit regression, chaotic dynamic systems, chaos, nonlinear dynamical systems.

<sup>1</sup> Daniel McCaffrey is a graduate student, Stephen Ellner is an Associate Professor. A. R. Gallant is a Professor and Douglas Nychka is an Associate Professor. All are members of the Department of Statistics, North Carolina State University, Box 8230, Raleigh, NC 27695-8203. The authors would like to acknowledge the support of the National Science Foundation and the North Carolina Experiment Station.

## 1. INTRODUCTION

Nonlinear dynamical systems (e.g., difference or differential equations) can behave in ways that are hard to distinguish from a random process. This phenomenon is called chaos, and it is now recognized as ubiquitous in the nonlinear equations used to model a variety of phenomena including fluid dynamics, chemical reactions, electrical circuits, physiological feedback mechanisms, and disease epidemics (e.g. Buchler and Eichorn 1987, Eckmann and Ruelle 1985, Glass and Mackey 1988, May 1987, Moon 1987, Olsen and Degn 1985, Schuster 1988). Consequently, methods of analyzing experimental or observational data for evidence of chaos have been applied to observational data in such diverse fields as physics, geology, astronomy, neurobiology, ecology, and economics (e.g. Albano et al. 1986, Babloyantz and Destexhe 1986, Brandstater et al. 1983, Brandstater and Swinney 1987, Brock and Sayers 1988, Guckenheimer and Buzyna 1983, Kot et al. 1988, Kurths and Herzog 1987, Mayer-Kress 1986, Mpitosos et al. 1988ab, Ramsay et al. 1989, Sugihara and May 1990). The available methods, developed over the last decade in theoretical physics (see Schuster 1988, Mayer-Kress 1986), are based on calculating a few key quantities that characterize the dynamics: in particular, fractal dimensions and the Lyapunov exponents (defined below). These methods give reliable results if the data are abundant (thousands or tens of thousands of values), if measurement error is near zero, and if the data really come from a deterministic system. With limited data, or a system subject to non-negligible stochastic perturbations, the results may be incorrect or ambiguous (Ruelle 1990).

We present here theoretical and simulation results on procedures to estimate Lyapunov exponents from time-series data, based on nonparametric nonlinear regression. An important new feature of our procedures, is that they are applicable to systems which may have stochastic as well as nonlinear components contributing to the unpredictability. We consider here only the simplest example of such systems, the nonlinear autoregressive model

$$(1.1) \quad x_t = f(x_{t-1}, x_{t-2}, \dots, x_{t-d}) + \sigma e_t,$$

with  $x_t \in \mathbb{R}^1$ ,  $\{e_t\}$  a sequence of *iid* random variables. It is useful to express this system in terms of a state vector  $X_t = (x_t, x_{t-1}, \dots, x_{t-d+1})^T$  and error vector  $\epsilon_t = (e_t, 0, 0, \dots, 0)^T$  in  $\mathbb{R}^d$ , and a function  $F: \mathbb{R}^d \mapsto \mathbb{R}^d$  such that

$$(1.2) \quad X_t = F(X_{t-1}) + \sigma \epsilon_t.$$

Most of the theory focuses on properties of the map  $F$ , and applies to (1.2) without requiring the special structure implied by (1.1).

Lyapunov exponents quantify how perturbations of the state vector affect the subsequent history of the system. For deterministic chaotic systems, the trajectories of the

system starting at two similar state vectors will diverge exponentially, until the trajectories are no longer similar. This is the “sensitive dependence on initial conditions” which is a hallmark of chaotic dynamics, and leads to the apparent unpredictability of chaotic systems. The almost-sure rate of divergence (in the limit of infinitesimally small perturbations), is given by the dominant Lyapunov exponent  $\lambda_1$ . If  $\lambda_1$  is negative rather than positive, nearby trajectories are converging rather than diverging. Thus, a common definition of chaos for deterministic systems is bounded solutions with  $\lambda_1 > 0$  (Eckmann and Ruelle 1985). In systems with stochastic components,  $\lambda_1 > 0$  indicates that the nonlinearity is contributing to the system’s unpredictability, so the value of  $\lambda_1$  is still the best operational indication of chaos.

In Section 2 of this paper, we review the definition and properties of Lyapunov exponents for (1.1) and (1.2). In Section 3 we discuss estimation of  $\lambda_1$  *via* nonparametric regression estimates of  $f$ , including consistency and rates of convergence. Section 4 surveys some nonparametric regression methods with potential value for estimating  $f$  while Section 5 reports the results of a numerical study that apply our procedures to data from chaotic systems. We discuss these numerical results in Section 6. This article establishes a framework for applying statistical analytic tools to an estimation problem which has been of concern for nearly a decade in a variety of scientific disciplines.

## 2. LYAPUNOV EXPONENTS FOR NONLINEAR STOCHASTIC SYSTEMS.

This section outlines the basic theory of Lyapunov exponents by relating the exponents to their more familiar counterparts for linear systems and indicates their significance for characterizing nonlinear dynamics. We also review the methods currently used to estimate Lyapunov exponents, and motivate our approach to the problem.

Lyapunov exponents are a generalization to nonlinear systems of the eigenvalues or roots of linear systems such as the linear difference equation

$$(2.1a) \quad x_t = \sum_{k=1}^d a_k x_{t-k}$$

or the linear autoregressive model

$$(2.1b) \quad x_t = \sum_{k=1}^d a_k x_{t-k} + e_t; \quad x_0, x_1, \dots, x_{d-1} \text{ given.}$$

In state-space form,

$$(2.2) \quad X_t = AX_{t-1} + \sigma e_t$$

where  $A$  is the  $d \times d$  matrix with  $[a_1, a_2, \dots, a_d]$  in the top row, 1’s on the sub-diagonal, and 0’s elsewhere. The characteristic polynomial for (2.2) is

$$(2.3) \quad p(\gamma) = \det(\gamma I - A) = \gamma^d - \sum_{k=1}^d a_k \gamma^{d-k}$$

whose (not necessarily distinct) complex roots  $\{\gamma_i\}_{i=1}^d$  are the eigenvalues of  $A$ .

If  $\sigma=0$ , the qualitative behavior of solutions to (2.2) is related to the magnitudes of the eigenvalues. Suppose for simplicity that the eigenvalues of  $A$  are distinct. Then for initial vector  $X_0 = \sum a_i v_i$ , where  $v_i$  is an eigenvector corresponding to  $\gamma_i$ , the solution to (2.2) is  $X_t = A^t X_0 = \sum a_i \gamma_i^t v_i$ . Hence

$$(2.4) \quad \|X_t\|^{\frac{1}{t}} = \|A^t X_0\|^{\frac{1}{t}} \rightarrow \max\{|\gamma_i| : a_i \neq 0\} \text{ as } t \rightarrow \infty.$$

Since all norms on a finite-dimensional vector space are equivalent, (2.4) holds in any norm.

The *Lyapunov exponents* for (2.2) are the numbers  $\lambda_i = \log|\gamma_i|$ , numbered so that  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ . In terms of the Lyapunov exponents, (2.4) says that

$$(2.5) \quad \frac{1}{t} \log \|X_t\| = \frac{1}{t} \log \|A^t v\| \rightarrow \max\{\lambda_i : a_i \neq 0\} \text{ as } t \rightarrow \infty.$$

The Lyapunov exponents give the possible asymptotic rates of exponential increase (or decrease) for solutions of (2.2) with  $\sigma=0$ . However, the asymptotic growth rate will be  $\lambda_1$  for almost all initial vectors  $v$  (with respect to Lebesgue measure on  $\mathbb{R}^d$ ), with the exceptional vectors lying in a subspace of dimension  $(d-1)$  or smaller. Note that  $\lambda_1$  also gives the growth rate for  $\|A^t\|$ .

The behavior of typical solutions consequently depends on whether the dominant exponent  $\lambda_1$  is positive or negative (unless  $\lambda_1=0$  exactly). If  $\lambda_1 > 0$  (corresponding to a root of (2.3) outside the complex unit circle) typical solutions of (2.1a) will diverge exponentially. If  $\lambda_1 < 0$  all roots lie inside the unit circle, and all solutions of (2.1a) converge exponentially to 0. This qualitative dichotomy carries over to the autoregressive model (2.1b): if  $\lambda_1 < 0$  solutions settle into stationary oscillations with bounded variance, while if  $\lambda_1 > 0$  the variance diverges and there is no convergence to a stationary distribution.

The theory of Lyapunov exponents generalizes these results to stationary random sequences of matrices  $\{A_t\}_{t=0}^{\infty}$ , and the behavior of the corresponding linear system  $X_{t+1} = A_t X_t$ . Let  $H_t = A_{t-1} \dots A_0$ , so that the system's solution is  $X_t = H_t X_0$ . The asymptotic behavior of  $\|H_t\|$  and  $\|X_t\|$  will in general vary over initial vectors  $X_0$  and over realizations of the process  $\{A_t\}$ . Under very general conditions, Oseledec's Multiplicative Ergodic Theorem states that if  $\{A_t\}$  is ergodic, then Lyapunov exponents are well-defined, nonrandom with probability 1, and give the possible rates of growth for  $\log \|X_t\|$ . The Lyapunov exponents can

be defined as

$$(2.6) \quad \lambda_i = \log \left( \lim_{t \rightarrow \infty} |\rho_{i,t}|^{1/2t} \right)$$

where  $\{\rho_{i,t}\}_{i=1}^d$  are the eigenvalues of  $H_t^* H_t$  ( $H^*$  = conjugate transpose of  $M$ ). See Cohen et al. (1986) for a precise statement and proof of Oseledec's Theorem.

We are mainly concerned with the dominant exponent  $\lambda_1$ . The existence of  $\lambda_1$  was first proved by Furstenberg and Kesten (1960): If  $\{A_t\}$  is stationary and ergodic, and  $\log^+ \|A_1\|$  has finite expectation ( $\log^+ x = \max(\log x, 0)$ ), then

$$(2.7) \quad \lambda_1 \stackrel{\text{def}}{=} \lim_{t \rightarrow \infty} \frac{1}{t} \log \|H_t\|$$

exists (as a random variable), and is constant with probability 1. This result is now an easy consequence of Kingman's Subadditive Ergodic theorem (Hall and Heyde 1980), applied to the subadditive process  $x_{st} = \log \|A_{t-1} A_{t-2} \cdots A_s\|$ ,  $s < t$ . As in the deterministic case,  $\lambda_1$  gives the growth rate of "typical" solutions, i.e., for all  $X_0$  lying outside a subspace of  $\mathbb{R}^d$  with dimension  $(d-1)$  or smaller, under the Furstenberg-Kesten assumptions (Cohen et al. 1986).

For nonlinear dynamical systems, consider the following generalization of our basic model (1.2):

$$(2.8) \quad X_t = F_{t-1}(X_{t-1}), \quad X_0 \text{ given,}$$

where  $X_t \in Q \subseteq \mathbb{R}^d$ , and  $\{F_t\}_{t=0}^\infty$  is a sequence of random variables taking values in the space of differentiable functions from  $Q \rightarrow Q$ . Stationary random matrices arise as the sequence of Jacobians for this system (also more generally for dynamics on abstract spaces: see Arnold & Wihstutz 1986, Kifer 1986).

Let  $X_t^{(1)}$  and  $X_t^{(2)}$  be solutions of (2.8) which differ only in their initial values, i.e.

$$(2.9) \quad X_t^{(i)} = F_{t-1}(X_{t-1}^{(i)}), \quad i=1,2, \quad X_0^{(i)} \text{ given.}$$

If the initial perturbation  $Y_0 = (X_0^{(1)} - X_0^{(2)})$  is sufficiently small, then the subsequent separation  $Y_t = (X_t^{(1)} - X_t^{(2)})$  can be approximated by linearization,

$$(2.10) \quad Y_t = F_{t-1}(X_{t-1}^{(1)}) - F_{t-1}(X_{t-1}^{(2)}) \approx J_{t-1} Y_{t-1},$$

where  $J_t = DF_t(X_t)$ , the Jacobian matrix of  $F_t$  at  $X_t$ . In the formal limit of infinitesimally small  $Y_0$ , (2.10) becomes exact and the solution is

$$(2.11) \quad Y_t = J_{t-1} J_{t-2} \cdots J_0 Y_0.$$

The Lyapunov exponents for (2.8) are defined to be the Lyapunov exponents for the linear system (2.11), whenever they exist. In particular, it must be assumed that the Markov chain (2.8) has an invariant measure  $\nu_1$ . If  $X_0$  is chosen at random from the distribution  $\nu_1$ , then  $\{J_t\}$  is a stationary process and under suitable assumptions on  $F_t$  the Multiplicative Ergodic Theorem will apply to  $\{J_t\}$  (see Kifer 1986, Chapter III; the pair  $(F, DF)$  are a “random bundle map” as defined by Kifer).  $\lambda_1 > 0$  indicates “sensitive dependence on initial conditions” in (2.8), since small perturbations to  $X_0$  will grow exponentially over time. As we noted above, this is a key indicator of chaotic dynamics, and is responsible for the extreme unpredictability of chaotic systems.

### 3.1 DIRECT METHODS FOR ESTIMATING THE LYAPUNOV EXPONENTS

There are two classes of methods in use for estimating  $\lambda_1$  from experimental or observational data. “Direct” methods, proposed by Guckenheimer (1985) and successfully implemented by Wolf et al. (1985), are based on the assumption that  $\|Y_t\|$  in (2.10) will grow exponentially at rate  $\lambda_1$ . The available data are searched for pairs of times  $(t_1, t_2)$  when  $\|X_{t_1} - X_{t_2}\|$  is sufficiently small. Then the growth of  $\tilde{Y}_t = \|X_{t_1+t} - X_{t_2+t}\|$  is recorded until the trajectories diverge beyond some pre-set limit (or, for some pre-set amount of time). The data are then searched for a replacement point  $X_{t_3}$  which is sufficiently near  $X_{t_1+t}$  and also sufficiently near the line connecting  $X_{t_1+t}$  and  $X_{t_2+t}$ , and the process is repeated. The average divergence rate over the entire data set is the estimate of  $\lambda_1$ .

In “Jacobian” methods, the data are used to estimate the Jacobians  $DF(X_t)$ , and  $\lambda_1$  is calculated from the estimated Jacobians. Our procedure falls into this category. Jacobian methods were proposed by Eckmann and Ruelle (1985), and an implementation based on linear regression was tested by Eckmann et al. (1986). To estimate  $DF(X_t)$ , Eckmann et al. (1986) identify all data points  $X_s$  lying within some specified distance of  $X_t$ . The approximation

$$X_{s+1} - X_{t+1} = F(X_s) - F(X_t) \approx DF(X_t)(X_s - X_t).$$

justifies a linear model  $X_{s+1} - X_{t+1} \approx A_t(X_s - X_t)$ , and the  $A_t$  from a least-squares fit is the estimate for  $DF(X_t)$ . The  $A_t$ 's are substituted into equation (2.6). Let  $\nu_i$  denote the  $i^{\text{th}}$  largest eigenvalue of  $H_N^* H_N$  then we obtain the estimated exponents  $\hat{\lambda}_i = (1/2N) / \log(\nu_i)$ .

In tests with simulated data from low-dimensional systems, the direct and linear Jacobian methods both appear to be adequate if there is an abundant supply of very accurate

data, the system is deterministic or nearly so ( $\sigma \approx 0$ ), and the correct dimension (number of lags in the model) is known (Mayer-Kries 1986). Since the correct dimension is unknown, a common strategy is to increase the dimension until the estimates of  $\lambda_1$  reach a plateau. The linear Jacobian method can generate spurious exponents when the dimension is too large, so this strategy may fail (see Vastano and Kostelich 1986). Eckmann et al. (1986) claim that the spurious exponents can be eliminated by proper choice of the method's free parameters, but it appears that discrimination between "real" and "spurious" solutions is essentially *ad hoc* and requires much trial-and-error tuning of the free parameters.

However, the direct methods are less suited to mixed stochastic/nonlinear dynamics. It is easy to see that direct method will be positively biased when  $\sigma > 0$ , and that measurement error will also inflate the estimate of  $\lambda_1$ . To eliminate these biases, the divergences will have to be fit to a more complicated growth model, which allows for the stochastic dynamics. If a model such as (1.1) is believed to represent the dynamics, then the appropriate model for the divergences would involve complicated  $X_t$  and  $Y_t$  dependence in the growth of  $Y_t$ . Consequently, we believe that Jacobian methods are preferable for stochastic nonlinear systems, but must either incorporate a criterion for choosing the dimension, or else employ an estimator for  $DF(X_t)$  that is less sensitive to extraneous lags in the model. Procedures for fitting nonlinear time-series models to chaotic data have been very successful at generating estimators of  $F$  (Abarbanel et al. 1989, Casdagli 1989, Farmer and Sidorowich 1987, 1988a, 1988b) despite *ad hoc* fitting methods. These successes were the motivation for our approach, in which the Jacobian estimates are derived from nonlinear estimation of  $F$ .

### 3.2 ESTIMATING $\lambda_1$ FROM THE MAP JACOBIANS.

We consider in this section the properties of Jacobian-method estimates for  $\lambda_1$  that are based on nonlinear regression estimates of  $f$  (for equation 1.1) and/or  $F$  (for equation 1.2). Under reasonable assumptions about the underlying process (1.1) and the regression estimates, we derive a rigorous upper bound for the estimation error. We also give a conjecture for the asymptotic behavior for the error.

Let  $\hat{F}$  denote an estimate of  $F$  based on  $N$  observations  $\{x_t\}_{t=1}^N$  of the time series, and let  $\hat{J}_k$  denote the matrix obtained by substituting estimated for exact partial derivatives,  $\hat{J}_k = D\hat{F}(X_k)$ . We require that the estimate  $\hat{J}_k$  be consistent, specifically:

**Assumption 1.** *There exists  $\beta_N \rightarrow 0$  such that  $\max_{1 \leq t \leq N} \|\hat{J}_t - J_t\| = O_p(\beta_N)$  as  $N \rightarrow \infty$ .*

From the theory of nonparametric regression,  $\beta_N$  can be expected to converge to zero at a rate of the form  $N^{-\delta}$ , where  $\delta$  depends on the form of the nonparametric estimate and the differentiability of  $f$ . For example if  $\hat{f}$  is the  $m^{\text{th}}$  order thin plate spline in  $d$  dimensions,  $f$  has  $m$  continuous partial derivatives and  $\sigma > 0$  in (1.1), then  $\delta \geq \frac{2m-2}{2m+d}$  (Cox 1984). Moreover, in the case that noise is not present ( $\sigma=0$ ), one can expect that  $\delta > 1$ .

A second requirement is ergodicity, so that a single realization of the process represents (with probability 1) the “typical” behavior. Formally,

**Assumption 2.** *The Markov chain (1.2) has a unique invariant measure  $\mu$ , and  $X_0$  is chosen by sampling from this measure. The stationary process  $\{J_t\} = \{DF(X_t)\}$  is ergodic, and  $\int \log^+ \|DF(x)\| d\mu(x) < +\infty$ .*

Under these assumptions, the sequence of estimated jacobians  $\{\hat{J}_t\}$ ,  $1 \leq t \leq N$  can be used to obtain consistent estimates of  $\lambda_1$ . To simplify notation, for the remainder of this section we write  $\lambda$  for  $\lambda_1$ . Also for any matrix  $A$  let  $\|A\|^2 =$  largest eigenvalue of  $A^T A$ . It will be necessary to distinguish between the “sample size”  $N$  (number of observations) used to estimate  $\hat{J}_t$ , and the “block length”  $M$  which will be the number of matrices  $\hat{J}_t$  used in estimating  $\lambda$ . Let  $\gamma_M = \|T_M\|^{1/M}$ , where  $T_M = J_M J_{M-1} \cdots J_1$ , and let  $\hat{\gamma}_M, \hat{T}_M$  be the estimates obtained by using  $\hat{J}$ 's in place of  $J$ 's. Then  $\lambda = \lim_{M \rightarrow \infty} \log(\gamma_M)$  with probability 1, and  $\hat{\lambda}_M = \log(\hat{\gamma}_M)$  is the Jacobian estimator for  $\lambda$  based on sample size  $N$ , and block length  $M$ . Define  $\gamma$  by  $\lambda = \log \gamma$ . The main error bounds are summarized by the following theorem.

**Theorem 3.1** *Under assumptions 1 and 2.*

$$(3.1) \quad \hat{\lambda}_M - \lambda_M = O_p(\beta_N^{1/M}) \text{ as } N, M \rightarrow \infty.$$

Let  $\alpha = \sup_{k \geq 1} \|J_k\|$  and suppose that there is a  $0 < \rho < \gamma$  such that

$$M \beta_N(\alpha/\rho)^M \rightarrow 0 \text{ as } M, N \rightarrow \infty.$$

Then

$$(3.2) \quad \hat{\lambda}_M - \lambda_M = O(\beta_N(\alpha/\rho)^M) \text{ as } M, N \rightarrow \infty.$$

The proof of these results is given in the Appendix.

For accurate estimation of  $\lambda_1$ ,  $M$  should be taken as large as possible. Subadditivity of  $x_{st}$  (as defined in Section 2) has the unfortunate implication that  $\gamma_M \geq \gamma$  with strict inequality

unless  $x_{st}$  is actually additive. Additivity is not to be expected in general, since  $x_{st}$  is additive only if  $\|J_r J_{r-1} \cdots J_k\|$  and  $\prod_{i=k}^r \|J_i\|$  have the same distribution for all  $0 \leq k \leq r < \infty$ .

Unfortunately, the error bounds in Theorem 3.1 imply that the growth in  $M$  must be unattractively slow in order to guarantee that  $|\hat{\lambda}_M - \lambda_M| \rightarrow 0$ . If  $\beta_N = \mathcal{O}(N^{-\delta})$  with  $\delta < 1$  as expected, then  $M = \mathcal{O}(\log N)$  is necessary. However, we conjecture that (3.1) and (3.2) are overly pessimistic, and the actual error is sufficiently small so that  $M = o(N^\delta)$  implies  $\hat{\lambda}_M \rightarrow \lambda$ .

**Conjecture:**

$$(3.3) \quad \lambda_M - \hat{\lambda}_M = \mathcal{O}_P(\beta_N) \quad \text{as } N, M \rightarrow \infty \text{ such that } M\beta_N \rightarrow 0.$$

**Rationale:** Expanding the leading-order term in  $T_M - \hat{T}_M$  ( see A.1 and A.2) we see that it is the sum of  $M$  products with the form

$$J_M J_{M-1} \cdots J_{k+1} (J_k - \hat{J}_k) J_{k-1} \cdots J_1.$$

Since  $\|J_k - \hat{J}_k\| = \mathcal{O}(\beta_N)$  and  $\|J_i\| \leq \alpha$  by assumption, these terms together are  $\mathcal{O}(M\alpha^{M-1}\beta_N)$ , which is the rate obtained rigorously in the proof of Theorem 3.1. However, except for a vanishingly small fraction of such products as  $M \rightarrow \infty$ , both  $k$  and  $M-k$  will be large.

Consequently, the asymptotic growth rates

$$(3.4) \quad \|J_M J_{M-1} \cdots J_{k+1}\| \sim \gamma^{(M-k)}, \quad \|J_{k-1} \cdots J_1\| \sim \gamma^{(k-1)}$$

should be sharper than the crude upper bounds used in the proof of (3.1), which have  $\alpha$  in place of  $\gamma$ . Assuming that (3.4) applies simultaneously to all terms in (A.2), and repeating the arguments used in the proof of Theorem 3.1 we obtain an argument to support the conjecture.

### 3.3 SIMULATIONS OF ESTIMATION ERROR

Some simulations supporting this conjecture are summarized in Figure 1. The “data” were generated by the Henon map (Schuster 1988)

$$(3.5) \quad x_t = 1 - 1.4x_{t-1}^2 + 0.3x_{t-2}$$

and the block size was varied over the range  $50 \leq M \leq 1000$ .

The “estimation error” matrices  $U_t$  were simulated by choosing all entries independently from the uniform distribution on  $[-\beta/2, \beta/2]$  where  $\beta = 5/\sqrt{M}$ . Uniform errors were chosen because  $\|U_t\| \leq \beta$  in the  $L_2$  norm. Because (3.3) is based on convergence in probability rather than almost-sure convergence, we compared the right hand side of (3.3) to the mean, the 95<sup>th</sup> percentile, and the maximum of  $|\hat{\lambda}_M - \lambda_M|$  over 1000 replicated values of  $\lambda_M$  and  $\hat{\lambda}_M$ . The bound (3.3) is well above the mean error, and mostly well above the 95<sup>th</sup> percentile. However, the maximum error was larger than the 95<sup>th</sup> percentile by an order of magnitude, and was not

bounded by (3.3).

The results above suggest that  $M \ll N$  may be necessary for obtaining a consistent estimate of  $\lambda_M$ . This raises the issue of optimal block length for estimating  $\lambda$ . In practice, one would presumably obtain  $\hat{\lambda}_M$ 's for each of the roughly  $N/M$  blocks and use their average as an estimate of  $\lambda$ . The error  $(\hat{\lambda}_{M,n} - \lambda)$  for block  $n$  then consists of

$$\begin{aligned} \text{Block bias:} & \quad b(M) = E(\lambda_M) - \lambda \\ \text{Block error:} & \quad \nu(n) = \lambda_{M,n} - E(\lambda_M) \\ \text{Estimation error:} & \quad \epsilon(M, n) = \hat{\lambda}_{M,n} - \lambda_{M,n} \end{aligned}$$

where  $E(\cdot)$  refers to the stationary measure for the true system.

Optimizing the choice of  $M$  requires information on the convergence rate for each of these error components. The block error and bias are properties of the stationary random matrix product generated by the true system. Unfortunately, the rates of convergence of these components can only be conjectured.

The most reasonable hypothesis for the block bias is

$$(3.6) \quad E(\lambda_M) - \lambda = O(1/M).$$

This is known to hold for positive matrices which satisfy a mixing condition (Heyde 1986), but the Jacobians for a nonlinear system with bounded trajectories cannot be positive everywhere.

The only reasonable conjecture for the block error is

$$(3.7) \quad \text{Std Dev}(\lambda_M) = O(1/\sqrt{M}).$$

The rationale for (3.7) is that  $\lambda_M$  can be written as  $\frac{1}{M} \sum_{j=1}^M \phi(X_j, S_j)$  where  $S_j = T_{j-1} / \|T_{j-1}\|$

(Furstenberg and Kesten 1960), and in our case  $\phi(X, S) = \log$

$\|DF(X)S\|$ . Since  $(X_j, S_j)$  is a Markov chain, central limit behavior is expected for  $\lambda_M$ . Such results are available for independent random matrices (Bougerol, P. and Lacroix J. 1986) and for positive matrices under mixing assumptions (Heyde and Cohen 1985). It will generally be reasonable to assume that  $\{X_t\}$  is a Harris-recurrent chain, hence the Jacobian matrices  $\{J_t\}$  will strong-mixing (Athreya and Pantula 1986).

Both (3.6) and (3.7) are consistent with simulation results for the Henon map (3.5) and also for the model

$$(3.8) \quad x_t = x_{t-1} + 10.5 \left[ \frac{0.2 x_{t-5}}{1 + (x_{t-5})^{10}} - 0.1 x_{t-1} \right].$$

Equation (3.8) is a discretized analog of the Mackey-Glass delay differential equation (Schuster 1988), which we devised to mimic financial market data. The results for the Henon map are summarized in Figure 2, and the results for (3.8) were very similar and thus are not reported.

Linear regression fits to (3.6) and (3.7), had  $r^2 > 0.98$  in all cases.

When the block size is smaller than the data series length several estimates of  $\lambda$  can be calculated based on nonoverlapping time intervals. In this case it seems most reasonable to take the average of these estimates as the overall estimate of  $\lambda$  for the system. Suppose that  $N$  factors as  $MB$  and let  $\hat{\lambda}(M, N)$ , denote the average estimate over  $B$  blocks. The block and estimation errors would presumably be reduced by order  $1/\sqrt{B}$  although the block bias would remain on the order of  $1/M$ . Adding this to our roster of conjectures and appropriately modifying (3.3) and (3.7), it appears that the block bias component of the overall error has the slowest asymptotic convergence rate. Determining the block bias for the estimated model, and subtracting it from  $\hat{\lambda}$ , might be helpful in reducing the overall bias.

#### 4. METHODS FOR ESTIMATING THE JACOBIAN OF THE MAP

From the theoretical results in Section 3 the consistency of the estimated Lyapunov exponent depends on the nonparametric estimates of the map.

In this section we describe four different approaches for estimating the map. These methods can be classified as to whether they give a global approximation to the map or are based on a local approximation to  $F$  at the lagged values of  $X$ . One important aspect of these methods is their ability to support a range of smoothing situations. This would include interpolating the observed data in a situation when the system is deterministic ( $\sigma=0$ ) to smoothing the data when there is a significant random component ( $\sigma \gg 0$ ) in (1.1). The introduction of some form of data smoothing is crucial for accurate derivative estimates based on noisy data.

The last part of this section discusses the relationship between the accuracy of the exponent estimate and mean squared prediction error.

##### 4.1 LOCAL THIN PLATE SPLINES

Spline functions have been identified by numerical analysts as an efficient and accurate way of approximating complicated functions. The fundamental definition of a spline is as the solution to a variational problem. For example consider the one dimensional curve fitting problem based on  $(x_t, y_t)$   $1 \leq t \leq N$  where in analogy to (1.1),  $y_t = f(x_t) + \sigma e_t$ . For a fixed value of  $\mu > 0$ , and let

$$J(h) = \int_{\mathbb{R}^1} [h^{(2)}(u)]^2 du.$$

A cubic smoothing spline approximation to  $f$  is defined as the function that minimizes:

$$(4.1) \quad \mathcal{L}(h) = \sum_{t=1}^N [Y_t - h(X_t)]^2 + \mu J(h)$$

over all  $h$  such that  $J(h) < \infty$ . The second component of  $\mathcal{L}$  can be interpreted as a measure of the amount of curvature or roughness of  $h$ . Thus the smoothing parameter,  $\mu$ , controls the relative weight between fitting the data well and the smoothness of the resulting spline. When measurement error is not present ( $\sigma=0$ ), the spline estimate should interpolate the data. This can be achieved by considering the limiting smoothing spline obtained as  $\mu \rightarrow 0$ . This limit will also be the solution to the minimizing  $J(h)$  subject the interpolation constraint:  $Y_t = h(x_t)$   $1 \leq t \leq N$ .

It should be emphasized that the quality of the resulting spline function depends strongly on the choice of  $\mu$  and it is good practice to investigate the sensitivity of the estimate to this parameter.

The generalization of splines for multivariate approximation involves replacing the one-dimensional roughness measure by a multivariate measure of curvature. Suppose  $h: \mathbb{R}^d \rightarrow \mathbb{R}$  and all mixed partial derivatives of  $h$  up to order  $m$  are contained in  $L^2(\mathbb{R}^d)$ . We will consider:

$$J_{m,d}(h) = \sum_{\alpha_1 + \dots + \alpha_d = m} \binom{d}{\alpha_1 \dots \alpha_d} \int_{\mathbb{R}^d} \left[ \frac{\partial^m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} h(x) \right]^2 dx$$

Unlike the one-dimensional roughness measure,  $J_{m,d}$  involves derivatives of higher order than second degree. This generalization is necessary to guarantee that the spline approximate will be consistent for higher dimensions. Although  $J_{m,d}$  has a complicated form, the mixed partials enter in a way so that this curvature measure is invariant to rotation of the coordinates axes. Replacing  $J$  by  $J_{m,d}$  in (4.1) a thin plate spline is defined as the minimizer of  $\mathcal{L}$  over all  $h$  such that  $J_{m,d}(h) < \infty$ . The same interpolation properties hold when  $\mu \rightarrow 0$ .

Although spline functions are defined abstractly as the solution to a minimization problem, they are readily computable. The solution will be a linear combination of the  $\binom{d+m-1}{d}$  monomials of degree less than  $d$  and a set of  $N$  radial basis functions (Wahba 1989). The coefficients for the spline are found by solving a system of  $N$  linear equations and for the special case of one-dimensional cubic smoothing splines efficient algorithms ( $\mathcal{O}(N)$ ) are available.

Thin plate spline approximations were used to give to give local estimates of the Jacobian of the map. The estimate at any particular point  $X_t$  was based on the  $L$  nearest state vectors with respect to Euclidian distance. There are two reasons for considering a local estimate rather fitting a single thin plate spline to the entire times series. The storage requirement for computing a thin plate spline is on the order of  $N^2$  and becomes prohibitive for the sample sizes typically encountered in the study of chaotic systems. Also, a global spline function uses a single smoothing parameter to smooth all parts of the surface. This is not desirable if the actual curvature of the map varies. A local fit to the surface has the potential to adapt to varying curvature and sharpen the accuracy of the approximation. One disadvantage of local estimates is that the variance of the resulting estimates are inflated because the effective sample size has been significantly reduced. In order for these local estimates to be consistent the number of nearest neighbors used to calculate the spline must increase as  $N \rightarrow \infty$ .

### 4.2 RADIAL BASIS FUNCTIONS

The form of a thin plate spline as a linear combination of polynomials and radial basis functions suggests a global approximation to the map. For any pair of  $m$  and  $d$  such that  $2m-d > 0$ , let  $\{\phi_j(\underline{y})\}$   $1 \leq j \leq \binom{d+m-1}{d}$  denote the set of all monomials of degree less than  $d$ . Let  $\{U_k\}$  for  $1 \leq k \leq K$  be a set of vectors in  $\mathbb{R}^d$  and consider the radial basis:

$$\psi_k(\underline{x}) = \begin{cases} \|\underline{x} - U_k\|^{2m-d} & \text{for } d \text{ odd} \\ \|\underline{x} - U_k\|^{2m-d} \log(\|\underline{x} - U_k\|^{2m-d}) & \text{for } d \text{ even} \end{cases}$$

Thus the estimate of  $f(\underline{X})$  has the form:

$$(4.2) \quad \hat{f}(\underline{X}) = \sum_j d_j \phi_j(\underline{X}) + \sum_k c_k \psi_k(\underline{X})$$

and the coefficients for this approximation can be computed by least squares. Note that each  $\psi_k$  will be a bowl-shaped function centered at the point  $U_k$ . The locations of these points should be adapted to the data and for our application this set was taken as a random sample from  $\{X_t\}$ . In this way the coverage of the basis follows the density of the the  $X$ 's.

### 4.3 NEURAL NETS

The use of neural nets to approximate the map of a chaotic process were found to be competitive with the best approximation methods studied by Casdagli (1989) and performed significantly better than several other methods considered by Lapedes and Faber (1987). We used a single hidden layer feedforward neural network. The functional form for this configuration is

$$\hat{f}(\underline{X}) = \sum_{j=1}^K \beta_j G(\gamma_j^T \underline{X} + \mu_j).$$

where  $G(u) = e^u / (1 + e^u)$  is the logistic distribution function and  $\gamma_j \in \mathbb{R}^d$ . The parameters are estimated by nonlinear least squares and the resulting function yields a global approximation to  $f$ . In comparison to the preceding functional approximations, the neural net form is not sensitive to increasing  $d$ . Although the lengths of the vectors  $\gamma_j$  increases the functional form remains a sum of simple univariate functions. This property contrasts sharply with the complexity of the thin plate spline where the number of polynomial terms grows exponentially with  $d$  and the spline order ( $m$ ). The quality of this approximation is influenced by the number of logistic functions in the sum. The results in Gallant and White (1989) however

suggest that performance may not be sensitive to the choice of  $K$  provided that the saturation ratio  $K(d+2)/N$  is approximately  $1/30$  and  $500 \leq N < 5000$ .

#### 4.4 PROJECTION PURSUIT

Projection pursuit approximation is an ambitious attempt to combine the projective power of neural nets with the flexibility of spline estimates. It was first proposed in the context of identifying interesting low dimensional structure in a high dimensional data set ( Friedman and Stuetzle 1981). Although this method yields a global approximation, storage requirements are minimal because the estimate is computed by solving a sequence of univariate problems. Note that if  $\alpha \in \mathbb{R}^d$  and  $\|\alpha\| = 1$  then  $(\alpha^T X)\alpha$  defines the projection of  $X$  onto the one dimensional subspace spanned by  $\alpha$ . Given  $K$  such projections indexed by the vectors  $\{\alpha_k\}$ ,  $1 \leq k \leq K$  then the projection pursuit approximation has the form:

$$(4.4) \quad h(X, \alpha_1, \dots, \alpha_K) = \sum_{k=1}^K G_k(\alpha_k^T X).$$

Unlike the neural net representation, the  $G_k$ , known as ridge functions, do not follow a rigid parametric form and are estimated. They are defined as the solution to minimizing

$$(4.5) \quad \sum_{t=1}^N [Y_t - h(X_t, \alpha_1, \dots, \alpha_K)]^2 + \mu \sum_{k=1}^K J(G_k)$$

over all  $G_k$  such that  $J(G_k) < \infty$ . In this manner any given set of projections imply a particular approximant,  $h$ . The specific set of projections are found by minimizing:

$$(4.6) \quad \sum_{t=1}^N [y_t - h(X_t, \alpha_1, \dots, \alpha_K)]^2$$

over all  $\{\alpha_k\} 1 \leq k \leq K \subset \mathbb{R}^d$ .

The projections and ridge functions are calculated by an iterative process known as the backfitting algorithm. The main feature of this algorithm is that the multivariate minimization of (4.5) and (4.6) is broken up into a series of one dimensional estimation problems. For example, suppose that  $\{G_k, \alpha_k\}$  for  $k \neq j$  are known. Given  $\alpha_j$  one can determine  $G_j$  based on fitting a one-dimensional smoothing spline to the regression data  $(\alpha_j^T X_t, y_t - \sum_{k \neq j} G_k(\alpha_k^T X_t))$   $1 \leq t \leq N$ . The idea of the backfitting algorithm is to estimate each ridge function sequentially. One then loops back and reestimates the individual ridge functions until the complete estimate converges. If the projections are kept fixed, the backfitting algorithm is related to the Gauss-Seidel method for solving linear systems ( Buja, Hastie and Tibishirani 1989) and the resulting

limit from backfitting will be the minimizer to (4.5). The complication for computing the projection pursuit approximation, however, is that the projections are also optimized in the backfitting steps. An outline of the algorithm used here can be found in Nychka (1988).

#### 4.5 PREDICTION ERROR OF THE ESTIMATED MAP

One measure of the accuracy of the map estimate is prediction error. Let  $\{x_s\}_{d \leq s \leq S}$  denote a time series generated according to (1.1) but independent of  $\hat{f}$ . An estimate of the error in using the estimated map to predict a subsequent observation of the series is given by

$$\hat{\sigma}^2 = \frac{1}{S} \sum_{s=1}^S [x_s - \hat{f}(X_{s-1})]^2.$$

Under Assumption 2 from Section 3 and the independence of the error process it follows that

$$(4.7) \quad \hat{\sigma}^2 \rightarrow \int_{\mathbb{R}^d} [f(X) - \hat{f}(X)]^2 d\mu(X) + \sigma^2 \quad \text{as } S \rightarrow \infty.$$

Thus for large  $S$ ,  $\hat{\sigma}^2$  differs by a constant from the integrated squared error of  $\hat{f}$  over the attractor.

The integrated squared error (ISE) is a useful measure of the closeness of  $\hat{f}$  to  $f$  because it can be estimated directly from the data by cross-validation. One problem with this norm is that it is not strong enough to imply convergence of the Lyapunov exponent. For example, it is possible that the ISE can converge to zero as  $N \rightarrow \infty$  but  $\hat{J}$  may not converge to the Jacobian matrix of  $F$ .

A practical concern is that “tuning” parameters of the map estimates that are selected by minimizing  $\hat{\sigma}^2$  may not yield the best values for estimating  $\lambda_1$ . For example, the embedding dimension is usually unknown in practice and an incorrect value for  $d$  will lead to biased estimates of the Lyapunov exponents. According to general theory for systems without a random component (Casdagli 1989),  $\hat{\sigma}^2$  should be large for small choices for  $d$  and decrease to zero and remain small as  $d$  is increased beyond the minimum embedding dimension. This property suggests that the embedding dimension might be chosen to minimize  $\hat{\sigma}^2$ . Note that by (4.7) minimizing  $\hat{\sigma}^2$  over  $d$  is asymptotically equivalent to minimizing the ISE. From the remarks above this may not be the appropriate choice for using the map to estimate  $\lambda_1$ . One reason for the numerical study was to investigate the relationship between the best choice for  $d$  and the dimension minimizing ISE.

## 5 SIMULATION RESULTS

A simulation was carried out to study the feasibility of estimating the largest Lyapunov exponent for moderate sample sizes. For a sample size in the range of 2000-2500 and no measurement error we evaluated the performance of the four approximation methods described in Section 4 ( Local Spline, Radial Basis, Neural Net, Projection Pursuit) for two simple chaotic systems (Henon, Rossler). Because no measurement error is present we can expect that  $\beta_N = o(1/N)$ . Under the assumption that the conjecture in Section 3.2 is true one may use  $M=N$  for  $\hat{\lambda}_1$  and still obtain a consistent estimate. To keep the results simple, this is what was done. The estimates of  $\lambda_1$  for these cases are summarized in Table 5.1 by the number of lags ( $d$ ) used in the approximation. The asterisk indicates the estimate associated with the smallest  $\hat{\sigma}^2$ .

A second set of simulations were run to investigate the sensitivity of the estimates when a random component is present. Attention was restricted to the Henon map and two approximation methods (Local splines and Neural Nets). Because the model is not deterministic, the consistency results from Section 3 suggest that the block size  $M$  must be smaller than the total number of observations. Accordingly, estimates of the Lyapunov exponent were calculated based on several block sizes (  $N=2000$ ,  $M=50, 100, 500, 2000$ ). The results of these simulations are reported in Table 5.3. A more useful summary of these results however is given by Figure 3. Here boxplots of the estimates of  $\lambda_1$  indicate how the distribution depends on  $M$ , the number of lags (embedding dimension) and the nonparametric method.

The remainder of this Section gives details concerning the different levels for the factors and ends with an overview of the simulation results.

### 5.1 HENON AND ROSSLER SYSTEMS

The Henon map is given at (3.5) and it is evident that it has the form of (1.1). The largest Lyapunov exponent is approximately  $.419 \pm .001$ .

The other system considered is derived from the Rossler continuous time, system of equations:

$$\begin{aligned}\dot{x} &= -(y+z) \\ \dot{y} &= x + .15y \\ \dot{z} &= .2 + z(x-10)\end{aligned}$$

These equations were numerically integrated with a fixed time step of  $\Delta t = .01$  and  $x$  was

sampled at every 50 steps. Although this time series is the result of sampling a single component of a continuous system, Takens' embedding theorem (Takens 1981) indicates that for  $d$  sufficiently large there exists an  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $x_t = f(x_{t-1}, \dots, x_{t-d-1})$ . Thus the Rossler time series also follows (1.1). Although the form of  $f$  is unknown, the value of  $\lambda_1$  is known to be approximately .04505.

To generate a sample from either of these systems first 5000 values were calculated based on a randomly selected starting values. The next 20000 values were taken to be points on the attractor of the map. To create a sample of size  $N$  the starting value  $t_0$  was randomly selected in the range  $[1, 20000-N]$  and the resulting index was taken as the starting point of the time series.

For the second simulation study, the random component added to each iteration of the Henon map was a uniform random variable on the range  $[-.012, .012]$ . This range was chosen because empirically it was found that perturbations with a larger variance eventually caused the Henon system to become unstable. From a qualitative point of view the shape of the attractor for this noisy system had the same overall shape as the strange attractor for the deterministic system. (Of course the fractal properties of these two sets are very different.) Since the invariant measure for the noisy system differs from the deterministic case the Lyapunov exponents will also differ. For the noisy system  $\lambda_1 = .408 \pm .001$ .

## 5.2 TUNING PARAMETERS OF THE MAP ESTIMATES

Local thin plate spline estimates were based on 20 nearest neighbors for  $d=[1,2,3]$  and 50 neighbors for  $d=[4,5]$ . Since no noise is present  $\mu=0$  and thus  $\hat{f}(X_{t-1})$  is equal to  $x_t$ . Due to computational efficiency  $\hat{\sigma}^2$  was computed using within sample cross-validation. Let  $\bar{x}_t$  denote the prediction of  $x_t$  based on the time series where  $x_t$  has been omitted. In this situation

$$\hat{\sigma}^2 = 1/(N-d-1) \sum_{k=d+1}^N (x_t - \bar{x}_t)^2.$$

For the radial basis estimates the value of  $m$  was varied as a function of the number of lags:  $m \in [d, d+3]$ . The estimate reported in Table 5.1 correspond to the value of  $m$  that minimized  $\hat{\sigma}^2$  for a given  $d$ .

The neural net estimates were computed using  $K=7$ . As is typical in nonlinear least squares estimation problems the minimization is not always an automatic procedure. Some outside intervention is required in the fitting procedure to eliminate spurious solutions

associated with local minima.

The projection pursuit estimates depend on the number of ridge functions,  $K$ , and the smoothing parameter,  $\mu$ . For the Henon map the estimate was computed for the 30 combinations of  $K=[2,3, \dots,6]$  and  $\log(\mu)=[-6, -10, -14, -18]$ . For the Rossler system the estimate was computed for the 24 combinations of  $K=[2,3, \dots,9]$  and  $\log(\mu)=[-2, -4, -6]$ . The map estimate was found using 20 backfits. The first time through the backfitting algorithm (4.6) was minimized over each projection using a coarse search of several thousand points on the  $d$  dimensional unit sphere followed by a simplex search with 200 iterations. Subsequent iterations of the backfitting algorithm only employed the simplex search.

### 5.3 SUMMARY OF RESULTS FOR DETERMINISTIC SYSTEMS

For these two chaotic systems the neural net estimator tends to yield estimates within 5% of the unknown exponent. Besides giving accurate estimates of  $\lambda_1$  the sequence of  $\hat{\lambda}_1$  as a function of the embedding dimension follows the characteristic pattern suggested by Casdagli (1989). After a certain point the estimates stabilize and remain unaffected by increasing the embedding dimension. The local thin plate splines also work well provide that one uses the value for  $d$  that minimizes the expected value of  $\hat{\sigma}^2$ . Unlike the neural net estimates, increasing the embedding dimension degrades the accuracy of  $\hat{\lambda}_1$ . Finally it should be noted that in comparison these two methods the projection pursuit and radial basis functions yield poor estimates of  $\lambda_1$  for the Rossler system. The radial basis functions gave the only case where the best estimate of the exponent did not coincide with the embedding dimension found by minimizing  $\hat{\sigma}^2$ .

### 5.4 SUMMARY OF RESULTS FOR THE NOISY HENON SYSTEM

The distribution of Lyapunov exponent estimates for a noisy Henon system are summarized by Table 5.2. Figure 3 gives a graphical display of the simulation results using boxplots. As a reference, estimates results based on the true Jacobians are also reported. For all the cases except local splines with  $d=2$  and  $M=100$ ,  $\hat{\lambda}$  estimates  $\lambda_M$  well. The estimates have a small bias relative to the standard deviation and the variability of these estimates is comparable to the estimates when the Jacobian is known. For this particular noisy system there is a large block bias relative to the the variability of  $\hat{\lambda}_M$ . Therefore while the nonparametric regression methods yield accurate estimates of  $\lambda_M$  for small  $M$  they differ from the actual Lyapunov exponent. As  $M$  increases the block bias decreases and we see that a block size on the order of the length of the series gives a nearly unbiased estimate of  $\lambda$ .

## 6 DISCUSSION

The simulation study of the estimated exponent yielded some expected results connected with estimating high dimensional surfaces and also some surprising results on the accuracy of neural net approximations.

The poor performance of the thin plate splines and the radial basis functions for increasing  $d$  is to be expected. This phenomenon, known as the “curse of dimensionality”, is related to the dramatic effect increasing dimension has on a convergence rate for nonparametric regression curve estimates. One symptom of this problem is the exponential increase in the number of monomials of a fixed degree as  $d$  increases. Intuitively the difficulty in recovering a high dimensional surface increases rapidly as a function of the dimension and this increase easily overwhelms the modest increases in sample sizes usually encountered in practice.

One way to avoid the problem of estimating an arbitrary high dimensional surface is to restrict attention to a subspace. This strategy is successfully employed by the neural net approximation. With little tuning and a relatively small number of parameters ( $7(d+2)$ ) the neural net functional form appears to nearly interpolate  $f$ . This might be expected for the Henon system where  $f$  is a simple polynomial but is surprising for the Rossler system. The question posed by these striking results is how well such a basis spans the space of chaotic maps typically encountered in the study of dynamic systems.

Although the projection pursuit approximation is similar to the neural net method it did not inherit the same accuracy for the Rossler system. There are several possible explanations for this difference in performance. Due to the iterative nature of the backfitting algorithm, projection pursuit does not optimize the objection function (4.6) simultaneously over all the projection vectors. This feature may reduce the flexibility of the approximation. Another problem is that the smoothing step by its very nature will always reduce the accuracy. For example, if the true ridge functions actually had the form of cubic smoothing splines, the projection pursuit estimate would still not be able to reproduce the ridge function exactly. One modification that would address this deficiency is to use one smoothing parameter to estimate the projections and another to determine the ridge functions once the best projections have been found. Finally, due to the inefficiency of the backfitting algorithm and the simplex optimization (and possible local minima in the objective function) it is possible that the backfitting algorithm was ended before a global minimum was reached.

The simulation results for a noisy system while very limited have promise. Both methods yield accurate estimates of  $\lambda$  provided that the block size is chosen correctly ( $M=N=2000$ ). Unlike the results for the deterministic Henon system, the accuracy of the local spline estimates are not sensitive to embedding dimensions larger than 2. This stability may be due to the smoothing of the observed data rather than interpolation. It is interesting to relate the results for a noisy system to the discussion of block bias in Section 3. Although we must have  $M \rightarrow \infty$  at a slower rate than  $N$  to obtain consistency, the simulation results indicate that the best block size is equal to  $N$ . This apparent contradiction can be resolved by noting that the estimation error associated with the Jacobians is much smaller than the bias due to small block sizes. This fits with our conjecture that the block bias will tend to be the dominant term in the total error of the estimate. Thus, in this particular case the best estimate is obtained by taking  $M$  as large as possible. There is no guarantee, however, that this is a good strategy for other noisy systems.

The numerical results suggests that the embedding dimension found by minimizing the expected prediction error is also good for exponent estimates. Note that the results are given with respect to the *expected* prediction error rather than the single estimate,  $\hat{\sigma}^2$ , found from a particular sample. One aspect that needs further study are better data-based estimates of  $\hat{\sigma}^2$ . The prediction error for the local spline was found from *within sample* using cross-validation. Such methods are known to yield nearly unbiased estimates but may have a large variance. For the estimates of  $\lambda$  not based on local splines,  $\hat{\sigma}^2$  was calculated *out of sample* using a large set of measurements independent from the data. This approach is not feasible if the amount of data is limited and within sample estimates need to be studied.

In conclusion we have demonstrated the potential of nonparametric regression estimates to extract an accurate estimate of the Lyapunov exponent. Although much work remains, we have identified the analysis of chaotic dynamical systems as a well-posed statistical problem.

APPENDIX

Proof of Theorem 3.1

$$\text{Let } U_t = \hat{J}_t - J_t, \quad \alpha = \sup_{k \geq 1} \|J_k\| \text{ and } \beta = \sup_{k \geq 1} \|\hat{J}_k - J_k\|.$$

Then

$$(A.1) \quad \hat{T}_M - T_M = (J_M + U_M)(J_{M-1} + U_{M-1}) \cdots (J_1 + U_1) - J_M J_{M-1} \cdots J_1.$$

To simplify the product, we need to index the possible combinations of terms involving  $J$  and  $U$ . Let  $\mathfrak{J}$  denote the set of all  $2^M$   $M$ -tuple  $\omega = (\omega_1, \omega_2, \dots, \omega_M)$  where each  $\omega_k = 0$  or  $1$ . Let  $A_k(\omega) = \omega_k J_k + (1 - \omega_k) U_k$ , and  $|\omega| = \sum_{k=1}^M \omega_k$ . Then (A.1) may be rewritten as

$$(A.2) \quad \hat{T}_M - T_M = \sum_{\omega \in \mathfrak{J}, |\omega| < M} A_M(\omega) A_{M-1}(\omega) \cdots A_1(\omega).$$

Hence, using the assumed bounds on  $\|J_k\|$  and  $\|U_k\|$ ,

$$\begin{aligned} \|\hat{T}_M - T_M\| &\leq \sum_{\omega \in \mathfrak{J}, |\omega| < M} \alpha^{|\omega|} \beta^{(M-|\omega|)} \\ &= \sum_{k=0}^{M-1} \binom{M}{k} \alpha^k \beta^{M-k} = (\alpha + \beta)^M - \alpha^M = \alpha^M [(1 + (\beta/\alpha))^M - 1]. \\ (A.3) \quad &= O_p(M \alpha^{M-1} \beta_N). \end{aligned}$$

To show (3.1) let  $\lambda_M = \log(\gamma_M)$  and  $\hat{\lambda}_M = \log(\hat{\gamma}_M)$ . By the elementary inequality  $|a^{1/M} - b^{1/M}| \leq |a - b|^{1/M}$  and (A.3) it follows that

$$|\hat{\gamma}_M - \gamma_M| \leq \|\hat{T}_M - T_M\|^{1/M} = O_p(\beta_N^{1/M}).$$

Now by the mean value theorem,

$$|\hat{\lambda}_M - \lambda_M| \leq (1/\gamma^*) |\hat{\gamma}_M - \gamma_M|$$

where  $\gamma^*$  lies between  $\hat{\gamma}_M$  and  $\gamma_M$ . With this bound, (3.1) now follows.

Now (3.2) will be considered. From the hypotheses on  $\beta_N$  and (A.3),  $\|\hat{T}_M - T_M\| = o_p(\rho^{M-1})$ . It follows from the Furstenburg-Kesten theorem that

$$\|T_M\| \rho^{-M} \rightarrow +\infty \text{ with probability 1,}$$

From this fact it follows that  $\|\hat{T}_M - T_M\| / \|T_M\| \rightarrow 0$  in probability and by the triangle inequality  $\|\hat{T}_M\| / \|T_M\| \xrightarrow{P} 1$ . Using the relation  $|\log(b) - \log(a)| \leq |b - a| / \min(b, a)$ ,

$$(A.4) \quad M |\hat{\lambda}_M - \lambda_M| = |\log \|\hat{T}_M\| - \log \|T_M\|| \leq \left| \frac{\|\hat{T}_M\| - \|T_M\|}{\|\hat{T}_M\| X_M} \right|$$

where  $X_M = \min(\|\hat{T}_M\| / \|T_M\|, 1)$ . From the remarks above it follows that  $X_M \xrightarrow{P} 1$  and thus

$$|\hat{\lambda}_M - \lambda_M| \leq \frac{\|\hat{T}_M - T_M\|}{M \|T_M\|} (1 + o_P(1)) \text{ as } M \rightarrow \infty.$$

The result now follows immediately from (A.2) and (A.4).  $\square$

Figure 1. Comparison of the conjectured convergence rate for  $\hat{\lambda}_M$  with simulations of the Henon map (3.5). The graph shows equation (3.3) (dashed line), and the mean (x), 95<sup>th</sup> percentile (box), and maximum (triangle) of the absolute error  $|\hat{\lambda}_M - \lambda_M|$  over 1000 simulations for each  $(M, \beta)$  pair. The block length was  $M = 5/\sqrt{\beta}$ , so that  $\beta \rightarrow 0$  corresponds to  $N, M \rightarrow \infty$  with  $M\beta_N = 5\sqrt{\beta} \rightarrow 0$ . For each replicate, the initial condition  $(x_0, x_1)$  was chosen at random from a file of 20,000 values on the Henon map's attractor. Error matrices  $U_t$  were chosen independently for each replicate at each  $(M, \beta)$  value. See text for further details.

Figure 2. Rates of convergence of the block bias and block error for the Henon map (3.8). Each replicate consists of 10000 iterations. For each block length  $M$  one value of  $\lambda_M$  was computed from the first  $M$  Jacobians. (a) Average  $\lambda_M$  vs.  $M$ ; the solid line was fit by linear regression of average  $\lambda_M$  on  $1/M$ . The actual exponent in this case is .419. (b) Standard deviation of  $\lambda_M$  vs.  $M$ ; the solid line was fit by linear regression of the standard deviation on  $1/\sqrt{M}$ . For both regressions  $R^2 > .99$ .

Figure 3. Distribution of Lyapunov exponent estimates from the noisy Henon system. Boxplots summarize the distribution of estimates as a function of method ( Local Spline, Neural Net), the embedding dimension ( 2,5) and the block length ( $M$ ). The horizontal line locates the true value of  $\lambda_1$  for this system. The boxplots are based on 14 observations for the local splines and 16 for the neural nets. The four arrows in the upper right plot indicate four points out of bounds.

## REFERENCES

- Abarbanel, H.D.I., R. Brown and J.B. Kadtko (1989), "Prediction and system identification in chaotic nonlinear systems: time series with broadband spectra," *Physics Letters A*, 138(8), 401-408.
- Albano, A.M., N.B. Abraham, G.C. deGuzman, M.F.H. Tarroja, D.K. Bandy, R.S. Goggia, P.E. Rapp, I.D. Zimmerman, N.N. Greenbaum, and T.R. Bashore (1986), "Lasers and brains: complex systems with low-dimensional attractors" in G. Mayer-Kress (ed.) *Dimensions and Entropies in Chaotic Systems*, Springer-Verlag, pp. 231-240.
- Arnold, L. and V. Wihstutz (1986), "Lyapunov exponents: a survey," in L. Arnold and V. Wihstutz (eds), *Lyapunov Exponents* (Lecture Notes in Mathematics 1186), Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, pp. 1-26.
- Athreya, K.B. and S.B. Pantula (1986), "Mixing properties of Harris chains and autoregressive processes," *Journal of Applied Probability*, 23, 880-892.
- Babloyantz, A. and A. Destexhe (1986), "Low-dimensional chaos in an instance of epilepsy," in *Proceedings of the National Academy of Science*, 53, 3513-3517.
- Bougerol, P., J. Lacroix (1986) "Products of random matrices with application to Schrodinger operators," Birkhaeuser-Verlag, Basel, MA, 283.
- Brandstater, A., J. Swift, H.L. Swinney, A. Wolf, J.D. Farmer, E. Jen, and P.J. Crutchfield (1983), "Low-dimensional chaos in a hydrodynamic system," *Physical Review Letters*, 51(6), 1442-1445.
- Brandstater, A. and H.L. Swinney (1987), "Strange attractors in weakly turbulent Couette-Taylor flow," *Physical Review A*, 35(5), 2207-2220.
- Brock, W.A. and C.L. Sayers (1988), "Is the business cycle characterized by deterministic chaos?," *Journal of Monetary Economics*, 22, 71-90.
- Buchler, J.R. and H. Eichorn (eds.) (1987), "Chaotic Phenomena in Astrophysics," *Annals of the New York Academy of Sciences*, (Vol. 497), New York Academy of Sciences, New York. New York.
- Buja, A., T. Hastie, and R. Tibshirani. (1989), "Linear smoothers and additive models." *Annals of Statistics*, 17, 453-510.
- Casdagli, M. (1989), "Nonlinear prediction of chaotic time series," *Physica*, 35D, 335-356.
- Cohen, J.E., H. Kesten and C.M. Newman (eds.) (1986), "Random Matrices and Their Applications," *Contemporary Mathematics*, (Vol. 50), American Mathematical Society, Providence R.I.
- Cox, D. D. (1984), "Multivariate smoothing splines," *SIAM Journal of Numerical Analysis*, 21, 789-813.
- 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 (1987), "Predicting chaotic time series," *Physical Review Letters*, 59, 845-848.

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

Farmer, J.D. and J.J. Sidorowich (1988b), "Predicting chaotic dynamics," in J.A.S. Kelso, A.J. Mandell, and M.F. Schlesinger (eds.) *Dynamic Patterns in Complex Systems*, World Scientific, Singapore, pp. 265-292.

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

Furstenberg, H. and H. Kesten (1960), "Products of random matrices," *Annals of Mathematical Statistics*, 31, 457-469.

Gallant, A.R. and H. White (1989), "On learning the derivatives of an unknown mapping with multilayer feedforward networks," (preprint).

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

Guckenheimer, J. (1985), "Noise in chaotic systems," *Nature*, 298, 358-361.

Guckenheimer, J. and G. Buzyna (1983), "Dimension measurements for geostrophic turbulence," *Physical Review Letters*, 51(6), 1438-1441.

Hall, P. and C.C. Heyde (1980), "Martingale Limit Theory and its Application," Academic Press, New York, p. 308.

Heyde, C.C. (1985), "An asymptotic representation for products of random matrices," *Stochastic Processes and their Applications*, 20, 307-314.

Heyde, C.C. and J.E. Cohen (1985), "Confidence intervals for demographic projections based on products of random matrices," *Theoretical Population Biology*, 27, 120-153.

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

Kifer, Y. (1986), *Ergodic Theory of Random Transformations*, Birkhauser, Boston, p. 210.

Kot, M., W.M. Schaffer, G.L. Truty, D.J. Graser, and L.F. Olsen (1988), "Changing criteria for imposing order," *Ecological Modelling*, 43, 75-110.

Kurths, J. and H. Herzel (1987), "An attractor in a solar time series," *Physica*, 25D, 165-172.

May, R.M. (1987), "Nonlinearities and complex behavior in simple ecological and

- epidemiological models," in S.H. Koslow, A.J. Mandell, and M.F. Schlesinger (eds.) *Perspectives in Biological Dynamics and Theoretical Medicine, Annals of the New York Academy of Sciences* (Vol. 504), New York Academy of Sciences, New York, New York, pp. 1-15.
- Mayer-Kress, G. (ed.) (1986), "Dimensions and Entropies in Chaotic Systems," Springer-Verlag, Berlin.
- Moon, F.C. (1987), "Chaotic Vibrations: An Introduction for Applied Scientists and Engineers," John Wiley & Sons, New York.
- Mpitsos, G.J., H.C. Creech, C.S. Cohan, and M. Mendelson (1988), "Variability and chaos: neurointegrative principles in self-organization of motor patterns," in J.A.S. Kelso, A.J. Mandell, and M.F. Schlesinger (eds.) *Dynamic Patterns in Complex Systems*, World Scientific, Singapore, pp. 162-190.
- Mpitsos, G.J., R.M. Burton Jr., H.C. Creech, and S.O. Soynila (1988), "Evidence for chaos in spike trains of neurons that generate rhythmic motor patterns," *Brain Research Bulletin*, 21, 529-538.
- Nychka, D. W. (1989), "Nonparametric regression and spatial data: some experiences collaborating with biologists," in E. Wegman, D. Grant and J. Miller (ed.) *Proceedings of the 20<sup>th</sup> Symposium on the Interface*, American Statistical Association, Alexandria, VA 731-736.
- Olsen, L.F. and H. Degn (1985), "Chaos in biological systems," *Quarterly Review of Biophysics*, 18, 165-225.
- Ramsay, J.B. and H.-J. Yuan (1989), "Bias and error bars in dimension calculations and their evaluation in some simple models," *Physics Letters A*, 134(5), 287-297.
- Ruelle, D. (1990), "Deterministic Chaos: the science and the fiction," *Proceedings of the Royal Society A*, London, 427, 241-248.
- Schuster, H.G. (1988), "Deterministic Chaos: An Introduction," (2nd revised edition), VCH Verlagsgesellschaft, Weinheim FRG
- Sugihara, G. and R.M. May (1990), "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series," *Nature*, 344, 734-741.
- Takens, F. (1981), "Detecting strange attractors in turbulence," in D. Rand and L.S. Young (ed.) *Dynamical Systems and Turbulence*, Warwick 1980, Lecture Notes in Mathematics, 898, Springer-Verlag, New York, 366-381.
- Vastano, J.E. and E.J. Kostelich (1986), "Comparison of algorithms for determining Lyapunov exponents from experimental data," in G. Mayer-Kress (ed.) *Dimensions and Entropies in chaotic systems*, Springer-Verlag, New York, pp. 100-107.
- Wahba, G. (1989), "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.

Table 5.1 Estimated Lyapunov Exponents

		<u>Henon System</u>				
<u>Map Estimate</u>	<u>N</u>	d				
		1	2	3	4	5
Local spline <sup>1</sup>	2500	5.7602 (.042)	.4188* (.005)	.0750 (.011)	-.0251 (.013)	.0259 (.013)
Neural Net	2000	.1147	.4106	.4227	.4236*	-
Projection Pursuit	2000	-	.4163	.4058	.4026*	-

		<u>Rossler System</u>						
<u>Map Estimate</u>	<u>N</u>	d						
		1	2	3	4	5	6	7
Local spline <sup>2</sup>	2500	7.1229 (.055)	.0992 (.004)	.0461* (.002)	1.7099 (.011)	1.567 (.021)	-	-
Radial Basis	2000	-	.0629	.7778*	10.24	10.26	-	-
Neural Net	2000	-	.0010	.1272	.6940	.0482	.0414	.0466*
Projection Pursuit	2000	-	-	.0966*	.0146	-.2792	-.0640	-

<sup>1</sup> Average of 5 estimates with standard deviation.

<sup>2</sup> Average of 10 estimates with standard deviation.

Table 5.2 Estimated Lyapunov exponents for the Henon map with noise as a function of block size<sup>1</sup>.

<u>Map Estimate</u>	<u>d</u>	<u>M</u>				
		50	100	500	2000	20000
Local spline <sup>2</sup>	2	.421 (.015)	.426 (.014)	.417 (.015)	.416 (.015)	
Neural Net <sup>3</sup>	2	.416 (.020)	.412 (.020)	.408 (.020)	.408 (.019)	
Exact Map <sup>4</sup>	2	.415 (.010)	.414 (.009)	.409 (.009)	.408 (.009)	.408 (.009)

<u>Map Estimate</u>	<u>d</u>	<u>M</u>			
		50	100	500	2000
Local spline	5	.417 (.009)	.431 (.009)	.406 (.008)	.404 (.008)
Neural Net	5	.417 (.007)	.411 (.009)	.406 (.009)	.405 (.009)

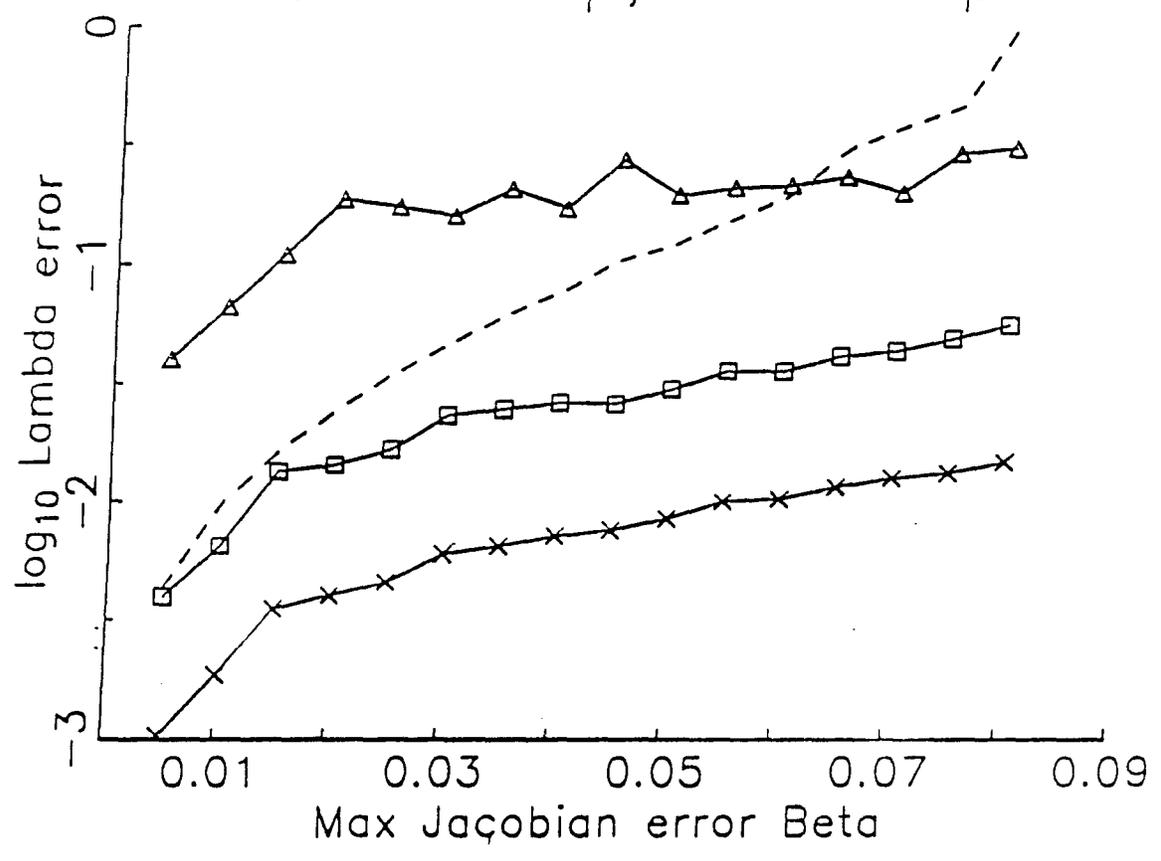
<sup>1</sup> Each Lyapunov exponent estimate is the average of the exponents obtained from N/M disjoint blocks of the data series.

<sup>2</sup> Average of 14 estimates with standard deviation.

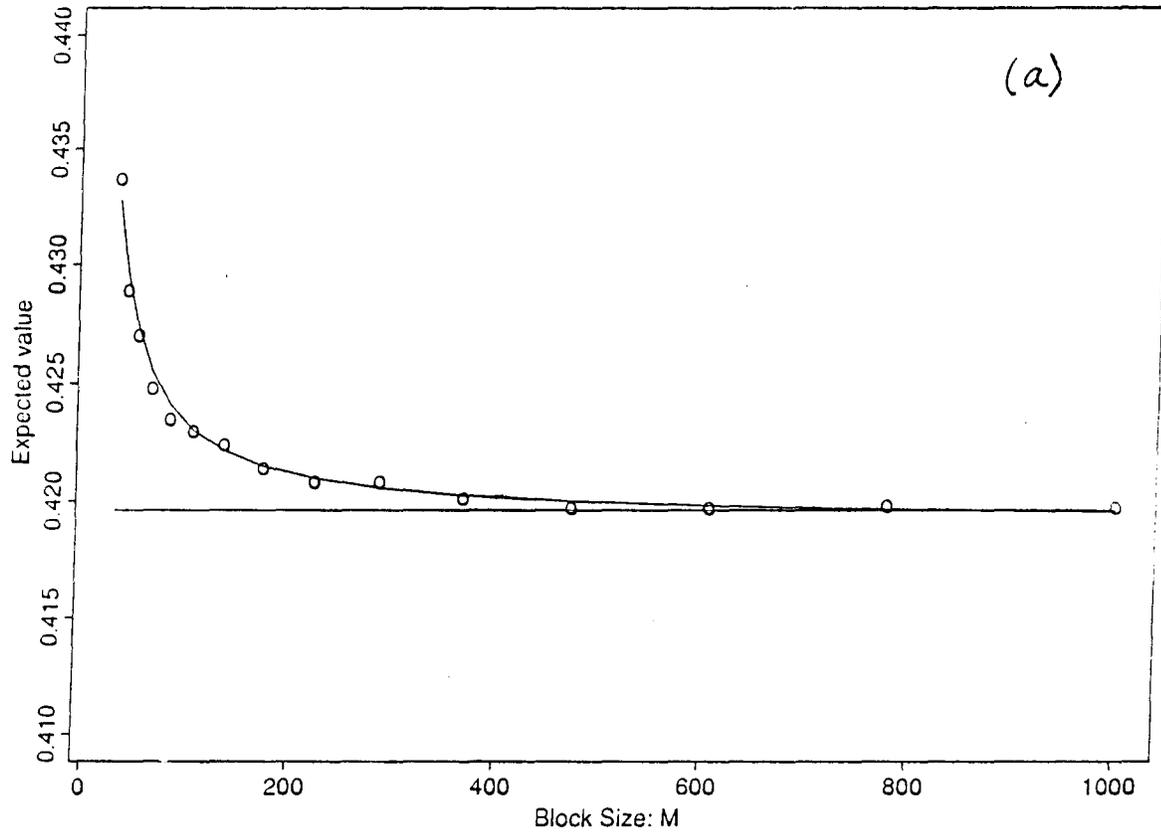
<sup>3</sup> Average of 16 estimates with standard deviation.

<sup>4</sup> Average of 200 estimates using the true Jacobian matrix. The standard deviation has been adjusted to be comparable with the other estimates. ( reported S.D.=sample S.D./ $\sqrt{B}$  where  $B=N/M$ .)

Henon map, 1000 reps



Expected Value of the Lyapunov exponent based on a finite block size



Variability of the Lyapunov exponent based on a finite block size

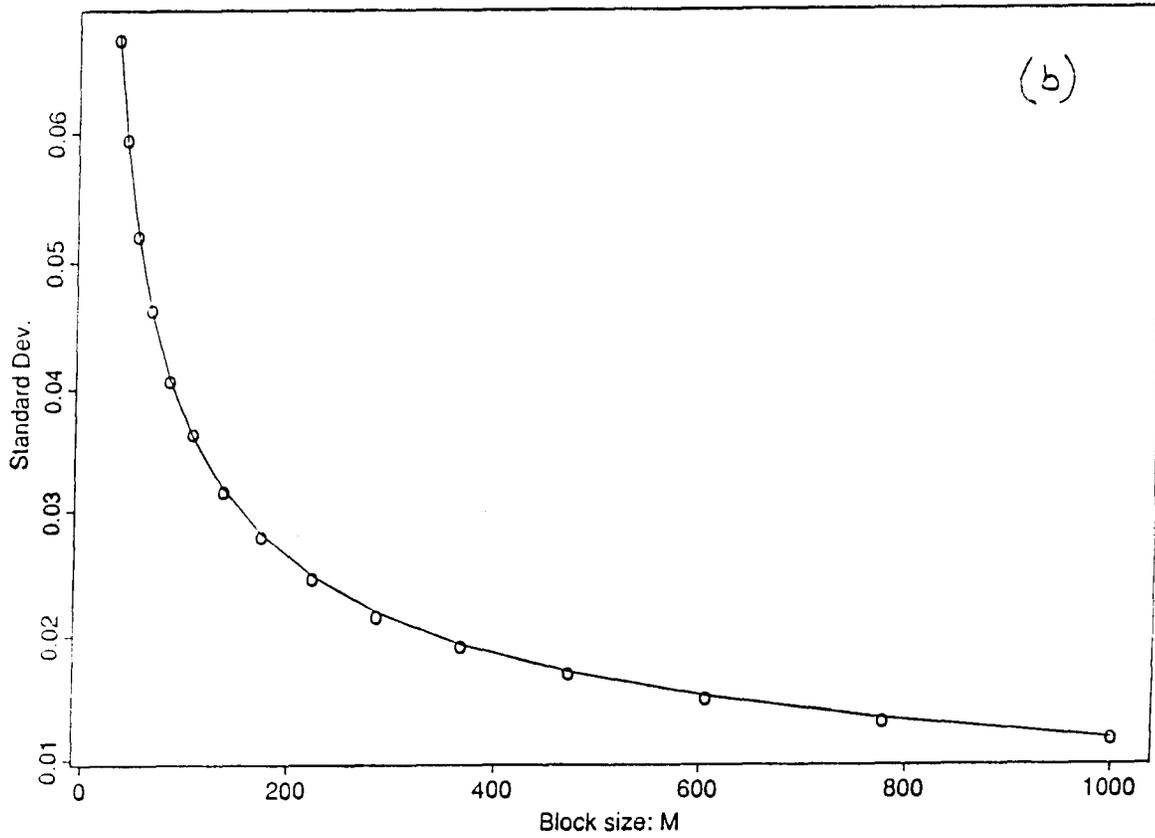


Figure 3

